RECOVERING A PHYLOGENETIC TREE USING PAIRWISE CLOSURE OPERATIONS

K. T. HUBER, V. MOULTON, C. SEMPLE, AND M. STEEL

Abstract. A fundamental task in evolutionary biology is the amalgamation of a collection $\mathcal{P}$ of leaf-labelled trees into a single parent tree. A desirable feature of any such amalgamation is that the resulting tree preserves all of the relationships described by the trees in $\mathcal{P}$. For unrooted trees, deciding if there is such a tree is NP-complete. However, two polynomial-time approaches that sometimes provide a solution to this problem involve the computation of the semi-dyadic and the split closure of a set of quartets that underlies $\mathcal{P}$. In this paper, we show that if a leaf-labelled tree $\mathcal{T}$ can be recovered from the semi-dyadic closure of some set $\mathcal{Q}$ of quartet subtrees of $\mathcal{T}$, then $\mathcal{T}$ can also be recovered from the split-closure of $\mathcal{Q}$. Furthermore, we show that the converse of this result does not hold, and resolve a closely related question posed in [1].

Keywords: supertree, quartets, splits, semi-dyadic closure, split-closure

1. Introduction

A *binary phylogenetic (X)-tree* is an unrooted tree in which every interior vertex has degree three and whose leaf set is X. In evolutionary biology, X is commonly a set of species and a binary phylogenetic X-tree is used to represent the evolutionary relationships between the species in X.

A natural and fundamental task in evolutionary biology is to amalgamate binary phylogenetic trees with different, but overlapping leaf sets into a single parent tree. This single parent tree is called a *supertree* and ways to perform such tasks are called *supertree methods*. A desirable property of any supertree method is that, if possible, the resulting supertree 'displays' all of the evolutionary relationships of the input trees. More precisely, let $\mathcal{T}$ and $\mathcal{T}'$ be binary phylogenetic trees with leaf sets X and X', respectively. Then $\mathcal{T}$ *displays* $\mathcal{T}'$ if $X' \subseteq X$ and, up to suppressing degree-two vertices, $\mathcal{T}'$ is the minimal subtree of $\mathcal{T}$ that connects the elements of X'. In general, a binary phylogenetic tree $\mathcal{T}$ *displays* a collection $\mathcal{P}$ of binary phylogenetic trees if $\mathcal{T}$ displays each tree in $\mathcal{P}$. This desirable property of a supertree method leads to the following algorithmic problem:

Date: March 17, 2004.

The first and second authors thank The Swedish Research Council (VR). The third and fourth authors thank the New Zealand Marsden Fund. All authors thank The Swedish Foundation for International Cooperation in Research and Education (STINT).

Problem: Tree Compatibility
Instance: A collection $\mathcal{P}$ of binary phylogenetic trees.
Question: Does there exist a binary phylogenetic tree that displays each of the trees in $\mathcal{P}$ and, if so, can we construct such a tree?

In general, this problem is NP-complete [5]. However, there are a number of polynomial-time approaches to this problem that may provide a solution. Two of these approaches are based on the closure operators 'semi-dyadic closure' and 'split closure'. The former is associated with a collection of quartets and the latter is associated with a collection of partial splits.

