# Comparison of Phylogenetic Trees

D. F. ROBINSON
*Mathematics Department, University of Canterbury, Christchurch, New Zealand*

AND

L. R. FOULDS
*Operations Research, University of Canterbury, Christchurch, New Zealand*

*Received 1 October 1979; revised 19 August 1980*

ABSTRACT

A metric on general phylogenetic trees is presented. This extends the work of most previous authors, who constructed metrics for binary trees. The metric presented in this paper makes possible the comparison of the many nonbinary phylogenetic trees appearing in the literature. This provides an objective procedure for comparing the different methods for constructing phylogenetic trees. The metric is based on elementary operations which transform one tree into another. Various results obtained in applying these operations are given. They enable the distance between any pair of trees to be calculated efficiently. This generalizes previous work by Bourque to the case where interior vertices can be labeled, and labels may contain more than one element or may be empty.

## 1. INTRODUCTION

It has been postulated that existing biological species have been linked in the past by common ancestors. A diagram showing these links is called a phylogenetic tree. In the past decade or so a number of such trees have been constructed using protein sequence data; for example: Fitch and Margoliash [3], Jardine and Sibson [7], (1971), Moore et al. [8], Sokal and Sneath [13], Waterman et al. [14], and Foulds et al. [5]. However it is evident that different methods often produce different trees when applied to the same data. It is important in comparing different methods to have an objective measure of how similar these different trees are. This problem of comparison has been studied by Robinson [11], Dobson [1], Rohlf [10], and Waterman and Smith [14]. However, all of these methods were developed for binary trees. The purpose of the present paper is to present a comparison method suitable for general trees, that is, trees whose internal points




52 Vanderbilt Ave., New York, NY 10017 0025-5564/81/01013117$02.50

have arbitrary degree. Many phylogenetic trees appearing in the literature are of this form, e.g. [5].

There are two different approaches to comparing phylogenetic trees. One can view them as weighted trees where each line has a weight equal to the number of mutations between the sequences it connects. These weights can then be taken into account in the comparison method. This has been done by Robinson and Foulds [12]. The second approach is to ignore weights and compare the structure or topology of the trees. This was the line of all the authors previously cited who presented comparison methods for binary trees, and we adopt it in this paper.

We present a metric $d$ on the set of all phylogenetic trees labeled with $n$ species. The metric defines a distance, $d(T_1, T_2)$ between any two trees $T_1, T_2$ in the set. This provides an objective measure for comparing trees produced by different methods. The distance $d(T_1, T_2)$ can be efficiently calculated for any pair of trees.

Bourque [2],[1] in Chapter 3 of his doctoral thesis on the Steiner problem in geometry, discusses operations on trees. He introduces the notion of a space of all tree topologies, the pendant vertices of whose trees are assigned nonempty distinct labels from some given finite set. [In the present paper the labels can be assigned to interior vertices as well and may contain more than one element or be empty.] He defines the "distance" between two topologies to be the smallest number of transformations required to obtain one topology from the other. He presents several metrics and shows that three of them are equivalent and that two of the others provide upper and lower bounds respectively on the distance.

## 2. OPERATIONS ON PHYLOGENETIC TREES

The graph-theoretic concepts used in this paper are explained in [6].

Let $S$ be the given set of $n$ ($> 1$) species.

*DEFINITION*

By a *phylogenetic tree* $T^*$ on $S$ we mean a tree $T$ with points $p_1, \ldots, p_m$, together with a partition of $S$ into disjoint, possibly empty, subsets $S_1, S_2, \ldots, S_m$, so that $S_i$ is assigned as the *label* of $p_i$; each member of $S$ appears in exactly one label. Points of degree 3 or more may have empty labels, but points of degree 1 or 2 must have nonempty labels.

The set of points of $T$ is $P$, and the set of *edges* (*lines* in [6]) is $E$. We then use the notation $T=(P, E)$.

The set of all phylogenetic trees on $S$ is $\gamma_S$.

---

[1]The authors wish to thank one of the referees for bringing to their attention Bourque's work, of which they were unaware when writing the first version of this paper.

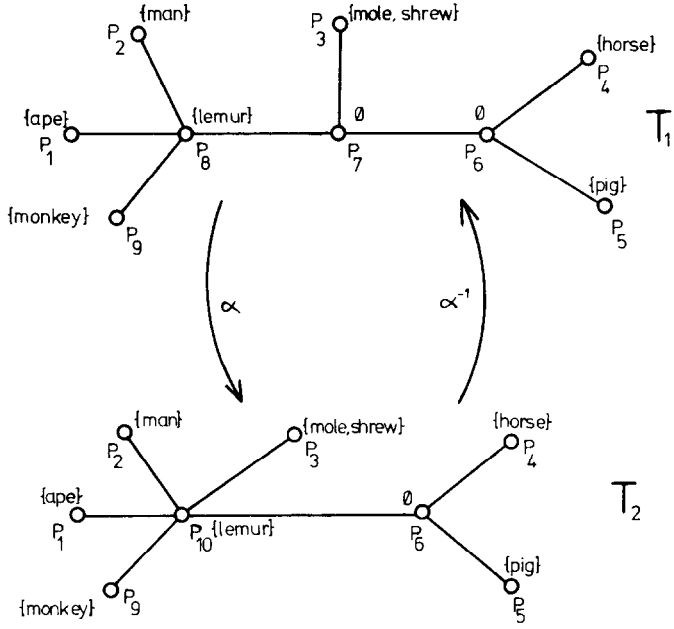


FIG. 1. The application of operations $\alpha$ and $\alpha^{-1}$.

