Syst. Biol. 42(2):126–141, 1993

DISTRIBUTIONS OF TREE COMPARISON METRICS—SOME NEW RESULTS

MIKE A. STEEL[1] AND DAVID PENNY[2]

[1]*Zentrum für Interdisziplinäre Forschung der Universität Bielefeld,*
Wellenberg 1, Bielefeld 1, D-4800, Germany
[2]*Molecular Genetics Unit, Massey University, Palmerston North, New Zealand*

Abstract.—Measures of dissimilarity (metrics) for comparing trees are important tools in the quantitative analysis of evolutionary trees, but many of their properties are incompletely known. The present paper reports formulae for the distributions of three classes of tree comparison metrics: the partition (or symmetric difference) metric, the quartet metric (which compares subsets of four taxa), and a metric based on path-length differences between pairs of taxa. The properties studied include the mean and variance for several underlying distributions of trees, the range, the effect of the number of taxa, and methods of calculation. Three basic theorems and their proofs are reported, one for each class of tree comparison metric. The partition metric generates an asymptotic Poisson distribution for most distributions of trees (its mean is given for three tree distributions). Exact expressions are derived for the variance of the quartet metric and the mean square value of a metric based on path differences. Factors that affect the choice of a metric for a particular study include the degree of similarity of the trees being compared and the type of hypothesis being tested (e.g., whether the trees estimate the same underlying phylogeny or are simply related in some, perhaps unknown, way). [Evolutionary trees; tree comparison metrics; quartet metric; partition metric; path-length differences.]

In classification, it is useful to have a metric (a measure of dissimilarity) for comparing the branching structure of phylogenetic trees. For example, such a metric allows a quantitative assessment of the similarity of trees reconstructed from different data sets. The distribution of the metric then provides an indication as to whether or not this measured similarity could have come about by chance. Without such an objective measure, one must rely on intuitive notions of whether trees are similar of not, and this has led to disagreements and controversies (see, for example, O'Grady et al., 1989, in reply to Cavalli-Sforza et al., 1988). Partly, this is because there is no "obvious" or "natural" way to measure the distance between two trees, unlike the comparison of two numbers (where one subtracts the smaller number from the larger) or points in space (where Euclidean distance can be applied).

A number of tree comparison metrics have been proposed (Swofford, 1991). Examples include the partition (or symmetric difference) metric (Bourque, 1978; Robinson and Foulds, 1979), the quartet metric (Estabrook et al., 1985; Day, 1986), the nearest neighborhood interchange metric (Waterman and Smith, 1978), and those metrics based on differences in the lengths of paths between pairs of taxa (Williams and Clifford, 1971). To interpret the significance of a measured value of one of these metrics on a pair of phylogenetic trees, it is useful to know the distribution of that metric on pairs of trees generated by some random process. We summarize a number of new and known results, extending previous studies such as those of Hendy et al. (1984), Day (1983, 1986), and Steel (1988).

In particular, we determine properties of the distribution of tree metrics that are invariant with respect to the underlying tree distribution. As examples of this invariance, we show that (under mild assumptions) the distribution of the partition metric is always described asymptotically by a Poisson distribution; we give its mean for three tree distributions. For the quartet metric, the mean of its distribution on pairs of binary trees is independent of the underlying distribution on tree topologies. We give an exact expression for the variance of this metric when all binary trees are equally probable. We also consider a third tree metric, which is based on the differences in the lengths of paths joining pairs of taxa, and derive an exact expression for its mean square value when all binary trees are equally probable. We do not consider the nearest neighbor interchange metric here, mainly because there is no known efficient method for its calculation but also because little is known analytically about its distribution (although Day [1983] obtained some limited information by simulation).

These results, along with simulations, allow us to make some broad conclusions about the distributions of these three metrics. Factors that affect the choice of a metric for a particular study are discussed briefly. These factors include the number of taxa, the underlying distribution of trees, whether the trees are highly similar or dissimilar, the number of trees being compared, whether the comparison of trees or the relative positions of taxa are of primary interest, and whether the trees are weighted or unweighted.

The categories under which results, such as those described here, fall are as follows.

1. Exact or explicit formulae.—These allow a direct calculation by substituting any value of a variable, *n*, into an explicit formula.
2. Recursive formulae.—These allow an indirect calculation by applying an algorithm to a given starting value.
3. Asymptotic formulae.—These give the limiting values as *n* becomes large. They are thus approximations, which become increasingly accurate as *n* grows.
4. Enumeration results.—These are found by an exhaustive search. They are usually only feasible for small values of *n* and so complement asymptotic formulae.
5. Simulation results.—These are useful in several ways, including how to estimate how large a value of *n* is required before an asymptotic formula is a good approximation.

For the three tree metrics discussed here, it is possible to either consider the metric directly or to normalize it by dividing by an appropriate scaling factor, which depends on the number of taxa. The most natural scaling factor is the diameter of the metric, i.e., the largest value the metric takes when applied to all pairs of trees. However, this factor is known only for the partition metric, for which the diameter is $2n - 6$, where *n* is the number of taxa. For the quartet metric, the diameter is not known in general, but a natural scaling factor is the total number of quartets (so that the range of the normalized distribution always lies between 0 and some number <1). For the path metric, the choice of an appropriate normalizing factors is less obvious; it should involve the number of pairs of taxa, but it may also include a further term relating to the length of the longest path possible in a tree. For all three tree metrics, normalization leads to a further distinction in the results presented here, depending on whether they relate to the metric or to its normalized value.

Basic Terminology

Graph theoretical terminology is useful for describing aspects of phylogenetic trees; definitions are illustrated in Figure 1 (for further graph theoretical terminology, see Bondy and Murty, 1976). A graph consists of vertices (or points, nodes) that are joined by edges (or lines). A tree is a graph that is connected but contains no cycles. The degree of a vertex is the number of edges that are incident with it; a leaf (pendent vertex) is a vertex of degree one. Edges that are not incident with a leaf are said to be internal. Vertices in a tree may be either labeled or unlabeled.

We consider trees that contain no unlabeled vertices of degree one or two and whose leaves are labeled in a one-to-one fashion by the taxa, which we number 1, . . . , *n*. These trees are often called phylogenetic trees; here we simply call them trees. If, in addition, each nonleaf vertex has degree three, the tree is said to be binary, or fully resolved. Trees that are equivalent except for the labeling of their leaves are said to have the same topology, which we designate by τ. An important

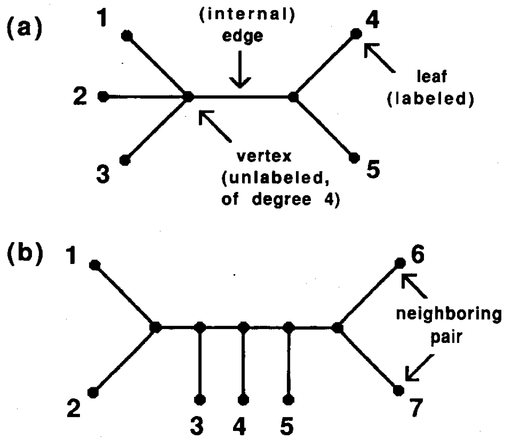


FIGURE 1. (a) A (nonbinary) tree consisting of six edges (one internal) and seven vertices, comprising five leaves (vertices of degree one) labeled 1–5 and two unlabeled vertices (of degree three and four). (b) A (binary) caterpillar tree with seven labeled leaves and two neighboring pairs (1 and 2, 6 and 7).

aspect of τ is the number of neighboring pairs it contains; these are leaves adjacent to a common vertex. A binary tree with just two neighboring pairs (as in Fig. 1b) is called a caterpillar tree.