A *quartet* is a binary phylogenetic tree with four leaves. The quartet with leaves a, b, c, d is denoted $ab|cd$ if the path from a to b does not intersect the path from c to d. A *(full) split* $A|B$ of X, also called an X*-split*, is a partition of X into two non-empty subsets A, B. Deleting any edge of a binary phylogenetic tree induces a split of X, namely the bipartition of X whose parts are the leaf sets of the two connected components of the resulting '2-tree forest'. For a binary phylogenetic tree $\mathcal{T}$, let $\mathcal{Q}(\mathcal{T})$ denote the set of quartets displayed by $\mathcal{T}$ and let $\Sigma(\mathcal{T})$ denote the set of splits of X induced by the interior edges of $\mathcal{T}$. It is well-known that $\mathcal{T}$ can be (efficiently) reconstructed from either $\mathcal{Q}(\mathcal{T})$ or $\Sigma(\mathcal{T})$. This means that possible solutions to Tree Compatibility can be sought by 'encoding' the input trees either as a set $\mathcal{Q}$ of quartets or as a set Σ of 'partial' X-splits (i.e., of splits of the various subsets of X constituting the leaf sets of the trees in $\mathcal{P}$), and then using these encodings either to construct an encoding of a binary phylogenetic tree that displays each of the original trees or to determine that no such tree exists. Two possible approaches in this regard are to compute the semi-dyadic closure of $\mathcal{Q}$ in case the encoding is done in terms of quartets or the split closure of Σ in case the encoding is done in terms of splits [3, 4]. The precise definitions are given in Section 2, but, roughly speaking, semi-dyadic closure and split closure are the end result of repeatedly applying a pairwise inference rule to collections of quartets or splits, respectively.

Any quartet can be viewed as partial split — simply take the split induced by the interior edge of the quartet — and so it is natural to ask how the semi-dyadic and the split closure of a set $\mathcal{Q}$ of quartets are related. In Section 3, we consider the relationship between the semi-dyadic and the split closure of $\mathcal{Q}$ when one or the other recovers a binary phylogenetic tree. In particular, we prove the following theorem:

Theorem 1.1. *Let $\mathcal{T}$ be a binary phylogenetic tree and let $\mathcal{Q}$ be a subset of $\mathcal{Q}(\mathcal{T})$. If the semi-dyadic closure of $\mathcal{Q}$ equals $\mathcal{Q}(\mathcal{T})$, then the split-closure of $\mathcal{Q}$ equals $\Sigma(\mathcal{T})$.*

Essentially, Theorem 1.1 states that if a binary phylogenetic tree $\mathcal{T}$ can be recovered from a subset $\mathcal{Q}$ of $\mathcal{Q}(\mathcal{T})$ using the semi-dyadic closure of $\mathcal{Q}$, then $\mathcal{T}$ can also be recovered from $\mathcal{Q}$ using the split-closure of $\mathcal{Q}$. Surprisingly, the converse of Theorem 1.1 is not true, a fact that we will also establish in Section 3.

The original motivation for Theorem 1.1 arose from an open question in [1, Remark 4] which relates semi-dyadic closure to minimum-sized sets of quartets that define a binary phylogenetic tree. In the last section, we resolve this question.

We end this section by noting that, throughout this paper, X is a finite set and, unless otherwise stated, the notation and terminology follows [4].

2. Semi-Dyadic Closure and Split Closure

The *semi-dyadic closure* of an arbitrary collection $\mathcal{Q}$ of quartets, denoted $\mathrm{scl}_2(\mathcal{Q})$, is the minimal set of quartets that contains $\mathcal{Q}$ and has the property that if $ab|cd$ and $bc|de$ are in $\mathrm{scl}_2(\mathcal{Q})$, then

$$ab|de, ab|ce, ac|de \in \mathrm{scl}_2(\mathcal{Q}).$$

The significance of this pairwise inference rule is highlighted in Proposition 2.1:

Proposition 2.1. [2] *Let $\mathcal{Q}$ be a set of quartets and let $\mathcal{T}$ be a binary phylogenetic tree. Then $\mathcal{T}$ displays $\mathcal{Q}$ if and only if $\mathcal{T}$ displays* $\mathrm{scl}_2(\mathcal{Q})$.

Let $\mathcal{S}_{part}(X)$ denote the set of all *partial splits* $A|B$ of X, i.e., of all splits of all subsets of X, considered as a poset relative to the partial order

$$A'|B' \leq A|B \iff (A' \subseteq A \text{ and } B' \subseteq B) \text{ or } (A' \subseteq B \text{ and } B' \subseteq A).$$

We will say that a partial split $A|B$ in $\mathcal{S}_{part}(X)$ *extends* a partial split $A'|B'$ in $\mathcal{S}_{part}(X)$ if $A'|B' \leq A|B$ holds.

