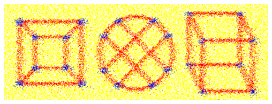

Journal of Graph Algorithms and Applications
http://jgaa.info/ vol. 18, no. 3, pp. 385–392 (2014)
DOI: 10.7155/jgaa.00327

The agreement problem for unrooted phylogenetic trees is FPT

Celine Scornavacca[1] *Leo van Iersel*[2] *Steven Kelk*[3] *David Bryant*[4]

[1]ISEM, CNRS – Université Montpellier II, Montpellier, France
[2]Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
[3]Department of Knowledge Engineering, Maastricht University, Maastricht, The Netherlands.
[4]Allan Wilson Centre for Molecular Ecology and Evolution, Department of Mathematics and Statistics, University of Otago, Otago, New Zealand

Abstract

A collection $\mathcal{T}$ of k unrooted phylogenetic trees on different leaf sets is said to be *strictly compatible* or *in agreement* if there exists a tree T such that each tree in $\mathcal{T}$ can be obtained from T by deleting leaves and suppressing degree-2 vertices. The problem of determining if a set of unrooted trees is in agreement has been proved NP-hard in 1992. Here, we show that an $f(k) \cdot n$ algorithm exists, for some computable function f of k, proving that strict compatibility of unrooted phylogenetic trees is fixed-parameter tractable with respect to the number k of trees. Designing a practical FPT algorithm remains an open problem.

Submitted:	Reviewed:	Revised:	Accepted:	Final:
March 2014	June 2014	June 2014	June 2014	June 2014

Published: June 2014

Article type: Regular paper

Communicated by: D. Wagner

Research supported by a Veni grant from The Netherlands Organisation for Scientific Research (NWO) and a French *Agence Nationale de la Recherche Investissements d'avenir / Bioinformatique* grant (ANR-10-BINF-01-02, Ancestrome).

E-mail addresses: celine.scornavacca@univ-montp2.fr (Celine Scornavacca) l.j.j.v.iersel@gmail.com (Leo van Iersel) steven.kelk@maastrichtuniversity.nl (Steven Kelk) david.bryant@otago.ac.nz (David Bryant)

1 Introduction

Phylogenetics is the research area that aims at reconstructing evolutionary histories from available data about present-day taxonomic units, usually called *taxa*. If X is a finite set of taxa, then an (unrooted) *phylogenetic tree* on X is a tree that has no degree-2 vertices and in which the leaves are bijectively labelled by the elements of the set X. The internal vertices of a phylogenetic tree model the divergence of lineages. A phylogenetic tree is said to be *binary* if each internal vertex has degree three, hence modelling the divergence of one lineage into two (sub)lineages. Non-binary phylogenetic trees, on the other hand, may have vertices of degree higher than three, which are called *polytomies*. There are two different interpretations of polytomies. When a polytomy is used to model uncertainty in the order of divergence events, we call it a *soft* polytomy, while polytomies that model simultaneous divergence of one lineage into three or more (sub)lineages are called *hard* polytomies.

When phylogenetic trees on several different, but overlapping, sets of taxa are available, a natural question to ask is whether there exists a single phylogenetic tree on the set X of all taxa that "agrees" with each of the input trees. The notion of "agreement" here depends on whether the polytomies of the input trees are assumed to be hard or soft. For the case of soft polytomies, one says that a set $\mathcal{T}$ of phylogenetic trees is *compatible* if there exists a phylogenetic tree T on X such that each tree in $\mathcal{T}$ can be obtained from a subtree of T by contracting edges. For the case of hard polytomies, one says that a set $\mathcal{T}$ of phylogenetic trees is *in agreement* or *strictly compatible* if there exists a phylogenetic tree T such that each tree in $\mathcal{T}$ can be obtained from a subtree of T by suppressing degree-2 vertices.

For binary input trees, the notions of compatibility and strict compatibility coincide. Unfortunately, it was shown that deciding whether a given set of binary phylogenetic trees is compatible (or, equivalently, strictly compatible) is NP-complete, even if each input tree contains exactly four taxa [12]. On the positive side, the compatibility problem was shown to be fixed-parameter tractable (FPT) in the number of input trees (not necessarily binary) [4]. However, prior to this article, it was not known whether it is fixed-parameter tractable to decide if a given set of non-binary trees is *strictly* compatible.