Let $b(n)$ and $p(n)$ be the number of binary trees and all trees, respectively, with n labeled leaves. Both these numbers, in a different context, were studied during the last century (Schröder, 1870), although Cavalli-Sforza and Edwards (1967) independently derived and popularized among taxonomists the following well-known expression for $b(n)$:

$$\begin{aligned} b(n) &= (2n - 5)!! \\ &= (2n - 5) \times (2n - 7) \\ &\quad \times \cdots \times 5 \times 3. \end{aligned}$$

No simple closed formula for $p(n)$ is known, although values can be calculated recursively (Felsenstein, 1978) by keeping track of the number $p(n, f)$ of trees with n leaves and with exactly f internal edges. These numbers are determined by $p(3, 0) = 1$, $p(3, f) = 0$ for $f > 0$ and by the recursion

$$\begin{aligned} p(n, f) = (n + f - 2)&p(n - 1, f - 1) \\ + (f + 1)&p(n - 1, f). \end{aligned}$$

The value for $p(n)$ can then be obtained by summing $p(n, f)$ over all values of f for which $\mathrm{p}(n, f) \neq 0$, namely $f = 0, \ldots, n - 3$. Foulds and Robinson (1984) provided a simple asymptotic estimate of $p(n)$, and an expression for $p(n)$ as a sum of terms, with alternating signs, that involved Stirling numbers was given by Steel (1990).

Distributions

The distribution of a metric for pairs of (unrooted) trees presupposes and depends upon an underlying probability distribution of the trees themselves. We denote any such distribution by D, and the probability of a tree T arising under D is written $P_D(T)$. We consider four distributions. The first three can be realized by Markovian processes in which a tree is built up randomly by attaching one new leaf at a time, although the details differ for each distribution.

D_{bin}.—Each binary tree has equal probability;

$$P_D(T) = \begin{cases} \dfrac{1}{b(n)}, & \text{if } T \text{ is a binary tree} \\ 0, & \text{otherwise.} \end{cases}$$

This model applies, for example, if a tree is randomly built up by starting with a subtree with just three leaves and recursively selecting a random edge of the current subtree and making the next leaf adjacent to the midpoint of that edge. The resulting distribution is the same whether random labeling or a predetermined labeling (for example, 1, 2, 3, . . .) of the first three and subsequently added leaves is used. An alternative way to obtain this distribution is to randomly generate numbers between 1 and $b(n)$ and apply a one-to-one correspondence between binary trees and the numbers $1, \ldots, b(n)$, such as that described by Rohlf (1983). This model leads asymptotically to an average of 50% of leaves belonging to a member of a neighboring pair.

D_{tip}.—This model was described by Harding (1971) and used by Simberloff (1987) and Slowinski (1990) to stochastically generate rooted binary trees. We can then regard these as unrooted binary trees by suppressing the root. This model also applies if we follow the same procedure described for D_{bin} but restrict our choice for the placement of a new leaf to those edges that are already incident with a leaf, using a random labeling at each step (the distribution differs from D_{tip} if a predetermined labeling scheme is used). This model tends to favor "bushy" trees over "linear" trees more than does D_{bin}; e.g., the asymptotic average proportion of leaves that belong to a neighboring pair under D_{tip} is 66.7%.

D_{all}.—Each tree (binary or nonbinary) has equal probability;

$$P_D(T) = \frac{1}{p(n)},$$

for each tree T. This model arises from a random recursive procedure, similar to that of the previous two models, by allowing for the attachment of a new leaf to a nonleaf vertex. Some care is required however to ensure that the resulting distribution is of the type claimed because the possible trees at any stage of a recursive construction have different numbers of places where a new leaf can be attached. This problem was solved by Oden and Shao (1984), who described an efficient algorithm to randomly generate trees according to D_{all}. This model is compared with the previous two by illustrating a method for generation of these distributions (Fig. 2).

D_R.—This distribution of trees arises from the application of some tree reconstruction procedure to genetic sequence data that have been randomized within each column. Such randomized data represent a "big bang" null hypothesis against which an evolutionary hypothesis can be tested (Thompson, 1975). The distribution D_R arises when one wishes to test whether trees constructed from different data sets are significantly more similar than if the tree reconstruction procedure was not extracting any hierarchical information from the sequences (see Penny et al., 1982, for such a study). In general, D_R will differ from a uniform distribution on binary trees, D_{bin}, because most tree reconstruction procedures applied to randomized data probably will favor certain topologies over others.

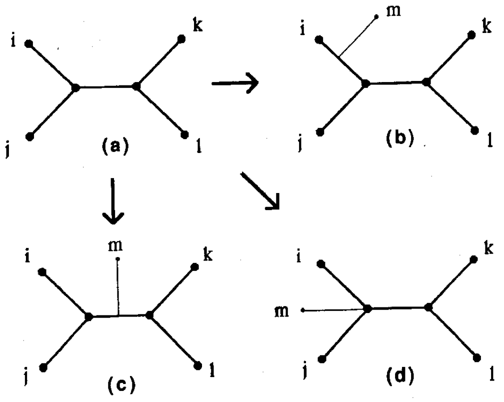


FIGURE 2. The models D_{bin} and D_{tip} arise from adding new leaves to randomly chosen edges of a tree (a). For D_{bin}, leaves may be added to any edge, as in (b) and (c), whereas for D_{tip}, leaves are added only to noninternal edges (b). The model D_{all}, which assigns equal probability to each (phylogenetic) tree, can also be generated by a similar process but by allowing for the attachment of leaves to unlabeled vertices (d).

D_R and the other three distributions all share the desirable property of being invariant under every permutation σ of the taxa $1, \ldots, n$:

$$P_D(T) = P_D(T^\sigma), \tag{1}$$

where T^σ is the tree obtained from T by replacing the leaf labeled i by that labeled $\sigma(i)$. Thus $P_D(T)$ depends only on the topology of T and not on its particular labeling. A distribution that satisfies Equation 1 is label invariant, and only these distributions will be considered here. Similarly, any metric comparing two trees should not distinguish the labels, i.e., for any permutation σ of the taxa,

$$d(T_1, T_2) = d(T_1^\sigma, T_2^\sigma), \tag{2}$$

which holds for all the tree metrics under consideration. Figure 3 illustrates the properties described by Equations 1 and 2.

A tree distribution D naturally induces a distribution on the set of distances (under some given metric) between pairs of trees. Specifically, let $P_D[d = k]$ be the probability that two trees randomly selected (with replacement) according to D are distance k

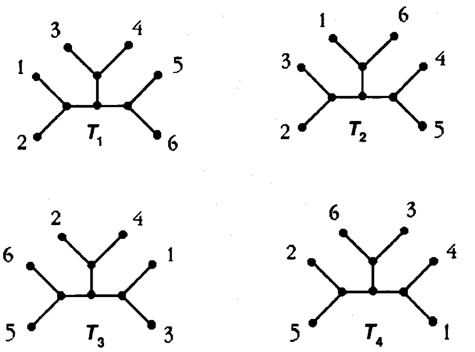


FIGURE 3. Desirable properties of a distribution on trees and of a tree comparison metric. $P_D(T_1) = P_D(T_2)$, the probability of selecting a tree, is independent of the permutation of the labels (taxa) (Equation 1). $d(T_1, T_2) = d(T_3, T_4)$, the distance between a pair of trees, is constant under a permutation of the labels, in this case (1, 2, 3, 4, 5, 6) to (6, 5, 2, 4, 1, 3) (Equation 2).

apart. Thus $P_D[d = k]$ is the sum of $P_D(T)P_D(T')$ over all pairs $\{T, T'\}$ that are distance k apart. We can simplify this summation by invoking Equations 1 and 2. Specifically, for a particular topology, τ, let $p(\tau)$ be the sum of $P(T)$ over all trees with this given topology, and let T_τ denote any labeled tree with this topology. Then,

$$P_D[d = k] = \sum_{\tau} p(\tau) \sum_{\{T':d(T_\tau,T')=k\}} P_D(T'). \quad (3)$$

The mean and variance of this distribution, which are denoted by $\mu_D(d)$ and $\sigma_D{}^2(d)$, respectively, and the range of the distribution are also of interest.

To state our results, some standard mathematical terms are used throughout. The symmetric difference of two sets is the set consisting of all those elements that are in one set but not in the other. Both the partition metric and the quartet metric are examples of symmetric difference metrics. Recall that $f(n) \sim g(n)$ means that $f(n)$ and $g(n)$ are asymptotically equal, i.e., $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 1$; and $\binom{n}{i}$ denotes the binomial coefficient, $\frac{n!}{i!(n-i)!}$, which is the number of ways of selecting a subset of i taxa from a set of n taxa.

