Quartets and Unrooted Phylogenetic Networks

Philippe Gambette, Vincent Berry, Christophe Paul

Journal of Bioinformatics and Computational Biology

QUARTETS AND UNROOTED PHYLOGENETIC NETWORKS

PHILIPPE GAMBETTE
Université Paris-Est, LIGM, 5 bd Descartes Champs sur Marne
Champs-sur-Marne 77454, France
philippe.gambette@univ-mlv.fr

VINCENT BERRY
LIRMM, Université Montpellier 2 / CNRS, 161 rue Ada
Montpellier 34392, France
vberry@lirmm.fr

CHRISTOPHE PAUL
LIRMM, Université Montpellier 2 / CNRS, 161 rue Ada
Montpellier 34392, France
paul@lirmm.fr

Received January 24, 2011
Revised January 19, 2012
Accepted (Day Month Year)

Phylogenetic networks were introduced to describe evolution in the presence of exchanges of genetic material between coexisting species or individuals. Split networks in particular were introduced as a special kind of abstract network to visualize conflicts between phylogenetic trees which may correspond to such exchanges. More recently, methods were designed to reconstruct explicit phylogenetic networks (whose vertices can be interpreted as biological events) from triplet data.

In this article, we link abstract and explicit networks through their combinatorial properties, by introducing the unrooted analogue of level-k networks. In particular, we give an equivalence theorem between circular split systems and unrooted level-1 networks. We also show how to adapt to quartets some existing results on triplets, in order to reconstruct unrooted level-k phylogenetic networks. These results give an interesting perspective on the combinatorics of phylogenetic networks and also raise algorithmic and combinatorial questions.

Keywords: phylogenetic networks; quartets; level-k networks; NP-hardness; exact algorithms.

1. Introduction

Phylogeny aims at reconstructing the evolution of a set of *taxa* (species for example), given information on this set of taxa, such as DNA sequences of some representatives. A tree, whose leaves are bijectively associated with the taxa, is often considered as the most appropriate model. This tree is rooted when the ancestor

Electronic version of an article published as
Journal of Bioinformatics and Computational Biology (*JBCB*), Vol. 10, No. 4 (2012) 1250004
DOI:10.1142/S0219720012500047

of the taxa can be located, unrooted otherwise. But sometimes, when exchanges of genetic material between coexisting species is suspected to occur, phylogenetic networks are preferred to trees. There exist two kinds of phylogenetic networks[1]: *abstract* networks are used to visualize evolutionary data but their vertices and edges cannot be interpreted as biological events. On the contrary, *explicit* phylogenetic networks model evolution with their vertices representing ancestral species. Abstract networks are usually faster to compute, but difficult to interpret and to visualize efficiently.

Methods based directly on sequences are usually slow, as sequences are very large, and approaches have been proposed to reconstruct unrooted phylogenetic networks from various input: distances, unrooted trees, quartets (i.e. unrooted phylogenetic trees on four taxa). Most of these methods work by computing a *split system*[2] on the taxa (except T-Rex which reconstructs reticulograms directly from a distance matrix[3]). After filtering some splits[4, 5], it is represented by a split network[6, 7, 8, 9], or by a galled network[10, 11]. These indirect reconstruction approaches, which first compute an abstract representation of the data, and then try to deduce an explicit phylogenetic network of a restricted subclass of phylogenetic networks, have drawbacks. If we choose to visualize the output with a split network, its number of edges may be quadratic in the number of splits, which results in a high dimensional network, or a grid, quite difficult to interpret biologically. If we choose galled networks, the output has topological constraints which may not reflect all the possibilities of biological evolution, if for example an ancestral species which appeared following hybridization events also gives rise to a new species through another hybridization event. Furthermore, the reconstruction of this network may be ambiguous, as a final step to choose among the possible networks which represent the split system may be necessary[12].

In a rooted context on the contrary, a lot of methods were designed to directly output explicit phylogenetic networks, in particular from triplet input, i.e. from rooted phylogenetic trees on three taxa. However, approaches where the root is chosen at the last step are usually preferred, because choosing the position of the root is a difficult task, which could cause important errors if done too early[13]. Unrooted tree data is most often available, for example in databases like Hogenom[14] or PhylomeDB[15].

In this article, we generalize to unrooted networks the *level* parameter proposed to describe the complexity of explicit rooted phylogenetic networks[16]. We focus on unrooted *binary* phylogenetic networks, defined as graphs whose vertices have either degree 3 (*internal vertices*), or degree 1 (the *leaves*, bijectively labeled by a set X of taxa). Note that as the network is unrooted and undirected, it is impossible to decide whether an internal vertex corresponds to a speciation event or a reticulation event. This interpretation task will however be possible after a rooting step (see Fig. 1), which can be done according to various criteria.

Methods were proposed to reconstruct rooted level-k networks from triplets[16, 17, 18, 19, 20, 21, 22]. In this article, we show how to translate some of them

to reconstruct unrooted level-k networks from quartets. Some of the results can be deduced easily with the appropriate changes in definitions for the unrooted context (SN-splits instead of SN-sets, for instance), while others involve very different algorithmic tools and ideas, such as testing whether a quartet is consistent with an unrooted level-k network.

The links we uncover here between abstract and explicit networks also provide a new point of view on combinatorial objects like circular split systems, quartets and unrooted trees. In the same way as the inclusion relationship between clusters consistent with a rooted level-1 network and weak hierarchies[23], the properties we give here may have interesting algorithmic consequences, and help better understand the combinatorics of the unrooted case, which is often more complex than the rooted one[24, 25].

Outline of the article. We first introduce some basic definitions in Section 2, then we detail the relationships between rooted and unrooted level-k networks in Section 3. Section 4 provides results about splits systems consistent with level-1 networks, which have consequences for quartet sets consistent with such networks. In Section 5, we address the complexity of two basic problems involving unrooted level-k networks and quartets. We show that deciding whether a quartet is consistent with an unrooted level-k network is polynomial time solvable. On the contrary, we prove that reconstructing an unrooted level-1 network consistent with an unrestricted quartet set is NP-complete. Therefore, we study a restricted case in Section 6: we show how to obtain a tree decomposition of an unrooted level-k network knowing its complete quartet set. Finally, in Section 7, we focus on the case of unrooted level-1 networks and show that it is possible to reconstruct such networks in polynomial time from their complete quartet set. We also study the dense case, to give puzzling properties and open problems.

2. Basic Definitions

Let us recall from graph theory that an *articulation vertex* is a vertex whose deletion disconnects the graph. A *biconnected component* of a graph $G = (V, E)$ is a maximal induced subgraph of G without articulation vertex. For any $E' \subset E$, we denote by $G - E'$ the graph $G' = (V, E - E')$, i.e. the graph obtained from deleting the edges of E' from G. We say that a *block of a directed graph* is a biconnected component of its underlying undirected graph[26]. Note that throughout this paper, the graphs and directed graphs we consider do not have multiple edges or arcs.

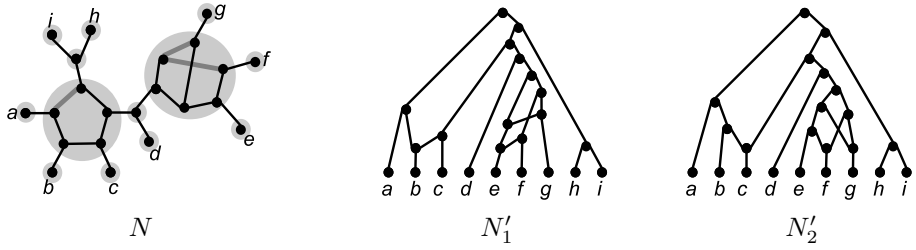
Definition 1. [16] A *rooted level-k network* N is a directed acyclic graph in which exactly one vertex has indegree 0 and outdegree 2 (the root) and all other vertices have either indegree 1 and outdegree 2 (*split vertices*), indegree 2 and outdegree ≤ 1 (*hybrid vertices*), or indegree 1 and outdegree 0 (leaves, distinctly labeled) and such that any block contains at most k hybrid vertices.

In fact, an equivalent definition can be obtained by imposing the same degree

conditions, and adding that in any block, it is possible to get a tree by removing at most k arcs (one arc per hybrid vertex).

In this paper, we extend the latter definition to unrooted phylogenetic networks. A *cut-edge* or *bridge* is an edge whose removal disconnects the graph, we say it is *trivial* if it is linked to a degree 1 vertex. We say that the *blobs of an undirected graph* are its maximal bridgeless components. Note that contrary to the biconnected components of a graph used to define blocks, there is no blob with exactly two vertices. A *minimal cut* is a set of edges whose removal disconnects the graph, which is minimal for inclusion.

