ALGORITHMS FOR MOLECULAR BIOLOGY

RESEARCH Open Access

Extracting conflict-free information from multi-labeled trees

Akshay Deepak[1*], David Fernández-Baca[1] and Michelle M McMahon[2]

Abstract

Background: A multi-labeled tree, or MUL-tree, is a phylogenetic tree where two or more leaves share a label, e.g., a species name. A MUL-tree can imply multiple conflicting phylogenetic relationships for the same set of taxa, but can also contain conflict-free information that is of interest and yet is not obvious.

Results: We define the information content of a MUL-tree T as the set of all conflict-free quartet topologies implied by T, and define the maximal reduced form of T as the smallest tree that can be obtained from T by pruning leaves and contracting edges while retaining the same information content. We show that any two MUL-trees with the same information content exhibit the same reduced form. This introduces an equivalence relation among MUL-trees with potential applications to comparing MUL-trees. We present an efficient algorithm to reduce a MUL-tree to its maximally reduced form and evaluate its performance on empirical datasets in terms of both quality of the reduced tree and the degree of data reduction achieved.

Conclusions: Our measure of conflict-free information content based on quartets is simple and topologically appealing. In the experiments, the maximally reduced form is often much smaller than the original tree, yet retains most of the taxa. The reduction algorithm is quadratic in the number of leaves and its complexity is unaffected by the multiplicity of leaf labels or the degree of the nodes.

Keywords: Phylogenetic trees, Evolutionary trees, Multi-labeled trees, Reduction, Singly-labeled trees

Background

Multi-labeled trees, also known as MUL-trees, are phylogenetic trees that can have more than one leaf with the same label [1-5] (Figure 1). MUL-trees arise naturally and frequently in data sets containing multiple gene sequences for the same species [6], but they can also arise in biogeographical studies or co-speciation studies where leaves represent individual taxa yet are labeled with their areas [7] or hosts [8].

MUL-trees, unlike singly-labeled trees, can contain conflicting species-level phylogenetic information due to biological processes such as whole genome duplications [9] or incomplete lineage sorting [10], to artifactual processes such as inferential error, or, frequently, an unknown combination of several factors. However, they can also contain substantial amounts of conflict-free information.

Here we provide a way to extract this information; specifically, we have the following results.

- We introduce a new quartet-based measure of the information content of a MUL-tree, defined as the set of conflict-free quartets that the tree displays (see **MUL-Trees and information content** on page 3).
- We introduce the concept of the maximally-reduced form (MRF) of a MUL-tree T, the smallest tree with the same information content as T (see **Maximally reduced MUL-Trees** on page 4), and show that any two MUL-trees with the same information content have the same MRF (Theorem 3).
- We present a simple algorithm to construct the MRF of a MUL-tree (see **The reduction algorithm** on page 7). Its running time is quadratic in the number of leaves and does not depend on the multiplicity of the leaf labels or the degrees of the internal nodes.
- We present computational experience with an implementation of our MRF algorithm (see **Results and discussion** on page 8). In our test data, the MRF

*Correspondence: akshayd@iastate.edu
[1]Department of Computer Science, Iowa State University, Ames, Iowa, USA
Full list of author information is available at the end of the article

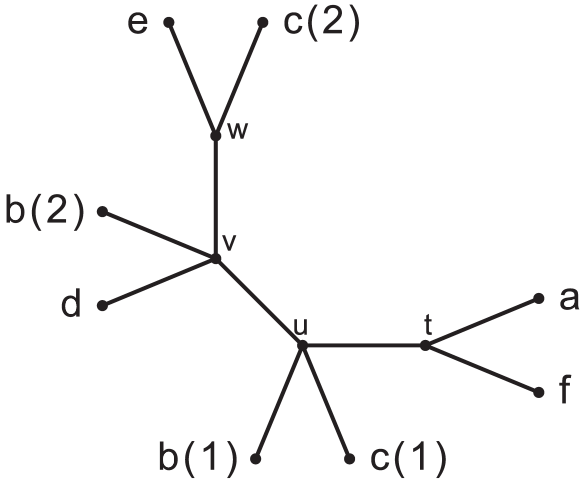

Figure 1 A MUL-tree. Numbers in parenthesis next to labels indicate the multiplicity of the respective labels and are not part of the labels themselves.

is often significantly smaller than the original tree, while retaining most of the taxa.

We now give the intuition behind our notion of information content, deferring the formal definitions of this and other concepts to the next section. Quartets (i.e., sets of four species) are a natural starting point, since they are the smallest subsets from which we can draw meaningful topological information. A singly-labeled tree implies exactly one topology on any quartet. More precisely, each edge e in a singly-labeled tree implies a bipartition (A, B) of the leaf set, where each part is the set of leaves on one of the two sides of e. From (A, B), we derive a collection of bipartitions $ab|cd$ of quartets, such that $\{a, b\} \subseteq A$ and $\{c, d\} \subseteq B$. Clearly, if one edge in a singly-labeled tree implies some bipartition $q = ab|cd$ of $\{a, b, c, d\}$, then there can be no other edge that implies a bipartition, such as $ac|bd$, that is in conflict with q. Indeed, the quartet topologies implied by a singly-labeled tree uniquely identify it [11].

The situation for MUL-trees is more complicated, as illustrated in Figure 1. Here, the presence of two copies of labels b and c — $b(1)$ and $b(2)$, and, $c(1)$ and $c(2)$ — leads to two conflicting topologies on the quartet $\{b, c, d, e\}$. Edge (u, v) implies the bipartition $bc|de$, corresponding to the labels $\{b(1), c(1), d, e\}$, while edge (v, w) implies $bd|ce$ corresponding to the leaves $\{b(2), c(2), d, e\}$. On the other hand, the quartet topology $af|bc$, implied by edge (t, u), has no conflict with any other topology that the tree exhibits on $\{a, b, c, f\}$. We show that the set of all such conflict-free quartet topologies is compatible (Theorem 1). That is, for every MUL-tree T there exists at least one singly-labeled tree that displays all the conflict-free quartets of T — and possibly some other quartets as well. Motivated by this, we only view conflict-free quartet topologies as informative, and define the information content of a MUL-tree as the set of all conflict-free quartet topologies it implies.

We should note that conflicting quartets may well provide valuable information, whether about paralogy, deep coalescence, or mistaken annotations. In some cases, species-level phylogenetic information can be recovered from conflicted quartets through application of, e.g., gene-tree species-tree reconciliation (generally an NP-hard problem [12]). However, this is not feasible when the underlying cause of multiplicity is unknown or when conducting large-scale analyses. Our definition of information content is deliberately designed to make no assumptions about the cause of conflict. It is also conservative with respect to species relationships, i.e., it does not introduce quartets not originally supported by the data. Further, knowing the information content of a MUL-tree allows us to easily identify its conflicting quartets as well.

