Available online at www.sciencedirect.com
ScienceDirect
Mathematical Biosciences 208 (2007) 521–537
Mathematical Biosciences
www.elsevier.com/locate/mbs

Closure operations in phylogenetics

Stefan Grünewald, Mike Steel *, M. Shel Swenson

Allan Wilson Centre for Molecular Ecology and Evolution, Biomathematics Research Centre, University of Canterbury, Christchurch, New Zealand

Received 21 July 2005; received in revised form 2 November 2006; accepted 7 November 2006
Available online 18 November 2006

Abstract

Closure operations are a useful device in both the theory and practice of tree reconstruction in biology and other areas of classification. These operations take a collection of trees (rooted or unrooted) that classify overlapping sets of objects at their leaves, and infer further tree-like relationships. In this paper we investigate closure operations on phylogenetic trees; both rooted and unrooted; as well as on X-splits, and in a general abstract setting. We derive a number of new results, particularly concerning the completeness (and incompleteness) and complexity of various types of closure rules.

Keywords: Phylogenetic tree; Closure operations; Compatibility; Supertree

1. Introduction

Phylogenetic trees are widely used to represent evolutionary relationships, particularly in biology. Such trees have labelled leaves, and unlabelled interior vertices, and may be rooted or unrooted. One technique for building phylogenetic trees – sometimes called the *supertree approach* – is to combine trees on overlapping sets of leaves. This has become a widely applied technique in systematic biology, and a central tool in the challenge to construct a 'tree of life' [2].

* Corresponding author.
E-mail addresses: stefan@picb.ac.cn (S. Grünewald), m.steel@math.canterbury.ac.nz (M. Steel), mswenson@math.utexas.edu (M.S. Swenson).

0025-5564/$ - see front matter
doi:10.1016/j.mbs.2006.11.005

The requirement that these trees are *compatible*, that is, fit together into a parent tree enforces further tree-like relations to hold. This allows one to infer new phylogenetic relationships from the input, and one can iterate this procedure, leading to the concept of a *closure operation* on a set of trees. Such operations have proved to be particularly useful, both for theory [4,5,8,12,17] and applications [9,7,13–15,20]. For example, a closure operation may produce a conflicting pair of trees, thereby showing that the initial set of input trees was inconsistent with any global tree, something which may not have been apparent to start with. Alternatively, a closure operation may produce a sufficiently rich set of additional trees, that a global tree is uniquely specified and easily constructed. Certain types of closure operations are also polynomial time, allowing for the solution of special cases of problems which in general are NP-complete [5,19].

For rooted phylogenetic trees, the basic building blocks are 'rooted triples' – induced rooted subtrees on three leaves. For unrooted phylogenetic trees, the building blocks are 'quartet trees' – induced subtrees on four leaves. In both cases the closure operations take a set of these small trees and produce further ones. For unrooted phylogenetic trees we also study a closure operation on a different type of building block – namely the 'splits' of the leaf set into two disjoint subsets.

Our main results can be summarized as follows (precise definitions and statements are given later):

- For any rooted phylogenetic tree we determine the minimum number of rooted triples whose closure gives all the induced rooted triples for that tree.
- In contrast to the rooted setting, closure rules for quartet trees do not suffice to detect incompatibility. That is, there exists an incompatible set of quartet trees for which every proper subset of the quartets is both compatible and closed, thereby settling a question raised in the literature.
- Two closure rules (defined more than 20 years ago) on pairs of splits of the leaf sets of trees are complete amongst pair-wise rules on partitions of subsets of X.
- We describe how some of the arguments presented can be rephrased in a more general setting.

1.1. Basic definitions

We mostly follow the notation of [18]. A *rooted phylogenetic X-tree* $\mathcal{T}$ is a rooted tree, in which X is the set of leaves, and the interior vertices are unlabelled and have at least two outgoing edges. In case each interior vertex has exactly two outgoing edges, $\mathcal{T}$ is said to be *binary*. We let $\mathring{E}(\mathcal{T})$ denote the set of (interior) edges of $\mathcal{T}$ that are not incident with a leaf. A binary rooted phylogenetic tree on three leaves is called a *rooted triple*.

The *clusters* of $\mathcal{T}$ are the subsets of X that consist of all the elements of X that are separated from the root vertex of $\mathcal{T}$ by some vertex of $\mathcal{T}$. It is a classical result that a rooted phylogenetic X-tree is determined up to isomorphism by its set of clusters. We denote the rooted triple with leaf set $\{x,y,z\}$ that contains the cluster $\{x,y\}$ by $xy|z$ or, equivalently, by $z|xy$.

For two phylogenetic X-trees $\mathcal{T},\mathcal{T}'$, if the clusters of $\mathcal{T}$ are a subset of the clusters of $\mathcal{T}'$ we say that $\mathcal{T}'$ *refines* $\mathcal{T}$, written $\mathcal{T} \leqslant \mathcal{T}'$. Given a subset X' of X, and a rooted phylogenetic X-tree, the *induced* tree $\mathcal{T}|X'$ is the rooted phylogenetic X'-tree that has as its set of clusters $\{A \cap X' : A \text{ is a cluster of } \mathcal{T}, A \cap X' \neq \emptyset\}$.

A rooted phylogenetic X-tree $\mathcal{T}$ is said to *display* another rooted phylogenetic X'-tree $\mathcal{T}'$ where $X' \subseteq X$ if $\mathcal{T}' \leqslant \mathcal{T}|X'$. We let $r(\mathcal{T})$ denote the set of all rooted triples displayed by $\mathcal{T}$. To illustrate this idea, Fig. 1 shows a rooted tree that displays the rooted triples 12|3 and 13|6 but not 13|4 nor 15|4.

Given a collection $\mathcal{R}$ of rooted triples, let $L(\mathcal{R})$ denote the set of leaf labels that appear in at least one tree and let $\mathrm{co}(\mathcal{R})$ denote the set of rooted phylogenetic trees on leaf set $L(\mathcal{R})$ that display all the trees in $\mathcal{R}$. We say $\mathcal{R}$ is *compatible* if $\mathrm{co}(\mathcal{R})$ is non-empty.

Similar definitions apply for unrooted phylogenetic X-trees, however in this section and the next we deal only with rooted trees.

1.2. Closure of a set of rooted triples

Given a compatible collection $\mathcal{R}$ of rooted triples, we write $\mathcal{R} \vdash ab|c$ if every rooted phylogenetic tree that displays $\mathcal{R}$ also displays $ab|c$ (this is equivalent to requiring that $\mathcal{R} \cup \{ac|b\}$ is incompatible, and $\mathcal{R} \cup \{bc|a\}$ is incompatible).

If $\mathcal{R}$ is a compatible set of rooted triples, we define the *closure of* $\mathcal{R}$ by

$$\mathrm{cl}(\mathcal{R}) = \bigcap_{\mathcal{T} \in \mathrm{co}(\mathcal{R})} r(\mathcal{T}).$$

Equivalently, $\mathrm{cl}(\mathcal{R})$ is the set $\{ab|c : \mathcal{R} \vdash ab|c\}$. This operation satisfies the usual three properties of a closure operator, namely: $\mathcal{R} \subseteq \mathrm{cl}(\mathcal{R})$; $\mathrm{cl}(\mathrm{cl}(\mathcal{R})) = \mathrm{cl}(\mathcal{R})$ and if $\mathcal{R}_1 \subseteq \mathcal{R}_2$ are compatible, then $\mathrm{cl}(\mathcal{R}_1) \subseteq \mathrm{cl}(\mathcal{R}_2)$.

If $\mathcal{R}$ is incompatible, then one can also define a closure of $\mathcal{R}$ as follows. We say that a set of rooted triples (compatible or not) $\mathcal{R}^*$ is *closed* if for every subset $\mathcal{R}' \subseteq \mathcal{R}^*$ such that $\mathcal{R}'$ is compatible, $\mathrm{cl}(\mathcal{R}') \subseteq \mathcal{R}^*$. In particular the set $\mathcal{R}(X)$ of all $3\binom{n}{3}$ rooted triples on X is closed, and so given a set $\mathcal{R} \subseteq \mathcal{R}(X)$ we can define the *closure* of $\mathcal{R}$, denoted $\mathrm{Cl}(\mathcal{R})$ to be the intersection of all closed sets containing $\mathcal{R}$. This also satisfies the three properties of a closure operator, and when $\mathcal{R}$ is compatible we have $\mathrm{Cl}(\mathcal{R}) = \mathrm{cl}(\mathcal{R})$.

The closure operation provides a neat characterization of compatibility as the following Lemma shows. The result is a slight strengthening of Proposition 9(2) of [4] and is established by the same argument used in that result.

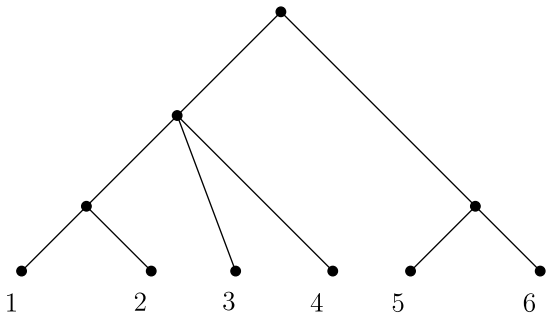

Fig. 1. A rooted phylogenetic tree $\mathcal{T}$ that displays 12|3 and 13|6 but not 13|4 nor 15|4.