Definition 2. An *unrooted phylogenetic network* N on a set X of taxa is a loopless graph whose vertices have either degree 3 (*internal vertices*), or degree 1 (the *leaves*), and such that its set $L(N)$ of leaves is bijectively labeled by X. An *unrooted level-k network* N on a set X of taxa is an unrooted phylogenetic network such that an unrooted tree connecting all vertices of N can be obtained by removing at most k edges per blob, as illustrated in the network N in Fig. 1. An unrooted phylogenetic network is *simple* if all its cut-edges are trivial.



N N'_1 N'_2

Fig. 1. An unrooted level-2 network N with leaf set $\{a, b, c, d, e, f, g, h, i\}$. All unlabeled vertices are internal vertices. The gray areas correspond to the blobs of N, and the bold arcs are such that their deletion transforms N into an unrooted tree. As we discuss in Section 3, there may be many possibilities (e.g. N'_1 and N'_2, whose arcs are directed downwards) to root an unrooted phylogenetic network, even if the root is chosen at the same position, which all provide a rooted phylogenetic network of the same level.

An unrooted level-0 network is usually called an *unrooted phylogenetic tree* and an unrooted level-1 network is simply an unrooted *galled tree*[27]. Note that an unrooted level-1 network is *outerplanar*, i.e. it has an embedding in the plane with no crossing edges, and all vertices on its outer face. For the sake of simplicity, in the following, we will identify each leaf with its label.

Note that, as usually done[20, 28], we consider that all rooted or unrooted level-k phylogenetic networks do not contain any block or blob with less than four vertices. This restriction is natural, as it prevents from adding superfluous edges which make the network more complex but do not have any influence on the set of triplets or quartets consistent with it. A *quartet* $ab|cd$ is an unrooted phylogenetic tree on four

leaves $a, b, c, d \in X$, where a and b (resp. c and d) share a common neighbor, as shown in Fig. 2.

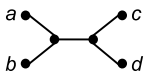


Fig. 2. The quartet $ab|cd$.

Definition 3. We say that an unrooted phylogenetic network N on a set X of taxa is *consistent* with the quartet $ab|cd$ (or equivalently $ab|cd$ is consistent with N) if N contains two distinct vertices u and v, four leaves a, b, c and $d \in L(N)$, and pairwise internally vertex-disjoint paths (i.e. which share no internal vertex) from a to u, from b to u, from u to v, from v to c and from v to d.

For example the unrooted level-2 network N of Fig. 1 is consistent with, amongst other, the quartets $cd|ef$, $fe|dg$, $fd|eg$ and $fg|ed$, and not with, amongst other, $ac|bd$ or $ai|dh$. This definition is equivalent to the following:

Definition 4. A quartet $ab|cd$ is *consistent* with an unrooted phylogenetic network N if there exist two vertex-disjoint paths in N, one from a to b and the other from c to d.

Indeed, Definition 3 trivially implies Definition 4, and the converse is true: as a, b, c and d are leaves, there has to be a path (whose extremities are u and v) not containing a, b, c nor d to join the two disjoint paths $a - b$ and $c - d$ in the connected graph N.

Also note that these definitions of quartet consistency are appropriate for the *completely resolved* unrooted phylogenetic networks we are studying here, i.e. networks with vertices of degree at most three. In the context of unresolved unrooted phylogenetic networks, with vertices of degree greater than 3, this definition should be adapted to allow multiple quartet resolutions for quartets involving these vertices.

A set Q of quartets is *consistent* with a network N if every quartet of Q is consistent with N. The set of all quartets consistent with N is denoted by $Q(N)$. We say that Q is *dense* if it contains at least one quartet on any subset of four leaves from X. For $A \subsetneq X$, we define the *restriction* of Q to A to be $Q_{|A} = \{ab|cd \text{ such that } a, b, c, d \in A\}$.

Given a set X of taxa, a *split* $S = A|\bar{A}$ (or equivalently $\bar{A}|A$) is a bipartition of X into two nonempty and complementary sets. We call S *trivial* if $|A| = 1$ or $|\bar{A}| = 1$. Two distinct splits $S_1 = A_1|A_1'$ and $S_2 = A_2|A_2'$ are *compatible* if one of the four intersections $A_1 \cap A_2$, $A_1 \cap A_2'$, $A_1' \cap A_2$ or $A_1' \cap A_2'$ is empty[2]. A *split system* $\mathcal{S}$ is a set of splits, it is *compatible* if its splits are all pairwise compatible.

It is *circular* if there exists an order σ on X such that for any split $A|\bar{A}$ of $\mathcal{S}$, A or $\bar{A}$ is an interval of σ (i.e. a set of consecutive elements of σ).

A split $A|\bar{A}$ is *consistent* with an unrooted phylogenetic tree T if T contains an edge which disconnects A and $\bar{A}$, i.e. A and $\bar{A}$ belong to two distinct connected components. An unrooted phylogenetic tree T on a set X of taxa is *contained* in an unrooted level-k network N if T can be obtained from N by a sequence of edge removals and edge contractions. We can now define split consistency with an unrooted phylogenetic network, following Woolley et al.[29]

Definition 5. A split $A|\bar{A}$ is *consistent* with an unrooted phylogenetic network N if it is consistent with an unrooted phylogenetic tree contained in N.

We finally call $\mathcal{S}(N)$ the set of all splits consistent with an unrooted phylogenetic networks, and say that a split system $\mathcal{S}$ is *consistent* with N if $\mathcal{S} \subset \mathcal{S}(N)$.

3. Rooted and Unrooted Level-*k* Networks

In this section, we illustrate the fact that there are many possible ways to root an unrooted level-k network by directing its edges. We first give a definition to formally explain how to root an unrooted level-k network to get a rooted level-k phylogenetic network.

Definition 6. *Rooting* an unrooted level-k network $N = (V, E)$ consists in obtaining a rooted phylogenetic network $N' = (V \cup \{r\}, A)$ in the following way:

(i) locating the root, i.e. choosing an edge xy of N and subdividing it to create a degree 2 vertex r (which will become the root of N', parent of x and y), which provides a graph N'';
(ii) orienting the edges of N'' to transform them into the arcs of N', i.e.:
   - choosing an order $\sigma : V \cup \{r\} \to [0..|V|]$ such that $\sigma(r) = 0$ and $\forall u \in V, \exists v \in V \cup \{r\}$ such that:
     * $uv \in E$,
     * and $\sigma(v) < \sigma(u)$,
     * and ensuring that every degree-3 vertex in N will have at least one parent and one child in N'. More formally, for each degree 3 vertex u of N, there is at least one vertex $v' \in V$ such that $uv' \in E$ and $\sigma(u) < \sigma(v')$;
   - setting the set of arcs $A = \{(u, v)$ such that $uv \in E$ and $\sigma(u) < \sigma(v)\}$.

An important remark is that, even when the position of the root is chosen at step (i), many rootings are still possible depending on the edge orientation chosen at step (ii), as illustrated by the networks N'_1 and N'_2 in Fig. 1, and in the following proposition.

Proposition 1. *For any integer $k \geq 2$, there exists an unrooted level-k network N_k with $2k$ leaves such that N_k has at least 2^k rootings where the root r is put on the same edge of N_k.*

Proof: We first describe how to recursively build N_k. To build N_1, we consider a cycle with two vertices v_0^0 and v_0^1 which are linked respectively to a leaf x_0^0 and a leaf x_0^1. To build N_{k+1} from N_k, call e_{k-1}^0 the edge incident to v_{k-1}^0 but not to v_{k-1}^1, and e_{k-1}^1 the edge incident to v_{k-1}^1 but not to v_{k-1}^0. Subdivide these two edges (to build N_2 from N_1, as both edges link v_0^0 and v_0^1, just subdivide one of these two edges twice) and connect the two vertices created by the subdivision, then subdivide this new edge twice to obtain two vertices : v_k^0 (the closest to v_{k-1}^0) and v_k^1. Finally, add two leaves x_k^0 and x_k^1 connected respectively to v_k^0 and v_k^1. For example N_3 is illustrated in Fig. 3.

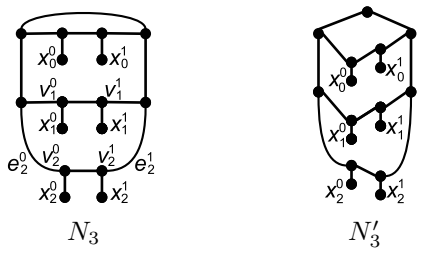


Fig. 3. Lower bound on the number of rootings: the unrooted level-3 network N_3 has at least 2^3 rootings. The level-3 network N_3', whose arcs are directed downwards, is an example of these rootings.