A MUL-tree may have leaves that can be pruned and edges that can be contracted without altering the tree's information content, i.e., without adding or removing conflict-free quartets. For example, in Figure 1, every quartet topology that edge (v, w) implies is either in conflict with some other topology (e.g., for set $\{b, c, d, e\}$) or is already implied by some other edge (e.g., $af|ce$ is also implied by (t, u)). Thus, (v, w) can be contracted without altering the information content. In fact, the information content remains unchanged if we also contract (u, v) and remove the leaves labeled $b(1)$ and $c(1)$. We define the MRF of a MUL-tree T as the tree that results from applying information-preserving edge contraction and leaf pruning operations repeatedly to T, until it is no longer possible to do so. The MRF of the tree in Figure 1 is shown in Figure 2. In this case, the MRF is singly-labeled; however, this is not true in general (see **An example** on page 8). If the MRF is itself a MUL-tree, it is not possible to reduce the original to a singly-labeled tree without either

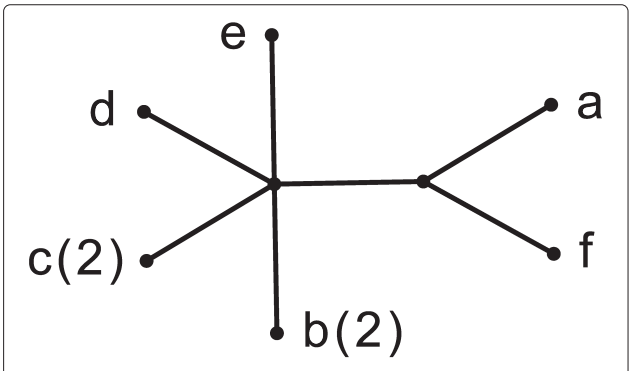

Figure 2 The MRF for the Mul-tree in Figure 1.

adding at least one quartet that did not exist conflict-free in T or by losing one or more conflict-free quartets.

Since any two MUL-trees with the same information content have the same MRF, rather than comparing MUL-trees directly, we can instead compare their MRFs. This is appealing mathematically, because it focuses on conflict-free information content, and also computationally, since an MRF can be much smaller than the original MUL-tree. Indeed, on our test data, the MRF was frequently singly-labeled. This reduction in input size is especially significant if the MUL-tree is an input to an algorithm whose running time is exponential in the label multiplicity, such as Ganapathy et al.'s algorithm to compute the contract-and-refine distance between two area cladograms [7] or Huber et al.'s algorithm to determine if a collection of "multi-splits" can be displayed by a MUL-tree [13].

For our experiments, we also implemented a post-processing step, which converts the MRF to a singly-labeled tree, rendering it available for analyses that require singly-labeled trees, including supermatrix [14,15] and supertree methods [16-19]. On the trees in our data set, the combined taxon loss between the MRF computation and the postprocessing was much lower than it would have been had we simply removed all duplicate taxa from the original trees.

Previous work on MUL-trees has concentrated on finding ways to reduce MUL-trees to singly-labeled trees (typically in order to provide inputs to supertree methods) [5], and to develop metrics and algorithms to compare MUL-trees [7,20-22]. In contrast to our approach — which is purely topology-based and is agnostic with respect to the cause of label multiplicity — the assumption underlying much of the literature on MUL-trees is that taxon multiplicity results from gene duplication. Thus, methods to obtain singly-labeled trees from MUL-trees usually work by pruning subtrees at putative duplication nodes. Although the proposed algorithms are polynomial, they are unsatisfactory in various ways. For example, in [5] if the subtrees are neither identical nor compatible, then the subtree with smaller information content is pruned, which seems to discard too much information. Further, the algorithm is only efficient for binary rooted trees. In [20] subtrees are pruned arbitrarily, while in [21] at each putative duplication node a separate analysis is done for each possible pruned subtree. Although the latter approach is better than pruning arbitrarily, in the worst case it can end up analyzing exponentially many subtrees.

MUL-Trees and information content

A *MUL-tree* is a triple (T, M, ψ), where (i) T is an unrooted tree[a] with leaf set $\mathcal{L}(T)$ all of whose internal nodes have degree at least three, (ii) M is a set of labels, and (iii) $\psi_T : \mathcal{L}(T) \to M$ is a surjective map that assigns each leaf of T a label from M. (Note that if ψ is a bijection, T is singly labeled; that is, singly-labeled trees are a special case of MUL-trees.) For brevity we often refer to a MUL-tree by its underlying tree T. In what follows, unless stated otherwise, by a tree we mean a MUL-tree.

An edge (u, v) in T is *internal* if neither u nor v belong to $\mathcal{L}(T)$, and is *pendant* otherwise. A *pendant node* is an internal node that has a leaf as its neighbor.

Let (u, v) be an edge in T and T' be the result of deleting (u, v) from T. Then T_u^{uv} (T_v^{uv}) denotes the subtree of T' that contains u (v). M_u^{uv} (M_v^{uv}) denotes the set of labels in T_u^{uv} (T_v^{uv}) but not in T_v^{uv} (T_u^{uv}). C^{uv} is the set of labels common to both T_u^{uv} and T_v^{uv}. Observe that M_u^{uv}, M_v^{uv} and C^{uv} partition M. For example, in Figure 1, $M_u^{uv} = \{a, f\}$, $M_v^{uv} = \{e, d\}$, $C^{uv} = \{b, c\}$.

A (resolved) *quartet* in a MUL-tree T is a bipartition $ab|cd$ of a set of labels $\{a, b, c, d\}$ such that there is an edge (u, v) in T with $\{a, b\} \in M_u^{uv}$ and $\{c, d\} \in M_v^{uv}$. We say that (u, v) *resolves* $ab|cd$. For example, in Figure 1, edge (t, u) resolves $af|bc$.

The *information content of an edge* (u, v) of a MUL-tree T, denoted $\Delta(u, v)$, is the set of quartets resolved by (u, v). An edge (u, v) in tree T is *informative* if $|\Delta(u, v)| > 0$; (u, v) is *maximally informative* if there is no other edge (u', v') in T with $\Delta(u, v) \subset \Delta(u', v')$. The *information content* of T, denoted $\mathcal{I}(T)$, is the combined information content of all edges in the tree; that is $\mathcal{I}(T) = \bigcup_{(u,v)\in E} \Delta(u, v)$, where E denotes the set of edges in T.

The next result shows that the quartets in $\mathcal{I}(T)$ are conflict-free.

Theorem 1. *For every MUL-tree T, there is a singly labeled tree T' such that $\mathcal{I}(T) \subseteq \mathcal{I}(T')$.*

Proof. Repeat the following step until T has no multiply-occurring labels. Pick any multiply-occurring label ℓ in T, select an arbitrary leaf labeled by ℓ, and relabel every other leaf labeled by ℓ, by a new, unique, label. The resulting tree T' is singly labeled, and all labels of T are also present in T'. Consider a quartet $ab|cd$ in T, that is resolved by edge (u, v). Assume that $\{a, b\} \in M_u^{uv}$ and $\{c, d\} \in M_v^{uv}$. Thus, T_u^{uv} contains all the occurrences of label a. Clearly, this also holds for the only occurrence of a in T'. Similar statements can be made about labels b, c, and d. Thus, the quartet $ab|cd$ is resolved by edge (u, v) in T', and, hence, T' displays all quartets of T. □

Note that there are examples where the containment indicated by the above result is proper.

