Patching Up X-trees

Sebastian Böcker[1*], Andreas W.M. Dress[1†], and Mike A. Steel [2‡]

[1]GK Strukturbildungsprozesse, FSP Mathematisierung, Universität Bielefeld, PF 100 131,
 33501 Bielefeld, Germany
 {boecker, dress}@mathematik.uni-bielefeld.de
[2]Biomathematics Research Centre, University of Canterbury, Private Bag 4800, Christchurch,
 New Zealand
 m.steel@math.canterbury.ac.nz

Abstract. A fundamental problem in many areas of classification, and particularly in biology, is the reconstruction of a leaf-labeled tree from just a subset of its induced subtrees. Without loss of generality, we may assume that these induced subtrees all have precisely four leaves. Of particular interest is the question of determining whether a collection of quartet subtrees uniquely defines a parent tree. Here, we solve this question in the case where the collection of quartet trees is of minimal size, by studying *encodings* of binary trees by such quartet trees. We obtain a characterization of minimal encodings that exploits an underlying "patchwork" structure. As we will show elsewhere, this allows one to obtain a polynomial time algorithm for certain instances of the problem of reconstructing trees from subtrees.

Keywords: trees, supertrees, tree amalgamation, quartet encodings of trees, hierarchies, patchworks

1. Introduction

Trees are widely used to represent evolutionary, historical, or hierarchical relationships in various fields of classification. In biology for example, such trees ("phylogenies") typically represent the evolutionary history of a collection of extant species or the line of descent of some gene [17]. They may also be used to classify individuals (or populations) of the same species [6]. In historical linguistics, trees have been used to represent the evolution of languages [18], while in the branch of philology known as stemmatology, trees may represent the way in which different versions of a manuscript arose through successive copying [12].

In most of these applications, the objects of interest occur at the tips (leaves) of the tree, and all other vertices of the tree correspond to a branching (or speciation) event. From a mathematical perspective, we have a finite set X [1] of objects of interest (species, languages, etc.), and we consider triples $T = (V, E; \phi)$ consisting of a tree (V, E) with finite vertex set $V = V_T$ and edge set $E = E_T \subseteq \binom{V}{2}$, and a map $\phi = \phi_T : X \to V$ such that

$$v \in \phi(X) \quad \text{holds for all} \quad v \in V \quad \text{with} \quad \deg_T(v) \leq 2, \tag{1.1}$$

where $\deg_T(v)$ denotes the degree $\#\{e \in E_T : v \in e\}$ of the vertex $v \in V_T$.

A triple $T = (V, E; \phi)$ that satisfies these conditions is henceforth called an *X-tree*. Denoting the set of vertices of degree i by V_i, we define an X-tree $T = (V, E; \phi)$ to be a *phylogenetic X-tree* if ϕ is a bijection from X onto the set V_1 of *leaves* of (V, E). In addition, if every vertex in V is of degree either 1 or 3 (in which case (V, E) is called a *binary tree*), then we will say that $T = (V, E; \phi)$ is a *binary X-tree*. Two X-trees $T = (V, E; \phi)$ and $T' = (V', E'; \phi')$ are *isomorphic* if there exists a bijection $\alpha : V \xrightarrow{\sim} V'$ that induces a bijection between E and E' and satisfies $\phi' = \alpha \circ \phi$ (in which case there is exactly one such map α).

As is well known, there is a canonical and useful one-to-one correspondence between (isomorphism classes of) X-trees and certain set systems due to P. Buneman (cf. [5]) that we shall now recall.

A *split* of X is a subset $\{A, B\} \subseteq P(X)$ such that A, B is a bipartition of X into two non-empty, disjoint subsets; a *partial split* of X is a split of some non-empty subset of X. We let $S(X)$ (resp. $S_{\mathrm{part}}(X)$) denote the set of all splits of X (resp. all partial splits of X). Two splits S_1, S_2 are called *compatible* if there exist $A_1 \in S_1$ and $A_2 \in S_2$ with $A_1 \cap A_2 = \emptyset$. A partial split $\{A, B\}$ is *trivial* if $\min\{\#A, \#B\} = 1$. A partial split $S_1 = \{A_1, B_1\}$ is said to *extend* a partial split $S_2 = \{A_2, B_2\}$ if $S_2 = \{(A_2 \cup B_2) \cap A_1, (A_2 \cup B_2) \cap B_1\}$ (that is, $A_2 \subseteq A_1$ and $B_2 \subseteq B_1$, or $A_2 \subseteq B_1$ and $B_2 \subseteq A_1$) holds in which case we will also write $S_2 \leq S_1$.

Now, each edge e in an X-tree $T = (V, E; \phi)$ gives rise to a split of X — simply delete e from E and apply ϕ^{-1} to the two connected components of the resulting graph $(V, E - \{e\})$ to obtain a split that we will call a *T-split* and denote by $S[e] = S_T[e]$. Let $S[T]$ denote the set of all T-splits. It is easily checked that distinct edges induce distinct splits and that the set $S[T]$ is *compatible*, that is, any two splits from $S[T]$ are compatible. Furthermore, we have $\#S[T] \leq 2\#X - 3$ with equality precisely if T is a binary X-tree. This follows easily from the following fundamental property of trees: If, for a tree (V, E), we denote the set of *inner edges* not incident with a leaf by $\mathring{E} \subset E$, then

$$\#\mathring{E} \leq \#V_1 + \#V_2 - 3 \tag{1.2}$$

holds for every finite tree (V, E), while equality holds for a tree with $\#V_2 = 0$ if and only if that tree is binary.