To check that N_k has at least 2^k rootings, we show how every integer in $a = \sum_{i=0}^{k-1} a_i 2^i \in [0..2^k - 1]$, where $a_i \in \{0, 1\}$, can be bijectively associated to a rooting of N_k. The idea is to root N_k such that the leaf $x_i^{a_i}$ is below a hybrid vertex and the leaf $x_i^{1-a_i}$ is below a split vertex. For example we show a rooting N_3' of N_3 corresponding to $4 = 0 \times 2^0 + 0 \times 2^1 + 1 \times 2^2$ in Fig. 3. □

Theorem 1. *Any rooting of an unrooted level-k network N provides a rooted level-k network N'.*

Proof: At step (ii) of the rooting process, orientations of the arcs are chosen such that if v can be reached by a directed path from u in N' then $\sigma(u) < \sigma(v)$. Hence N' is acyclic, otherwise it would contain two vertices u and v such that $\sigma(u) < \sigma(v)$ and $\sigma(v) < \sigma(u)$, a contradiction.

We now check that N' respects the degree condition of level-k networks. Step (i) of the rooting process ensures that N' has a root. Step (ii) guarantees that every degree 1 vertex in N is an indegree 1 vertex in N'. The remaining vertices have degree 3 in N. Step (ii) forces them to have indegree at least 1 and outdegree at least 1 in N'. Then, depending on the orientation of their third incident edge, they become split or hybrid vertices in N'.

Finally, we prove that the block B' of N' corresponding to an unrooted level-x blob B of N, for $x > 0$, contains x hybrid vertices. In the following, we focus on the graph B and the digraph B' themselves, without considering the rest of the graph N or the digraph N'. Recall that as B has unrooted level x, removing x edges from B provides a tree T.

We denote by $e(B')$ the number of arcs of B' and $v(B')$ its number of vertices. We also call respectively h and s its number of hybrid and split vertices (excluding the root of B'). Then we have $v(B') = 1 + h + s$. If we consider the arcs of B' as incoming arcs, we have $e(B') = 2h + s$. Thus, $h = e(B') - v(B') + 1$.

Now, note that the number $v(B)$ (respectively $e(B)$) of vertices (resp. edges) of B can be $v(B') - 1$ (resp. $e(B') - 1$) or $v(B')$ (resp. $e(B')$) depending on whether the root of N' belongs to B' or not. Thus, we have $e(B) - v(B) = e(B') - v(B')$. As the tree T covers all vertices of B, it has $v(B) - 1$ edges. We also know that B has x edges more than T, so $e(B) = v(B) - 1 + x$. Finally, we conclude that $h = e(B') - v(B') + 1 = e(B) - v(B) + 1 = x$, so B' has x hybrid vertices, so the level is the same in the rooted and unrooted context. □

4. Splits and Unrooted Level-*k* Networks

In this section, we give properties which link abstract and explicit phylogenetic networks.

The following definition of split consistency with an unrooted network, based on minimal cuts, is similar to the one given by Brandes and Cornelsen[30]. Although they claim that explicit networks "represent sets of splits differently" than the minimal cut representation they propose for the networks they consider, we show that the following definition and Definition 5 are equivalent.

Definition 7. A split $A|\bar{A}$ is consistent with an unrooted phylogenetic network N if there is a minimal cut of N which disconnects A and $\bar{A}$.

Proposition 2. *Definitions 5 and 7 are equivalent.*

Proof: Suppose N is disconnected by a minimal cut E into N_A which contains all leaves of A and $N_{\bar{A}}$ which contains $\bar{A}$. Then let uv be an edge of E, and consider two spanning trees T_u of N_A and T_v of $N_{\bar{A}}$ such that $u \in T_u$ and $v \in T_v$. Then $A|\bar{A}$ is consistent with a tree T' contained in N, obtained by the union of T_u, T_v and uv, and contraction of degree-2 vertices with one of their neighbors (avoiding of course the contraction of edge uv).

Now suppose $A|\bar{A}$ is consistent with a tree contained in N. Then there exists an edge e_x of T which disconnects T into two subtrees, T_A which contains A and $T_{\bar{A}}$ which contains $\bar{A}$. Reversing the edge deletion and edge contraction operations performed to obtain T from N, we call x an edge of N corresponding to e_x, and N_A (respectively $N_{\bar{A}}$) the induced subgraph of N corresponding to the vertices of T_A (respectively $T_{\bar{A}}$).

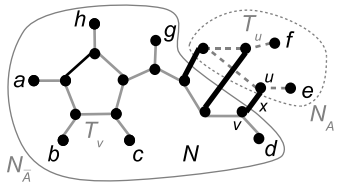


Fig. 4. An unrooted level-2 phylogenetic network N, separated in two networks N_A and $N_{\bar{A}}$ by a cut shown in bold edges, representing bipartition $A|\bar{A}$, for $A = \{e, f\}$. Edges of the spanning tree T_u of N_A are shown in solid gray lines, and edges of the spanning tree T_v of $N_{\bar{A}}$ in dotted gray lines.

We consider the set E of edges of N with one vertex in N_A and the other one in $N_{\bar{A}}$. This set E is a minimal cut of N which disconnects A from $\bar{A}$. It is a cut because when we delete the edges of E, we disconnect N_A from $N_{\bar{A}}$. If it was not minimal, then there would exist another cut $E' \subsetneq E$. Consider an edge $x' \in E - E'$: it connects N_A and $N_{\bar{A}}$, which are both connected, so E' is not a cut of N. Thus, we have found a minimal cut, E, which disconnects A and $\bar{A}$. □

Note that the fact that all splits consistent with an unrooted phylogenetic tree T are consistent with an unrooted phylogenetic network N does not necessarily imply that T is contained in N. For example, the splits of the tree T of Fig. 5(iii) are consistent with the network N' of Fig. 5(ii) but T is not contained in N'.

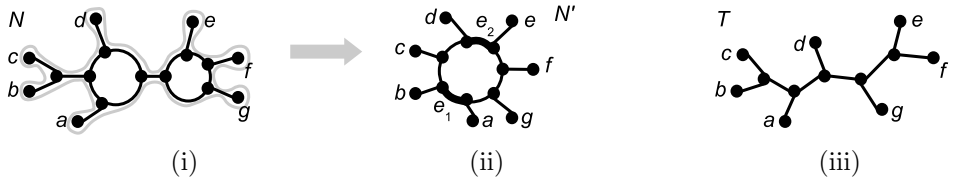


Fig. 5. An unrooted level-1 network N (i) and a simple unrooted level-1 network $N' = Simple(N)$ built from the order $\sigma = abcdefg$ (ii) such that $\mathcal{S}(N) \subset \mathcal{S}(N')$ and $Q(N) \subset Q(N')$ (see Lemmas 1 and 2). However the tree T is contained in N but not in N'.

We now introduce a transformation operation to obtain a simple unrooted level-1 network from an unrooted level-1 network, illustrated in Fig. 5.

Definition 8. Given an unrooted level-1 network N, we define as $Simple(N)$ a network obtained from N in the following way:

- as N is outerplanar, we consider the order σ of its leaves around the outer face of a planar embedding of N,
- the graph $Simple(N)$ is obtained by attaching the leaves of X to a cycle respecting the order σ.

As $Simple(N)$ contains only one cycle and cut-edges leading to the leaves, it is clearly a simple unrooted level-1 network.

Lemma 1. *Let N be an unrooted level-1 network and $N' = Simple(N)$, then $\mathcal{S}(N) \subset \mathcal{S}(N')$.*

Proof: Let $A|\bar{A}$ be a split consistent with N, represented by a minimal cut of N, i.e. either a single cut-edge e, or a pair of edges $\{e_1, e_2\}$ of the same cycle of N. Now consider the embedding of N used to build N', and the order σ of leaves around the outer face in this embedding.

If $A|\bar{A}$ is represented by a single cut-edge e in N as in Fig. 6(i), then we can draw a closed curve intersecting only edge e of N. Otherwise, we draw a closed curve intersecting only edges e_1 and e_2 of N, as in Fig. 6(ii). In both cases, this curve splits the outerface of N in two parts, one containing A and the other $\bar{A}$, and the set of leaves contained in one of these parts appears as an interval of σ.

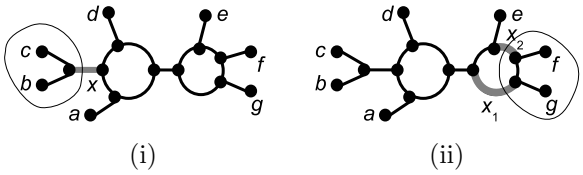


Fig. 6. Bipartitions in an unrooted level-1 network whose leaves are in the order $\sigma = abcdefg$ around the outer face: $\{b, c\}|\{a, d, e, f, g\}$ is represented by edge x and $\{f, g\}|\{a, b, c, d, e\}$ is represented by the cut $\{x_1, x_2\}$.