We now consider three metrics and present new results for each, including two results that apply to any label-invariant distribution D. Figure 4 illustrates calculations of these three metrics.

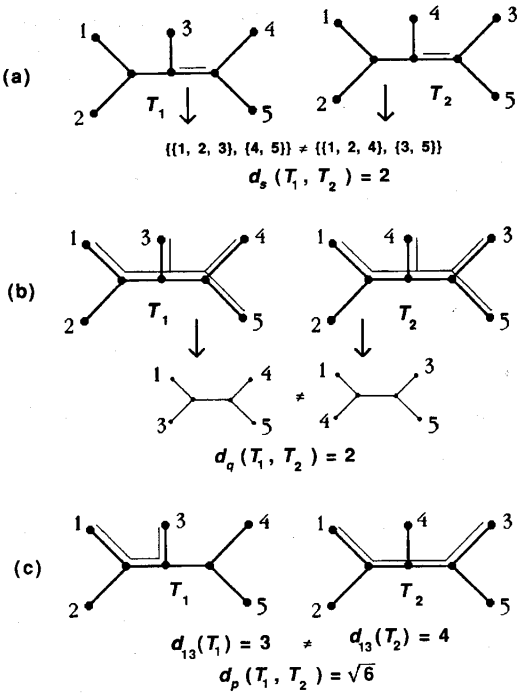


FIGURE 4. (a) The partition metric, d_s, measures the number of splits in one tree but not the other. (b) The quartet metric, d_q, measures the number of quartets that induce different subtrees. (c) The path-length metric, d_p, measures the sum of the squares of the differences in the lengths of paths between pairs of taxa.

PARTITION METRIC (d_s)

When an edge of a tree is deleted, the taxa are partitioned into two sets. This partition is usually called a split or bipartition, and the partition metric, d_s, measures how many splits are in one tree but not the other. More formally, $d_s(T_1, T_2)$ is the size of the symmetric difference of the sets of splits induced by two trees T_1 and T_2. Equivalently,

$$d_s(T_1, T_2) = i(T_1) + i(T_2) - 2v_s(T_1, T_2), \quad (4)$$

where $i(T)$ denotes the number of edges of T that are internal (i.e., not incident with a leaf) and $v_s(T_1, T_2)$ denotes the number of pairs of identical splits of the taxon set induced by deleting an internal edge from each of T_1 and T_2. Equation 4 illustrates that d_s is an example of a valuation metric, which has been studied by Monjardet (1981).

The partition metric was proposed by Bourque (1978) (see also Robinson and Foulds, 1981), and an extension to weighted trees was described by Robinson and Foulds (1979). This metric is easy to calculate, and Day (1985) described a linear-time algorithm. Also, the range of this metric is well known; the maximum value of d_s across all pairs of trees with n leaves is $2n - 6$. Furthermore, the distribution of this metric for up to 16 taxa, and asymptotically, for D_{bin} is known (Hendy et al., 1984, 1988; Steel, 1988). Here, we extend this analysis to other tree distributions.

In view of Equation 4, the distribution of d_s is determined by the distribution of $v_s(T_1, T_2)$ and $i(T)$, but $i(T)$ is trivial if distributions on just binary trees are being considered. Let $\mu_D(I)$ and $\mu_D(v_s)$ denote the expected value of $i(T)$ and $v_s(T_1, T_2)$, respectively. Then, for any label-invariant distribution D,

$$\mu_D(d_s) = 2[\mu_D(I) - \mu_D(v_s)]$$

and

$$\mu_D(v_s) = \sum_{i=2}^{n/2} \binom{n}{i} \times P(i)^2, \tag{5}$$

where $P(i)$ is the probability that a tree randomly generated by D has the split $\{\{1, \ldots, i\}, \{i+1, \ldots, n\}\}$.

For example, if $D = D_{bin}$, the uniform distribution on binary trees, then

$$\begin{aligned}\mu_D(v_s) &= \sum_{i=2}^{n/2} \binom{n}{i} \frac{b^2(i+1)b^2(n-i+1)}{b^2(n)} \\ &= \frac{n(n-1)}{2!(2n-5)^2} \\ &\quad + \frac{n(n-1)(n-2)}{3!(2n-5)^2(2n-7)^2} + \ldots \\ &\to \frac{1}{8}, \qquad \text{as } n \to \infty. \end{aligned} \tag{6}$$

Similarly, if $D = D_{all}$, then replacing $b(j)$ by $p(j)$ and using the asymptotic formula for $p(j)$ from Foulds and Robinson (1984) gives

$$\mu_D(v_s) \sim \frac{[2 \log_e(2) - 1]^2}{2} \simeq 0.0746. \tag{7}$$

Moreover, for any label-invariant distribution D satisfying a simple technical condition (described later), the induced distribution on v_s is asymptotically Poisson. This condition has already been established for the particular distribution $D = D_{bin}$ (Steel, 1988). Here, using a quite different approach, we first show that as the number of taxa grows it becomes increasingly certain that the only splits shared by two randomly generated trees consist of a neighboring pair and its complement. By concentrating on these splits, one can then show that under only mild restrictions the number of shared splits is asymptotically Poisson. Finally, the normalized mean, in which d_s is divided by $(2n - 6)$, always approaches its maximum value, namely 1, for any label-invariant distribution on binary trees. The proof of this and later theorems and the corollary that follows Theorem 1 are contained in the Appendix.

To state the theorem, we make one further definition. Let μ_D' and σ_D' denote the mean and standard deviation for the number of neighboring pairs of a tree randomly generated by D.

Theorem 1

(a) As $n \to \infty$, the probability tends to 0 that two trees, randomly generated by a label-invariant tree distribution, have an equivalent split in which the smaller set in the split has three or more labels.

(b) For any label-invariant distribution D, $\mu_D(v_s)$ is bounded, and its limit, as $n \to \infty$, is at most ½. Thus, for any label-invariant distribution D on binary trees, the normalized mean $\frac{\mu_D(d_s)}{2n-6}$ tends to 1.

(c) Provided $\frac{\mu_D'}{n} \to \mu$ (for some μ) and $\frac{\sigma_D'}{n} \to 0$ as $n \to \infty$, then $P_D[v_s = k]$ is as-

TABLE 1. Rate of convergence to asymptotic values under the partition metric, d_s. To evaluate the rate of convergence of the asymptotic equations, pairs of trees were randomly generated for two distributions of trees, D_{bin} and D_{tip} (see Fig. 2), and the distance between pairs of trees was measured by the partition metric. Each entry of this table is from at least 100,000 simulations. The mean and variance are normalized by the number of internal edges on the trees. For each distribution of trees, the expected frequency of observing two trees the maximum, maximum − 1, and maximum − 2 distance apart are shown. For D_{bin}, exact values are known for $n \leq 16$ (Hendy et al., 1984). Under D_{bin}, the convergence is quite rapid, but even with D_{tip} with $n = 11$ taxa, at least 95% of the pairs of trees are expected to have no more than one edge in common. The asymptotic formulae described here (Equation 8), together with the λ values (Equations 6 and 9), will thus provide useful approximations for $n > 30$.

