SIAM J. DISC. MATH.
Vol. 1, No. 4, November 1988

010

DISTRIBUTION OF THE SYMMETRIC DIFFERENCE METRIC ON PHYLOGENETIC TREES*

M. A. STEEL†

Abstract. The symmetric difference metric has been useful in comparing phylogenetic trees derived from DNA sequence data. The main result shown here is that the frequency of pairs of binary trees a given distance apart is described by a limiting Poisson distribution, with $e^{-1/8} \approx 88$ percent of all pairs maximally distant. Asymptotic bounds on the distribution are derived, and the asymptotic mean and variance of the normalized metric on the class of all phylogenetic trees is also calculated. The results rely on simple combinatorial constructions and analytic properties of appropriate generating functions.

Key words. trees, binary trees, phylogenetic trees, probability, symmetric difference metric, asymptotic distribution

AMS(MOS) subject classifications. 05C05, 60C05, 92A12

Introduction. The symmetric difference metric defined on phylogenetic trees is a special case of symmetric difference metrics on sets studied in [9], [12]. The tree metric has been useful in testing evolutionary hypotheses and in examining the methods used to build evolutionary trees [11]. An optimally efficient algorithm has been developed by Day [4] to compute the metric, and its distribution among pairs of small trees is described in [3] and [8].

This paper extends those results to obtain bounds on the distribution of pairs of arbitrarily-large binary trees a given distance apart. As a result, the asymptotic distribution is shown to be Poisson, which answers a conjecture in [8]. The distribution of the normalized metric on the full class of phylogenetic trees is also examined. In particular, the asymptotic mean of the normalized metric is derived, and confirms a second conjecture in [8].

A description of the distribution in the binary case is also given in terms of a generating polynomial, the coefficients of which can be computed directly from the tree. Some consequences of this representation are given. The properties of the metric on binary trees make it useful for hypothesis testing involving trees derived from homologous DNA sequences, as in [10]. The resulting trees may be expected to be similar and it is useful to have a metric for which most trees are far apart.

DEFINITIONS. Let L be a set of $n \geq 2$ labels. A *phylogenetic tree on* L is a tree with n vertices of degree one, each labeled with a distinct element from L, and with the remaining (internal) vertices of degree at least three, and unlabeled. For such a tree, the n edges incident with a pendant vertex are called *pendant edges*, and the remaining edges are *internal*.

Two phylogenetic trees on L are considered equivalent if there is a graph isomorphism between them that preserves the labeling on the pendant vertices. More generally, if two phylogenetic trees are graph isomorphic with their labelings suppressed, we say they are *topologically equivalent*.

As in [8], let $PT(n,f)$ denote the set of *phylogenetic trees* with n pendant vertices and f internal edges on the label set $\{1, \cdots, n\}$. For $n \geq 3$, let $PT(n) = \cup\{PT(n,f); 0 \leq f \leq n-3\}$, and $BPT(n) = PT(n, n-3)$, the set of *binary phylogenetic trees*, for which each internal vertex has degree three. The following result is from [5, p. 29].

* Received by the editors July 13, 1987; accepted for publication (in revised form) May 5, 1988.
† Department of Mathematics and Statistics, Massey University, Palmerston North, New Zealand.

LEMMA 1. *The size of* $PT(n,f)$ *is determined recursively as follows*:

$$|PT(n,f)| = (n+f-2)|PT(n-1,f-1)| + (f+1)|PT(n-1,f)|, \qquad n \geq 4,$$

$$|PT(3,0)| = 1, \quad |PT(3,f)| = 0, \quad f > 0.$$

For $n \geq 3$, Lemma 1 gives $|BPT(n)| = (2n-5)!! = 1.3.5 \cdots (2n-5)$. For convenience, we let $b(n) = (2n-5)!!$ and $p(n) = |PT(n)|$.

The symmetric difference metric d, which Bourque [2], and Robinson and Foulds [13] applied to phylogenetic trees, is defined on $PT(n)$, and so on $BPT(n)$, as follows: For $T \in PT(n)$, deletion of an internal edge e induces a two-set partition $\pi(T, e)$ of $\{1, \cdots, n\}$ corresponding to the labels on the two connected components of T with e deleted. For $T_1 \in PT(n_1, f_1)$, $T_2 \in PT(n_2, f_2)$, and $\pi(T_1, e_1) = \pi(T_2, e_2)$ we call e_1, e_2 an *equivalent pair* of *edges*. If T_1, T_2 have exactly m equivalent pairs of edges then $d(T_1, T_2) = f_1 + f_2 - 2m$. In particular for $T_1, T_2 \in BPT(n)$,

$$d(T_1, T_2) = 2(n-3-m).$$

For $T \in PT(n)$ we recall from [8] the generating polynomials

$$P(T) = P(T,x) = \sum_{m \geq 0} p_m(T)x^m, \qquad Q(T) = Q(T,x) = \sum_{m \geq 0} q_m(T)x^m$$

where $p_m(T)$ (respectively, $q_m(T)$) is the number of trees in $PT(n)$ (respectively, $BPT(n)$) at distance m from T.

