Pergamon

Appl. Math. Lett. Vol. 9, No. 2, pp. 55–59, 1996

0893-9659/96 $15.00 + 0.00

S0893-9659(96)00012-2

Finding a Maximum Compatible Tree is NP-Hard for Sequences and Trees

A. M. HAMEL* AND M. A. STEEL
Department of Mathematics and Statistics, University of Canterbury
Christchurch, New Zealand

(Received July 1995; accepted August 1995)

Abstract—We show that the following two related problems arising in phylogenetic analysis are NP-hard:

(i) given a collection of aligned 2-state sequences, find the largest subset of sequences compatible with some tree,

(ii) given six trees, leaf-labelled by S, find the largest subset S' of S so that the six subtrees induced by S' are compatible.

Keywords—Trees, Sequences, NP-hard, Compatibility.

1. INTRODUCTION

A tree that has its leaves labelled by a set S and its remaining vertices unlabelled and of degree at least 3 is a useful model for representing evolutionary relationships in biology. Such an object is called a phylogenetic tree on S. Here we refer to it simply as a *tree on* S, and it is *binary* if all nonleaf vertices have degree 3. Note that a tree T on S determines a collection Σ_T of bipartitions (i.e., partitions of a set into two nonempty subsets) of S, called the *splits* of T—where each split is obtained by deleting an edge of T and recording which leaves lie in the two resulting components. We say a split is *trivial* if one of the sets contains just one element. A collection Σ of bipartitions is said to be *compatible* if $\Sigma \subseteq \Sigma_T$ for some tree T on S.

A fundamental theorem in [1] states that Σ is compatible if and only if Σ is pairwise compatible, and this is equivalent to requiring that for each pair $\{A, A'\}$, $\{B, B'\} \in \Sigma$, at least one of the four intersections $A \cap B$, $A \cap B'$, $A' \cap B$, $A' \cap B'$ is empty. Thus determining compatibility of Σ can be achieved in polynomial time (indeed in linear time, see [2]).

Day and Sankoff [3] showed that the problem of determining whether Σ has a subset of size at least k which is compatible is NP-complete (for variable k). Here we consider the following dual problem, which we show later is NP-complete.

Problem: Subcharacter compatibility (SCC).
Instance: A collection Σ of bipartitions of a set S, integer k.
Question: Is there a subset S' of S of size at least k, such that the bipartitions Σ' of S' induced by Σ are compatible?

It follows that the following problem in phylogenetic analysis is, in general, NP-hard: given a collection of aligned DNA sequences, determine the largest subset of these sequences that can have

*Supported by a postdoctoral fellowship from the Natural Sciences and Engineering Research Council of Canada. We wish to thank A. Dress, whose comments have helped to clarify our presentation.

evolved on a tree from some (unknown) ancestral sequence without reverse or parallel mutations. SCC is a particular case of this problem since (i) any collection Σ of bipartitions can be realized by sites in a collection of aligned DNA sequences (using just two of the four states available), and (ii) compatibility for Σ corresponds to fitting the corresponding sequences to a tree in the manner prescribed.

A related problem takes as its input a collection of $P = \{T_1, \ldots, T_k\}$ of trees on S, rather than bipartitions. Given a subset S' of S, and a tree T on S, take the subtree of T which connects just the leaves of T labelled by S' and make this subtree homeomorphically irreducible (i.e., suppress vertices of degree two) to obtain a tree on S', denoted $T_{|S'}$. The *maximum agreement subtree* (MAST) *problem* is to find a largest subset S' of S for which $T_{i|S'}$, $i = 1, \ldots, k$ all agree (this common tree which we call a MAS tree is called a *maximum agreement subtree* in [4], or a *maximum homeomorphic subtree* in [5]). This problem, posed in [6], is solvable in polynomial time when either $k = 2$ [4], or when the degree of the vertices of the trees in P is bounded [5]; however, without this last restriction it is NP-hard when $k = 3$ [5].

One problem with MAST in phylogenetic applications is that it is overly severe. This is because a vertex v of degree $d > 3$ in a reconstructed phylogenetic tree does not necessarily represent the simultaneous creation of $(d-1)$ descendants from the ancestral species represented by v, but may represent rather that the exact phylogenetic details of the descent of these $(d-1)$ descendants are unclear. This leads us to the following definitions.