Buneman established the following fundamental correspondences in [5]:

Lemma 1.1. *The map $T \rightsquigarrow S[T]$ induces bijections between*:

(i) *the set of* (*isomorphism classes of*) *X-trees and the set of compatible split systems $S \subseteq S(X)$*

---

[1] Throughout this paper, let $\#X$ denote the cardinality of a finite set X, and $P(X)$ the set of its subsets.

(ii) *the set of* (*isomorphism classes of*) *binary X-trees and the set of compatible split systems* $S \subseteq S(X)$ *for which* #$S = 2$#$X - 3$ *holds, or equivalently, the set of maximal compatible split systems.*

This correspondence between X-trees and compatible split systems provides a convenient partial order on the set of (isomorphism classes of) X-trees: We write $T' \leq T$ precisely if $S[T'] \subseteq S[T]$.

Now, given an X-tree $T = (V, E; \phi)$ and a non-empty subset $Y \subseteq X$, we obtain an induced Y-tree $T|_Y$ as follows: First, construct the minimal subtree (V', E') of (V, E) that connects all vertices from $\phi(Y)$. Then make this tree "homeomorphically irreducible" by replacing each maximal path running (except for its two end points) through degree-two vertices from $V' - \phi(Y)$ only, by a single edge (and deleting the superfluous vertices and edges) to obtain a tree (V_Y, E_Y). The restriction $\phi|_Y =: \phi_Y$ maps Y into V_Y and satisfies condition (1.1) (with X, V, and T replaced by Y, V_Y, and $T|_Y$, respectively). We call the resulting Y-tree $T|_Y = (V_Y, E_Y; \phi_Y)$ the *induced* (*Y*-)subtree of T. We will say that an X-tree T *displays* a Y-tree T' if $T' \leq T|_Y$ holds.

A more succinct but less visual description of $T|_Y$ is, in view of Lemma 1.1, to describe its set of splits:

$$S[T|_Y] = \{S' \in S(Y) : S' \leq S \text{ holds for some } S \in S[T]\}$$

$$= \{\{A \cap Y, B \cap Y\} : \{A, B\} \in S[T] \text{ and } A \cap Y, B \cap Y \neq \emptyset\}. \qquad (1.3)$$

Our interest lies in the reverse reconstruction problem: Given as input a collection of subtrees, we wish to determine whether we can *amalgamate* them into a common *supertree*, that is, whether there exists some X-tree that displays these subtrees, and if so, whether there exists exactly one such tree. Formally, let $(Y_1, \ldots, Y_k)$ be a family of non-empty subsets of X, and consider a family $F := (T_1, \ldots, T_k)$ where T_j is a Y_j-tree for $j = 1, \ldots, k$. We may wish to consider the set $T(F)$ of all (isomorphism classes of) phylogenetic X-trees T that display every tree in F.

As we will see in the following section, the related and seemingly more special task of reconstructing trees from partial splits is actually equivalent to the task described above; so, the problem of computing $T(F)$ can always be reduced to this particular version of the reconstruction problem.

There are several reasons why such reconstruction problems arise naturally in applications such as biology. Firstly, we may wish to combine trees that have been reconstructed using distinct, though overlapping, collections of species (usually by different researchers using different data and, as often as not, different reconstruction methods). A second reason is that, in general, it is difficult to accurately reconstruct large trees directly, and we may choose instead to reconstruct trees for small subsets and then combine these in a parent tree (or parent trees) (see [1, 10, 11, 16, 19]). A third reason is that, for genetic data, the number of sites that can be accurately aligned across a small number of closely related sequences is generally much larger than the corresponding number of sites for a set that is large and includes rather diverse sequences.

If we were to construct F by estimating a tree for *every* subset of size 4, then (as #$T(F) \leq 1$ must hold, cf. [1]) we can easily compute $T(F)$ and will usually find that $T(F) = \emptyset$ holds, that is, some of the subtrees must have been incorrectly estimated (this was already known to Colonius and Schulze [7, 8], see also [1, 19]). Thus, we may wish to use only those subtrees that are strongly supported by the data (usually

involving closely related objects), and so we will generally have available trees for only a small number of subsets of X. This makes the reconstruction problem more difficult computationally but, of course, also more gratifying whenever one is led this way to a simultaneously non-empty and well-supported set of trees.

Of course, we could examine all X-trees to determine which (if any) of them display (every tree in) F; however, this is computationally infeasible since even the number of non-isomorphic binary X-trees grows super-exponentially with the number of leaves. Indeed, this number is precisely the product $1 \cdot 3 \cdots (2\#X - 5)$ of the first $(\#X - 2)$ odd numbers, a result that dates back to 1870 (see [13]). This motivates the results described below.

We wish to thank David Torney for some encouraging and helpful comments on this work.