Thus, for $T \in PT(n,f)$, $Q(T,x)$ has degree $n+f-3$ and is an even or odd polynomial of parity equal to the numerical parity of its degree. For $T \in PT(n)$, $s \geq 0$ let $q(s,T)$ denote the number of binary trees having s equivalent edge pairs with T, and let $q(s,n)$ be the average value of $b(n)^{-1}q(s,T)$ over $BPT(n)$. Thus, $q(s,n) = b(n)^{-2}\sum_{T \in BPT(n)} q(s,T)$ is the probability that two trees randomly chosen from $BPT(n)$ have exactly s equivalent pairs of edges.

The main result (Theorem 3) shows that $q(s) = \lim_{n \to \infty} q(s,n)$ has a Poisson distribution in s with mean $\frac{1}{8}$. Consequently, the probability that two randomly chosen trees are a maximal distance apart tends asymptotically to $e^{-1/8}$, answering a question raised in [8].

We begin by noting that for $T \in PT(n)$, $P(T)$ and $Q(T)$ do not depend on the labeling of T, but only on its topology. We shall frequently write these and other tree-valued functions that are invariant under topological equivalence without specifying the labeling of the tree. With this in mind from [8], we now state the following theorem.

THEOREM 1. *Let* e *be an internal edge of* $T \in PT(n)$. *Let* T/e *be the tree formed by contracting* e, *and let* T_1, T_2 *be the maximal subtrees of* T *with* e *as a pendant edge. Then*

$$P(T) = xP(T/e) + (1-x^2)P(T_1)P(T_2),$$

$$Q(T) = xQ(T/e) + (1-x^2)Q(T_1)Q(T_2).$$

We now give a constructive description of $Q(T,x)$. Let $T \in PT(n,f)$ and let E be a set of internal edges of T. For each edge $e \in E$ cut e in half and place new pendant vertices on each of the two "ends" of e. In this way E defines a collection of trees T_i, having n_i pendant vertices, for $i = 1, \cdots, |E|+1$ (with $T_1 = T$, if $E = \emptyset$). Clearly, $\sum_{1 \leq i \leq |E|+1} n_i = n + 2|E|$. Let $\Phi(E)$ be the sequence $(n_1, \cdots, n_{|E|+1})$, taken in some

order, and let $\langle \Phi(E) \rangle = \prod_{1 \le i \le |E|+1} b(n_i)$. Define $r(s, T)$ to be the sum of $\langle \Phi(E) \rangle$ over all sets of internal edges E, with $|E| = s$. Finally, let

$$R(T) = R(T, x) = \sum_{s \ge 0} r(s, T)x^s, \qquad q(T) = q(T, x) = \sum_{s \ge 0} q(s, T)x^s$$

(so that $q(T, x) = x^{(n-3+f)/2} Q(T, x^{-1/2})$).

LEMMA 2. *For $T \in$ PT (n), $T_n \in$ PT $(n, 0)$, we have the following*:

(a) $q(T, x) = R(T, x - 1)$;

(b) *In the notation of Theorem* 1,

(i) $R(T) = R(T/e) + xR(T_1)R(T_2)$,

(ii) $q(T) = q(T/e) + (x - 1)q(T_1)q(T_2)$;

(c) $R(T, 0) = R(T_n, x) = b(n)$.

Proof. (a) Let $T \in$ BPT (n) and let E be a set of s internal edges of T. Under the above construction, for each edge $e \in E$, new pendant vertices v_1, v_2 are attached to the ends of a bisection of e.

For $i = 1, 2$, label v_i with the *set* of labels of those pendant vertices of T that are no longer joined by a path to v_i when e is cut. Each tree T_i $(i = 1, \cdots, s + 1)$ defined by E, thus has a natural label set L_i for its pendant vertices, so that $T_i \in BPT(L_i)$. This process is illustrated for $s = 2$ in Fig. 1 by the tree J_6 with two distinguished edges.

Now, let $B(T, E)$ be the set of trees in $BPT(n)$ equivalent to T on E and possibly other internal edges. We construct a bijection F from $B(T, E)$ to $\prod_i BPT(L_i)$. Given $T' \in B(T, E)$, performing the above edge splitting and labeling procedure on T' produces the label sets $L_1, \cdots, L_{s+1}$, and hence an element of $F(T') \in \prod_i BPT(L_i)$.

The inverse of F, takes $(T_1, \cdots, T_{s+1}) \in \prod_i BPT(L_i)$ and identifies all pairs of pendant vertices v_1, v_2 labeled with sets A_1, A_2 such that $A_1 \cup A_2 = L$, the identified vertices then being suppressed to give a tree in $B(T, E)$. Now, $|L_i| = n_i$ implies $|BPT(L_i)| = b(n_i)$, so that the bijection gives $|B(T, E)| = \langle \Phi(E) \rangle$. By the principle of inclusion and exclusion [6] $r(T, x - 1)$ is then the (ordinary) generating function for the number of binary trees equivalent to T on an exact number of internal edges, thus we establish (a).