We say that a tree T' on S *refines* a tree T on S if, by collapsing certain edges of T', one obtains T. More generally, given a collection $P = \{T_1, \ldots, T_k\}$ of trees on S, a tree T' on $S' \subseteq S$ is *compatible* with P if T' refines $T_{i|S'}$ for $i = 1, \ldots, k$. A *maximum compatible tree* (MC tree) for P is a tree T' on a maximum cardinality subset S' of S which is compatible with P. Note that an MC tree can have more vertices than any of the input trees, while a MAS tree is a subtree of each input tree.

Note also that if all the trees in P are binary then MCT is equivalent to MAST. Thus, in this case, finding an MC tree can be achieved in polynomial time by using an algorithm described in [5]. However, in general this problem is NP-hard, as we will shortly show. First, we state the problem more precisely.

Problem: Maximum Compatible Tree (MCT).
Instance: A collection P of trees on a set S, integer k.
Question: Is there a subset S' of S of size at least k, and a tree T' on S' which is compatible with P?

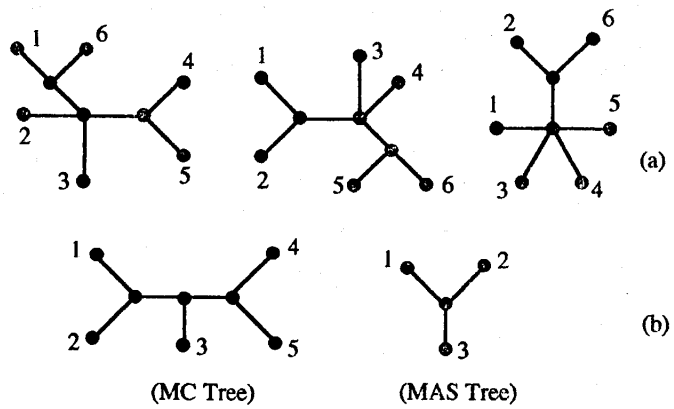

Figure 1. The MC tree and a MAS tree for three trees on $\{1, 2, 3, 4, 5, 6\}$.

2. RESULTS

Note that SCC and MCT are both in NP, and although superficially different, they are actually (polynomially) equivalent by the following reasoning. Given an instance (Σ, k) of SCC, we can

replace each $\sigma \in \Sigma$ by the tree on S whose only nontrivial split is $\sigma = \{A, A'\}$. In this way we obtain a collection $P = P(\Sigma)$ of trees and thereby an instance (P, k) of MCT, for which the corresponding question has answer "yes" precisely if it is "yes" for (Σ, k) in SCC [7, Theorem 1 (3a)]. Conversely, given an instance $(P, k), P = \{T_1, \ldots, T_r\}$ of MCT, let $\Sigma = \Sigma_P = \bigcup_{i=1,\ldots,r} \Sigma_{T_i}$, the union of the splits of the T_i. This gives an instance (Σ, k) of SCC for which the corresponding question has answer "yes" precisely if it is "yes" for (P, k) in MCT [7, Theorem 1 (3a)]. Note that the two constructions can be implemented in polynomial time, so that both problems are NP-complete once we show that either one of them is. In fact, we show a stronger result, namely that the following specialization of MCT is NP-complete.

Problem: Maximum Compatible Tree for six trees (MCT6).
Instance: A collection P of six trees on S, integer k.
Question: Same as for MCT.

THEOREM 2.1. *MCT6 and SCC are NP-complete.*

PROOF. Our proof is a modification of the NP-completeness proof of MAST (for 3 trees) [7]. By the comments preceding the theorem it suffices to show that MCT6 is NP-complete. The MCT6 problem is clearly in NP. We will reduce the three dimensional matching problem (3DM) [8] to MCT6. The 3DM is as follows:

Problem: 3DM.
Instance: Let $M \subseteq X \times Y \times Z$ where X, Y, and Z are disjoint sets, $|X| = |Y| = |Z| = q$.
Question: Does there exist a set M' such that $M' \subseteq M$, $|M'| = q$, and any two elements of M' differ in all three coordinates?

Define a *caterpillar tree* on n leaves ($n > 3$) to be a binary tree for which there are exactly two vertices that are each adjacent to precisely two leaves (note that, in any binary tree with more than three leaves, there are at least two such vertices). Examine these two vertices and, for each, distinguish one of the two leaves. Call one of these leaves the *root*; call the other the *summit* (see Figure 2a).