To conclude this section, we give some results that are useful for the MUL-tree reduction algorithm (see **The reduction algorithm**, beginning on page 7). In the next lemmas, (u, v) and (w, x) denote two edges in tree T that lie on the path $P_{u,x} = (u, v, \ldots, w, x)$ as shown in Figure 3.

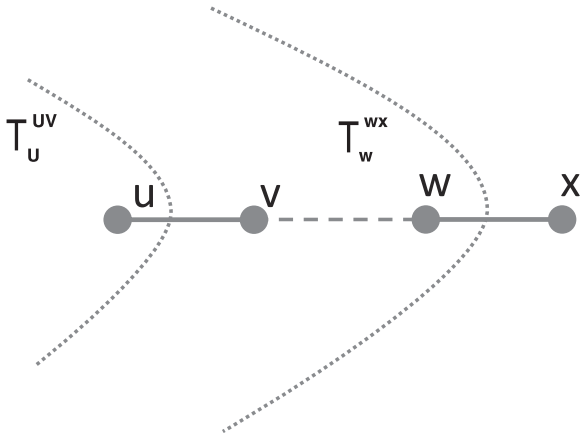

Figure 3 Supportive illustration for the proof of Lemma 1.

Lemma 1. *If* $|M_u^{uv}| = |M_w^{wx}|$ *then* $M_u^{uv} = M_w^{wx}$. *Otherwise,* $M_u^{uv} \subset M_w^{wx}$.

Proof. Refer to Figure 3. Since T_u^{uv} is a subtree of T_w^{wx}, $M_u^{uv} \subseteq M_w^{wx}$ by definition of M_u^{uv}. Thus, if $|M_u^{uv}| = |M_w^{wx}|$, we must have $M_w^{wx} = M_u^{uv}$ and, if $|M_u^{uv}| \neq |M_w^{wx}|$, we must have $M_u^{uv} \subset M_w^{wx}$. □

Together with Lemma 1, the next result allows us to check whether the information content of an edge is a subset of that of another based solely on the cardinalities of the M_u^{uv}s.

Lemma 2. $\Delta(u,v) \subseteq \Delta(w,x)$ *if and only if* $M_v^{uv} = M_x^{wx}$.

Proof. (Only if) Suppose $\Delta(u,v) \subseteq \Delta(w,x)$; therefore, $M_v^{uv} \subseteq M_x^{wx}$. By definition, $M_v^{uv} \supseteq M_x^{wx}$; hence, $M_v^{uv} = M_x^{wx}$.
(If) Suppose $M_v^{uv} = M_x^{wx}$. By definition, $M_u^{uv} \subseteq M_w^{wx}$, which implies that $\Delta(u,v) \subseteq \Delta(w,x)$. □

Lemma 3. *Suppose* $\Delta(u,v) \subseteq \Delta(w,x)$. *Then, for any edge* (y,z) *on* $P_{u,x}$ *such that* v *is closer to* y *than to* z, $\Delta(u,v) \subseteq \Delta(y,z) \subseteq \Delta(w,x)$.

Proof. By Lemma 2, since $\Delta(u,v) \subseteq \Delta(w,x)$, we have $M_v^{uv} = M_x^{wx}$. Now consider an edge (y,z) on $P_{u,x}$. By definition $M_v^{uv} \supseteq M_z^{yz} \supseteq M_x^{wx}$. But $M_v^{uv} = M_x^{wx}$, therefore $M_v^{uv} = M_z^{yz} = M_x^{wx}$. By definition $M_u^{uv} \subseteq M_y^{yz} \subseteq M_w^{wx}$. Hence, by Lemma 2, $\Delta(u,v) \subseteq \Delta(y,z) \subseteq \Delta(w,x)$. □

Maximally reduced MUL-Trees

Our goal is to provide a way to reduce a MUL-tree T as much as possible, while preserving its information content. Our reduction algorithm uses the following operations.

Prune(v)*:* Delete leaf v from T. If, as a result, v's neighbor u becomes a degree-two node, connect the former two neighbors of u by an edge and delete u.
Contract(e)*:* Delete an internal edge e and identify its endpoints.

A leaf v in T is *prunable* if the tree that results from pruning v has the same information content as T. An internal edge e in T is *contractible* if the tree that results from contracting e has the same information content as T. T is *maximally reduced* if it has no prunable leaf and no contractible internal edge.

Theorem 2. *Every internal edge in a maximally reduced tree* T *resolves a quartet that is resolved by no other edge.*

Proof. We rely on two facts. First, every internal node in the tree has degree at least three. Second, every internal edge in the tree resolves a quartet; otherwise, the edge would be contractible and the tree would not be maximally reduced.

Consider any edge (u,v) in the tree. To prove that (u,v) resolves a quartet not resolved by any other edge, we need to show that there exists a quartet $ab|cd$ of the form shown in Figure 4. First, we describe how to select leaves a and b. Consider the following cases:

1. u has at least two neighbors i and j, apart from v, that are internal nodes. Then, we select any $a \in M_i^{ui}$ and any $b \in M_j^{uj}$.
2. u has only one neighbor $i \neq v$ that is an internal node. Then, at least one of u's neighboring leaves must participate in a quartet that (u,v) resolves. Without such a leaf, (u,v) would resolve the same set of quartets as (u,i), so one of these two edges would

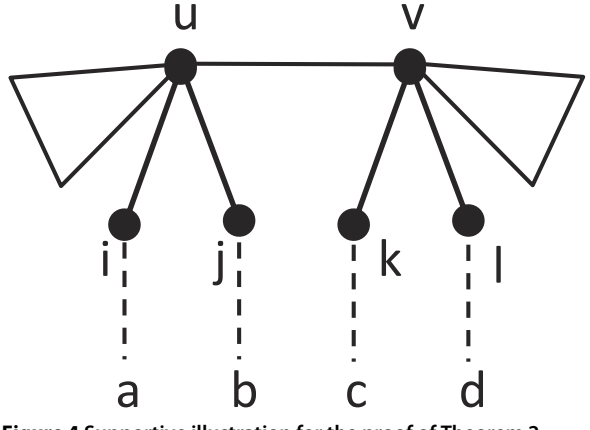

Figure 4 Supportive illustration for the proof of Theorem 2. Quartet $ab|cd$ is resolved only by edge (u,v). Here, $a \in M_i^{ui}$, $b \in M_j^{uj}$, $c \in M_k^{vk}$ and $d \in M_l^{vl}$.

be contractible, contradicting the assumption that the tree is maximally reduced. We select this leaf as b and we select any $a \in M_i^{ui}$.

