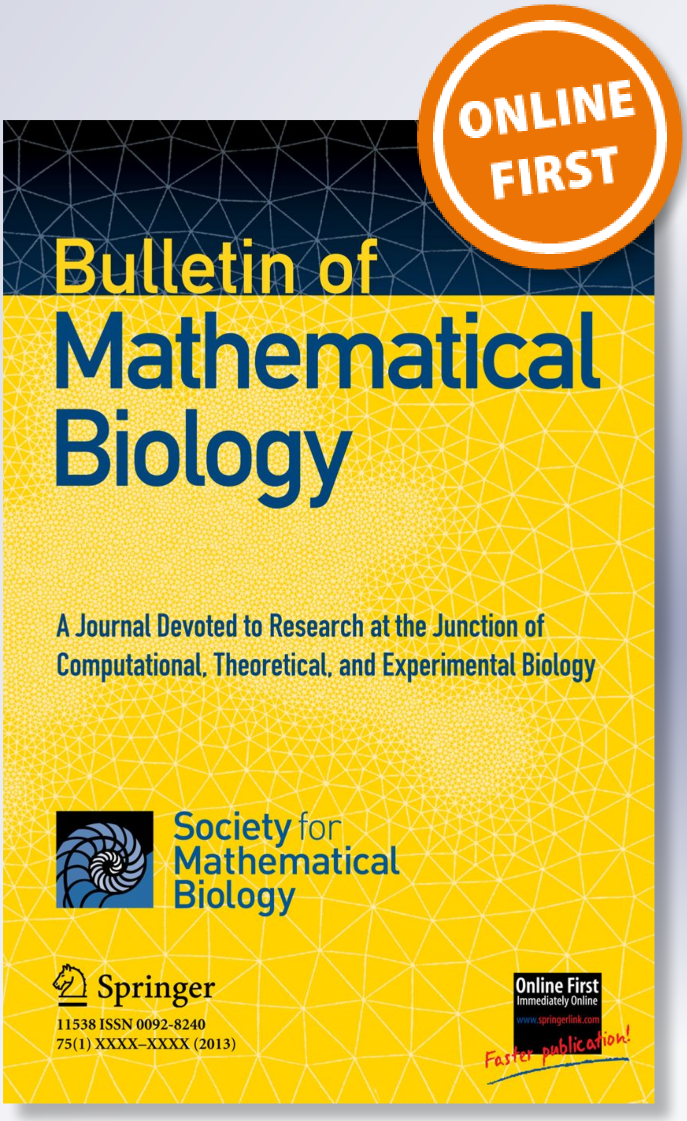
Identifying Species Network Features from Gene Tree Quartets Under the Coalescent Model

Hector Baños

Bulletin of Mathematical Biology
A Journal Devoted to Research at the Junction of Computational, Theoretical and Experimental Biology Official Journal of The Society for Mathematical Biology

ISSN 0092-8240

Bull Math Biol
DOI 10.1007/s11538-018-0485-4



ONLINE FIRST

Bulletin of Mathematical Biology

A Journal Devoted to Research at the Junction of Computational, Theoretical, and Experimental Biology

Society for Mathematical Biology

Springer
11538 ISSN 0092-8240
75(1) XXXX–XXXX (2013)

Online First
Immediately Online
www.springerlink.com
Faster publication!

Springer

Bulletin of Mathematical Biology
https://doi.org/10.1007/s11538-018-0485-4

Society for Mathematical Biology

SPECIAL ISSUE: ALGEBRAIC METHODS IN PHYLOGENETICS

Identifying Species Network Features from Gene Tree Quartets Under the Coalescent Model

Hector Baños[1]

Received: 27 November 2017 / Accepted: 30 July 2018

Abstract

We show that many topological features of level-1 species networks are identifiable from the distribution of the gene tree quartets under the network multi-species coalescent model. In particular, every cycle of size at least 4 and every hybrid node in a cycle of size at least 5 are identifiable. This is a step toward justifying the inference of such networks which was recently implemented by Solís-Lemus and Ané. We show additionally how to compute quartet concordance factors for a network in terms of simpler networks, and explore some circumstances in which cycles of size 3 and hybrid nodes in 4-cycles can be detected.

Keywords Coalescent theory · Phylogenetics · Networks · Concordance factors

1 Introduction

As phylogenetic analysis of DNA data has progressed, more evidence has appeared showing that hybridization is often an important factor in evolution. As surveyed in Nakhleh (2011), hybridization has played a very important role in the evolutionary history of plants, some groups of fish and frogs (Ellstrand et al. 1996; Linder and Rieseberg 2004; Mallet 2005; Noor and Feder 2006; Rieseberg et al. 2000). Other biological processes, such as introgression, lateral gene transfer and gene flow, also require moving beyond a simple treelike view of species relationships.

Phylogenetic networks are the objects used to represent the relationships between species that admit such events (Arnold 1997; Bapteste et al. 2013). These networks are often thought of as obtained from phylogenetic trees by adding additional edges,

This research was supported in part by the National Institutes of Health Grant R01 GM117590, awarded under the Joint DMS/NIGMS Initiative to Support Research at the Interface of the Biological and Mathematical Sciences.

✉ Hector Baños
hdbanoscervantes@alaska.edu

[1] University of Alaska Fairbanks, P.O. Box 756660, Fairbanks, AK 99775-6660, USA

Published online: 09 August 2018

so that some nodes in the tree have two parents. Nodes with two parents, called *hybrid nodes*, represent species whose genome arises from two different ancestral species. Inference of phylogenetics networks from biological data presents new challenges, with methods still being developed, as shown by recent works including Ané et al. (2007), Meng and Kubatko (2009), Solís-Lemus and Ané (2016), Zhang et al. (2018), Yu et al. (2014) and Yu et al. (2011).

Another challenge in inferring evolutionary history arises from the fact that many multi-locus data sets exhibit gene tree incongruence, even without suspected hybridization. One possible reason is incomplete lineage sorting (ILS), which is described in the tree setting by the multi-species coalescent model Pamilo and Nei (1988). See, for example, Carstens et al. (2007), Pollard et al. (2006), and Syring et al. (2005) where ILS is explained in the biological setting.

Meng and Kubatko (2009) formulated a model of gene tree production, based on the multi-species coalescent model, incorporating both hybridization and ILS. We refer to this model as the *network multi-species coalescent model*, which is further developed in Yu et al. (2012), Zhu et al. (2016), and Solís-Lemus et al. (2016), to mention some. The model determines the probability of observing any rooted gene tree given a metric rooted phylogenetic species network.

Solís-Lemus and Ané (2016) recently presented a novel statistical method, based on the network multi-species coalescent model, to infer phylogenetic networks from gene tree quartets in a pseudolikelihood framework. The quartets themselves might come from larger gene trees inferred by standard phylogenetic methods. The pseudolikelihood in this work is built on quartet frequencies, or concordance factors, extending an idea of Liu et al. (2010) from the tree setting. The pseudolikelihood approach is simpler and faster than computing the full likelihood and makes large-scale data analysis more tractable. They demonstrate positive results in reconstructing the evolutionary relationships among swordtails and platyfishes.

However, the theoretical underpinnings of the method of Solís-Lemus and Ané (2016) are not complete. In using a model for statistical inference it is important to know whether it is theoretically possible to uniquely recover the parameters from the data the model predicts. In more precise terms, for model-based statistical inference to have a solid basis, we need that the probability distribution for data which arises under the model uniquely determines the parameters. This is known as *identifiability* of the model parameters.

While Solís-Lemus and Ané (2016) showed that any particular hybridization in a level-1 network with h hybridizations and n taxa can be generically detected under certain assumptions, their study never addressed the full identifiability of the network topology, only the detectability of a specific hybridization event. Working in the setting of level-1 networks, which is also adopted here, their arguments do not include investigations on network properties such as cycle sizes, and the structure of the whole network. These properties are crucial to determine, for example, whether two networks with different cycle sizes, or different number of cycles, could produce the same set of gene tree quartet probabilities.

The primary purpose of this work is to begin to address some of these identifiability questions raised in Solís-Lemus and Ané (2016). That is, we study the question: given information on gene quartet probabilities for some unknown level-1 network $\mathcal{N}$, what can be determined about the topology of $\mathcal{N}$?

Although others have considered the problem of constructing large networks from small ones, these works do not seem to be applicable to the question studied here. Most of these works, including Huber et al. (2017a, b) and Keijsper and Pendavingh (2014), are primarily combinatorial in nature. In particular, these studies do not address semidirected networks, ILS through the network multi-species coalescent model, nor the types of inputs that might be obtained from biological data.

The main result of this work, Theorem 4 of Section 8, is that under the network multi-species coalescent model on level-1 networks, we can generically identify from gene quartet distributions "most" of the unrooted topological network, including all cycles of size at least 4, and hybrid nodes in the cycles of size greater than 4. "Generically" here means for all values of numerical parameters except those in a set of measure zero. The methods used are a mix of the semialgebraic study of quartet gene tree frequencies (in terms of linear equalities and inequalities they satisfy) with combinatorial approaches to combining this knowledge for many quartets. As a side benefit the proofs suggest combinatorial methods for reconstructing networks, as opposed to just showing identifiability. However, we do not explore how such methods might be implemented in the presence of the noise that any collection of inferred gene trees will have.

Another result of this work, in Sect. 5, is a rigorous derivation of how gene quartet probabilities can be computed for large networks under the coalescent model. Although this parallels some of the results in Solís-Lemus and Ané (2016), the arguments given here are more rigorous, as is necessary for them to form the basis of our main results. Our approach is to express quartet frequencies as convex combinations of those on simplified networks, ultimately leading to expressions in terms of trees, as is done in other situations Zhu and Degnan (2017). This is different from the approach in Solís-Lemus and Ané (2016) of finding networks with less hybridizations displaying the same gene quartet probabilities.

The outline of this work is as follows: Sect. 2 introduces basic definitions and establishes some terminology on graphs and networks. Section 3 sets forth insights and tools for studying the structure of level-1 networks. Section 4 reviews the network multi-species coalescent model of Meng and Kubatko (2009), as well as quartet concordance factors and some of their properties. In Sect. 5 we show how concordance factors of quartet networks can be expressed in terms of simpler networks. Section 6 introduces the "Cycle property" of concordance factors, and Sect. 7 defines the "Big Cycle" property of concordance factors. In Sect. 8, the main result on topological network identifiability is proved using the Big Cycle property, and in Sect. 9 some extended results on the "Cycle property" are shown.

2 Phylogenetic Networks

We adopt standard terminology for graphs and networks, as used in phylogenetics; see, for example, Semple and Steel (2005) and Steel (2016). All undirected, directed, or semidirected graphs will not contain loops. If G is a directed or semidirected graph, the *undirected graph of* G, denoted by $U(G)$, is the graph G with all directions omitted.

2.1 Rooted Networks

To set terminology, we begin with some fundamental definitions.

Definition 1 A *topological binary rooted phylogenetic network* $\mathcal{N}^+$ on taxon set X is a connected directed acyclic graph with vertices V and edges E, where V is the disjoint union $V = \{r\} \sqcup V_L \sqcup V_H \sqcup V_T$ and E is the disjoint union $E = E_H \sqcup E_T$, and a bijective leaf-labeling function $f : V_L \to X$ with the following characteristics:

1. The *root* r has indegree 0 and outdegree 2.
2. A *leaf* $v \in V_L$ has indegree 1 and outdegree 0.
3. A *tree node* $v \in V_T$ has indegree 1 and outdegree 2.
4. A *hybrid node* $v \in V_H$ has indegree 2 and outdegree 1.
5. A *hybrid edge* $e \in E_H$ is an edge whose child is a hybrid node.
6. A *tree edge* $e \in E_T$ is an edge whose child is a tree node or a leaf.

Definition 2 Let $\mathcal{N}^+$ be a topological binary rooted phylogenetic network with $|E| = m$ and $|E_H| = 2h$. A *metric for* $\mathcal{N}^+$ is a pair (λ, γ), where $\lambda : E \to \mathbb{R}_{>0}$ and $\gamma : E_H \to (0, 1)$ satisfies that if two edges h_1 and h_2 have the same hybrid node as child, then $\gamma(h_1) + \gamma(h_2) = 1$.

If (λ, γ) is a metric for $\mathcal{N}^+$, then we refer to $(\mathcal{N}^+, (\lambda, \gamma))$ as a *metric binary rooted phylogenetic network*.

Note that Definition 1 differs from that of Steel (2016) in that it allows up to two edges between a pair of nodes. An edge weight $\lambda(e)$ is interpreted as the time (in coalescent units) between speciation events represented by the ends of edge e. For any hybrid edge h with child v, the value $\gamma(h) = \gamma_h$ is the probability that a lineage at v has ancestral lineage in h and is often called *hybridization parameter or inheritance probability*. Since we are focusing on parameter identifiability, we will use the term hybridization parameter.

2.2 Lowest Stable Ancestor

We review and show some properties of the lowest stable ancestor, a network analog of the most recent common ancestor on a tree.

Definition 3 Let $\mathcal{N}^+$ be a (metric or topological) binary rooted phylogenetic network. We say that a node v is *above* a node u, and u is *below* v, if there exists a non-empty directed path in $\mathcal{N}^+$ from v to u. We also say that an edge with parent node x and child y is above (below) a node v if y is above or equal to v (x is below or equal to v).

Note that since $\mathcal{N}^+$ has no directed cycles, u cannot be both above and below v.