2. Partial Splits

Given a partial split $S = \{A, B\} \in S_{\text{part}}(X)$, we define the *support* of S by $\underline{S} := A \cup B$, and for every $x \in \underline{S}$, we define $S(x) := A$ if $x \in A$ and $S(x) := B$ otherwise; for $S \subseteq S_{\text{part}}(X)$, we define $\underline{S} := \bigcup_{S \in S} \underline{S}$.

For natural numbers i, j, let

$$S_{i,j}(X) := \left\{ \{A, B\} \in S_{\text{part}}(X) : \{\#A, \#B\} = \{i, j\} \right\}. \tag{2.1}$$

A partial split $Q \in Q(X) := S_{2,2}(X)$ is called a *quartet split*, and we use $Q = xy|wz$ as shorthand for $Q = \{\{x, y\}, \{w, z\}\}$.

Given a subset $S \subseteq S_{\text{part}}(X)$ of partial splits of X and an X-tree T, we say that T is *concordant* with S if, for every $\{A, B\} \in S$, there exists at least one edge e of T that separates $\phi(A)$ from $\phi(B)$, that is, with $\{A, B\} \leq S[e]$. Let $T_X(S) = T(S)$ denote the set of all (isomorphism classes of) phylogenetic X-trees concordant with S.

Note that $T \in T(S)$ and $T \leq T'$ implies $T' \in T(S)$, hence any X-tree T with $\{T\} = T(S)$ for some set $S \subseteq S_{\text{part}}(X)$ must be a binary X-tree.

Note also that a collection S of partial splits with $T(S) = \{T\}$ for some binary X-tree $T = (V, E; \phi)$ must contain at least one partial split for every inner edge that specifically "fits" this edge and, hence, S must contain at least $\#\mathring{E} = \#X - 3$ distinct non-trivial splits in view of inequality (1.2). It is easily shown [1, 14] that, for $S \subseteq S_{\text{part}}(X)$ and

$$Q(S) := \{Q \in Q(X) : Q \leq S \text{ for some } S \in S\}, \tag{2.2}$$

the relation

$$T(S) = T(Q(S))$$

must hold. Thus, there is no loss of generality in restricting one's attention to quartet splits when reconstructing (phylogenetic) trees from partial splits. Similarly, we have

$$T(F) = T\left(\bigcup_{i=1,\dots,k} S[T_i] \right) = T\left(Q\left(\bigcup_{i=1,\dots,k} S[T_i] \right) \right) \tag{2.3}$$

for any family $F = (T_1, \dots, T_k)$ as above, so the problem of reconstructing trees from subtrees also reduces to the problem of reconstructing them from (quartet) splits.

3. Quartet Encodings

In this section, we analyze conditions under which a binary *X*-tree is uniquely deter-mined by selecting, for each inner edge e, a corresponding representative quartet split Q with $Q \leq S[e]$. In order to describe our main result (Theorem 3.11), we must introduce more terminology and preliminary results.

Definition 3.1. *Let* $T = (V, E; \phi)$ *denote a binary X-tree. A map*

$$q : \mathring{E} \to \mathcal{Q}(X)$$

is called a quartet encoding of T if $S[e]$ *extends* $q(e)$ *for each* $e \in \mathring{E}$. *For every* $e \in \mathring{E}$ *and* $F \subseteq \mathring{E}$, *we define*

$$\underline{q}(e) := \underline{q(e)} \subseteq X, \quad q(F) := \{q(e) : e \in F\} \subseteq \mathcal{Q}(X), \quad and \quad \underline{q}(F) := \bigcup_{e \in F} \underline{q}(e).$$

We will say that q defines T if T is the only phylogenetic X-tree concordant with $q(\mathring{E})$. *A quartet encoding q is called tight if, for each edge* $e \in \mathring{E}$, *there exists no other edge in* $\mathring{E}$ *separating the two subsets in* $\underline{q}(e)$.

It is easy to see (cf. [14]) that a quartet encoding that defines a tree T is tight; furthermore, given a binary *X*-tree T and a tight quartet encoding q of T, then $\underline{q}(\mathring{E}) = X$ and $\#q(\mathring{E}) = \#X - 3$ holds. It is also easy to see that $T(\mathcal{Q}) = \{T\}$ holds for some set of quartet splits $\mathcal{Q}$ with $\#\mathcal{Q} = \#X - 3$ and some (necessarily binary) *X*-tree T if and only if there exists a quartet encoding q of T with $\mathcal{Q} = q(\mathring{E})$ that defines T.

We now present three instructive examples of tight encodings.

Example 3.2. For $X := \{1, \dots, 6\}$, consider the binary *X*-tree $T_1 = (V, E; \phi_1 := Id_X)$ with $E := \{e_1, \dots, e_9\}$ having the nontrivial splits

$$S[e_1] = \{\{1, 2\}, \{3, 4, 5, 6\}\},$$
$$S[e_2] = \{\{3, 4\}, \{1, 2, 5, 6\}\},$$

and

$$S[e_3] = \{\{5, 6\}, \{1, 2, 3, 4\}\},$$

plus the six trivial splits as depicted in Figure 1(a). Consider the quartet encoding $q : \mathring{E} = \{e_1, e_2, e_3\} \to \mathcal{Q}(X)$ defined by

$$q(e_1) := 12|45, \quad q(e_2) := 34|16 \quad and \quad q(e_3) := 56|23. \tag{3.1}$$

Then q is a tight encoding of T_1, but does not define T_1, as it also encodes the X-tree T_2 depicted in Fig. 1 (b). A construction generalizing this example can be used to prove the non-existence of consensus methods that are equivariant and Pareto on subtrees, see [15].

