Bull Math Biol
https://doi.org/10.1007/s11538-018-0419-1

SPECIAL ISSUE: ALGEBRAIC METHODS IN PHYLOGENETICS

Phylogenetic Flexibility via Hall-Type Inequalities and Submodularity

Katharina T. Huber[1] · **Vincent Moulton**[1] · **Mike Steel**[2]

Received: 23 October 2017 / Accepted: 19 March 2018

Abstract Given a collection τ of subsets of a finite set X, we say that τ is *phylogenetically flexible* if, for any collection R of rooted phylogenetic trees whose leaf sets comprise the collection τ, R is compatible (i.e. there is a rooted phylogenetic X-tree that displays each tree in R). We show that τ is phylogenetically flexible if and only if it satisfies a Hall-type inequality condition of being 'slim'. Using submodularity arguments, we show that there is a polynomial-time algorithm for determining whether or not τ is slim. This 'slim' condition reduces to a simpler inequality in the case where all of the sets in τ have size 3, a property we call 'thin'. Thin sets were recently shown to be equivalent to the existence of an (unrooted) tree for which the median function provides an injective mapping to its vertex set; we show here that the unrooted tree in this representation can always be chosen to be a caterpillar tree. We also characterise when a collection τ of subsets of size 2 is thin (in terms of the flexibility of total orders rather than phylogenies) and show that this holds if and only if an associated bipartite graph is a forest. The significance of our results for phylogenetics is in providing precise and efficiently verifiable conditions under which supertree methods that require consistent inputs of trees can be applied to any input trees on given subsets of species.

✉ Katharina T. Huber
K.Huber@uea.ac.uk

Vincent Moulton
V.Moulton@uea.ac.uk

Mike Steel
mike.steel@canterbury.ac.nz

[1] School of Computing Sciences, University of East Anglia, Norwich, UK

[2] Biomathematics Research Centre, University of Canterbury, Christchurch, New Zealand

Published online: 27 March 2018

Keywords Phylogenetic tree · Set systems · Partial taxon coverage · Bipartite graph · Hall's marriage theorem · Submodularity

1 Introduction

In phylogenomics, biologists often encounter the following problem: Given a collection τ of different subsets of species, the corresponding phylogenetic trees—each one reconstructed from the genomic data available for the corresponding subset—cannot be consistently combined into a single phylogenetic tree for all the species. When this occurs, various heuristic and somewhat ad hoc 'supertree' methods (such as 'matrix recoding with parsimony') are often applied to provide some estimate of a parent tree (Felsenstein 2004). However, when the collection of subsets of species has sufficiently sparse overlap (in a sense we will make precise shortly), then *any* phylogenetic tree assignment for τ will lead to a set of trees that can be consistently combined into a parent tree. Figure 1i provides an example of this.

In this paper, we investigate the conditions under which the existence of a consistent parent tree can be guaranteed regardless of the tree structure for each subset. Here 'parent' tree means that the leaf set of the tree is the union of the leaf sets of the input trees. For example, given a set of input trees, if there is a parent tree that displays each tree, then a simple, fast and well-known algorithm due to Aho et al. (1981) constructs such a tree in a canonical way. However, this method will fail to return any phylogenetic tree when presented with input trees that are incompatible (i.e. cannot be displayed by any parent tree). In this paper, we characterise when such a method will always be safe to use on any set of input trees, given the sets of taxa that form the leaf sets of those trees. Thus, we consider as input just subsets of species and develop mathematical characterisations and algorithms for this combinatorial question in the special case where each subset has a fixed (small) size. Later in the paper, we consider how the results extend to more general set systems. Our approach throughout is to

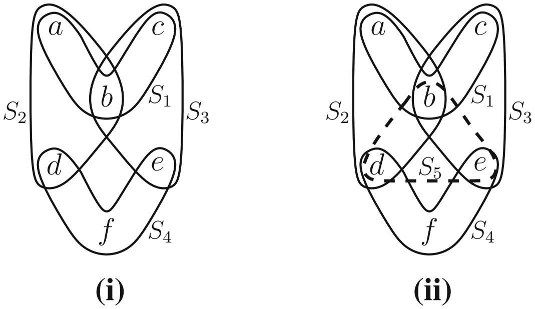

Fig. 1 A collection $\tau = \{\{a, b, c\}, \{a, b, d\}, \{b, c, e\}, \{d, e, f\}\}$ of four sets that is phylogenetically flexible and (**ii**) a collection $\tau' = \tau \cup \{\{b, d, e\}\}$ that fails to be phylogenetically flexible. In (**i**) all of the $3^4 = 81$ choices of rooted triples (one for each of the four leaf sets) give a set of rooted triples that is displayed by at least one rooted phylogenetic tree on the six leaves $a, b, \ldots, f$. However, in (**ii**) this fails, for example, the set of rooted triples $ab|c$, $bd|a$, $bc|e$, $df|e$ together with $be|d$ (for the fifth set) is not displayed by any tree on the six leaves. The set τ is thin, but τ' is not, since it has a subset (namely τ' itself) which has strictly negative excess (equal to -1)

reduce certain combinatorial questions in phylogenetics to the study of systems of inequalities involving linear expressions and related submodularity properties.

In discussion section, we mention a further biological context where the results may be relevant. Note that there are many reasons why phylogenetic trees are constructed on different subsets of species, and a particularly topical one is that genes used to estimate a given phylogeny may only be present (or have been sequenced) in a given subset of the species, and these subsets vary from gene to gene (Sanderson et al. 2011).

Our work is motivated in part by a remarkable combinatorial result by Grünewald (2012) involving unrooted binary trees. In that paper, a set $\mathcal{P}$ of binary trees having leaves labelled from some set X is said to be 'slim' if for every non-empty subset $\mathcal{P}'$ of $\mathcal{P}$, the number of leaves appearing in at least one tree in $\mathcal{P}'$ is at least the total number of interior edges of T plus 3. Theorem 1.1 of Grünewald (2012) then states that for any such thin collection $\mathcal{P}$ there is a tree with leaf set X that 'displays' each of the trees in $\mathcal{P}$. In particular, this leads to the rather striking consequence that 'the property of being slim only depends on the involved leaf sets of the trees and not on which phylogenetic tree is chosen for a fixed leaf set' (Grünewald 2012, p. 324). In this paper, we explore this notion further, and by working with rooted trees (rather than unrooted ones) we are able to establish precise characterisations of the analogous 'slim' property.

Our work is also partly motivated by results from Dress and Steel (2009) where slim-type properties also arise in a tree-based setting, but for a quite different question involving 'median' vertices. To explain this, given a tree $T = (V, E)$ and a subset S of V of size 3, say $S = \{x, y, z\}$, consider the path in T connecting x, y, the path connecting x, z and the path connecting y, z. There is a unique vertex that is shared by these three paths, the *median vertex* of S in T, denoted $\mathrm{med}_T(S)$. In Dress and Steel (2009), the authors show that 'slim'-type properties characterise when a set of triples from X can be realised as providing an encoding of the interior vertices of a (unrooted) tree with leaf set X. (An extension of this to sets of subset of X of size greater than 3 is also described.) In this paper, we extend this result further by showing that the tree that provides this encoding can be chosen to have a particular special type of structure (a 'caterpillar').

The phylogenetic combinatorics of subsets of a species set is a topic that has also been explored recently in the setting of 'phylogenetic decisiveness' (Steel and Sanderson 2010). However, the questions that we consider here are quite different from that setting; rather than requiring a dense overlap of the species subsets in the phylogenetic decisiveness setting, here we investigate sparse overlap.

We begin with some definitions. Throughout this paper, X will denote a fixed finite set.

1.1 Thin Set Systems

Suppose τ is a non-empty subset of $\binom{X}{r}$, $r \geq 2$. Let $L(\tau) = \bigcup_{s \in \tau} s$ (i.e. the set of elements of X that appear in at least one set in τ) and define the *excess* of τ, denoted $\mathrm{exc}(\tau)$, by:

$$\mathrm{exc}(\tau) = |L(\tau)| - |\tau| - (r-1).$$

We say that τ is *thin* if, for all non-empty subsets τ' of τ, we have:

$$\mathrm{exc}(\tau') \geq 0.$$

This notion appears in related but slightly different settings, namely for the leaf sets of unrooted trees in Grünewald (2012), in the median representation of sets of triples in Dress and Steel (2009), and as sparse triplet covers in Grünewald et al. (2017).