We consider two phylogenetic trees $T_1, T_2$ to be the same in all respects (we use the term *identical*) if their trees $T_1, T_2$ are isomorphic, and the isomorphism also preserves the labeling. That is, if $T_1=(P_1, E_1)$ and $T_2=(P_2, E_2)$ and the labels are $(S_1^1, S_2^1, \ldots, S_m^1)$ on $T_1$ and $(S_1^2, S_2^2, \ldots, S_m^2)$ on $T_2$, there is a one-one correspondence $H: P_1 \to P_2$ such that

$$\text{if} \quad p_i p_j \in E_1 \quad \text{then} \quad h(p_i)h(p_j) \in E_2, \text{ and conversely,}$$

and if $h(p_i)=q_j$ then $S_i^1 = S_j^2$. We shall not normally distinguish in the notation between the phylogenetic tree $T^*$ and the underlying tree $T$, using the symbol $T$ for both.

In Fig. 1 are shown two examples of phylogenetic trees in which $S=\{\text{man, ape, monkey, lemur, shrew, mole, horse, pig}\}$.

We now describe two operations, $\alpha$ and $\alpha^{-1}$, which may be applied to a tree in $\gamma_S$ to construct new trees on the same species set.

*OPERATION* $\alpha$ ("Contraction" of Bourque)

Let $T_1$ be a phylogenetic tree on set $S$, and let $p_r p_s$ be an edge of $T_1$. Then we can form a new tree $T_2$ on $S$ by "shrinking" $p_r p_s$ to a single point, the label of the new point being the union of the labels of $p_r$ and $p_s$. An

example is given in Fig. 1, in which $p_7p_8$ is collapsed to form the new point $p_{10}$. Formally,

Let $T_1=(P_1, E_1)\in\gamma_S$, let $P_1=\{p_1, p_2, \ldots, p_m\}$, and let $p_rp_s\in E_1$. Let

$$E^r=\{p_rp_i: p_rp_i\in E_1, i\neq s\},$$
$$E^s=\{p_sp_j: p_sp_j\in E_1, j\neq r\},$$
$$E^{rs}=\{p_rp_s\}.$$

Then we define

$$P_2=(P_1\backslash\{p_r, p_s\})\cup\{p_{m+1}\}$$
$$E_2=[E_1\backslash(E^r\cup E^s\cup E^{rs})]\cup\{p_{m+1}p_i: p_rp_i\in E^r\}$$
$$\cup\{p_{m+1}p_i: p_sp_i\in E^s\}.$$

If $S_{i1}$ is the label of $p_i$ in $T_1$, and $S_{i2}$ the label of $p_i$ in $T_2$, then

$$S_{i2}=S_{i1} \quad \text{if} \quad i\in P_1\cap P_2,$$
$$S_{(m+1)2}=S_{r1}\cup S_{s1}.$$

To indicate that $T_2$ is obtained from $T_1$ by shrinking the edge $p_rp_s$ we write

$$T_2=\alpha(T_1, p_rp_s).$$

Given a tree $T\in\gamma_S$ with $m$ points, $T$ has $m-1$ edges, each of which produces a different tree. Hence $m-1$ trees are obtainable by single $\alpha$-operations acting on $T$.

*OPERATION* $\alpha^{-1}$ ("Decontraction" of Bourque)

We also wish to reverse this operation, by an operation $\alpha^{-1}$, which takes a point of a tree and divides it into two parts, the new points being joined by an edge. In Fig. 1 the upper tree is formed by the lower by replacing the point $p_{10}$ with two points $p_7$ and $p_8$.

This operation is not defined completely by specifying which point is to be split. The edges incident with the chosen point $p_k$ may be partitioned arbitrarily between the two new points $p_{m+1}$ and $p_{m+2}$, and the label of $p_k$ partitioned arbitrarily between $p_{m+1}$ and $p_{m+2}$, so long as each receives at least one species if its degree is 2 or 1. [Since deg $(p_{m+1})+$deg $(p_{m+2})=$deg $(p_k)+2$, they cannot both have degree 1 or 2 unless $p_k$ has degree 1 or 2.]

A number of results concerning these operations will now be presented. It will be assumed throughout that $|S|=n$.