n	Normalized mean		Normalized variance		Probability of two trees being a given distance apart					
					D_{bin}			D_{tip}		
(taxa)	D_{bin}	D_{tip}	D_{bin}	D_{tip}	Max	Max − 1	Max − 2	Max	Max − 1	Max − 2
7	0.915	0.901	0.0241	0.0288	0.7258	0.2169	0.0477	0.6935	0.2346	0.0568
8	0.940	0.926	0.0136	0.0172	0.7549	0.2001	0.0375	0.7082	0.2289	0.0500
9	0.956	0.942	0.00822	0.01089	0.7774	0.1858	0.0314	0.7179	0.2258	0.0465
10	0.965	0.952	0.00549	0.00759	0.7918	0.1769	0.0268	0.7277	0.2211	0.0422
11	0.972	0.960	0.00383	0.00545	0.8030	0.1703	0.0234	0.7347	0.2172	0.0414
12	0.977	0.965	0.00275	0.00420	0.8155	0.1616	0.0203	0.7392	0.2154	0.0387
13	0.980	0.969	0.00208	0.00329	0.8246	0.1550	0.0186	0.7416	0.2147	0.0378
15	0.985	0.975	0.00133	0.00216	0.8334	0.1497	0.01551	0.7501	0.2104	0.0345
20	0.990	0.983	0.000578	0.00102	0.8506	0.1363	0.0122	0.7564	0.2076	0.0319
25	0.993	0.987	0.000324	0.000585	0.8572	0.1315	0.0105	0.7618	0.2046	0.0301
30	0.995	0.990	0.000206	0.000380	0.8631	0.1263	0.0100	0.7625	0.2052	0.0293
...										
∞	1.000	1.000	0.000	0.000	0.8825	0.1103	0.0069	0.8007	0.1779	0.0132

ymptotically a Poisson distribution with mean $\lambda = 2\mu^2$:

$$P_D[v_s = k] \sim \frac{e^{-\lambda}\lambda^k}{k!}. \tag{8}$$

Corollary

If D is one of the three distributions D_{all}, D_{bin}, and D_{tip}, then $P_D[v_s = k]$ is asymptotically Poisson with mean λ. For D_{bin} and D_{all}, the corresponding values of λ (given by Equations 6 and 7) are ⅛ and $\frac{[2\log_e(2) - 1]^2}{2}$, respectively, whereas for the distribution D_{tip},

$$\lambda = \frac{2}{9}. \tag{9}$$

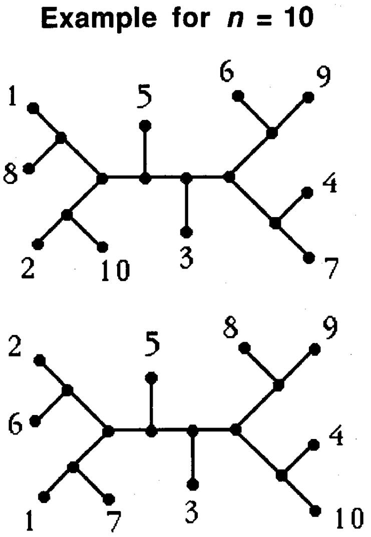
The condition $\lim_{n\to\infty} \frac{\sigma_D'}{n} = 0$ is essential for the theorem. For example, if D is label invariant and chosen so that the sum of $P_D(T)$ over all binary trees T with just two neighboring pairs (linear or caterpillar trees) is ½ and the sum of $P_D(T)$ over all binary trees with $n/2$ neighboring pairs (fully branched trees) is also ½, then it can be checked (by using Theorem 1 and considering moment generating functions) that v_s is not asymptotically Poisson. Clearly, for this contrived distribution the required condition on $\frac{\sigma_D'}{n}$ fails.

The results presented here for the partition metric are asymptotic and do not specify how large values of n need to be for the formulae to give useful approximations. We have used simulations for D_{bin} and D_{tip} to estimate the rate of convergence (see Table 1). These results show that convergence to the asymptotic values is relatively fast. By $n = 20$, the values are already within a few percent of their limiting value, and convergence is faster with D_{bin} than with D_{tip}. The large number of binary trees, even for relatively small values of n, probably aids this rapid convergence.

QUARTET METRIC (d_q)

Given a tree T, a subset S of the taxon set defines in a natural way a subtree of T, namely the minimal subtree of T, with all

n	min shared	$\max\{d_q\}$
4	0	1
5	0	5
6	1	14
7	4	31
8	9	61
9	18	108
10	34	176



Example for n = 10

FIGURE 5. The diameter of the quartet metric. Min shared = the fewest possible number of quartets shared by two trees for n = 4–10 taxa, with an example of such a pair for n = 10. For n = 11 and 12, a great deluge search (Dueck, 1991) found pairs for which min shared = 56 and 89, respectively.

vertices of degree two removed, which connects the leaves of the tree whose labels lie in S. For a subset of four taxa (a quartet), there are four subtrees possible, three of which are binary and one is degenerate. Denote by $d_Q(T_1, T_2)$ the size of the symmetric difference of the two sets of subtrees of T_1 and T_2 obtained by selecting all $\binom{n}{4}$ quartets of taxa. Because d_Q is always an even integer, we define the quartet metric d_q as $\frac{1}{2} d_Q$.

Variations on the quartet metric were introduced by Estabrook et al. (1985), and some properties of the metric were studied by Bandelt and Dress (1986) (whose definition is equivalent to d_Q) and Day (1986) (where d_q would be written as "$D + R$"). d_q is a metric rather than just a pseudometric because every tree is characterized by its quartet subtrees (see Bandelt and Dress, 1986).

The calculation of d_q is more time consuming than that of the partition metric; explicitly listing and comparing the quartets shared by two trees requires order $n^4\log(n)$ time. Doucette (unpubl.) developed a faster, implicit method for calculating d_q that requires order n^3 time. By restricting attention to just those (order n^3) quartets containing a fixed taxon, one also obtains a metric, but it has the undesirable property of violating Equation 2.

Less is known about the distribution of the quartet metric than about that of the partition metric. For example, no formula is known for the maximum possible distance between two trees. Figure 5 gives these values for up to 10 taxa and shows a representative pair of trees that realizes this value for n = 10 (these data extend an earlier table from Bandelt and Dress [1986], which considered only pairs of caterpillar trees, i.e., those possessing just two neighboring pairs, as in Fig. 1b). An interesting feature of Figure 5 is that pairs of trees that are furthest apart under the quartet metric cannot both be caterpillar trees for $8 \leq n \leq 10$. At the other end of the distribution, the minimal possible value of d_q on two distinct trees with n taxa is $n - 3$ (this may be shown by a simple inductive argument).

Bandelt and Dress (1986) showed that the maximal value of d_q, when normalized by dividing by the total number of quartets for n taxa, $\binom{n}{4}$, is monotone increasing with n. They conjectured that the limiting value of this ratio is precisely ⅔, which is the normalized mean value of d_q on binary trees under the uniform distribution, D_{bin}.

For any tree distribution, this normalized mean is also easily found. By a slight modification of the argument given in Day (1986),

$$\frac{\mu_D(d_q)}{\binom{n}{4}} = \frac{2}{3}X_n^2 + 2X_n(1 - X_n), \quad (10)$$

where X_n is the probability that for a tree randomly generated by D, a pregiven quartet, say $\{1, 2, 3, 4\}$, is resolved (i.e., has an internal edge). In particular, if D is any label-invariant distribution on binary trees (i.e., if $P_D(T) = 0$ for nonbinary trees T),

$$\frac{\mu_D(d_q)}{\binom{n}{4}} = \frac{2}{3}. \quad (11)$$

In the case $D = D_{all}$, the uniform distribution on all trees, we must know the behavior of X_n, a question raised by Day (1986). In this case (Steel, 1990),

$$1 - X_n = \sum_{i=0}^{n-3} \binom{n-3}{i} \frac{p(i+2)p(n-i-1)}{2p(n)}$$

$$\sim \sqrt{\frac{\pi[2\log_e(2) - 1]}{4n}}.$$

Thus, a tree chosen with uniform probability is increasingly likely to resolve any particular quartet as the number of taxa grows. It therefore follows from Equation 10 that the normalized mean of the quartet metric on such trees approaches ⅔, the same value as for binary trees.

The following result for the variance of d_q implies that the normalized value of d_q is increasingly certain to lie within any interval of the form ⅔ ± ϵ, as n grows, for any label-invariant tree distribution on binary trees. We also calculate the variance exactly for the uniform distribution on binary trees, D_{bin}.

Theorem 2

(a) For any label-invariant distribution D on binary trees, the variance of the normalized quartet metric, $\frac{d_q}{\binom{n}{4}}$, tends to zero as $n \to \infty$. In fact,

$$\sigma_D^2\left(\frac{d_q}{\binom{n}{4}}\right) \sim \gamma n^{-2},$$

where γ is a constant dependent on D.