Example 3.3. Let $X := \{1, \dots, 7\}$, and suppose $T = (V, E; \phi := Id_X)$ is a caterpillar with seven leaves as depicted in Figure 2, and $q : \mathring{E} \to \mathcal{Q}(\{1, \dots, 7\})$ is the following quartet encoding:

$$q(e_1) := 12|36, \ q(e_2) := 13|46, \ q(e_3) := 24|57, \ and \ q(e_4) := 25|67.$$

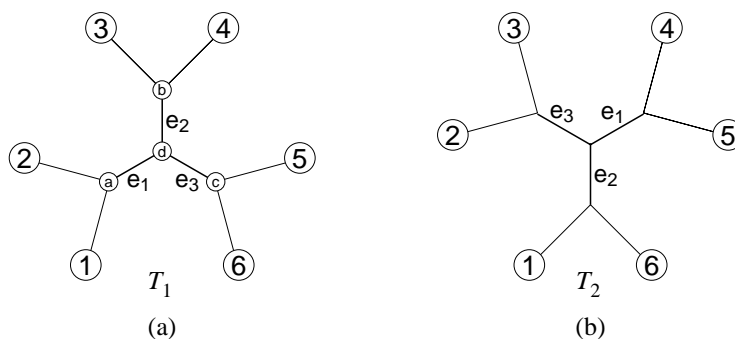Then it is easy to check that q defines T.

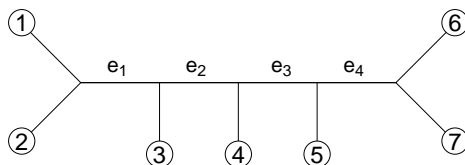Figure 1: Two possible binary trees for Example 3.2.

Figure 2: Caterpillar with seven leaves.

Example 3.4. Suppose q is any tight quartet encoding of a binary X-tree with

$$\bigcap_{e\in\mathring{E}} \underline{q}(e) \neq \emptyset. \tag{3.2}$$

Then it has been observed in [14] — and it follows immediately from Theorem 3.11 below — that q defines T.

To exploit inequality (1.2), we now introduce the following definition:

Definition 3.5. *Given a binary X-tree $T = (V, E; \phi)$, a subset $F \subseteq \mathring{E}$, and a quartet encoding $q : \mathring{E} \to \mathcal{Q}(X)$, the (q-)excess of F is given by*

$$\mathrm{exc}(F) = \mathrm{exc}_q(F) := \#\underline{q}(F) - \#F - 3. \tag{3.3}$$

We say F is (q-)excess-free if $\mathrm{exc}(F) = 0$ holds.

Note that $\mathrm{exc}(\{e\}) = 0$ holds for every $e \in \mathring{E}$, and that $\mathrm{exc}(\emptyset) = -3$.

Lemma 3.6. *Suppose q is a tight quartet encoding of a binary X-tree $T = (V, E; \phi)$. Then*

  (i) $\mathrm{exc}(\mathring{E}) = 0$;

  (ii) *for every non-empty subset $F \subseteq \mathring{E}$, one has $\mathrm{exc}(F) \geq 0$;*

 (iii) *if F is excess-free, then F is a connected subset of edges in T, and F equals the set of inner edges of $T\big|_{\underline{q}(F)}$.*

Sketch of proof. (i) This is merely restating that a binary tree with n leaves has $n-3$ inner edges, cf. [5] for example.

(ii) Consider $T|_Y = (V_Y, E_Y; \phi_Y)$ with $Y := \underline{q}(F)$ and note that, as q is tight, we have $F \subseteq (E_Y)°$. Hence, (1.2) implies

$$\#F \le \#(E_Y)° \le \#Y - 3 = \#\underline{q}(F) - 3,$$

as claimed.

(iii) As above, we consider $T|_Y = (V_Y, E_Y; \phi_Y)$ for $Y := \underline{q}(F)$. If F were not connected, then $F \subsetneq (E_Y)°$ since $(E_Y)°$ is connected. From (1.2), we would conclude $\#F < \#(E_Y)° \le \#Y - 3$ and, hence, $\mathrm{exc}(F) > 0$. ∎

To illustrate the usefulness of these concepts, we now show how they provide further constructions of encodings that define a binary X-tree T. First, we make a further definition: Let us say that two distinct elements x, y of X form a pair of *twins* of T if the two edges incident with $\phi(x)$ and $\phi(y)$, respectively, share a vertex $v = v(x, y)$, in which case the third edge incident with v is denoted by $e(x, y)$. In Example 3.2, for instance, ① and ② form a pair of twins for the tree depicted in Figure 1(a) with $e(①, ②) = e_1$. If $\#X \ge 4$, every binary X-tree has at least two pairs of twins, and one has $e(x, y) \in \mathring{E}$ for every pair x, y of twins.

