The complexity of the median procedure for binary trees

F. R. McMorris[1] and Michael A. Steel[2]

[1] Department of Mathematics, University of Louisville,
Louisville, KY 40292, USA
[2] Department of Mathematics and Statistics, University of Canterbury,
Private Bag, Christchurch, New Zealand

Summary: The median procedure for trees has been nicely characterized in a way that allows it to be efficiently implemented. However, when the problem is restricted to binary trees, we will show that computing the median binary tree is NP-hard. This provides another reason to not always insist that a "consensus tree" be fully resolved.

1. Introduction and Definitions

A *phylogenetic tree* on a label set L is a tree which exactly $|L|$ leaves (vertices of degree one), no vertices of degree two, and each leaf labeled with a distinct element from L. A *binary phylogenetic tree* is a phylogenetic tree in which every non-leaf has degree three. Phylogenetic trees get their name because they are often appropriate models for evolutionary history. However, for simplicity, in this note we will refer to (binary) phylogenetic trees as *(binary) trees.*

Removing an edge from a tree T with leaf set L results in a bipartition $\{A, B\}$ of L, which is called a *split* of T. Let $\sigma(T)$ denote the set of all splits of a tree T. We say that a split $\{A, B\}$ is *nontrivial* if $min\{|A|, |B|\} > 1$, and we let $\widehat{\sigma}(T)$ denote the set of nontrivial splits of T. For two trees T_1 and T_2, the (*partition*, or *symmetric difference*) *distance* between T_1 and T_2 is defined by

$$d(T_1, T_2) = |\ [\sigma(T_1) \cup \sigma(T_2)] - [\sigma(T_1) \cap \sigma(T_2)]\ |$$

This function d, which is indeed a metric, has been analyzed and applied frequently (Steel and Penny (1993)). For example, Penny et al. (1982) propose defining a consensus of a collection P of binary trees as a median tree of P in the space of binary trees with metric d. That is, given a *profile* $P = (T_1, \ldots, T_k)$ of binary trees, a *median binary tree* for P is a binary tree T which minimizes

$$D(T, P) = \sum_{i=1}^{k} d(T, T_i).$$

Let $M(P)$ denote the set of median binary trees of P. (In general, the *median procedure* on an arbitrary finite metric space is the function M with domain the set of all profiles and $M(P) = \{x : D(x, P)\}$ is minimum.)

Note that if we do not insist that T be binary, then minimizing $D(T, P)$ is simple, and $M(P)$ can be nicely characterized (Barthélemy and McMorris (1986)) and constructed in polynomial time. Essentially it is the "majority rule" tree for splits with some additional splits adjoined when k is even. This characterization has an elegant generalization to semilattices in Barthélemy and Janowitz (1991), and is also discussed in Barthélemy and Monjardet (1988).

However, insisting that T be binary changes the character of the problem considerably and it is not immediately clear whether $M(P)$ can even be constructed in polynomial time. In fact, we show that this is unlikely by showing that the following problem is NP-complete (see Garey and Johnson (1979) for definition of the classes P, NP, etc.).

MEDIAN BINARY TREE (MBT)

INSTANCE: A profile $P = (T_1, \ldots, T_k)$ of binary trees on leaf set L; integer ℓ.

QUESTION: Does there exist a binary tree T on L such that $D(T, P) \leq \ell$?

2. Results

We will show MBT is NP-complete by establishing a polynomial transformation to MBT from the problem of determining whether or not a collection of binary qualitative characters has a compatible subset of size at least ℓ (for variable ℓ), which was shown to be NP-complete by Day and Sankoff (1986).

In this latter problem a *binary qualitative character* is just a bipartition of L, and two such characters, $\{A, B\}, \{A', B'\}$ are *compatible* precisely if $\emptyset \in \{A \cap A', A \cap B', B \cap A', B \cap B'\}$. A collection of pairwise compatible binary qualitative characters is said to be *compatible*. Notice, for example, that the splits of a phylogenetic tree form a compatible set. Buneman (1971) established the fundamental converse result: Any compatible collection of splits which, in addition, includes all the trivial bipartitions $\{\{i\}, L - \{i\}\}$ is $\sigma(T)$ for a unique (not necessarily binary) tree T. We can now state the problem in two equivalent forms:

COMPATIBILITY OF BINARY QUALITATIVE CHARACTERS (CBQC)

INSTANCE: A collection $\sum$ of bipartitions of L; an integer ℓ'.