Definition 4 Steel (2016) Let $\mathcal{N}^+$ be a (metric or topological) binary rooted phylogenetic network on X and let $Z \subseteq X$. Let D be the set of nodes which lie on every

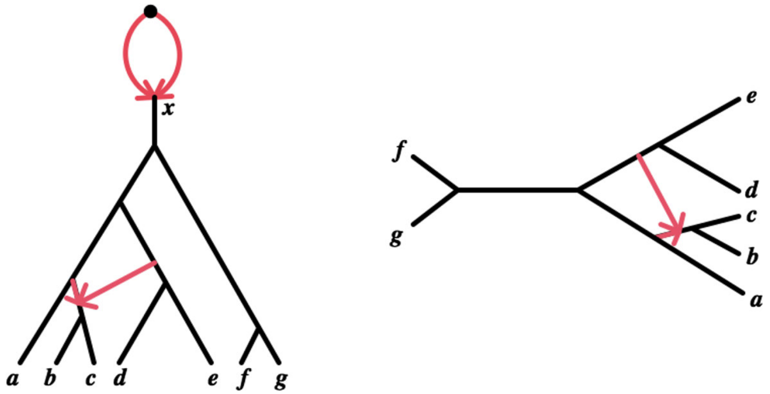


Fig. 1 (Left) A binary rooted phylogenetic network on X, with LSA(X) the node labeled x, and (Right) its induced unrooted semidirected network. In a depiction of a rooted network, all edges are directed downward, from the root, but arrowheads are shown only on hybrid edges. For the unrooted network, all edges except hybrid ones are undirected

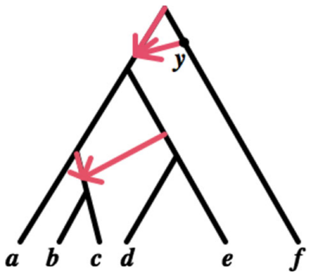


Fig. 2 A binary rooted phylogenetic network where the node labeled y is ancestral to all taxa in X but is not LSA(X). LSA(X) here is the root of the network

directed path from the root r of $\mathcal{N}^+$ to any $z \in Z$. Then the lowest stable ancestor of Z of $\mathcal{N}^+$*, denoted by* $LSA(Z, \mathcal{N}^+)$*, is the unique node $v \in D$ such that v is below all $u \in D$, $u \neq v$.*

When $\mathcal{N}^+$ is clear from context, we write LSA(Z) for LSA$(Z, \mathcal{N}^+)$. To see that LSA(Z) is well defined for any $Z \subseteq X$, note first that $D \neq \emptyset$ since $r \in D$. Also, since every pair of nodes $u, v \in D$ both lie on a path, we have a notion of above and below for u and v, i.e., a total order on D, and hence a minimal element.

While the definition of LSA agrees with the most recent common ancestor for trees, it is more subtle. In particular, if $\mathcal{N}^+$ is a network on X, LSA(X) need not to be the root of the network, as Fig. 1 (left) shows. Furthermore, there can be nodes below LSA(X) which are ancestral to all of X, as Fig. 2 shows.

Lemma 1 *Let $\mathcal{N}^+$ be a (metric or topological) binary rooted phylogenetic network on X with root r, and let $Z \subseteq Y \subseteq X$. Then*

(i) *the indegree of LSA(Z) is at most one for any $Z \subset X$;*
(ii) *at most one of the out edges of LSA(Z) is hybrid;*
(iii) *if $Z \subseteq Y \subseteq X$ then LSA(Z) is below or equal to LSA(Y).*

Proof To see (i), suppose that the indegree of LSA(Z) is two. Then the outdegree would be one, and the child of LSA(Z) would be in any path from the root to any taxa in Z, contradicting the definition of LSA(Z).

For (ii), suppose the out edges of LSA(Z), e_1 and e_2, are both hybrid. If e_1 and e_2 have the same child, then every path from r to any $z \in Z$ would contain that node, contradicting the definition of LSA(Z).

Now denote by $x_1 \neq x_2$ the child nodes of e_1 and e_2, respectively. If both x_1 and x_2 had parents below LSA(Z), then x_1 has a parent below x_2 and x_2 has a parent below x_1 giving a directed cycle. Thus, without loss of generality, assume x_1 has parents LSA(Z) and v with v not below LSA(Z). Let $z \in Z$ with z below x_1. If we remove the LSA(Z) from $\mathcal{N}^+$ there is still a path from r to z (which goes from r to v to x_1 to z). This contradicts the fact that LSA(Z) is on all paths from r to any $z \in Z$.

For (iii) we observe that since $Z \subseteq Y$, LSA(Y) must be equal or above LSA(Z) since the set of paths from r to any taxa in Y contains the set of paths from r to any taxon in Z. □

Lemma 2 *Let $\mathcal{N}^+$ be a (metric or topological) binary rooted phylogenetic network on X and let $Z \subset X$, $|Z| \geq 2$. For every $x \in Z$, there is a $y \in Z$ such that $LSA(x, y) = LSA(Z)$.*

Proof Let $m =$ LSA(Z), fix $x \in Z$ and let P be a path from m to x. By definition of LSA, for all $y \in Z$, LSA(x, y) is a node in P and is below or equal to m by Lemma 1. Suppose that LSA(x, y) is below m for all $y \in Z$. Let $z \in Z$ be such that LSA(x, z) is above or equal to LSA(x, y) for all $y \in Z \setminus \{z\}$.

We claim that any path from m to $y \in Z$ passes through LSA(x, z). Suppose there exists taxon y with path P' from m to y that does not pass through LSA(x, z). But P' must pass through LSA(x, y). Since LSA(x, y) is below LSA(x, z), there is a path from m to LSA(x, y) to x that does not contain LSA(x, z). This is a contradiction.

But every path from m to any $y \in Z$ passes through LSA(x, z), contradicting that LSA(x, z) is below m. □

By this lemma we can characterize LSA(Z) as the highest node of the form LSA(x, y) for some $x, y \in Z$, or the highest node of that form for fixed $x \in Z$.

2.3 Unrooted Networks

Let G be a directed or semidirected graph with z a degree two node. Let x and y be the two nodes adjacent to z. Then, up to isomorphism, the subgraph on x, y and z must be one of the graphs shown on the left of Fig. 3, which we denote by H. By *suppressing* z we mean replacing H in G by the graph to the right of it in Fig. 3.

Definition 5 Let $\mathcal{N}^+$ be a binary topological rooted phylogenetic network on a set of taxa X. Then $\mathcal{N}^-$ is the semidirected network obtained by (1) keeping only the edges and nodes below LSA(X); (2) removing the direction of all tree edges; (3) suppressing LSA(X). We refer to $\mathcal{N}^-$ as the *topological unrooted semidirected network induced from $\mathcal{N}^+$*.

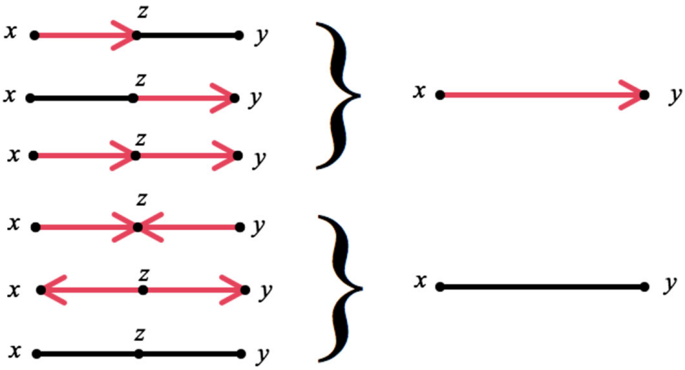


Fig. 3 On the left are all the semidirected graphs, up to isomorphism, on a degree two node z and its adjacent vertices x and y. On the right are the corresponding graphs obtained by suppressing z

Figure 1 shows an example of a network $\mathcal{N}^+$ and its induced $\mathcal{N}^-$. We now introduce a metric on $\mathcal{N}^-$ induced from one on $\mathcal{N}^+$.

Definition 6 Let $(\mathcal{N}^+, (\lambda, \gamma))$ be a metric binary rooted phylogenetic network and let $\mathcal{N}^-$ be the topological unrooted semidirected network induced from $\mathcal{N}^+$. Denote by e^* the edge of $\mathcal{N}^-$ introduced in place of the edges e_1 and e_2 in $\mathcal{N}^+$ when $\mathrm{LSA}(X)$ is suppressed. Define $\lambda' : E(\mathcal{N}^-) \to \mathbb{R}_{>0}$ such that $\lambda'(e^*) = \lambda(e_1) + \lambda(e_2)$ and $\lambda'(e) = \lambda(e)$ for $e \in \mathcal{N}^-$, $e \neq e^*$. If e^* is not hybrid, $\gamma' = \gamma$, else let $\gamma'(h) = \gamma(h)$ for all hybrid edges of $\mathcal{N}^-$ other than e^* and $\gamma'(e^*) = \gamma(e_i)$, where e_i is, by Lemma 1, the single hybrid edge in $\{e_1, e_2\}$. We refer to $(\mathcal{N}^-, (\lambda', \gamma'))$ as the *metric unrooted semidirected network induced from* $(\mathcal{N}^+, (\lambda, \gamma))$.

The networks considered in this work are always induced from a rooted binary metric phylogenetic network. To simplify language, we refer to a (metric or topological) binary rooted phylogenetic network as a *(metric or topological) rooted network* and to a induced (metric or topological) unrooted semidirected phylogenetic network as a *(metric or topological) unrooted network*.

We note that not all binary semidirected graphs are topological unrooted networks, since some graphs are not compatible with suppressing the root on any rooted network. Moreover, $\mathcal{N}^-$ might be induced from several rooted networks $\mathcal{N}^+$. See Fig. 4.

Although an unrooted network $\mathcal{N}^-$ does not have a root specified, since hybrid edges are directed, the suppressed $\mathrm{LSA}(X)$ of $\mathcal{N}^+$ must have been located 'above' them. Thus, in $\mathcal{N}^-$, we still have a well-defined notion of which taxa are descendants of a hybrid node v. These are the taxa x such that there exists a semidirected path from v to x in $\mathcal{N}^-$. In this case we say that x *descends from* v.

2.4 Induced Networks on Subset of Taxa

Since later arguments require an understanding of the behavior of the network multispecies coalescent model on a subset of taxa, we introduce some needed definitions.

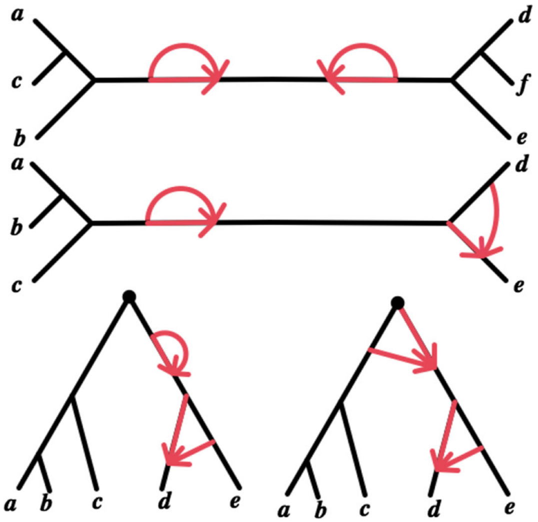


Fig. 4 The top graph is not a topological unrooted semidirected phylogenetic network, since its directed edges cannot be obtained by suppressing the root of any 6-taxon topological binary rooted phylogenetic network. The middle graph is the induced topological unrooted network from either of the bottom rooted networks, as well as others

Definition 7 Let $\mathcal{N}^+$ be a (metric or topological) rooted network on X and let $Z \subset X$. The *induced rooted network* $\mathcal{N}^+_Z$ *on* Z is the network obtained from $\mathcal{N}^+$ by (1) retaining only edges and nodes in paths from the root to any taxa in Z; (2) suppressing all degree two nodes except the root; (3) in the case the root then has outdegree one, contracting the edge incident to the root.

Note that $\mathrm{LSA}(Z, \mathcal{N}^+_Z) = \mathrm{LSA}(Z, \mathcal{N}^+)$. If $|Z| = 4$ then $\mathcal{N}^+_Z$, the *induced rooted quartet network on* Z, will also be denoted by $\mathcal{Q}^+_Z$ to emphasize it involves only 4 taxa.

Definition 8 Let $\mathcal{N}^+$ be a (metric or topological) rooted network on X and let $Z \subset X$. The *induced LSA network of* Z, denoted $\mathcal{N}^\oplus_Z$, is the rooted network obtained from $\mathcal{N}^+_Z$ by deleting everything above $\mathrm{LSA}(Z, \mathcal{N}^+)$.

In particular, we note that $\mathcal{N}^\oplus_Z$ has root $\mathrm{LSA}(Z, \mathcal{N}^+)$. If $|Z| = 4$ then $\mathcal{N}^\oplus_Z$, the *induced LSA quartet network on* Z, is also denoted by $\mathcal{Q}^\oplus_Z$.

Definition 9 Let G be a semidirected graph and let x, y be two nodes in G. A *trek* in G from x to y is an ordered pair of semidirected paths (P_1, P_2) where P_1 has terminal node x, P_2 has terminal node y, and both P_1 and P_2 have starting node v. The node v is called the *top* of the trek, denoted $\mathrm{top}(P_1, P_2)$. A trek (P_1, P_2) is *simple* if the only common node among P_1 and P_2 is v.

This definition is adopted from non-phylogenetic studies of statistical models on graphs, such as Sullivant et al. (2010).

Definition 10 Let $\mathcal{N}^-$ be a (metric or topological) unrooted network on X and let $Z \subseteq X$. The *induced unrooted network* $(\mathcal{N}^-)_Z$ *on a set of taxa* Z is the network obtained from $\mathcal{N}^-$ by retaining only edges in simple treks between pairs of taxa in Z, and then suppressing all degree two nodes.

Note that it is not immediately clear that for a network $\mathcal{N}^+$, the networks $(\mathcal{N}^-)_Z$ and $(\mathcal{N}^+_Z)^-$ are isomorphic. Proposition 1 shows that the operations of unrooting and inducing a network on a subset of taxa commute. While this statement is intuitively plausible, its rather technical proof is in "Appendix."

Proposition 1 *Let $\mathcal{N}^+$ be a (metric or topological) rooted network on X and let $Z \subseteq X$. Then $(\mathcal{N}^-)_Z$ and $(\mathcal{N}^+_Z)^-$ are isomorphic.*

If $|Z| = 4$ then $(\mathcal{N}^-)_Z$, the *induced unrooted quartet network on Z*, is also denoted by $\mathcal{Q}^-_Z$.

2.5 Cycles

Although the networks $\mathcal{N}^+$, $\mathcal{N}^-$ are acyclic (in both, the directed and semidirected settings), their undirected graphs $U(\mathcal{N}^+)$, $U(\mathcal{N}^-)$ may contain a cycle. Thus, the term 'cycle' may be used to unambiguously refer to cycles in the undirected graphs. We formalize this with the following definition:

Definition 11 Let $\mathcal{N}$ be a (metric or topological, rooted or unrooted) network. A *cycle* in $\mathcal{N}$ is a non-empty path from a node to itself, allowing edges to be traversed without regard to their possible direction. The *size* of the cycle is the number of edges in the path. A *k-cycle* is a cycle of size k.

By *contracting or shrinking* a cycle C in a graph we mean removing all edges in C and identifying all nodes in C.