Example 3.7. Suppose $T = (V, E; \phi)$ is a binary X-tree such that the inner edges of T are labeled $\mathring{E} = \{e_1, \ldots, e_{n-3}\}$, and that q is a tight quartet encoding of T with

$$\#\left(\underline{q}(e_i) \setminus \bigcup_{j<i} \underline{q}(e_j)\right) = 1 \quad \text{for} \quad i = 2, \ldots, n-3. \tag{3.4}$$

Then q defines T.

Proof. We apply induction on n. The result holds for $n = 4$, so suppose it holds for $4, \ldots, n-1$ and that $\#X = n$. Let T' denote another binary X-tree that is concordant with $q(\mathring{E})$. By assumption, there exists $x \in X$ with $x \in \underline{q}(e_{n-3})$, but $x \notin \underline{q}(e_j)$ for $j = 1, \ldots, n-4$. We define $Y := X - \{x\}$ and $F := \{e_1, \ldots, e_{n-4}\}$, then $q|_F$ defines $T|_Y$ as well as $T'|_Y$ and, hence, $T|_Y \cong T'|_Y$ by the induction hypothesis. It remains to show that x is attached to the same edge of T and T'. To this end, we infer from Lemma 3.6 (iii) that F is connected, since q is tight and $\mathrm{exc}(F) = 0$. So x must have a twin in T, denoted $y \in Y$, and $\{x, y\} \in q(e_{n-3})$ must hold. But the same holds true for T' which indeed implies $T \cong T'$. ∎

Example 3.8. For a binary X-tree $T = (V, E; \phi)$, define $\mathrm{clus}(T) := \bigcup_{e \in E} S[e]$, the set of *clusters* of T. Suppose $f : \mathrm{clus}(T) \to X$ is any function that satisfies the condition

$$f(A) \in A \quad \text{for all} \quad A \in \mathrm{clus}(T). \tag{3.5}$$

Then f defines a tight quartet encoding of T, denoted q_f, as follows: For each inner edge $e = \{v_1, v_2\}$, deletion of v_1 and v_2 and their incident edges partition T into four connected components and, thereby, it partitions X into four sets $\{A_1, A_2, B_1, B_2\}$,

where we may suppose, without loss of generality, that $S[e] = \{\{A_1 \cup A_2\}, \{B_1 \cup B_2\}\}$. Let

$$q_f(e) := f(A_1)f(A_2) \mid f(B_1)f(B_2).$$

Clearly, not every tight quartet encoding is of this form; furthermore, q_f does not necessarily define T. However, if f satisfies the condition

$$f(A) \in B \subseteq A \Longrightarrow f(A) = f(B) \quad \text{for all} \quad A, B \in \text{clus}(T) \tag{3.6}$$

then q_f satisfies the condition described in Example 3.7; in particular, q_f defines T.

Sketch of proof. It suffices to show that the edges of T can be labeled as described in Example 3.7 (so as to satisfy (3.4)) which is again achieved by induction on $n := \#X$. Let $\{x, x'\}$ be a twin in T, and suppose $f(\{x, x'\}) = x'$. We define $Y := X - \{x\}$ and $T' := T|_Y$. Then, by the condition placed on f, we obtain a corresponding function $f' : \text{clus}(T') \to Y$ that satisfies (3.5) with f, T replacing by f', T', and hence, the edges of T' can be ordered $\{e_1, \ldots, e_{n-4}\}$ so as to satisfy the condition (3.6). We then label the edge $e(x, x')$ of T incident with the twin $\{x, x'\}$ as e_{n-3}, and verify that this also satisfies (3.4) for $i = n - 3$. ∎

This last example generalizes a result from [9] where a specific function f satisfying (3.6) is considered.

The excess-free subsets of $\mathring{E}$ for a tight quartet encoding q have a useful "patchwork" structure that we now discuss. Following [3], a collection $\mathcal{C}$ of subsets of a set M is called an *M-patchwork* if it satisfies the following condition:

$$\emptyset \neq \mathcal{C}' \subseteq \mathcal{C} \quad \text{and} \quad \bigcap \mathcal{C}' \neq \emptyset \quad \Longrightarrow \quad \bigcap \mathcal{C}', \bigcup \mathcal{C}' \in \mathcal{C} \tag{3.7}$$

There are quite a number of distinct characterizations of patchworks that are *ample*, that is (cf. [3]), patchworks $\mathcal{C}$ that satisfy the condition

$$A, B \in \mathcal{C} \quad \text{and} \quad \#\{C \in \mathcal{C} : A \subseteq C \subseteq B\} = 2 \Longrightarrow B - A \in \mathcal{C}, \tag{3.8}$$

in particular,

(i) a patchwork $\mathcal{C} \subseteq P(M)$ is ample if and only if, for every cluster $C \in \mathcal{C}$ with $\mathcal{C}' := \{C' \in \mathcal{C} : \emptyset \neq C' \subset C\} \neq \emptyset$, there exist either two clusters $A, B \in \mathcal{C}'$ with $C = A \cup B$, or there exists a chain $\mathcal{C}'' \subset \mathcal{C}'$ with $\bigcup \mathcal{C}'' = C$;

(ii) a patchwork $\mathcal{C} \subseteq P(M)$ with $\{m\} \in \mathcal{C}$ for all $m \in M$ and $\emptyset, M \in \mathcal{C}$ is ample if and only if $\mathcal{C}$ contains a *maximal hierarchy* $\mathcal{C}'$, that is, a maximal subset $\mathcal{C}'$ of $P(M)$ for which $A \cap B \in \{\emptyset, A, B\}$ holds for all $A, B \in \mathcal{C}'$.