Here, we answer this question affirmatively, showing that there exists an algorithm with running time $f(k) \cdot |X|$ that decides whether a given set of k non-binary phylogenetic trees is strictly compatible, with X the union of the taxon sets of the input trees. We do so by formulating this problem in monadic second order logic (MSOL) on a graph of which the treewidth is bounded by k, a (non-trivial) extension of the approach by Bryant and Lagergren for the case of (non-strict) compatibility [4].

Our approach can also be extended to a more general setting where the input trees are allowed to have both hard as well as soft polytomies. Such situations occur for example in practical applications in biology [7, 9–11].

We note that the problem becomes polynomial-time solvable, both for compatibility as well as for strict compatibility, when the input trees are rooted [1,8]

(or all contain some common taxon). However, in practical applications it is not always possible to identify the root locations, e.g. due to a lack of data [6]. Hence, the unrooted variant is equally important.

2 Preliminaries

In this section, we give some preliminary definitions that are used throughout this paper. An unrooted phylogenetic tree T consists of vertices connected by edges, in which any two vertices are connected by exactly one path and with no degree-2 vertex. A rooted phylogenetic tree is defined similarly, except that it has exactly one vertex, called the *root* of the tree, that can have degree two. Leaves, defined as vertices with degree one, are labeled by the label set $\mathcal{L}(T)$, while other vertices, called internal vertices, are usually not labelled.

Given an unrooted phylogenetic tree T and a subset $Y \subseteq L(T)$, we denote with $T|_Y$ the tree obtained from the minimal subgraph of T connecting Y when suppressing vertices of degree two. We say that $T|_Y$ is the subtree of T induced by Y. Induced subtrees are defined in the same way for rooted trees, except that the root of $T|_Y$ becomes the vertex in the minimal connecting subgraph that is closest to the root of T, and we suppress all degree-2 vertices except the new root.

Given a rooted tree T and a set of three labels $\{u, v, w\}$ in $\mathcal{L}(T)$, $T|_{\{u,v,w\}}$ can be any of the three possible rooted binary trees on $\{u, v, w\}$ or the unique non-binary tree on $\{u, v, w\}$. The binary trees on $\{u, v, w\}$ are called triplets and are denoted respectively by $uv|w$, $uw|v$ and $wv|u$, depending on the unique non-trivial cluster in $T|_{\{u,v,w\}}$ (respectively $\{u, v\}$, $\{u, w\}$, and $\{v, w\}$). Representing rooted trees in Newick notation, we say that T displays the *triplet* $uv|w$ if $T|_{\{u,v,w\}} = ((u, v), w)$, while it displays the *fan* (u, v, w) if $T|_{\{u,v,w\}} = (u, v, w)$.

Let $\mathcal{T} = \{T_1, T_2, \ldots, T_k\}$ be a collection of strictly compatible, unrooted phylogenetic trees, not necessarily on the same taxon set. The *display graph* G for $\mathcal{T}$ is obtained from the disjoint graph union of all trees in $\mathcal{T}$ by identifying vertices with the same label. We denote by R^G the vertex-edge incidence relation in G.

3 Fixed-parameter tractability of Agreement

We say that a collection of phylogenetic trees $\mathcal{T} = \{T_1, T_2, \ldots, T_k\}$ is *strictly compatible* if there exists a phylogenetic tree T such that for each tree $T_i \in \mathcal{T}$ we have that $T|_{L(T_i)} = T_i$.

The problem we consider in this paper is thus the following:

Agreement of unrooted phylogenetic trees
Instance: A set of unrooted phylogenetic trees $\mathcal{T} = \{T_1, T_2, \ldots, T_k\}$.
Parameter: The number of trees k.

Question: Does there exist an unrooted phylogenetic tree T such that for each tree $T_i \in \mathcal{T}$ we have that $T|_{L(T_i)} = T_i$?

Recall that the Agreement problem is polynomial for a set of rooted phylogenetic trees [8].