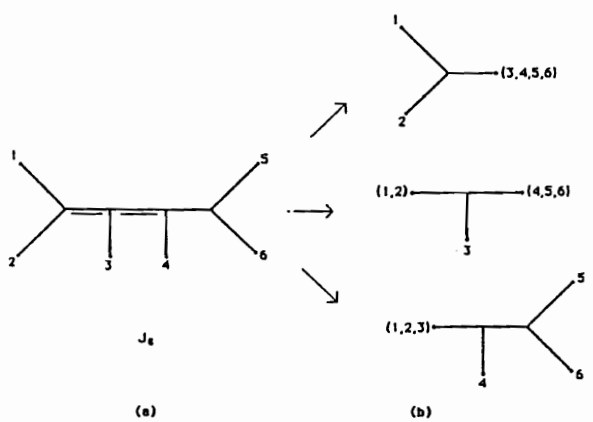

FIG. 1

Part (b)(i) can be proved directly, or from Theorem 1 by noting $Q(T, x) = x^{n-3+f} R(T, x^{-2} - 1)$. Part (b)(ii) follows from (a), while part (c) follows from the definition of $R(T, x)$.

Example 1. For J_6 in Fig. 1(a), we have

$$R(J_6, x) = b(6) + (2b(3)b(5) + b(4)^2)x + 3b(3)^2 b(4)x^2 + b(3)^2 x^3$$
$$= 105 + 39x + 9x^2 + x^3,$$
$$q(J_6, x) = R(T, x - 1) = 74 + 24x + 6x^2 + x^3,$$
$$Q(J_6, x) = 1 + 6x^2 + 24x^4 + 74x^6.$$

Example 2. Let $T \in PT(n, f)$. Then $q(s, T) = 0$, for $s > f$, and $q(f, T) = \prod_i b(\partial_i)$, where $(\partial_1, \cdots, \partial_{f+1})$ is the degree sequence of the internal vertices of T.

DEFINITION. For $n \ge 2a > 4$, let $T_n(a) \in PT(n)$ be a tree obtained by attaching pairs of pendant vertices to "a" pendant vertices of a (star) tree $T \in PT(n - a, 0)$. The resulting tree is unique up to topology so that $q(s, T_n(a))$ is well defined. We call such trees *binary semistars*. The tree J_6, in Fig. 1(a), with its central edge contracted, is a $T_6(2)$.

To calculate $q(s, T_n(a))$, let F be the set of internal edges of $T_n(a)$, so that $|F| = a$, and for $E \subseteq F$, $\Phi(E)$ consists of $|E|$ copies of three and one copy of $n - |E|$, giving $\langle \Phi(E) \rangle = b(n - |E|)$. Thus, $r(s + i, T) = {}^aC_{(s+i)} b(n - s - i)$, and so by Lemma 2,

$$q(s, T_n(a)) = \sum_{i \ge 0} (-1)^i \cdot {}^{(s+i)}C_s \cdot {}^aC_{(s+i)} \cdot b(n - s - i).$$

Rearranging terms, we obtain Lemma 3.

LEMMA 3. $q(s, T_n(a)) = {}^aC_s \sum_{i \ge 0} (-1)^i \cdot {}^{(a-s)}C_i \cdot b(n - s - i)$.

DEFINITIONS. For $T \in PT(n)$, $n > 4$, e is a *binary edge* of T if e is an internal edge adjacent to a pair of adjacent pendant edges. For $T \in PT(n)$, let $a(T)$ be the number of binary edges of T.

For $T \in PT(n)$, $n > 4$, if e is a nonbinary internal edge of T, we say the contracted tree T/e is a *σ-reduction* of T. Equivalently it is a contraction for which the maximal subtrees T_1, T_2 both have at least four pendant vertices.

LEMMA 4. *For $T \in PT(n, f)$, $n > 4$, $a(T) = a$, T can be successively σ-reduced to a $T_n(a)$ in $f - a$ steps.*

Proof. If $T' \in PT(n)$ is not a binary semistar, T' has a subtree topologically equivalent to either J_6 in Fig. 1(a), or to J_6 with a contracted noncentral internal edge. Since these trees are σ-reducible, T' is also. By induction, T can thus be reduced to a binary semistar $T_n(a)$. Since the internal edges that survive under all possible σ-reductions are precisely the binary edges, we have $a(T) = a$, the number of binary edges of $T_n(a)$. Finally since each σ-reduction eliminates one internal edge, $f - a$ steps are required.

Now for positive integers t, s, let $f(x)$ be a positive, real-valued function defined on integers greater than or equal to s, with the property that $t < x \le y$ implies $f(x)f(y) \le f(x - 1)f(y + 1)$. Thus, for positive integers N, k, with $N \ge kt$, induction on N for each k gives

$$\max \left\{ \prod_{1 \le i \le k} f(x_i) : \sum_i x_i = N, x_i \ge t \right\} = f(t)^{k-1} f(N - (k - 1)t).$$

Taking $t \ge 3$, $f(x) = b(x)$, we have, for $x \le y$, $b(x)b(y) = (2x - 5)b(x - 1)b(y) \le (2y - 5)b(x - 1)b(y) = b(x - 1)b(y + 1)$, so that $b(x)$ satisfies the above property.

This gives the following lemma.