Lemma 3.9. *Suppose q is an arbitrary quartet encoding of a binary X-tree $T = (V, E; \phi)$ and assume $F_1, F_2 \subseteq \mathring{E}$. Then*

$$\text{exc}(F_1 \cup F_2) + \text{exc}(F_1 \cap F_2) \leq \text{exc}(F_1) + \text{exc}(F_2). \tag{3.9}$$

Proof. In view of $\underline{q}(F_1 \cap F_2) \subseteq \underline{q}(F_1) \cap \underline{q}(F_2)$, we have

$$\mathrm{exc}(F_1 \cup F_2) + \mathrm{exc}(F_1 \cap F_2) + 6$$

$$= \#\underline{q}(F_1 \cup F_2) - \#(F_1 \cup F_2) + \#\underline{q}(F_1 \cap F_2) - \#(F_1 \cap F_2)$$

$$\leq \#\big(\underline{q}(F_1) \cup \underline{q}(F_2)\big) + \#\big(\underline{q}(F_1) \cap \underline{q}(F_2)\big) - \big(\#(F_1 \cup F_2) + \#(F_1 \cap F_2)\big)$$

$$= \#\underline{q}(F_1) + \#\underline{q}(F_2) - \#F_1 - \#F_2$$

$$= \mathrm{exc}(F_1) + \mathrm{exc}(F_2) + 6.$$

∎

Lemma 3.10. *If q is a tight quartet encoding of $T = (V, E; \phi)$, then the excess-free subsets of $\mathring{E}$ form a patchwork denoted by $C(q)$.*

Proof. Let $F_1, F_2 \subseteq \mathring{E}$ denote subsets with $\mathrm{exc}(F_1) = \mathrm{exc}(F_2) = 0$ and $F_1 \cap F_2 \neq \emptyset$. By Lemma 3.6(ii), we have

$$\mathrm{exc}(F_1 \cap F_2) \geq 0 \quad \text{and} \quad \mathrm{exc}(F_1 \cup F_2) \geq 0,$$

yet by Lemma 3.9,

$$\mathrm{exc}(F_1 \cap F_2) + \mathrm{exc}(F_1 \cup F_2) \leq 0.$$

Consequently, $\mathrm{exc}(F_1 \cap F_2) = \mathrm{exc}(F_1 \cup F_2) = 0$, so $C(q)$ must be a patchwork. ∎

It is tempting to conjecture that if q defines a binary X-tree T, then there exists always an edge $e \in \mathring{E}$ such that $\mathring{E} - \{e\}$ is q-excess-free. In some cases (e.g., for $\#X \leq 6$ and for encodings constructed as in Examples 3.7 and 3.8 above), this is indeed the case, but in general it is not (Example 3.3 provides a counterexample with seven leaves). Nevertheless, we can still hope that the excess-free subsets of $\mathring{E}$ do form at least an ample patchwork, and this turns out to be indeed the case, as we state now as part of our main result:

Theorem 3.11. *Given a quartet encoding q of a binary X-tree $T = (V, E; \phi)$, the following three statements are equivalent:*

 (i) *q defines T;*
(ii) *q is tight, and the patchwork $C(q)$ of excess-free subsets of $\mathring{E}$ is ample;*
(iii) *if Q^* denotes the smallest subset of $Q(X)$ containing $q(\mathring{E})$ and containing all quartet splits $ab|cd$ for which some $x \in X$ exists with either $ab|cx, ab|dx \in Q^*$ or $ax|cd, ab|cx \in Q^*$, then Q^* contains exactly one of the three quartet splits $ab|cd$, $ac|bd$, or $ad|bc$ for any four distinct elements $a, b, c, d \in X$.*

The implications (ii) $\Rightarrow$ (iii) and (iii) $\Rightarrow$ (i) follow easily from combining the lines of thought used already in the previous examples with standard results from [1], see also [7,8]. The implication (i) $\Rightarrow$ (ii) — except for the fact that q must be tight — is far

from trivial: Of course, one proceeds by induction relative to $n := \#X$. And this allows us to assume that

$$\mathcal{C}(q)_{\subseteq F} := \{F' \in \mathcal{C}(q) : F' \subseteq F\}$$

is an ample patchwork for all subsets F of X contained in

$$\mathcal{C}(q)_{\subset X} := \{F' \in \mathcal{C}(q) : F' \subseteq X \text{ and } F' \neq X\}.$$

It then follows easily (cf. [3], Lemma 4) that

$$\max(\mathcal{C}(q)_{\subset X}) := \{F \in \mathcal{C}(q)_{\subset X} : F \subseteq F' \in \mathcal{C}(q)_{\subset X} \text{ implies } F = F'\}$$

must be a partition of X into at least three distinct subsets.