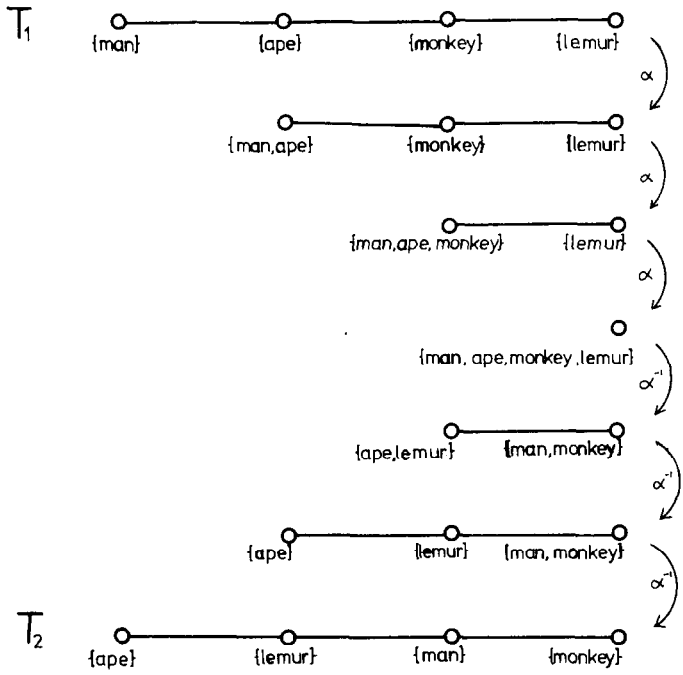


FIG. 2. A sequence of $\alpha$- and $\alpha^{-1}$-operations.

$U_S$ is used to denote a tree in $\gamma_S$ with $E=\varnothing$ and just a single point, which is labeled with $S$. $U_S$ can be obtained from any tree $T\in\gamma_S$ by a sequence of $m-1$ $\alpha$-operations, where $T$ has $m$ points.

Now suppose $T_1$ and $T_2\in\gamma_S$. Then we can convert $T_1$ into $U_S$ by a sequence of $m_1-1$ $\alpha$-operations, and $T_2$ into $U_S$ by a sequence of $m_2-1$ $\alpha$-operations. By reversing the latter sequence we can convert $U_S$ into $T_2$ by a sequence of $m_2-1$ $\alpha^{-1}$-operations. Combining the two, we can convert $T_1$ to $T_2$ by a sequence of at most $m_1+m_2-2$ operations of the two types, via $U_S$. It may of course be possible to reach $T_2$ from $T_1$ by a shorter route, but the number of steps must be at least $|m_1-m_2|$, the difference in the number of edges between the two trees.

Figure 2 shows such a sequence of operations in a case in which the upper bound is achieved.

## 3. A METRIC ON PHYLOGENETIC TREES

If $T_1, T_2\in\gamma_S$, then the minimum number of applications of operations of either type ($\alpha$ or $\alpha^{-1}$) necessary to convert $T_1$ into $T_2$ is denoted by

$d(T_1, T_2)$. It can easily be shown that $d$ has the following properties:

$d(T_1, T_2) > 0$ if $T_1, T_2 \in \gamma_S$, $T_1$ not identical to $T_2$,
$d(T_1, T_2) = 0$ if $T_1$ is identical to $T_2$,
$d(T_1, T_2) = d(T_2, T_1)$, $T_1, T_2 \in \gamma_S$,
$d(T_1, T_3) \leqslant d(T_1, T_2) + d(T_2, T_3)$, $T_1, T_2, T_3 \in \gamma_S$.

Hence $d$ is a metric.

*THEOREM 1*

*If $T \in \gamma_S$ and $|S| = n > 1$, then $T$ has at most $2n-2$ points.*

*Proof.*

(a) If $T$ is a tree in which some point $u$ has a label with two or more members, then $u$ may be split by an $\alpha^{-1}$-operation to provide a new tree $T'$ with one more point, the label of $u$ being partitioned into two nonempty sets.

(b) If $T$ is a tree in which some point $u$ with nonempty label has degree greater than 1, then $u$ may be split by an $\alpha^{-1}$-operation to provide a new tree $T'$ with one more point, the new point having degree 1 and carrying the label of $u$.

(c) If $T$ is a tree in which some point $u$ has degree greater than 3, then $u$ may be split by an $\alpha^{-1}$-operation to provide a new tree $T'$ with one more point, both new points having degree at least 3. The label of $u$ may be shared between them in any way.

Thus a tree $T^*$ with given $S$ and maximal number of points consists of $n$ points of degree 1, each with a singleton label, and a number of points of degree 3 with empty labels. Let this number be $q$.

By relating $e$ the number of edges to the degree,

$$2e = 3q + n.$$

On the other hand, in any tree the number of edges is one less than the number of points:

$$e = n + q - 1,$$

so

$$q = n - 2.$$

Hence $T^*$ has $2n-2$ points, and for any tree $T \in \gamma_S$, $T$ has at most $2n-2$ points. ■

Consider a phylogenetic tree $T = (P, E) \in \gamma_S$. The removal of an edge $e$ creates two subtrees $T_e'$ and $T_e''$ of $T$. Let the union of all the labels of points in $T_e'$ be $S_e'$, and the union of all the labels of points in $T_e''$ be $S_e''$. Since both $T_e'$ and $T_e''$ contain pendant points of $T$, $S_e'$ and $S_e''$ are both nonempty.