Lemma 1.1. *Let $\mathcal{R}$ be a set of rooted triples. Then $\mathcal{R}$ is incompatible if and only if there exists a set $\mathcal{R}' \subset \mathcal{R}$ such that for every rooted triple $ab|c \in \mathcal{R} - \mathcal{R}'$ either $\mathcal{R}' \vdash ac|b$ or $\mathcal{R}' \vdash bc|a$.*

We will use this lemma later in the paper; however we pause to note an application of it now that is relevant to supertree reconstruction. Given two sets of rooted triples $\mathcal{R}_1, \mathcal{R}_2$, let

$$[\mathcal{R}_1, \mathcal{R}_2] := \{ab|c \in \mathcal{R}_1 : \text{there does not exist } \mathcal{R}' \subseteq \mathcal{R}_2 : \ \mathcal{R}' \vdash ac|b \text{ or } \mathcal{R}' \vdash bc|a\}.$$

Proposition 1.2. *Let $\mathcal{R}_1$ and $\mathcal{R}_2$ be two sets of rooted triples (compatible or not) for which $\mathcal{R}_1 \subseteq \mathcal{R}_2$. Then $[\mathcal{R}_1, \mathcal{R}_2]$ is compatible. In particular $[\mathcal{R}_1, \mathcal{R}_1]$ is compatible.*

Proof. Suppose $[\mathcal{R}_1, \mathcal{R}_2]$ were incompatible. By Lemma 1.1 there would exist a set $\mathcal{R}' \subseteq [\mathcal{R}_1, \mathcal{R}_2] \subseteq \mathcal{R}_1 \subseteq \mathcal{R}_2$ and a rooted triple $ab|c \in [\mathcal{R}_1, \mathcal{R}_2] - \mathcal{R}' \subseteq \mathcal{R}_2$ such that either $\mathcal{R}' \vdash ac|b$ or $\mathcal{R}' \vdash bc|a$. However, this implies that $ab|c \notin [\mathcal{R}_1, \mathcal{R}_2]$, a contradiction. □

For example, for any set $\mathcal{R}$ of rooted triples (compatible or not) we could take $\mathcal{R}_1 = \mathcal{R}_2 = \mathrm{Cl}(\mathcal{R})$, and Proposition 1.2 would ensure that $[\mathrm{Cl}(\mathcal{R}), \mathrm{Cl}(\mathcal{R})]$ is compatible. This is relevant for a desired property for supertree methods, described semi-formally in [11] as: 'the property of [the output tree] displaying $x|yz$ if it is found in some input tree or implied by some combination of input trees and no input tree or combination of input trees displays or implies $y|xz$ or $z|xy$'.

2. Minimal sets whose closure gives all the information in a tree

For every phylogenetic tree $\mathcal{T}$, the set $r(\mathcal{T})$ of all rooted triples displayed by $\mathcal{T}$ is closed. However, in general there exist subsets $\mathcal{R}$ of $r(\mathcal{T})$ with $\mathrm{cl}(\mathcal{R}) = r(\mathcal{T})$. For example, the set $\{12|3,12|4,13|5,34|5,56|1\}$ has this property for the tree depicted in Fig. 1. In this section, we will compute a tight lower bound for the cardinality of such a set $\mathcal{R}$.

Before exploring this further, we first note that this question has an equivalent reformulation.

Definition. A collection of rooted triples *identifies* a rooted phylogenetic X-tree $\mathcal{T}$ if $\mathcal{T}$ displays $\mathcal{R}$ and every other tree that displays $\mathcal{R}$ is a refinement of $\mathcal{T}$. That is, $\mathrm{co}(\mathcal{R}) = \{\mathcal{T}' : \mathcal{T} \leqslant \mathcal{T}', L(\mathcal{T}) = L(\mathcal{T}')\}$. In view of the following Lemma our problem is to determine the smallest number of rooted triples needed to identify $\mathcal{T}$.

Lemma 2.1. *For any subset $\mathcal{R}$ of $r(\mathcal{T})$, $\mathrm{cl}(\mathcal{R}) = r(\mathcal{T})$ iff $\mathcal{R}$ identifies $\mathcal{T}$.*

Proof. We have $\mathrm{cl}(\mathcal{R}) = \bigcap_{\mathcal{T}' \in \mathrm{co}(\mathcal{R})} r(\mathcal{T}')$. Thus, $\mathrm{cl}(\mathcal{R}) = r(\mathcal{T})$ precisely if $r(\mathcal{T}) \subseteq r(\mathcal{T}')$ for all $\mathcal{T}' \in \mathrm{co}(\mathcal{R})$. But $r(\mathcal{T}) \subseteq r(\mathcal{T}')$ iff $\mathcal{T} \leqslant \mathcal{T}'$, and so $\mathrm{cl}(\mathcal{R}) = r(\mathcal{T})$ iff $\mathcal{T} \leqslant \mathcal{T}'$ for all $\mathcal{T}' \in \mathrm{co}(\mathcal{R})$, precisely the requirement for $\mathcal{R}$ to identify $\mathcal{T}$. □

To proceed further we need to introduce some further definitions. Regarding a rooted phylogenetic tree $\mathcal{T}$ as a directed graph (with arcs oriented away from the root), and given $v \in V(\mathcal{T})$, the *descendents of* v, denoted $\mathrm{des}_{\mathcal{T}}(v)$, is the set of leaves that can be reached via a directed path in $\mathcal{T}$ starting at v. We simply write $\mathrm{des}(v)$ when $\mathcal{T}$ is clear from context.

A rooted triple $ab|c$ *distinguishes* an edge (u,v) in $\mathcal{T}$ if and only if $a,b,c \in \mathrm{des}(u)$, $a,b \in \mathrm{des}(v)$, and $c \notin \mathrm{des}(v)$. For example, in Fig. 1 the rooted triple $12|4$ distinguishes the interior edge of $\mathcal{T}$ that is not incident with the root.

If $\mathcal{R}$ identifies $\mathcal{T}$, then it is clearly necessary that for each internal edge of $\mathcal{T}$, $\mathcal{R}$ contains at least one rooted triple that distinguishes that edge. For a rooted binary phylogenetic tree, this condition is also sufficient, thus for a binary tree on n leaves, a set of cardinality $n-2$ (one rooted triple for each of the $n-2$ internal edges) is enough to identify the tree [19]. As noted in [20], 'calculating the number of absolutely independent triples for non-binary trees is more complex, depending on the degree and level of resolution of the tree.' We will establish a lower bound on the number of rooted triples needed to identify a tree, and show that this lower bound can actually be realized for any tree. First we recall a useful construction in classical phylogenetics.

Given a compatible collection $\mathcal{R}$ of rooted triples, there is a well-known and canonical construction of a tree denoted $\mathcal{A}_{\mathcal{R}}$ which displays $\mathcal{R}$ due to Aho et al. [1]. There is a polynomial-time procedure which constructs the clusters of this tree recursively from $\mathcal{R}$ (readers unfamiliar with this construction may wish to consult [18]). The basis of this algorithm is the following graph, which can be constructed from any set $\mathcal{R}$ of rooted triples (compatible or not) and which we denote as $G(\mathcal{R})$. The set of vertices of this graph is $L(\mathcal{R})$, the set of leaf labels of the elements of $\mathcal{R}$. There is an edge between two vertices a and b, if there is $c \in L(\mathcal{R})$ such that $ab|c \in \mathcal{R}$. This graph, $G(\mathcal{R})$, is called the clustering graph in [18]. The components of this graph form the maximal clusters of the tree $\mathcal{A}_{\mathcal{R}}$, and the algorithm for constructing the clusters in the remainder of $\mathcal{A}_{\mathcal{R}}$ proceeds recursively by restricting $\mathcal{R}$ to the leaf labels within each component (for details see [18]). The following result is from [16].

Lemma 2.2. *If $\mathcal{R}$ identifies $\mathcal{T}$ then $\mathcal{A}_{\mathcal{R}} = \mathcal{T}$.*

As well as $G(\mathcal{R})$ we will require a further graph in the arguments that follow. Let $\mathcal{R}$ be a set of rooted triples, and let V and U be sets of subsets of $L(\mathcal{R})$. We define an edge-labelled graph $G(\mathcal{R}, V, U)$ as follows. Take the vertices of the graph to be the elements of V. Add an edge between two vertices v and v' if there is a rooted triple $ab|c \in \mathcal{R}$ such that $a \in v$, $b \in v'$, and $c \in u$ for some $u \in U$. Label each edge $\{v,v'\}$ with the set $\{u \in U : \exists\ ab|c \in \mathcal{R}$ such that $a \in v, b \in v',$ and $c \in u\}$. If $V = U = \{\{x\} : x \in L(\mathcal{R})\}$, then $G(\mathcal{R},V,U)$ is simply $G(\mathcal{R})$ with edge labels as defined in [4].