The next step consists of applying the induction hypothesis to trees one derives from T by identifying pairs of twins $x, y \in X$. This way, one is led to study decompositions of $\mathring{E}$ into two disjoint and connected subsets $F_1 = F_1(x, y)$ and $F_2 = F_2(x, y)$ with $e(x, y) \in F_1$,

$$\#\Big(\underline{q}\big(F_1 - \{e(x, y)\}\big) \cup \{x, y\}\Big) = \#F_1 + 3,$$

and $\mathrm{exc}(F_2) = 0$ or $\#(\underline{q}(F_2) - \{x, y\}) = \#F_2 + 2$. Next, one shows — and this is the most tricky part of the whole proof — that (i) by choosing $F_1 = F_1(x, y)$ maximal subject to these conditions, one can always assume $\mathrm{exc}(F_2) = 0$, and (ii) that $\mathrm{exc}(F_2) = 0$ and $\#F_2 \geq 2$ would in turn imply the existence of a connected subset $F_2' \in \mathcal{C}(q)$ with $F_2 \subseteq F_2'$ and $\mathring{E} - F_2' \in \mathcal{C}(q)$ in contradiction to $\#\max(\mathcal{C}(q)_{\subset X}) \geq 3$. Consequently, we can assume, for every pair x, y of twins in T, there exists a single edge $f(x, y) \in \mathring{E} - \{e(x, y)\}$ such that $F_1 := \mathring{E} - \{f(x, y)\}$ and $F_2 := \{f(x, y)\}$ is a pair as above, that is, with

$$\#\Big(\underline{q}(\mathring{E} - \{e(x, y), f(x, y)\}) \cup \{x, y\}\Big) = \#\big(\mathring{E} - \{f(x, y)\}\big) + 3$$
$$= (\#X - 4) + 3 = \#X - 1$$

which in turn implies that $f(x, y)$ must be of the form $e(x', y')$ for some pair x', y' of twins (because $F_1 = \mathring{E} - \{f(x, y)\}$ is connected) and with, say, x' the unique element in X missing in $\underline{q}(\mathring{E} - \{e(x, y), f(x, y)\})$. Applying the same argument now to the twins x', y' and repeating this process iteratively will therefore eventually produce a sequence of distinct pairs $x_1, y_1; x_2, y_2; \ldots; x_k, y_k$ of twins such that, for $i = 1, \ldots, k \pmod{k}$, we have

$$\underline{q}(\mathring{E} - \{e(x_i, y_i), e(x_{i+1}, y_{i+1})\}) \cup \{x_i, y_i\} = X - \{x_{i+1}\}$$

which in turn implies easily that we can construct a second X-tree $(V', E'; \phi')$ that is concordant with $q(\mathring{E})$: Note first that our assumptions imply that, for each $i = 1, \ldots, k$, there must exist some $z_i \in X$ with $x_i y_i | x_{i+1} z_i = q(e(x_i, y_i))$, and that $z_i \in \{x_1, \ldots, x_k\}$ cannot hold because, according to our construction, $e(x_i, y_i)$ is the *only* edge in $\mathring{E} - \{e(x_{i+1}, y_{i+1})\}$ with $x_{i+1} \in \underline{q}(e(x_i, y_i))$. Hence, we can cut off, for every $i = 1, \ldots, k$
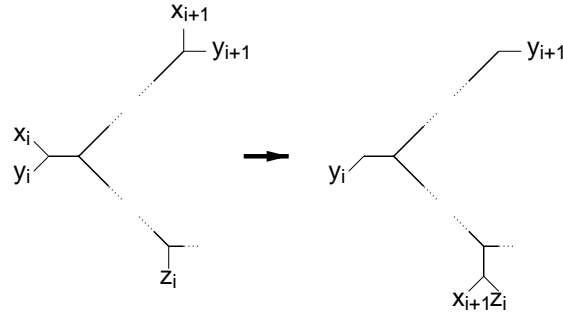
Figure 3: Cutting off and re-implanting edges.

(mod k), the edge incident with $\phi(x_{i+1})$ and implant it instead into the edge incident with $\phi(z_i)$ as depicted in Figure 3.

If $z_i = z_{i+1}$, we have to make sure that the edge leading to $\phi(x_{i+2})$ is implanted closer to $\phi(z_i)$ than the edge leading to $\phi(x_{i+1})$; no further special care needs to be taken (and because we cannot have $z_1 = z_2 = \cdots = z_k$, this requirement can always be fulfilled).

Finally, defining V', E' and ϕ' accordingly, we find a second *X*-tree $(V', E'; \phi')$ in $T(\underline{q}(\mathring{E}))$ that is clearly non-isomorphic with $(V, E; \phi)$, the final contradiction. ∎

Remark. Note that, for the tree T_1 considered in Example 3.2 (see also Figure 1), we can choose $f(2i - 1, 2i) \in \{e_j, e_k\}$ for all $\{i, j, k\} = \{1, 2, 3\}$; so we can use the twin sequence $1, 2; 4, 3$ as well as the twin sequence $2, 1; 4, 3; 6, 5$, leading to $z_1 = 5$ and $z_2 = 6$ or $z_1 = 5$, $z_2 = 1$, and $z_3 = 3$, respectively; indeed, both rearrangements lead to a tree isomorphic to T_2.