Theorem 1 *Let $\mathcal{T} = \{T_1, T_2, \ldots, T_k\}$ be a collection of strictly compatible, unrooted phylogenetic trees, not necessarily on the same taxon set. The display graph of $\mathcal{T}$ has treewidth at most k.*

Proof: Theorem 1 in [4] states that, if $\mathcal{T}$ is compatible, then the display graph of $\mathcal{T}$ has treewidth at most k. Since strict compatibility implies compatibility, the theorem follows. □

To root an (unrooted) phylogenetic tree T on an edge $e = \{u, v\}$ consists in first creating a new node w and two new edges $\{w, u\}$ and $\{w, v\}$, then deleting the edge e, and finally defining w as the root of T.

Proposition 1 *A collection of unrooted phylogenetic trees $\mathcal{T} = \{T_1, T_2, \ldots, T_k\}$ is strictly compatible if and only if each tree $T_i \in \mathcal{T}$ can be rooted at some edge or vertex in such a way that the resulting rooted trees are strictly compatible.*

Proof: Suppose that each $T_i \in \mathcal{T}$ can be rooted (at a vertex or an edge) such that the resulting forest of rooted trees $\mathcal{T}^r = \{T_1^r, T_2^r, \ldots, T_k^r\}$ is strictly compatible. This implies that there exists a rooted tree T^r such that $T^r|_{L(T_i^r)} = T_i^r$. Let T be the unrooted version of T^r. Since, when unrooting trees the degree two vertices are suppressed, the unrooted version of T_i^r coincides with T_i for each $T_i \in \mathcal{T}$, and thus $T|_{L(T_i)} = T_i$ holds.
Now, suppose that $\mathcal{T}$ is strictly compatible. Then there exists an unrooted tree T such that $T|_{L(T_i)} = T_i$ that we can root at some vertex $r \in V(T)$ to obtain a rooted tree T^r. Let $\mathcal{T}^r = \{T_1^r, T_2^r, \ldots, T_k^r\}$ be such that, $\forall 1 \leq i \leq k$, $T_i^r = T^r|_{L(T_i)}$. It is easy to see that each T_i^r is a rooted version of T_i. Moreover, $\mathcal{T}^r$ is strictly compatible by construction. This concludes the proof.
□

Note that in the case of compatibility, it is sufficient to root the trees in $\mathcal{T}$ on edges [4]. This is not the case for strict compatibility. For example, the unrooted trees in Figure 1 are strictly compatible but none of the 288 possible ways of rooting them on edges gives rise to a set of strictly compatible rooted trees. In Figure 2, we also show an example for which it is not enough to root on vertices: indeed, the unrooted trees are strictly compatible but none of the 8 possible ways of rooting them on vertices gives rise to a set of strictly compatible rooted trees.

We denote by $\mathcal{R}(\mathcal{T})$ and $\mathcal{F}(\mathcal{T})$ respectively the set of triplets and fans displayed by a set of rooted trees $\mathcal{T}$.

Lemma 1 *Let $\mathcal{T} = \{T_1, T_2, \ldots, T_k\}$ be rooted phylogenetic trees on subsets of a leaf set $L(\mathcal{T})$. Then $\mathcal{T}$ is not strictly compatible if and only if there exists*

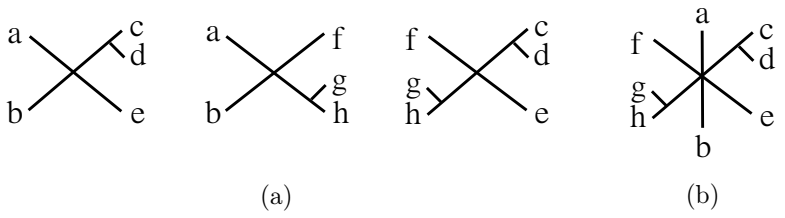

Figure 1: The unrooted trees in figure (a) are strictly compatible since the tree in (b) contains them all.

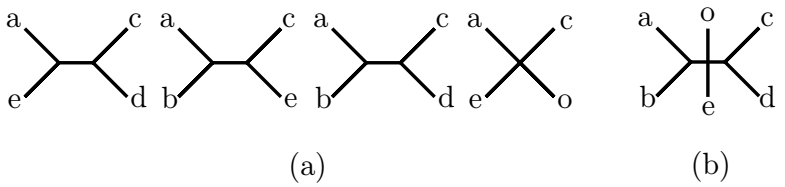

Figure 2: The unrooted trees in figure (a) are strictly compatible since the tree in (b) contains them all.