Thus, these leaves also appear consecutively around the cycle of N', which implies that $A|\bar{A}$ also belongs to $S(N')$. □

Theorem 2. *Let $\mathcal{S}$ be a split system on a set X of taxa. Then $\mathcal{S}$ is circular if and only if there exists an unrooted level-1 network N such that $\mathcal{S} \subset \mathcal{S}(N)$.*

Proof: $\Rightarrow$: As $\mathcal{S}$ is circular, consider one of the orders σ such that for each split $A|\bar{A} \in \mathcal{S}$, either A or $\bar{A}$ appears as an interval in σ. We build the simple unrooted level-1 network N by attaching the leaves of X to a cycle respecting the order σ. Then, for any split $A|\bar{A}$, N restricted to A and their neighbors is a connected graph, which is connected through two edges e_1 and e_2 to the rest of the graph, as illustrated in Fig. 5(ii) for $A = \{b, c, d\}$. Thus $\{e_1, e_2\}$ is a minimal cut of N which disconnects A and $\bar{A}$, so $A|\bar{A}$ is consistent with N. Finally, $\mathcal{S} \subset \mathcal{S}(N)$.

$\Leftarrow$: Suppose there exists an unrooted level-1 network N such that $\mathcal{S} \subset \mathcal{S}(N)$. Then there also exists a simple unrooted level-1 network $N' = Simple(N)$ such that $\mathcal{S}(N) \subset \mathcal{S}(N')$ from Lemma 1, so $\mathcal{S}$ is also consistent with N'. The order σ obtained by considering the leaves around the cycle of N', turning clockwise, and starting from any leaf, shows that $\mathcal{S}(N')$ is circular, so $\mathcal{S}$ is circular. □

As the QNet algorithm is a heuristic algorithm to reconstruct a circular split system $\mathcal{S}$ from a set Q of weighted quartets[8] (aiming at optimizing the weights of consistent quartets), thanks to Theorem 2, it would be possible to find an unrooted level-1 network consistent with $\mathcal{S}$. This would provide the first heuristic method to reconstruct explicit networks from quartets. We now focus on direct explicit network reconstruction from quartets, and give other quartet properties and algorithms in this perspective.

5. Quartets and Unrooted Level-k Networks

The first natural algorithmic question concerning quartets and unrooted phylogenetic networks is the time complexity for checking consistency. The dynamic programming approach used to compute the set of all triplets consistent with a rooted phylogenetic network in $O(n^3)$ time[22] does not extend to the unrooted case. However the problem is solvable in polynomial time.

Theorem 3. *The set of all quartets consistent with an unrooted level-k network N can be computed in $O\left(n^5(1+\alpha(n,n))\right)$ time, where α is the inverse of the Ackermann function.*

Proof: Using Definition 4, we just apply the best currently known algorithm for the 2-Vertex-Disjoint Paths Problem, which decides whether there exist two vertex-disjoint paths, one between a and b and the other between c and d in $O(n + n\alpha(n,n))$ time[31], for each of the $O(n^4)$ quartets $ab|cd$. Hence, the overall complexity of the algorithm is $O\left(n^5(1+\alpha(n,n))\right)$ to retrieve all quartets consistent with N. □

We now focus on reconstructing an unrooted level-k network consistent with a set of quartets.

Problem 1 (Level-k Quartet Consistency).
Input: a quartet set Q on a set X of taxa.
Output: decide whether there exists an unrooted level-k network on X, which is consistent with every quartet in Q.

We recall that in the rooted case, for level 0, this problem for triplets is polynomial time solvable[24]. However, for upper levels it is NP-complete, by reduction form Set Splitting[18, 28]. Steel proved in 1992 that the problem Level-0 Quartet Consistency is NP-complete[25], by reduction from Betweenness [32].

We first prove that for level 1, this problem is equivalent to the same problem with the restriction that the network to reconstruct is simple. In this case, we refer to the problem as Simple Level-1 Quartet Consistency.

Lemma 2. *Let Q be a quartet set. There exists an unrooted level-1 network N consistent with Q if and only if there exists a simple unrooted level-1 network N' consistent with Q.*

Proof: $\Rightarrow$: Consider an arbitrary unrooted level-1 network N consistent with Q, and a simple unrooted level-1 network $N' = Simple(N)$ (see Fig. 5). For any quartet $ad|bc$ in Q, as Q is consistent with N, there exists an unrooted tree which is contained in N and is consistent with $ad|bc$. In particular, this tree is consistent with a split $A|\bar{A}$ such that $a, d \in A$ and $b, c \in \bar{A}$. So $A|\bar{A}$ is consistent with N, and with N' from Lemma 1, so $ad|bc$ is consistent with N'.

$\Leftarrow$: As N' is a simple unrooted level-1 network, then in particular it is an unrooted level-1 network, and it is consistent with Q. □

Theorem 4. Level-1 Quartet Consistency *is NP-complete.*

Proof: As it is possible to check in polynomial time that a quartet set is contained in an unrooted level-1 network, thanks to Theorem 3, the Level-1 Quartet Consistency problem is in NP.

To prove that it is NP-hard, as Lemma 2 shows that the Level-1 Quartet Consistency problem is strictly equivalent to the Simple Level-1 Quartet Consistency problem, it is sufficient to show that the latter is NP-complete. However, it is an equivalent formulation of the QCIRC problem[33], which is NP-complete, by reduction from Betweenness. We recall that the QCIRC problem takes as input a set Q of quartets on a set X of taxa, and asks whether there exists an order σ on X such that for each quartet $ab|cd \in Q$, there exists a split $A|\bar{A}$, where A is an interval of σ, such that $a, b \in A$ and $c, d \in \bar{A}$ (or $a, b \in \bar{A}$ and $c, d \in A$). Considering that σ is the order of the leaves around the cycle of a simple level-1 network (starting from any point, and turning clockwise, for example), it is straightforward to see that the problem is strictly equivalent to Simple Level-1 Quartet Consistency. □

6. Finding the blobs from the quartets

We now focus on the Level-k Quartet Consistency problem when the input is a dense quartet set Q, i.e. when Q contains at least one quartet on every set of four leaves. In this section, we show how to find the blobs of an unrooted phylogenetic network N consistent with Q. To this purpose, we will introduce the concept of SN-split, which is the unrooted analogue of SN-set[16, 21].

Note that thanks to Lemma 2, if we know that an unrooted level-1 network is consistent with a quartet set Q, then there also exists a simple unrooted level-1 network N', which is more parsimonious in terms of edges and also contains Q. However, this simple network N' is biologically less interesting, as it does not optimize the quantity $|Q(N') - Q|$, i.e. the number of quartets which could be considered "false positive" as they are consistent with N', but not present in the input set Q. This explains why in practice we will first try to deduce from Q the blobs of the network to reconstruct, and then focus on reconstructing each blob of the level-1 network if possible.

6.1. *Building the SN-splits*

Definition 9. Let Q be a set of quartets on a set X of taxa, $A \subseteq X$. A split $A|\bar{A}$ of the taxa is an *SN-split* of Q if it is either a trivial split, or it satisfies the following property: for any $x, y \in A$, $z, t \in \bar{A}$, the only quartet on $\{x, y, z, t\}$ in Q, if there is any, is $xy|zt$.

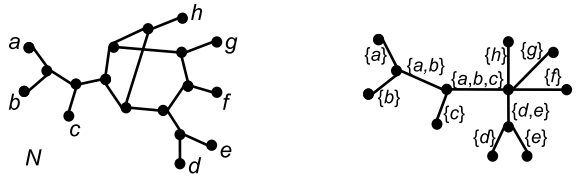
This definition of SN-split is similar to a definition of SN-sets[21]. The original SN-set definition[16] can also be adapted to define SN-splits in the unrooted context as the closure of some set completion operation. However we do not describe it, as it is more complex than the one used here.

We now give an important property of SN-splits before describing an approach to efficiently compute them.

Proposition 3. *For a dense set Q of quartets, the set of SN-splits of Q is a compatible split system.*

Proof: We consider two SN-splits $S_1 = A_1|A'_1$ and $S_2 = A_2|A'_2$. Suppose by contradiction that none of the four intersections $A_1 \cap A_2$, $A_1 \cap A'_2$, $A'_1 \cap A_2$ and $A'_1 \cap A'_2$ is empty. Then $\exists a \in A_1 \cap A_2$, $b \in A_1 \cap A'_2$, $c \in A'_1 \cap A_2$ and $d \in A'_1 \cap A'_2$. As Q is dense, it must contain a quartet on $\{a, b, c, d\}$. As S_1 is an SN-split, $a, b \in A_1$ and $c, d \in \bar{A}_1$, this quartet should be $ab|cd$. But as S_2 is also an SN-split, $a, c \in A_2$ and $b, d \in \bar{A}_2$, this quartet should be $ac|bd$: this contradicts the definition of an SN-split. □

We recall the classical property that a compatible split system can be represented by an unrooted tree, which we call the *unrooted SN-tree* of a dense quartet set, whose set of edges is bijectively labeled by the set of SN-splits, as illustrated in Fig. 7(ii). We now show how to build this unrooted SN-tree in $O(n^4)$ time.



(i) (ii)

Fig. 7. An unrooted level-2 network N (i) and the SN-tree of its quartet set $Q(N)$ (ii), where we label by the leaf set A the edge corresponding to the SN-split $A|\bar{A}$.