In the following lemma, recall that a collection of (not necessarily distinct) sets $\{B_1, B_2, \ldots, B_m\}$ has a *system of distinct representatives* if one can select an element $x_i \in B_i$ for each $i \in \{1, \ldots, m\}$ so that the elements $x_1, x_2, \ldots, x_m$ are all distinct. For τ a non-empty subset of $\binom{X}{r}$, $r \geq 2$ with $L(\tau) = X$ and for $x \in X$ let $n_\tau(x)$ be the number of elements in τ that contain x.

Lemma 1 *Let τ be a non-empty subset of $\binom{X}{r}$, $r \geq 2$ and $L(\tau) = X$. If τ is thin, then the following properties hold:*

(i) *$|\tau| \leq n - r + 1$ where $n = |X|$.*
(ii) *For some $x \in X$, $n_\tau(x) \leq r - 1$.*
(iii) *For any subset B of X of size $r - 1$, the collection of sets $\{S - B : S \in \tau\}$ has a system of distinct representatives.*

Proof Part (i) follows from the defining condition for thin upon taking $\tau' = \tau$.

Part (ii) can be established by the following double-counting argument. Suppose that there is no element $x \in X$ with $n_\tau(x) \leq r - 2$, so that $n_\tau(x) \geq r - 1$ for all $x \in X$. Let $\Omega = \{(x, S) : x \in S \in \tau\}$. We then have:

$$|\Omega| = \sum_{x \in X} n_\tau(x) \geq (r-1)k + r(n-k) \tag{1}$$

where $k = |\{x \in X : n_\tau(x) = r - 1\}|$. On the other hand:

$$|\Omega| = r|\tau| \leq r(n - (r-1)), \tag{2}$$

where the inequality is from Part (i). Combining (1) and (2) gives $k \geq r(r-1)$ and, so, $k \geq 2$. By the definition of k, (ii) follows.

For Part (iii), consider the union of any l sets $A_1, A_2, \ldots, A_l$ where $A_i = S_i - B$ and $S_i \in \tau$ for $i = 1, \ldots, l$. (Note that these sets may have different sizes and a set may occur more than once.) Since τ is thin, $|\bigcup_{i=1}^{l} S_i| \geq l + (r-1)$, and so, since B has size $r - 1$, $|\bigcup_{i=1}^{l} A_i| = |\bigcup_{i=1}^{l} S_i| - (r-1) \geq l$. Since the inequality $|\bigcup_{i=1}^{l} A_i| \geq l$ holds for all $1 \leq l \leq |\tau|$, Hall's marriage theorem (Hall 1935) ensures that τ has a system of distinct representatives. □

For the first part of this paper, we will deal with the case where $r = 3$. However, the main theorem in this setting (Theorem 1) will be used in Sect. 4 to derive a result for the more general case where the sets have different sizes. When $r = 3$, notice that if $|\tau'| = 1$, then $\text{exc}(\tau') = 3 - 1 - 2 = 0$; however, if $|\tau'| = 2$, then $\text{exc}(\tau') \geq 4 - 2 - 2 = 0$, so it suffices, in the definition of thin, to consider subsets of τ' of τ of size at least 3.

A simple way to generate a thin set is to take any ordered sequence of subsets of X of size 3, for which the ordered sequence has the property that each member contains at least one element of X that is not present in any earlier member of the sequence. However, not all thin sets can be obtained in this way. For example, consider the collection $\{\{a, b, c\}, \{c, d, e\}, \{b, e, f\}, \{a, d, f\}\}$ of four subsets sets of $X = \{a, b, \ldots, f\}$. This collection of subsets is thin, yet these four sets cannot be ordered so as to satisfy the property described.

1.2 Phylogenetic Trees and Flexible Sets

Following Semple and Steel (2003), a *rooted phylogenetic tree* T is a rooted tree having a set $L(T)$ of labelled leaves (vertices of out-degree 0) and for which every non-leaf vertex is unlabelled and has out-degree at least 2. We let ρ_T, or more briefly ρ denote the root vertex of T, which has in-degree 0. In case each non-leaf vertex has out-degree exactly 2, we say that T is *binary*. If $L(T) = X$, we will also say that T is a *rooted phylogenetic X-tree*. We let $\mathring{V}(T)$ denote the set of interior (i.e. non-leaf) vertices of T. Similarly, an *unrooted phylogenetic tree* T is an unrooted tree having a set $L(T)$ of labelled leaves (vertices of degree 1) and for which every non-leaf vertex is unlabelled and has degree at least 3. In case each non-leaf vertex has degree exactly 3, we say that T is *binary*. If $L(T) = X$, we will also say that T is a *unrooted phylogenetic X-tree*.

A *rooted triple* is a rooted binary phylogenetic tree on three leaves, and we denote such a tree as $ab|c$ if it has leaf set $\{a, b, c\}$ with leaf c adjacent to the root. A rooted phylogenetic X-tree T is said to *display* the rooted triple $ab|c$ if some subdivision of the tree $ab|c$ is a subgraph of T.

A *cherry* in a (rooted or unrooted) phylogenetic tree is a pair of leaves that is adjacent to the same vertex. A *rooted (respectively, unrooted) caterpillar* tree on X is a rooted (resp. unrooted) binary phylogenetic X-tree for which the number of cherries is at most 1 (respectively, 2).

These notions are illustrated in Fig. 2.

A set R of rooted triples chosen from X is said to be *compatible* if there is a rooted phylogenetic X-tree T that displays each rooted triple in R (in which case, we say

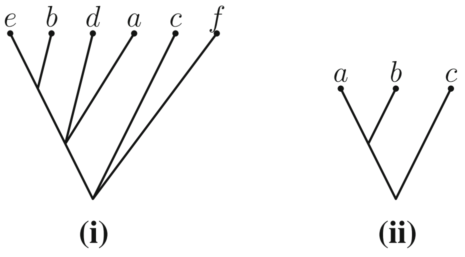

Fig. 2 (**i**) A rooted phylogenetic tree on leaf set $\{a, b, c, \ldots, f\}$. This tree is not binary, as it has a vertex of out-degree 3 (adjacent to a and d). (**ii**) The rooted triple $ab|c$ for which a, b forms a cherry. This rooted triple is also a rooted caterpillar and it is displayed by the tree in (**i**)

that T *displays* R). Note that if R is compatible, then T can always be chosen to be a binary tree and R can contain at most one tree for any triplet (i.e. at most one of $ab|c$, $ac|b$, and $bc|a$ can be present in R).

Suppose that we have a set R of rooted triples with leaves chosen from X. We will let $||R||$ denote the subset of $\binom{X}{3}$ consisting of the leaf sets of the trees in R. We say that a non-empty subset τ of $\binom{X}{3}$ is *phylogenetically flexible* if every set R of rooted triples for which $||R|| = \tau$ holds is compatible. An example to illustrate this notion is provided in Fig. 1.

The following observation that phylogenetic flexibility is hereditary is straightforward to check.

Lemma 2 *Suppose τ is a non-empty subset of $\binom{X}{3}$ that is phylogenetically flexible. If τ' is a non-empty subset of τ, then τ' is phylogenetically flexible.*

2 Characterisation Result

We can now state our first main result.

Theorem 1 *Suppose that τ is a non-empty subset of $\binom{X}{3}$. Then τ is phylogenetically flexible if and only if τ is thin.*

The 'if' direction of Theorem 1 can be established by applying Theorem 1.1 of Grünewald (2012); however, we give a shorter and more direct proof of this direction here (as well as establishing the converse). We begin with some preliminary results, which are required for the argument.

Given a rooted phylogenetic tree $T = (V, E)$ with leaf set X and every vertex in $\mathring{V}(T) - \{\rho_T\}$ having degree three. We say that a rooted triple $xy|z$ *supports* a vertex v in T if $xy|z$ is displayed by T and $v = \mathrm{lca}_T(x, y)$.