*THEOREM 2*

*If $e_1, e_2$ are two edges in $T=(P, E)$, then $S$ may be partitioned into three sets $X, Y, Z$ such that with appropriate assignments,*

$$S'_{e_1}=X, \qquad S''_{e_1}=Y\cup Z,$$
$$S'_{e_2}=Y, \qquad S''_{e_2}=X\cup Z.$$

*Proof.* Suppose $e_1$ and $e_2$ deleted from $T$. Then $T$ falls into three components $A, B, C$. In general $e_1=ab$ and $e_2=cd$ with $a, b, c, d$ all different. Then as at least one of the four belongs to each component, some component contains two of them (arising from different edges) and the others one each. Suppose $A$ contains $a$, $B$ contains $d$, and $C$ contains $b$ and $c$. It is however possible for $e_1$ and $e_2$ to have a point in common. When this happens we let $e_1=ab$ with $a\in A$ and $e_2=bd$, with $d\in B$ as before, and $b\in C$.

Let $X, Y, Z$ be the sets of species associated with points in $A$, $B$ and $C$ respectively.

Then we may set

$$S'_{e_1}=X, \qquad S''_{e_1}=Y\cup Z,$$
$$S'_{e_2}=Y, \qquad S''_{e_2}=X\cup Z,$$

as required. ∎

For a given tree $T\in\gamma_S$ with edge set $E$, we define a function $f$ from $E$ to the set $Z_S$ of all partitions of $S$ into two nonempty subsets by

$$f(e)=\{S'_e, S''_e\}$$

and call $f$ the *partitioning function* of $T$. The deletion of $e$ would create subtrees with species sets $S'_e$ and $S''_e$ respectively. Bourque [2, p. 55] introduces this concept of label partitioning and discusses a number of its properties complementing those given here.

*THEOREM 3*

*If $e_1, e_2$ are edges in $T=(P, E)$ such that*

$$\{S'_{e_1}, S''_{e_1}\}=\{S'_{e_2}, S''_{e_2}\},$$

*then*

$$e_1=e_2.$$

*That is, $f$ is one-to-one.*

*Proof.* Assume $e_1 \neq e_2$. Then by Theorem 2 we may put

$$S'_{e_1} = X, \quad S'_{e_2} = Y, \quad S''_{e_1} = Y \cup Z, \quad S''_{e_2} = X \cup Z.$$

The components $A$ and $B$, each being cut off by the deletion of a single edge, must each contain a pendant point of $T$. The corresponding labels must be nonempty; hence $X \neq Y$. Thus the sets must be paired:

$$S'_{e_1} = S''_{e_2} \quad \text{and} \quad S''_{e_1} = S'_{e_2},$$

so that

$$Z = \varnothing. \tag{*}$$

Hence every point in $C$ has an empty label. Therefore every point in $C$ has degree 3 or more in $T$. But as $C$ is a tree, there must be at least two points pendant in $C$. None of these can be pendant in $T$, for then they would have nonempty labels. Hence $C$ can have as pendant points only the point(s) $b$ and $c$ incident with $e_1$ and $e_2$. As they are pendant in $C$, such points have degree 2 in $T$. Thus they must have nonempty labels. This contradicts (*). Hence the hypothesis that $e_1 \neq e_2$ is false. ■

Consider now $T_1 = (P_1, E_1)$ and $T_2 = (P_2, E_2)$ in $\gamma_S$ with partitioning functions $f_1, f_2$ respectively.

*DEFINITION*

Edges $e_1 \in E_1$ and $e_2 \in E_2$ are said to be *matched* if and only if

$$f_1(e_1) = f_2(e_2).$$

For example, in Fig. 1 edges $p_6p_7$ and $p_6p_{10}$ in $T_1$ and $T_2$ respectively are matched.

Because each edge $e_1 \in E$ creates a unique partition of $S$, $e_1$ will be matched to no more than one edge of $T_2$. Also the matching relation is symmetric; hence there will be a one-one correspondence between matched edges in $T_1$ and $T_2$.

Define the sets

$$E'_1 = \{e_1 \in E_1 : \exists e_2 \in E_2 \text{ s.t. } f_1(e_1) = f_2(e_2)\},$$

$$E'_2 = \{e_2 \in E_2 : \exists e_1 \in E_1 \text{ s.t. } f_2(e_2) = f_1(e_1)\}.$$

The edges in $E_1 \backslash E'_1$ and $E_2 \backslash E'_2$ create partitions of $S$ which do not correspond to edges in $T_2$ and $T_1$ respectively. It is possible for $E'_1$ and $E'_2$ to be empty.

*THEOREM 4*

*$T_1=(P_1, E_1)$, $T_2=(P_2, E_2)\in\gamma_S$ are identical if and only if there is a one-one correspondence $h: E_1\to E_2$ such that $e\in E_1\Rightarrow e$ and $h(e)$ are matched.*

*Proof.* $\Rightarrow$: As $T_1$ and $T_2$ are identical, there exist one-one correspondences:

$$p: E_1\to E_2,$$
$$q: P_1\to P_2,$$

such that

$$uv\in E_1 \quad\Rightarrow\quad p(uv)=q(u)q(v) \tag{1}$$

and

$$w\in P_1 \quad\Rightarrow\quad S_w=S_{q(w)}, \tag{2}$$

where $S_i$ is the subset of $S$ assigned to point $i$. Consider any edge $uv\in E_1$. Let the partition of $S$ created by $uv$ be $\{S', S''\}$. As $T_1$ and $T_2$ are identical, by (1) and (2) each subgraph of $T_2$ created by the removal of $p(uv)$ is identical with a subgraph of $T_1$ created by the removal of $uv$ from $T$. Hence $uv$ and $p(uv)$ are matched. Hence

$$h=p$$

is the desired one-one correspondence.