For a rooted phylogenetic tree $\mathcal{T}$ with $L(\mathcal{T}) \subseteq L(\mathcal{R})$ and $(u,v) \in \mathring{E}(\mathcal{T})$, let

$$G(\mathcal{R}, \mathcal{T}, (u,v)) := G(\mathcal{R}, V, U),$$

where $V = \{\mathrm{des}(x) : (v,x) \in E(\mathcal{T})\}$ and $U = \{\mathrm{des}(w) : (u,w) \in E(\mathcal{T}),\ w \neq v\}$. Furthermore, for a vertex w of $\mathcal{T}$ such that $(u,w) \in \mathring{E}(\mathcal{T})$, and $w \neq v$, we let $G_w(\mathcal{R}, \mathcal{T}, (u,v))$ denote the subgraph of $G(\mathcal{R}, \mathcal{T}, (u,v))$ with the same vertex set and only those edges which have w in their label set. For a subset $L' \subseteq L(\mathcal{R})$, we denote the set of all triples in $\mathcal{R}$ that have all leaves in L' by $\mathcal{R}|_{L'}$ and for a graph G and a subset V' of its vertex set, $G[V']$ is the subgraph of G induced by V'.

Lemma 2.3. *If $\mathcal{R}$ is a set of rooted triples and $(u,v) \in \mathring{E}(\mathcal{A}_{\mathcal{R}})$, then $G(\mathcal{R}, \mathcal{A}_{\mathcal{R}}, (u,v))$ is connected.*

Proof. By the construction of $\mathcal{A}_{\mathcal{R}}$, $\mathrm{des}(v)$ is the vertex set of a connected component of $G(\mathcal{R}|_{\mathrm{des}(u)})$. We will show that, ignoring edge labels, $G(\mathcal{R}, \mathcal{A}_{\mathcal{R}}, (u,v))$ can be obtained from $G(\mathcal{R}|_{\mathrm{des}(u)})[\mathrm{des}(v)]$ by simply identifying vertices. Let G^* be the graph obtained from $G(\mathcal{R}|_{\mathrm{des}(u)})[\mathrm{des}(v)]$ by identifying all vertices that are in the same connected component of $G(\mathcal{R}|_{\mathrm{des}(v)})$. Clearly, G^* and

$G(\mathcal{R}, \mathcal{A}_\mathcal{R}, (u,v))$ have the same vertex set, say $B_1, B_2, \ldots, B_{d^+(v)}$ where $d^+(v)$ denotes the outdegree of v in $\mathcal{A}_\mathcal{R}$. If B_iB_j is an edge of $G(\mathcal{R}, \mathcal{A}_\mathcal{R}, (u,v))$, then there exists $a \in B_i$, $b \in B_j$, and $c \in \mathrm{des}(u)$ such that $ab|c \in \mathcal{R}$. Therefore, B_iB_j is also an edge of G^*. Suppose that B_iB_j is an edge of G^* but is not in the edge set of $G(\mathcal{R}, \mathcal{A}_\mathcal{R}, (u,v))$. Then there must be some $a \in B_i$, $b \in B_j$, and $c \in \mathrm{des}(u)$ such that $ab|c \in \mathcal{R}$ and $c \notin \mathrm{des}(u) - \mathrm{des}(v)$. Thus $c \in \mathrm{des}(v)$. This contradicts the fact that a and b are in distinct connected components of $G(\mathcal{R}|_{\mathrm{des}(v)})$. Therefore, G^* and $G(\mathcal{R}, \mathcal{A}_\mathcal{R}, (u,v))$ have the same edge set and we conclude that $G(\mathcal{R}, \mathcal{A}_\mathcal{R}, (u,v))$ is connected. □

Lemma 2.4. *If $\mathcal{R}$ is a set of rooted triples that identifies $\mathcal{A}_\mathcal{R}$, then, for every two edges $(u,v) \in \mathring{E}(\mathcal{A}_\mathcal{R})$ and $(u,w) \in E(\mathcal{A}_\mathcal{R})$ with $w \neq v$, the graph $G_w(\mathcal{R}, \mathcal{A}_\mathcal{R}, (u,v))$ is connected.*

Proof. If $d^+(u) = 2$, then $G_w(\mathcal{R}, \mathcal{A}_\mathcal{R}, (u,v)) = G(\mathcal{R}, \mathcal{A}_\mathcal{R}, (u,v))$. Thus, by Lemma 2.3, $G_w(\mathcal{R}, \mathcal{A}_\mathcal{R}, (u,v))$ is connected.

Now consider the case where $d^+(u) > 2$. Suppose $G_w(\mathcal{R}, \mathcal{A}_\mathcal{R}, (u,v))$ is not connected and let $C_1, \ldots, C_k$ be its components with more than one vertex. Note that $k = 0$ if all vertices are isolated. Let $\mathcal{T}$ be the tree obtained from $\mathcal{A}_\mathcal{R}$ by adding vertices $x_1, \ldots, x_k$, replacing all edges (v,y_i) for which $\mathrm{des}(y_i)$ is a vertex of C_i by an edge (x_i,y_i) and adding an edge (v,x_i) for every $i \in \{1,\ldots,k\}$, and replacing the edge (u,w) by (v,w). Suppose that $\mathcal{T}$ does not display $\mathcal{R}$. Then there is a rooted triple $ab|c \in \mathcal{R}$ which is displayed by $\mathcal{A}_\mathcal{R}$ but not by $\mathcal{T}$. This implies that $c \in \mathrm{des}(w)$, $a,b \in \mathrm{des}(v)$, and a and b are contained in vertices of different components of $G_w(\mathcal{R}, \mathcal{A}_\mathcal{R}, (u,v))$ which is impossible in view of the definition of that graph. Thus, $\mathcal{T}$ displays $\mathcal{R}$. This is a contradiction since $\mathcal{T}$ is not a resolution of $\mathcal{A}_\mathcal{R}$ but we assumed that $\mathcal{R}$ identifies $\mathcal{A}_\mathcal{R}$. Therefore, $G_w(\mathcal{R}, \mathcal{A}_\mathcal{R}, (u,v))$ is connected. □

Theorem 2.5. *Given a rooted phylogenetic X-tree $\mathcal{T}$, and a set of rooted triples $\mathcal{R}$ with $L(\mathcal{R}) = X$, we have $\mathcal{A}_\mathcal{R} = \mathcal{T}$ if and only if the following two conditions hold*:

(i) $\mathcal{R} \subseteq r(\mathcal{T})$ *and*
(ii) $\forall (u,v) \in \mathring{E}(\mathcal{T})$, $G(\mathcal{R}, \mathcal{T}, (u,v))$ *is connected.*
Furthermore, $\mathcal{R}$ identifies $\mathcal{T}$ if and only if in addition to (i) and (ii), the following condition holds.
(iii) $\forall (u,v) \in \mathring{E}(\mathcal{T})$ *and for each* $(u,w) \in E(\mathcal{T})$ *with* $w \neq v$, $G_w(\mathcal{R}, \mathcal{T}, (u,v))$ *is connected.*

Proof. Assume that $\mathcal{A}_\mathcal{R} = \mathcal{T}$. Then $\mathcal{R} \subseteq r(\mathcal{A}_\mathcal{R}) = r(\mathcal{T})$, satisfying condition (i). By Lemma 2.3, condition (ii) is also satisfied. To prove the converse (i.e. conditions (i) and (ii) imply $\mathcal{A}_\mathcal{R} = \mathcal{T}$) we will use induction on the number of internal edges of $\mathcal{T}$. The result clearly holds for trees with exactly one internal edge. Let $\mathcal{T}$ be a tree with $|\mathring{E}(\mathcal{T})| > 1$. Assume the result holds for any tree with less than $|\mathring{E}(\mathcal{T})|$ leaves.

We assume that $\mathcal{A}_\mathcal{R} \neq \mathcal{T}$. Hence, there are edges (u,v) of $\mathcal{T}$ and (u',v') of $\mathcal{A}_\mathcal{R}$ such that $\mathrm{des}_\mathcal{T}(u) = \mathrm{des}_{\mathcal{A}_\mathcal{R}}(u')$ and $\mathrm{des}_\mathcal{T}(v) \neq \mathrm{des}_{\mathcal{A}_\mathcal{R}}(v')$ and $\mathrm{des}_\mathcal{T}(v) \cap \mathrm{des}_{\mathcal{A}_\mathcal{R}}(v') \neq \emptyset$. Let $\mathcal{T}_v$ be the subtree of $\mathcal{T}$ with root v and leaf set $\mathrm{des}(v)$. Clearly, we have $\mathcal{R}|_{\mathrm{des}(v)} \subseteq r(\mathcal{T}_v)$ and that $G(\mathcal{R}_{\mathrm{des}(v)}, \mathcal{T}_v, (w_1,w_2))$ is connected for every edge $(w_1,w_2) \in \mathring{E}(\mathcal{T}_v)$. By induction hypothesis, we have $\mathcal{T}_v = \mathcal{A}_{\mathcal{R}|_{\mathrm{des}(v)}}$. Hence, for every edge (v,w) of $\mathcal{T}$, $\mathrm{des}(w)$ is contained in one connected component of $G(\mathcal{R}|_{\mathrm{des}(v)})$ and therefore in one component of $G(\mathcal{R}|_{\mathrm{des}(u)})$. Further, since

$G(\mathcal{R},\mathcal{T},(u,v))$ is connected, even $\mathrm{des}(v)$ is contained in one connected component of $G(\mathcal{R}|_{\mathrm{des}(u)}) = G(\mathcal{R}|_{\mathrm{des}(u')})$, thus we have $\mathrm{des}(v) \subseteq \mathrm{des}(v')$. Since $\mathrm{des}(v) \neq \mathrm{des}(v')$ there is an edge $\{x_1,x_2\}$ in the connected graph $G(\mathcal{R}|_{\mathrm{des}(u)})[\mathrm{des}(v')]$ with $x_1 \in \mathrm{des}(v)$ and $x_2 \notin \mathrm{des}(v)$. Hence, there is $y \in \mathrm{des}(u)$ with $x_1x_2|y \in \mathcal{R}$ but this rooted triple is not displayed by $\mathcal{T}$, a contradiction. This proves that $\mathcal{T} = \mathcal{A}_{\mathcal{R}}$, hence the first result of this theorem.