Proposition 4. *For a dense set Q of quartets, there are $O(n)$ SN-splits, and the unrooted SN-tree can be reconstructed in $O(n^4)$ time.*

Proof: Proposition 3 implies that the number of SN-splits is linear in the size of X.

The algorithm to reconstruct the unrooted SN-tree works as follows. We first partition the input set Q into a set Q_1 where there is exactly one quartet on each set of four leaves, and a set of remaining quartets Q_2. Then, using the "Q^* algorithm[34]", we reconstruct from Q_1 an unrooted tree T^*, not necessarily binary, whose leaves are labeled by X, which satisfies the following property: $Q(T^*) \subseteq Q_1$ and $Q(T^*)$ is of maximum size.

Finally, for each remaining quartet $q \in Q' = Q_2 \cup (Q_1 - Q(T^*))$, if q is not consistent with T^*, then we modify the tree T^* in the following way, to finally obtain an unrooted tree T'^*.

Let $\{a, b, c, d\}$ be the set of leaves of q, $ab|cd$ be the quartet on $\{a, b, c, d\}$ consistent with T^*, u be the intersection vertex of paths $a - b$, $a - c$ and $b - d$, and v be the intersection vertex of paths $c - d$, $c - a$ and $d - b$. Then contract all vertices in the path between u and v.

We can perform this step efficiently in total $O(n^4)$ time in the following way: after a $O(n)$ preprocessing, it is possible to decide in constant time, for each of the $O(n^4)$ quartets of Q', whether it is consistent with T^*. The trick is to use constant time lowest common ancestor queries[35] in a rooted version of T^*. For each of the $O(n)$ edge contractions in the tree T^*, it takes $O(n)$ time to recompute a rooted version of the tree and the data structure for lowest common ancestors queries. Hence, the total time complexity is $O(n^4 + n^2) = O(n^4)$.

It remains to prove that the algorithm is correct, i.e. that the T'^* tree obtained in the end is the SN-tree of Q.

Consider an edge e of T'^* which partitions the taxa of X into A and $\bar{A}$. As $Q(T'^*) \subseteq Q(T^*) \subseteq Q_1 \subseteq Q$, $\forall a, b \in A, b, c \in \bar{A}$, $ab|cd \in Q$. Furthermore, suppose by contradiction that $q_1 = ac|bd \in Q$ or $q_2 = ad|bc \in Q$. As neither q_1 nor q_2 are consistent with T^*, e would have been contracted during the algorithm: impossible as e is an edge of T'^*, so $A|\bar{A}$ is an SN-split of Q.

Conversely, given an SN-split $A|\bar{A}$ of Q, $\forall a, b \in A, c, d \in \bar{A}$, $ab|cd \in Q$, $ac|bd \notin Q$ and $ad|bc \notin Q$ so $ab|cd \in Q(T^*)$, so T^* contains an edge e which separates A from $\bar{A}$. Suppose by contradiction that this edge is contracted by the algorithm and is no more present in T'^*. Then there exists a set $\{a, b, c, d\}$ of four leaves such that $a, b \in A$, $c, d \in \bar{A}$, and Q contains a quartet on $\{a, b, c, d\}$ which is not consistent with T^*. As $A|\bar{A}$ is an SN-split, this is impossible, so there exists an edge of T'^* which separates A from $\bar{A}$.

Finally, the splits of T'^* are exactly the SN-sets of Q, so T'^* is the SN-tree of Q. □

As a corollary, the set of all SN-splits of a dense quartet set Q can be computed in $O(n^4)$ time. Note that this time is optimal when Q has at least one non-trivial SN-split $A|\bar{A}$: in this case, $|A| > 1$ and $|\bar{A}| > 1$, so, to ensure that $A|\bar{A}$ is indeed an SN-split of Q, it must be checked that none of the $O(n^4)$ quartets of type $ab|cd$ (with $a, c \in A$ and $b, d \in \bar{A}$) belongs to Q.

6.2. *The link between blobs and SN-splits*

We now prove two lemmas before showing the link between the SN-splits of Q and the blobs of N.

Definition 10. For each blob B of an unrooted phylogenetic network N, we define $E(B)$ as the set of cut-edges $\{e_1, \ldots, e_t\}$ having one vertex in B. For any edge $e_i = b_i c_i \in E(B)$ where $b_i \in B$ and $c_i \notin B$, we denote by $L_B(e_i)$ the set of leaves of the connected component of $N - \{e_i\}$ containing c_i.

Lemma 3. *Given an unrooted level-k phylogenetic network N, and a dense quartet set Q consistent with N, then for each SN-split $A|\bar{A}$ of Q, there exists a blob B in N such that $E(B)$ is partitioned into two disjoint sets E_A and $E_{\bar{A}}$ such that $A = \bigcup_{e \in E_A} L_B(e)$ and $\bar{A} = \bigcup_{e \in E_{\bar{A}}} L_B(e)$.*

Proof: Suppose this property is false, and let $A|\bar{A}$ be an SN-split of Q such that for each blob B in N, $E(B)$ cannot be partitioned into two disjoint sets E_A and $E_{\bar{A}}$ such that $A = \bigcup_{e \in E_A} L_B(e)$ and $\bar{A} = \bigcup_{e \in E_{\bar{A}}} L_B(e)$, i.e. there exists a cut-edge $e \in E(B)$ such that neither $L_B(e) \subseteq A$ nor $L_B(e) \subseteq \bar{A}$.

Let B_1 and e_1 be respectively a blob and an edge of N such that $A \subseteq L_{B_1}(e_1)$, and $L_{B_1}(e_1)$ is minimal for inclusion. The existence of such elements is guaranteed by the fact that the incident edge e_x of any leaf $x \in \bar{A}$ satisfies $A \subseteq L_{\{x\}}(e_x)$. We call v the vertex incident of e_1 not in B_1, and call B_2 the blob containing v, as shown in Fig. 8.

As stated in the beginning of the proof, there exists a cut-edge $e_2 \in E(B_2)$ such that neither $L_{B_2}(e_2) \subseteq A$ nor $L_{B_2}(e_2) \subseteq \bar{A}$, so there exist $a_2 \in A \cap L_{B_2}(e_2)$ and $x_2 \in \bar{A} \cap L_{B_2}(e_2)$. As $A \subseteq L_{B_1}(e_1)$, we know that no leaf of A is contained in $L_{B_2}(e_1)$, so $L_{B_2}(e_1) \subseteq \bar{A}$, therefore $e_1 \neq e_2$, and there exists $x_1 \in \bar{A} \cap L_{B_2}(e_1)$.

Let us consider the edges of $E(B_2)$ other than e_1 and e_2 (there is at least one). If $\forall e \in E(B_2) - \{e_1, e_2\}$, $L_{B_2}(e) \subseteq \bar{A}$, then e_2 and B_2 also satisfy $A \subseteq L_{B_2}(e_2)$ and $L_{B_2}(e_2) \subsetneq L_{B_1}(e_1)$, which contradicts the minimality of $L_{B_1}(e_1)$. Otherwise, there exists $e_2' \in E(B_2) - \{e_1, e_2\}$, $a_1 \in L_{B_2}(e_2') \cap A$. Then, edge e_2 implies that $x_1 a_1 | x_2 a_2 \in Q(N)$, and $x_1 x_2 | a_1 a_2 \notin Q(N) \supseteq Q$, which contradicts the fact that $A|\bar{A}$ is an SN-split of Q.

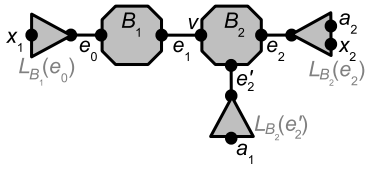


Fig. 8. An impossible configuration if $a_1, a_2 \in A$, $x_1, x_2 \in \bar{A}$, and $A|\bar{A}$ is an SN-split of Q.

□

Lemma 4. *Given an unrooted level-k phylogenetic network N, if there exists a blob B, four different edges e_1, e_2, e_3 and $e_4 \in E(B)$ and four different leaves $a \in L_B(e_1)$, $b \in L_B(e_2)$, $c \in L_B(e_3)$, $d \in L_B(e_4)$, then at least two of the three quartets $ab|cd$, $ac|bd$ and $ad|bc$ are consistent with N.*

Proof: The blob B is bridgeless, and the vertices of N have maximum degree 3, so B contains no articulation vertex, so for any two pairs of distinct vertices $\{x, y\}$ and $\{z, t\}$ of B, by Menger's Theorem, there exists at least two vertex-disjoint paths in B between $\{x, y\}$ and $\{z, t\}$.

Let us call respectively a', b', c' and d' the vertices of B adjacent to e_1, e_2, e_3 and e_4. By Menger's Theorem, there exist two vertex-disjoint paths P_1 and P_2 in B between $\{a', b'\}$ and $\{c', d'\}$.