$\Leftarrow$: We observe first that if $h$ exists, then $T_1$ and $T_2$ have the same number of edges.

We now restate Theorem 3. If $e=uv$, we may, instead of writing $\{S'_e, S''_e\}$, write $\{S^u_e, S^v_e\}$. As $S^u_e$ and $S^v_e$ are complementary, any nonempty subset $W$ of $S$ corresponds to at most one edge-point pair $(e, u)$ of $T_1$ by

$$W=S^u_e.$$

The given one-one correspondence $h$ can therefore be converted into a one-one correspondence between edge-point pairs, where the edge and point must be incident. We may write

$$k(e, u)=(h(e), w),$$

and then define $v$ and $x$ by

$$e=uv, \qquad h(e)=wx.$$

It will then follow that

$$k(e, v)=(h(e), x).$$

We must next show that $k$ acts consistently at points. That is, if $e'$ is another edge of $T_1$ incident with $u$, and $k(e', u)=(h(e'), w')$, then $w=w'$. The alternative is that $h(e')=wx'$ and $k(e', u)=(h(e'), x')$.

But looking at Theorem 2 for $e$ and $e'$ in $T_1$, we may set

$$S_e^u = Y \cup Z, \qquad S_e^v = X, \qquad S_e^{u'} = X \cup Y,$$

and $k$ then implies that in $T_2$

$$S_{h(e)}^w = Y \cup Z, \qquad S_{h(e)}^x = X, \qquad S_{h(e')}^{x'} = X \cup Y,$$

so $S_{h(e')}^w = Z$, which conflicts with the form established in Theorem 2.

The function $k$ is thus a one-one correspondence consistent on both edges and points, so that we may define one-one correspondences

$$p: E_1 \to E_2, \qquad q: P_1 \to P_2$$

with $k(e, u)=(p(e), q(u))$ and

$$p(uv)=q(u)q(v).$$

The isomorphism is thus established. That $q$ preserves labels follows from the relationship that the label $S_u$ of $u$ is the intersection over all edges $e$ incident with $u$ of $S_e^u$.

Then $T_1$ and $T_2$ are identical. ■

Theorem 4 implies that in order to convert $T_1$ to $T_2$, operation $\alpha$ must be used to remove all the unmatched edges in $E_1$, and $\alpha^{-1}$ must be used to create all the unmatched edges in $E_2$. Thus

$$d(T_1, T_2) \geqslant |E_1 \backslash E_1'| + |E_2 \backslash E_2'|. \tag{3}$$

*THEOREM 5*

*If* $T_1, T_2 \in \gamma_S$ *then* $d(T_1, T_2)=|E_1 \backslash E_1'| + |E_2 \backslash E_2'|$.

*Proof.* Consider the tree $\bar{T}_1$ ($\bar{T}_2$) obtained from $T_1(T_2)$ by removing all the edges of $E_1 \backslash E_1'$ ($E_2 \backslash E_2'$) with $\alpha$-operations. $\bar{T}_1$ ($\bar{T}_2$) has edges $E'$ ($E_2'$). By the nature of the $\alpha$-operation each edge in $\bar{T}_1$ ($\bar{T}_2$) will partition $S$ in the same way as the corresponding edge in $T_1$ ($T_2$).

If $E_1'=E_2'=\varnothing$, then

$$\bar{T}_1 = \bar{T}_2 = U_S, \tag{4}$$

and $T_1$ and $T_2$ are identical. However, if (4) does not hold, there is, by definition, a one-one correspondence between $E_1'$ and $E_2'$, the edge sets of $\bar{T}_1$ and $\bar{T}_2$. Hence by Theorem 4, $\bar{T}_1$ and $\bar{T}_2$ are identical and it is possible to

convert $T_1$ into $T_2$ by $|E_1 \backslash E_1'|$ $\alpha$-operations followed by $|E_2 \backslash E_2'|$ $\alpha^{-1}$-operations, the reverse of those $\alpha$-operations required to transform $T_2$ into $\bar{T}_2$. Therefore

$$d(T_1, T_2) \leqslant |E_1 \backslash E_1'| + |E_2 \backslash E_2'|.$$

So (3) implies the result. ■

The significance of Theorem 5 is that $d(T_1, T_2)$ can be calculated without having to find the actual sequence of $\alpha$- and $\alpha^{-1}$-operations contributing to $d(T_1, T_2)$. The identical trees $\bar{T}_1$ and $\bar{T}_2$ are denoted by $T_1 \wedge T_2$.