Now we will prove the second result that conditions (i)–(iii) are necessary and sufficient for $\mathcal{R}$ to identify $\mathcal{T}$. If $\mathcal{R}$ identifies $\mathcal{T}$ then $\mathcal{T} = \mathcal{A}_{\mathcal{R}}$. Thus, conditions (i) and (ii) follow from the first part of this theorem and condition (iii) follows from Lemma 2.4.

Assume that conditions (i)–(iii) hold. By the first part of this theorem, $\mathcal{T} = \mathcal{A}_{\mathcal{R}}$. Suppose that $\mathcal{R}$ does not identify $\mathcal{T}$. It was shown in [6], p. 45, that then there are edges $(u,v),(u,w) \in E(\mathcal{T})$ such that $v \neq w$ and $G(\mathcal{R}|_{\mathrm{des}(v)\cup\mathrm{des}(w)})$ has more than two connected components. We know that for each connected component C of $G(\mathcal{R}|_{\mathrm{des}(v)\cup\mathrm{des}(w)})$ either $V(C) \subseteq \mathrm{des}(v)$ or $V(C) \subseteq \mathrm{des}(w)$. Assume without loss of generality that the vertices of at least two of the connected components of $G(\mathcal{R}|_{\mathrm{des}(v)\cup\mathrm{des}(w)})$ are subsets of $\mathrm{des}(v)$. Then the graph $G_w(\mathcal{R},\mathcal{A}_{\mathcal{R}},(u,v))$ can not be connected, in contradiction to the assumption that (iii) holds. Therefore, $\mathcal{R}$ must identify $\mathcal{T}$. □

Notice that the graphs in conditions (ii) and (iii) in Theorem 2.5 depend only on those rooted triples that distinguish an edge. Therefore, the following corollary is immediate.

Corollary 2.6. *If $\mathcal{R}$ is a minimal set of rooted triples identifying $\mathcal{T}$ then each element of $\mathcal{R}$ distinguishes an internal edge of $\mathcal{T}$.*

We are now ready to establish a lower bound on the number of rooted triples needed to identify a tree. For a rooted phylogenetic tree $\mathcal{T}$, we define

$$lb(\mathcal{T}) = \sum_{(u,v)\in\mathring{E}(\mathcal{T})} (d^+(v)-1)(d^+(u)-1).$$

Theorem 2.7. *If $\mathcal{R}$ is a set of rooted triples that identifies $\mathcal{T}$, then $|\mathcal{R}| \geqslant \mathrm{lb}(\mathcal{T})$.*

Proof. Let $\mathcal{R}$ be minimal set of rooted tiples identifying $\mathcal{T}$ by Corollary 2.6, each element of $\mathcal{R}$ distinguishes exactly one internal edge of $\mathcal{T}$. For each internal edge (u,v) of $\mathcal{T}$, let $\pi_{(u,v)}$ be the set of elements of $\mathcal{R}$ that distinguish (u,v). Then $\{\pi_{(u,v)} : (u,v) \in \mathring{E}(\mathcal{T})\}$ is a partition of $\mathcal{R}$ and

$$|\mathcal{R}| = \sum_{(u,v)\in\mathring{E}(\mathcal{T})} |\pi_{(u,v)}|.$$

We will show that for every internal edge $(u,v) \in \mathcal{T}$,

$$|\pi_{(u,v)}| \geqslant (d^+(v)-1)(d^+(u)-1).$$

By Lemma 2.4, for $(u,w) \in \mathring{E}(\mathcal{T})$ such that $w \neq v$, the graph $G_w = G_w(\mathcal{R},\mathcal{T},(u,v))$ is connected. Hence,

$$|E(G_w)| \geqslant |V(G_w)| - 1 = d^+(v) - 1.$$

Let $w_1, w_2, \ldots, w_{d^+(u)-1}$ be the vertices of $\mathcal{T}$ such that $(u,w) \in \mathring{E}(\mathcal{T})$ and $w \neq v$. Then we have

$$|\pi_{(u,v)}| \geqslant \sum_{i=1}^{d^+(u)-1} |E(G_{w_i}(\mathcal{R},\mathcal{T},(u,v)))| \geqslant (d^+(v)-1)(d^+(u)-1).$$

This establishes the lower bound on $|\mathcal{R}|$. □

Now we will show that the lower bound from Theorem 2.7 can be attained for every tree $\mathcal{T}$.

Theorem 2.8. *For every rooted phylogenetic tree $\mathcal{T}$, there is a set $\mathcal{R}$ of rooted triples such that $\mathcal{R}$ identifies $\mathcal{T}$ and $|\mathcal{R}| = \mathrm{lb}(\mathcal{T})$.*

Proof. We prove the theorem by constructing a set $\mathcal{R}$ of rooted triples with the desired property. For each $(u,v) \in \mathring{E}(\mathcal{T})$, we choose a set of rooted triples $\pi_{(u,v)}$ in the following manner. Let $w_1, w_2, \ldots, w_{d^+(v)}$ be the children of v and, for $i \in \{1,\ldots,d^+(v)\}$, let $x_i \in \mathrm{des}(w_i)$. Further, let $y_1, \ldots, y_{d^+(u)}$ be the children of u with $y_{d^+(u)} = v$ and, for $i \in \{1,\ldots,d^+(u)-1\}$, let $z_i \in \mathrm{des}(y_i)$.

$$\pi_{(u,v)}| = \{x_i x_i + 1 | z_j : 1 \leqslant i \leqslant d^+(v) - 1 \text{ and } 1 \leqslant j \leqslant d^+(u) - 1\}.$$

Let $\mathcal{R} = \bigcup_{(u,v)\in\mathring{E}(\mathcal{T})} \pi_{(u,v)}$. By construction, $\mathcal{R}$ fulfils conditions (i)–(iii) of Theorem 2.5, thus $\mathcal{R}$ identifies $\mathcal{T}$ and $|\mathcal{R}| = \mathrm{lb}(\mathcal{T})$. □

For the phylogenetic tree depicted in Fig. 1, the construction above yields the minimum identifying set {12|3,12|4,13|5,34|5,56|1} of rooted triples.

3. The closure operation for unrooted trees

Up to this point we have considered the closure operation on rooted trees. When we move to unrooted trees, many of the results one might expect to carry over do not. Perhaps most surprising is that Lemma 1.1 is no longer true in the unrooted setting, thereby settling a question posed in [5]. To explain this result we begin with some terminology.

Following [16], an *unrooted phylogenetic X-tree* $\mathcal{T}$ is a tree with leaf set X and whose interior vertices are unlabelled and of degree at least 3 (in case all these degrees equal 3 we say that $\mathcal{T}$ is *binary*).

An unrooted phylogenetic tree $\mathcal{T}$ is said to be *induced* by the tree $\mathcal{T}'$ if the leaf set of $\mathcal{T}$ is a subset of the leaf set of $\mathcal{T}'$ and $\mathcal{T}$ is obtained from the maximal subgraph of $\mathcal{T}'$ containing the leaf set of $\mathcal{T}$ by suppressing vertices of degree 2. An unrooted binary phylogenetic tree with four leaves is called a *quartet tree*. A phylogenetic tree $\mathcal{T}$ *displays* a set Q of quartet trees if every quartet tree in Q is induced by $\mathcal{T}$ and we let $\mathrm{co}(Q)$ denote the set of phylogenetic X-trees that display Q where X is the set of labels appearing at leaves in Q.

The quartet tree with leaf set $\{a,b,c,d\}$ that contains an inner edge separating a, b from c, d is denoted by $ab|cd$ and $\{a,b\}$ and $\{c,d\}$ are called its *quartet halfs*. A set of quartet trees is said to be *compatible* if there is an unrooted phylogenetic tree inducing all of those rooted quartet trees. Furthermore, we say that a set Q of quartet trees is *closed* if it has the property that Q contains every quartet tree that is displayed by every tree in $\mathrm{co}(Q')$ for each compatible non-empty subset Q' of Q. If each subset of Q is closed we call Q *strongly closed*. For example, for a given unrooted phylogenetic tree $\mathcal{T}$ having at least two internal edges, the set of all quartets displayed by $\mathcal{T}$ is a closed set but it is not strongly closed.

The following result is a slight re-statement of Proposition 9(2) of [4]. It follows easily from Lemma 1.1.

Proposition 3.1. *Every closed non-compatible rooted triple set contains a conflict, i.e. two different rooted triples of the same set of three leaves.*

Surprisingly, the analogue of Proposition 3.1 fails for quartet trees; that is, there exists a non-compatible closed quartet tree set which does not contain a conflict (two different quartet trees of the same set of four leaves). This resolves a question raised in [4], and disproves a conjecture from [3]. In fact we can establish a slightly stronger result by replacing 'closed' by 'strongly closed'.

