Finding a maximum compatible tree is NP-hard for sequences and trees

by

A.M. Hamel and M.A. Steel
Department of Mathematics and Statistics
University of Canterbury, Christchurch, New Zealand.

No. 114 October, 1994

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

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

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 non-leaf 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 $\Sigma = \Sigma_T$ for some tree T on S (this is equivalent to requiring $\Sigma \subseteq \Sigma_{T'}$ for some tree T' on S).

A fundamental theorem, due to Buneman [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 Gusfield [2]).

*Supported by a postdoctoral fellowship from the Natural Sciences and Engineering Research Council of Canada

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 a largest subset of these sequences that can have evolved on a tree from some (unknown) ancestral sequence without reverse or parallel mutations. SCC is a particular case of this problem since (i) a site in a collection of aligned DNA sequences induces a partition of the species set into at most four parts, (ii) any collection Σ of bipartitions can be realized in this way, and (iii) 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 is called a *maximum agreement subtree* in Steel and Warnow [4], or a *maximum homeomorphic subtree* in Amir and Kesselman [5]). This problem, posed by Finden and Gordon [6], is solvable in polynomial time when either $k = 2$ (Steel and Warnow [4]), or when the degree of the vertices of the trees in P is bounded (Amir and Kesselman [5]); however, without this last restriction it is NP-hard when $k = 3$ (Amir and Kesselman [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* (MCT) for P is a tree T' on a maximum cardinality subset S' of S which is compatible with P.

For example, consider the set P of the three trees in Fig. 1 (a). The unique MCT for P is the tree in Fig. 1 (b). A MAST is shown in Fig. 1 (c). Note that

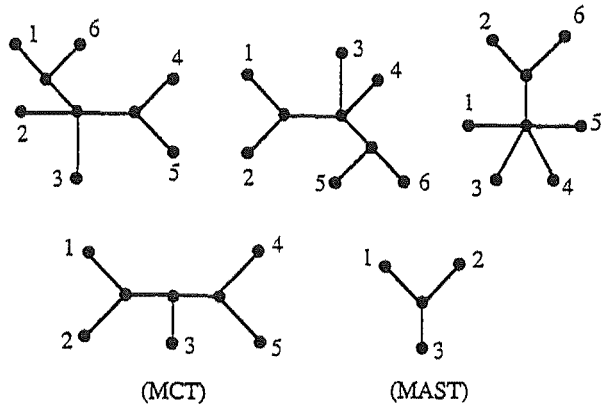

Figure 1: The MCT and a MAST for three trees on $\{1, 2, 3, 4, 5, 6\}$.

a MCT can have more vertices than any of the input trees, while a MAST 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 a MCT can be achieved in polynomial time by using an algorithm described by Amir and Kesselman [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?

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 (by Dress and Steel [7, Them. 1 (3a)]). Conversely, given an instance (P, k), $P = \{T_1, \ldots, T_r\}$ of MCT, let $\Sigma = \Sigma_P = \bigcup_{i=1,\ldots,r} \sum_{T_i}$, 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 (by Dress and Steel [7, Thm. 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, 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) given by Amir and Kesselman [5]. 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) (Karp [8]) to MCT6. The 3DM is as follows:

Problem: 3DM

Instance: Let $M \subseteq W \times X \times Y$ where W, X, and Y are disjoint sets, $|W| = |X| = |Y| = q$.

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 $n > 3$ leaves to be a binary tree for which exactly two vertices are each adjacent to two leaves. 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.*

Given an instance of 3DM, construct six trees, $\mathcal{T} = \{T_1, T_1', T_2, T_2', T_3, T_3'\}$ as follows. Let T_i and T_i' have a root r_i, $i = 1, 2, 3$. Order the 3–tuples arbitrarily in M. Let $w_i \in W$ for some $i = 1, \ldots, r$. Consider all 3–tuples $e_{i_1}, \ldots, e_{i_t} \in M$ whose first coordinate is w_i. Construct two caterpillar trees with $5t+1$ leaves as follows. Let the root of each be unlabelled and let the leaves of one be labelled $e_{i_1}^1, e_{i_1}^2, e_{i_1}^3, e_{i_1}^4, e_{i_1}^5, \ldots, e_{i_t}^1, \ldots, e_{i_t}^5$ in order with $e_{i_1}^1$ labelling the leaf closest the root and $e_{i_t}^5$ labelling the summit. Let the leaves of the other tree be labelled in the opposite subscript order but the same superscript order with $e_{i_t}^1$ labelling the leaf closest to the root, and $e_{i_1}^5$ labelling the summit. Identify the root of the first caterpillar with r_1 in T_1; identify the root of the second with r_1 in T_1'. Repeat this procedure for each x_i and y_i, $i = 1, \ldots, q$.