3 Structure of level-1 Networks

The class of all phylogenetic networks is often too large to obtain strong mathematical results (Steel 2016), so it is common to restrict to networks that have a simpler structure, for instance, the class of *level-1* phylogenetic networks.

Definition 12 Let $\mathcal{N}$ be a (rooted or unrooted) topological network. If no two cycles in $\mathcal{N}$ share an edge, then $\mathcal{N}$ is *level-1*.

If $\mathcal{N}$ is a level-1 network, any subnetwork or induced network of $\mathcal{N}$ is also level-1.

Given a hybrid node v, denote the hybrid edges whose child is v by h_v and h'_v. Then h_v and h'_v are called the *hybrid edges of v*.

Lemma 3 *Let $\mathcal{N}$ be a (topological or metric, rooted or unrooted) level-1 network and let C be a cycle of $\mathcal{N}$. Then C contains exactly one hybrid node v, and the associated hybrid edges h_v, h'_v. Furthermore, each node of $\mathcal{N}$ is in at most one cycle and, as a result, v, h_v and h'_v are in exactly one cycle of $\mathcal{N}$.*

The proof of each statement of this lemma, using different terminology, is given by Rosselló and Valiente (2009).

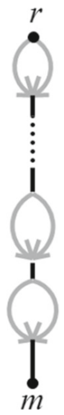


Fig. 5 In a level-1 network on X, the structure between the root and $m = \mathrm{LSA}(X)$ is a chain of two cycles. The number of two cycles in the chain could be zero

Proposition 2 *Let $\mathcal{N}^+$ be a topological level-1 rooted network on X. The structure of all the nodes and edges above LSA(X) in $\mathcal{N}^+$ is a (possibly empty) chain of 2-cycles connected by edges, as depicted in Fig. 5.*

Proof Let $m = \mathrm{LSA}(X)$, and denote by r the root of $\mathcal{N}^+$. The proof is by induction on the number of the edges above m. If there are no edges above m, then $m = r$ and the result is trivially true. By Lemma 1, one easily sees that there cannot be only 1 or 2 edges above m in a binary phylogenetic network. That is, if there were just 1 edge above m the outdegree of the root would be 1, contradicting the definition of binary phylogenetic network. Suppose there are 2 edges above m. By definition of binary phylogenetic network the outdegree of r is 2 and by definition of $\mathrm{LSA}(X)$ all paths from the root to $x \in X$ contain m. Therefore, m has indegree 2, contradicting Lemma 1 part (i).

Now assume the claim holds when there are at most k edges above m and suppose there are $k + 1$ edges above m. Note that r has outdegree 2 by the definition of $\mathcal{N}^+$.

Suppose that edges incident to r have different children, x and y. Note neither x nor y can be m. The outdegree of one of x or y must be 2, otherwise both would be hybrid nodes, which would require x above y and y above x. Without loss of generality suppose x has outdegree 2, and denote by e_1 and e_2 its out edges, and denote by e_3 the edge (r, y). Since every path from r to a leaf goes through m, there are at least 3 distinct paths P_1, P_2, P_3 from r to m, where P_i contains e_i.

This contradicts the level-1 condition. Thus, $x = y$, and the edges from r form a 2-cycle.

Now since x is a hybrid node, it has outdegree 1, with child v. Also, there are $k - 3$ edges above m that are also below v. Applying the inductive hypothesis to $\mathcal{N}^+$ with edges above v removed, the result follows. □

Proposition 2 applied to $\mathcal{N}_Z^+$ illustrates the structure of the common ancestry of a subset Z of taxa. When we pass to a LSA network or an induced unrooted network, we "throw away" this structure. We show in Sect. 5 that under the network multi-species coalescent model this structure has no effect on the formation of quartet gene trees.

Let v be a hybrid node in a level-1 (rooted or unrooted, metric or topological) network $\mathcal{N}$ on X and let C_v be the cycle containing v. By removing the edges of C_v

from $\mathcal{N}$ we obtain a partition of X according to the connected components of the resulting graph. We refer to this partition as the *v-partition* and its partition sets as *v-blocks*.

Note that each node in C_v can be associated with a v-block. That is, a v-block B_u is associated with a node u in C_v if by removing u from the network (and therefore the edges adjacent to u), the induced partition of taxa is $\{B_u, X \setminus B_u\}$. We refer to the v-block B_v, whose elements descend from v, as the *v-hybrid block*. Two distinct v-blocks B_u, B_w are *adjacent* if the nodes $u, w \in C_v$ are adjacent.

Let $\mathcal{D} = \{C_1, \ldots, C_n\}$ be a collection of cycles in $\mathcal{N}$. The partition of X obtained by removing all the edges in the cycles of $\mathcal{D}$ is the *network partition induced by $\mathcal{D}$* and its blocks are *network blocks induced by $\mathcal{D}$*. When $\mathcal{D}$ is the set of all cycles in $\mathcal{N}$ of size at least k, the partition is the *k-network partition* and its blocks are *k-network blocks*. The 4-network blocks play an important role in Sect. 8. For now and on, we will refer to removing all edges of a cycle C from a network $\mathcal{N}$ as *removing the cycle C from $\mathcal{N}$*.

The following is straightforward to prove.

Lemma 4 *Let $\mathcal{N}$ be a level-1 (rooted or unrooted) topological network on X. Let $\mathcal{D} = \{C_1, \ldots, C_n\}$ be a collection of cycles in $\mathcal{N}$. For any two taxa a and b in different network blocks induced by $\mathcal{D}$, there exists a hybrid node v of some cycle in $\mathcal{D}$ such that a and b are in different v-blocks.*

If two taxa a and b are in the same network block induced by $\mathcal{D}$, then they are connected when all cycles in $\mathcal{D}$ are removed. As a result they are connected when a single cycle in $\mathcal{D}$ is removed. This comment together with Lemma 4 yields the following.

Corollary 1 *Let $\mathcal{N}$ be a level-1 (rooted or unrooted) topological network on X. Let $\mathcal{D} = \{C_1, \ldots, C_n\}$ be a collection of cycles in $\mathcal{N}$, with v_i the hybrid node associated with C_i. The network partition induced by $\mathcal{D}$ is the common refinement of the v_i-partitions for $1 \leq i \leq n$.*

Since contracting cycles in level-1 networks does not introduce loops or multi-edges, we can define a notion of a tree of cycles which is useful for the proof of Theorem 4.

Definition 13 Let $\mathcal{N}^-$ be a topological unrooted level-1 network. Let $\mathcal{T}$ be the graph obtained from $\mathcal{N}^-$ by (1) removing all pendant edges, repeatedly, until no pendant edges remain; (2) suppressing all vertices of degree two that are not part of a cycle; (3) contracting each cycle in the network obtained from steps 1 and 2. We refer to $\mathcal{T}$ as the *tree of cycles of $\mathcal{N}^-$*.

In the tree of cycles of $\mathcal{N}^-$ certain nodes, including all the leaves, represent a cycle of the original network $\mathcal{N}^-$. The notion of tree of cycles is different from "tree of blobs" of Gusfield et al. (2007), as there is no deletion of the non-cycle edges in the tree of blobs. In Fig. 6 we see an example of a tree of cycles.

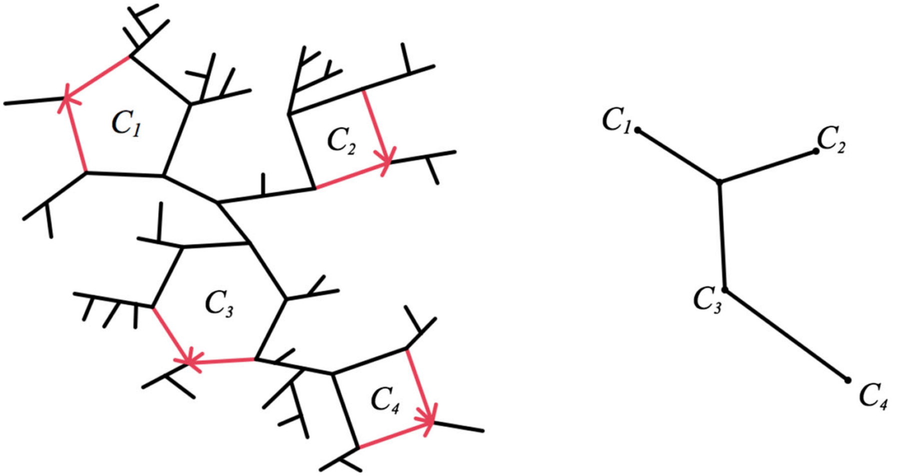


Fig. 6 (Left) A level-1 unrooted network $\mathcal{N}^-$ and (Right) the tree of cycles of $\mathcal{N}^-$

4 The Network Multi-Species Coalescent Model and Quartet Concordance Factors

Coalescent theory models the formation of gene trees within populations of species. The coalescent model for a single population traces (backward in time) the ancestries of a finite set of individual copies of a gene as the lineages *coalesce* to form ancestral lineages (see Wakeley 2008). The *multi-species coalescent* (*MSC*) *model* is a generalization of the coalescent model, formulated by applying it to multiple populations connected to form a rooted population tree, or species tree. It is commonly used to obtain the probabilities of gene trees in the presence of incomplete lineage sorting.

Meng and Kubatko (2009) extended the MSC by introducing phenomena such as hybridization or other horizontal gene transfer across the species-level and Nakhleh et al. further developed it Yu et al. (2012); Zhu et al. (2016). This model describes any situation in which a gene lineage may "jump" from one population to another at a specific time. The model parameters are specified by a metric binary rooted phylogenetic network as defined in Sect. 2. Different from models such as the structured coalescent with continuous gene flow (see Wakeley 2008), the network model approach assumes the gene transfer occurs at a single point in time along hybrid edges. We refer to this extended version of the MSC as the *network multi-species coalescent (NMSC) model*.

The NMSC model assumes that speciation by hybridization results in what Meng and Kubatko refer to as a mosaic genome. One assumption of the NMSC model, inherited from the MSC model, is that all gene lineages present at a specific point on the species tree behave identically above this point. That is, the probability of any event conditioned on a set of lineages being present at a certain point on the species tree is invariant under permutation of those lineages. This feature is known as the *exchangeability* property.

Example 1 We illustrate how to compute the probability of a gene tree topology under the NMSC with an example. Suppose we have the rooted metric species network given in Fig. 7. Let A, B, C and D be genes sampled from species a, b, c and d, respectively. We compute the probability that a gene tree has the unrooted topology $((A, B), (C, D))$ under the NMSC model.

First observe that until B and C trace back to the edge with length z there cannot be a coalescent event. In that edge these lineages cannot coalesce if the gene tree $((A, B), (C, D))$ is to be formed. The probability of no coalescence on this edge is e^{-z}. Now there are 4 cases, illustrated in Fig. 8:

1. with probability γ^2, lineages B and C enter the edge of length w;
   A.
2. with probability $(1-\gamma)^2$, B and C enter the edge of length v;
   D.
3. with probability $\gamma(1-\gamma)$, B enters the edge of length w and C enters the edge of length v;
   with the edge with lineage A and C enter the edge that joins with the edge with lineage D.
4. with probability $(1-\gamma)\gamma$, B enters the edge of length v and C enters the edge of length w.

Observe that each case is now reduced to a standard MSC scenario with several samples per population (see Degnan 2010). Let P_i the probability of observing $((A, B), (C, D))$ under the MSC of case i. Then the probability of observing $((A, B), (C, D))$ is $e^{-z}(\gamma^2 P_1 + (1-\gamma)^2 P_2 + \gamma(1-\gamma)P_3 + \gamma(1-\gamma)P_4)$.

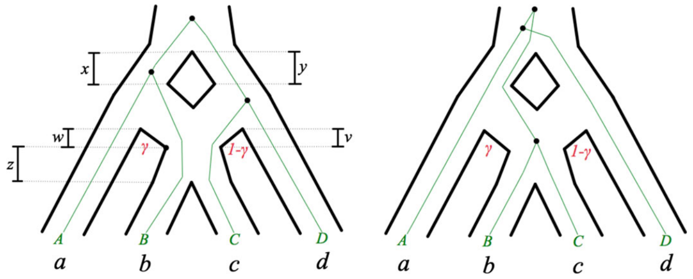


Fig. 7 Two gene trees within a species network with one hybrid node

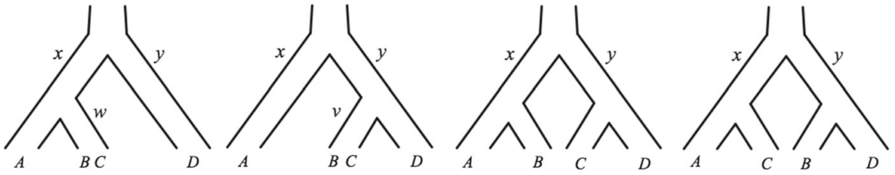


Fig. 8 Cases 1-4 (Left-Right) of Example 1, of how lineages may behave under the NMSC model on the network of Fig. 7

Following Solís-Lemus and Ané (2016), we are interested in the probability that a species network produces various gene quartets under the NMSC. This motivates the following definition.

Definition 14 Let $\mathcal{N}^+$ be a metric rooted network on a taxon set X. Let A, B, C, D be genes sampled from species a, b, c, d, respectively. Given a gene quartet $AB|CD$, the *quartet concordance factor* $CF_{AB|CD}$ is the probability under the NMSC on $\mathcal{N}^+$ that a gene tree displays the quartet $AB|CD$, and

$$CF_{abcd} = (CF_{AB|CD}, CF_{AC|BD}, CF_{AD|BC})$$

is the ordered triple of concordance factors of each quartet on the taxa a, b, c, d.

When a, b, c, d are clear from context, we write CF for CF_{abcd}.

In the particular case where $\mathcal{N}^+$ has no hybrid edges, so the network is a tree, it is known that the quartet concordance factors do not depend on the root placement Allman et al. (2011). For example, let a, b, c, d be taxa and consider any root placement in the unrooted species tree with topology $ab|cd$ and internal edge of length t. Then

$$CF_{abcd} = \left(1 - \frac{2}{3}e^{-t}, \frac{1}{3}e^{-t}, \frac{1}{3}e^{-t}\right). \tag{1}$$

As mentioned in Solís-Lemus and Ané (2016), for unrooted species networks the concordance factors do not depend on the placement of the root in the species network, as long as the root is placed in a way consistent with the direction of the hybrid edges. This fact is shown in Sect. 5, as we explore quartet concordance factors more thoroughly.