Theorem 3.2. *Let*

$$W := \{12|78, 23|58, 15|37, 14|67, 26|48, 34|56\}.$$

Then W is a non-compatible strongly closed set of quartet trees without a conflict.

Proof. The proof is divided into three parts: First we show that W is not compatible, then we prove that W is closed, and finally we show that W is even strongly closed.

Assume there is a tree $\mathcal{T}_W$ that displays W. Since $\mathcal{T}_W$ displays $\{12|78, 23|58, 15|37\}$ the tree with leaf set $\{1,2,3,5,7,8\}$ that is induced by $\mathcal{T}_W$ must be either $\mathcal{T}_1$ or $\mathcal{T}_2$, as shown in Fig. 2. Further, since $\mathcal{T}_W$ displays $\{12|78, 14|67, 26|48\}$ the tree with leaf set $\{1,2,4,6,7,8\}$ that is induced by $\mathcal{T}_W$ has to be either $\mathcal{T}_3$ or $\mathcal{T}_4$ of Fig. 2. However, every tree that induces $\mathcal{T}_1$ and $\mathcal{T}_3$ or $\mathcal{T}_2$ and $\mathcal{T}_4$ displays the quartet tree 35|46 while every tree that induces $\mathcal{T}_1$ and $\mathcal{T}_4$ or $\mathcal{T}_2$ and $\mathcal{T}_3$ displays 45|36. Hence, a tree that induces one of $\mathcal{T}_1$ and $\mathcal{T}_2$ and one of $\mathcal{T}_3$ and $\mathcal{T}_4$ and displays 34|56 can not exist.

Obviously, W contains at most one quartet tree of every quadruple of $\{1,\ldots,8\}$, so W does not contain a conflict and it remains to show that W is strongly closed. It suffices to prove that $W - q$ is compatible and strongly closed for every element $q \in W$. Every quartet tree in W can be interpreted as a pair of opposite edges of the cube where the vertices are labeled by $1,\ldots,8$ (see Fig. 3). For every two pairs $\{e_1, e_2\}$ and $\{f_1, f_2\}$ of opposite edges of the cube, there is a graph isomorphism that maps e_1 to f_1 and e_2 to f_2. Hence, it suffices to prove that $W - q$ is compatible and strongly closed for $q = 34|56$. This can be done by applying Proposition 5 of [4] which states that a rooted triple set (respectively, quartet tree set) W is compatible and closed if and only if

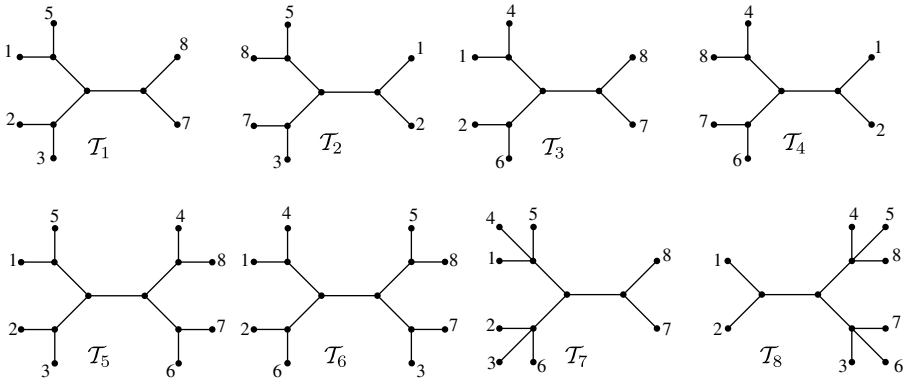

Fig. 2. The example trees used for the proof of Theorem 3.2.

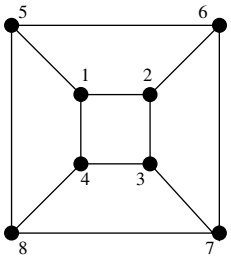

Fig. 3. The quartet set W represented as pairs of edges of the cube.

there is a collection C of phylogenetic X-trees such that W is the set of triples (respectively, quartet trees) displayed by all of the trees in C.

It can easily be checked by hand that

$$W' = \{15|26, 15|37, 23|14, 23|58, 48|26, 48|37, 67|14, 67|58, 12|78, 35|46\}$$

is the set of quartet trees displayed by $\mathcal{T}_5$ and $\mathcal{T}_6$ and the subset of those quartet trees of W' which are also displayed by $\mathcal{T}_7$ and $\mathcal{T}_8$ is exactly $W - 34|56$. Hence, $W - 34|56$ is closed.

For every quartet half $\{a,b\}$ of a quartet $q_1 \in W - 34|56$, there is no other quartet $q_2 \in W - 34|56$ such that $\{a,b\}$ is also a quartet half of q_2. Moreover, for every quartet $q_1 \in W - 34|56$, there are a quartet half $h(q_1)$ of q_1, a tree $\mathcal{T}(q_1) \in \{\mathcal{T}_5, \mathcal{T}_6, \mathcal{T}_7, \mathcal{T}_8\}$, and an edge $e(q_1)$ of $\mathcal{T}(q_1)$ such that $e(q_1)$ separates $h(q_1)$ from $X - h(q_1)$. For a fixed subset W' of $W - 34|56$, we identify the vertices incident with the edge $e(q')$ of $\mathcal{T}(q')$ for every quartet $q' \in W'$ and the set of quartets displayed by all obtained trees is $W - 34|56 - q'$. Hence, $W - 34|56$ and therefore W are strongly closed. □

4. Completeness of Meacham's rules for pairwise closure of characters

We turn now to a closure operation for partial X-splits, where the 'building blocks' are no longer small subtrees but rather splits of the leaf set of the input trees obtained by deleting edges of those trees. In this section we establish a completeness result for two closure rules described in 1983 by Christopher Meacham [14]. Informally, we show that, not only do these two rules produce all the 'information' that can be obtained from any pair of X-splits, but moreover they produce all the 'information' that can be generated from any pair of partitions of X.

To explain this more formally (and following the notation of [8]), we define a *character* (on X) to be a partition of X and a *split* to be a bipartition of X. A character χ is *displayed* by a phylogenetic tree $\mathcal{T}$ if there is a set E_χ of edges of $\mathcal{T}$ such that every part of χ is the set of labelled vertices of a component of the graph obtained from $\mathcal{T}$ by removing E_χ. A set $\mathcal{C}$ of characters is *displayed* by $\mathcal{T}$ if each of its elements is displayed by $\mathcal{T}$ and $\mathcal{C}$ is *compatible* if there is an X-tree that displays $\mathcal{C}$. A *partial split* is an unordered pair of disjoint non-empty subsets of X and the partial split $\{A,B\}$ is also denoted by $A|B$ or $B|A$. A partial split $A'|B'$ *refines* $A|B$ if $A \subseteq A'$ and $B \subseteq B'$ (or $A \subseteq B'$ and $B \subseteq A'$). A partial split χ is *displayed* by an X-tree $\mathcal{T}$ if $\mathcal{T}$ displays a split that refines χ and a set Σ of partial splits is *displayed* by $\mathcal{T}$ if every element of Σ is displayed by $\mathcal{T}$. Again, Σ is called *compatible* if there is an X-tree that displays Σ. Let $\mathcal{C}$ be a compatible set

of characters and Σ be a compatible set of partial splits. We say that $\mathcal{C}$ (respectively, Σ) *infers* a partial split $A|B$ and we write $\mathcal{C} \vdash A|B$ (respectively, $\Sigma \vdash A|B$) if every phylogenetic tree that displays $\mathcal{C}$ (respectively, Σ) also displays $A|B$.

Meacham [14] described two inference rules (referred to below as (M_1) and (M_2)) for the case that Σ contains exactly two partial splits, say $\Sigma = \{A_1|B_1, A_2|B_2\}$. These rules can be stated as follows:

(M_1) If $A_1 \cap A_2 \neq \emptyset$ and $B_1 \cap B_2 \neq \emptyset$ then
$\Sigma \vdash A_1 \cap A_2|B_1 \cup B_2$ and $\Sigma \vdash A_1 \cup A_2|B_1 \cap B_2$.

(M_2) If $A_1 \cap A_2 \neq \emptyset$ and $B_1 \cap B_2 \neq \emptyset$ and $A_1 \cap B_2 \neq \emptyset$ then
$\Sigma \vdash A_2|B_1 \cup B_2$ and $\Sigma \vdash A_1 \cup A_2|B_1$.

A set $\mathcal{C}$ of characters canonically defines a set

$$\Sigma(\mathcal{C}) := \{A|B : A, B \in \chi \in \mathcal{C}\}$$

of partial splits. It has been shown in [19] that all partial splits inferred by $\mathcal{C}$ are also inferred by $\Sigma(\mathcal{C})$, that is,

$$\mathcal{C} \vdash A|B \text{ if and only if } \Sigma(\mathcal{C}) \vdash A|B.$$

The main result of this section is that every partial split inferred by a compatible set $\mathcal{C}$ of two characters can be obtained by consecutively applying Meacham's inference rules to $\Sigma(\mathcal{C})$. Let θ be a non-empty subset of $\{1,2\}$ and let Σ be a compatible set of partial splits. We define $\mathrm{spcl}_\theta(\Sigma)$ to be the smallest set of partial splits Σ' such that every partial split in Σ is refined by a partial split in Σ' and every partial split $A|B$ that can be obtained from two partial splits in Σ' by applying a rule (M_i) for $i \in \theta$ is refined by a partial split in Σ'. It has been proved in [17] that all split closures $\mathrm{spcl}_\theta(\Sigma)$ for $\emptyset \neq \theta \subseteq \{1,2\}$ are well defined.