*COROLLARY*

*For $T_1, T_2 \in \gamma_S$ there exists a unique tree, $T_1 \wedge T_2 \in \gamma_S$ such that:*

(1) *$T_1 \wedge T_2$ can be obtained from both $T_1$ and $T_2$ by $|E_1 \backslash E_1'|$ and $|E_2 \backslash E_2'|$ $\alpha$-operations respectively.*

(2) $d(T_1, T_2) = d(T_1, T_1 \wedge T_2) + d(T_1 \wedge T_2, T_2)$

(3) *If*

$$E_2 = E_2',$$

*that is, every edge of $T_2$ is matched to an edge of $T_1$, then*

$$|E_2| = |E_2'| = |E_1'|$$

*and*

$$d(T_1, T_2) = |E_1 \backslash E_1'| = n_1 - n_2,$$

*where $T_1$ and $T_2$ have $n_1$ and $n_2$ points respectively.*

(4) $d(T_1, T_2) = d(T_1, U_S) + d(U_S, T_2) - 2d(T_1 \wedge T_2, U_s)$

This last equation is given in [2, p. 62].

*DEFINITION*

For each positive integer $n$ we define $d_n$ to be the maximum distance between two phylogenetic trees on $n$ species.

*THEOREM 6*

*If $n = 2$, then $d_n = 1$. If $n > 2$, then $d_n = 3n - 6$.*

*Proof.* If $n = 2$, there are only two possible trees, and these are 1 apart.

If $n = 3$ the argument below does not work, but the formula can be verified by constructing all 8 possible phylogenetic trees, and considering the effects of the operations.

We may therefore assume that $n > 3$. We proceed by discovering features of pairs of trees $T_1, T_2$ which are a maximal distance apart.

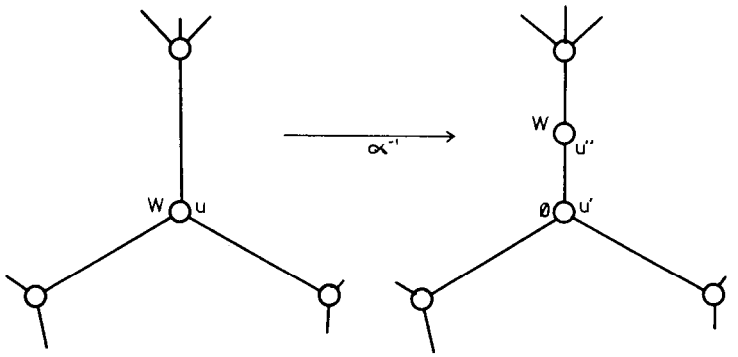


FIG. 3. Creation of an empty label for a point of degree 3 or more.

(a) If $T_1, T_2$ have matched edges, we can construct new trees $T_1', T_2'$ by collapsing the matched edges. The remaining edges are unmatched and

$$d(T_1', T_2') = d(T_1, T_2).$$

We may therefore assume that $T_1$ and $T_2$ have no matched edges.

(b) Suppose $T_1$ has a point $u$ with degree greater than 2 and a nonempty label. Then by an $\alpha^{-1}$-operation $u$ can be split into two points: $u'$, which has the same degree as $u$ and empty label, and $u''$, which has degree 2 and the same label as $u$. This is shown in Fig. 3. The problem to be guarded against is that the new edge might match an edge of $T_2$. It can be shown, using Theorem 2, that if a trial splitting with $u''$ in one edge of $T_1$ yields such a matching, then it suffices to put $u''$ into another edge to avoid a matching. The new tree $T_3$ thus created has one more edge unmatched with $T_2$ than $T_1$ has, so that

$$d(T_3, T_2) = d(T_1, T_2) + 1.$$

Hence if $T_1, T_2$ are at maximal separation, then in both $T_1$ and $T_2$ all points of degree 3 or more have empty labels.

(c) Suppose a point $u$ of tree $T_1$ has degree 2 and its label contains more than one member. Then $u$ may be split into two points $u', u''$ by an $\alpha^{-1}$-operation. Each of the new points has degree 2 and may be assigned a nonempty label by a partition of the label of $u$. This is shown in Fig. 4. If the new edge $u'u''$ should match an edge in $T_2$ it suffices to interchange the labels of these points to avoid such a matching. Following a similar argument to that in (b), we may assume that in $T_1$ and $T_2$ all points of degree 2 are labeled with a single species.

(d) If a pendant point $u$ of the tree $T_1$ has a label containing more than one species, we split $u$ into two new points: a new pendant point $u'$ with

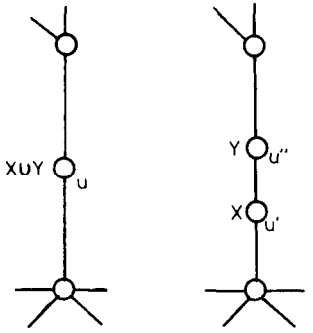


FIG. 4. Splitting a point of degree 2.

label one of the species in the label of $u$, and a point $u''$ with degree 2 and labeled by the rest of the species set of $u$. It may not be possible to avoid matching in this case, but such matching as does occur can be confined to edges pendant in both $T_1$ and $T_2$.

(e) We thus have that $T_1$ and $T_2$ may be assumed to have all labels consisting of a single species for points of degree 1 or 2, and empty for points of higher degree. If some species is assigned to points of degree 2 in both trees, we may split the corresponding point $u$ in one of them into $u'$ of degree 3 and empty label, and $u''$ a pendant point labeled with the species. We may therefore assume that every species forms the label of a pendant point in at least one of the trees. Matching occurs only between pendant edges.