Definition 15 Let $\mathcal{N}^+$ be a metric rooted level-1 network on X. Given a set of distinct taxa $\{a, b, c, d\}$, we define the *ordering of* CF_{abcd} *on* $\mathcal{N}^+$ as the natural decreasing order of $CF_{AB|CD}$, $CF_{AC|BD}$, $CF_{AD|BC}$ in the real line.

For example, if $t > 0$ the ordering of the concordance factors in Eq. (1) is given by

$$CF_{AB|CD} > CF_{AC|BD} = CF_{AD|BC}.$$

Many arguments toward the main result of this work use the ordering of CF_{abcd}, and not its precise values.

5 Computing Quartet Concordance Factors

In this section we show how to express the concordance factors arising on a LSA quartet network as a linear combination of the concordance factors arising on quartet trees using a similar approach as in Yu et al. (2014). This enables us to see how the ordering of concordance factors reflects the network topology, and how the precise root location does not matter.

The final results of this section are largely in Solís-Lemus and Ané (2016). However, we provide formal arguments and take in consideration some matters that were left unaddressed. For example, we address the possibility that an induced 4-taxon network does not contain the root of the original network.

Let $\mathcal{N}^+$ be a (metric or topological) rooted level-1 network on X and let $\{a, b, c, d\}$ be a set of distinct taxa of X. Then the induced unrooted network on 4 taxa $\mathcal{Q}^-_{abcd}$ is a (metric or topological) unrooted level-1 network. By Proposition 1, $\mathcal{Q}^-_{abcd}$ is the same graph as $(\mathcal{N}^+_{abcd})^-$ and $(\mathcal{N}^\oplus_{abcd})^-$, where $\mathcal{N}^\oplus_{abcd}$ is the LSA network of Definition 8. Any cycle in $\mathcal{N}^\oplus_{abcd} = \mathcal{Q}^\oplus_{abcd}$ induces a cycle in $\mathcal{Q}^-_{abcd}$. A cycle C in $\mathcal{Q}^\oplus_{abcd}$ of size k, induces a cycle in $\mathcal{Q}^-_{abcd}$ of either size k (when C does not contain LSA(a, b, c, d)) or size $k - 1$ (otherwise). For convenience when we refer to the size of a cycle C in $\mathcal{Q}^\oplus_{abcd}$ we mean the size of the induced cycle in $\mathcal{Q}^-_{abcd}$.

Lemma 5 *Let $\mathcal{Q}^-_{abcd}$ be a metric unrooted level-1 quartet network. The number of k-cycles in $\mathcal{Q}^-_{abcd}$ is* 0 *for $k \geq 5$, at most* 1 *for $k = 4$ in which case there is no* 3*-cycle, and at most* 2 *for $k = 3$.*

Proof Suppose that $\mathcal{Q}^-_{abcd}$ has a cycle $C = C_v$ of size k. Then there is an associated partition of taxa into k v-blocks. Trivially none of these blocks can be empty, so $k \leq 4$.

Suppose that there are two cycles, a cycle C_1 of size k_1 and C_2 of size k_2 with $k_i \geq 3, i = 1, 2$. Since $\mathcal{Q}^-_{abcd}$ is level-1, by removing these two cycles we induce a partition of the taxa into at least $k_1 + k_2 - 2$ blocks. None of the blocks of this partition can be empty, so $k_1 + k_2 - 2 \leq 4$. Hence there is a most one cycle of size 4 or at most two cycles of size 3. Moreover, there cannot be a cycle of size 3 and a cycle of size 4 in the same unrooted quartet network.

Suppose that there are three cycles, a cycle C_1 of size k_1, C_2 of size k_2, and C_3 of size k_3 with $k_i \geq 3, i = 1, 2, 3$. By removing these three cycles we induce a partition of the taxa into at least $k_1 + k_2 + k_3 - 3$ blocks, so $k_1 + k_2 + k_3 - 3 \leq 4$ which is a contradiction since $k_i \geq 3$. □

Our arguments will depend on the number of descendants on the hybrid node of a cycle, so we introduce additional terminology. An n-cycle with exactly k taxa descending from the hybrid node is referred to as a n_k*-cycle*. Figure 9 shows the 6 different types of 2-, 3-, and 4-cycles possible in an unrooted quartet network.

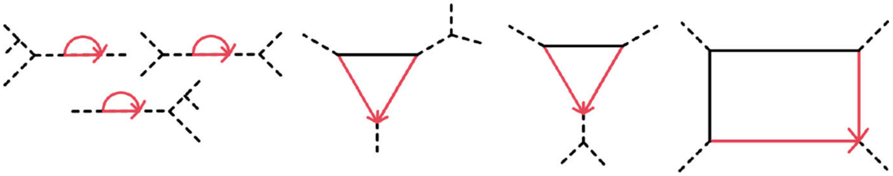


Fig. 9 (Left) The three types of 2-cycles in an unrooted quartet network (2_1-,2_2- and a 2_3-cycle); (Center) The two types of 3-cycles in the unrooted quartet network (3_1- and a 3_2-cycle). (Right) The only type of 4-cycle in an unrooted quartet network (a 4_1-cycle). The dashed lines represent subgraphs that may contain other cycles

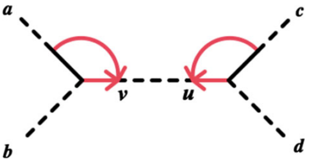


Fig. 10 A graph with two 3_2 cycles. Each dashed edge represents a chain of 2-cycles with, possibly, other cycles

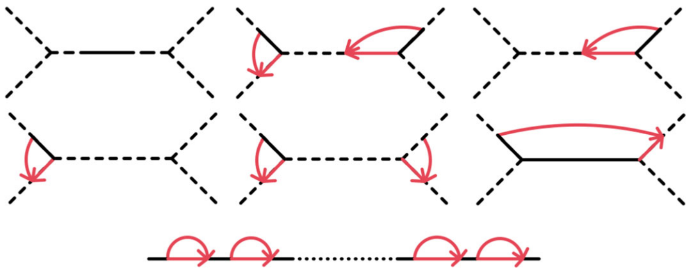


Fig. 11 Possible structures for unrooted quartet networks. Every dashed arrow represents a chain of an arbitrary number of 2-cycles, as the one in the bottom of the figure. The direction of these 2-cycles must be such that the obtained graph is induced from a rooted network

Lemma 6 *Let $\mathcal{Q}^{-}_{abcd}$ be a metric unrooted level-1 unrooted quartet network. Then $\mathcal{Q}^{-}_{abcd}$ cannot have two 3_2-cycles, or a 2_2-cycle and a 4_1-cycle.*

Proof Suppose $Q = \mathcal{Q}^{-}_{abcd}$ has two distinct 3_2-cycles, C_u and C_v. Suppose C_u has u-hybrid block $\{a, b\}$ and u-blocks $\{c\}$ and $\{d\}$. If we remove C_u from Q, by the level-1 assumption C_v is in one on the connected components. This implies that 2 of the 3 v-blocks must be contained in one of $\{a, b\}$, $\{c\}$ or $\{d\}$. This is only possible if the v-hybrid block is $\{c, d\}$, and the other v-blocks are $\{a\}$ and $\{b\}$. Thus, Q must be as the network in Fig. 10, where u is below v and v is below u, contradicting that Q is induced from a rooted network.

Now suppose that Q has a 4-cycle and a 2_2-cycle. The 4-cycle induces 4 singleton blocks. By the level-1 condition at least one of the blocks induced by the 2_2-cycle has to be contained in a singleton block. That is impossible since the blocks induced by the 2_2-cycle have size 2. □

Lemmas 5 and 6 determine all possible topological structures for unrooted quartet networks which are shown in Fig. 11.

5.1 Concordance Factor Formulas for Quartet Networks

Next we prove a number of "reduction" lemmas relating concordance factors for quartet networks to those for networks with fewer cycles. This allows us to express the network concordance factors as a linear combination of concordance factors of trees. The following observation is useful through this section.

Observation 1 *Given a rooted metric species quartet network, under the NMSC model the first coalescent event (going backward in time) determines the unrooted topology of a quartet gene tree.*

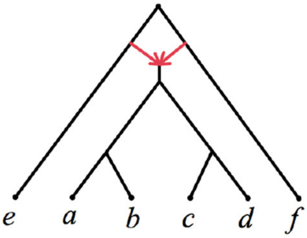


Fig. 12 A level-1 rooted network where the root differs from the LSA (a, b, c, d)

As illustrated in Fig. 12, in passing from a rooted network on X to a rooted induced network on $Z \subset X$, $\mathcal{N}_Z^+$, we may find there is a network structure above LSA(Z), a chain of 2-cycles by Proposition 2. *A priori*, this could have an impact on the behavior of the NMSC model on $\mathcal{N}_Z^+$. For quartet concordance factors, however, this additional structure has no impact, and we effectively snip it off. Formally, we have the following.

Theorem 2 *Let $\mathcal{N}^+$ be a level-1 rooted metric network on X and let a, b, c, d be distinct taxa of X. Under the NMSC model, CF_{abcd} can be computed from the LSA network $\mathcal{Q}_{abcd}^{\oplus}$.*

Proof In any realization of the coalescent process if there are fewer than 4 lineages at the LSA(a, b, c, d) in $\mathcal{N}_{abcd}^+ = \mathcal{Q}_{abcd}^+$, then a coalescent event has occurred below and therefore the unrooted gene tree topology has been determined. Thus, we condition on 4 lineages being present at LSA(a, b, c, d).

There are 2 rooted shapes for 4-taxon gene trees, the caterpillar and balanced trees. Regardless of the ancestral chain of 2-cycles above LSA(a, b, c, d), conditioned on one of these shapes, exchangeability of lineages under the coalescent tells us all labeled versions of that specific shape will have equal probability. While the rooted shapes might have different probability, since there is only 1 unrooted shape, all labelings of it must be equally probable. This is the same as if there were no ancestral cycles. Therefore, $CF_{abcd}(\mathcal{Q}_{abcd}^{\oplus}) = CF_{abcd}(\mathcal{Q}_{abcd}^{+})$. □

This argument can be modified to apply to 5 taxa, but not 6 or more, since then there is more than 1 unrooted shape.

Let $Q^{\oplus} = \mathcal{Q}_{abcd}^{\oplus}$ be a level-1 LSA quartet network and let C_v be a cycle in $Q^{\oplus}$, with hybrid node v and hybrid edges h_1 and h_2, where $\gamma = \gamma_{h_1}$. The following notation is used throughout this section:

- $Q_1^{\oplus}$ denotes the rooted quartet network obtained from $Q^{\oplus}$ by removing h_2.
- $Q_2^{\oplus}$ denotes the rooted quartet network obtained from $Q^{\oplus}$ by removing h_1.
- $Q_0^{\oplus}$ denotes the rooted quartet network obtained from $Q^{\oplus}$ by contracting C_v; if the root of $Q^{\oplus}$ is in C_v, the node obtained in the contraction process is the root of $Q_0^{\oplus}$.

Note that $Q_i^{\oplus}$, for $i = 1, 2$ have degree 2 nodes, and thus are not binary. This does not affect the coalescent process in any way and by suppressing such nodes we obtain a binary LSA network. In a slight abuse of notation, we use $Q_i^{\oplus}$ to denote both of these networks, as needed in our arguments.

To compute concordance factors we often need to designate how many lineages are present at a hybrid node in a realization of the coalescent process. To handle this formally, given a rooted metric species network $\mathcal{N}^+$ on X, we define the random variable K_v to be the number of lineages at node v, where K_v takes values in $\{1, ..., l_v\}$, where l_v is the number of taxa below v. We can extend this concept to hybrid nodes in $\mathcal{N}^-$, since a hybrid node in $\mathcal{N}^-$ induces an orientation of the nodes that are descending from it.

Let $Q^{\oplus} = \mathcal{Q}^{\oplus}_{abcd}$ be a level-1 LSA quartet network and let C_v be a cycle in $Q^{\oplus}$, with hybrid node v, which induces a cycle C'_v in $\mathcal{Q}^{-}_{abcd}$. If C'_v has size 2, then $1 \leq l_v \leq 3$; if C'_v has size three, then $1 \leq l_v \leq 2$; and if C'_v has size four then $l_v = 1$. For example, let $Q^{\oplus}$ be the LSA network shown in the left of Fig. 14 and let C_v be the cycle in $Q^{\oplus}$. By unrooting $Q^{\oplus}$ note that C_v induces a 3-cycle C'_v. Note also that Q^- is isomorphic to the network in Fig. 18.

We show that cycles in $\mathcal{Q}^{\oplus}_{abcd}$ that induce 2_1-cycles or 2_3-cycles in $\mathcal{Q}^{-}_{abcd}$ have no impact on concordance factors. But first we state Propositions 3 and 4, proven in Allman et al. (2011), which are useful in arguments to come.

Proposition 3 *Let $\mathcal{T}^+$ be a binary rooted metric species tree on X. For $|X| = 4$, $\mathcal{T}^-$ is identifiable from the unrooted topological gene tree distribution under the multispecies coalescent model on $\mathcal{T}^+$, but $\mathcal{T}^+$ is not.*

Proposition 4 *Proposition 3 remains valid when $\mathcal{T}^+$ is not binary.*

Lemma 7 *Let $Q^{\oplus} = \mathcal{Q}^{\oplus}_{abcd}$ be a metric level-1 LSA quartet network and let C_v be a cycle in $Q^{\oplus}$ that induces a 2_1-cycle in $\mathcal{Q}^{-}_{abcd}$. Then $CF(Q^{\oplus}) = CF(Q^{\oplus}_0)$.*

Proof Let $K = K_v$. Since C_v induces a 2_1-cycle in $\mathcal{Q}^{-}_{abcd}$, $P(K = 1) = 1$. Then

$$\begin{aligned} CF(Q^{\oplus}) &= P(K = 1) CF\left(Q^{\oplus} \mid K = 1\right) \\ &= P(K = 1)\left[\gamma CF\left(Q^{\oplus}_1 \mid K = 1\right) + (1-\gamma) CF\left(Q^{\oplus}_2 \mid K = 1\right)\right] \\ &= \gamma CF\left(Q^{\oplus}_1\right) + (1-\gamma) CF\left(Q^{\oplus}_2\right) \end{aligned}$$

If the root of $Q^{\oplus}$ is not in C_v, no lineages can coalesce on the edges that differ in $Q^{\oplus}_1$ and $Q^{\oplus}_2$ since there is only one lineage in such edges. Thus,

$$CF\left(Q^{\oplus}_1\right) = CF\left(Q^{\oplus}_2\right) = CF\left(Q^{\oplus}_0\right),$$

and the claim is established in this case.