To describe the split closure of a collection of partial splits, we need one further concept: A binary phylogenetic tree $\mathcal{T}$ *displays* a partial X-split σ if there is an X-split in $\Sigma(\mathcal{T})$ that extends σ. More generally, we say that $\mathcal{T}$ *displays* a collection Σ of partial X-splits if $\mathcal{T}$ displays each member of Σ.

For a collection Σ of partial X-splits, let $\overline{\Sigma}$ denote the (uniquely determined) minimal set of partial X-splits that contains Σ and has the property that if $A_1|B_1$ and $A_2|B_2$ are elements of $\overline{\Sigma}$ that satisfy

$$\emptyset \notin \{A_1 \cap A_2, A_1 \cap B_2, B_1 \cap B_2\} \text{ and } B_1 \cap A_2 = \emptyset,$$

then $(A_1 \cup A_2)|B_1$ and $A_2|(B_1 \cup B_2)$ are also elements of $\overline{\Sigma}$. We define the *split closure* of Σ, denoted $\mathrm{spcl}(\Sigma)$, to be the collection of maximal elements (with respect to the above partial order) in $\overline{\Sigma}$ in case any two partial splits in $\overline{\Sigma}$ are *compatible*, i.e., if one of the four sets $A_1 \cap A_2, A_1 \cap B_2$, $B_1 \cap A_2, B_1 \cap B_2$ is empty for any two splits $A_1|B_1$ and $A_2|B_2$ in $\overline{\Sigma}$, and to be the empty set otherwise.

The next lemma and corollary will be used in the proof of Theorem 1.1. For a partial X-split $A|B$, let

$$\mathcal{Q}(A|B) = \{aa'|bb' : a, a' \in A; b, b' \in B; a \neq a'; b \neq b'\}$$

and, for a set Σ of partial X-splits, let $\mathcal{Q}(\Sigma) = \bigcup_{A|B \in \Sigma} \mathcal{Q}(A|B)$. Observe that, for all binary phylogenetic trees $\mathcal{T}$, we have $\mathcal{Q}(\Sigma(\mathcal{T})) = \mathcal{Q}(\mathcal{T})$. Part (i) of Lemma 2.2 is due to Meacham [2] and Part (ii) is shown in [3, Proposition 2].

Lemma 2.2. *Let Σ be a set of partial X-splits. Then*

(i) *A binary phylogenetic tree $\mathcal{T}$ displays Σ if and only if $\mathcal{T}$ displays* $\mathrm{spcl}(\Sigma)$.
(ii) *If there exists a binary phylogenetic tree that displays Σ, then* $\mathrm{scl}_2(\mathcal{Q}(\Sigma)) \subseteq \mathcal{Q}(\mathrm{spcl}(\Sigma))$.

An immediate consequence of Lemma 2.2 is Corollary 2.3.

Corollary 2.3. *Let $\mathcal{T}$ be a binary phylogenetic tree and let $\mathcal{Q} \subseteq \mathcal{Q}(\mathcal{T})$. If* $\mathrm{scl}_2(\mathcal{Q}) = \mathcal{Q}(\mathcal{T})$, *then* $\mathcal{Q}(\mathrm{spcl}(\mathcal{Q})) = \mathcal{Q}(\mathcal{T})$.

3. Proof of Theorem 1.1

Before proving Theorem 1.1, we require one more concept. Let $\mathcal{T}$ be a binary phylogenetic tree and let e be an interior edge of $\mathcal{T}$. A quartet $q \in \mathcal{Q}(\mathcal{T})$ *distinguishes* e if e is the unique interior edge of $\mathcal{T}$ for which the quartet q is extended by the X-split in $\Sigma(\mathcal{T})$ induced by e. Also, a partial X-split σ *distinguishes* e if there is a quartet in $\mathcal{Q}(\sigma)$ that distinguishes e.