(b) If $D = D_{bin}$, then

$$\sigma_D^2(d_q) = \frac{1}{9}(an^2 + bn + c) \times \binom{n}{4},$$

where $a = \frac{192}{1{,}225}$, $b = \frac{-32}{245}$, and $c = \frac{18}{1{,}225}$.

In general the calculation of the variance depends on the probability that a tree, randomly generated by D, induces the branched binary topology rather than the caterpillar topology for a given subset of six taxa. The proportion of branched to caterpillar topologies for a subset of six taxa is 1:6 for D_{bin} and 1:4 for D_{tip}. A more detailed analysis of the calculation of γ than that given in the Appendix (in which the cases leading to these two topologies are distinguished) reveals that D_{tip} gives a higher value for $\sigma_D^2(d_q)$ than does D_{bin}. This result is consistent with the results of simulations carried out for $n \leq 30$, where the normalized variance under D_{tip} is higher than that for D_{bin} for $7 \leq n \leq 30$.

Path Difference Metric (d_p)

Let $d_{ij}(T)$ denote the number of edges in T in the path joining the leaves labeled i and j, and let $d(T)$ be the associated vector obtained by a fixed ordering of the pairs $\{i, j\}$. We denote by $d_p(T_1, T_2)$ the Euclidean distance between the two vectors $d(T_1)$ and $d(T_2)$:

$$d_p(T_1, T_2) = \|d(T_1) - d(T_2)\|_2.$$

Equivalently, $d_p(T_1, T_2)$ is the square root of the sum of the squares of the differences $d_{ij}(T_1) - d_{ij}(T_2)$. Williams and Clifford (1971) defined a similar dissimilarity measure on trees, except using an L^1-norm rather than an L^2-norm. For both measures to form a metric rather than simply a pseudometric, the classical theorem, first stated by Zaretskii (1965), is required; the vector $d(T)$ characterizes T.

The metric d_p differs fundamentally from the previous two metrics in that it is not the symmetric difference of sets generated by the two trees being compared. Also, unlike the quartet metric, the mean of d_p for a distribution on binary trees depends significantly on the tree distribution, as might be expected given the wide variation in maximum path lengths between different topologies.

The complexity of calculating d_p lies between that of the other two metrics; it can be calculated in $O(n^2)$ steps. Of all metrics considered here, d_p generalizes most naturally to a metric on weighted trees; one simply lets $d_{ij}(T)$ be the sum of the weights of the edges in T on the path joining the leaves labeled i and j. However, d_p has the least natural choice of a normalizing factor of the three metrics. Division of $d_p(T_1, T_2)$ by $\sqrt{\binom{n}{2}}$ is one possibility, which is equivalent to evaluating the root mean square value of the components of $d(T_1) - d(T_2)$. However, the range of this normalized metric will continue to grow with the number of taxa. Division of $d_p(T_1, T_2)$ by $\binom{n}{2}$ gives a metric that always lies in a bounded range ([0, 2], for example); however, by Theorem 3(b) the mean value of this normalized metric tends to 0 as n grows.

We now present an exact result describing the distribution of d_p under the uniform distribution on binary trees, $D = D_{bin}$. Let $\mu(n)$ denote the mean value of $d_p{}^2$ under this model. This value is closely related to the mean and variance of $d_{ij}(T)$ (the number of edges between a pair of taxa, i and j) on a binary tree chosen with uniform probability. We denote this mean and variance by $E[d_{ij}]$ and $Var[d_{ij}]$, respectively.

Theorem 3

$$\text{Let } \alpha(n+2) = \frac{2^{2n}}{\binom{2n}{n}} \sim \sqrt{\pi n}.$$

(a) $$E[d_{ij}] = \alpha(n)$$

$$Var[d_{ij}] = 4n - 6 - \alpha(n) - \alpha^2(n).$$

(b) $$\frac{\mu(n)}{\binom{n}{2}} = 2(Var[d_{ij}])$$

$$\sim 2[(4-\pi)n - \sqrt{\pi n}].$$

We do not have an exact expression for $\mu(n)$ under D_{tip}, but results have been simulated for $7 \leq n \leq 30$. As expected from the lower proportion of caterpillar trees, the values of both the mean and variance of d_p increase more slowly than do those for D_{bin}. The diameter of d_p, normalized by dividing by $\sqrt{\binom{n}{2}}$ can also be estimated by a great deluge search (Dueck, 1991). Examples of the resulting lower bounds on the diameter are 9.2 ($n = 7$), 21.7 ($n = 10$), and 32.6 ($n = 12$).

Figure 6 illustrates the distributions of the three tree metrics d_s, d_q, and d_p for $n = 12$ under the model D_{bin}.

Discussion

We do not advocate the use of any particular tree metric; instead we provide information about the distributions so that an investigator can make a more informed choice for a particular study.

The partition metric, d_s, is the fastest to calculate, and more is known of its distribution. The Poisson limit results in a very skewed distribution so that the metric is only of use when the trees being compared are very similar. We have found this metric useful when comparing trees from different data sets and when studying how fast a tree reconstruction method is converging as longer sequences are used (Penny et al., 1982; Hendy et al., 1988). In this application, the smooth distribution at the tail is particularly useful. It has intermediate sen-

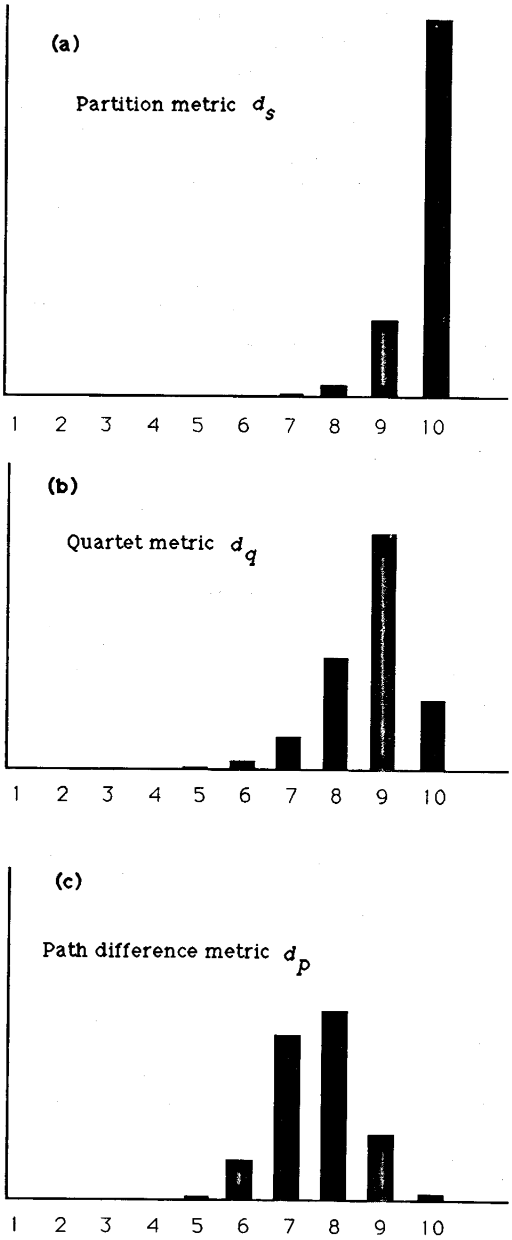


FIGURE 6. Comparison of the normalized distribution of the three tree metrics d_s (a), d_q (b), and d_p (c), for $n = 12$, obtained from a simulation involving 100,000 pairs of binary trees (D_{bin}). Each metric is normalized by its range in the simulation, which has then been subdivided into 10 equal intervals.

sitivity to the underlying distributions of trees, D_{tip} and D_{bin}. In case of doubt, it may be safest to assume the worst distribution.