3. All neighbors of u, except v, are leaves. Then, at least two of its neighbors must participate in a quartet, because (u, v) must resolve a quartet. We select the two neighbors as a and b.

In every case, we can select the desired leaves a and b. By a similar argument, we can also select the desired c and d. This proves the existence of the desired quartet $ab|cd$. Therefore, each internal edge of T uniquely resolve a quartet. □

The next result shows that the set of quartets resolved by a maximally reduced tree uniquely identifies the tree.

Theorem 3. *Let T and T' be two maximally reduced trees such that $\mathcal{I}(T) = \mathcal{I}(T')$. Then, T and T' are isomorphic.*

The *maximally reduced form* (MRF) of a MUL-tree T is the tree that results from repeatedly pruning prunable leaves and contracting contractible edges from T until this is no longer possible. Theorem 3 shows that we can indeed talk about "the" MRF of T. Before proving Theorem 3, we mention some of its consequences.

Corollary 1. *Every MUL-tree has a unique MRF.*

Corollary 2. *Any two MUL-trees with the same information content have the same MRF.*

Corollary 3. *If a maximally reduced MUL-tree T is not singly-labeled, there does not exist a singly-labeled tree T' such that $\mathcal{I}(T) = \mathcal{I}(T')$.*

Note that Corollary 3 does not contradict Theorem 1. If the MUL-tree in Theorem 1 is maximally reduced and not singly-labeled, the containment is proper; i.e., $\mathcal{I}(T) \neq \mathcal{I}(T')$, which is the claim of Corollary 3. Figure 5 illustrates this. Any singly-labeled tree resolving the same set of quartets must be obtained by removing one of the leaves labeled with *f*. However, doing so will also introduce quartets that are not resolved by the maximally reduced MUL-tree.

Corollary 4. *The relation "sharing a common MRF" is an equivalence relation on the set of MUL-trees.*

The last result implies that MUL-trees can be partitioned into equivalence classes, where each class consists of the set of all trees with the same information content.

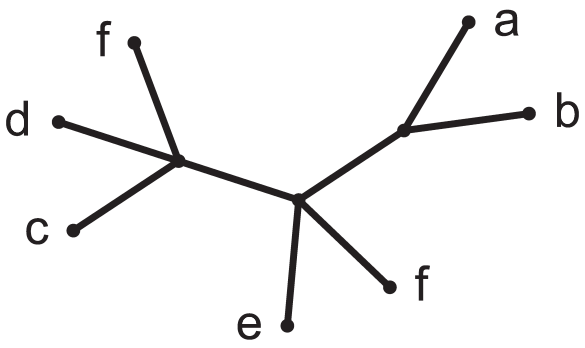

Figure 5 A maximally reduced MUL-tree.

Thus, instead of comparing MUL-trees directly, we can compare their maximally reduced forms.

We now proceed to the proof of Theorem 3. We need two lemmas.

Lemma 4. *There is a bijection ϕ between the respective sets of internal edges of T and T' with the following property. Let (u, v) be an internal edge in T and let $(u', v') = \phi(u, v)$. Then, $M_u^{uv} = M_{u'}^{u'v'}$ and $M_v^{uv} = M_{v'}^{u'v'}$. Therefore, $\Delta(u, v) = \Delta(u', v')$.*

Proof. Consider an edge (u, v) in T. By Theorem 2, (u, v) must resolve a quartet $ab|cd$ not resolved by any other edge as shown in Figure 4. We claim that this quartet must be resolved uniquely by an edge (u', v') in T'. Suppose not. Using arguments similar to those in the proof of Lemma 3, we can show that all edges that resolve $ab|cd$ in T' form a path $(u', x', \ldots, w', v')$, where possibly $x' = w'$, as shown in Figure 6. Here, $\{a, b\} \subseteq M_{u'}^{u'x'}$ and $\{c, d\} \subseteq M_{v'}^{w'v'}$.

Since (w', v') resolves a quartet not resolved by any other edge, by Theorem 2 there exists a label ℓ as shown, where $\ell \in M_m^{w'm}$. Since $ab|\ell d$ is a quartet in T' and $\mathcal{I}(T) = \mathcal{I}(T)$, it must be true that $\ell \in M_v^{uv}$ in T. Clearly, T does not resolve the quartet on $\{a, \ell, d, c\}$ in the same way, $a\ell|cd$, as

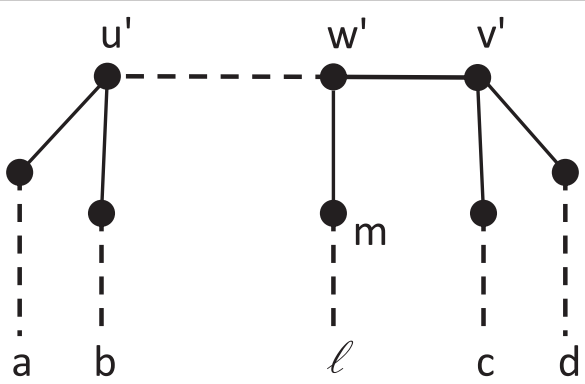

Figure 6 Supportive illustration for the proof of Lemma 4.

T'. This contradicts the assumption that $\mathcal{I}(T) = \mathcal{I}(T')$. Thus, (u', v') must be an edge. Moreover, only one such edge exists in T' as it uniquely resolves the quartet $ab|cd$.

Now consider any label $f \in M_u^{uv}$ such that $f \notin \{a, b, c, d\}$. Label f must be in $M_{u'}^{u'v'}$; otherwise, T and T' would resolve the quartet $\{a, f, c, d\}$ differently. Similarly, any such $f \in M_{u'}^{u'v'}$ must be in M_u^{uv} as well. Thus $M_u^{uv} = M_{u'}^{u'v'}$. In the same way, we can prove that $M_v^{uv} = M_{v'}^{u'v'}$. Thus, $\Delta(u, v) = \Delta(u', v')$.

We have shown that there is a one-to-one mapping ϕ from edges in T to edges of T' such that $\Delta(e) = \Delta(\phi(e))$. To complete the proof, we show that ϕ is onto. Suppose that for some edge e' in T' there is no edge e in T such that $\phi(e) = e'$. But then e' must resolve a quartet not resolved by any other edge in T'. This quartet cannot be in $\mathcal{I}(T)$, contradicting the assumption that $\mathcal{I}(T) = \mathcal{I}(T')$. □

Let ϕ be the bijection between the edge sets of T and T' from the preceding lemma.