Theorem 4.1. *Let $\chi_1 = \{A_1, \ldots, A_k\}$ and $\chi_2 = \{B_1, \ldots, B_l\}$ be two compatible characters on X. Suppose that every X-tree that displays both χ_1 and χ_2, also displays the partial X-split $A|B$. Then there exists a partial X-split $A'|B'$ that refines $A|B$ such that $A'|B' \in spcl_{1,2}(\Sigma(\{\chi_1, \chi_2\}))$.*

Proof. An outline of the proof is as follows: First we define a graph from χ_1 and χ_2 that enables us to construct many different X-trees which all display χ_1 and χ_2. We will use those trees to show that every partial split inferred by χ_1 and χ_2 must belong to one of two disjoint special classes of partial splits. We conclude the proof by showing that all partial splits in one class can be obtained from $\Sigma(\{\chi_1, \chi_2\})$ by repeatedly applying (M_1) while all partial splits in the other class can be obtained by repeatedly applying (M_2).

Let G_I be the partition intersection graph of χ_1 and χ_2, i.e. the graph with vertex set $\{A_1, \ldots, A_k, B_1, \ldots, B_l\}$ where two vertices are connected by an edge if and only if they have a non-empty intersection. Since χ_1 and χ_2 are compatible characters it follows from [10] that the graph G_I does not contain a cycle. Let G_S be the graph obtained from G_I by subdividing every edge A_iB_j by a new vertex $A_i \cap B_j$. We denote the vertex set of G_I by V_I and the set of vertices of G_S which are not contained in V_I by V_S. We define ϕ: $X \to V(G_S)$ to be the mapping that maps every element of

X that is contained in an element of V_S to that vertex and every element $x \in X$ that is not contained in any element of V_S to the vertex in V_I that contains x. By definition, the pair $(T;\phi)$ is an X-tree that displays χ_1 and χ_2 for every tree T that is obtained from G_S by adding edges and then removing unlabelled leaves and suppressing unlabelled vertices of degree 2.

We will now prove that there are connected components C_A and C_B of G_S such that $\phi(A) \subseteq V(C_A)$ and $\phi(B) \subseteq V(C_B)$ hold. We assume that there are $a_1,a_2 \in A$ such that $\phi(a_1)$ and $\phi(a_2)$ are contained in different components of G_S and that $b \in B$. Then we can construct an X-tree from G_S which contains edges $\phi(a_i)\phi(b)$ for every $i \in \{1,2\}$ for which $\phi(a_i)$ and $\phi(b)$ are contained in different components of G_S. However, that tree can not display $A|B$ since the path from a_1 to a_2 contains b.

We have to distinguish whether the components C_A and C_B are equal. We start with the case $C_A = C_B := C$. We claim that there is a vertex $v \in V(C) \cap V_I$ such that there are different components C_A^v and C_B^v of $C - v$ with $\phi(A) \subseteq V(C_A^v)$ and $\phi(B) \subseteq V(C_B^v)$. Since every X-tree obtained from G_S by adding edges and then removing unlabelled leaves and suppressing unlabelled vertices of degree 2 displays $A|B$ there must be a cut edge in C separating $\phi(A)$ from $\phi(B)$. Without loss of generality we can assume that such an edge connects the vertices A_i and $A_i \cap B_j$ for some $i \in \{1,\ldots,k\}$ and $j \in \{1,\ldots,l\}$. We define G'_S to be the graph obtained from G_S by identifying the vertices $A_i,A_i \cap B_j,B_j$ where the new vertex is called $A_i \cup B_j$ and $\phi' : X \to V(G'_S)$ to be the mapping with $\phi'(x) = \phi(x)$ if $\phi(x) \notin \{A_i, A_i \cap B_j, B_j\}$ and $\phi'(x) = A_i \cup B_j$, else. Then every X-tree (T',ϕ') where T' is obtained from G'_S by adding edges and then removing unlabelled leaves and suppressing unlabelled vertices of degree 2 displays χ_1 and χ_2, thus it also displays $A|B$. Hence there is a cut edge of G'_S between $A_i \cup B_j$ and a vertex $u \in V_S \cap V(G'_S)$ that separates $\phi'(A)$ from $\phi'(B)$. Let $v \in \{A_i,B_j\}$ be the vertex that is in G_S adjacent to u and let w be the other vertex in $\{A_i, B_j\}$. Then one of the sets $\phi(A),\phi(B)$ is contained in the component C_1 of $C - v$ that contains u and the other one is contained in the component C_2 of $C - v$ that contains w. This proves the claim. Further, the partial split $\bigcup_{y \in V(C_1) \cap V_I} y | \bigcup_{z \in V(C_2) \cap V_I} z$ refines $A|B$.

Now we consider the case $C_A \neq C_B$. We claim that there is a vertex $u \in V_S$ with $\phi(A) = u$ or $\phi(B) = u$. We assume the contrary. We define P_A to be the set of all vertices v of $V(C_A) \cap V_I$ for which there are $a_1,a_2 \in A$ such that the path from $\phi(a1)$ to $\phi(a_2)$ (possibly of length 0) contains v. We define P_B correspondingly. By assumption, P_A and P_B are non-empty. If there is $i \in \{1,2\}$ such that $P_A \cup P_B \subseteq \chi_i$, then $|P_A| = |P_B| = 1$ holds since every path in G_S connecting two different elements of χ_1 contains an element of χ_2 and vice versa. Therefore, $P_A \cup P_B \subseteq \chi_1$ implies that there are $i,j \in \{1,\ldots,k\}$ with $i \neq j$ and $A \subseteq A_i$, $B \subseteq A_j$, but then $A_i|A_j \in \Sigma(\chi_1)$ refines $A|B$. Correspondingly, $P_A \cup P_B \subseteq \chi_2$ implies that there are $i,j \in \{1,\ldots,l\}$ with $i \neq j$ and $B_i|B_j \in \Sigma(\chi_2)$ refines $A|B$. Hence, we can assume that there are $u \in P_A$ and $v \in P_B$ such that each of χ_1 and χ_2 contains exactly one of the vertices u and v. Let G' be the graph obtained from G_S by identifying u and v where the new vertex is called w, and let $\phi' : X \to V(G'_S)$ be the mapping with $\phi'(x) = \phi(x)$ if $\phi(x) \notin \{u,v\}$ and $\phi'(x) = w$, else. Then every X-tree (T',ϕ') where T' is obtained from G'_S by adding edges and then removing unlabelled leaves and suppressing unlabelled vertices of degree 2 displays χ_1 and χ_2 but not $A|B$, a contradiction. This proves the claim. Let $i \in \{1,\ldots,k\}$ and $j \in \{1,\ldots,l\}$ such that $u = A_i \cap B_j$ and let $C' \in \{C_A,C_B\}$ such that C' does not contain u. Then the partial split $A_i \cap B_j | \bigcup_{z \in V(C') \cap V_I} z$ refines $A|B$.

We have shown that $A|B$ is either refined by a partial split $\bigcup_{y \in V(C_1)} y | \bigcup_{z \in V(C_2)} z$ such that C_1 and C_2 are two different components of $C - v$ where v is a vertex of a component C of G_I, or $A|B$ is

refined by a partial split $A_i \cap B_j|\bigcup_{z\in V(C')}z$ where A_iB_j is an edge of G_I and C' is a component of G_I that does not contain A_iB_j. The last step to prove the theorem is to show that every partial X-split of either of those two kinds is refined by a partial split in $\mathrm{spcl}_{1,2}(\Sigma(\{\chi_1,\chi_2\}))$.

Let C be a component of G_I, and let T be a subtree of C that contains a vertex v of degree 2. Let T_1 and T_2 be the two components of $C - v$. We claim that the partial split $\bigcup_{y\in V(T_1)}y|\bigcup_{z\in V(T_2)}z$ can be derived from $\Sigma(\{\chi_1,\chi_2\})$ by applying the second split closure rule. The claim is true if $|V(T_1)| = |V(T_2)| = 1$ since then the vertex in C_1 and the vertex in C_2 are contained in the same character. Assume that the claim is wrong and that $m := |V(T_1)| + |V(T_2)|$ is minimal with that property. Further, we assume $|V(T_1)| > 1$. Let w be the vertex of T_1 that is in C adjacent to v, and let $x \neq w$ be a leaf of T_1. The minimality of m implies that the partial split $\bigcup_{y\in V(T_1)-x}y|\bigcup_{z\in V(T_2)}z$ can be derived from $\Sigma(\{\chi_1,\chi_2\})$ by applying the second split closure rule. Let T_1' be the component of $T_1 - w$ that contains x. Since w has degree 2 in the subtree of C with vertex set $V(T_1') \cup \{v,w\}$ and $|V(T_1')| + 1 < m$ the partial split $v|\bigcup_{y\in V(T_1')}y$ can be derived from $\Sigma(\{\chi_1,\chi_2\})$ by applying the second split closure rule. Applying the second split closure rule to $\bigcup_{y\in V(T_1)-x}y|\bigcup_{z\in V(T_2)}z$ and $v|\bigcup_{y\in V(T_1')}y$ infers $\bigcup_{y\in V_1}y|\bigcup_{z\in V_2}z$, contradicting the assumption.