The quartet metric, d_q, has many useful features for a general tree comparison metric. It has a much larger range and so is more discriminating than the partition metric; also its mean is independent of the distribution on binary trees, and for larger n, the mean is essentially the same as that for D_{all}. Its variance also has only a relatively low sensitivity to the tree distribution. Although its calculation is more time consuming than that of the partition metric, Doucette's (unpubl.) algorithm has been used to calculate distances between trees with 95 leaves in less than 17 sec on a Vax 11/780 computer (W. H. E. Day, pers. comm.). An alternative approach for comparing trees with a large number of leaves would be to evaluate a random subset of, for example, 1,000 quartets. The behavior of the quartet metric at the tails of its distribution is poorly known, and this metric probably should not yet be used in cases where this knowledge is important. A further desirable feature of the quartet metric is that quartets have a natural place in some approaches to tree reconstruction (see Eigen and Winkler-Oswatitsch, 1981; Dress et al., 1986).

The path difference metric, d_p, has several interesting features that suggest that it merits more study and consideration for use when studying evolutionary trees. It is fast to calculate and generates a wide range of values over which to compare trees. These features will make it particularly attractive when studying large trees. It is sensitive to the tree distribution, and the largest values are found with caterpillar trees (those with two neighboring pairs). It is worth studying the effect of modifying the path difference metric by dividing by the lengths of the longest paths through the trees, as this would reduce the effect of the tree distribution. Another useful application of d_p is when the topic of interest is the relative position of subsets of taxa (Penny and Hendy, 1985) rather than the comparison of trees themselves. The d_p metric may be the method of choice when trees are more dissimilar than expected by chance.

It is also necessary to decide which tree distribution is appropriate for a particular study. For example, the tip distribution seems particularly suitable for a simulation study where the tree was generated randomly during the simulation. However, in cases where a subset of taxa has been selected by the investigator, the D_{bin} distribution seems preferable. This is because such a deliberate selection of taxa usually includes a higher proportion from less numerous groups than would be expected if the taxa were picked at random. Penny et al. (1991) demonstrated that a random subset of mammalian species would be composed almost entirely of rodents and bats! Systematists who select (nonrandom) subsets of taxa for study will generally select fewer bushy trees than if random subsets were selected under the D_{tip} model, therefore in this case D_{bin} seems more appropriate.

Thus, there is no one "best" metric; the choice depends on the application, and the use of more than one metric may often be desirable. Tree comparison metrics are an important aid in the study of evolution and should make the study of trees more objective, informative, and useful to biologists.

Acknowledgments

This work was made possible by grants from the Alexander von Humboldt-Stiftung (Steel) and from the Zentrum für interdisziplinäre Forschung der Universität Bielefeld (Penny). The stimulation and encouragement of Professor A. Dress is gratefully acknowledged. We also thank the referees for their constructive comments.

References

Bandelt, H.-J., and A. Dress. 1986. Reconstructing the shape of a tree from observed dissimilarity data. Adv. Appl. Math. 7:309-343.

Bondy, J. A., and U. S. R. Murty. 1976. Graph theory with applications. Macmillan, London.

Bourque, M. 1978. Arbres de Steiner et reseaux dont varie l'emplagement de certains sommets. Ph.D. Thesis, Department d'Informatique et de Recherche Operationnelle, Univ. Montréal, Montréal.

Cavalli-Sforza, L. L., and A. W. F. Edwards. 1967. Phylogenetic analysis. Models and estimation procedures. Am. J. Hum. Genet. 19:233-257.

Cavalli-Sforza, L. L., A. Piazza, P. Menozzi, and J. Mountain. 1988. Reconstruction of human evolution: Bringing together genetic, archaeological, and linguistic data. Proc. Natl. Acad. Sci. USA 85: 6002-6006.

Day, W. H. E. 1983. Distribution of distances between pairs of classifications. Pages 127-131 *in* Numerical taxonomy (J. Felsenstein, ed.). Springer-Verlag, Berlin.

Day, W. H. E. 1985. Optimal algorithms for comparing trees with labeled leaves. J. Classif. 2:7-28.

Day, W. H. E. 1986. Analysis of quartet dissimilarity measures between undirected phylogenetic trees. Syst. Zool. 35:325-333.

Dress, A., A. von Haeseler, and M. Krueger. 1986. Reconstructing phylogenetic trees using variants of the "four-point-condition." Stud. Klassif. 17:299-305.

Dueck, G. 1991. New optimization heuristics: The great deluge algorithm and the record-to-record travel. IBM Heidelberg Scientific Centre Tech. Rep. 89.06.011. IBM, Heidelberg.

Eigen, M., and R. Winkler-Oswatitsch. 1981. Transfer-RNA: The early adaptor. Naturwissenschaften 68:217-228.

Estabrook, G. F., F. R. McMorris, and C. A. Meacham. 1985. Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. Syst. Zool. 34:193-200.

Felsenstein, J. 1978. The number of evolutionary trees. Syst. Zool. 27:27-33.

Foulds, L. R., and R. W. Robinson. 1984. Enumeration of phylogenetic trees without points of degree two. Ars Combin. 17A:169-183.

Harding, E. F. 1971. The probabilities of rooted tree-shapes generated by random bifurcations. Adv. Appl. Probab. 3:44-77.

Hendy, M. D., C. H. C. Little, and D. Penny. 1984. Comparing trees with pendant vertices labelled. SIAM J. Appl. Math. 44:1054-1065.

Hendy, M. D., M. A. Steel, D. Penny, and I. M. Henderson. 1988. Families of trees and consensus. Pages 355-362 *in* Classification and related methods of data analysis (H. H. Bock, ed.). Elsevier Science B.V., Amsterdam.

Monjardet, B. 1981. Metrics on partially ordered sets—A survey. Discr. Math. 35:173-184.

Oden, N. L., and K.-T. Shao. 1984. An algorithm to equiprobably generate all directed trees with k labeled terminal nodes and unlabeled interior nodes. Bull. Math. Biol. 6:379-387.

O'Grady, R. T., I. Goddard, R. M. Bateman, W. A. DiMichele, V. A. Funk, W. J. Kress, R. Mooi, and P. F. Cannell. 1989. Genes and tongues. Science 243:1651.

Penny, D., L. R. Foulds, and M. D. Hendy. 1982. Testing the theory of evolution by comparing phylogenetic trees constructed from five different protein sequences. Nature 297:197-200.

Penny, D., and M. D. Hendy. 1985. The use of tree comparison metrics. Syst. Zool. 34:75-82.

Penny, D., M. D. Hendy, and M. A. Steel. 1991. Testing the theory of descent. Pages 155-183 *in* Phylogenetic analysis of DNA sequences (M. M. Miyamoto and J. Cracraft, eds.). Oxford Univ. Press, New York.

ROBINSON, D. F., AND L. R. FOULDS. 1979. Comparison of weighted labelled trees. Pages 119–126 *in* Lecture notes in mathematics, Volume 748. Springer-Verlag, Berlin.

ROBINSON, D. F., AND L. R. FOULDS. 1981. Comparison of phylogenetic trees. Math Biosci. 53:131–147.

ROHLF, F. J. 1983. Numbering binary trees with labelled terminal vertices. Bull. Math. Biol. 45:33–40.

SCHRÖDER, E. 1870. Vier combinatorische Probleme. Z. Math. Phys. 15:361–376.

SIMBERLOFF, D. S. 1987. Calculating probabilities that cladograms match: A method of biogeographical inference. Syst. Zool. 36:175–195.

SLOWINSKI, J. B. 1990. Probabilities of n-trees under two models: A demonstration that asymmetrical interior nodes are not improbable. Syst. Zool. 39:89–94.

STEEL, M. A. 1988. Distribution of the symmetric difference metric on phylogenetic trees. SIAM J. Discr. Math. 1:541–551.

STEEL, M. A. 1990. Distributions on bicoloured evolutionary trees. Bull. Aust. Math. Soc. 41:159–160.

SWOFFORD, D. L. 1991. When are phylogeny estimates from molecular and morphological data incongruent? Pages 295–333 *in* Phylogenetic analysis of DNA sequences (M. M. Miyamoto and J. Cracraft, eds.). Oxford Univ. Press, New York.

THOMPSON, E. A. 1975. Human evolutionary trees. Cambridge Univ. Press, Cambridge, England.

WATERMAN, M. S., AND T. F. SMITH. 1978. On the similarity of dendrograms. J. Theor. Biol. 73:789–800.

WILLIAMS, W. T., AND H. T. CLIFFORD. 1971. On the comparison of two classifications of the same set of elements. Taxon 20:519–522.