LEMMA 5. *For positive integers* $t \geq 3$, N, k;

$$\max\left\{\prod_{1 \leq i \leq k} b(x_i) : \sum_i x_i = N, x_i \geq t\right\} = b(t)^{k-1} b(N-(k-1)t).$$

Example 3. For $T \in PT(n, f)$, $n \geq 3$, $s \geq 0$, $q(s, T) \leq {}^{f}C_s b(n-s)$, so that

$$\lim_{n \to \infty} b(n)^{-1} q(s, T) \leq 2^{-s}/s!.$$

Proof of Example 3. By Lemma 2(a), $q(s, T) \leq \sum_{i \geq 0} {}^{(s+i)}C_s . q(s+i, T) = r(s, T)$. From Example 2 we may assume $s \leq f \leq n-3$, so that for a set E of s internal edges of T, if $\Phi(E) = (n_1, \cdots, n_{s+1})$, then $\sum_i n_i = n + 2s \geq 3(s+1) \geq \min\{n_i\}.s$. Thus, we can apply Lemma 5 with $t = 3$, $N = n+2s$, $k = s+1$, to obtain $\langle \Phi(E) \rangle \leq b(3)^t b(n-s) = b(n-s)$, and since there are ${}^{f}C_s$ possible choices for E, $r(s, T) \leq {}^{f}C_s . b(n-s)$, which completes the proof.

THEOREM 2. *For* $T \in PT(n)$, $s \geq 0$, $n > 4$,

$$b(n)^{-1} q(s, T) = b(n)^{-1} q(s, T_n(a)) + \delta(s, T),$$

where $a = a(T)$, $|\delta(s, T)| < 3(s+1)/2(2n-7)$.

Proof. If $n = 5$, T is a binary semistar and the theorem holds, then suppose $n \geq 6$. By Lemmas 2(a) and 2(b),

$$(1) \quad q(s, T) = q(s, T/e) + \sum_{\alpha+\beta=s-1} q(\alpha, T_1) q(\beta, T_2) - \sum_{\alpha+\beta=s} q(\alpha, T_1) q(\beta, T_2).$$

Now if T/e is a σ-reduction of T, each of the two summation terms in (1) does not exceed $(s+1)b(4)b(n-2)$ (since for $T \in PT(m)$, $q(s, T) \leq b(m)$, and for $n \geq 6$, Lemma 5 applies with $k = 2$, $t = 4$, $N = n+2$). Thus, by Lemma 4, $b(n)^{-1} q(s, T) = b(n)^{-1} q(s, T_n(a)) + \delta(s, T)$; $a = a(T)$, where

$$\begin{aligned} |\delta(s, T)| &\leq (s+1)(n-3-a)b(4)b(n-2)/b(n) \\ &= 3(s+1)(n-3-a)/(2n-5)(2n-7) \\ &< 3(s+1)/2(2n-7). \end{aligned}$$

LEMMA 6. *Let*

$$\eta_n(a) = |\{T \in BPT(n) : a(T) = a\}|,$$

$n > 4$, *and* $T_n(x) = \sum_a \eta_n(a) x^a$, $t_n(s) = D^s(T_n(x))|_{x=1}$, *where* D *denotes differentiation with respect to* x. *Then*, $t_n(s) = 2^{-s} b(n-s) n!/(n-2s)!$.

Proof. Given a set of labels L, let $\pi_2(L, k)$ be the collection of all sets of k disjoint sets of size two drawn from L, and let $\pi_2(n, k) = \pi_2(\{1, \cdots, n\}, k)$. For $S \in \pi_2(n, k)$, let $A(S) \subseteq BPT(n)$ be the set of binary trees for which any pair of pendant vertices labeled by a pair from S have adjacent edges. If $A(n, k) = \sum_{S \in \pi_2(n,k)} |A(S)|$, and $A_n(x) = \sum_j A(n, k) x^j$, then by the principle of inclusion and exclusion [6] $T_n(x) = A_n(x-1)$, giving $t_n(s) = s! A(n, s)$.

Let $V(n, s) = \{(T, S) : T \in A(S), S \in \pi_2(n, s)\}$, and let

$$W(n, k) = \{(T, G, H) : T \in BPT(n-s), G \in \pi_2(H, s), H \subseteq \{1, \cdots, n\}, |H| = 2s\}.$$

Then $A(n, s) = |V(n, s)|$, and we construct a bijection F from $V(n, s)$ to $W(n, s)$ as follows. Given $(T, S) \in V(n, s)$, let $H(S) = \cup\{X : X \in S\}$, so that $|H(S)| = 2s$, and $S \in \pi_2(H(S), s)$.

Let $T(S)$ be the tree obtained by deleting each pair of pendant vertices (and their pendant edges) with labels x_1, x_2 in S, and relabeling the exposed internal vertex by $\min\{x_1, x_2\}$. Relabeling $T(S)$ again by $\{1, \cdots, n-s\}$ so as to preserve the order of labels gives a tree $T'(S) \in BPT(n-s)$. Let $F(T, S) = (T'(S), S, H(S))$. Then F has the following inverse (and hence is a bijection). Given $(T, G, H) \in W(n, s)$, if $G = \{\{x_1, y_1\}, \cdots, \{x_s, y_s\}\}$, let $L(G) = \{1, \cdots, n\} - \cup_i \{\max\{x_i, y_i\}\}$. Given $T \in BPT(n-s)$, relabel the pendant vertices of T with $L(G)$ so as to preserve the order of the labels. Then, for $i = 1, \cdots, s$, join new pendant vertices labeled x_i, and y_i to the pendant vertex of T labeled $\min\{x_i, y_i\}$, and we obtain a tree $T' \in A(G)$. Thus, $F^{-1}((T, G, H)) = (T', G)$.