If P_1 is a path between a' and c' and P_2 is a path between b' and d', then we apply Menger's Theorem in B between $\{a', c'\}$ and $\{b', d'\}$ and thus find two quartets on leaves $\{a, b, c, d\}$.

If P_1 is a path between a' and d' and P_2 is a path between b' and c', then we apply Menger's Theorem in B between $\{a', d'\}$ and $\{b', c'\}$ and thus find two quartets on leaves $\{a, b, c, d\}$. □

Theorem 5. *Let N be an unrooted level-k phylogenetic network. Its set of cut-edges is in bijective correspondence with the SN-splits of the set $Q(N)$.*

Proof: For each cut-edge e in N, as $Q(N)$ is consistent with N, the bipartition of leaves induced by e is clearly an SN-split of $Q(N)$. Suppose there exists an SN-split $A|\bar{A}$ which is not represented by any cut-edge of N. By Lemma 3, there exists a blob B of N such that $E(B)$ is partitioned into E_A and $E_{\bar{A}}$, where $A = \bigcup_{e \in E_A} L_B(e)$ and $\bar{A} = \bigcup_{e \in E_{\bar{A}}} L_B(e)$. As $A|\bar{A}$ is not represented by any cut-edge of N, $|E_A| \geq 2$ and $|E_{\bar{A}}| \geq 2$, so there exist four different cut-edges $e_1, e_2 \in E_A$ and $e'_1, e'_2 \in E_{\bar{A}}$, and four leaves a_1 and a_2 in A, x_1 and x_2 in $\bar{A}$, such that $a_1 \in L_B(e_1)$, $a_2 \in L_B(e_2)$, $x_1 \in L_B(e'_1)$ and $x_2 \in L_B(e'_2)$. Then, by Lemma 4, two different quartets on $\{a_1, a_2, x_1, x_2\}$ are consistent with N, which contradicts the fact that $A|\bar{A}$ is an SN-split of $Q(N)$. □

Thanks to this theorem, we can consider the SN-tree of $Q(N)$ as a summary of N, as both have the same set of cut edges, and only differ in their blobs, which are simple vertices in the SN-tree, and bridgeless components in N. However the structure inside the blobs of N remains unknown, we will now study this structure in case of unrooted level-1 networks.

7. Reconstructing Unrooted Level-1 Networks from Quartets

7.1. *From the set of all quartets of a network*

Given the set of all quartets consistent with an unrooted level-1 network, it is possible to reconstruct it in polynomial time. We first show this for simple level-1 networks, after introducing the *quartet ordering graph*.

Definition 11. Given a set Q of quartets, we define the *quartet ordering graph* as $G(Q) = (\{\{a,b\} \mid a \neq b \in X\}, \{\{\{a,b\},\{b,c\}\} \mid \forall d \in X, ac|bd \notin Q\})$. For any edge $\{\{a,b\},\{b,c\}\}$ of this graph, we label it by b.

This definition is illustrated in Fig. 9(ii). Note that $G(Q)$ is an undirected graph because a and c have a symmetric role in the definition of the edges of $E(G(Q))$.

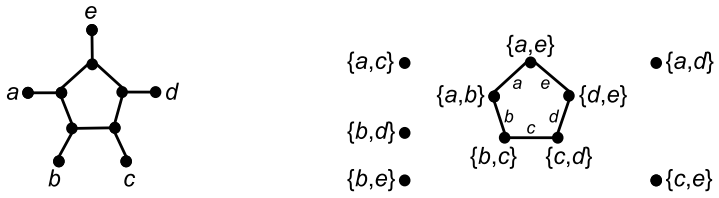


Fig. 9. A simple unrooted level-1 network N (i) and the quartet ordering graphs of his quartets $G(Q(N))$ (ii).

Lemma 5. *For a quartet set Q, we can decide in optimal $O(n^4)$ time whether there exists a simple unrooted level-1 network N such that $Q = Q(N)$, i.e. the set of all quartets consistent with N is exactly Q. Moreover, such a network can be computed in $O(n^4)$ time for any positive instance.*

Proof: In a simple unrooted level-1 network N, the n leaves hang around one cycle. We label the leaves by $[1..n]$ according to their position along the cycle. Our goal is to find this ordering given $Q(N)$.

Any three leaves a, b and c are consecutive iff there is no other leaf d hanging between a and c on the same side of the cycle as b, which is equivalent to $ac|bd \notin Q$ for any leaf d of X.

Hence, the quartet ordering graph $G(Q(N))$ is composed of one cycle of length n as well as isolated vertices, so the ordering of the leaves around the cycle of N corresponds to the ordering of the labels of the cycle $\{\{a,b\},\{b,c\}\} - \ldots - \{\{x,y\},\{y,a\}\} - \{\{y,a\},\{a,b\}\}$ of $G(Q(N))$.

To find this ordering, we build the quartet ordering graph $G(Q(N))$ in $O(n^4)$ (for any set of three leaves $\{a,b,c\}$ we test in $O(n)$ whether an edge should be added between $\{a,b\}$ and $\{b,c\}$). Then we extract the ordering in $O(n)$ starting from any edge of $G(Q(N))$: if we do not obtain a cycle of length n we answer NO. We finally check that the input quartet set Q is indeed equal to $Q(N)$, otherwise we answer NO. □

Theorem 6. *For a quartet set Q, we can decide in optimal $O(n^4)$ time whether there exists an unrooted level-1 network N such that $Q = Q(N)$, i.e. the set of all quartets consistent with N is exactly Q. Moreover, such a network can be computed in $O(n^4)$ time for any positive instance.*

Proof: We first build the unrooted SN-tree of Q thanks to Proposition 4. By Theorem 5, the cut-edges of any solution N are in bijective correspondence with the edges of the SN-tree of Q. So, as N has maximum degree three, any vertex of degree $\delta \geq 4$ in the SN-tree corresponds to a blob B with at least four vertices in N.

Let u be a vertex of degree $\delta \geq 4$ of the SN-tree, let $B(u)$ be the associated blob, and let $E(B(u)) = \{e_1, \ldots, e_\delta\}$. Then, for any distinct $a, b, c, d \in [1..\delta]$, for any leaves $l_a \in L_B(e_a)$, $l_b \in L_B(e_b)$, $l_c \in L_B(e_c)$ and $l_d \in L_B(e_d)$, $Q_{|\{l_a,l_b,l_c,l_d\}}$ must be consistent with the network reconstructed for blob $B(u)$.

So, for each vertex u of degree $\delta \geq 4$ in the SN-tree, let us call $A_i|\bar{A}_i$ the SN-splits corresponding to the incident edges of u (for $i \in [1..\delta]$), such that $A_x \cap A_y = \emptyset$ for any $x, y \in \delta$. We pick one leaf $l_i \in A_i$ for each $i \in [1..\delta]$, then we build in $O(\delta^4)$ time the simple unrooted level-1 network consistent with $Q_{|\{l_1,\ldots,l_\delta\}}$, thanks to Lemma 5. If this fails for one of the vertices of degree at least four of the SN-tree, then we answer NO. Otherwise we build N by replacing every such vertex in the SN-tree by the reconstructed network for $Q_{|\{l_1,\ldots,l_\delta\}}$. We finally check that $Q = Q(N)$, and answer NO if it is not the case. The overall time complexity of this algorithm is $O(n^4)$. □

7.2. *From a dense quartet set*

We show that in the case of a dense quartet set consistent with an unrooted level-1 network (i.e. a weaker condition than knowing all the quartets of the network), the SN-splits of Q are still related to the cut-edges of one of the solutions.

Lemma 6. *If a dense quartet set Q is consistent with an unrooted level-1 network N, then it is consistent with an unrooted level-1 network N' whose cut-edges are in bijective correspondence with the SN-splits of Q.*

Proof: For each cut-edge e in N, as Q is consistent with N, the bipartition of leaves induced by e clearly is an SN-split of Q. Now suppose there is a non-trivial SN-split $A|\bar{A}$ of Q which does not correspond to any cut-edge in N.

By Lemma 3, we know that there exists a blob C of N (in fact a cycle, as N has level 1) such that $E(C)$ is partitioned into E_A and $E_{\bar{A}}$, where $A = \bigcup_{e \in E_A} L_C(e)$ and $\bar{A} = \bigcup_{e \in E_{\bar{A}}} L_C(e)$. We label by X any vertex of C incident to a cut-edge in E_X, as shown in gray on Fig. 10.

We prove that the set of vertices labeled by A appears consecutively on C, i.e. the subgraph of C induced by all vertices labeled by A is a connected path. Suppose that it is not the case, then we can find two vertices a'_1 and a'_2 labeled by A such that on both paths between them in C, there is a vertex labeled by $\bar{A}$ (called x'_1 and x'_2 respectively), as illustrated in Fig. 10. We call e_1, e_2, e'_1 and e'_2 the cut-edges incident respectively to a'_1, a'_2, x'_1 and x'_2. So there exist two leaves a_1, $a_2 \in A$, x_1, $x_2 \in \bar{A}$ such that $a_1 \in L_C(e_1)$, $a_2 \in L_C(e_2)$, $x_1 \in L_C(e'_1)$ and $x_2 \in L_C(e'_2)$. This is impossible because in this case, $a_1a_2|x_1x_2$ does not belong to Q, so $A|\bar{A}$ is not an

SN-split of Q.