(f) We may therefore count off the species as follows:

$x$ species labeling pendant points in both $T_1$ and $T_2$,
$y$ species labeling pendant points in $T_1$ only,
$z$ species labeling pendant points in $T_2$ only.

We have that

$$x+y+z=n.$$

We also know from Theorem 1 that a tree with $m$ pendant points has $m-2$ points of degree 3. Thus in $T_1$ there are

$x+y$ points of degree 1,
$z$ points of degree 2,
$x+y-2$ points of degree 3.

Then if $T_1$ has $e_1$ edges,

$$2e_1=(x+y)+2z+3(x+y-2),$$
$$e_1=2x+2y+z-3,$$

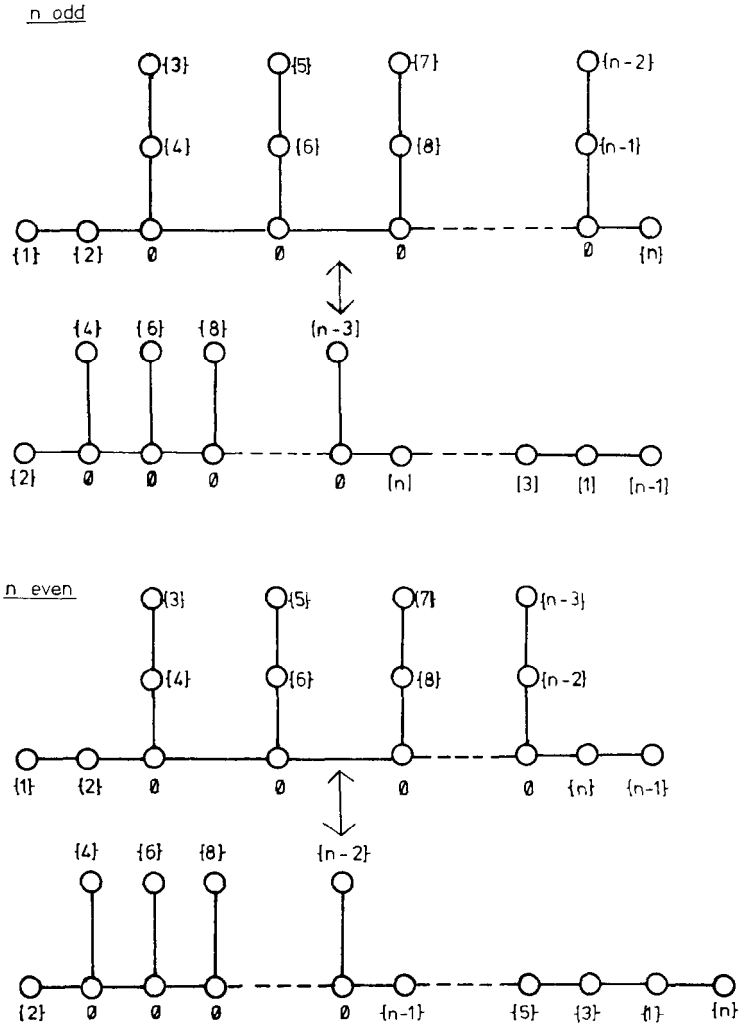


FIG. 5. Phylogenetic trees with maximum separation for $n \geq 5$ species.

of which $e_1 - x$ are unmatched. In the same way $T_2$ has

$$e_2 = 2x + 2z + y - 3$$

edges of which $e_2 - x$ are unmatched. Hence

$$\begin{aligned} d(T_1, T_2) &= (e_1 - x) + (e_2 - x) \\ &= (x + 2y + z - 3) + (x + 2z + y - 3) \\ &= (3x + 3y + 3z) - x - 6 \\ &= 3n - x - 6. \end{aligned}$$