Lemma 5. *Let (u, v) and (v, x) be any two neighboring internal edges in T, and let $(p, q) = \phi(u, v)$ and $(r, s) = \phi(v, x)$ be the corresponding edges in T' such that $M_u^{uv} = M_p^{pq}$ and $M_v^{vx} = M_r^{rs}$. Then, (p, q) and (r, s) are neighbors in T' with $q = r$.*

Proof. Since (u, v) and (v, x) are neighbors, and each resolves a quartet that is not resolved by the other, $M_u^{uv} \subset M_v^{vx}$ and $M_v^{uv} \supset M_x^{vx}$. By Lemma 4, this implies that $M_p^{pq} \subset M_r^{rs}$ and $M_q^{pq} \supset M_s^{rs}$. Thus, the only way (p, q) and (r, s) can exist in T' is as part of the path $P_{p,s} = (p, q, \ldots, r, s)$. If $q \neq r$, then consider the edge (t, r) on P_{ps} such that p is closer to t than to r. Then, the following must hold:

$$M_p^{pq} \subset M_t^{tr} \subset M_r^{rs} \tag{1}$$

and

$$M_q^{pq} \supset M_r^{tr} \supset M_s^{rs} \tag{2}$$

Let $(z, w) = \phi^{-1}(t, r)$ be the edge in T corresponding to (t, r). Irrespective of the position of (z, w) in T, (1) and (2) cannot be simultaneously true with respect to edges (u, v), (v, x) and (z, w) in T. Therefore, $q = r$, which proves the desired result. □

Proof of Theorem 3. Lemmas 4 and 5 show that T and T' are isomorphic with respect to their internal edges. It remains to show a one-to-one correspondence between their leaf sets. For this, we match up the leaves attached at every pendant node in T and T'. We start with pendant nodes to which only one internal edge is attached. For example, consider an internal edge (u, v) in T such that v is a pendant node and T_v^{uv} has only leaves. Let $(u', v') = \phi(u, v)$ be the corresponding edge in T' such that $M_u^{uv} = M_{u'}^{uv}$. By Lemma 4, $C^{uv} = C^{u'v'}$. Moreover, neither T nor T' have prunable leaves. Thus, the same set of leaves must be attached at v and v' respectively. In subsequent steps, we select an internal edge (u, v) in T such that v is a pendant node and all the other pendant nodes in T_v^{uv} have already been matched up in previous iterations. Again, let $(u', v') = \phi(u, v)$ such that $M_u^{uv} = M_{u'}^{uv}$. Using similar arguments, the same set of leaves must be attached at v and v' respectively. Proceeding this way, each pendant node in T can be paired with the corresponding pendant node in T', and be shown to have the same set of leaves attached to them. This shows that T and T' are isomorphic, as claimed. □

Identifying contractible edges and prunable leaves

In preparation for the MUL-tree reduction algorithm of the next section, we give some results that help to identify contractible edges and prunable leaves.

The setting for the next result is the same as for Lemmas 2 and 3: (u, v) and (w, x) are two edges in tree T that lie on the path $P_{u,x} = (u, v, \ldots, w, x)$ (see Figure 3). We say that subtree T_z^{yz} *branches out* from the path $P_{u,x}$ if $y \in P_{u,x} - \{u, x\}$, and $z \notin P_{u,x}$.

Lemma 6. *Suppose $\Delta(u, v) \subseteq \Delta(w, x)$ then*

1. *every internal edge on a subtree branching out from $P_{u,x}$ is contractible, and*
2. *if $\Delta(u, v) = \Delta(w, x)$, every leaf on a subtree branching out from $P_{u,x}$ is prunable. Thus, the entire subtree can be deleted without changing the information content of the tree.*

Proof. Refer to Figure 7.

1. Consider any edge (a, b) in a subtree branching out of $P_{u,x}$, as shown. We claim that $M_a^{ab} \cup C^{ab} = M$; i.e., all the labels in M appear in T_a^{ab}. This means that $M_b^{ab} = \emptyset$, so (a, b) is uninformative.
   To prove the claim, observe first that, by definition, $M_u^{uv} \cup C^{uv} \cup M_v^{uv} = M$. By Lemma 2, since $\Delta(u, v) \subseteq \Delta(w, x)$, we have $M_x^{wx} = M_v^{uv}$, so
   $$M_u^{uv} \cup C^{uv} \cup M_x^{wx} = M. \tag{3}$$
   Now, $M_u^{uv} \cup C^{uv}$ is the set of labels on the leaves of T_u^{uv}, while every label in M_x^{wx} appears in T_x^{wx}. Hence, T_u^{uv} and T_x^{wx} jointly contain every label in M. Since T_u^{uv} and T_x^{wx} are subtrees of T_a^{ab}, this completes the proof of the claim.
2. Suppose $\Delta(u, v) = \Delta(w, x)$. By an argument similar to the one used in the proof of Lemma 3, we can show that any edge (y, z) on the path $P_{v,w} = (v \ldots w)$ (see Figure 7) satisfies $M_v^{uv} = M_z^{yz} = M_x^{wx}$ and $M_u^{uv} = M_y^{yz} = M_w^{wx}$. Consider a leaf c as shown; let ℓ

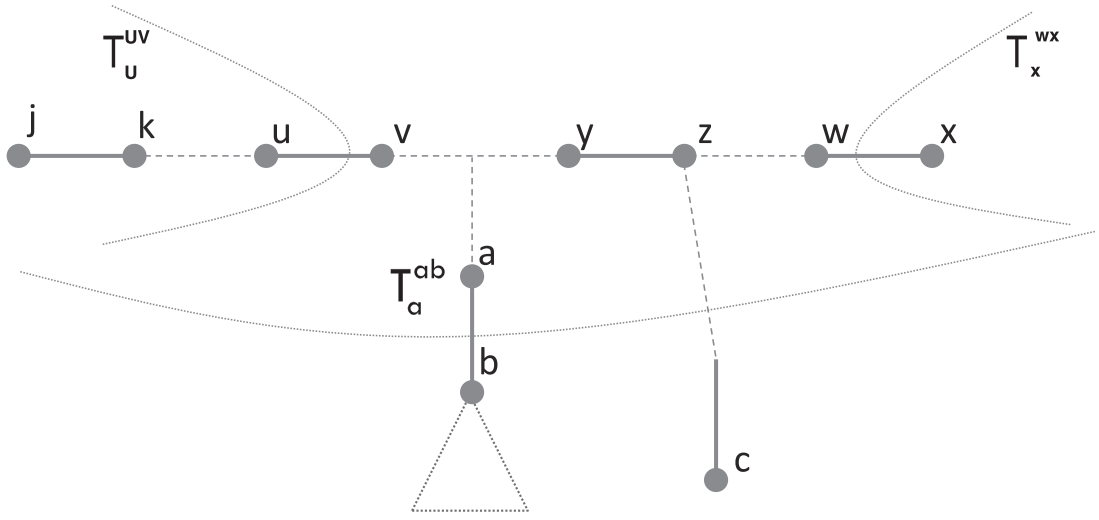

Figure 7 Supportive illustration for the proof of Lemma 6.

be its label. Then, ℓ appears in T_x^{wx}, for else $M_y^{yz} \neq M_w^{wx}$, a contradiction. Similarly, ℓ appears in T_u^{uv}.

Now, let S be the tree obtained after pruning leaf c.