Summarizing, we have $t_n(s) = s! A(n, s) = s! |V(n, s)| = s! |W(n, s)|$. But $|W(n, s)| = b(n-s) |\pi_2(2s, s)| {}^{n}C_{2s}$, and $|\pi_2(2s, s)| = (2s)!/s! 2^s$ [1], which proves the lemma.

Remark 1. An argument in [7] shows that

$$\eta_n(a) = \begin{cases} n!(n-4)!/(n-2a)!a!(a-2)!2^{2a-2} & \text{for } 2 \leq a \leq [n/2], \\ 0 & \text{otherwise.} \end{cases}$$

The proof relies on a recurrence for $\eta_n(a)$ which can be written

$$T_n(x) = (n-4+nx-x)T_{n-1}(x) + 2(x-x^2)d/dx T_{n-1}(x).$$

An inductive argument based on this result gives an alternative proof of Lemma 6.

THEOREM 3. *For* $n > 4$, $s \geq 0$, $q(s, n) = \psi(s, n) + \delta_n(s)$, *where* $\psi(s, n) = (n!/b(n)2^s s!) \sum_i b(n-s-i)^2/(-2)^i i!(n-2s-2i)!$, $|\delta_n(s)| < 3(s+1)/2(2n-7)$.

Proof.

$$\begin{aligned} q(s, n) &= b(n)^{-2} \sum_{T \in BPT(n)} q(s, T) \\ &= b(n)^{-2}\left\{\sum_a q(s, T_n(a))\eta_n(a)\right\} + \delta_n(s) \quad \text{by Theorem 2.} \end{aligned}$$

By Lemma 3,

$$\begin{aligned} \sum_{a \geq 0} q(s, T_n(a))\eta_n(a) &= \sum_{a \geq 0}\left\{{}^{a}C_s \sum_{i \geq 0} (-1)^i . {}^{(a-s)}C_i . b(n-s-i)\right\}\eta_n(a) \\ &= \sum_{i \geq 0}\left\{\sum_{a \geq 0} {}^{a}C_s {}^{(a-s)}C_i \eta_n(a)\right\}(-1)^i b(n-s-i). \end{aligned}$$

Now, $\sum_{a \geq 0} {}^{a}C_s {}^{(a-s)}C_i \eta_n(a) = t_n(s+i)/s!i!$, so that

$$\begin{aligned} b(n)^{-2} \sum q(s, T_n(a))\eta_n(a) &= (s!)^{-1} \sum_{i \geq 0} (-1)^i t_n(s+i) b(n-s-i)/b(n)^2 i! \\ &= \psi(s, n) \quad \text{by Lemma 6, as required.} \end{aligned}$$

COROLLARY 1. $q(s) = \lim_{n \to \infty} q(s, n) = e^{-1/8}/8^s s!$.

Proof. The proof is obtained by Theorem 3, observing that

$$\lim_{n \to \infty} b(n-x)^2 n!/b(n)^2 (n-2x)! = 4^{-x} \quad \text{and} \quad \lim_{n \to \infty} \delta_n(s) = 0.$$

COROLLARY 2. *If* $v(n)$ *is the expected distance between two trees in* $BPT(n)$, *and* $\sigma^2(n)$ *is the variance, then we have the following*:

(a) $\lim_{n \to \infty} (2n-6) - v(n) = 0.25$;

(b) $\lim_{n\to\infty} \sigma^2(n) = 0.5$.

Proof. (a) $v(n) = \sum_s q(s, n)(2n - 6 - 2s)$; thus,

$$(2n-6) - v(n) = (2n-6)\left(1 - \sum_s q(s,n)\right) + 2\sum_s sq(s,n) = 2\sum_s sq(s,n),$$

since $\sum_s q(s, n) = 1$, and $\lim_{n\to\infty} \sum_s sq(s, n) = \sum_s sq(s) = \frac{1}{4}$, by Corollary 1.

(b)

$$\sigma^2(n) = \sum_s q(s,n)((2n-6-2s)-v(n))^2$$

$$= ((2n-6)-v(n))^2 \sum_s q(s,n) - 4(2n-6-v(n)) \sum_s sq(s,n)$$

$$+ 4\sum_s s^2 q(s,n).$$

Letting $n \to \infty$, using (a) and noting that $\sum_s s^2 q(s) = 0.125 + (0.125)^2$, we obtain the result.

Remark 2. Using a different type of counting argument it can be shown that $q(0, n)$ is monotone increasing in n, for $n \geq 3$. Table 4 of [8], which gives $q(s, n)$ for $4 \leq n \leq 16$, further suggests that for each $s > 0$, $q(s, n)$ is monotone decreasing in n.

Having found the asymptotic average value over $BPT(n)$ of $b(n)^{-1}q(s, T)$, we now calculate its asymptotic range.