For a set R of rooted triples on X, put $L(R) = \bigcup_{t \in R} L(t)$. Furthermore, for a non-empty subset S of X, let $[R, S]$ be the graph with vertex set S and with an edge $\{a, b\}$ if and only if there exists a rooted triple $ab|c \in R$ for at least one element $c \in S$. By Bryant and Steel (1995, Theorem 2), R is compatible if and only if the graph $[R, S]$ is disconnected for all subsets S of X of size at least 2.

Lemma 3 *Suppose that T is a rooted binary phylogenetic X-tree, that R is a set of rooted triples with $L(R) = X$, and that each rooted triple supports a unique (interior non-root) vertex in T. Then, the graph $[R, X]$ has precisely two connected components.*

Proof For $v \in \mathring{V}(T) - \{\rho_T\}$, let X_v be the leaf set of the rooted subtree of T with root v. We claim that for every such v, the graph induced by $[R, X]$ on X_v is connected. The lemma then follows immediately by considering the graphs induced by $[R, X]$ on X_u, X_w for u and w the children of the root of T.

To prove the claim, for u (a child of the root ρ_T of T), we consider the following set:

$$X^u = \{X_v \,:\, v \text{ is an internal vertex of } T \text{below or equal to } u\},$$

where v is said to be *below* u if u lies on the path from ρ_T to v. Note that since $|X| \geq 3$, there must exist a child u of ρ_T such that $X^u \neq \emptyset$ and also there exists some vertex $v \in V(T)$ below or equal to u such that $|X_v| \geq 2$. We use induction on $|X_v|$ for X_v in X^u. If $|X_v| = 2$, then both children of v are leaves and the lemma holds because if $X_v = \{p, q\}$, then, by assumption, there exists a rooted triple in R of the form $r|pq$ for some $r \in X - \{p, q\}$ that supports v. Hence, there is an edge $\{p, q\}$ in $[R, X]$, and therefore, the graph induced by $[R, X]$ on X_v is connected.

Now suppose that v is an internal vertex of T below or equal to u such that $|X_v| \geq 3$. Then at least one of the two children v_1 and v_2 of v is not a leaf of T. Without loss of generality, we may assume that v_1 is that child. Therefore, $2 \leq |X_{v_1}| < |X_v|$ and so, by induction, the graph induced by $[R, X]$ on X_{v_1} is connected. If v_2 is not a leaf of T, then the same arguments as before imply that the graph induced by $[R, X]$ on X_{v_2} is also connected. If v_2 is a leaf of T, then the graph $[R, X]$ on v_2 is a vertex and therefore is (trivially) connected. Since, by assumption, there exists a rooted triple in R that supports v, there is an edge $\{y, z\}$ in $[R, X]$ with $y \in X_{v_1}$ and $z \in X_{v_2}$. Hence, the graph induced by $[R, X]$ on X_v is connected. □

Proof of Theorem 1 We first establish the ‘if’ direction. Suppose that τ is thin, and let R be a set of rooted triples with leaves chosen from X with $||R|| = \tau$. We show that any such choice of R is compatible.

We will establish the compatibility of R via the aforementioned characterisation that R is compatible if and only if $[R, S]$ is disconnected for all subsets S of X of size at least 2. To that end, let S be a subset of X of size at least two.

Notice that $[R, S] = [R_S, S]$ where R_S is the subset of those rooted triples in R that have all three of their leaves in S. Let $\tau' = ||R_S||$. Since τ is thin, we have $\mathrm{exc}(\tau') \geq 0$, in other words:

$$|L(\tau')| - |\tau'| \geq 2. \tag{3}$$

Now (i) the number of vertices of $[R, S]$ is $|S|$ and $|S| \geq |L(\tau')|$; (ii) the number of edges of $[R, S]$ is at most $|R_S| = |\tau'|$. Thus, by Inequality (3), the number of vertices of $[R, S]$ minus the number of edges of this graph is at least 2. But any finite graph with this property must be disconnected. Since this holds for all subsets S of X of size at least two, it follows that R is compatible.

We turn now to the ‘only if’ direction.

We use induction on $|\tau|$. If $|\tau| = 1$, then τ is clearly thin. So, suppose the ‘only if’ direction holds for all $\tau' \subset \binom{X}{3}$ with $1 \leq |\tau'| < m$, some $m \geq 2$, and let $\tau \subseteq \binom{X}{3}$ such that $|\tau| = m$. Without loss of generality, we may assume that $X = L(\tau)$.

Suppose that τ' is a non-empty proper subset of τ. By Lemma 2, τ' is phylogenetically flexible. Hence by induction, τ' is thin. Thus, $|L(\tau')| \geq |\tau'| + 2$. To show that τ is thin, it therefore suffices to prove that $|L(\tau)| \geq |\tau| + 2$.

Suppose for the purposes of obtaining a contradiction that $|L(\tau)| < |\tau| + 2$. Let $\{x, y, z\} \in \tau$ and set $\tau' = \tau - \{\{x, y, z\}\}$. Then, as τ' is thin by induction,

$$|\tau|+2 > |L(\tau)| = |L(\tau')|+(3-|L(\tau')\cap\{x, y, z\}|) \geq |\tau|+4-|L(\tau')\cap\{x, y, z\}|. \tag{4}$$

Hence $|L(\tau') \cap \{x, y, z\}| > 2$ and, so, $\{x, y, z\} \subseteq L(\tau')$. Thus, $L(\tau') = X$.

Now, since τ' is thin, there exists a (unrooted) phylogenetic tree $T = (V, E)$ with leaf set X, and all vertices in $\mathring{V}(T)$ of degree 3, for which the map $\mathrm{med}_T : \tau' \to \mathring{V}(T)$ is one to one (Dress and Steel 2009) (see also Sect. 3). We claim that the map med_T must in fact be bijective. Suppose that this is not the case. Then, there exists some $v \in \mathring{V}(T)$ such that $\mathrm{med}_T(s) \neq v$, for all $s \in \tau'$. Hence, $|X|-2 = |\mathring{V}(T)| > |\tau'|$ and, so, $|X| - 1 > |\tau|$. But then $|X| + 1 > |\tau| + 2 > |L(\tau)| = |X|$, which is impossible as $|\tau| + 2$ is an integer. Hence, med_T is a bijection as claimed.

Now, root the tree T by inserting a root vertex ρ into an edge which separates x, y from z, when the edge is removed from T. Let R' be a set of rooted triples induced by the map med_T (for each element $\{a, b, c\}$ in τ', med_T maps to some $v \in \mathring{V}(T)$ so that we get a rooted triple with leaf set $\{a, b, c\}$ which supports v in the rooted version of T) with $||R'|| = \tau'$ and $L(R') = X$. Since med_T is a bijection, R' satisfies the conditions of Lemma 3 for the rooted version of T. Hence, the graph $[R', X]$ has two connected components, one that contains x, y in its vertex set and the other that contains z.

Now consider the set of rooted triples $R = R' \cup \{y|zx\}$. Then $L(R) = X$, $[R, X]$ is connected and so R is not compatible, and $||R|| = \tau$. But this is impossible, since τ is phylogenetically flexible. □

The following corollary of Theorem 1 is now immediate from Lemma 1(i).

Corollary 1 *If a non-empty subset τ of $\binom{X}{3}$ is phylogenetically flexible, then $|\tau| \leq n - 2$ where $n = |X|$.*

We end this section by considering how many trees can display a set of rooted triples R when $||R||$ is phylogenetically flexible. It might be suspected that since the overlap between the leaf sets of the trees in R is sparse, the number of trees displaying R would need to be large. Indeed, this is sometimes the case; for example, suppose that the leaf sets in R are all disjoint, so the total number of leaves is given by $n = 3k$, where $k = |R|$. In this case, the number N of rooted binary trees on n leaves that display R is given by:

$$N = \frac{(2n-3)!!}{3^{n/3}}, \tag{5}$$

which grows exponentially with n. The proof of Eq. (5) is to observe that each of the 3^k ways to select a rooted triple from the k triples in $||R||$ provides a set of rooted triples that is displayed by at least one rooted phylogenetic tree [by the algorithm from Aho et al. (1981)] and hence by at least one rooted binary tree, and these rooted binary trees are pairwise distinct, since any two of them display a different rooted triple for at least one triple in $||R||$.