ZARETSKII, K. A. 1965. Postroenie dereva po naboru rasstoianii mezhdu visiacimi vershinami. Usp. Mat. Nauk 20:94–96.

Received 8 September 1991; accepted 16 November 1992

APPENDIX

PROOFS OF THEOREMS 1–3

Proof of Theorem 1

To simplify notation, we will drop the letter D from probabilities that depend on the tree distribution. Also, a split in which the two sets have size k and $n - k$ is a split of type $(k, n - k)$.

(a) First consider the conditional probability, denoted $P(k \mid \tau)$, that a tree T has the split $\{\{1, \ldots, k\}, \{k + 1, \ldots, n\}\}$, given that T has a given topology τ. By the label-invariance condition of Equation 1, $P(k \mid \tau)$ is just the proportion of trees in τ that have this property. The number of trees in τ is

$$\frac{n!}{|S(\tau)|}, \tag{A1}$$

where $S(\tau)$ is the symmetry group of τ (see Hendy et al., 1984). By considering the symmetry groups of the (at most $\frac{n}{k}$) rooted subtrees of τ that have k leaves and applying Equation A1,

$$P(k \mid \tau) \le \frac{n}{k} \times \binom{n}{k}^{-1}, \tag{A2}$$

from which the probability that two trees have a split of the type described in the theorem is at most

$$\sum_{k>2} \left(\frac{n}{k}\right)^2 \times \binom{n}{k}^{-1}, \tag{A3}$$

which tends to 0 as $n \to \infty$, thereby establishing part (a).

(b) Clearly, $P(i)$ from Equation 5 is at most $\max_\tau P(i \mid \tau)$, where $P(i \mid \tau)$ is defined as in the proof of (a), and because $P(i \mid \tau) \le \frac{n}{i} \times \binom{n}{i}^{-1}$ from Equation A2, we have from Equation 5

$$\mu_D(v_s) \le \sum_{i=2}^{n/2} \frac{n^2}{i^2}\binom{n}{i}^{-1} \to \frac{1}{2}, \qquad \text{as } n \to \infty, \tag{A4}$$

and the second claim in (b) then follows immediately.

(c) From part (a), $P[v_s = k]$ asymptotically equals the probability that two trees have exactly k splits of type $(2, n - 2)$. To calculate this probability, which we denote by $P'[v_s = k]$, we make the following observation. Let $P(S \mid j)$ denote the conditional probability that a tree, T, randomly generated by D, has among its splits of type $(2, n - 2)$ a fixed collection S, given that T has j neighboring pairs. Then, by the label-invariance condition (Equation 1), $P(S \mid j)$ is just the proportion of ways of selecting and pairing $2j$ labels from a set of n labels, for which the pairings include the set S:

$$P(S \mid j) = \frac{\binom{n - 2|S|}{2j - 2|S|}}{\binom{n}{2j}} \times \frac{\Pi(2j - 2|S|)}{\Pi(2j)}, \tag{A5}$$

where $\Pi(2j) = \frac{(2j)!}{j!2^j}$ is the number of ways of pairing $2j$ objects. The (unconditional) probability that any two trees, randomly generated by D, agree, at least on any given r-tuple S of splits of type $(2, n - 2)$, is

$$\sum_{j,k} P(S \mid j)P(S \mid k)P[j]P[k], \tag{A6}$$

where $P[i]$ is the probability that a tree has exactly i neighboring pairs. We denote the probability described in Equation A6 as P_S. Let

$$N_r = \sum_{S:|S|=r} P_S, \tag{A7}$$

where the summation is taken over all collections, S, consisting of r splits of type $(2, n - 2)$. Now P_S is 0 if S contains two splits for which the two corresponding sets of size two have nonempty intersection. Otherwise, applying label invariance, P_S takes a value, denoted by P_r, that depends only on r and n. By combining these two observations, we deduce that

$$N_r = \binom{n}{2r}\Pi(2r)P_r. \qquad \text{(A8)}$$

By combining these equations and letting n tend to infinity for r fixed in Equation A8, we obtain

$$N_r \sim \frac{2^r}{r!}\times\left(\sum_j \frac{j^r}{n^r}P[j]\right)^2. \qquad \text{(A9)}$$

If $\nu(T)$ denotes the number of neighboring pairs of T, then Chebychev's inequality (applied to the hypothesis regarding $\frac{\sigma_D{}'}{n}$) implies that the probability $P\left[\left|\frac{\nu(T)}{n} - \mu\right| > \epsilon\right]$ tends to 0 as $n \to \infty$. Because $\frac{\nu(T)}{n}$ lies between 0 and 1, it follows that the bracketed term in Equation A9 tends to μ^r as $n \to \infty$. Thus, $N_r \to \frac{(2\mu^2)^r}{r!}$ as n $\to \infty$. It can then be checked, using the principle of inclusion and exclusion applied to the cumulative moments described by Equation A7, that $P'[v_s = k]$ converges to a Poisson distribution with mean $2\mu^2$.

Proof of corollary.—For D_{bin}, this result was established by Steel (1988) using a different approach. The results for both D_{bin} and D_{all} follow from the theorem if one checks the conditions on $\frac{\mu_D{}'}{n}$ and $\frac{\sigma_D{}'}{n}$. This may be achieved by considering the exponential generating function

$$\sum_{n,k}\frac{T_{n,k}}{n!}x^n y^k,$$

where $T_{n,k}$ is the number of leaf-rooted (planted) binary trees (respectively, edge-rooted and leaf-rooted phylogenetic trees, as in Foulds and Robinson, 1984) with n nonroot leaves and k neighboring pairs. If these generating functions are denoted by $B(x, y)$ and $P(x, y)$, respectively,

$$B(x, y) = \frac{1}{2}B(x, y)^2 + \frac{x^2}{2}(y - 1) + x$$

$$P(x, y) = e^{P(x,y)} - 1 - P(x, y) + \frac{x^2}{2}(y - 1) + x,$$

from which one can show that $\frac{\mu_D{}'}{n}$ converges and $\frac{\sigma_D{}'}{n} \to 0$ as $n \to \infty$.

Regarding D_{tip}, let $p(n, k)$ denote the probability that a tree with n leaves generated under the model D_{tip} has exactly k neighboring pairs. We then have the recursion

$$p(n+1,k) = \frac{2k}{n}p(n,k) + \frac{[2n-2(k-1)]}{n}\,p(n,k-1).$$

Thus, if we let $P_n(x) = \sum_k p(n, k)x^k$, we have

$$P_{n+1}(x) = xP_n(x) + \frac{2x}{n}(1 - x)\frac{\partial}{\partial x}P_n(x).$$

If $\epsilon(n)$ denotes the expected number of neighboring pairs in a tree with n leaves generated by the model D_{tip}, then because

$$\epsilon(n) = \frac{\partial}{\partial x}P_n(x)\bigg|_{x=1},$$

the previous recursion gives

$$\epsilon(n + 1) = 1 + \epsilon(n)\left(1 - \frac{2}{n}\right).$$

From this, $\epsilon(n) \sim \frac{n}{3}$ and so $\frac{\mu_D{}'}{n} \to \frac{1}{3}$ as $n \to \infty$. Similarly, applying second derivatives to the above recursion for $P_n(x)$ gives $\frac{\sigma_D{}'}{n} \to 0$ as $n \to \infty$.