QUESTION: Does there exist a compatible set $\sum_0 \subseteq \sum$ with $|\sum_0| \geq \ell'$?

or, equivalently,

QUESTION': Does there exist a tree T on the leaf set L with $|\sigma(T) \cap \sum| \geq \ell'$?

In our reduction, we take an instance of CBQC and replace each bipartition of L by a carefully chosen collection of binary trees on leaf set $L' \supseteq L$, where $L' - L$ are additional, new leaves, required to allow the desired construction. We then show that the answer to the CBQC question is "yes" for ℓ' if and only if the corresponding answer to MBT is "yes" for a value ℓ determined polynomially by ℓ' and $(|L|, |\sum|)$. Since the transformation is of polynomial time, it follows that if MBT was in P, then so also would CBQC.

The replacement of each bipartition π of L by a set of binary trees on L' is actually a two-step process. Firstly, π is extended to a bipartition π' of L', by associating with each label $j \in L$ a number of new labels, which always "follow j around", i.e., occur in the same set of each bipartition as j. In this way the resulting collection $\sum'$ of bipartitions of L' has a pairwise compatible subset $\sum'_0 \subseteq \sum', |\sum'_0| \geq \ell$, if and only if a pairwise compatible subset $\sum_0 \subseteq \sum$ exists with $|\sum_0| \geq \ell$. The second step is to replace each bipartition π' of L' by a collection $\mathcal{B}_{\pi'}$ of binary trees with leaf set L' such that

(1) each tree in $\mathcal{B}'_\pi$ has π' as a split

(2) no two trees in $\mathcal{B}'_\pi$ have any other nontrivial split in common.

In order for this association to be useful it is necessary that $|\mathcal{B}_\pi{}'|$ grow sufficiently fast compared to $|L'|$. Thus we first require the following lemma.

Lemma 1: Let $f(s)$ denote the maximum possible number of binary trees on a common leaf set S of size s which have no nontrivial splits in common. Then, $f(s) > \lfloor cs^2 \rfloor$, for some $c > 0$. Furthermore, a set of $\lfloor cs^2 \rfloor$ binary trees can be constructed in polynomial (in s) time.

Proof: We will apply a version of Turan's theorem to the following graph $G = (V, E)$. The vertex set V consists of linear ("caterpillar") binary trees, leaf labelled by S (as in Figure 1). An edge exists between two vertices precisely if the associated trees have at least one shared non-trivial split. Since a linear tree has exactly three (2-fold) symmetrics,

$$|V| = s!/(2!)^3$$

Furthermore, for each $T \in V$, there exists at most $2 \times \frac{r!}{2!} \times \frac{(s-r)!}{2!}$ other such trees which share at least one of the (at most) two splits $\{A, B\}$ of T for which $r = min\{|A|, |B|\} > 1$. Thus the degree of each vertex in G is at most

$$\frac{1}{2!} \sum_{r=2}^{s/2} (s-r)!r!$$

and so $|E| \leq \frac{s!}{(2!)^2} \sum_{r=2}^{s/2} (s-r)!r!$

Now, a simple version of Turan's theorem (Spencer (1987)) states that if $|V| = n$ and $|E| = e$, then G has at least $\frac{n^2}{(2e+n)}$ vertices which form an independent set (i.e., no two vertices are adjacent). Applying this in the present context, G has at least

$$\frac{[s!/(2!)^3]^2}{(\frac{s!}{2!}[(s-2)!2! + (s-3)!3! + \cdots] + s!/(2!)^3)} \geq cs^2$$

vertices (for some $c > 0$), no two of which are adjacent. Since vertices are trees and adjacent means to share a nontrivial split, the proof is complete. We leave the efficient construction of such a set of trees to the interested reader. □