At the other extreme, if R has the maximum possible size for a phylogenetically flexible set on n leaves (namely $n - 2$ by Corollary 1), then it is possible for there to be just a single rooted phylogenetic tree that displays R; this is stated more precisely in the next proposition.

Proposition 1 *(i) For every rooted binary phylogenetic X-tree T on $n \geq 3$ leaves, there exists a set R_T of $n-2$ rooted triples for which (a) T is the only phylogenetic X-tree that displays R_T and (b) $||R_T||$ is thin.*

(ii) *There exist phylogenetically flexible sets of triples of size $n-2$ on n leaves ($n \geq 6$) for which each assignment of a tree structure to these triples leads to a set of rooted triples that can be displayed by more than one rooted phylogenetic tree.*

Proof (i) We use induction on n. For $n = 3$, we can write $T = ab|c$, in which case $R_T = \{ab|c\}$ satisfies Conditions (a) and (b). Suppose now that Proposition 1 holds for $k \leq n$ where $n \geq 3$ and that T is a rooted binary phylogenetic X-tree with $n+1$ leaves. Select a pair of leaves a, b that are adjacent to the same vertex (say v) of T (i.e. $\{a, b\}$ is a cherry of T), let vertex u be the parent of vertex v in T, and let c be any leaf of T present in the component of $T - u$ (the graph obtained by deleting u from T) that contains neither the root, nor the leaves a, b. Put $X' = X - \{a\}$ and let T' be the rooted binary phylogenetic X'-tree obtained from T by deleting leaf a and its incident edge and suppressing the resulting vertex of degree 2. Since T' has n leaves, the induction hypothesis ensures that there is a set $R_{T'}$ of $n-2$ rooted triples for which T' is the only phylogenetic X'-tree that displays $R_{T'}$ and that $||R_{T'}||$ is thin. If we now let $R_T = R_{T'} \cup \{ab|c\}$, then R_T is a set of $(n+1)-2$ rooted triples and R_T satisfies Conditions (a) and (b) for the tree T. This establishes the induction step and thereby the proposition.

(ii) Let $\tau = \{\{1, 2, j\} : 2 < j \leq n\}$. In this case, τ is a thin (and therefore phylogenetically flexible) set of size $n-2$. Now, for $n \geq 6$, it can be checked that any assignment of a tree structure to these triples leads to a set of rooted triples that can be displayed by more than one rooted phylogenetic tree. □

3 Median Characterisations

Given a phylogenetic tree T with leaf set X and a set $s \in \binom{X}{3}$, let $\mathrm{med}_T(s)$ refer to the vertex that is the unique median vertex of T for the three elements of s.

The following result was established in Dress and Steel (2009, Theorem 1.1). Suppose that τ is a subset of $\binom{X}{3}$ with $L(\tau) = X$. The following are equivalent:

(i) τ is thin.

(ii) There exists a binary unrooted phylogenetic X-tree $T = (V, E)$ for which the function $\mathrm{med}_T : \tau \to \mathring{V}(T)$: $s \mapsto \mathrm{med}_T(s)$ from the elements s of τ to the set of interior vertices of T is one to one.

When (ii) holds, we say that T provides a *median representation* of τ. Figure 3i illustrates how this equivalence applies.

We now strengthen this result from Dress and Steel (2009) by showing that the tree T can always be chosen to be an unrooted caterpillar tree. For example, for the thin collection of sets considered in Fig. 3, we may select the caterpillar tree shown in Fig. 3ii.

Theorem 2 *Suppose τ is a non-empty subset of $\binom{X}{3}$, where $|X| \geq 4$. If τ is thin, then there exists an unrooted caterpillar tree $T = (V, E)$ with leaf set X for which the function* $\mathrm{med}_T : \tau \to \mathring{V}(T)$ *is one to one.*

Proof We adapt the proof of (3) ⇒ (2) of Dress and Steel (2009, Theorem 1.1) and use induction on the size of X. If $|X| = 4$, the theorem clearly holds in view of

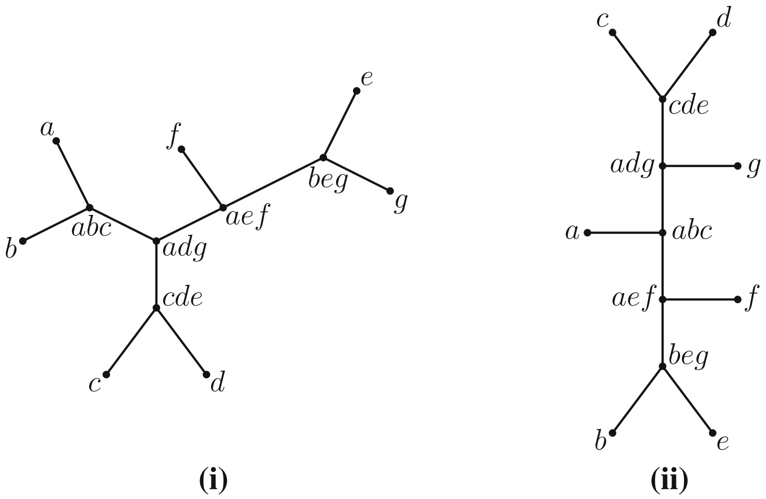

Fig. 3 (**i**) Associating each member of the thin collection of sets $\{\{a, b, c\}, \{c, d, e\}, \{a, e, f\}, \{b, e, g\}, \{a, d, g\}\}$ with its median vertex in the tree shown provides a one-to-one mapping. (**ii**) A caterpillar tree that also provides a median representation of this thin collection of sets

Lemma 1(i). Let us suppose that it holds whenever $4 \leq |X| \leq n - 1$, for some $n \geq 5$. Let X be such that $|X| = n$. By Lemma 1(ii), we may assume that one of the following two cases hold:

(A) There is an element x of X with $n_\tau(x) = 1$.
(B) There is an element x of X with $n_\tau(x) = 2$.

In case (A), there is some triple $\{a, b, x\} \in \tau$ such that for $\tau' = \tau - \{\{a, b, x\}\}$ we have that τ' is thin. Put $X' = L(\tau')$. By induction, there is an unrooted caterpillar tree T' with leaf set X' and the function $\mathrm{med}_{T'} : \tau' \to \mathring{V}(T')$ is one to one. Now we can create a tree T by inserting an edge $\{x, u\}$ where u is a new vertex subdividing an interior edge of T' on the path between a and b. The resulting tree T is clearly a unrooted caterpillar tree on X and $\mathrm{med}_T : \tau \to \mathring{V}(T)$ is one to one. This establishes the induction step in this case.

In Case (B), there is an element x in X with $n_\tau(x) = 2$. Then there exist two distinct triples $t, t' \in \tau$ each of which contains x. We consider the following two possible cases: (i) $|t \cap t'| = 2$ and (ii) $|t \cap t'| = 1$.

Case (i): $|t \cap t'| = 2$. In this case, there exist $a, b, b' \in X$ with $b \neq b'$ such that $t = \{a, b, x\}$ and $t' = \{a, b', x\}$. Since τ is thin, it follows that

$$\tau' = \tau - \{\{a, b, x\}, \{a, b', x\}\} \cup \{\{a, b, b'\}\}$$

is also thin. Put $X' = L(\tau')$. Then, by induction, there is a unrooted caterpillar tree T' with leaf set X' and $\mathrm{med}_{T'} : \tau' \to \mathring{V}(T')$ is one to one.

Consider the leaf b' of T'. Let $b'' \in \mathring{V}(T')$ denote the vertex adjacent to b'. As T' is an unrooted caterpillar tree, it suffices to consider the following two subcases:

Subcase (a): The leaves a and b are on the same side of T' relative to b' (i.e. they are in the same connected component of $T' - b''$ as b'). Without loss of generality, assume that the distance from a to b' in T' is less than or equal to distance from b to b in T'. Note that in this case $\mathrm{med}_{T'}(a, b, b')$ is the vertex in T' that is adjacent to a. Now create a tree T with leaf set X by inserting a new vertex u and a new edge $\{u, x\}$ into T' such that $\{u, b''\}$ is an edge on the path connecting b' and a. The tree T is again an unrooted caterpillar tree on X. Furthermore, $\mathrm{med}_T : \tau \to \mathring{V}(T)$ is one to one since (i) $\mathrm{med}_{T'}$ is one to one, and (ii) $\mathrm{med}_T(x, a, b') = u$ and the median of $\{x, a, b\}$ in T corresponds to the median vertex of $\{a, b, b'\}$ in T' and therefore is a vertex of T that is different from any other median vertex of an element in τ.