$S \subseteq L(\mathcal{T}), |S| \geq 3$, *such that for all non-empty, proper subsets* U *of* S *there exists either (i)* $uv|w \in R(\mathcal{T})$ *with* $u \in U$, $v \in S \setminus U$ *and* $w \in S$ *or (ii)* $(u, v, w) \in F(\mathcal{T})$ *with* $u, v \in U$ *and* $w \in S \setminus U$.

Proof: Let $[\mathcal{R}(\mathcal{T}), S]$ be the Aho graph [1] for the triplet set $\mathcal{R}(\mathcal{T})$ on a leaf set S, i.e. the undirected graph with vertices S such that there is an edge connecting two vertices u and v if and only if there exists $uv|w \in \mathcal{R}(\mathcal{T})$ and $w \in S$. Let $[\mathcal{R}(\mathcal{T}), \mathcal{F}(\mathcal{T}), S]$ be the graph obtained from the graph $[\mathcal{R}(\mathcal{T}), S]$ by repeatedly adding an edge between any pair of vertices u and v in two different connected components B_i, B_j if there exists $w \in B_j$ such that $(u, v, w) \in F(\mathcal{T})$. Similarly to what has been done in [4], we can prove the claim by proving that there exists $S \subseteq L(\mathcal{T}), |S| \geq 3$ such that $[\mathcal{R}(\mathcal{T}), \mathcal{F}(\mathcal{T}), S]$ is connected if and only if $\mathcal{T}$ is not strictly compatible.

If $\mathcal{T}$ is strictly compatible, then $\exists T$ such that $R(\mathcal{T}) \subseteq R(T)$ and $F(\mathcal{T}) \subseteq F(T)$. Let $S \subseteq L(\mathcal{T})$, $|S| > 1$ and consider the subtree $T|_S$ and its root x. Each direct descendant y of x in $T|_S$ determines a subset of S given by those leaves that are descendants of y. Then, the collection of these subsets partitions S into two or more blocks $B_1, \ldots, B_s$. If u and v are in two different blocks, say B_i and B_j, then there exists no $uv|w$ in $R(T)$, and thus in $R(\mathcal{T})$, with $w \in S$. Moreover, there exists no (u, v, w) in $F(T)$, and thus in $F(\mathcal{T})$, such that $w \in (B_i \cup B_j)$. Thus there are no edges between the blocks $B_1, \ldots, B_s$ in $[\mathcal{R}(\mathcal{T}), \mathcal{F}(\mathcal{T}), S]$, thus $[\mathcal{R}(\mathcal{T}), \mathcal{F}(\mathcal{T}), S]$ cannot be connected.

Suppose now that $\mathcal{T}$ is not strictly compatible. Then the OneTree algorithm presented in [8] will conclude that a supertree does not exist, and this

happens only if there exists a leaf set $S \subseteq L(\mathcal{T}), |S| \geq 3$ such that the graph $[\mathcal{R}(\mathcal{T}), \mathcal{F}(\mathcal{T}), S]$ is connected. □

Similarly to what has been done in [4], in the following we shall translate the choice of roots, together with Lemma 1, into monadic second order logic on the display graph G. The main difference between our characterization and that presented in [4] is the fact that we need to encode the possibility of rooting trees on edges and vertices (and not only on edges as done in [4]) and condition (ii) of Lemma 1, needed only for strict compatibility.