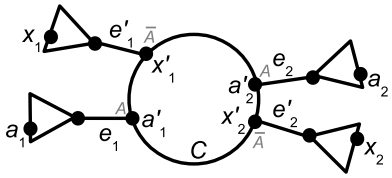


Fig. 10. A configuration which cannot happen in an unrooted level-1 network if $a_1, a_2 \in A$, $x_1, x_2 \in \bar{A}$ and $A|\bar{A}$ is an SN-split.

We now show that as the set of vertices labeled by A is contiguous, then it is possible to transform N into another level-1 network which is still consistent with Q. As shown in Fig. 11(b), we cut the cycle into two different cycles linked by a cut-edge. On one cycle, C_A, we hang, in the same order as in C, the cut-edges hanging from C which have a vertex labeled by A, while on the other cycle $C_{\bar{A}}$, we hang the cut-edges hanging from C which have a vertex labeled by $\bar{A}$, in the same order as in C. If one of these two cycles has less than four vertices, then we contract it into one vertex, like $C_{\bar{A}}$ in Fig. 11(b).

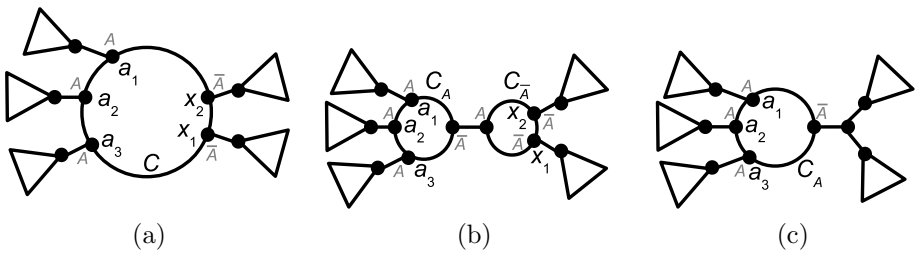


Fig. 11. The cycle splitting operation : the only cycle of the unrooted level-1 network N whose vertices are labeled, in gray, by both A and $\bar{A}$ (a), can be split into two parts (b) to represent the SN-split $A|\bar{A}$ of Q with a cut-edge, then the cycle $C_{\bar{A}}$, which has less than four vertices, is contracted (c).

We can now check that the new network, in which $A|\bar{A}$ is represented by a cut-edge, is consistent with Q. Note that the cycle splitting operation does not affect the quartets which contain zero or just one leaf of A, or, symmetrically, of $\bar{A}$, as the order of the vertices on the cycle has been conserved. For quartets having two leaves a_1, a_2 in A and two leaves x_1, x_2 in $\bar{A}$, we know that they have to be $a_1a_2|x_1x_2$ as $A|\bar{A}$ is an SN-split. Those quartets are consistent with the network after the cycle splitting operation.

So, finally, after applying to N the cycle splitting operation for each SN-split which is not represented by a cut-edge, we obtain a level-1 network N' whose cut-

edges correspond bijectively to the SN-splits of Q. □

7.3. *Simple unrooted level-1 networks*

Even though we are able to deduce the global structure of a level-1 solution consistent with a dense quartet set Q, if there exists one, the complexity of reconstructing a simple unrooted level-1 network from Q remains unknown. We will give properties indicating that this task may be difficult, as some dense quartet sets lead to solutions with very different structures. This suggests that the density restriction may be too weak to reconstruct unrooted level-1 networks, as it does not necessarily fix the structure of cut-edges around the cycles of the network.

Note that the problem of reconstructing a simple unrooted level-1 network from a quartet set Q is equivalent to the following problem: finding an order σ of the leaves such that for any quartet $ab|cd \in Q$ such that $\sigma(a) < \sigma(b)$ and $\sigma(c) < \sigma(d)$, neither $\sigma(a) < \sigma(c) < \sigma(b) < \sigma(d)$ nor $\sigma(c) < \sigma(a) < \sigma(d) < \sigma(b)$. This formulation is similar to the Non-Betweenness problem, which is known to be NP-complete in the general case[36] but whose complexity is unknown in the dense case.

Proposition 5. *For any two distinct simple unrooted level-1 networks N_1 and N_2, there exists a dense quartet set Q such that $Q \subset Q(N_1)$ and $Q \subset Q(N_2)$.*

Proof: For any set of four leaves $\{a, b, c, d\}$, both N_1 and N_2 are consistent with two among the three possible quartets on $\{a, b, c, d\}$. Hence, they share at least one common quartet. □

It is even possible to build a dense quartet set which is consistent with an exponential number of simple unrooted level-1 networks.

Proposition 6. *For any integer $n \geq 3$, there exists a dense quartet set on $2n$ leaves which is consistent with 2^n non-isomorphic simple unrooted level-1 networks.*

Proof: We consider the set of leaves $\{x_i, i \in [1..2n]\}$. Let us define some *leaf pairs* $P_i = \{x_{2i-1}, x_{2i}\}$, and the simple unrooted level-1 network N obtained by hanging the leaves around a cycle in the order $x_1 \dots x_{2n}$.

We now consider the following quartet set Q, which is consistent with N. For each set of four leaves $a, b, c, d \in [1..2n]$:

- case 1) if the four leaves belong to different leaf pairs P_i, say that $a < b < c < d$, we add $ab|cd$ and $bc|ad$ to Q.
- case 2) if exactly two leaves (say a and b) belong to a same leaf pair P_i, we add $ab|cd$ to Q.
- case 3) otherwise, two leaves (say a and b) belong to a leaf pair P_i and 2 others (c and d) belong to a pair P_j with $i < j$, then we add $ab|cd$ to Q.

Note that in this construction, two leaves belonging to a same leaf pair P_i have a symmetric position in the quartet added to Q. Hence, any other network obtained

by hanging the leaves in an order equal to the one of N, up to transpositions inside the leaf pairs P_i, is still consistent with Q. As there are n leaf pairs P_i, there are n possible transpositions, thus 2^n simple unrooted level-1 networks consistent with Q. □

To supplement these results, we have tried an approach based on "obstructions" to decide whether a dense quartet set Q is consistent with a simple level-1 network: this consists in identifying a finite size set of quartet sets of finite size (the obstructions) such that Q is consistent with a simple level-1 network if and only if it does not contain any of those obstructions. We have enumerated all 7 "minimally dense" quartet sets on five leaves (i.e. quartet sets with exactly one quartet for each set of four leaves) and have observed that each one is consistent with at least one simple unrooted level-1 network. Hence, looking for obstructions of size 5, it is necessary to consider quartet sets with at least one quartet with two conflicting resolutions.

8. Open Problems

Problems about quartets and unrooted phylogenetic networks are of interest from a graph-theoretical point of view, because they show more symmetry than triplets, deal with undirected graphs, thus seem to be more directly related to classical problems in graph theory, which could in turn be a way to understand better the combinatorics of triplets and rooted phylogenetic networks.

The time complexity of computing the set of all quartets consistent with an unrooted level-k network may be improved with appropriate preprocessing, it should be possible to get an optimal $O(n^4)$ bound.

The most puzzling open problem about unrooted phylogenetic networks and quartets is whether it is possible to reconstruct a simple unrooted level-1 network from a dense quartet set. This problem is similar to the Non-Betweenness problem with a dense set of constraints, whose complexity is also unknown.

Finding a dense quartet set which is consistent with a unique unrooted level-k network, for each k, could lead, like for rooted level-k networks and triplets[28], to an NP-completeness proof of the Level-k Quartet Consistency problem, for $k > 1$. Also, the approach to know how to partition a dense triplet set into different blocks of the level-k network to reconstruct[21] does not directly translate to the quartet context. Hence, the strategy needs to be adapted or changed more deeply in order to find a polynomial time algorithm to solve the Level-k Quartet Consistency problem, for a fixed k and a dense quartet set. The same applies for simple unrooted level-k networks, where the reconstruction algorithms for triplets in the rooted case cannot be adapted.

Finally, results on the structure of rooted level-k phylogenetic networks[37] hold for unrooted networks, which can also be decomposed as unrooted trees of unrooted level-k generators, defined as 3-regular biconnected multigraphs with $2k-2$ vertices. As generators seem a promising approach to finding a fixed-parameter algorithm for

rooted level-k network reconstruction from a dense triplet set[38], the same question is open for unrooted level-k network reconstruction from a dense quartet set.