Proof of Theorem 1.1. Let $\mathcal{T}$ be a binary phylogenetic tree and let $\mathcal{Q}$ be a subset of $\mathcal{Q}(\mathcal{T})$. Suppose that $\mathrm{scl}_2(\mathcal{Q}) = \mathcal{Q}(\mathcal{T})$. Evidently, the theorem holds if $\mathcal{T}$ has exactly one interior edge. Therefore we may assume that $\mathcal{T}$ has at least two interior edges. Now assume that $\mathrm{spcl}(\mathcal{Q}) \neq \Sigma(\mathcal{T})$.

We first show that there is an interior edge of $\mathcal{T}$ for which there is a partial X-split in $\mathrm{spcl}(\mathcal{Q})$ that distinguishes this edge, but it is not full. Let e be an interior edge of $\mathcal{T}$ and let q be a quartet in $\mathcal{Q}(\mathcal{T})$ that distinguishes e. Then, by Corollary 2.3, $q \in \mathcal{Q}(\mathrm{spcl}(\mathcal{Q}))$ and so there exists a partial X-split σ in $\mathrm{spcl}(\mathcal{Q})$ that extends q. This means that σ distinguishes e. It follows that, for all interior edges e of $\mathcal{T}$, there is a partial X-split in $\mathrm{spcl}(\mathcal{Q})$ that distinguishes e. Furthermore, not all such partial X-splits are full, for otherwise $\mathrm{spcl}(\mathcal{Q}) = \Sigma(\mathcal{T})$.

Let $\sigma_1 = A_1|B_1$ be a partial X-split in $\mathrm{spcl}(\mathcal{Q})$ that is not full and distinguishes an interior edge, e_1 say, of $\mathcal{T}$. Let $aa'|bb'$ be a quartet in $\mathcal{Q}(A_1|B_1)$ that distinguishes e_1 with $a, a' \in A_1$ say, and let $A|B$ denote the full split in $\Sigma(\mathcal{T})$ that distinguishes e_1. Evidently, $A|B$ extends σ_1. Since σ_1 is not full, we may assume without loss of generality that A_1 is a proper subset of A. Let $c \in A - A_1$. As $\mathcal{T}$ is binary, it now follows that either (i) $ac|bb'$ but not $a'c|bb'$ distinguishes e_1, or (ii) $a'c|bb'$ but not $ac|bb'$ distinguishes e_1. First assume that Case (i) holds. Then $a'c|ab$ must be contained in $\mathcal{Q}(\mathcal{T})$. By Corollary 2.3, there is a partial X-split $\sigma_2 = A_2|B_2$ in $\mathrm{spcl}(\mathcal{Q})$ that extends $a'c|ab$. Clearly, $\sigma_1 \neq \sigma_2$. Without loss of generality, we may assume that $a', c \in A_2$ and $a, b \in B_2$. As $\mathcal{T}$ displays σ_1 and σ_2 and $\emptyset \notin \{A_1 \cap A_2, A_1 \cap B_2, B_1 \cap B_2\}$, it follows that $B_1 \cap A_2 = \emptyset$ (this is a well-known property of binary phylogenetic trees, see [4]). By the definition of the set $\overline{\mathcal{Q}}$ associated to $\mathcal{Q}$, this implies that $(A_1 \cup A_2)|B_1$ is contained in $\overline{\mathcal{Q}}$. But A_1 is a proper subset of $A_1 \cup A_2$ and so σ_1 is not a maximal element of $\overline{\mathcal{Q}}$. This contradicts the assumption that $\sigma_1 \in \mathrm{spcl}(\mathcal{Q})$. This completes the argument for Case (i). The argument for Case (ii) is similar and omitted. The theorem now follows. □

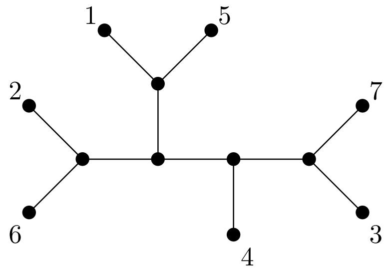

FIGURE 1. A binary phylogenetic tree.