The detailed proof will take up more than 25 pages of reasoning, and so it will be published elsewhere (see [4]).

Corollary 3.12. *Suppose a set of quartet splits Q with $\#Q = \#\underline{Q} - 3 \geq 2$ is concordant with a unique tree. Then Q is the disjoint union of two proper subsets Q_1 and Q_2 with $\#Q_i = \#\underline{Q}_i - 3$ for $i = 1, 2$ such that $\#T(Q_i) = 1$ holds for $i = 1, 2$.*

It is worth mentioning that the above Theorem implies the result mentioned in Example 3.4: To this end, let $q : \mathring{E} \to Q(X)$ denote a tight quartet encoding of a binary *X*-tree $T = (V, E; \phi)$. Suppose that condition (3.2) is satisfied. We assume $x \in \bigcap_{Q \in \mathcal{Q}} \underline{Q}$, define $v \leq u$ for $u, v \in V$ if the path from $\phi(x)$ to u passes through v, and put

$$V(v) := \{u \in V : v \leq u\} \quad \text{and} \quad \mathring{E}(v) := \{e \in \mathring{E} : e \subseteq V(v)\}.$$

Now, it is easy to see that $C' := \{\mathring{E}(v) : v \in V\}$ is a maximal $\mathring{E}$-hierarchy, and that every non-empty element from C' is in $C(q)$. In view of the results in [3], this implies that $C(q)$ is ample which in turn, according to Theorem 3.11, implies that q defines T.

Applications regarding tree amalgamation algorithms will be discussed in a separate paper [2].

References

1. H.-J. Bandelt and A. Dress, Reconstructing the shape of a tree from observed dissimilarity data, Adv. in Appl. Math. **7** (1986) 309–343.

2. S. Böcker, D. Bryant, A.W. Dress, and M. Steel, Algorithmic aspects of tree amalgamation, in preparation.

3. S. Böcker and A.W. Dress, Patchworks, 1999, to appear in Adv. in Math.

4. S. Böcker, A.W. Dress, and M. Steel, Most parsimonious quartet encodings of binary trees, in preparation.

5. P. Buneman, The recovery of trees from measures of dissimilarity, In: Mathematics in the Archaeological and Historical Sciences, F. Hodson, D. Kendall, and P. Tautu, Eds., Edinburgh University Press, Edinburgh, 1971, pp. 387–395.

6. R.L. Cann, M. Stoneking, and A.C. Wilson, Mitochondrial DNA and human evolution, Nature **325** (1987) 31–36.

7. H. Colonius and H.-H. Schulze, Tree structures for proximity data, British J. Math. Statist. Psych. **34** (1981) 167–180.

8. H. Colonius and H.-H. Schulze, Repräsentation nichtnumerischer Ähnlichkeitsdaten durch Baumstrukturen, Psych. Beitr. **21** (1979) 98–111.

9. P.L. Erdős, L.A. Székely, M. Steel, and T. Warnow, A few logs suffice to build (almost) all trees, Random Structures Algorithms **14** (1999) 153–184.

10. D. Huson, S. Nettles, L. Parida, T. Warnow, and S. Yooseph, The disk-covering method for tree reconstruction, In: Proceedings of "Algorithms and Experiments" (ALEX98), Trento, Italy, Feb. 9–11, 1998, R. Battiti and A. Bertossi, Eds., 1998, pp. 62–75. Available from WWW site `http://rtm.science.unitn.it/alex98/proceedings.html`.

11. D. Huson and T.J. Warnow, Obtaining highly accurate topology and evolutionary estimates of evolutionary trees from very short sequences, accepted for RECOMB 99.

12. P.M. Robinson, Computer-assisted stemmatic analysis and 'best-text' historical editing, In: Studies in Stemmatology, P. Reenen and M. van Mulken, Eds., John Benjamins Publishing, Amsterdam, 1996, pp. 71–103.

13. E. Schröder, Vier Kombinatorische Probleme., Z. Math. Phys. **15** (1870) 361–376.

14. M. Steel, The complexity of reconstructing trees from qualitative characters and subtrees, J. Classification **9** (1992) 91–116.

15. M. Steel, A.W. Dress, and S. Böcker, Some simple but fundamental limitations on supertree and consensus tree methods, 1999, submitted.

16. K. Strimmer and A. von Haeseler, Quartet puzzling: A quartet maximum likelihood method for reconstructing tree topologies, Mol. Biol. Evol. **13** (1996) 964–969.

17. D.L. Swofford, G.J. Olsen, P.J. Waddell, and D.M. Hillis, Phylogenetic inference, Molecular Systematics, D. Hillis, C. Moritz, and B. Marble, Eds., Sinauer Associates, second ed., 1996, pp. 407–514.

18. T. Warnow, Mathematical approaches to comparative linguistics, Proc. Nat. Acad. Sci. U.S.A. **94** (1997) 6585–6590.

19. S.J. Willson, Measuring inconsistency in phylogenetic trees, J. Theor. Biol. **190** (1998) 15–36.