Now suppose the root r of $Q^{\oplus}$ is in C_v, and C_v has nodes r, u, v, and edges (r, v), (r, u), (u, v). Without loss of generality suppose that the taxon below v is d. Since u is a tree node, it has another descendant y. Note that $Q^{\oplus}_1$ and $Q^{\oplus}_2$ have the same topology; moreover, they just differ in the edge length from the root to y. Define a random variable K', by $K' = 1$ if there has been a coalescent event before a, b, and c trace back to y and $K' = 0$ otherwise. If $K' = 1$, the unrooted topology has been determined and thus

$$CF\left(Q^{\oplus}_1 \mid K' = 1\right) = CF\left(Q^{\oplus}_2 \mid K' = 1\right) = CF\left(Q^{\oplus}_0 \mid K' = 1\right).$$

Also, by Proposition 4,

$$CF\left(Q_1^{\oplus} \mid K' = 0\right) = CF\left(Q_2^{\oplus} \mid K' = 0\right) = CF\left(Q_0^{\oplus} \mid K' = 0\right).$$

Thus, $CF(Q^{\oplus}) = CF(Q_0^{\oplus})$. □

Lemma 8 *Let $Q^{\oplus} = \mathcal{Q}^{\oplus}_{abcd}$ be a metric level-1 LSA quartet network and let C_v be a cycle in $Q^{\oplus}$, that induces a 2_3-cycle in $\mathcal{Q}^{-}_{abcd}$. Then $CF(Q^{\oplus}) = CF(Q_0^{\oplus})$.*

Proof Let $K = K_v$, so K takes values in $\{1, 2, 3\}$. Therefore,

$$\begin{aligned} CF(Q^{\oplus}) = P(K=1)CF(Q^{\oplus} \mid K=1) + P(K=2)CF(Q^{\oplus} \mid K=2) \\ + \ P(K=3)CF(Q^{\oplus} \mid K=3). \end{aligned} \tag{2}$$

If $K = 1$ or 2 then at least one coalescent event has occurred, so the unrooted gene tree topology is already determined, and

$$CF(Q^{\oplus} \mid K = k) = CF\left(Q_0^{\oplus} \mid K = k\right) \text{ for } \ k = 1, 2.$$

The case $K = 3$ requires more argument. Without loss of generality suppose that the three taxa descending from v are a, b, and c. Denote by $\mathfrak{D}$ the random variable defined by $\mathfrak{D} = 1$ if the lineage d is involved in the first coalescent event and $\mathfrak{D} = 0$ otherwise. Thus,

$$\begin{aligned} CF(Q^{\oplus} \mid K=3) = P(\mathfrak{D}=1)CF(Q^{\oplus} \mid K=3, \mathfrak{D}=1) \\ + \ P(\mathfrak{D}=0)CF\left(Q^{\oplus} \mid K=3, \mathfrak{D}=0\right). \end{aligned} \tag{3}$$

If d is in the first coalescent event, by the exchangeability property of the NMSC, a, b or c are equally likely to be the other lineage involved in that event. This is the same as if the cycle was contracted, so

$$CF(Q^{\oplus} \mid K=3, \mathfrak{D}=1) = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) = CF\left(Q_0^{\oplus} \mid K=3, \mathfrak{D}=1\right)$$

If d is not in the first coalescent event, this event involves only two of $a, b,$ and c, with each pair equally likely by exchangeability. This is also the same as if the cycle was contracted, so

$$CF(Q^{\oplus} \mid K=3, \mathfrak{D}=0) = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) = CF\left(Q_0^{\oplus} \mid K=3, \mathfrak{D}=0\right)$$

Thus, by Eqs. (2) and (3), $CF(Q^{\oplus}) = CF(Q_0^{\oplus})$. □

Together, the preceding lemmas yield the following.

Corollary 2 *Let $Q^{\oplus} = \mathcal{Q}^{\oplus}_{abcd}$ be a metric level-1 LSA quartet network and let $\widetilde{Q}^{\oplus}$ be the LSA network obtained by contracting all cycles that induce either 2_3- or a 2_1-cycles in $\mathcal{Q}^{-}_{abcd}$. Then $CF(Q^{\oplus}) = CF(\widetilde{Q}^{\oplus})$.*

While 2_1- and 2_3-cycles have no impact on concordance factors, things are not quite so simple for other types of cycles.

Lemma 9 *Let $Q^{\oplus} = \mathcal{Q}^{\oplus}_{abcd}$ be a metric level-1 LSA quartet network and let C_v be a cycle in $Q^{\oplus}$, that induces a 2_2-cycle in $\mathcal{Q}^{-}_{abcd}$. Then*

$$CF(Q^{\oplus}) = \gamma^2 CF\left(Q^{\oplus}_1\right) + (1-\gamma)^2 CF\left(Q^{\oplus}_2\right) + 2\gamma(1-\gamma) CF\left(Q^{\oplus}_0\right).$$

Proof Let $K = K_v$ with values in $\{1, 2\}$, so that

$$CF(Q^{\oplus}) = P(K=1) CF(Q^{\oplus} \mid K=1) + P(K=2) CF(Q^{\oplus} \mid K=2).$$

Suppose the root r of $Q^{\oplus}$ is not in C_v, so C_v is also a 2_2-cycle in $Q^{\oplus}$. Note that

$$\begin{aligned} CF(Q^{\oplus} \mid K=2) = \gamma^2 CF\left(Q^{\oplus}_1 \mid K=2\right) + (1-\gamma)^2 CF\left(Q^{\oplus}_2 \mid K=2\right) \\ + 2\gamma(1-\gamma) CF\left(Q^{\oplus}_0 \mid K=2\right). \end{aligned}$$

Thus, we will express $CF(Q^{\oplus} \mid K=1)$ in a similar fashion. If $K=1$ the gene tree topology has been determined before the lineages enter v. Thus, $CF(Q^{\oplus}_i \mid K=1) = CF(Q^{\oplus} \mid K=1)$ for $i \in \{0, 1, 2\}$ and

$$\begin{aligned} CF(Q^{\oplus} \mid K=1) = \gamma^2 CF\left(Q^{\oplus}_1 \mid K=1\right) + (1-\gamma)^2 CF\left(Q^{\oplus}_2 \mid K=1\right) \\ + 2\gamma(1-\gamma) CF\left(Q^{\oplus}_0 \mid K=1\right); \end{aligned} \tag{4}$$

by summing the result holds when r is not in C_v.

Now suppose that r is in C_v, and C_v has nodes r, v, u. Without loss of generality suppose that the taxa below v are c and d. Since u is a tree node, it has another descendant y. Define a random variable K_y to be the number of lineages at y. Note that K and K_y are independent, with values in $\{1, 2\}$. If either K or K_y is 1, one coalescent event has occurred and the unrooted gene tree topology has been determined so $CF(Q^{\oplus}_i \mid K=1 \text{ or } K_y=1)$ are equal for $i \in \{0, 1, 2\}$, and

$$\begin{aligned} CF\left(Q^{\oplus} \mid K=1 \text{ or } K_y=1\right) = \gamma^2 CF(Q^{\oplus}_1 \mid K=1 \text{ or } K_y=1) \\ + (1-\gamma)^2 CF\left(Q^{\oplus}_2 \mid K=1 \text{ or } K_y=1\right) \\ + 2\gamma(1-\gamma) CF\left(Q^{\oplus}_0 \mid K=1 \text{ or } K_y=1\right) \end{aligned} \tag{5}$$

Even though Eq. (5) is equal to $CF(Q^{\oplus}_0 \mid K=1 \text{ or } K_y=1)$, we express it in a similar fashion to the claimed result. Now suppose that K and K_y are both 2. Let T_c and T_d be the trees shown in Fig. 13. Therefore,

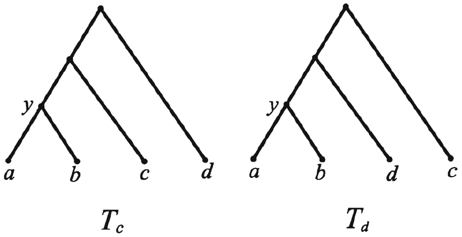


Fig. 13 The two trees T_d and T_c in the proof of Lemma 9, obtained when $K=2$, $K_y=2$ and the lineages c and d trace different hybrid edges

$$\begin{aligned} CF(Q^{\oplus} \mid K=2, K_y=2) = \gamma^2 CF\left(Q_1^{\oplus} \mid K=2, K_y=2\right) \\ + (1-\gamma)^2 CF\left(Q_2^{\oplus} \mid K=2, K_y=2\right) \\ + \gamma(1-\gamma) CF(T_c \mid K_y=2) \\ + \gamma(1-\gamma) CF(T_d \mid K_y=2). \end{aligned}$$

By Proposition 3, $CF(T_d \mid K_y=2) = CF(T_c \mid K_y=2)$, and in fact they equal $CF(Q_0^{\oplus} \mid K=2, K_y=2)$. This is because in $Q_0^{\oplus}$ the contraction of the cycle identifies the nodes r, u, and v, so conditioned on $K=2$, $K_y=2$ we may view the coalescent process on $Q_0^{\oplus}$ as that in the 4-taxon tree $((a,b):l,(c,d):0)$ where l is the length of (u,y). By Proposition 4, $CF(T_c \mid K_y=2) = CF(Q_0^{\oplus} \mid K=2, K_y=2)$. Therefore,

$$\begin{aligned} CF(Q^{\oplus} \mid K=2, K_y=2) = \gamma^2 CF\left(Q_1^{\oplus} \mid K=2, K_y=2\right) \\ + (1-\gamma)^2 CF\left(Q_2^{\oplus} \mid K=2, K_y=2\right) + 2\gamma(1-\gamma) CF\left(Q_0^{\oplus} \mid K=2, K_y=2\right). \end{aligned}$$

This together with Eq. (5) implies the claim. □

Lemma 10 *Let $Q^{\oplus} = \mathcal{Q}_{abcd}^{\oplus}$ be a metric level-1 LSA quartet network and let C_v be a cycle in $Q^{\oplus}$, that induces either a 4-cycle or a 3_1-cycle in $\mathcal{Q}_{abcd}^{-}$. Then*

$$CF(Q^{\oplus}) = \gamma CF\left(Q_1^{\oplus}\right) + (1-\gamma) CF\left(Q_2^{\oplus}\right).$$

Proof Letting $K = K_v$, then $P(K=1) = 1$. Thus,

$$\begin{aligned} CF(Q^{\oplus}) &= P(K=1) CF\left(Q^{\oplus} \mid K=1\right) \\ &= P(K=1)\left(\gamma CF\left(Q_1^{\oplus} \mid K=1\right) + (1-\gamma) CF\left(Q_2^{\oplus} \mid K=1\right)\right) \\ &= \gamma CF\left(Q_1^{\oplus}\right) + (1-\gamma) CF\left(Q_2^{\oplus}\right). \end{aligned}$$

□

It remains to consider a 3_2-cycle. For this case it helps to introduce new terminology. Let G be a semidirected graph and v be a node in G with indegree 2 and outdegree 0. Let h_v and h'_v be the edges incident to v and let u and u' the parent nodes in h_v and h'_v, respectively. We refer to *disjointing* h_v and h'_v from v as the process of (1) deleting

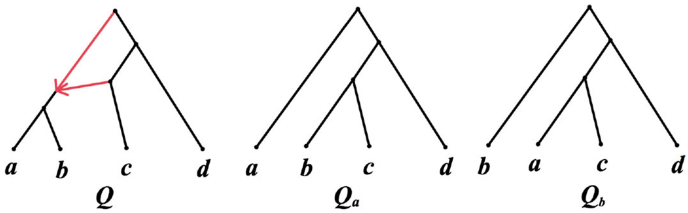


Fig. 14 A LSA quartet $Q^{\oplus}$ with a cycle C that induces a 3_2-cycle in the unrooted quartet and the graphs obtained by deleting everything below the hybrid node, disjointing, and labeling the leaves

v from G; (2) introducing nodes w and w'; (3) introducing directed edges (u, w) and (u', w').

Let $Q^{\oplus} = \mathcal{Q}^{\oplus}_{abcd}$ be a metric level-1 LSA quartet network, and C_v a cycle in $Q^{\oplus}$, that induces a 3_2-cycle in $\mathcal{Q}^{-}_{abcd}$. Without loss of generality suppose that a and b are the taxa below v. Let $Q^{\oplus}_a$ be the network obtained from $Q^{\oplus}$ by (1) deleting everything below v; (2) disjointing h_1 and h_2 from v; (3) labeling a leaf that is currently unlabeled by a and the other unlabeled leaf by b. We construct $Q^{\oplus}_b$ by swapping the labels a and b in $Q^{\oplus}_a$. Figure 14 depicts an particular example of this.

Lemma 11 *Let $Q^{\oplus} = \mathcal{Q}^{\oplus}_{abcd}$ be a metric level-1 LSA quartet network, C_v be a cycle in $Q^{\oplus}$, that induces a 3_2-cycle in $\mathcal{Q}^{-}_{abcd}$ and let $K = K_v$. Suppose that the two taxa below v are a and b, then*

$$\begin{aligned} CF(Q^{\oplus}) = \gamma^2 CF\left(Q^{\oplus}_1\right) + (1-\gamma)^2 CF\left(Q^{\oplus}_2\right) \\ + P(K=1) 2\gamma(1-\gamma) CF\left(Q^{\oplus}_0 \mid K=1\right) \\ + P(K=2)\gamma(1-\gamma)\left[CF\left(Q^{\oplus}_a\right) + CF\left(Q^{\oplus}_b\right)\right]. \end{aligned}$$

Proof By hypothesis K takes values in $\{1, 2\}$ and

$$CF(Q^{\oplus}) = P(K=1) CF\left(Q^{\oplus} \mid K=1\right) + P(K=2) CF\left(Q^{\oplus} \mid K=2\right).$$

If $K = 1$ the unrooted tree topology has been determined and $CF(Q^{\oplus} \mid K = 1)$ is given by the expression in equation (4). If $K = 2$,

$$\begin{aligned} CF\left(Q^{\oplus} \mid K=2\right) = \gamma^2 CF\left(Q^{\oplus}_1 \mid K=2\right) + (1-\gamma)^2 CF\left(Q^{\oplus}_2 \mid K=2\right) \\ + \gamma(1-\gamma) CF\left(Q^{\oplus}_a\right) + \gamma(1-\gamma) CF\left(Q^{\oplus}_b\right). \end{aligned}$$

Therefore,

$$\begin{aligned} CF(Q^{\oplus}) = P(K=1)(\gamma^2 CF\left(Q^{\oplus}_1 \mid K=1\right) + (1-\gamma)^2 CF\left(Q^{\oplus}_2 \mid K=1\right) \\ + 2\gamma(1-\gamma) CF\left(Q^{\oplus}_0 \mid K=1\right) \\ + P(K=2)\Big[\gamma^2 CF\left(Q^{\oplus}_1 \mid K=2\right) + (1-\gamma)^2 CF\left(Q^{\oplus}_2 \mid K=2\right) \\ + \gamma(1-\gamma) CF\left(Q^{\oplus}_a\right) + \gamma(1-\gamma) CF\left(Q^{\oplus}_b\right)\Big], \end{aligned}$$

which yields the claim. □

These lemmas together imply that concordance factor for rooted quartet networks actually depend only on the unrooted network. This is formalized in the following.