THEOREM 4.

$$\lim_{n\to\infty} \sup_{m \geq n, T \in BPT(m)} \{ b(m)^{-1}q(s,T) \} = \begin{cases} 1, & s = 0, \\ e^{-1/4}/4^s s!, & s > 0, \end{cases}$$

$$\lim_{n\to\infty} \inf_{m \geq n, T \in BPT(m)} \{ b(m)^{-1}q(s,T) \} = \begin{cases} e^{-1/4}, & s = 0, \\ 0, & s > 0. \end{cases}$$

Proof. For $s = 0$, (1) gives $q(0, T) \geq q(0, T_n(a))$, which together with Lemma 3 and Theorem 2 gives $q(0, T) \geq 1 - a(T)/(2n - 5) + \delta(0, T)$.

For each positive integer n, choose $J_n \in \mathrm{BPT}(n)$ with $a(J_n) = 2$. For a given n, any two such trees are topologically equivalent and are often called caterpillar trees. Then $b(n)^{-1}q(0, J_n) \geq 1 - 2/(2n - 5) + \delta(0, J_n) \to 1$, as $n \to \infty$, and since

$$b(m)^{-1}q(0, T) \leq 1$$

for all $T \in BPT(m)$, lim sup $\{ b(m)^{-1}q(0, T) \} = 1$. A similar argument using J_n gives lim inf $\{ b(m)^{-1}q(0, T) \} = 0$ for $s > 0$. To obtain the other two results, consider the family of binary trees, $K_n \in BPT(2n)$, obtained by attaching pairs of pendant edges to every pendant vertex of a caterpillar tree $T \in J_n$. For fixed n the trees obtained are again topologically equivalent. An example of a K_5 is given in Fig. 2. Clearly $a(K_n) = n$, so that by Lemma 3 and Theorem 2,

$$b(2n)^{-1}q(s, K_n) = {}^nC_s \sum_{i \geq 0} (-1)^i \cdot {}^{(n-s)}C_i \cdot b(2n - i - s)/b(2n) + \delta(s, K_n).$$

Rearranging, we have

$$(s!)^{-1} \sum_{i \geq 0} (-1)^i (i!)^{-1} \{ n(n-1) \cdots (n-s-i+1)$$

$$/(2(2n-5) \cdots (2(2n) - 2i - 2s - 3)) \} + \delta(s, K_n).$$

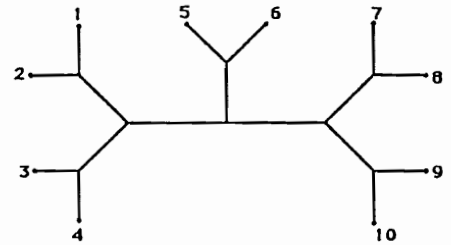

FIG. 2

Now the bracketed term has $s + i$ factors in the numerator and denominator so that as $n \to \infty$, the bracketed term approaches 4^{-s-i}. Hence,

$$\lim_{n\to\infty} b(2n)^{-1}q(s, K_n) = (4^s s!)^{-1} \sum_{i \geq 0} \left(-\frac{1}{4}\right)^i \Big/ i!$$

$$= e^{-1/4}/4^s s!.$$

Thus,

$$\limsup \{ b(m)^{-1}q(s,T) \} \geq e^{-1/4}/4^s s!,$$

$$\liminf \{ b(m)^{-1}q(0,T) \} \leq e^{-1/4}.$$

We now show $\alpha = \liminf \{ b(m)^{-1}q(0, T) \} \geq e^{-1/4}$ and by a similar argument, the other inequality for lim sup can be derived, thus establishing the theorem. Let $T_i \in BPT(n_i)$ be a sequence with $\lim_{i\to\infty} b(n_i)^{-1}q(0, T_i) = \alpha$. Then since $a(T_i)/n_i$ is bounded, it has a convergent subsequence $a(T_{i(j)})/n_{i(j)}$. Let $\gamma = \lim_{j\to\infty} a(T_{i(j)})/n_{i(j)}$. Since $T_{i(j)}$ is a subsequence of T_i, $\lim_{j\to\infty} b(n_{i(j)})^{-1}q(0, T_{i(j)}) = \alpha$. A calculation similar to the above result on $b(2n)^{-1}q(s, K_n)$ shows $\alpha = e^{-\gamma/2}$. But for any $T \in BPT(n)$, $a(T) \leq n/2$, so that $\gamma \leq \frac{1}{2}$. Thus $\alpha \geq e^{-1/4}$, as required.

We now consider the distribution of the symmetric difference metric on $PT(n)$. The normalized distance between two trees $T, T' \in PT(n)$ is $d(T, T')$ divided by the maximum possible distance, $2n - 6$. Theorem 5 shows that, asymptotically, the normalized distance becomes increasingly peaked about its mean $\mu(n)$, which is shown to be less than one, confirming a conjecture in [8]. The symbols O and $\sim$ have their usual meaning: $f(n) \sim g(n)$ means $\lim_{n\to\infty} f(n)/g(n) = 1$, and $f(n) = O(g(n))$ means $f(n)/g(n)$ is bounded as $n \to \infty$. Let $\rho = 2 \ln 2 - 1$. Lemma 7 follows from similar results in [14].

LEMMA 7.

(a) $p(n)/n! \sim \rho^{1-n} n^{-5/2} \sqrt{(\rho/4\pi)}$.