Subcase (b): The leaves a and b are on different sides of T' relative to b'. Note that in this case, $\mathrm{med}_{T'}(a, b, b') = b''$. Now create a tree T with leaf set X by inserting a new vertex u and a new edge $\{x, u\}$ into T' such that $\{u, b''\}$ is an edge on the path connecting b' and b. T is then clearly an unrooted caterpillar tree on X. Since $\mathrm{med}_T(x, b', b) = u$ and the median of $\{x, a, b'\}$ in T corresponds to the median vertex of $\{a, b, b'\}$ in T', the same arguments as in the previous case imply that $\mathrm{med}_T : \tau \to \mathring{V}(T)$ is one to one.

Case (ii): $|t \cap t'| = 1$. In this case, there exist pairwise distinct elements a, a', b, b' in X such that $t = \{a, b, x\}$ and $t' = \{a', b', x\}$. We may assume that τ does not contain both $\{a, a', b\}$ and $\{a, a', b'\}$ as, otherwise, the claim follows from Case (B)(i)(a). By symmetry, we can assume without loss of generality that $\{a, b', b\}$ is not in τ. Let $\tau' = \tau - \{\{a, b, x\}, \{a', b', x\}\} \cup \{\{a, a', b\}\}$. Then since τ is thin it follows that τ' is thin. Put $X' = L(\tau')$. Then, by induction, there is an unrooted caterpillar tree T' with leaf set X' and $\mathrm{med}_{T'} : \tau' \to \mathring{V}(T')$ is one to one.

Consider the leaf a'. As T' is a caterpillar tree on X', we can again consider two subcases ((a) and (b)), the first of which involves two further subcases:

Case (a): The leaves a and b are on the same side of T' relative to b'. Without loss of generality, assume that the distance from a to b' in T' is less than or equal to distance from b to b' in T'. We now have two subcases to consider for this subcase:

Case (a1): The leaf a' is on the same side of the caterpillar T' as a and b relative to b'. If a and b are on the same side of T' relative a', then the same arguments as in the Case (B)(i)(a) apply with a' playing the role of b'. If a and b are on different sides of T' relative a', then the same arguments apply as in Case (B)(i)(b) with a' playing the role of b'.

Case (a2): The leaf a' is on a different side of the caterpillar T' from a and b relative to b'. Now create a tree T on X by inserting a new vertex u and a new edge $\{x, u\}$ into T' such that with $b'' \in \mathring{V}(T)$ the vertex adjacent with b' we have that $\{b'', u\}$ is an edge on the path connecting a' and b'. Then, T is clearly a unrooted caterpillar tree with leaf set X. Since $\mathrm{med}_T(x, a', b')$ is u and the median of $\{x, a, b\}$ in T corresponds to the median of $\{a, a', b\}$ in T', the same arguments as in Case (B)(i)(a) imply that $\mathrm{med}_T : \tau \to \mathring{V}(T)$ is one to one.

Case (b): The leaves a and b are on different sides of T' relative to b'. If a' lies on the same side of T' as a relative b' and a' and b' lie on different sides of T' relative a, then the same arguments as in the Case (B)(i)(a) apply with a' playing the role of b'. In all other cases, the same arguments as in the Case (B)(i)(b) apply with a' playing the role of b' □

3.1 The Case $r = 2$

The concept of phylogenetic flexibility does not directly carry over to the case where $r = 2$, since in this case, there is just a single rooted phylogenetic tree. Instead, we use a stronger notion of tree structure (namely, total order) to obtain an analogue of Theorem 1.

We say that a non-empty subset τ of $\binom{X}{2}$ is *total-order flexible* if every choice of a total order on the set s, for each $s \in \tau$, is compatible with a total order on X. More formally, for every $s = \{x, y\} \in \tau$, if we declare that either $x \prec y$ or $y \prec x$, then for any such selection of choices (one for each $s \in \tau$), there is a total order on X that agrees with these inequalities. For example, $\tau = \{\{a, b\}, \{b, c\}\}$ is total-order flexible but $\{\{a, b\}, \{b, c\}, \{a, c\}\}$ is not, since the orderings $a \prec b, b \prec c, c \prec a$ are not compatible with any total order on a, b, c. The following result is the analogue of Theorem 1 for the case where $r = 2$.

Theorem 3 *Suppose that τ is a non-empty subset of $\binom{X}{2}$. Then τ is thin if and only if τ is total-order flexible.*

Proof We first show that if τ is not thin, then τ is not total-order flexible. Suppose that τ is not thin. Then there exists a non-empty subset τ' of τ for which $|L(\tau')| \leq |\tau'|$. Let $G_{\tau'}$ be the graph $(L(\tau'), \tau')$ that has vertex set $L(\tau')$ and edge set τ'. Since $G_{\tau'}$ has at least as many edges as vertices, this graph has a connected component that contains a cycle. If the edges of this cycle are $\{x_1, x_2\}, \ldots, \{x_i, x_{i+1}\}, \ldots, \{x_r, x_1\}$, then the total orders $x_1 \prec x_2, \ldots, x_i \prec x_{i+1}, \ldots, x_r \prec x_1$ on these pairs are not compatible with any total order on X (since transitivity would imply that $x_1 \prec x_1$).

We now show that the thin property implies total-order flexibility by using induction on $k = |\tau|$. The result clearly holds for $k = 1$ so suppose that the result holds for subsets of $\binom{X}{2}$ of size $k \geq 1$ and that $\tau \subseteq \binom{X}{2}$ is a thin set of size $k+1$. By Lemma 1(ii), there is an element x in X that is present in precisely one set, say $\{x, y\}$, in τ. Let τ' be the set obtained from τ by deleting $\{x, y\}$. Then τ' is thin and, since $|\tau'| = k$, the induction hypothesis implies that τ' is total-order flexible. Then any choice of a total order on the set s for each $s \in \tau'$ is compatible with a total order $\prec$ on $X - \{x\}$ (recall that $x \notin L(\tau')$ by the choice of x). If we now introduce a total order on $\{x, y\}$, then we can extend the total order $\prec$ to X by placing x after y if $\{x, y\}$ is ordered as x, y, and placing x after y otherwise. □

We now present some characterisations for when a non-empty set $\tau \subseteq \binom{X}{2}$ is thin. We begin with an analogue of Dress and Steel (2009, Theorem 1.1), which was stated in the last section.

Given a rooted tree T with leaf set X and a set $s \in \binom{X}{2}$, let $\mathrm{lca}_T(s)$ refer to the vertex that is the unique vertex of T that is the least common ancestor of the elements in the set s.

Theorem 4 *Suppose that τ is a subset of $\binom{X}{2}$ with $L(\tau) = X$. The following are equivalent:*

(i) τ is thin.

(ii) There exists a rooted binary phylogenetic X-tree $T = (V, E)$ for which the function $s \mapsto \mathrm{lca}_T(s)$ from the elements of τ to the set of interior vertices of T is one to one.

(iii) As for (ii) but with T a rooted caterpillar tree.

Proof (iii) $\Rightarrow$ (ii) is trivial.

(i) $\Rightarrow$ (iii) Suppose τ is thin. We use induction on the size of X. If $|X| = 3$, then clearly (iii) holds. Therefore, suppose it holds whenever $3 \leq |X| \leq n - 1$, for some $n \geq 4$. Let $|X| = n$.

By Lemma 1(ii), there is some x with $n_\tau(x) = 1$. Let $X' = X - \{x\}$. It is then straightforward to see that there is some pair $\{a, x\} \in \tau$ with $\tau' = \tau - \{\{a, x\}\}$. Clearly τ' is thin as τ is thin and either $L(\tau') = X - \{x\}$ or $L(\tau') = X - \{x, a\}$.