Proposition 5 *Let $Q = \mathcal{Q}^{\oplus}_{abcd}$ and $\widetilde{Q} = \widetilde{\mathcal{Q}}^{\oplus}_{abcd}$ be metric level-1 LSA quartet networks which induce the same unrooted network $\mathcal{Q}^{-}_{abcd} = \tilde{\mathcal{Q}}^{-}_{abcd}$. Then $CF(Q) = CF(\widetilde{Q})$.*

Proof We prove this by induction on the number of cycles in $\mathcal{Q}^{-}_{abcd}$. When there are no cycles in $\mathcal{Q}^{-}_{abcd}$, Q and $\widetilde{Q}$ are trees, and by Proposition 3, $CF(Q) = CF(\widetilde{Q})$. Assume now the result is true when there are fewer than $k+1$ cycles and that $\mathcal{Q}^{-}_{abcd}$ has $k+1$ cycles. Let C_v be a cycle in $\mathcal{Q}^{-}_{abcd}$ with hybrid edges h_1 and h_2, by Lemmas 7, 8, 9, 10, and 11, we can express the concordance factors of Q and $\widetilde{Q}$ in terms of networks with one fewer cycle. Note that these networks for Q and $\widetilde{Q}$ have the same unrooted metric structure. Thus, by the induction hypothesis $CF(\widetilde{Q}_i) = CF(Q_i)$, for $i = 0, 1, 2$, and therefore $CF(\widetilde{Q}) = CF(Q)$. □

Corollary 3 *Let $\mathcal{N}^{+}$ be a level-1 rooted metric network on X and let a, b, c, d be distinct taxa of X. Under the NMSC, $CF_{abcd} = CF(\mathcal{Q}^{\oplus}_{abcd})$ can be computed from the unrooted network $\mathcal{Q}^{-}_{abcd}$.*

We indicate how to compute the concordance factors of a LSA network $\mathcal{Q}^{\oplus}_{abcd}$ from the unrooted quartet network $Q = \mathcal{Q}^{-}_{abcd}$ without having to introduce a root. For $Q = \mathcal{Q}^{-}_{abcd}$ a unrooted metric level-1 quartet network, where using Corollary 3 we define $CF(Q) = CF(\mathcal{Q}^{\oplus}_{abcd})$:

(i) Let Q' be the graph obtained from Q by contracting all 2_3- and 2_1- cycles. By Corollary 2, $CF(Q) = CF(Q')$. If Q has a 4-cycle go to step (ii), otherwise go to step (iii).

(ii) By Lemmas 5 and 6 there are no 3_1-, 3_2- or 2_2-cycles in Q, and thus none in Q'. Then Q' only has a 4-cycle so apply Lemma 10 to Q'. Since Q'_1 and Q'_2 are quartet trees, use the formula in Eq. (1) to complete the calculation.

(iii) There are at most two 3_1-cycles in Q'. Choose one arbitrarily and apply Lemma 10. If Q'_1 and Q'_2 still have a 3_1-cycle, apply Lemma 10 again to Q'_1 and Q'_2.

(iv) We have now expressed concordance factors of Q in terms of concordance factors of unrooted quartet networks with no 2_1-,2_3-,3_1-, or 4-cycles. Apply Lemma 9 to these networks, by for instance choosing a 2_2-cycle with smallest graph theoretical distance from its hybrid node to a leaf, repeating until no 2-cycle remains.

(v) We have now an expression of the concordance factors of Q in terms of concordance factors of unrooted quartet networks with at most one 3_2-cycle. Apply Lemma 11. Then we have suppressed all cycles, and the concordance factors are now in terms of unrooted quartet trees. The formula of Eq. (1) completes the calculation.

The use of these lemmas and theorem is illustrated by a few examples.

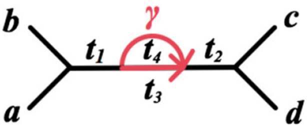


Fig. 15 An unrooted quartet with a single 2_2-cycle

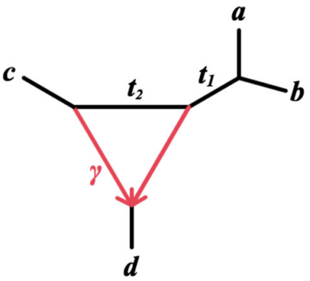


Fig. 16 An unrooted quartet with a single 3_1-cycle

Example 2 Consider the unrooted quartet network shown in Fig. 15. By Lemma 9, with $x_i = e^{-t_i}$, the quartet concordance factors are given by:

$$\begin{aligned} CF_{AB|CD} &= (1-\gamma)^2\left(1-\frac{2}{3}x_1x_2x_3\right) + 2\gamma\,(1-\gamma)\left(1-\frac{2}{3}x_1x_2\right) \\ &\quad + \gamma^2\left(1-\frac{2}{3}x_1x_2x_4\right), \\ CF_{AC|BD} &= CF_{AD|BC} \\ &= (1-\gamma)^2\left(\frac{1}{3}x_1x_2x_3\right) + 2\gamma\,(1-\gamma)\left(\frac{1}{3}x_1x_2\right) + \gamma^2\left(\frac{1}{3}x_1x_2x_4\right). \end{aligned} \tag{6}$$

Example 3 Consider the unrooted quartet network shown in Fig. 16. By Lemma 10, with $x_i = e^{-t_i}$, the quartet concordance factors are given by:

$$\begin{aligned} CF_{AB|CD} &= (1-\gamma)\left(1-\frac{2}{3}x_1\right) + \gamma\left(1-\frac{2}{3}x_1x_2\right), \\ CF_{AC|BD} &= CF_{AD|BC} = (1-\gamma)\left(\frac{1}{3}x_1\right) + \gamma\left(\frac{1}{3}x_1x_2\right). \end{aligned} \tag{7}$$

Example 4 Consider the unrooted quartet network shown in Fig. 17. By Lemma 10, with $x_i = e^{-t_i}$, the quartet concordance factors are given by:

$$\begin{aligned} CF_{AB|CD} &= (1-\gamma)\left(1-\frac{2}{3}x_1\right) + \gamma\left(\frac{1}{3}x_2\right), \\ CF_{AC|BD} &= (1-\gamma)\left(\frac{1}{3}x_1\right) + \gamma\left(\frac{1}{3}x_2\right), \\ CF_{AD|BC} &= (1-\gamma)\left(\frac{1}{3}x_1\right) + \gamma\left(1-\frac{2}{3}x_2\right). \end{aligned} \tag{8}$$

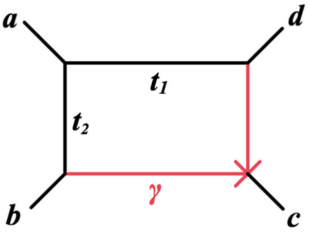


Fig. 17 An unrooted quartet with a single 4_1-cycle

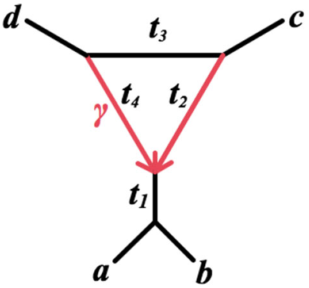


Fig. 18 An unrooted quartet with a single 3_2-cycle

Example 5 Consider the unrooted quartet network shown in Fig. 18. Given $K = 1$, one coalescent event has occurred below the hybrid node, so a and b coalesced. Therefore, $CF(Q_0 \mid K = 1) = (1, 0, 0)$. By Lemma 11, with $x_i = e^{-t_i}$, the quartet concordance factors are given by:

$$\begin{aligned} CF_{AB|CD} &= (1-\gamma)^2\left(1-\frac{2}{3}x_1x_2\right) + 2\gamma\,(1-\gamma)\left(1-x_1+\frac{1}{3}x_1x_3\right) \\ &\quad + \gamma^2\left(1-\frac{2}{3}x_1x_4\right), \\ CF_{AC|BD} &= CF_{AD|BC} \\ &= (1-\gamma)^2\left(\frac{1}{3}x_1x_2\right) + \gamma\,(1-\gamma)\,x_1\left(1-\frac{1}{3}x_3\right) + \gamma^2\left(\frac{1}{3}x_1x_4\right). \end{aligned} \tag{9}$$

Examples 1–5 agree with those in Solís-Lemus and Ané (2016).

6 The Cycle Property

In this section we focus on the ordering by magnitude of the concordance factors.

Proposition 6 *Let $Q = Q^{-}_{abcd}$ be a metric unrooted level-1 quartet network with no 3_2-cycle. The ordering of $CF_{abcd}(Q)$ is the ordering of $CF_{abcd}(Q')$ where Q' is obtained from Q by contracting all 2-cycles and all 3_1-cycles.*

Proof By Corollary 2, $CF(Q) = CF(Q^*)$, where Q^* is obtained from Q by contracting all 2_1- and 2_3-cycles. Therefore, we can assume Q has no 2_1- or 2_3-cycles. If Q has a 4-cycle, it has no 3_1- and no 2_2-cycles and the claim is established.

So suppose Q has only 2_2-cycles and 3_1-cycles. We proceed by induction in the number of cycles, with the base case of 0 cycles trivial. Assume the result is true for unrooted quartet networks with k 3_1- and 2_2-cycles and suppose Q has $k+1$. Picking one cycle and applying one of Lemmas 9 or 10 to Q, we can express the concordance factors of Q as a convex combination of $CF(Q_0)$, $CF(Q_1)$ and $CF(Q_2)$. Note that Q_0, Q_1 and Q_2 have the same topology and by induction hypothesis, $CF(Q_0)$, $CF(Q_1)$ and $CF(Q_2)$ have the same ordering as the concordance factors of Q_0', Q_1' and Q_2', respectively, the networks obtained after contracting all 2_2- and 3_1-cycles from Q_0, Q_1 and Q_2. Since Q_0', Q_1', Q_2 and Q' are trees with the same topology, their concordance factors have the same ordering by Eq. (1). Thus, $CF(Q_0)$, $CF(Q_1)$ and $CF(Q_2)$ have the same ordering, and ergo so does $CF(Q)$. □

One consequence of Proposition 6 is that for any unrooted metric level-1 quartet network Q without a 3_2- or a 4-cycle, the ordering of the concordance factors is the same as the ordering of the concordance factors of a quartet tree. That is, the two smallest elements of the concordance factors are equal. When this happens we say that Q is *treelike*, since we could use Eq. (1) to find a quartet tree with appropriate edge lengths and concordance factors equal to $CF(Q)$. However, not all unrooted quartet networks are treelike.

Example 6 Let $\mathcal{Q}^-_{abcd}$ be the unrooted 3_2-cycle quartet in Fig. 18, where $\gamma = \frac{1}{2}$, $t_1 = -\log\left(\frac{6}{7}\right)$, $t_2 = -\log\left(\frac{6}{7}\right)$, $t_3 = -\log\left(\frac{1}{14}\right)$ and $t_4 = -\log\left(\frac{13}{14}\right)$. By the equations in (9) we observe that the concordance factors are:

$$CF_{AB|CD} = \frac{32}{98},\ CF_{AC|BD} = \frac{33}{98},\ CF_{AD|BC} = \frac{33}{98}.$$

The fact that such a quartet network cannot be treelike was identified in Solís-Lemus et al. (2016), where it was pointed out that this may cause species tree methods not to be robust to the presence of gene flow.

This motivates the following definition.

Definition 16 Let $\mathcal{N}^+$ be a metric rooted level-1 network on X. We say that a set of four distinct taxa $s = \{a, b, c, d\}$ satisfies the *Cycle property* if $\mathcal{Q}^-_s$ is not treelike, that is, if the two smallest values of $CF_s = CF(\mathcal{Q}^-_s)$ are not equal.

The Cycle property is best understood geometrically. Denote by Δ_2 the two-dimensional probability simplex, the set of points in $\mathbb{R}^3$ with nonnegative entries adding to 1. Observe that $CF_{abcd} \in \Delta_2$ for any distinct taxa a, b, c, d. Figure 19 (left) depicts the simplex where the black lines are the points where the Cycle property is not satisfied; that is, the treelike unrooted quartet networks are those with concordance factors (x, y, z) satisfying $x > \frac{1}{3}, y = z$ or $y > \frac{1}{3}, x = z$ or $z > \frac{1}{3}, x = y$. All points off these segments satisfy the Cycle property. For simplicity in arguments to come, note that we can interpret concordance factors, CF_{abcd}, as a function that depends on a metric network on $\{a, b, c, d\}$ and has for image points in Δ_2.