(b) $p(n)^{-1}(\sum_{f \geq 0} f \,|PT(n, f)|) - n\rho^{-1}(1 - \ln 2) = O(1)$.

(c) $p(n)^{-1}(\sum_{f \geq 0} f(f-1)\,|PT(n, f)|) - n^2\rho^{-2}(1 - \ln 2)^2 = O(n)$.

THEOREM 5. *Let $\mu(n)$ and $\sigma^2(n)$ denote, respectively, the mean and variance of the normalized distance between two trees in $PT(n)$. Then, we have the following:*

(a) $\mu(n) \sim (1 - \ln 2)/\rho \approx .7943$;

(b) $\sigma^2(n) = O(n^{-1})$.

Proof. A straightforward argument using Lemma 7(a) gives a constant C_1 such that:

(2) if $n_1, n_2 \geq 3$, $n_1 + n_2 = n + 2$, $p(n_1)p(n_2)/p(n) < C_1/n$.

Let $T \in PT(n, f)$ and $\mu(T) = p(n)^{-1} DP(T, x)|_{x=1}$, $(D = d/dx)$ be the expected distance between T and trees in $PT(n)$.

By Theorem 1,

$$DP(T, x)|_{x=1} = P(T/e, 1) + DP(T/e, x)|_{x=1} - 2P(T_1, 1)P(T_2, 1).$$

Hence, $\mu(T) = 1 + \mu(T/e) - 2p(n_1)p(n_2)/p(n)$, where $T_1 \in PT(n_1)$, $T_2 \in PT(n_2)$. By contracting T to $T_n \in PT(n, 0)$, (2) gives

$$\mu(T) = f + \mu(T_n) - E(T) \quad \text{where } 0 < E(T) < 2C_1 f/n. \tag{3}$$

Averaging $\mu(T)$ over $PT(n)$, and dividing by $(2n-6)$ to obtain $\mu(n)$, we obtain

$$\mu(n) = (p(n)^{-1} \sum_f f\,|PT(n, f)|) + (\mu(T_n) - \epsilon(n))/(2n-6) \tag{4}$$

with $0 < \epsilon(n) < 2C_1$.

Now the trees at distance f from any $T_n \in PT(n, 0)$ are precisely those with f internal edges. Thus,

$$\mu(T_n) = p(n)^{-1} \sum_{f \geq 0} f\,|PT(n, f)|, \qquad \mu(n) = (2\mu(T_n) - \epsilon(n))/(2n-6). \tag{5}$$

Result (a) now follows by Lemma 7.

The variance $\sigma^2(n)$ of the normalized distance is the average value of

$$((d(T, T'))/(2n-6) - \mu(n))^2$$

over all pairs $T, T' \in PT(n)$. Thus,

$$\begin{aligned} \sigma^2(n) &= \Big(\sum_{T \in PT(n)} \sum_k k^2 P_k(T) \Big) \Big/ p(n)(2n-6)^2 - \mu^2(n) \\ &= \Big(\sum_{T \in PT(n)} D^2 P(T, x)|_{x=1} + \sum_{T \in PT(n)} DP(T, x)|_{x=1} \Big) \Big/ p(n)(2n-6)^2 - \mu^2(n). \end{aligned} \tag{6}$$

From Theorem 1,

$$\begin{aligned} D^2 P(T, x)|_{x=1} &= D^2 P(T/e, x)|_{x=1} + 2DP(T/e, x)|_{x=1} \\ &\quad - 4D((P(T_1, x))P(T_2, x))|_{x=1} - 2P(T_1, 1)P(T_2, 1). \end{aligned} \tag{7}$$

Now, $p(n)^{-1} D((P(T_1, x))P(T_2, x))|_{x=1} \leq 2(2n-8)p(n_1)p(n_2)/p(n)$ (where $T_1 \in PT(n_1)$, $T_2 \in PT(n_2)$) since for $T' \in PT(m, f)$, $DP(T', x)|_{x=1}$ is clearly bounded above by $(m + f - 3)p(m) \leq (2m-6)p(m)$.

Let $T \in PT(n, f)$. Dividing (7) by $p(n)$, we obtain $p(n)^{-1} D^2 P(T, x)|_{x=1} = p(n)^{-1} D^2 P(T/e, x)|_{x=1} + 2\mu(T/e) - c(T)$, with $0 < c(T) < \epsilon(n) = O(1)$ by (2). Reducing T to $T_n \in PT(n, 0)$, and using (3), we have

$$\begin{aligned} p(n)^{-1} D^2 P(T, x)|_{x=1} &= p(n)^{-1} D^2 P(T_n, x)|_{x=1} \\ &\quad + 2\Big(\sum_{0 \leq f' \leq f-1} (f' + \mu(T_n)) - c_1(T) \Big) \end{aligned} \tag{8}$$

$$\text{with } 0 < c_1(T) < c_1(n) = O(n).$$

Now, $D^2 P(T_n, x)|_{x=1} = \sum_f f(f-1)|PT(n, f)|$, while the second term in (8) is $f(f-1) + 2f\mu(T_n)$. Averaging (8) over $PT(n)$, we obtain

$$2p(n)^{-1} \sum_f f(f-1)|PT(n, f)| + 2\mu^2(T_n) - c_2(n) \quad \text{with } c_2(n) = O(n).$$