(a) $\mathcal{I}(T) \subseteq \mathcal{I}(S)$: Suppose pruning c removes a quartet from $\mathcal{I}(T)$. If such a quartet exists in T, it must be resolved by an edge $(j,k) \in T_u^{uv}$ (say). But then (j,k) still resolves the same quartet in S because $\ell \in M_x^{wx}$, and the labels in T_x^{wx} are a subset of those in T_k^{jk}. This is a contradiction.

(b) $\mathcal{I}(S) \subseteq \mathcal{I}(T)$: Suppose pruning c adds a quartet to $\mathcal{I}(S)$ that is not in $\mathcal{I}(T)$. Such a quartet in S must be resolved by an edge (j,k) in S_u^{uv} (say), that before pruning satisfied $\ell \in C^{jk}$, but now has $\ell \notin M_k^{jk}$. However $\ell \in M_x^{wx}$; therefore we still have $\ell \in C^{jk}$ and the edge still cannot resolve the quartet, a contradiction.

Hence, c is prunable. □

Lemma 7. *Suppose that T is a MUL-tree where no pendant node is adjacent to two or more leaves with the same label. Let ℓ be any multiply-occurring label in T and let T' be the minimal subtree of T that spans all the leaves labeled by ℓ. Then, any leaf in T labeled ℓ attached to a pendant node of degree at least three in T' is prunable.*

Proof. Refer to Figure 8. Consider any pendant node v of degree at least three in T' attached to a leaf labeled ℓ. Clearly deleting the leaf does not change the information content of any edge in T_u or T_y. Now consider an edge (w,x) in T' as shown. Note that $\ell \in C^{wx}$, so ℓ does not contribute to $\Delta(w,x)$. After deleting the leaf, we still have $\ell \in C^{wx}$, so $\Delta(w,x)$ remains unchanged. Therefore, the leaf is prunable. □

The reduction algorithm

We now describe a $O(n^2)$ algorithm to compute the MRF of an n-leaf MUL-tree T. In the previous section, the MRF was defined as the tree obtained by applying information-preserving pruning and contraction operations to T, in any order, until it is no longer possible. For efficiency,

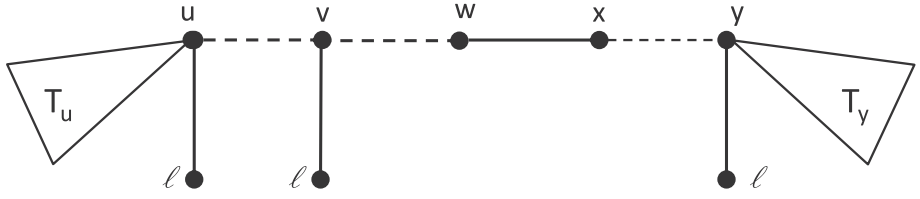

Figure 8 Supportive illustration for the proof of Lemma 7. The leaves attached to pendant nodes u, v, and y are labeled by ℓ, and the subtrees indicated by T_u and T_y do not contain a leaf labeled with ℓ. Nodes u and y have degree two in T', while v has degree three.

however, the sequence in which these steps are performed is important. Our algorithm has three distinct phases: a preprocessing step, redundant edge contraction, and pruning of redundant leaves. We describe these next and then give an example.

Preprocessing

For every edge (u, v) in T, we compute $|M_u^{uv}|$ and $|M_v^{uv}|$. This can be done in $O(n^2)$ time as follows. First, traverse subtrees T_u^{uv} and T_v^{uv} to count number of distinct labels n_u^{uv} and n_v^{uv} in each subtree. Then, $|M_u^{uv}| = |M| - n_v^{uv}$ and $|M_v^{uv}| = |M| - n_u^{uv}$. We then contract non-informative edges; i.e., edges (u, v) where $|M_u^{uv}|$ or $|M_v^{uv}|$ is at most one.

Edge contraction and subtree pruning

Next, we repeatedly find pairs of adjacent edges (u, v) and (v, w) such that $\Delta(u, v) \subseteq \Delta(v, w)$ or vice-versa, and contract the less informative of the two. By Lemmas 1 and 2, we can compare $\Delta(u, v)$ and $\Delta(v, w)$ in constant time using the precomputed values of $|M_u^{uv}|$ and $|M_v^{uv}|$. Lemma 6(1) implies that we should also contract all internal edges incident on v or in the subtrees branching out of v. Further, by Lemma 6(2), if $\Delta(u, v) = \Delta(v, w)$, we can in fact delete these subtrees entirely, since their leaves are prunable. Lemma 3 implies that all such edges must lie on a path, and hence can be identified in linear time. The total time for all these operations is linear, since at worst we traverse every edge twice.

Pruning redundant leaves

The tree that is left at this point has no contractible edges; however, it can still have prunable leaves. We first prune any leaf with a label ℓ that does not participate in any resolved quartet. Such an ℓ has the property that for every edge (u, v), $\ell \notin M_u^{uv}$ and $\ell \notin M_v^{uv}$. All such leaves can be found in $O(n^2)$ time and $O(n)$ space.

Next, we consider sets of leaves with the same label ℓ that share a common neighboring pendant node. Such leaves can be found in linear time. For each such set, we delete all but one element. Let T be the tree that results from removing such leaves. Now, the only prunable leaves with a given label ℓ that might remain are leaves attached to different pendant nodes. By Lemma 7, we can identify and prune such leaves by performing the following steps.

1. For each label ℓ, consider the subgraph on the leaves labeled by ℓ.
2. In this subgraph, delete any leaf not attached to a degree 2 pendant node as it is a redundant leaf.

This takes $O(n)$ time per label and $O(n^2)$ time total. The space used is $O(n)$. Hence, the overall time and space complexities are $O(n^2)$ time and $O(n)$, respectively.

The resulting tree has no contractible edges nor prunable leaves. Therefore, it is the MRF of the orginal MUL-tree.

An example

We illustrate the reduction of the unrooted MUL-tree shown in Figure 9(a) to its MRF.