Now adjoin $5q^2 + 1$ children $z_1, \ldots, z_{5q^2+1}$ to each root r_1, r_2, r_3 in $\mathcal{T}$ and let $S = \{z_1, \ldots, z_{5q^2+1}, e_1^1, e_1^2, e_1^3, e_1^4, e_1^5, \ldots, e_l^4, e_l^5\}$ where $l = |M|$ and $e_i \in M$ for $i = 1, \ldots, l$. These new leaves will force the roots r_1, r_2, and r_3 to be in a MCT since a caterpillar can have at most $5q^2$ leaves.

We claim that the leaf set A of any MCT has the property that if $e_i^\alpha, e_j^\beta \in A$ for $i \neq j$, $i, j \in \{1, 2, \ldots, l\}$, $\alpha, \beta \in \{1, 2, 3, 4, 5\}$, then e_i and e_j differ in all three coordinates. Otherwise, suppose that e_i^α and e_j^β appear in the leaf set A of some MCT and share a coordinate, say the first coordinate (the other two cases are similar). Thus e_i^α and e_i^β appear on the same caterpillar of T_1. Since

the MCT refines both $T_{1|_A}$ and $T'_{1|_A}$ and since r_1 lies in A, it follows that e_i^α and e_j^β are the only two leaves of A on this caterpillar. If in addition e_i^α appears with another leaf, e_r^γ from A on the same caterpillar of T_2 (resp. e_s^δ on the same caterpillar of T_3), then e_r^γ (resp. e_s^δ) is the only leaf from A which appears with e_i^α on that caterpillar.

Let A' be the set obtained from A by deleting e_j^β and, if they exist, e_r^γ and e_s^δ and then adding the four other leaves e_i^ρ, $\rho \in \{1,2,3,4,5\}$, $\rho \neq \alpha$. Note that $|A'| > |A|$, and that A' is the leaf set of a tree which refines the six trees, $\{T|_{A'} : T \in \mathcal{T}\}$, hence A cannot have been the leaf set of a MCT. Note that in general if a MCT contains e_i^α for some $i \in \{1,\ldots,l\}$ and some $\alpha \in \{1,2,3,4,5\}$, it must also contain the other four e_i^ρ, $\rho \in \{1,2,3,4,5\}$, $\rho \neq \alpha$, as they appear in the same order on caterpillars in $\mathcal{T}$ and as the only way to prevent them from appearing in a MCT is if some e_j^β, $i \neq j$, from one of the same caterpillars is in the MCT. By the argument above this is impossible. Hence if the six trees have an MCT of size $5q^2 + 5q + 1$, then taking the leaves of this MCT and replacing each set $\{e_i^1,\ldots,e_i^5\}$, by e_i to obtain a subset M' of M, then M' has size q and any two elements of M' differ in all three coordinates, so M' is a three dimensional matching.

Conversely, any three dimensional matching set M' gives rise to a MCT of size $5q^2+5q+1$, namely the tree obtained from the (star) tree consisting of just a root and $5q^2 + 1$ leaves, $\{z_1,\ldots,z_{5q^2+1}\}$, by adjoining to the root, for each $e_i \in M'$, a caterpillar tree with six leaves, one of which is identified with the root and the other five of which are labelled $e_i^1,\ldots,e_i^5$.

Hence there is a set 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 a MCT is of size $5q^2+5q+1$. ◇

References

[1] P. Buneman, The recovery of trees from measures of dissimilarity, in *Mathematics in the Archaeological and Historical Sciences*, (F.R. Hodson, D.G. Kendall, and P. Tautu, Editors) Edinburgh University Press, Edinburgh, pp. 387-395, (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, *Proc. 35th IEEE FOCS, Santa Fe, 1994.*

[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] Karp, R.M. Reducibility among combinatorial problems, in *Complexity of Computer Computations* (R.E. Miller and J.W. Thatcher, Editors) Plenium Press, New York, pp. 85–103, (1972).