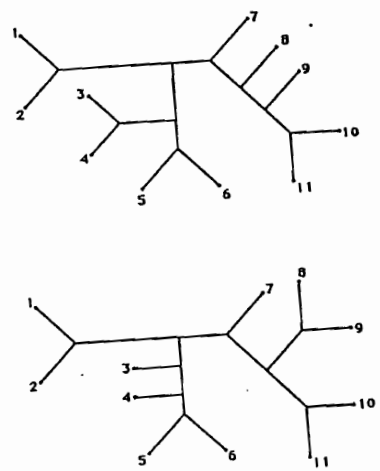

FIG. 3

Substituting this into (6), and noting that $\sum_{T \in PT(n)} DP(T, x)|_{x=1}/p(n)(2n-6) = \mu(n)/(2n-6)$, and $2\mu^2(T_n)/(2n-6)^2 - 0.5\mu(n) = O(n^{-1})$ (from (5)), we have $\sigma^2(n) = 2(\sum_f f(f-1)|PT(n, f)|)/(2n-6)^2 - 0.5\mu^2(n) + c_3(n)$, with $c_3(n) = O(n^{-1})$. The result now follows from part (a) and Lemma 7.

Remark 3. By Chebyshev's inequality, Theorem 5 shows that for any number k, the probability two trees in $PT(n)$ have k or less equivalent edges tends to zero as n becomes large. This is in contrast to the binary case for which most trees are a maximal distance apart.

For $T \in BPT(n)$ an analysis similar to the first part of Theorem 5 can be applied to $v(T)$—the expected distance between T and trees in $BPT(n)$. Differentiating $Q(T, x)$ and setting $x = 1$, we obtain $v(T) = v(T/e) + 1 - 2b(n_1)b(n_2)/b(n)$. By induction, $v(T) = v(T_n) + (n-3) - 2r(1, T)/b(n)$, for $T_n \in PT(n, 0)$.

Now $Q(T_n, x) = b(n)x^{n-3}$ so that $v(T_n) = n-3$. Thus, $v(T)/(2n-6) = 1 - 2r(1, T)/b(n)(2n-6)$. As in Example 3 $r(1, T) \leq (n-3)b(n-1)$, giving $v(T)/(2n-6) \geq 1 - 1/(2n-5)$. In particular, $v(T)/(2n-6) \to 1$ as $n \to \infty$.

Remark 4. The expected distance of binary trees from a given binary tree $T \in BPT(n)$ does not characterize T (up to topological equivalence) in $BPT(n)$. Indeed a counting argument shows that for all integers $k \geq 1$, there exists a positive integer n and a set $S \subset BPT(n)$ of k topologically distinct trees, on which $v(T)$ is constant. For $k = 2$ this is realized for $n = 11$, with the two trees given in Fig. 3. Although $v(T)$ does not characterize the topology of T, it is not known whether or not $Q(T, x)$ (equivalently $R(T, x)$ by Lemma 2) does.

REFERENCES

[1] I. ANDERSON, *A First Course in Combinatorial Mathematics*, Clarendon Press, Oxford, 1974.
[2] M. BOURQUE, *Arbres de Steiner et reseaux dont varie l'emplagement de certains sommets*, Ph.D. thesis, Department d'Informatique et de Recherche Operationnelle, Universite de Montreal, Montreal, Quebec, Canada, 1978.

[3] W. H. E. DAY, *Distribution of distances between pairs of classifications*, in Numerical Taxonomy, J. Felsenstein, ed., Springer-Verlag, Berlin, Heidelberg, 1983, pp. 127–131.
[4] ———, *Optimal algorithms for comparing trees with labeled leaves*, J. Classification, 2 (1985), pp. 7–28.
[5] J. FELSENSTEIN, *The number of evolutionary trees*, Syst. Zool., 27 (1978), pp. 27–33.
[6] I. P. GOULDEN AND D. M. JACKSON, *Combinatorial Enumeration*, John Wiley, New York, 1983.
[7] M. D. HENDY AND D. PENNY, *Branch and bound algorithms to determine minimal evolutionary trees*, Math. Biosci., 59 (1982), pp. 277–290.
[8] M. D. HENDY, C. H. C. LITTLE, AND D. PENNY, *Comparing trees with pendant vertices labelled*, SIAM J. Appl. Math., 44 (1984), pp. 1054–1065.
[9] E. MARCZEWSKI AND H. STEINHAUS, *On a certain distance of sets and the corresponding distance of functions*, Colloq. Math., 6 (1958), pp. 319–327.
[10] 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), pp. 197–200.
[11] D. PENNY AND M. D. HENDY, *The use of tree comparison metrics*, Syst. Zool., 34 (1985), pp. 75–82.
[12] F. RESTLE, *A metric and an ordering on sets*, Psychometrica, 24 (1959), pp. 207–220.
[13] D. ROBINSON AND L. R. FOULDS, *Comparison of phylogenetic trees*, Math. Biosci., 53 (1981), pp. 131–147.
[14] ———, *Enumeration of phylogenetic trees without points of degree two*, Ars Combin., 17A (1984), pp. 169–183.