For this, we construct the following relational structure [2, 4]:

$$\mathbf{G} = (V(G), E(G), L(\mathcal{T}), V(T_1), ..., V(T_k), E(T_1), ..., E(T_k), R^G)$$

for which we define a formula $\Phi(A)$ such that $\mathbf{G} \models \Phi(A)$ if and only if A is a set of *edges and vertices* (one from each tree $T_i \in \mathcal{T}$) in which we can root each tree in $\mathcal{T}$ to make the forest *strictly* compatible in its rooted version (Proposition 1). For each $1 \leq i \leq k$, we define $\Psi_i(u, v, X)$ to express that there is a path with vertex set $Y \subseteq X \subseteq V(T_i)$ between the vertices u and v, and similarly $\Psi'_i(u, e, X)$ to express that there is a path with vertex set $Y \subseteq X \subseteq V(T_i)$ between the vertex u and the edge e (Note that the edge e does not need to lie entirely inside X, one endpoint is sufficient). More in detail, we have:

$$\Psi_i(u, v, X) = C_i(X) \wedge u \in X \wedge v \in X,$$

$$\Psi'_i(u, e, X) = C_i(X) \wedge u \in X \wedge (\exists v \in X(R^G(v, e))).$$

We use $C_i(X)$ to express that $X \subseteq V(T_i)$ and that X induces a connected subgraph of G:

$$C_i(X) = X \subseteq V(T_i) \wedge (\forall Y, Z \subseteq X((Y \cup Z = X) \rightarrow (\exists y \in Y, z \in Z, e \in E(R^G(y, e) \wedge R^G(z, e))))).$$

To simplify notations, we define $\bar{\Psi}_i(u, j, X)$ as follows:

$$\bar{\Psi}_i(u, j, X) = (j \in V(T_i) \rightarrow \Psi_i(u, j, X)) \wedge (j \in E(T_i) \rightarrow \Psi'_i(u, j, X)).$$

Then we define $\Phi(A)$ as the conjunction of

$$\bigwedge_{1 \leq i \leq k} |A \cap (V(T_i) \cup E(T_i))| = 1$$

and

$$\forall S \subseteq L(\mathcal{T})(|S| \geq 3 \rightarrow \exists U \subseteq S(U \neq \emptyset \wedge U \neq S \wedge (\forall u \in U, v \in (S \setminus U), w \in S(\neg\mathcal{R}(u, v, w, A))) \wedge (\forall u, v \in U, w \in (S \setminus U)(\neg\mathcal{F}(u, v, w, A))))).$$

We use $\mathcal{R}(u, v, w, A)$ to express that there is a tree $T_i \in \mathcal{T}$ with $\{u, v, w\} \subseteq V(T_i)$ such that, with the rooting implied by A, the path from u to v is vertex disjoint from the path from w to the root, and this is true if and only if $uv|w$ is a rooted triple in one of the trees rooted according to A. In monadic second order logic, $\mathcal{R}(u, v, w, A)$ can be expressed as follows:

$$\bigvee_{1 \leq i \leq k} \exists Y, Z \subset V(T_i), x \in A \cap (V(T_i) \cup E(T_i))(\Psi_i(u, v, Y) \wedge \bar{\Psi}_i(w, x, Z) \wedge (Y \cap Z = \emptyset)).$$

Finally, $\mathcal{F}(u, v, w, A)$ is used to express that there is a tree $T_i \in \mathcal{T}$ with $\{u, v, w\} \subseteq V(T_i)$ such that, with the rooting implied by A, $uv|w$, $uw|v$ and $vw|u$ are not rooted triples in T_i, and this is true if and only if (u, v, w) is a fan in one of the trees rooted according to A. Thus $\mathcal{F}(u, v, w, A)$ can be expressed as follows:

$$\bigvee_{1 \leq i \leq k} \forall Y, Z \subset V(T_i), x \in A \cap (V(T_i) \cup E(T_i))(\neg(\Psi_i(u, v, Y) \wedge \bar{\Psi}_i(w, x, Z) \wedge (Y \cap Z = \emptyset)) \wedge \neg(\Psi_i(u, w, Y) \wedge \bar{\Psi}_i(v, x, Z) \wedge (Y \cap Z = \emptyset)) \wedge \neg(\Psi_i(v, w, Y) \wedge \bar{\Psi}_i(u, x, Z) \wedge (Y \cap Z = \emptyset))).$$

Then, since the problem of determining how to root the unrooted trees to give strictly compatible rooted trees can be translated into second order monadic logic on the display graph G, and G has treewidth at most k, the following result holds:

Theorem 2 *Let $\mathcal{T} = \{T_1, T_2, \ldots, T_k\}$ be unrooted phylogenetic trees on subsets of a leaf set $L(\mathcal{T})$. Then it takes $f(k) \cdot n$ time to solve the Agreement problem for $\mathcal{T}$, where f is some computable function of k.*