Acknowledgments

We thank Sylvain Guillemot for improving the proof of Lemma 4, and Stéphan Thomassé for pointing out the 2-Disjoint Path Problem. We also thank Steven Kelk and anonymous referees whose comments helped to improve this paper. This work was supported by the French ANR projects ANR-06-BLAN-0148-01 (GRAAL) and ANR-08-EMER-011-01 (PhylARIANE).

References

1. DA Morrison. Networks in phylogenetic analysis: new tools for population biology. *International Journal for Parasitology*, 35:567–582, 2005.
2. P Buneman. The recovery of trees from measures of dissimilarity. In FR Hodson, DG Kendall, and P Tautu, editors, *Mathematics in Archeological and Historical Sciences*, pages 387–395. Edimburgh University Press, 1971.
3. V Makarenkov. T-REX: reconstructing and visualizing phylogenetic trees and reticulation networks. *Bioinformatics*, 17(7):664–668, 2001.
4. JB Whitfield, SA Cameron, DH Huson, and M Steel. Filtered Z-closure supernetworks for extracting and visualizing recurrent signal from incongruent gene trees. *Systematic Biology*, 57(6):939–947, 2008.
5. BR Holland, S Benthin, PJ Lockhart, V Moulton, and KT Huber. Using supernetworks to distinguish hybridization from lineage-sorting. *BMC Evolutionary Biology*, 8(202), 2008.
6. AWM Dress and DH Huson. Constructing splits graphs. *IEEE/ACM Transactions in Computational Biology and Bioinformatics*, 1(3):109–115, 2004.
7. BR Holland, F Delsuc, and V Moulton. Visualizing conflicting evolutionary hypotheses in large collections of trees: Using consensus networks to study the origins of placentals and hexapods. *Systematic Biology*, 54(1):66–76, 2005.
8. S Grünewald, K Forslund, AWM Dress, and V Moulton. QNet: An agglomerative method for the construction of phylogenetic networks from weighted quartets. *Molecular Biology and Evolution*, 24(2):532–538, 2007.
9. S Grünewald, A Spillner, K Forslund, and V Moulton. Constructing phylogenetic supernetworks from quartets. In *Proc* 8^{th} *Workshop Algorithms Bioinformatics (WABI'08)*, volume 5251 of *LNCS*, pages 284–295. Springer Verlag, 2008.
10. DH Huson, T Kloepper, PJ Lockhart, and M Steel. Reconstruction of reticulate networks from gene trees. In *Proc* 9^{th} *Annu Int Conf on Research in Computational Mol Biol (RECOMB'05)*, volume 3500 of *LNCS*, pages 233–249, 2005.
11. DH Huson and T Kloepper. Beyond galled trees - decomposition and computation of galled networks. In *Proc* 11^{th} *Annu Int Conf on Research in Computational Mol Biol (RECOMB'07)*, volume 4453 of *LNCS*, pages 211–225, 2007.
12. T Kloepper. *Algorithms for the Calculation and Visualisation of Phylogenetic Networks*. PhD thesis, Eberhard-Karls-Universität Tübingen, Germany, 2008.
13. ORP Bininda-Emonds, RMD Beck, and A Purvis. Getting to the roots of matrix representation. *Systematic Biology*, 54(4):668–672, 2005.
14. J-F Dufayard, L Duret, S Penel, M Gouy, F Rechenmann, and G Perrière. Tree pattern matching in phylogenetic trees: Automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics*, 21:2596–2603, 2005.

15. J Huerta-Cepas, S Capella-Gutierrez, LP Pryszcz, I Denisov, D Kormes, M Marcet-Houben, and T Gabaldón. PhylomeDB v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. *Nucleic Acids Research*, 39(supp. 1):D556–D560, 2011.
16. J Jansson and W-K Sung. Inferring a level-1 phylogenetic network from a dense set of rooted triplets. *Theoretical Computer Science*, 363(1):60–68, 2006.
17. Y-J He, TND Huynh, J Jansson, and W-K Sung. Inferring phylogenetic relationships avoiding forbidden rooted triplets. *Journal of Bioinformatics and Computational Biology*, 4(1):59–74, 2006.
18. J Jansson, NB Nguyen, and W-K Sung. Algorithms for combining rooted triplets into a galled phylogenetic network. *SIAM Journal on Computing*, 35(5):1098–1121, 2006.
19. L van Iersel, J Keijsper, S Kelk, L Stougie, F Hagen, and T Boekhout. Constructing level-2 phylogenetic networks from triplets. *IEEE/ACM Transactions in Computational Biology and Bioinformatics*, 6(4):667–681, 2009.
20. L van Iersel and S Kelk. Constructing the simplest possible phylogenetic network from triplets. *Algorithmica*, 60(2):207–235, 2011.
21. T-H To and M Habib. Level-k phylogenetic networks are constructable from a dense triplet set in polynomial time. In *Proc* 20^{th} *Annu Symp Combinatorial Pattern Matching (CPM'09)*, volume 5577 of *LNCS*, pages 275–288, 2009.
22. J Byrka, P Gawrychowski, KT Huber, and S Kelk. Worst-case optimal approximation algorithms for maximizing triplet consistency within phylogenetic networks. *Journal of Discrete Algorithms*, 8(1):65–75, 2010.
23. P Gambette and KT Huber. On encodings of phylogenetic networks of bounded level. *Journal of Mathematical Biology*, 2012. to appear.
24. AV Aho, Y Sagiv, TG Szymanski, and JD Ullman. Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. *SIAM Journal on Computing*, 10(3):405–421, 1981.
25. M Steel. The complexity of reconstructing trees from qualitative characters and subtrees. *Journal of Classification*, 9:91–116, 1992.
26. D Gusfield, V Bansal, V Bafna, and YS Song. A decomposition theory for phylogenetic networks and incompatible characters. *Journal of Computational Biology*, 14(10):1247–1272, 2007.
27. D Gusfield, S Eddhu, and C Langley. Optimal, efficient reconstruction of phylogenetic networks with constrained recombination. *Journal of Bioinformatics and Computational Biology*, 2(1):173–213, 2004.
28. L van Iersel, S Kelk, and M Mnich. Uniqueness, intractability and exact algorithms: reflections on level-k phylogenetic networks. *Journal of Bioinformatics and Computational Biology*, 7(4):597–623, 2009.
29. SM Woolley, D Posada, and KA Crandall. A comparison of phylogenetic network methods using computer simulation. *PLoS ONE*, 3(4):e1913, 2008.
30. U Brandes and S Cornelsen. Phylogenetic graph models beyond trees. *Discrete Applied Mathematics*, 157(10):2361–2369, 2010.
31. T Tholey. Improved algorithms for the 2-vertex disjoint paths problem. In *Proc* 35^th *Int Conf Current Trends in Theory and Practice of Computer Science (SOFSEM'09)*, volume 5404 of *LNCS*, pages 546–557, 2009.
32. J Opatrny. Total ordering problem. *SIAM Journal on Computing*, 8(1):111–114, 1979.
33. Stefan Grünewald, Vincent Moulton, and Andreas Spillner. Consistency of the QNet algorithm for generating planar split networks from weighted quartets. *Discrete Applied Mathematics*, 157:2325–2334, 2009.
34. V Berry and O Gascuel. Inferring evolutionary trees with strong combinatorial evi-

dence. *Theoretical Computer Science*, 240(2):271–298, 2000.
35. D Harel and RE Tarjan. Fast algorithms for finding nearest common ancestors. *SIAM Journal on Computing*, 13(2):338–355, 1984.
36. W Guttmann and M Maucher. Variations on an ordering theme with constraints. In *Proc* 4^{th} *IFIP Int Conf on Theor Comp Sci (TCS'06)*, volume 209 of *IFIP*, pages 77–90, 2006.
37. P Gambette, V Berry, and C Paul. The structure of level-k phylogenetic networks. In *Proc* 20^{th} *Annu Symp Combinatorial Pattern Matching (CPM'09)*, volume 5577 of *LNCS*, pages 289–300, 2009.
38. S Kelk, C Scornavacca, and L van Iersel. On the elusiveness of clusters. *IEEE/ACM Transactions in Computational Biology and Bioinformatics*, 2012. to appear.

Philippe Gambette received his PhD at Université Montpellier 2 in 2011. After a postdoctoral position at Institut de Mathématiques de Luminy, he is now an associate professor at Université Paris-Est Marne-la-Vallée. He works on graph algorithms applied to bioinformatics or natural language processing in the AlgoB research team of LIGM.

Vincent Berry received the PhD degree from the Université Montpellier, France, in 1997. After a postdoctoral position at the University of Warwick, United Kingdom, and an assistant professor position, he is now a full professor at Université Montpellier, performing research in the bioinformatics group of the LIRMM. His research interests are combinatorial algorithms and statistical methods in the field of phylogenetics.

Christophe Paul received the PhD degree in computer science from Montpellier University, France, in 1998. He is a full-time CNRS researcher. His research interests are mainly algorithms and graph theory.