The converse of Theorem 1.1 holds if $\mathcal{T}$ has at most six leaves, but fails in general. To see this, consider the binary phylogenetic tree $\mathcal{T}$ on $X = \{1, \ldots, 7\}$ shown in Fig. 1 and the set $\mathcal{Q} = \{26|57, 16|47, 15|34, 15|23, 14|37\}$ of quartets. Now $\mathcal{Q} \subseteq \mathcal{Q}(\mathcal{T})$, and it is easily verified that $\mathrm{spcl}(\mathcal{Q})$ equals $\Sigma(\mathcal{T})$. However,

$$\mathrm{scl}_2(\mathcal{Q}) = \mathcal{Q} \cup \{16|37, 46|37, 16|34, 15|37, 45|37, 15|47\} \neq \mathcal{Q}(\mathcal{T}).$$

4. Tight Sets

Let $\mathcal{P}$ be a collection of binary phylogenetic trees. We say that $\mathcal{P}$ *defines* a binary phylogenetic tree $\mathcal{T}$ if $\mathcal{T}$ displays $\mathcal{P}$ and $\mathcal{T}$ is the only such tree with this property. Furthermore, the *excess of* $\mathcal{P}$, denoted $\mathrm{exc}(\mathcal{P})$, is the quantity

$$\mathrm{exc}(\mathcal{P}) = |\mathcal{L}(\mathcal{P})| - 3 - \sum_{\mathcal{T} \in \mathcal{P}} i(\mathcal{T}),$$

where $\mathcal{L}(\mathcal{P})$ is the union of the leaf sets of the trees in $\mathcal{P}$ and $i(\mathcal{T})$ is the number of interior edges of $\mathcal{T}$. For a binary phylogenetic tree $\mathcal{T}$, we say that $\mathcal{P}$ is $\mathcal{T}$*-tight* if $\mathcal{P}$ defines $\mathcal{T}$ and $\mathrm{exc}(\mathcal{P}) = 0$. In particular, if a collection $\mathcal{Q}$ of quartets is $\mathcal{T}$-tight, then $\mathcal{Q}$ has size $|\mathcal{L}(\mathcal{T})| - 3$, the smallest sized subset of $\mathcal{Q}(\mathcal{T})$ that defines $\mathcal{T}$. Loosely speaking, a collection of binary phylogenetic trees is $\mathcal{T}$-tight if it contains the absolute minimum amount of information that is required to recover a binary phylogenetic tree $\mathcal{T}$.

It is shown in [1, Theorem 3] that if $\mathcal{P}$ is a collection of binary phylogenetic trees that defines a binary phylogenetic tree $\mathcal{T}$ and contains a $\mathcal{T}$-tight subset $\mathcal{P}'$, then

$$\mathrm{scl}_2\left(\bigcup_{\mathcal{T}' \in \mathcal{P}} \mathcal{Q}(\mathcal{T}')\right) = \mathcal{Q}(\mathcal{T}).$$

Moreover, in the remark directly following this theorem, it is stated that the converse of this result does not hold for arbitrary collections $\mathcal{P}$ of binary phylogenetic trees. However, the authors also state that they do not know if this is the case when $\mathcal{P}$ is a collection of quartets. In other words, the following question remained unanswered: if $\mathcal{T}$ is a binary phylogenetic tree and $\mathcal{Q} \subseteq \mathcal{Q}(\mathcal{T})$ with $\mathrm{scl}_2(\mathcal{Q}) = \mathcal{Q}(\mathcal{T})$, does it follow that $\mathcal{Q}(\mathcal{T})$ contains a $\mathcal{T}$-tight subset? Observe that $\mathcal{Q}$ satisfies the assumptions of Theorem 1.1. We conclude this paper by providing an example which shows that this is not necessarily the case.