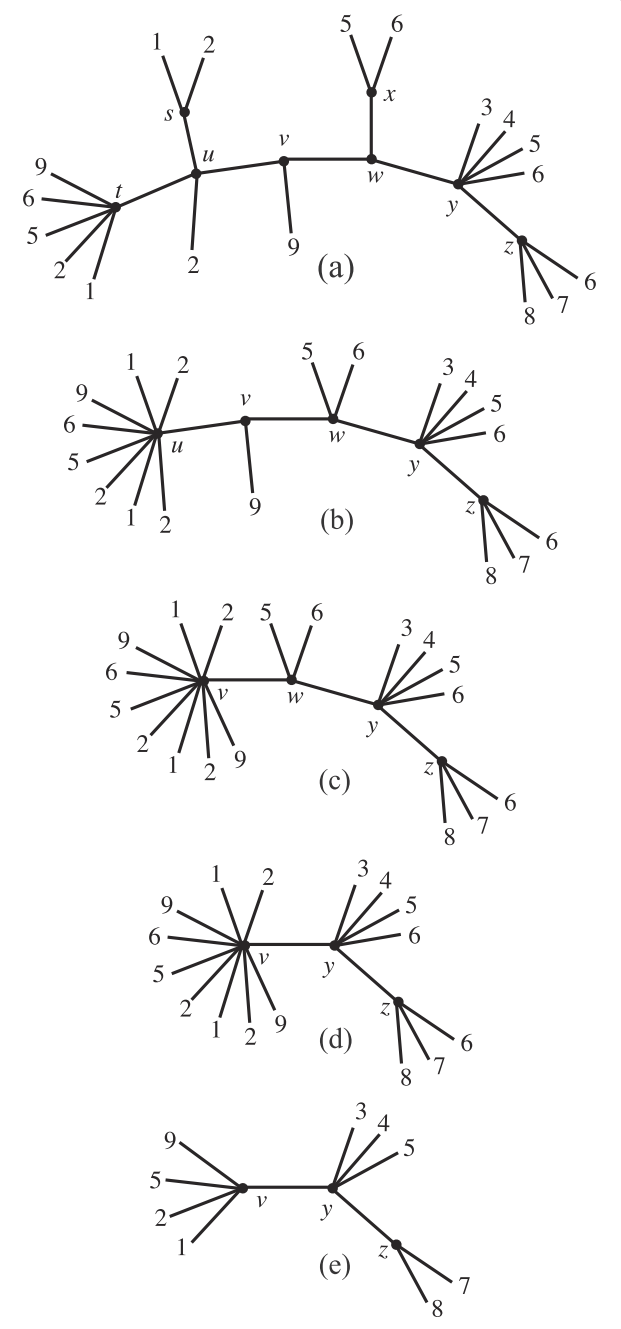

Figure 9 Reduction to the MRF. Supportive illustrations for the example on page 8.

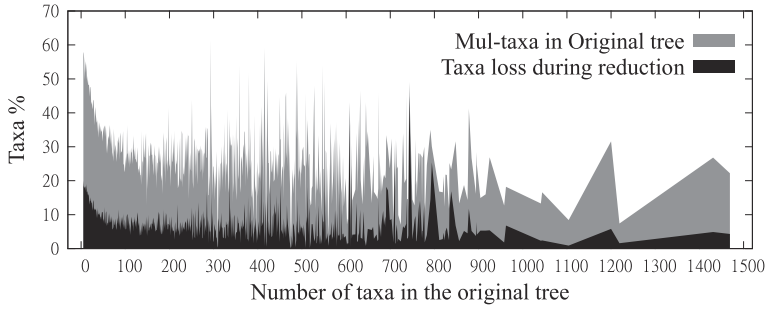

Figure 10 Experimental results: Taxon loss in the second step.

1. In the preprocessing step, we find that $M_t^{tu} = \emptyset$, $M_s^{su} = \emptyset$ and $M_x^{wx} = \emptyset$, so edges (t, u), (s, u) and (w, x) are uninformative. They are therefore contracted, resulting in the tree shown in Figure 9(b).
2. Since $\Delta(u, v) \subset \Delta(v, w)$, contract (u, v). The result is shown in Figure 9(c).
3. Since $\Delta(v, w) = \Delta(w, y)$, delete the subtree branching out at w from the path from v to y and contract (v, w). The result is shown in Figure 9(d).
4. Prune taxon 6, which does not participate in any quartet, and all duplicate taxa at the pendant nodes. The result, shown in Figure 9(e), is the MRF of the original tree.

Results and discussion

We implemented our MUL-tree reduction algorithm, as well as a second step that restricts the MRF to the set of labels that appear only once, which yields a singly-labeled tree. We tested our two-step program on a set of 110,842 MUL-trees obtained from the PhyLoTA database [6] (http://phylota.net/; GenBank eukaryotic nucleotide sequences, release 184, June 2011), which included a broad range of label-set sizes, from 4 to 1500 taxa.

There were 8,741 trees (7.8%) with essentially no information content; these lost all resolution either when reduced to their MRFs, or in the second step. The remaining trees fell into two categories. Trees in set A had a singly-labeled MRF; 65,709 trees (59.3%) were of this kind. Trees in set B were reduced to singly-labeled trees in the second step; 36,392 trees (32.8%) were of this kind. Reducing a tree to its MRF (step 1), led to an average taxon loss of 0.83% of the taxa in the input MUL-tree. The total taxon loss after the second step (reducing the MRFs in set B to singly-labeled trees), averaged 12.81%. This taxon loss is not trivial, but it is far less than the 41.27% average loss from the alternative, naïve, approach in which all MUL-taxa (taxa that label more than one leaf) are removed at the outset. Note that, by the definition of MRFs, taxa removed in the first step do not contribute to the information content, since all non-conflicting quartets are preserved. On the other hand, taxa removed in the second step do alter the information content, because each such taxon participated in some non-conflicting quartet. Information content, in this case, will be lost but new information is never introduced, so the algorithm can be considered conservative.

Taxon loss is sensitive to the number of total taxa and, especially, MUL-taxa, as demonstrated in Figure 10. The grey function shows the percentage of MUL-taxa in the original input trees, which is the taxon loss if we had restricted the input MUL-trees to the set of singly-labeled leaves. The black function shows the percentage of MUL-taxa lost after steps 1 and 2 of our reduction procedure.

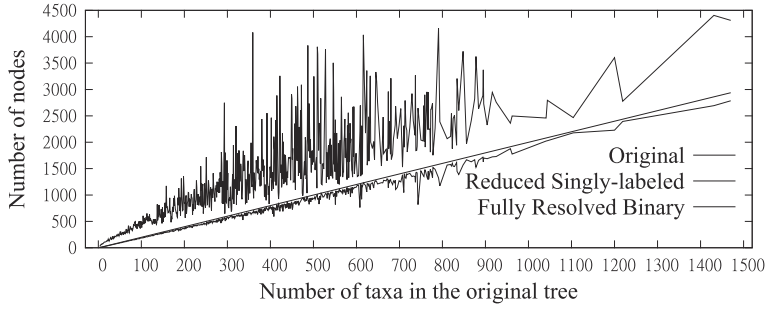

Figure 11 Experimental results: Quality of reduced singly-labeled trees.

In addition to the issue of taxon loss, we investigated the effect of our reduction on edge loss, i.e., the level of resolution within the resulting singly-labeled tree. Input MUL-trees were binary and therefore had more nodes than twice the number of taxa (Figure 11, solid line), whereas a binary tree on singly labeled taxa would have approximately as many nodes as twice the number of taxa (Figure 11, dashed line). We found that, although there was some edge loss, the number of nodes in the reduced singly-labeled trees (Figure 11, dotted line) corresponded well to the total possible, indicating low levels of edge loss. Note that each point on the dotted or solid lines represents an average over all trees with the same number of taxa.