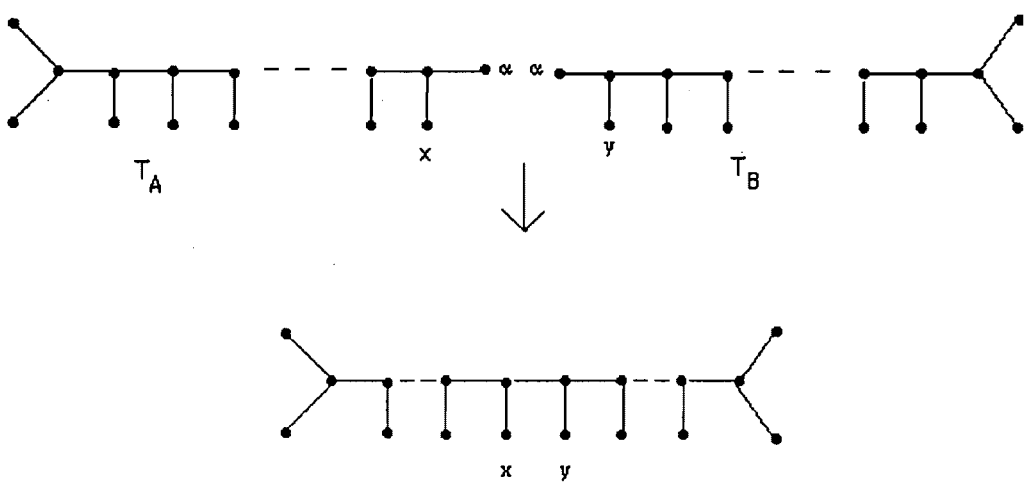

Figure 1

Lemma 2: If $P = (T_1, \ldots, T_k)$, where each T_i is a binary tree on the same set of n leaves, then

$$D(T,P) = (2n-6)k - 2 \sum_{\pi \in \widehat{\sigma}(T)} f_\pi$$

where $f_\pi = \mid \{i : \pi \in \widehat{\sigma}(T_i)\} \mid$.

Proof: First note that the T_i's each have n trivial splits and each T_i has $n-3$ nontrivial splits. Thus, from the definition of d, $d(T,T_i) = 2n - 6 - 2 \mid \widehat{\sigma}(T) \cap \widehat{\sigma}(T_i) \mid$ and

$$D(T,P) = (2n-6)k - 2\sum_{i=1}^{k} \mid \widehat{\sigma}(T) \cap \widehat{\sigma}(T_i) \mid = (2n-6)k - 2 \sum_{\pi \in \widehat{\sigma}(T_i)} f_\pi \quad \square$$

We can now state our main result.

Theorem: MBT is NP-complete.

Proof: First note that MBT is in NP, since if we are given any binary tree T we can efficiently calculate $\sum_{i=1}^{k} d(T,T_i)$ (see Day (1985)).

Now suppose we are given an instance of CBQC, i.e., a collection $\sum$ of bipartitions of L and an integer ℓ. Let $L' = L \dot{\cup} \{x_i^j : j = 1, \ldots, t\}$ where x_i^j are new labels not found in L. ($\dot{\cup}$ denotes disjoint union)

Define a collection $\sum'$ of bipartitions of L' by placing $\{x_i^j : j = 1, \ldots, t\}$ in the same set of each bipartition in $\sum$, as i. That is, $\{A,B\} \in \sum'$ if and only if

$$\{A \cap L, B \cap L\} \in \sum \tag{1}$$

$$A = A \cap L \dot{\cup} \{x_i^j : i \in A \cap L, j = 1, \ldots t\} \tag{2}$$

$$B = B \cap L \cup \{x_i^j : i \in B \cap L, j = 1, \ldots t\}$$

To each $\{A,B\} \in \sum'$ we then construct a collection of linear binary trees on L' as follows: Take $\alpha \notin L'$, and set

$$A_\alpha = A \dot{\cup} \{\alpha\}, \quad B_\alpha = B \dot{\cup} \{\alpha\}$$

Applying Lemma 1, construct efficiently a set $\mathcal{J}_A$ of $\lfloor ct^2 \rfloor$ linear binary trees on leaf set A_α which share no nontrivial splits. Similarly, for B_α, an analogous set $\mathcal{J}_B$ exists of size $\lfloor ct^2 \rfloor$. Thus, by any matching between these two set, there exists $\lfloor ct^2 \rfloor$ pairs (T_A, T_B) where $T_A \in \mathcal{J}_A, T_B \in \mathcal{J}_B$ and each T_A or T_B appears in at most one pair. For each such pair, identify the leaf of T_A labelled α with the leaf of T_B labelled α and supress this vertex to obtain a linear binary tree on L', as indicated in Fig. 1.

In this way $\lfloor ct^2 \rfloor$ trees are created which all possess the split $\{A,B\} \in \sum'$, but which share no other nontrivial split.