Figure 2. Two binary phylogenetic trees.

Let $\mathcal{T}$ be the binary phylogenetic tree on $X = \{1, \ldots, 6\}$ shown in Fig. 2(a) and let

$$\mathcal{Q} = \{14|56, 15|36, 23|45, 12|36\}.$$

Note that $\mathcal{Q} \subseteq \mathcal{Q}(\mathcal{T})$. It is straightforward to check that $\mathrm{scl}_2(\mathcal{Q}) = \mathcal{Q}(\mathcal{T})$. Now, each quartet in $\mathcal{Q} - \{15|36\}$ distinguishes a distinct interior edge of $\mathcal{T}$, while $15|36$ does not distinguish any interior edge of $\mathcal{T}$. This means that the only possibility for a $\mathcal{T}$-tight subset of $\mathcal{Q}$ is $\mathcal{Q} - \{15|36\}$ as every interior edge of $\mathcal{T}$ needs to be distinguished by a quartet in $\mathcal{Q}$ (see [4, Theorem 6.8.7]). But the binary phylogenetic tree shown in Fig. 2(b) also displays $\mathcal{Q} - \{15|36\}$. Thus $\mathcal{Q} - \{15|36\}$ does not define $\mathcal{T}$ and so $\mathcal{Q}$ does not contain a $\mathcal{T}$-tight subset.

Acknowledgments. We thank Tobias Dezulian for the use of his software for verifying the example given in Section 3. We also thank Andreas Dress for his constructive comments (especially in regards to the presentation of the proof of Theorem 1.1) and suggestions for future research based around more general pairwise closure operators.

References

[1] S. Böcker, D. Bryant, A. Dress, and M. Steel, Algorithmic aspects of tree amalgamation, *Journal of Algorithms* **37** 522-537 (2000).
[2] C. A. Meacham, Theoretical and computational considerations of the compatibility of qualitative taxonomic characters, In *Numerical Taxonomy*, (Edited by J. Felsenstein), pp. 304-314, NATO ASI Series Vol. G1, Springer-Verlag (1983).
[3] C. Semple and M. Steel, Tree reconstruction via a closure operation on partial splits, In *Proceedings of Journées Ouvertes: Biologie, Informatique et Mathématiques*, (Edited by O. Gascuel and M.-F. Sagot), pp. 129-134, Lecture Notes in Computer Science, **2066**, Springer-Verlag, Berlin (2001).
[4] C. Semple and M. Steel, *Phylogenetics*, Oxford University Press (2003).
[5] M. Steel, The complexity of reconstructing trees from qualitative characters and subtrees, *Journal of Classification* **9** 91-116 (1992).

Department of Biometry and Engineering, The Swedish University of Agricultural Sciences, Box 7013, 750 07 Uppsala, Sweden, and, The Linnaeus Centre for Bioinformatics, Uppsala University, Box 598, 751 24 Uppsala, Sweden

E-mail address: katharina.huber@lcb.uu.se

The Linnaeus Centre for Bioinformatics, Uppsala University, Box 598, 751 24 Uppsala, Sweden

E-mail address: vincent.moulton@lcb.uu.se

Biomathematics Research Centre, Department of Mathematics and Statistics, University of Canterbury, Private Bag 4800, Christchurch, New Zealand

E-mail address: c.semple@math.canterbury.ac.nz

Biomathematics Research Centre, Department of Mathematics and Statistics, University of Canterbury, Private Bag 4800, Christchurch, New Zealand

E-mail address: m.steel@math.canterbury.ac.nz