We have integrated our reduction algorithm into STBase (available at http://stbase.org/), a phylogenetic tree search engine that takes a user-provided list of species names and finds matches with a precomputed collection of phylogenetic trees, more than half of which are MUL-trees, assembled from GenBank sequence data. The trees returned are ranked by a tree quality criterion that takes into account overlap with the query set, support values for the branches, and degree of resolution. We have added functionality to provide reduced singly-labeled trees as well as the MUL-trees based on the full leaf set and the label sets from the reduced singly-labeled trees are used in downstream supermatrix construction.

Conclusions

We introduced an efficient algorithm to reduce a multi-labeled MUL-tree to a maximally reduced form with the same information content, defined as the set of non-conflicting quartets it resolves. We showed that the information content of a MUL-tree uniquely identifies the MUL-tree's maximally reduced form. This has potential application in comparing MUL-trees by significantly reducing the number of comparisons as well as in extracting species-level information efficiently and conservatively from large sets of trees, irrespective of the underlying cause of multiple labels. Our algorithm can easily be adapted to work for rooted trees.

Further work investigating the relationship of the MRF to the original tree under various biological circumstances is also underway. We might expect, for example, that well-sampled nuclear gene families reduce to very small MRF trees, and that annotation errors in chloroplast gene sequences (in which we expect little gene duplication), result in relatively large MRF trees. Comparing the MRF to the original MUL-tree may well provide a method for efficiently assessing and segregating data sets with respect to the causes of multiple labels.

It would be interesting to compare our results with some of the other approaches for reducing MUL-trees to singly-labeled trees (e.g., [5]) or, indeed, to evaluate if our method can benefit from being used in conjunction with such approaches.

Endnote

[a]The results presented here can be extended to rooted trees, using triplets instead of quartets, exploiting the well-known bijection between rooted and unrooted trees ([23], p. 20).

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

MMM and DFB conceived the problem. AD, DFB and MMM designed the experiments and drafted the manuscript. AD designed and implemented the algorithms, and implemented the experiments. DFB coordinated the project. All authors read and approved the final manuscript.

Acknowledgements

This work was supported in part by National Science Foundation grant DEB-0829674. We thank Mike Sanderson for helping to motivate this work, for many discussions about the problem formulation, and for our ongoing collaboration in the STBase project. Sylvain Guillemot listened to numerous early versions of our proofs and offered many insightful comments.

Author details

[1]Department of Computer Science, Iowa State University, Ames, Iowa, USA. [2]School of Plant Sciences, University of Arizona, Tucson, Arizona, USA.

Received: 11 December 2012 Accepted: 29 June 2013
Published: 9 July 2013

References

1. Fellows M, Hallett M, Stege U: **Analogs & duals of the MAST problem for sequences & trees.** *J Algorithms* 2003, **49:**192–216. [1998 European Symposium on Algorithms]
2. Grundt H, Popp M, Brochmann C, Oxelman B: **Polyploid origins in a circumpolar complex in Draba (Brassicaceae) inferred from cloned nuclear DNA sequences and fingerprints.** *Mol Phylogenet Evol* 2004, **32**(3):695–710.
3. Huber K, Moulton V: **Phylogenetic networks from multi-labelled trees.** *J Math Biol* 2006, **52:**613–632.
4. Popp M, Oxelman B: **Inferring the history of the Polyploid Silene aegaea (Caryophyllaceae) using Plastid and Homoeologous nuclear DNA sequences.** *Mol Phylogenet Evol* 2001, **20**(3):474–481.
5. Scornavacca C, Berry V, Ranwez V: **Building species trees from larger parts of phylogenomic databases.** *Inf Comput* 2011, **209**(3):590–605. [Special Issue: 3rd International Conference on Language and Automata Theory and Applications (LATA 2009)]
6. Sanderson M, Boss D, Chen D, Cranston K, Wehe A: **The PhyLoTA Browser: processing GenBank for molecular phylogenetics research.** *Syst Biol* 2008, **57**(3):335.
7. Ganapathy G, Goodson B, Jansen R, Le H, Ramachandran V, Warnow T: **Pattern identification in biogeography.** *IEEE/ACM Trans Comput Biol Bioinformatics* 2006, **3:**334–346.
8. Johnson K, Adams R, Page R, Clayton D: **When do parasites fail to speciate in response to host speciation?** *Syst Biol* 2003, **52:**37–47.
9. Lott M, Spillner A, Huber K, Petri A, Oxelman B, Moulton V: **Inferring polyploid phylogenies from multiply-labeled gene trees.** *BMC Evol Biol* 2009, **9:**216.
10. Rasmussen M, Kellis M: **Unified modeling of gene duplication, loss, and coalescence using a locus tree.** *Genome Res* 2012, **22:**755–765.
11. Steel M: **The complexity of reconstructing trees from qualitative characters and subtrees.** *J Classif* 1992, **9:**91–116.
12. Stolzer M, Lai H, Xu M, Sathaye D, Vernot B, Durand D: **Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees.** *Bioinformatics* 2012, **28**(18):i409—i415.

13. Huber K, Lott M, Moulton V, Spillner A: **The complexity of deriving multi-labeled trees from bipartitions.** *J Comput Biol* 2008, **15**(6):639–651.
14. de Queiroz A, Gatesy J: **The supermatrix approach to systematics.** *Trends Ecol Evol* 2007, **22:**34–41.
15. Wiens JJ, Reeder TW: **Combining data sets with different numbers of Taxa for Phylogenetic analysis.** *Syst Biol* 1995, **44**(4):548–558.
16. Baum BR: **Combining trees as a way of combining data sets for Phylogenetic inference, and the desirability of combining gene trees.** *Taxon* 1992, **41:**3–10.
17. Ragan M: **Phylogenetic inference based on matrix representation of trees.** *Mol Phylogenet Evol* 1992, **1:**53–58.
18. Bansal M, Burleigh JG, Eulenstein O, Fernández-Baca D: **Robinson-foulds supertrees.** *Algorithms Mol Biol* 2010, **5:**18.
19. Swenson M, Suri R, Linder C, Warnow T: **SuperFine: fast and accurate supertree estimation.** *Syst Biol* 2012, **61**(2):214–227.
20. Puigbò P, Garcia-Vallvé S, McInerney J: **TOPD/FMTS: a new software to compare phylogenetic trees.** *Bioinformatics* 2007, **23**(12):1556.
21. Marcet-Houben M, Gabaldón T: **TreeKO: a duplication-aware algorithm for the comparison of phylogenetic trees.** *Nucleic Acids Res* 2011, **39:**e66.
22. Huber K, Spillner A, Suchecki R, Moulton V: **Metrics on multilabeled trees: interrelationships and diameter bounds.** *Comput Biol Bioinformatics, IEEE/ACM Trans* 2011, **8**(4):1029–1040.
23. Semple C, Steel M: *Phylogenetics*. Oxford: Oxford University Press; 2003.

doi:10.1186/1748-7188-8-18
Cite this article as: Deepak *et al.*: Extracting conflict-free information from multi-labeled trees. *Algorithms for Molecular Biology* 2013 **8**:18.