Letting $k = \mid \sum \mid$, now consider the profile P, consisting of all the $\lfloor ct^2 \rfloor$ binary trees on $n(1+t)$ leaves formed from $\sum'$ in this way. Applying Lemma 2, with

$$\ell = 2\lfloor ct^2 \rfloor [(n(1+t) - 3)k - \ell']$$

we have $D(T,P) \leq \ell$ if and only if

$$\sum_{\pi' \in \widehat{\sigma}(T)} f_{\pi'} \geq \lfloor ct^2 \rfloor \ell'$$

Now, $\sum_{\pi' \in \widehat{\sigma}(T)} f_{\pi'} = A + B$, where

$$A = \sum_{\pi' \in \widehat{\sigma}(T) \cap \sum'} f_{\pi'} \text{ and } B = \sum_{\pi' \in \widehat{\sigma}(T) - \sum} f_{\pi'}$$

Note that $|A| = \lfloor ct^2 \rfloor \mid \sum' \cap \widehat{\sigma}(T) \mid$, since $f_{\pi'} = \lfloor ct^2 \rfloor$ for any $\pi' \in \sum' \cap \widehat{\sigma}(T)$. Also $|B| \leq \mid \widehat{\sigma}(T) \mid = n(1+t) - 3$, since if $\pi' \notin \sum'$, then $f_{\pi'} \leq 1$. Finally, note that $\mid \sum' \cap \widehat{\sigma}(T) \mid = \mid \sum \cap \widehat{\sigma}(T^*) \mid$, where T^* is the tree obtained from T by pruning the leaves in $L' - L$ from T. Thus, $D(T,P) \leq \ell$, if and only if

$$\mid \sum \cap \widehat{\sigma}(T^*) \mid + \frac{B}{\lfloor ct^2 \rfloor} \geq \ell'$$

But since $|B| \leq n(1+t) - 3$, we can choose t sufficiently large so that $B/\lfloor ct^2 \rfloor < 1$, and then we have $D(T,P) \leq \ell$ if and only if

$$\mid \sum \cap \widehat{\sigma}(T^*) \mid \geq \ell' \quad (*)$$

Thus, given a polynomial (in $n, \ell, |P|$) algorithm for deciding whether $D(T,P) \leq \ell$, for variable ℓ, one obtains a polynomial time algorithm for deciding whether or not $\sum$ has a compatible subset of cardinality of least ℓ', for variable ℓ', since the relationship between ℓ and ℓ' is polynomial, and the value of t required to satisfy (*) is polynomial in n. Thus we obtain the required polynomial time transformation, thereby completing the proof. □

Acknowledgement: FRM was supported by Grant N00014-89-J-1643 from the US Office of Naval Research. MAS was supported by the New Zealand Lotteries Commission.

References:

BARTHÉLEMY, J. P., and MONJARDET, B. (1988): The median procedure in data analysis: New results and open problems. In: H. H. Bock (ed.): *Classification and Related Methods of Data Analysis.* Elsevier, Amsterdam, 309-316.

BARTHÉLEMY, J. P., and JANOWITZ, M. (1991): A formal theory of consensus. *SIAM Journal on Discrete Mathematics, 4, 305-322.*

BARTHÉLEMY, J. P., and MCMORRIS, F. R. (1986): The medain procedure for n-trees. *Journal of Classification, 3, 329-334.*

BUNEMAN, P. (1971): The recovery of trees from measures of dissimilarity. In: F. R. Hodgson, D. G. Kendall, and P. Tauta (eds.): *Mathematics in Archaeological and Historical Sciences.* Edinburgh University Press, Edinburgh, 387-395.

DAY, W. H. E. (1985): Optimal algorithm for comparing trees with labeled leaves. *Journal of Classification, 2, 7-28.*

DAY, W. H. E. and SANKOFF, D. (1986): Computational complexity of inferring phylogenies by compatibility. *Systematic Zoology, 35, 224-229.*

GAREY, M. R. and JOHNSON, D. S. (1979): Computers and intractability. W. H. Freeman, San Francisco.

PENNY, D., FOULDS, L. R., and HENDY, M. D. (1982): Testing the theory of evolution by comparing phylogenetic trees constructed from five different portein sequences. *Nature, 297, 197-200.*

SPENCER, J. (1987): Ten lectures on the probabilistic method. *SIAM*, Philadelphia.

STEEL, M. A., and PENNY, D. (1993): Distributions of tree comparison metrics - some new results. *Systematic Biology, 42, 126-141.*