Proof: First, note that the display graph G for $\mathcal{T}$ can be computed in $O(n \cdot k)$ time, where $n = |L(\mathcal{T})|$. Moreover, determining whether or not G has treewidth at most k requires $O(n \cdot f(k))$ for some computable f function of k [3]. Then, by Theorem 1, we have that if the treewidth of G is greater than k, then $\mathcal{T}$ cannot be strictly compatible. Then, by the result of Courcelle [5] and Arnborg et al. [2], we have that all problems in second order monadic logic can be solved in linear time (with respect to the number of vertices) on graphs with bounded treewidth. Since the number of vertices in G is $O(n \cdot k)$, the theorem follows. □

Note that our approach can also be extended to a more general setting where the input trees are allowed to have both hard as well as soft polytomies: it suffices to use $\mathcal{F}(u, v, w, A)$ for hard polytomies and $\mathcal{R}(u, v, w, A)$ for soft polytomies.

4 Conclusion

In this paper we prove that deciding whether a given set of k (not necessarily binary) phylogenetic trees is strictly compatible, is fixed-parameter tractable (FPT) in k. Note, however, that our algorithm does not directly lead to an efficient implementation. Therefore, it remains an important challenge to develop a practical FPT algorithm with a running time $c^k \cdot p(n)$ with c a small constant and p a low-degree polynomial.

Acknowledgments

This publication is contribution no. 2014-069 of the Institut des Sciences de l'Evolution de Montpellier (ISEM, UMR 5554).

References

[1] A. V. Aho, Y. Sagiv, T. G. Szymanski, and J. D. Ullman. Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. *SIAM Journal of Computing*, 10:405–421, 1981. doi:10.1137/0210030.

[2] S. Arnborg, J. Lagergren, and D. Seese. Easy problems for tree-decomposable graphs. *Journal of Algorithms*, 12:308 – 340, 1991. doi:10.1016/0196-6774(91)90006-K.

[3] H. Bodlaender. A linear-time algorithm for finding tree-decompositions of small treewidth. *SIAM Journal of Computing*, 25:1305–1317, 1996. doi:10.1137/S0097539793251219.

[4] D. Bryant and J. Lagergren. Compatibility of unrooted phylogenetic trees is FPT. *Theoretical Computer Science*, 351:296 – 302, 2006. doi:10.1016/j.tcs.2005.10.033.

[5] B. Courcelle. The monadic second-order logic of graphs. i. recognizable sets of finite graphs. *Information and Computation*, 85:12–75, 1990. doi:10.1016/0890-5401(90)90043-H.

[6] J. Felsenstein. *Inferring Phylogenies*. Sinauer Associates, Incorporated, 2004. URL: http://books.google.fr/books?id=GI6PQgAACAAJ.

[7] W. Maddison. Reconstructing character evolution on polytomous cladograms. *Cladistics*, 5(4):365–377, 1989. doi:10.1111/j.1096-0031.1989.tb00569.x.

[8] M. Ng and N. C. Wormald. Reconstruction of rooted trees from subtrees. *Discrete Applied Mathematics*, 69:19 – 31, 1996. doi:10.1016/0166-218X(95)00074-2.

[9] S. Poe and A. Chubb. Birds in a bush: five genes indicate explosive evolution of avian orders. *Evolution*, 58(2):404–415, 2004. doi:10.1554/03-037.

[10] O. Seehausen. African cichlid fish: a model system in adaptive radiation research. *Proceedings of the Royal Society B: Biological Sciences*, 273(1597):1987–1998, 2006. doi:10.1098/rspb.2006.3539.

[11] E. L. Stanley, A. M. Bauer, T. R. Jackman, W. R. Branch, and P. Mouton. Between a rock and a hard polytomy: rapid radiation in the rupicolous girdled lizards (squamata: Cordylidae). *Molecular phylogenetics and evolution*, 58(1):53–70, 2011. doi:10.1016/j.ympev.2010.08.024.

[12] M. Steel. The complexity of reconstructing trees from qualitative characters and subtrees. *Journal of Classification*, 9:91–116, 1992. doi:10.1007/BF02618470.