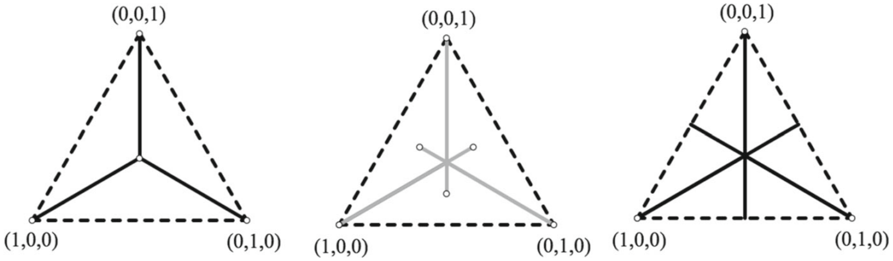


Fig. 19 On the left a planar projection of the simplex Δ_2, where the black lines represent concordance factors that are treelike. In the center, the gray segments in Δ_2 represent all the concordance factors arising from unrooted quartet networks with a 3_2-cycle. On the right, the black lines represent the variety $V((x-z)(y-z)(x-y), x+y+z-1)$, these are all concordance factors not satisfying the BC property of Definition 17

Proposition 7 *Let $Q = \mathcal{Q}^-_{abcd}$ be a metric unrooted level-1 quartet network with a 3_2-cycle. Then $CF(Q)$ lies in the set $\mathcal{I}$ defined by $x > \frac{1}{6}$, $y = z$ or $y > \frac{1}{6}$, $x = z$ or $z > \frac{1}{6}$, $x = y$, shown on the middle of Fig. 19. Furthermore, for any point (x, y, z) in this set there is such a Q with $(x, y, z) = CF(Q)$.*

Proof Let $s = \{a, b, c, d\}$ be a set of four distinct taxa and suppose that $\mathcal{Q}^-_s$ contains only a 3_2-cycle, as in Fig. 18. Then $CF(\mathcal{Q}^-_s)$ is given by Eq. (9) with $x_i = e^{-t_i}$, and in particular $CF_{AC|BD} = CF_{AD|BC}$. To maximize $CF_{AD|BC}$ in (9), let $t_i \to 0$ for $i \in \{1, 2, 4\}$ and $t_3 \to \infty$ to obtain a quadratic polynomial in γ,

$$CF_{AD|BC} \to \frac{1}{3}(1-\gamma)^2 + \gamma(1-\gamma) + \frac{1}{3}\gamma^2,$$

whose maximum value is $\frac{5}{12}$ and it is attained at $\gamma = \frac{1}{2}$. For these values, we obtain $CF(\mathcal{Q}^-_s) \to \left(\frac{2}{12}, \frac{5}{12}, \frac{5}{12}\right)$. To minimize $CF_{AD|BC}$ it is enough to let $t_1 \to \infty$, so $CF(\mathcal{Q}^-_s) \to (1, 0, 0)$.

Let $\mathcal{L}$ be the open line segment with endpoints $(1, 0, 0)$ and $\left(\frac{2}{12}, \frac{5}{12}, \frac{5}{12}\right)$. Since $CF(\mathcal{Q}^-_s)$ is continuous in t_i and γ, its image is a connected set on the line (x, y, y) containing points arbitrarily close to the endpoints of $\mathcal{L}$. Thus, the image of $CF(\mathcal{Q}^-_s)$ is $\mathcal{L}$. Permuting taxon names shows every point in the set $\mathcal{I}$ is a concordance factor for a network with a 3_2-cycle.

Now suppose $\mathcal{Q}^-_s$ has a 3_2 cycle with a, b descending from the hybrid node, and possibly other cycles. We may contract all 2_1- and 2_3-cycles by Corollary 2 without affecting $CF(\mathcal{Q}^-_s)$. By Lemmas 9 and 10, we may suppress 2_2- and 3_1-cycles by expressing $CF(\mathcal{Q}^-_s)$ as a convex sum of networks with a 3_2-cycle, but one fewer cycle. Thus, $CF(\mathcal{Q}^-_s)$ is a convex sum of points in $\mathcal{L}$, which lies in $\mathcal{L}$. □

In the supplementary materials of Solís-Lemus and Ané (2016) it is stated that an unrooted quartet network Q_{abcd} with a 3_2-cycle can be always reduced to an unrooted quartet tree with some adjustment in the edge lengths. This is not true in general; that is, when $\{a, b, c, d\}$ satisfies the Cycle property it is not treelike. However, Proposition 7 indicates that sometimes unrooted quartet networks with 3_2-cycles are treelike.

To conclude this section, we show the Cycle property can give positive information about a network.

Proposition 8 *Let $\mathcal{Q}_s^-$ be an unrooted level-1 quartet network on a set of taxa $s = \{a, b, c, d\}$. If s satisfies the Cycle property, the unrooted quartet network $\mathcal{Q}_s^-$ contains either a 3_2-cycle or a 4-cycle.*

Proof Proposition 6 shows that if $\mathcal{Q}_s^-$ has neither a 3_2-cycle nor a 4-cycle, the concordance factors of $\mathcal{Q}_s^-$ are those of a tree. □

7 The Big Cycle Property

In this section we investigate how to detect 4-cycles in a network from quartet concordance factors.

Even though the Cycle property gives us some information about an unrooted quartet network, it is not sufficient to tell us what the unrooted quartet network is. This is shown by the following example, where a 4-cycle network lead to identical concordance factors as those in Example 6.

Example 7 Let $\widetilde{\mathcal{Q}}_{abcd}^-$ be the 4-cycle unrooted quartet in Fig. 17, where $\gamma = \frac{1}{2}$, $t_1 = -\log\left(\frac{48}{49}\right) = t_2$. By the equations in (8) the concordance factors are:

$$CF_{AB|CD} = \frac{32}{98},\ CF_{AC|BD} = \frac{33}{98},\ CF_{AD|BC} = \frac{33}{98},$$

These agree with those of $\mathcal{Q}_{abcd}^-$ in Example 6.

This motivates the following definition.

Definition 17 Let $\mathcal{N}^+$ be a metric rooted level-1 network on X. We say that a subset of four distinct taxa $\{a, b, c, d\} \subset X$ satisfies the *Big Cycle* property (denoted BC) if all the entries of CF_{abcd} are different.

Let $\{a, b, c, d\}$ be a subset of taxa satisfying the BC property. Denote by q_{abcd}^{BC} the unrooted quartet corresponding to the smallest entry of CF_{abcd}.

For example, if $CF_{AB|CD} < CF_{AC|BD} < CF_{AD|BC}$, then $q_{abcd}^{BC} = AB|CD$.

Note that if s satisfies the BC property then s satisfies the Cycle property but the Cycle property is weaker than the Big Cycle property.

Proposition 9 *Let $\mathcal{Q}_s^-$ be an unrooted level-1 quartet network on a set of taxa $s = \{a, b, c, d\}$. If s satisfies the BC property, then the unrooted quartet network $\mathcal{Q}_s^-$ contains a 4-cycle.*

Proof By Proposition 8, $\mathcal{Q}_s^-$ contains either a 3_2-cycle or a 4-cycle, and by Proposition 7, $\mathcal{Q}_s^-$ cannot have a 3_2-cycle. □

A converse of Proposition 9 also holds, provided we include an assumption of generic parameters.

Proposition 10 *Let $\mathcal{N}^+$ be a metric rooted level-1 on X with $|X| \geq 4$. Let $\{a, b, c, d\} \subset X$ such that $\mathcal{Q}^-_{abcd}$ has a 4-cycle. Then $\{a, b, c, d\}$ satisfies the Cycle property. Moreover, for generic numerical parameters on $\mathcal{N}^+$, $\{a, b, c, d\}$ satisfies the BC property. That is, for all numerical parameters except those in a set of measure zero, the BC property holds.*

Proof Let $s = \{a, b, c, d\} \subset X$ be such that $\mathcal{Q}^-_s$ has a 4-cycle. Without loss of generality suppose that c is the descendant of the hybrid node and the hybrid block $\{c\}$ of $\mathcal{Q}^-_s$ is adjacent to the v-blocks containing b and d. Since $\mathcal{N}^-$ is level-1, the only other possible cycles in $\mathcal{Q}^-_s$ are 2_1 or 2_3-cycles. By Corollary 2, $CF(\mathcal{Q}^-_s) = CF(Q')$, where Q' is the network obtained after contracting all cycles other than the 4-cycle. Note that Q' is the network shown in Fig. 17, and by Eq. (8), $CF(Q')$ depends only on the length of the non-hybrid edges in the 4-cycle and the γ parameter of the hybrid edges of $\mathcal{Q}^-_s$. Moreover, Eq. (8) shows that $\{a, b, c, d\}$ satisfies the Cycle property.

When $\mathcal{Q}^-_s$ is obtained from $\mathcal{N}^-$, the lengths of the edges of $\mathcal{Q}^-_s$ are the sum of edge lengths from $\mathcal{N}^-$. Let $\Theta_{\mathcal{N}^-} = (0, \infty)^m \times [0, 1]^h$ be the numerical parameter space for $\mathcal{N}^-$ and let $\Theta'_s = (0, \infty)^2 \times [0, 1]$. Thus, we can define a map $\nu_s : \Theta_{\mathcal{N}^-} \to \Theta'_s$ such that for any metric (λ, γ) of $\mathcal{N}^-$, $\nu_s((\lambda, \gamma))$ encodes the edge length of the non-hybrid edges in the 4-cycle and the γ parameter of the hybrid edges. In particular, this map is linear and surjective.

With $\chi_s = (0, 1)^2 \times [0, 1]$, let $\eta : \Theta'_s \to \chi_s$ be defined as $\eta(l_1, l_2, \gamma) = (e^{-l_1}, e^{-l_2}, \gamma)$, so η is a biholomorphic function. Defining $f : \chi_s \to \Delta_2$ by

$$f((L_1, L_2, \gamma)) = (1 - \gamma)(1 - 2L_1/3, L_1/3, L_1/3) + \gamma(L_2/3, L_2/3, 1 - 2L_2/3),$$

the quartet concordance factor map can be viewed as a composition

$$\Theta_{\mathcal{N}^-} \xrightarrow{\nu_s} \Theta'_s \xrightarrow{\eta} \chi_s \xrightarrow{f} \Delta_2.$$

It is straightforward to see that the image of f restricted to $\gamma = 0$ and $\gamma = 1$ is the red (skewed) and blue (vertical) segments shown on the right of Fig. 20.

Let $V = V((x - z)(y - z)(x - y), x + y + z - 1)$, that is, let V be the algebraic variety composed of the points on which $(x - z)(y - z)(x - y)$ and $x + y + z - 1$ are zero, as depicted on the right of Fig. 19. Observe that V is the points in Δ_2 that, if interpreted as concordance factors, would *not* satisfy the BC property.

Since f is a polynomial map whose image is not contained in V, the pre-image of V under f is contained in a proper sub-variety of χ_s, and therefore, $f^{-1}(V)$ has measure zero in χ_s. Since η is biholomorphic, then $\eta^{-1}(f^{-1}(V))$ has measure zero. Since ν is linear surjective, then $\nu^{-1}(\eta^{-1}(f^{-1}(V)))$ has measure zero. Thus, generic points in $\Theta_{\mathcal{N}^-}$ are mapped to concordance factors satisfying the BC property. □

To better understand the geometry of the map f in this proof, let $s = \{a, b, c, d\}$ be a subset of four distinct taxa satisfying the BC property. Figure 20 depicts the subset of χ_s that is mapped by f to those segments of the shaded triangle inside Δ_2. The interior of χ_s is mapped to the interior of the shaded triangle.

The following theorem follows immediately from Propositions 10 and 9.

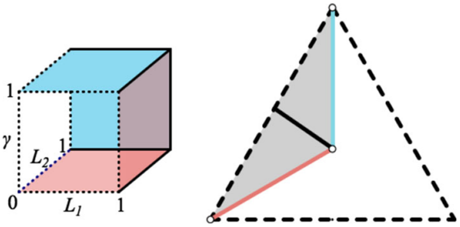


Fig. 20 The function f maps the cube χ_s (left) to Δ_2 (right). The blue facets (rear and top) of the cube are mapped by f to the blue (vertical) segment and the red facets (bottom and right) to the red (skewed) segment. The full cube is mapped onto the shaded triangle with all the concordance factor displayed by a network with a 4-cycle. The three line segments, two on the boundary of and one within the shaded triangle, are comprised of points not satisfying the BC property

Theorem 3 *Let $\mathcal{N}^+$ be a metric rooted level-1 network on X with $|X| \geq 4$ and $\{a, b, c, d\} \subset X$. For generic numerical parameters, $\{a, b, c, d\}$ satisfies the BC property if and only if $\mathcal{Q}^-_{abcd}$ has a 4-cycle.*

Theorem 3 and Proposition 8 yield the following.

Corollary 4 *Let $\mathcal{N}^-$ be a metric unrooted level-1 network on X and let $s = \{a, b, c, d\}$ be a set of distinct taxa in X. Then if s satisfies the Cycle property but not the BC property for generic parameters, then $\mathcal{Q}^-_s$ contains a 3_2-cycle.*

The converse of Corollary 4 does not hold, as pointed out by Proposition 7.

If a set of 4 taxa satisfy the BC property, we can deduce some finer information about the 4-cycle on the unrooted quartet network and a larger network, as proved in the following.

Proposition 11 *Let $\mathcal{N}^-$ be a metric unrooted level-1 network on X and let $\{a, b, c, d\} \subseteq X$ satisfy the BC property, so $\mathcal{Q}^-_{abcd}$ contains a 4-cycle C_v. Then $q^{BC}_{abcd} = AC|BD$ if and only the v-blocks of $\mathcal{Q}^-_{abcd}$ containing a and c are not adjacent.*

Proof Let $\mathcal{Q} = \mathcal{Q}^-_{abcd}$. Since $\mathcal{N}^-$ is level-1 the only possible cycles in $\mathcal{Q}$, other than C_v, are 2_1 and 2_3-cycles. Let $\mathcal{Q}'$ be the network obtained after contracting all 2_1 and 2_3-cycles, so $\mathcal{Q}'$ has only a four cycle. By Corollary 2, $CF(\mathcal{Q}) = CF(\mathcal{Q}')$. Example 4 shows that if the v-blocks of $\mathcal{Q}^-_{abcd}$ containing a and c are not adjacent then $q^{BC}_{abcd} = AC|BD$. Interchanging taxon labels in this example shows that when $q^{BC}_{abcd} = AC|BD$, then a and c are not adjacent. □

Lemma 12 *Let $\mathcal{N}^-$ be a metric unrooted level-1 network on X with generic numerical parameters. There exists $\{a, b, c, d\} \subseteq X$ satisfying the BC property if and only if $\mathcal{N}^-$ contains a cycle C_v of size $k \geq 4$ with one of these taxa is in the hybrid block, and the others in distinct v-blocks on $\mathcal{N}^-$.*

Proof Suppose that $\mathcal{N}^-$ has a cycle of size k for some $k \geq 4$ with hybrid node v. Choose four taxa $\{a, b, c, d\}$, such that a is in the hybrid block and a, b, c and d are in distinct v-blocks. This set of taxa induces a unrooted quartet network with a 4-cycle, and so by

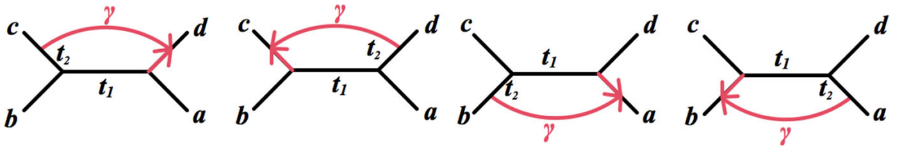


Fig. 21 Four unrooted metric level-1 quartet networks with the same concordance factors

Theorem 3 this set of taxa satisfies the BC property for generic parameters. Suppose conversely, that there exists $\{a, b, c, d\}$ satisfying the BC property. By Theorem 3, $\mathcal{Q}^{-}_{abcd}$ has a 4-cycle, so $\mathcal{N}^{-}$ has a cycle of at least size four and one of these taxa is a descendant of the hybrid node. Since the other taxa are in distinct v-blocks of $\mathcal{Q}^{-}_{abcd}$, they must be in distinct v-blocks of $\mathcal{N}^{-}$. □

For a level-1 metric unrooted network $\mathcal{N}^{-}$, let S be the collection of sets of 4 distinct taxa satisfying the BC property and V_H be the set of hybrid nodes. We observe that for any $s \in S$, there is a natural map $\psi : S \mapsto V_H$, where $\psi(s) = v$ if v is the hybrid node associated with the cycle of size 4 in $\mathcal{Q}^{-}_{s}$. In this case we say that s *determines* the hybrid node v.