Assume first that $L(\tau') = X'$, where $X' = X - \{x\}$. By induction, there is a rooted caterpillar tree T' with leaf set X' and root ρ' for which the function $\mathrm{lca}_{T'} : \tau' \to \mathring{V}(T')$ is one to one. Now, we can create a new rooted tree T with root ρ by adding two new edges $\{\rho, \rho'\}$ and $\{\rho, x\}$ to T where ρ is a new vertex that is not in T'. T is then clearly a rooted caterpillar tree on X; every vertex in $\mathring{V}(T)$ has out-degree 2 and $\mathrm{lca}_T : \tau \to \mathring{V}(T)$ is one to one. This establishes the induction step, and so (iii) holds.

Assume next that $L(\tau') = X'$, where $X' = X - \{x, a\}$. By induction, there exists a rooted caterpillar tree T' on X'. Let ρ' denote the root of T'. Let T be rooted caterpillar tree obtained from T' via the following two-step process. First, add a new root ρ and a new edge $e = \{\rho, \rho'\}$ to T'. In the resulting tree subdivide e by a vertex c and add the edges $\{c, a\}$ and $\{\rho, x\}$. Clearly, $\mathrm{lca}_T : \tau \to \mathring{V}(T)$ is one to one. This establishes again the induction step, and so (iii) holds too in this case.

(ii) $\Rightarrow$ (i) Suppose x is an element which is not in $L(\tau) = X$. Given a non-empty subset ω of τ, let $\omega^* = \{t \cup \{x\} : t \in \omega\}$.

Suppose a rooted phylogenetic tree T on X satisfies the conditions in Part (ii) of the theorem. Add a new leaf x that is not in $L(\tau)$ to T by adding the edge $\{\rho_T, x\}$ and regard the resulting tree as an unrooted phylogenetic tree T' on $X \cup \{x\}$. In T', the map $\mathrm{med}_{T'}$ from τ^* to the internal vertices of T' is then one to one. Hence, by Dress and Steel (2009, Theorem 1.1), τ^* is thin, and thus, for any non-empty subset ω of τ, we have:

$$|L(\omega)| + 1 = |L(\omega^*)| \geq |\omega^*| + 2 = |\omega| + 2.$$

It immediately follows that τ is thin. □

Interestingly, we can give an alternative characterisation of thin subsets τ of $\binom{X}{2}$ in terms of bipartite graphs.

We first recall some results from matching theory. For a graph G and v a vertex in G, we let $\deg_G(v)$ denote the degree of v in G. Given a bipartite graph $G = (A \cup B, E)$ and a non-empty set $Y \subseteq A$, we let $N_G(Y)$ denote the set of vertices in B that are adjacent to some vertex in Y, and we define the *surplus* $\sigma_G(Y)$ of Y to be:

$$\sigma_G(Y) = \sigma(Y) = |N_G(Y)| - |Y|.$$

We also define the *surplus* $\sigma(G)$ of G, to be the minimum surplus over all non-empty sets of A. We say that a bipartite graph $G = (A \cup B, E)$ has *positive surplus (as viewed from A)* if $\sigma(G) > 0$. The following result is from Lovász and Plummer (1986, Theorem 1.3.8).

Theorem 5 *A bipartite graph $G = (A \cup B, E)$ has positive surplus (as viewed from A) if and only if G contains a forest F such that* $\deg_F(u) = 2$ *for all* $u \in A$.

We now apply this result to the setting of thin sets. Let τ be a non-empty collection of non-empty subsets of X. We associate a bipartite graph $G(\tau)$ to τ that has the vertex set $\tau \cup L(\tau)$ and the edge set given by containment (i.e. $\{t, x\}$ is an edge in $G(\tau)$ if and only if $x \in t$ with $x \in L(\tau)$ and $t \in \tau$). Thus, we are representing our set τ by a bipartite graph $G = (A \cup B, E)$ with $A = \tau$, $B = X$ and E given by containment.

Since τ is thin if and only if $G(\tau)$ has positive surplus, by Theorem 5, the following corollary is straightforward.

Corollary 2 *Suppose that τ is a subset of $\binom{X}{2}$ with $L(\tau) = X$. Then τ is thin if and only if $G(\tau)$ is a forest.*

For $r = 3$, it might also be interesting to characterise those graphs $G(\tau)$ for which τ is thin.

4 The General Case (Slim Set Systems)

Suppose we have a non-empty collection τ of subsets of X, each of size at least 3. Consider the modified notion of excess, denoted exc$'$ and defined as follows: Define

$$\mathrm{exc}'(\tau) = |L(\tau)| - 2 - \sum_{s \in \tau}(|s| - 2).$$

Notice that when $\tau \subseteq \binom{X}{3}$ this notion of excess agrees with the earlier one.

Given a non-empty collection τ of subsets of X, each of size at least 3, we say that τ is *slim* if for every non-empty subset τ' of τ, we have $\mathrm{exc}'(\tau') \geq 0$. The next result relates slim to thin; the two notions coincide when $\tau \subseteq \binom{X}{3}$; however, slim is a more restrictive notion than thin when $\tau \subseteq \binom{X}{r}$, for $r > 3$. Note, however, that (unlike the thin property) the slim property does not require the sets in τ to all have the same size.

Lemma 4 *Suppose that $\tau \subseteq \binom{X}{r}$, $r \geq 3$. If τ is slim, then τ is thin. Moreover, for $r = 3$, τ is thin if and only if τ is slim.*

Proof If $|s| = r$ for each $s \in \tau$, then $\sum_{s \in \tau'}(|s| - 2) = |\tau'|(r - 2)$; therefore, τ is thin if and only if $|L(\tau')| \geq |\tau'|(r - 2) + 2$ for every non-empty subset τ' of τ. We now impose the assumption that $r \geq 3$. First, if $r > 3$, then the required inequality: $|\tau'|(r-2)+2 \geq |\tau'|+(r-1)$ is equivalent to the condition that $|\tau'| \geq 1$, which holds by the assumption that τ' is non-empty. Thus if τ is slim, it is also thin. Moreover, when $r = 3$, the inequality $|\tau'|(r - 2) + 2 \geq |\tau'| + (r - 1)$ becomes an equality; in this case, τ is slim if and only if it is thin. □

We can extend the notion of phylogenetic flexibility introduced in Sect. 1 to arbitrary collections of subsets of X as follows. We first need to extend the earlier definitions of 'display' and 'compatibility' from sets of rooted triples to arbitrary collections of rooted trees, as follows.

Given a rooted phylogenetic X-tree T and a binary phylogenetic tree T' with leaf set $Y \subseteq X$, T is said to *display* T' if T contains a subdivision of T' as a (directed) subtree. [This is equivalent to the condition that each rooted triple displayed by T' is also displayed by T (Bryant and Steel 1995).] A set R of rooted binary phylogenetic trees is said to be *compatible* if there is rooted phylogenetic tree T that displays each of the trees in R.

For a set R of rooted binary phylogenetic tree, let $||R||$ denote the collection of their leaf sets. Thus, $||R||$ is a set of sets. Given a non-empty collection τ of subsets of X, each of size at least 3, we say that τ is *phylogenetically flexible* if every set R of rooted binary phylogenetic trees for which $||R|| = \tau$ holds is compatible.

This notion agrees with the earlier notion of phylogenetic flexibility in the case where each set in τ has size exactly 3. Moreover, as before, we can assume without loss of generality that the tree T (in the definition) is binary.

The following result is a strengthening of our earlier Theorem 1; one direction follows from that theorem, the other direction is a consequence of a result from Grünewald (2012) (which dealt with unrooted trees).

Theorem 6 *Suppose that τ is a collection of sets, each of size at least 3. Then τ is phylogenetically flexible if and only if τ is slim.*

Proof We first establish the 'only if' direction. Suppose that τ is phylogenetically flexible. For each set $s \in \tau$, select two elements $x, y \in s$ and let:

$$A(s) = \{\{x, y, z\} : z \in s, z \neq x, y\}\},$$

and for any non-empty subset τ' of τ let

$$\alpha(\tau') = \bigcup_{s \in \tau'} A(s).$$

Thus, $A(s)$ is a set of $|s| - 2$ triples, and $\alpha(\tau)$ is also a set of triples. □

Claim 1 $A(s) \cap A(s') = \emptyset$ *for each* $s, s' \in \tau, s \neq s'$.