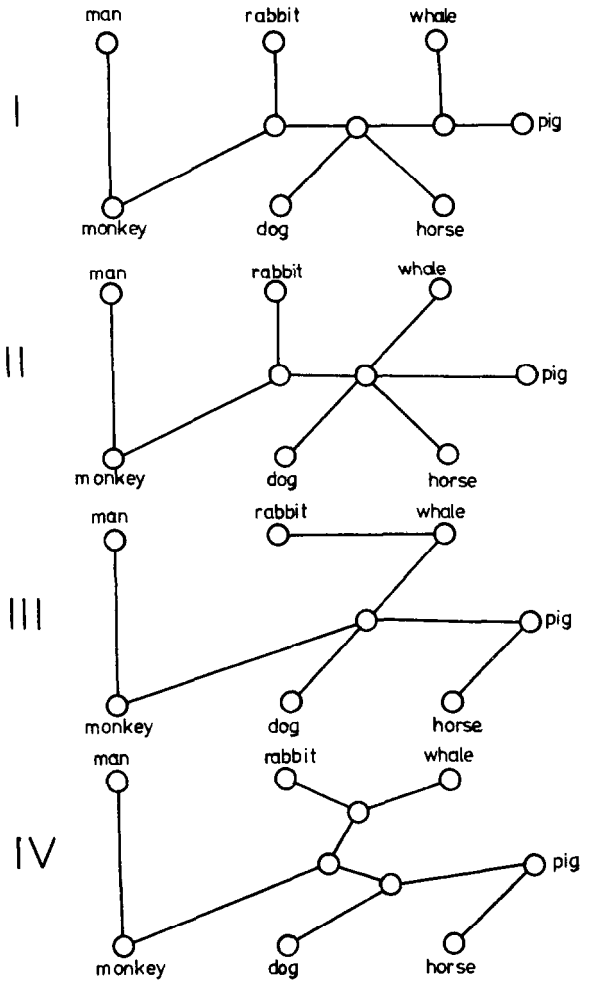


FIG. 6. Four phylogenetic trees.

As $x \geqslant 0$, $d(T_1, T_2) \leqslant 3n-6$ as required.

To see that the bound is exact for $n \geqslant 5$, consider the pairs of phylogenetic trees for $n$ respectively odd and even shown in Fig. 5. No edges are matched, and the number of edges is $3n-6$ in both cases. The case $n=4$ is covered by Fig. 2. ■

## 4. AN EXAMPLE

As an example of the use of this distance consider the four trees in Fig. 6. These are obtained from figures given by:

| | |
|---|---|
| I | Fitch [4] |
| II | Penny [9] |
| III, IV | Foulds et al. [5] |

by suppression of the weights associated with the edges in those papers.

The distances are as follows:

| | I | II | III | IV |
|---|---|---|---|---|
| I | 1 | | | |
| II | 6 | 5 | | |
| III | 6 | 5 | 2 | |
| IV | | | | |

Thus I and II are similar, as are III and IV, but the pairs are relatively dissimilar. On the other hand with seven species the maximum possible distance is 15, so there is still considerable agreement. This agreement could be assessed in terms of the probability of two phylogenetic trees being at most a certain distance apart. This probability depends on the number of phylogenetic trees on a given set of species, and on detailed knowledge of the numbers of trees obtainable from a given tree by single $\alpha$- and $\alpha^{-1}$-operations and by combinations. These are at present unsolved problems.

## 5. CONCLUSION

We have presented a method for expressing the agreement between two phylogenetic trees on the same species. We have established a quick way of calculating the distance by means of testing edges for matching and counting the unmatched edges. The calculation of distance therefore presents no difficulties for practical-sized problems.

By comparing distances it is possible to say that a tree $T_1$ is closer to a tree $T_2$ than it is to $T_3$. We have not yet reached the position of being able to say that agreement between trees is significant at a given probability level.

## REFERENCES

1 A. J. Dobson, Comparing the shapes of trees, in *Combinatorial Mathematics III*, Lecture Notes in Mathematics, 452 (A. Dold and B. Eckmann, Eds.), Springer, Berlin, 1975, pp. 95–100.

2 M. Bourque, Arbres de Steiner et reseaux dont varie l'emplagement de certain sommets, Ph.D. Thesis, Departement l'Informatique et de Recherche Operationelle, Université de Montréal, Montréal, Québec, Canada, Sept. 1978.

3 W. M. Fitch and E. Margoliash, Construction of phylogenetic trees, *Science* 155:279–284 (1967).

4 W. M. Fitch, Is the fixation of observable mutations distributed randomly among the three nucleotide positions of the codon?, *J. Mol. Evol.* 2:128–136 (1973).
5 L. R. Foulds, M. D. Hendy, and David Penny, A graph theoretic approach to the development of optimal phylogenetic trees, *J. Mol. Evol.* 13:127–150 (1979).
6 F. Harary, *Graph Theory*, Addison-Wesley, Reading, Mass., 1969.
7 N. Jardine and R. Sibson, *Mathematical Taxonomy*, Wiley, New York, 1971.
8 G. Moore, M. Goodman, and J. Barnabas, An iterative approach from the standpoint of the additive hypothesis to the dendrogram problem posed by molecular data sets, *J. Theoret. Biol.* 38:423–457 (1973).
9 David Penny, Evolutionary clock: the rate of evolution of rattlesnake cytochrome *c*, *J. Mol. Evol.* 3:179–188 (1974).
10 F. J. Rohlf, Methods of comparing classifications, *Ann. Rev. Ecology and Syst.* 5:101–113 (1974).
11 D. F. Robinson, Comparison of labeled trees with valency three, *J. Combinatorial Theory Ser. B* 11:105–119 (1971).
12 D. F. Robinson and L. R. Foulds, Comparison of weighted labelled trees, in *Combinatorial Mathematics VI*, Lecture Notes in Mathematics 748, Springer, Berlin, 1979, pp. 119–126.
13 R. R. Sokal and P. H. Sneath, *Principles of Numerical Taxonomy*, Freeman, San Francisco, 1963.
14 M. S. Waterman, T. F. Smith, M. Singh, and W. A. Beyer, Additive evolutionary trees, *J. Theoret. Biol.* 64:199–209 (1977).
15 M. S. Waterman and T. F. Smith, On the Similarity of Dendrograms, *J. Theoret. Biol.* 73:789–800 (1978).