Let A_iB_j be an edge of G_I and let C' be a connected component of G_I that does not contain A_iB_j. If C' contains only one vertex x, then one of the partial splits $A_i|x$ and $B_j|x$ is contained in $\Sigma(\{\chi_1,\chi_2\})$ and refines $A_i \cap B_j|x$. Hence, we can assume that there is at least one edge uv in C'. We claim that the partial X-split $A_i \cap B_j|\bigcup_{z\in V(C')}z$ can be derived from $\Sigma(\{\chi_1,\chi_2\})$ by applying the first split closure rule. We assume that the claim is wrong and that U is a subtree of C' containing uv such that the partial split $A_i \cap B_j|\cup_{z\,\in\, V(U)}z$ can not be derived from $\Sigma(\{\chi_1,\chi_2\})$ by applying the first split closure rule and $|V(U)|$ is minimal under all subtrees with that property. If $|V(U)| = 2$, then we can assume that $u \in \chi_1$ and applying the first split closure rule to $A_i|u \in \Sigma(\{\chi_1\})$ and $B_j|v \in \Sigma(\{\chi_2\})$ infers $A_i \cap B_j|u \cup v$. Let $|V(U)| \geqslant 3$ and let $w \notin \{u,v\}$ be a leaf of U. By the minimality of $|V(U)|$, the partial split $A_i \cap B_j|\cup_{z\,\in\, V(U)-w}z$ can be derived from $\Sigma(\{\chi_1,\chi_2\})$ by applying the first split closure rule. Without loss of generality we can assume $w \in \chi_1$, thus $A_i|w \in \Sigma(\{\chi_1\})$. The first split closure rule applied to $A_i \cap B_j|\cup_{z\,\in\, V(U)-w}z$ and $A_i \cap B_j|w$ infers $A_i \cap B_j|\cup_{z\,\in\, V(U)}z$, in contradiction to the assumption. □

We remark that Theorem 4.1 also holds for *partial partitions*, i.e. χ_1 and χ_2 are partitions of subsets X_1 and X_2 of X and refining, displaying, and compatibility are defined as for partial splits. In that case we can assume $X = (\bigcup_{i=1}^{k}A_i) \cup (\bigcup_{j=1}^{l}B_j)$ and leave the proof unchanged. Further, the constructive proof of Theorem 4.1 provides the following result.

Corollary 4.2. *The set* $spcl_{1,2}(\Sigma(\{\chi_1,\chi_2\}))$ *can be computed in polynomial time*; *moreover* $spcl_{1,2}(\Sigma(\{\chi_1,\chi_2\})) = spcl_1(\Sigma\ (\{\chi_1,\chi_2\})) \cup spcl_2(\Sigma(\{\chi_1,\chi_2\}))$.

5. Closure operations in a general setting

Some of the concepts we have discussed concerning compatibility, inference rules and closure operations in the phylogenetic setting can extended to a more general setting, which may be useful for other applications (for example in problems concerning the reconstruction of linear orderings). We describe this viewpoint in this section.

Let Y be a (finite or infinite) set, let $r > 1$ be a positive natural number. We say that a subset W of $Y \times \{1,\ldots,r\}$ *contains no conflict* if it has the property:

$$(y,i),(y,j) \in W \Rightarrow i = j.$$

Suppose that $\mathcal{J}$ is a collection of subsets of $Y \times \{1,\ldots,r\}$ such that, for every $J \in \mathcal{J}$, each element $y \in Y$ occurs in exactly one element of J (thus J contains no conflict). We say that a subset W of $Y \times \{1,\ldots,r\}$ is $\mathcal{J}$-compatible if there exists a set $J \in \mathcal{J}$ so that $W \subseteq J$.

Example 5.1. To illustrate these ideas in a setting outside of phylogenetics (but more relevant to gene ordering on chromosomes) and with $r = 2$ let $Y = \{(i,j)\colon i,j \in \{1,\ldots,n\},\ i < j\}$ and for a bijection $f\colon\{1,\ldots,n\} \to \{1,\ldots,n\}$ let

$$W(f) := \{((i,j),s) \in Y \times \{1,2\} : f(i) < f(j) \iff s = 1\}.$$

Let $\mathcal{J}$ be the union of the sets $W(f)$ over all bijections f. Thus a subset $W \subset Y \times \{1,2\}$ is $\mathcal{J}$-compatible precisely if the pairwise orderings provided by W is consistent with a linear ordering of $\{1,\ldots,n\}$.

Given a $\mathcal{J}$-compatible set W define the *closure of* W (relative to $\mathcal{J}$) to be the collection of pairs $(y,i) \in Y \times \{1,\ldots,r\}$ for which, for all $j \in \{1,\ldots,r\} - \{i\}$ the set $W \cup \{(y,j)\}$ is not $\mathcal{J}$-compatible (this implies, in particular, that $W \cup \{(y,i)\}$ is $\mathcal{J}$-compatible). We say that a subset W of $Y \times \{1,\ldots,r\}$ is *closed* if the closure of every $\mathcal{J}$-compatible subset of W is contained in W. Finally, given a subset W of $Y \times \{1,\ldots,r\}$ we define the (generalized) *closure* of W (relative to $\mathcal{J}$) denoted $\mathrm{Cl}_{\mathcal{J}}(W)$ to be the intersection of all closed subsets of $Y \times \{1,\ldots,r\}$ that contain W (this is well-defined since $Y \times \{1,\ldots,r\}$ is closed).

Note that when W is $\mathcal{J}$-compatible, we have $\mathrm{cl}_{\mathcal{J}}(W) = \mathrm{Cl}_{\mathcal{J}}(W)$, so $\mathrm{Cl}_{\mathcal{J}}$ is an extension of $\mathrm{cl}_{\mathcal{J}}$. Also, $\mathrm{Cl}_{\mathcal{J}}$ satisfies the three properties one would expect of a closure operation, namely: $W \subseteq \mathrm{Cl}_{\mathcal{J}}(W)$, $V \subseteq W \Rightarrow \mathrm{Cl}_{\mathcal{J}}(V) \subseteq \mathrm{Cl}_{\mathcal{J}}(W)$ and $\mathrm{Cl}_{\mathcal{J}}(\mathrm{Cl}_{\mathcal{J}}(W)) = \mathrm{Cl}_{\mathcal{J}}(W)$.

Note that if Y is finite, then we can generate $\mathrm{Cl}_{\mathcal{J}}(W)$ as follows. Construct a sequence $W^{(1)}, W^{(2)}, \ldots,$ where, $W^{(1)} = W$ and for each $k \geqslant 1$, $W^{(k+1)}$ is the union of $\mathrm{cl}_{\mathcal{J}}(A)$ over all subsets A of $W^{(k)}$ that are $\mathcal{J}$-compatible. Then it is easily checked that $\mathrm{Cl}_{\mathcal{J}}(W) = \cup_{k=1}^{s} W^{(k)}$ for the first number s for which $W^{(s+1)} = W^{(s)}$.

The main question that we consider in this section (motivated by earlier results in this paper) is the following: how is the condition 'W is $\mathcal{J}$–compatible' related to the condition '$\mathrm{Cl}_{\mathcal{J}}(W)$ has no conflict'? The next lemma shows that the former condition implies the latter. We will then consider the reverse implication.

Lemma 5.2. *If W is $\mathcal{J}$-compatible then $\mathrm{Cl}_{\mathcal{J}}(W)$ contains no conflict.*

Proof. First note that if $J \in \mathcal{J}$ then J is $\mathcal{J}$-compatible, and the closure of J is J; in particular J is closed. Now if W is $\mathcal{J}$-compatible, then, by definition there exists a set $J \in \mathcal{J}$ with $W \subseteq J$. Since J is closed, the (generalized) closure of W is a subset of J. Finally, since $J \in \mathcal{J}$ and elements of $\mathcal{J}$ contain no conflict, $\mathrm{Cl}_{\mathcal{J}}(W)$ contains no conflict. □

We now describe two examples to show that the converse to Lemma 5.2 does not hold.

Example 5.3. The first example shows how this abstract framework is related to the quartet set from Theorem 3.2. We define $Y = \{\{i,j,k,l\} \subseteq \{1,\ldots,n\} : |\{i,j,k,l\}| = 4\}$ and, for $y = \{i,j,k,l\} \in Y$ with

$i < j < k < l$, we define $(y,1),(y,2),(y,3)$ to represent the quartet $ij|kl$, $ik|jl$, $il|jk$, respectively. Further, we define $\mathcal{J}$ to be the set of all quartet sets of binary phylogenetic trees on $\{1,\ldots,n\}$. With these definitions, $\mathrm{Cl}_{\mathcal{J}}(W)$ is the set of all closed quartet sets and the quartet set W from Theorem 3.2 has the property that W is not $\mathcal{J}$-compatible yet $\mathrm{Cl}_{\mathcal{J}}(W)(=W)$ so that W is closed and contains no conflict.