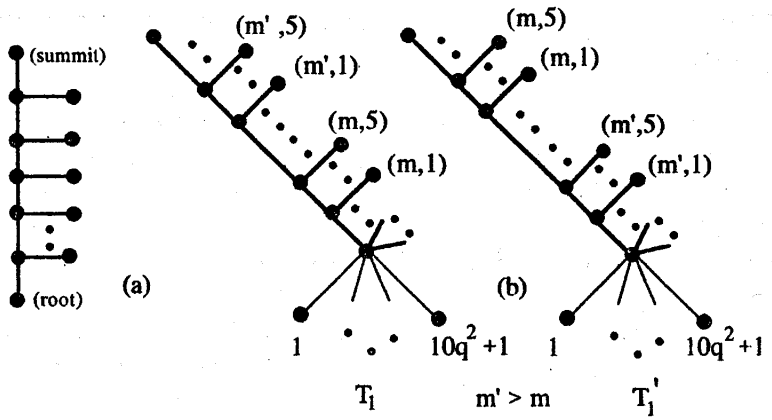

Figure 2. A caterpillar tree (a), and the ordering of its leaves in the trees T_1 and T_1' (b).

Given an instance of 3DM, we construct six trees, $\mathcal{T} = \{T_1, T_1', T_2, T_2', T_3, T_3'\}$ each on leaf set $S = \{1, 2, \ldots, 10q^2 + 1\} \cup (M \times \{1, \ldots, 5\})$, which have the property that $\mathcal{T}$ has an MC tree of size $10q^2 + 5q + 1$ if and only if the answer to the 3DM question is "yes." Each tree in $\mathcal{T}$ has leaves $1, \ldots, 10q^2 + 1$ adjacent to a vertex, and this vertex is identified with each root of a family of q caterpillar trees. Furthermore, each of the q caterpillar trees in T_1 and T_1' corresponds (bijectively) to an element of X (similarly, each of the caterpillar trees in T_2 and T_2' (respectively, T_3, T_3') correspond to an element of Y (respectively, Z)). We now describe precisely how the leaves of these caterpillar trees are labelled. First, we (arbitrarily) order M, and we use this order

to define two orders, $\leq$ and $\leq'$, on $M \times \{1,\ldots,5\}$ as follows: we write $(m,i) \leq (m',i')$ if $m < m'$ or if $m = m'$ and $i \leq i'$; and we write $(m,i) \leq' (m',i')$ if $m > m'$ or $m = m'$ and $i \leq i'$. Then in T_1 (respectively, T_1') we label the leaves of the caterpillar corresponding to $x \in X$ with the pairs $\{(m,i) : m_X = x;\ i = 1,\ldots,5\}$ (where $m = (m_X, m_Y, m_Z)$), arranged in increasing order under $\leq$ (respectively, $\leq'$) from the leaf closest to the root, to the summit (see Figure 2b; note also that, without loss of generality, we may assume that $\{m \in M : x_M = x\}$ is nonempty). We do this for all $x \in X$ (for which we have an associated caterpillar tree in T_1 and T_1'). We then repeat this procedure, with Y in place of X, to obtain the trees T_2 and T_2', and again, with Z in place of X, to obtain the trees T_3, T_3'.

We first show that any MC tree for $\mathcal{T}$ must contain all the leaves $1,\ldots,10q^2+1$. Suppose that T is compatible with $\mathcal{T}$ and that T uses leaves from i caterpillars of T_1 and j leaves from $1,\ldots,10q^2+1$. If $i+j > 2$, then we have a distinguished vertex in T to which we can attach the remaining leaves from $1,\ldots,10q^2$ to obtain a tree which is still compatible with $\mathcal{T}$. If however $i+j \leq 2$, then T can have at most $10q^2$ leaves (as there are at most q^2 elements of M which have x as their first coordinate, and so each caterpillar in T_1 has at most $5q^2$ leaves), and so the star tree with leaves $1,\ldots,10q^2+1$ is a tree that is compatible with $\mathcal{T}$, yet has more leaves than T. This establishes our claim that any MC tree for $\mathcal{T}$ must contain all the leaves $1,\ldots,10q^2+1$.