To see this, suppose that a triple, say $\{a, b, c\}$, lies in $A(s)$ and $A(s')$ for two distinct elements s and s' of τ. Then, we can select a rooted binary phylogenetic tree T_s with leaf set s that displays the rooted triple $ab|c$ and select a rooted binary phylogenetic tree $T_{s'}$ with leaf set s' that displays the rooted triple $ac|b$. But no rooted binary phylogenetic tree can display both T_s and $T_{s'}$ (since such a tree would also simultaneously display two different rooted triples with leaf set $\{a, b, c\}$). This contradicts the assumption that τ is phylogenetically flexible, so such a shared triple $\{a, b, c\}$ in $A(s) \cap A(s')$ cannot exist. This establishes Claim 1.

Claim 2 $\alpha(\tau)$ *is phylogenetically flexible.*

To see this, suppose that for each triple $t \in \alpha(\tau)$, we have an associated rooted triple T_t with leaf set t. We need to show that there is a rooted binary phylogenetic tree that displays $\{T_t : t \in \alpha(\tau)\}$. Observe that $A(s)$ is thin for each $s \in \tau$, since if A is a non-empty subset of $A(s)$ of size k (say), then $|\bigcup A| = k + 2$, and so $|\bigcup A| = |A| + 2$. Theorem 1 (the 'if' direction) then ensures that for each s in τ, there is a rooted phylogenetic tree T_s with leaf set s that displays $\{T_t : t \in \alpha(\tau) \cap A(s)\}$. Moreover, since τ is phylogenetically flexible, there is a rooted binary phylogenetic tree T that displays T_s for each $s \in \tau$. It follows that the tree T displays $\{T_t : t \in \alpha(\tau)\}$, and so $\alpha(\tau)$ is phylogenetically flexible, as claimed.

Claim 2 implies that $\alpha(\tau)$ is thin, by Theorem 1 (the 'only if' direction). We now show that this implies that τ is slim. Let τ' be a non-empty subset of τ and consider $\alpha(\tau')$. Since $L(\tau') = L(\alpha(\tau'))$, we have:

$$|L(\tau')| = |L(\alpha(\tau'))| \geq |\alpha(\tau')| + 2, \tag{6}$$

where the inequality holds because $\alpha(\tau)$ (and thereby its subset $\alpha(\tau')$) is thin.

Now:

$$|\alpha(\tau')| = \sum_{s \in \tau'} |A(s)| = \sum_{s \in \tau'} (|s| - 2),$$

where the first equality holds by Claim 1. Combining this last equation with Inequality (6) gives:

$$|L(\tau')| - 2 \geq \sum_{s \in \tau'} (|s| - 2),$$

which shows that τ is slim as claimed.

This establishes the 'only if' direction. Notice in doing so that we have used *both* directions of Theorem 1 in different places in this proof.

We turn now to the 'if' direction. Given τ, select a new element, say x, that is not present in any of the sets in τ, and add this to each of the sets in τ to produce a set τ_{+x}. Notice that if τ is slim, then τ_{+x} satisfies the property that for each non-empty set τ' of τ_{+x}, we have:

$$|L(\tau')| - 3 \geq \sum_{s \in \tau'} (|s| - 3).$$

It follows from Theorem 1.1 of Grünewald (2012) that for any assignment of unrooted binary phylogenetic trees with leaf sets that correspond to the sets in τ_{+x}, there is a binary phylogenetic tree T_x that displays each of these unrooted trees. Suppose now that we have an assignment of rooted binary phylogenetic trees having leaf sets that correspond to the sets in τ. By attaching x as a leaf adjacent to the root of each of these trees, we obtain an assignment of unrooted binary phylogenetic trees with leaf sets that correspond to the sets in τ_{+x}. Hence, by the result just stated, there is an unrooted binary phylogenetic tree T_x that displays each of these unrooted trees. If we now let T be the rooted binary phylogenetic tree obtained from T_x by deleting the leaf x and rooting the resulting tree on the vertex adjacent to x, then T displays the original assignment of rooted binary phylogenetic trees. Since this holds for all possible assignments of rooted phylogenetic trees to the sets in τ, it follows that τ is phylogenetically flexible. □

5 Polynomial-Time Algorithms for Thin and Slim

Given finite set S, a function $f : 2^S \rightarrow \mathbb{R}$ is called *submodular* if for all $A, B \subseteq S$:

$$f(A) + f(B) \geq f(A \cup B) + f(A \cap B).$$

Submodular functions play an important role in optimisation and matroid theory (see, e.g. Lovász and Plummer 1986; Bixby and Cunningham 1995; Welsh 1995). In this section, we exploit these connections to show that there are polynomial-time algorithms to decide whether sets are thin or slim.

Suppose that τ is a subset of 2^X. For $\tau' \subseteq \tau$, we define

$$\sigma(\tau') = |L(\tau')| - |\tau'|,$$

and

$$\gamma(\tau') = |L(\tau')| - \sum_{s \in \tau'} (|s| - 2)$$

Note that $\sigma(\emptyset) = \gamma(\emptyset) = 0$ (since the summation term is then empty). Although the following result is straightforward to show by using results concerning submodular functions in the literature, for completeness we give a direct proof.

Theorem 7 *For τ a non-empty subset of 2^X, the functions $\sigma : 2^\tau \rightarrow \mathbb{R}; \tau' \mapsto \sigma(\tau')$ and $\gamma : 2^\tau \rightarrow \mathbb{R}; \tau' \mapsto \gamma(\tau')$ are submodular.*

Proof Suppose that τ is a non-empty subset of 2^X and that $\tau_1, \tau_2 \subseteq \tau$ are non-empty. Clearly, $|L(\tau_1 \cup \tau_2)| = |L(\tau_1) \cup L(\tau_2)|$ and $|L(\tau_1 \cap \tau_2)| \leq |L(\tau_1) \cap L(\tau_2)|$. Hence:

$$\begin{aligned} |L(\tau_1 \cup \tau_2)| + |L(\tau_1 \cap \tau_2)| &\leq |L(\tau_1) \cup L(\tau_2)| + |L(\tau_1) \cap L(\tau_2)| \\ &= (|L(\tau_1)| + |L(\tau_2)| - |L(\tau_1) \cap L(\tau_2)|) \\ &\quad + |L(\tau_1) \cap L(\tau_2)| \\ &= |L(\tau_1)| + |L(\tau_2)|. \end{aligned}$$

The fact that σ is submodular now follows, since $|\tau_1| + |\tau_2| = |\tau_1 \cup \tau_2| + |\tau_1 \cap \tau_2|$ and thus, by the above inequality, we have:

$$\begin{aligned} \sigma(\tau_1) + \sigma(\tau_2) &= |L(\tau_1)| + |L(\tau_2)| - (|\tau_1| + |\tau_2|) \\ &\geq |L(\tau_1 \cup \tau_2)| + |L(\tau_1 \cap \tau_2)| - |\tau_1 \cup \tau_2| - |\tau_1 \cap \tau_2| \end{aligned}$$

$$= \sigma(\tau_1 \cup \tau_2) + \sigma(\tau_1 \cap \tau_2).$$

Similarly, γ is submodular, since

$$\sum_{s\in\tau_1}(|s|-2) + \sum_{s\in\tau_2}(|s|-2) = \sum_{s\in\tau_1\cup\tau_2}(|s|-2) + \sum_{s\in\tau_1\cap\tau_2}(|s|-2)$$

and therefore:

$$\begin{aligned}\gamma(\tau_1)+\gamma(\tau_2) &= |L(\tau_1)|+|L(\tau_2)| - \left(\sum_{s\in\tau_1}(|s|-2) + \sum_{s\in\tau_2}(|s|-2)\right)\\ &\geq |L(\tau_1\cup\tau_2)|+|L(\tau_1\cap\tau_2)| - \sum_{s\in\tau_1\cup\tau_2}(|s|-2) - \sum_{s\in\tau_1\cap\tau_2}(|s|-2)\\ &= \gamma(\tau_1\cup\tau_2)+\gamma(\tau_1\cap\tau_2).\end{aligned}$$

□

For $\tau \subseteq 2^X$, we define:

$$\sigma^*(\tau) = \min\{\sigma(\tau') \,:\, \tau' \subseteq \tau, \tau' \neq \emptyset\}, \text{ and } \gamma^*(\tau) = \min\{\gamma(\tau') \,:\, \tau' \subseteq \tau, \tau' \neq \emptyset\}.$$

Lemma 5 *(i) Suppose that $\tau \subseteq \binom{X}{r}$ where $r \geq 3$. Then τ is thin if and only $\sigma^*(\tau) \geq 2$.*

(ii) Suppose $\tau \subseteq 2^X$ such that each element in τ has size at least three. Then τ is slim if and only $\gamma^*(\tau) \geq 2$.