Example 5.4. The phenomenon described in Example 5.3 – whereby a set that is closed with respect to $\mathcal{J}$ and contains no conflict can fail to be $\mathcal{J}$-compatible – can be demonstrated by a more contrived example, though one for which the verification is much easier. Let Y be defined as in Example 5.3. We define a quartet tree $ab|cd$ to be a *crossing* for a cyclic ordering if the lines from a to b and from c to d cross each other in a cycle realizing that cyclic ordering. Let $\mathcal{J}$ contain the sets of all crossings for some cyclic ordering. W represents an arbitrary subset of quartet trees, and W is $\mathcal{J}$-compatible precisely if there is a cyclic ordering of $\{1,\ldots,n\}$ such that every element of W is a crossing. Let $n=5$ and $W=\{12|34,12|35,12|45\}$.

Proposition 5.5. *The set W defined in Example 5.4 is $\mathcal{J}$-compatible, yet W is closed and contains no conflict.*

Proof. For a quartet tree $ab|cd$ and a cyclic ordering of $\{1,\ldots,n\}$, the straight lines from a to b and from c to d cross each other if one of the two paths from a to b contains c and the other one contains d. Hence, if the straight line from 1 to 2 crosses the lines from 3 to 4 and from 3 to 5 then there must be a path from 1 to 2 containing 4 and 5 implying that the line from 4 to 5 does not cross the line from 1 to 2. This proves W is not $\mathcal{J}$-compatible. On the other hand, for both cyclic orderings 13245 and 13254, the quartet trees 12|34 and 12|35 are crossings, but different quartet trees are crossings for each of the remaining quadruples $\{1,2,4,5\}$, $\{1,3,4,5\}$, and $\{2,3,4,5\}$. Therefore, the set $\{12|34,12|35\}$ is closed, and, by symmetry, the other subsets of W are closed, too. □

We will shortly provide a partial converse to Lemma 5.2. First we present a lemma that is required for its proof.

Lemma 5.6. *Suppose that W is $\mathcal{J}$-compatible and $y \in Y$ satisfies*

$$W \cap \{(y,i) : i = 1, \ldots, r\} = \emptyset.$$

Then, there exists $i \in \{1,\ldots,r\}$ such that $W \cup \{(y,i)\}$ is $\mathcal{J}$-compatible.

Proof. Since W is $\mathcal{J}$-compatible there exists a set $J \in \mathcal{J}$ with $W \subseteq J$. Since $J \in \mathcal{J}$ there exists for y one value $i_y \in \{1,\ldots,r\}$ for which $(y,i_y) \in J$. Then $W \cup \{(y,i_y)\}$ is a subset of J, and hence $W \cup \{(y,i_y)\}$ is $\mathcal{J}$-compatible. □

For $y \in Y$ and $W \subset Y \times \{1,\ldots,r\}$, let

$$S_W(y) := \{i \in \{1,\ldots,r\} : W \cup \{(y,i)\} \text{ is } \mathcal{J}\text{-compatible}\}.$$

Proposition 5.7. *If $\#S_W(y) \in \{0,1,r\}$ for every subset W of $Y \times \{1,\ldots,r\}$ that contains no conflict, and every $y \in Y$, then*

$$W \text{ is } \mathcal{J}\text{-compatible} \iff \mathrm{Cl}_{\mathcal{J}}(W) \text{ has no conflict} \tag{1}$$

Proof. The $\Rightarrow$ direction follows from Lemma 5.2. Conversely, suppose that W is not $\mathcal{J}$-compatible. Let W_1 be a maximal $\mathcal{J}$-compatible subset of W (note that this might be $W_1 = \emptyset$). There exists some element $(y,i) \in W - W_1$. Then, $(y,i) \in W \subseteq \mathrm{Cl}_{\mathcal{J}}(W)$. Since $(y,i) \notin \in W_1$, it follows from Lemma 5.6 that there is $j \in \{1,\ldots,r\}$ with $W_1 \cup \{(y,j)\}$ $\mathcal{J}$-compatible, implying $\#S_{W_1}(y) > 0$. Yet by the maximality assumption on W_1 we have $W_1 \cup \{(y,i)\}$ is not $\mathcal{J}$-compatible, implying $\#S_{W_1}(y) < r$. Consequently, we have $S_{W_1}(y) = \{j\}$ and, therefore, $(y,j) \in \mathrm{cl}_{\mathcal{J}}(W_1) \subseteq \mathrm{Cl}_{\mathcal{J}}(W)$. Thus we see that $\mathrm{Cl}_{\mathcal{J}}(W)$ contains both (y,i) and (y,j) and so $\mathrm{Cl}_{\mathcal{J}}(W)$ contains a conflict. □

We illustrate an application of Proposition 5.7 by deriving Proposition 3.1.

Let Y denote the set of subsets of $\{1,\ldots,n\}$ of size 3 and, for $y = \{i,j,k\} \in Y$ with $i < j < k$, let $(y,1)$, $(y,2)$, $(y,3)$ represent the rooted triple with leaf set y that groups together i and j, i and k, and j and k, respectively. Let $\mathcal{J}$ contain the sets of all induced rooted triples of some parent tree. W represents an arbitrary subset of rooted triples, and W is $\mathcal{J}$-compatible precisely if W is compatible in the phylogenetic sense. Given a compatible and closed rooted triple set W and an element $y \in Y$ such that $(y,i) \notin W$ for $i \in \{1,2,3\}$, it can easily be checked that $W \cup (y,i)$ is $\mathcal{J}$-compatible for every $i \in \{1,2,3\}$ (for details see [4], Proposition 9(1)). Hence, Proposition 5.7 implies that every closed non-compatible rooted triple set contains a *conflict*. This result was also proved in [3] and [4].

Note that Proposition 5.7 has the following consequence for settings such as Example 5.1.

Corollary 5.8. *If* $r = 2$, *then* W is $\mathcal{J}$-compatible *if and only if* $\mathrm{Cl}_{\mathcal{J}}(W)$ *contains no conflict.*

Acknowledgments

We thank the two referees for their helpful comments.

References

[1] A.V. Aho, Y. Sagiv, T.G. Szymanski, J.D. Ullman, Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions, SIAM J. Comput. 10 (1981) 405.
[2] O.R.P. Bininda-Emonds, Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life, Kluwer Academic Publishers, Dordrecht, 2004.
[3] D. Bryant, Building trees, hunting for trees, and comparing trees: theory and methods in phylogenetic analysis. Unpublished Ph.D. thesis. University of Canterbury, 1997.
[4] D. Bryant, M. Steel, Extension operations on sets of leaf-labelled trees, Adv. Appl. Math. 16 (1995) 425.
[5] D. Bryant, S. Böcker, A.W.M. Dress, M. Steel, Algorithmic aspects of tree amalgamation, J. Algor. 37 (2000) 522.
[6] P. Daniel, Supertree methods: some new approaches, Masters thesis, University of Canterbury, New Zealand, 2003.
[7] M.C.H. Dekker, Reconstruction methods for derivation trees, Masters thesis, Vrije Universiteit, Amsterdam, Netherlands, 1986.
[8] T. Dezulian, M. Steel, Phylogenetic closure operations and homoplasy-free evolution, In Classification, clustering, and data mining applications, in: D. Banks, L. House, F.R. McMorris, P. Arabie, W. Gaul (Eds.), Proceedings of the Meeting of the International Federation of Classification Societies (IFCS) 2004, Springer, Berlin, 2004, p. 395.

[9] P.L. Erdös, M.A. Steel, L.A. Székely, T. Warnow, A few logs suffice to build (almost) all trees (Part 1), Random Struct. Algor. 14 (1999) 153.
[10] G.F. Estabrook, F.R. McMorris, When are two qualitative taxonomic characters compatible? J. Math. Biol. 4 (1977) 195.
[11] P.A. Golobof, D. Pol, Semi-strict supertrees, Cladistics 18 (2002) 514.
[12] K. Huber, V. Moulton, C. Semple, M. Steel, Recovering a phylogenetic tree using pairwise closure operations, Appl. Math. Lett. 18 (2005) 361.
[13] D.H. Huson, T. Dezulian, T. Kloepper, M.A. Steel, Phylogenetic super-networks from partial trees, IEEE/ACM Trans. Comput. Biol. Bioinform. 1 (2004) 151.
[14] C.A. Meacham, Theoretical and computational considerations of the compatibility of qualitative taxonomic characters, in: J. Felsenstein (Ed.), Numerical taxonomy, NATO ASI Series, vol. 1, Springer, Berlin, 1983, p. 304.
[15] E. Mossel, M. Steel, A phase transition for a random cluster model on phylogenetic trees, Math. Biosci. 187 (2004) 189.
[16] C. Semple, Reconstructing minimal rooted trees, Discrete Appl. Math. 127 (2003) 489.
[17] C. Semple, M. Steel, Tree reconstruction via a closure operation on partial splits, in: O. Gascuel, M.-F. Sagot (Eds.), Proceedings of Journées Ouvertes: Biologie, Informatique et Mathématique, Lecture Notes in Computer Science, Springer, Berlin, 2001, p. 126.
[18] C. Semple, M. Steel, Phylogenetics, Oxford University, Oxford, 2003.
[19] M. Steel, The complexity of reconstructing trees from qualitative characters and subtrees, J. Classif. 9 (1992) 91.
[20] M. Wilkinson, J.A. Cotton, J.L. Thorley, The information content of trees and their matrix representation, Syst. Biol. 53 (2004) 989.