We next claim that any MC tree T for $\mathcal{T}$ has the property that if (m,i) and (m',j), $m \neq m'$, are leaves of T, then m and m' differ in all three coordinates. Suppose, on the contrary, that (m,i) and (m',j) both appear in the leaf set A of T and that these two leaves share a coordinate in m, say the first (the other two cases are similar)—we will derive a contradiction and show that T is not an MC tree by constructing a tree compatible with $\mathcal{T}$ but with more leaves than T. By assumption, (m,i) and (m',j) appear on the same caterpillar of T_1. Since T refines both $T_{1|A}$ and $T'_{1|A}$, and since, as we have just shown, $\{1,\ldots,10q^2+1\}$ lies in A, it follows that (m,i) and (m',j) are the only two leaves of A on this caterpillar. Let T^* be the tree obtained from T by deleting (m',j) and deleting any other pair (m'',k) which appears together with (m,i) on the same caterpillar of T_2 or T_3 (note that, as before, there can only be at most one such pair (m'',k) for T_2 and at most one such pair for T_3) and then making the resulting tree homeomorphically irreducible. Note that T^* is still compatible with $\mathcal{T}$ but it may have up to three leaves fewer than T. However if we now delete from T^* the leaf (m,i) and identify its adjacent vertex with the root of a caterpillar tree which has five other leaves, labelled (m,j), $j = 1,\ldots,5$, and arranged in increasing order (under $\leq$) from the leaf nearest the root to the summit, we obtain a tree which is both compatible with $\mathcal{T}$ and which has more leaves than T. This shows that T was not an MC tree for $\mathcal{T}$, thereby providing the required contradiction. Thus, m and m' differ in all three coordinates, as claimed.

Furthermore, if an MC tree contains leaf (m,i) for some $m \in M$ and $i \in \{1,\ldots,5\}$ then it must also contain the four other leaves (m,j): $j = 1,\ldots,5; j \neq i$, as all five leaves appear in the same order on the caterpillars in trees from $\mathcal{T}$, and we have already shown that no other leaves of this MC tree can lie on the same caterpillar as these leaves. Hence, if the six trees in $\mathcal{T}$ have an MC tree of size $10q^2+5q+1$, then, by taking the leaves of this MC tree and replacing each of the five leaves $\{(m,i);\ i = 1,\ldots,5\}$ by m we obtain a subset M' of M of size q, for which (as we have shown) any two elements differ in all three coordinates, so M' is a three-dimensional matching set.

Conversely, any three dimensional matching set M' gives rise to an MC tree for $\mathcal{T}$ of size $10q^2+5q+1$, namely the tree obtained from the (star) tree consisting of just a central vertex and the leaves $1,\ldots,10q^2+1$ as follows: for each $m \in M'$, identify with v the root of a caterpillar tree with six leaves, and label the other five leaves of this caterpillar tree $(m,1),\ldots,(m,5)$ in increasing order (under $\leq$) from the leaf closest to the root to the summit leaf.

Hence, M has a subset M' of size q such that any two elements of M' differ in all three coordinates iff, for the six trees $\mathcal{T}$, there is an MC tree of size $10q^2+5q+1$. ∎

REFERENCES

1. P. Buneman, The recovery of trees from measures of dissimilarity, In *Mathematics in the Archaeological and Historical Sciences*, (Edited by F.R. Hodson, D.G. Kendall, and P. Tautu), pp. 387-395, Edinburgh University Press, Edinburgh, (1971).
2. D. Gusfield, Efficient algorithms for inferring evolutionary trees, *Networks* **21**, 19–28 (1991).
3. W.H.E. Day and D. Sankoff, Computational complexity of inferring phylogenies by compatibility, *Syst. Zool.* **35** (2), 224–229 (1986).
4. M. Steel and T. Warnow, Kaikoura tree theorems: Computing the maximum agreement subtree, *Inform. Proc. Lett.* **48**, 77–82 (1993).
5. A. Amir and D. Kesselman, Maximum agreement subtree in a set of evolutionary trees—Metrics and efficient algorithms, (submitted).
6. C.R. Finden and A.D. Gordon, Obtaining common pruned trees, *J. Classification* **2**, 255–176 (1985).
7. A. Dress and M. Steel, Convex tree realizations of partitions, *Appl. Math. Lett.* **5** (3), 3–6 (1992).
8. R.M. Karp, Reducibility among combinatorial problems, In *Complexity of Computer Computations*, (Edited by R.E. Miller and J.W. Thatcher), pp. 85–103, Plenum Press, New York, (1972).