Proof of Theorem 2

Given quartets $q_1, \ldots, q_i$, let $\pi(q_1, \ldots, q_i)$ denote the probability that a pair of binary trees with n leaves agree, at least on $q_1, \ldots, q_i$. Let

$$N(x) = \sum_{i\geq 0} N_i x^i,$$

where $N_0 = 1$, and for $i > 0$, N_i is the sum of $\pi(q_1, \ldots, q_i)$ over all unordered i-tuples of distinct quartets $\{q_1, \ldots, q_i\}$. Then, by the principle of inclusion and exclusion, $E(x) = N(x - 1)$ is the generating function for the probability that pairs of binary trees with n leaves agree on a given number (marked by the exponent of x) of quartets. Thus $\sigma_D{}^2(d_q) = E''(1) + E'(1) - E'(1)^2$, where $'$ denotes differentiation with respect to x. Because $E'(1)$ is just $\mu(d_q)$, given in Equation 11, $E''(1) = 2N_2$ remains to be determined. By definition, $2N_2$ is the sum of $\pi(q_1, q_2)$ over all *ordered* pairs (q_1, q_2) of distinct quartets. Because $\pi(q_1, q_2)$ depends only on the size of $q_1 \cap q_2$, we must distinguish four cases, depending on whether this intersection has cardinality 0, 1, 2, or 3. The number of ordered pairs of quartets that give rise to these cardinalities is, respectively,

$$\binom{n}{4}\times\binom{n-4}{4},\ 4\binom{n}{4}\times\binom{n-4}{3},$$

$$6\binom{n}{4}\times\binom{n-4}{2},\ \text{and}\ 4\binom{n}{4}\times\binom{n-4}{1}.$$

To calculate $\pi(q_1, q_2)$ for these cases, by definition

$$\pi(q_1, \ldots, q_i) = \sum_{t_1,\ldots,t_i} \nu(t_1, \ldots, t_i)^2, \qquad \text{(A10)}$$

where the summation is over all of the 3^i possible choices of binary trees $t_1, \ldots, t_i$, for which t_i is a resolution of the quartet q_i, and $\nu(t_1, \ldots, t_i)$ is the probability that a randomly generated tree has $t_1, \ldots, t_i$ as subtrees. We can rewrite Equation A10 as

$$\pi(q_1, \ldots, q_i) = \sum_{t_1,\ldots,t_i}\left[\sum_T \nu(T)\right]^2, \qquad \text{(A11)}$$

where the second summation is over all binary trees, T, that have leaf set $\cup_i q_i$ and that have $t_1, \ldots, t_i$ as subtrees. By symmetry arguments based on the label

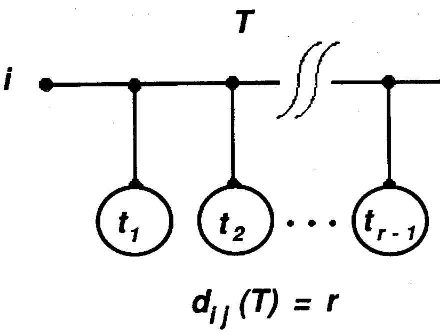


FIGURE 7. The path decomposition of a binary tree. By considering the path joining leaves i and j, T can be regarded as an ordered forest of leaf-rooted (planted) subtrees.

invariance described by Equation 1, for any label-invariant tree distribution Equation A11 gives

$$\pi(q_1, q_2) = \begin{cases} \dfrac{1}{9} & \text{if } |q_1 \cap q_2| = 0 \text{ or } 1 \\ \dfrac{33}{(15)^2} & \text{if } |q_1 \cap q_2| = 3. \end{cases}$$

In case $|q_1 \cap q_2| = 3$, $\nu(T) = \dfrac{3}{15}$ for three choices of (t_1, t_2) and $\dfrac{1}{15}$ for six choices of (t_1, t_2).

If $|q_1 \cap q_2| = 2$, write $\pi(q_1, q_2)$ as $\dfrac{1+\alpha}{9}$ (we calculate this shortly for D_{bin}) and write $\dfrac{33}{(15)^2} = \dfrac{1+\beta}{9}$, where $\beta = \dfrac{8}{25}$. Then, applying the combinatorial identity

$$\sum_{i=0}^{3} \binom{n-4}{4-i} \times \binom{4}{i} = \binom{n}{4} - 1,$$

we deduce

$$E''(1) = \frac{1}{9}\left[\binom{n}{4} - 1 + 6\alpha\binom{n-4}{2} + 4\beta\binom{n-4}{1}\right] \times \binom{n}{4},$$

and so

$$\sigma_D{}^2(d_q) = \frac{1}{9}\left[2 + 6\alpha\binom{n-4}{2} + 4\beta\binom{n-4}{1}\right] \times \binom{n}{4} \quad \text{(A12)}$$

so that

$$\sigma_D{}^2\left[\frac{d_q}{\binom{n}{4}}\right] \sim 8\alpha n^{-2},$$

which proves part (a).

For part (b), we must determine $\pi(q_1, q_2)$ when $|q_1 \cap q_2| = 2$. For this, we may suppose that $q_1 = \{1, 2, 3, 4\}$ and $q_2 = \{1, 2, 5, 6\}$. From Equation A11, $\nu(T) = \dfrac{1}{105}$ (because $|q_1 \cup q_2| = 6$, $b(6) = 105$). By considering the nine possible choices for the pair (t_1, t_2) for Equation A11, 17 trees T arise in the one case where t_1 and t_2 both have $\{1, 2\}$ as a neighboring pair, whereas 9 trees arise in four other cases, and 13 trees arise in another four cases. Thus, Equation A11 gives

$$\pi(q_1, q_2) = 1 \times \frac{17^2}{105^2} + 4 \times \frac{9^2}{105^2} + 4 \times \frac{13^2}{105^2}$$

so that

$$\alpha = \frac{64}{1{,}225}.$$

The claimed expression for $\sigma_D{}^2(d_q)$ then follows from Equation A12 with $a = 3\alpha$, $b = -27\alpha + 4\beta$, and $c = 60\alpha - 16\beta + 2$.

Proof of Theorem 3

For $s = 1, 2$, let E_S denote, respectively, the expected value of $d_{ij}(T)$ and $d_{ij}(T)[d_{ij}(T) - 1]$ under D_{bin}. Thus,

$$E[d_{ij}] = E_1$$
$$Var[d_{ij}] = E_2 + E_1 - E_1{}^2. \quad \text{(A13)}$$

(a) Every binary tree T having n labeled leaves and for which $d_{ij}(T) = r$ has a unique representation as in Figure 7, where the trees $t_1, \ldots, t_{r-1}$ are binary trees with a distinguished (root) leaf. This representation provides a bijection between the set of trees T with $d_{ij}(T) = r$ and the collection of ordered forests consisting of $r - 1$ planted (i.e., leaf-rooted) binary trees with a total of $n - 2$ nonroot labeled leaves.

Let $B(x) = \sum_{n>0} \dfrac{b(n+1)}{n!} x^n$, which is the exponential generating function for the number of planted binary trees, $b(n + 1)$, with n labeled nonroot leaves. By the standard decomposition for planted binary trees,

$$B(x) = \frac{1}{2}B(x)^2 + x,$$

and so

$$B(x) = 1 - \sqrt{1 - 2x}.$$

Because

$$F(x, y) = yB(x) + y^2B(x)^2 + \ldots = \frac{1}{[1 - yB(x)]} - 1$$

is the exponential generating function for the number of ordered forests consisting of a given number of rooted trees (marked by y) and a given number of leaves (marked by x), by applying the above bijection, we have

$$E_1 = \frac{(n-2)!}{b(n)} \times [x^{n-2}]\frac{\partial}{\partial y} yF(x, y)|_{y=1}$$

$$E_2 = \frac{(n-2)!}{b(n)} \times [x^{n-2}]\frac{\partial^2}{\partial y^2} yF(x, y)|_{y=1},$$

where $[x^k]f(x)$ denotes the coefficient of x^k in $f(x)$. Using the above equation for $B(x)$ to simplify these two expressions, $E_1 = \alpha(n)$, whereas $E_2 = 4n - 6 - 2\alpha(n)$, and part (a) now follows by substituting these values into Equation A13.

(b) By definition,

$$\mu(n) = \frac{1}{b(n)^2} \times \sum_{T,T'} \sum_{i<j} [d_{ij}(T) - d_{ij}(T')]^2.$$

Expanding the squared terms in this sum and then inverting the order of summation, noting that $\sum_{T} d_{ij}(T)^s$ is the same for all i, j, gives

$$\mu(n) = 2 \times \binom{n}{2} \times \left[\sum_{T} \frac{d_{ij}(T)^2}{b(n)} - \left(\sum_{T} \frac{d_{ij}(T)}{b(n)}\right)^2\right],$$

from which the result follows immediately.