Proof (i) τ is thin if and only if $\sigma(\tau') \geq 2$ for all non-empty $\tau' \subseteq \tau$ if and only if $\sigma^*(\tau) \geq 2$.

(ii) τ is slim if and only if $\gamma(\tau') \geq 2$ for all non-empty $\tau' \subseteq \tau$ if and only if $\gamma^*(\tau) \geq 2$. □

The following result (Lovász 1983, Theorem 4.4) is originally due to Grötschel et al. (1981) (see also Lovász and Plummer 1986, pp. 417–418).

Theorem 8 *Let f be a submodular function defined on the subsets of some finite set S. A set minimising f over all non-empty subsets of S can then be found in polynomial time.*

In light of this theorem and Theorem 7, it follows that we can determine $\sigma^*(\tau)$ and $\gamma^*(\tau)$ for a given set $\tau \subseteq 2^X$ in polynomial time. Therefore, by Lemma 5, we can determine whether or not a given set τ for which each element has size at least three is thin or slim in polynomial time.

Note that although this shows that polynomial-time algorithms exist for determining whether or not a set is thin or slim, these are likely to be impracticable (Lovász and Plummer 1986, pp. 417–418). However, for the case of determining whether or not a set τ is thin a more explicit algorithm can be given. More specifically, in Fritzilas et al. (2013, Theorem 2) a polynomial-time algorithm is presented for computing the surplus $\sigma(G)$ of a bipartite graph G. Since for a set $\tau \subseteq \binom{X}{r}$, $r \geq 3$, the surplus of the bipartite graph $G(\tau)$ as defined in Sect. 3.1 is equal to $\sigma^*(\tau)$, we can therefore apply this algorithm to determine if τ is thin. It would be interesting to find an explicit algorithm for determining whether a set is slim.

Theorem 7 has another consequence that relates to phylogenetics. Recall that a *patchwork* is a non-empty collection $\mathcal{P}$ of sets that satisfies the property: if $A, B \in \mathcal{P}$ and $A \cap B \neq \emptyset$, then $A \cap B, A \cup B \in \mathcal{P}$. A combinatorial theory of patchworks, relevant to phylogenetics, was developed in Böcker and Dress (2001). Patchworks were also referred to as 'intersecting families' in earlier work by Lovász (1983, p. 240). The following is a generalisation of Dress and Steel (2009, Lemma 1.2), and the proof follows a similar argument to that result.

Corollary 3 *If τ is slim, then the collection $\mathcal{P}$ of non-empty subsets τ' of τ such that $|s| \geq 3$ for all $s \in \tau'$ and* $\mathrm{exc}'(\tau') = 0$ *forms a patchwork.*

Proof Suppose $\tau_1, \tau_2 \in \mathcal{P}$ satisfy $\tau_1 \cap \tau_2 \neq \emptyset$. For $i = 1, 2$, notice that $\mathrm{exc}'(\tau_i) = \gamma(\tau_i) - 2$ and so, by the submodularity property of γ from Theorem 7, we have:

$$\mathrm{exc}'(\tau_1) + \mathrm{exc}'(\tau_2) \geq \mathrm{exc}'(\tau_1 \cup \tau_2) + \mathrm{exc}'(\tau_1 \cap \tau_2), \tag{7}$$

noting that $\mathrm{exc}'(\tau_1 \cap \tau_2)$ is well defined by the condition that $\tau_1 \cap \tau_2 \neq \emptyset$. Since $\mathrm{exc}'(\tau_1) = \mathrm{exc}'(\tau_2) = 0$, Inequality (7) gives:

$$0 \geq \mathrm{exc}'(\tau_1 \cup \tau_2) + \mathrm{exc}'(\tau_1 \cap \tau_2).$$

It follows that the terms $\mathrm{exc}'(\tau_1 \cup \tau_2)$ and $\mathrm{exc}'(\tau_1 \cap \tau_2)$ on the right of this inequality must both be zero since $\tau_1 \cup \tau_2$ and $\tau_1 \cap \tau_2$ are non-empty subsets of the slim set τ and so each has non-negative excess. Thus, $\tau_1 \cup \tau_2, \tau_1 \cap \tau_2 \in \mathcal{P}$, as required. □

6 Discussion

When an evolutionary biologist compares a number of trees on different, but overlapping, leaf sets, it is typically very rare that these trees are found to be compatible, due mainly to errors in the estimation of phylogenetic trees. Thus, in cases where the trees are compatible this fact alone may provide the biologist with some heightened confidence in the accuracy of the input trees. However, such confidence should clearly depend, in part, on the pattern of taxon coverage. In the extreme case where the subsets of species on which the input trees were built from a phylogenetically flexible collection, it is clear that compatibility provides absolutely no hint of accuracy of the input trees, since *any* trees that had been considered for those subsets would be compatible. For applications, it might therefore be useful to quantify how close to 'phylogenetically flexible' a given pattern of taxon coverage is.

Our results also suggest a second possible future research direction. Since submodular functions are connected to matroid theory, are there relevant connections between thin/slim sets and matroids? Other matroid structures in phylogenetics have been recently been described, in different contexts, by Dress et al. (2014) and Hellmuth and Seemann (2017).

Acknowledgements We thank the organisers of the Algebraic and Combinatorial Phylogenetics Workshop (Barcelona, 26–30 June 2017) where some of the ideas in this paper were conceived, and the London Mathematical Society for supporting the visit of KTH and VM to visit MS in New Zealand. We also thank the two anonymous reviewers of this paper for numerous helpful suggestions.

References

Aho AV, Sagiv Y, Szymanski TG, Ullman JD (1981) Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. SIAM J Comput 10:405–421

Bixby RE, Cunningham WH (1995) Matroid optimization and algorithms. In: Graham RL et al (eds) Handbook for combinatorics, vol 1. Elsevier, New York, pp 551–609

Böcker S, Dress AWM (2001) Patchworks. Adv Math 157:1–21

Bryant DJ, Steel M (1995) Extension operations on sets of leaf-labelled trees. Adv Appl Math 16(4):425–453

Dress A, Steel M (2009) A Hall-type theorem for triplet set systems based on medians in trees. Appl Math Lett 22:1789–1792

Dress A, Huber KT, Steel M (2014) A matroid associated with a phylogenetic tree. Discret Math Theor Comput Sci 16(2):41–56

Felsenstein J (2004) Inferring phylogenies. Sinauer Associates, Sunderland

Fritzilas E, Milanič M, Monnot J, Rios-Solis YA (2013) Resilience and optimization of identifiable bipartite graphs. Discret Appl Math 161(4):593–603

Grötschel M, Lovász L, Schrijver A (1981) The ellipsoid method and its consequences in combinatorial optimization. Combinatorica 1:169–197

Grünewald S (2012) Slim sets of binary trees. J Comb Theory A 119:323–330

Grünewald S, Huber KT , Moulton V, Steel M (2017) Combinatorial properties of triplet covers for binary trees. arXiv:1707.07908

Hall P (1935) On representatives of subsets. J Lond Math Soc 10(1):26–30

Hellmuth M, Seemann CR (2017) The matroid structure of representative triple sets and triple-closure computation. arXiv:1707.01667

Lovász L (1983) Submodular functions and convexity. In: Mathematical programming the state of the art. Springer, pp 235–257

Lovász L, Plummer MD (1986) Matching theory. Elsevier, New York

Sanderson MJ, McMahon MM, Steel M (2011) Terraces in phylogenetic tree space. Science 333:448–450

Semple C, Steel M (2003) Phylogenetics. Oxford University Press, Oxford

Steel M, Sanderson MJ (2010) Characterizing phylogenetically decisive taxon coverage. Appl Math Lett 23:82–86

Welsh DJA (1995) Matroids: fundamental concepts. In: Graham RL et al (eds) Handbook for combinatorics. Elsevier, New York, pp 481–550