Lemma 13 *Let $\mathcal{N}^{-}$ be a metric unrooted level-1 network and let $\{a, b, c, d\}$ and $\{a, b, c, e\}$ be subsets of the taxa satisfying the BC property. The set $\{a, b, c, d\}$ determines v if and only if $\{a, b, c, e\}$ determines v.*

Proof Let $\{a, b, c, d\}$ determine v, $\{a, b, c, e\}$ determine u, and suppose that $u \neq v$. Let C_v and C_u the cycles in $\mathcal{N}^{-}$ containing v and u, respectively, so C_u and C_v do not share edges. Since $\{a, b, c, d\}$ satisfies the BC property, by Lemma 12, a, b, c, and d belong to different v-blocks, so that in $\mathcal{N}^{-} \setminus E(C_v)$ the taxa a, b and c are in different connected components. Since $\mathcal{N}^{-}$ is level-1, C_u is in one of the connected components of $\mathcal{N}^{-} \setminus E(C_v)$, say $\mathcal{K}$. In particular, note that all the taxa not in $\mathcal{K}$ are in the same u-block. But at least two of a, b and c are not in $\mathcal{K}$, so at least two of a, b and c are in the same u-block. This contradicts Lemma 12, so $u = v$. □

Interestingly, under the NMSC the ordering of quartet concordance factors is insufficient to identify the hybrid node of cycles of size 4. For example, the networks shown in Fig. 21 all have the same ordering of their concordance factors despite different hybrid nodes. The concordance factors for all those networks have the same values:

$$CF_{AB|CD} = (1-\gamma)\left(1 - \frac{2}{3}e^{-t_1}\right) + \gamma\left(\frac{1}{3}e^{-t_2}\right),$$

$$CF_{AC|BD} = (1-\gamma)\left(\frac{1}{3}e^{-t_1}\right) + \gamma\left(\frac{1}{3}e^{-t_2}\right),$$

$$CF_{AD|BC} = (1-\gamma)\left(\frac{1}{3}e^{-t_1}\right) + \gamma\left(1 - \frac{2}{3}e^{-t_2}\right).$$

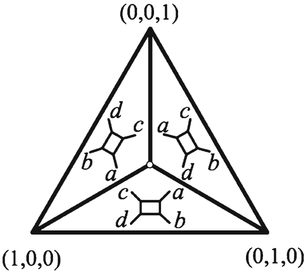


Fig. 22 Each section of the simplex is depicted with an unrooted quartet network topology whose image under the concordance factor map fills that region, independent of the placement of the hybrid node

Figure 22 shows the 4-cycle network topologies drawn in the regions of Δ_2 which their concordance factors fill. In each case it does not matter which of the cycle nodes is the hybrid node; all those unrooted quartet networks define concordance factors that fill that region.

8 Identifying Cycles in Networks

Having shown that the BC property can detect the existence of 4-cycles in networks, for generic parameters, we are poised to prove our main result. Our arguments now are mainly combinatorial.

Given a network $\mathcal{N}^+$ on X, let S denote the set of 4-taxon subsets of X satisfying the BC property. Recall that for a unrooted level-1 network $\mathcal{N}^-$ on X, the 4-network partition is the partition of X according to the connected components of the graph obtained after removing all cycles of size at least 4 from $\mathcal{N}^-$. Recall also that the blocks of such partition are referred to as 4-network blocks.

Lemma 14 *Let $\mathcal{N}^+$ be a metric rooted level-1 network on X. Then under the NMSC model with generic parameters the* 4*-network blocks of $\mathcal{N}^+$ can be determined from the set S.*

Proof If $|X| < 3$ there is nothing to prove. The case $|X| = 4$ follows from Proposition 9, so we assume $|X| \geq 5$. By Lemma 12, for any $\{a, b, c, d\} \in S$ each taxon a, b, c, d must belong to a different 4-network block. Let

$$Y_a = \bigcup_{\{s \in S | a \in s\}} s \smallsetminus \{a\}$$

Then Y_a is the complement of the 4-network block containing a. To see this, note that for any taxon b that does not belong to the 4-network block of a, by Lemma 4, there exists a cycle C_v of size at least 4 such that a and b are in different v-blocks. Now choose any two different taxa c and d, such that all taxa a, b, c, d are in different v-blocks and one of a, b, c or d is in the v-hybrid block. Then $\{a, b, c, d\} \in S$, and thus $b \in Y_a$.

It follows that $X \smallsetminus Y_x$ is the 4-network block containing taxon x. Since x was arbitrary, all 4-network blocks can be determined. □

Lemma 15 *Let $\mathcal{N}^+$ be a metric rooted level-1 network on X with cycle C_v of size $k_v \geq 4$. Then for generic parameter choices, the v-blocks and the size k_v can be identified from the set S. If $k_v \geq 5$ the v-hybrid block can also be identified.*

Proof Let $\{a, b, c, d\} \in S$ and let v be the hybrid node determined by it. By Lemma 12, each of these taxa belongs to a different v-block, and hence to a different 4-network block. Denote by A, B, C, D the v-blocks containing a, b, c and d, respectively.

Let Z_{abc} be the set of all taxa e such that $\{a, b, c, e\} \in S$. By Lemma 13, all such $\{a, b, c, e\} \in S$ determine the same hybrid node v. Consider now Z_{bcd}, Z_{acd} and Z_{abd}. If $k_v = 4$, then, by the last statement of Lemma 12, $Z_{abc} = D$, $Z_{bcd} = A$, $Z_{acd} = B$ and $Z_{abd} = C$, so all pairwise intersections of Z_{abc}, Z_{bcd}, Z_{acd}, Z_{abd} are empty. If $k_v > 4$, then, again by Lemma 12, for some distinct taxa $i, j, k \in \{a, b, c, d\}$, Z_{ijk} is the v-hybrid block, and for any $l, m, n \in \{a, b, c, d\}$ with $\{l, m, n\} \neq \{i, j, k\}$, $Z_{lmn} = (L \cup M \cup N)^c$. Note that $Z_{ijk} \cap Z_{lmn} = \emptyset$ since one of L, M, N is the v-hybrid block. Since Z_{lmn} contains at least one v-block other than A, B, C or D, for any $l', m', n' \in \{a, b, c, d\}$, with $\{l', m', n'\} \neq \{i, j, k\}$, $Z_{lmn} \cap Z_{l'm'n'} \neq \emptyset$. Hence we can determine whether $k_v > 4$ or $k_v = 4$: if all pairwise intersection of Z_{abc}, Z_{bcd}, Z_{acd}, Z_{abd} are empty then $k_v = 4$, else $k_v > 4$. If $k_v > 4$ we can determine the hybrid block, by noting which of the sets Z_{abc}, Z_{bcd}, Z_{acd}, Z_{abd} has empty intersection with any other set in this family. At this point we have determined either that $k_v = 4$ and all v-blocks, or that $k_v > 4$ and the hybrid block.

In the case $k_v > 4$, without loss of generality, suppose that A is the v-hybrid block. Let $y \notin Z_{abc} = (A \cup B \cup C)^c$, so y is in one of A, B and C. For some $u, w \in \{a, b, c\}$, $s' = \{y, u, w, d\} \in S$, which shows y and the taxon $g \in \{a, b, c\} \smallsetminus \{u, w\}$ are in the same v-block. Thus, we can determine A, B and C.

Note that for any taxon x that is not in any of A, B or C, then $s = \{a, x, b, c\} \in S$. Since s determines v, following the steps of the last paragraph identifies the v-block that contains x. Therefore, all v-blocks can be determined, and thus k_v as well. □

Lemma 16 *Let $\mathcal{N}^+$ be a metric rooted level-1 network on X. Then for any hybrid node v with $k_v \geq 4$ the order of the v-blocks in the cycle can be determined from the ordering of the concordance factors.*

Proof If $k_v = 4$, the claim is established by Proposition 11. Now suppose that $k_v > 4$, so by Lemma 15 we know the v-hybrid block. Let $A_1, \ldots, A_{k_v}$ be the v-block partition with A_1 the v-hybrid block. Let $a_i \in A_i$ be an element of the i-th v-block. By Proposition 11, A_1 and A_j are adjacent if and only if $q^{BC}_{a_1 a_j xy} \neq a_1 a_j | xy$ for any distinct $x, y \in \{a_2, \ldots, a_{k_v}\} \smallsetminus \{a_j\}$. Thus, we can identify the two v-blocks adjacent to A_1. Suppose that such v-blocks are A_p and A_q. We find the other v-block adjacent to A_q from $\{q^{BC}_{a_1 a_p a_j a_m}\}$ for all distinct $j, m \in \{2, 3, 4, \ldots, k_v\} \smallsetminus \{p, q\}$. This is, A_q and A_j are adjacent if and only if $q^{BC}_{a_1 a_j a_p x} \neq a_1 a_j | x a_p$ for any distinct $x \in \{a_2, \ldots, a_{k_v}\} \smallsetminus \{a_p, a_q, a_j\}$ and $j \neq 1, p, q$. Continuing in this way, the full order of blocks around the cycle can be determined. □

We reach the main result.

Theorem 4 *Let $\mathcal{N}^+$ be a metric rooted level-1 network on X. Then under the NMSC model, for generic parameters, the collection of orderings of quartet concordance* *factors identifies the unrooted semidirected topological network $\widetilde{\mathcal{N}}$ obtained from $\mathcal{N}^-$ by contracting all 2- and 3-cycles, and directions of hybrid edges in 4-cycles, while retaining directions of hybrid edges of k-cycles for $k \geq 5$.*

Proof We proceed by induction in the number of cycles of size ≥ 4. Suppose there are no such cycles. Then every induced quartet tree will have no cycle of size 4, and the ordering of the concordance factors determines the topology of the quartet tree obtained by contracting all 2- and 3-cycles. These then determine the topology $\widetilde{\mathcal{N}}$ by a standard result Semple and Steel (2005).

Suppose there is exactly one cycle of size at least 4. Then there is just one hybrid node v in $\mathcal{N}^-$ with $k_v \geq 4$. By Lemmas 15 and 16 we can determine the size k_v of the cycle, the v-blocks and the order of the v-blocks in the cycle. If $k_v \geq 5$ we can identify the hybrid node v and thus identify the direction of the hybrid edges. Let P_u be a v-block where u is a node in C_v, and $q \in X \smallsetminus P_u$. Let $\mathcal{K}$ be the induced network on $P_u \cup \{q\}$ with all 2-cycles and 3-cycles contracted. Note that $\mathcal{K}$ is a tree, and the quartet concordance factors for taxa in $P_u \cup q$ identify its topology. Viewing q as an outgroup of P_u induces a rooted tree on P_u. The root can then be joined with an edge to u. Doing this for all v-blocks establishes the claim.

Now suppose that the result is true for networks with l cycles of size at least 4, and $\mathcal{N}^-$ contains $l+1$ such cycles. We can first determine all 4-network blocks and the v-blocks and its cycle order for every cycle of size at least 4 by Lemmas 14, 15, and 16. Following Definition 13, consider $\mathcal{T}$, the tree of cycles of $\widetilde{\mathcal{N}}$. A leaf of $\mathcal{T}$ arises from a cycle C_v on $\mathcal{N}^-$ if and only if all v-blocks but one are 4-network blocks. We may therefore determine the v-blocks of some cycle C_v that is a leaf of $\mathcal{T}$.

Let u be the vertex in C_v associated with the v-block that is not a 4-network block. Note that $\widetilde{\mathcal{N}} \smallsetminus \{u\}$ is a disconnected graph, with two connected components $\widetilde{\mathcal{N}}_1$ and $\widetilde{\mathcal{N}}_2$. Let $\widetilde{\mathcal{N}}_1$ be the component containing all nodes of C except u, and S_i the set of taxa on $\widetilde{\mathcal{N}}_i$, $i \in \{1, 2\}$. Let $s_i \in S_i$. Then $\mathcal{N}^-_{S_i \cup \{s_j\}}$ for $i, j \in \{1, 2\}$, $i \neq j$, has at most l cycles of size at least 4. By the induction hypothesis we can determine the semidirected topological network $\mathcal{N}_i$ obtained from $\mathcal{N}^-_{S_i \cup \{s_j\}}$ by contracting all 2- and 3-cycles, and directions of the hybrid edges in 4-cycles, while retaining directions of the hybrid edges of k-cycles for $k \geq 5$. We obtain $\widetilde{\mathcal{N}}$ by identifying s_1 in $\mathcal{N}_2$ with s_2 in $\mathcal{N}_1$ and suppressing that node. □

Figure 23 shows a phylogenetic metric rooted network $\mathcal{N}^+$ and $\widetilde{\mathcal{N}}$, the unrooted semidirected topological network which is identified by Theorem 4. The cycle colored in green is a 4-cycle and, though, its hybrid node is not identified from quartet concordance factors. However, its hybrid node has to be such that $\widetilde{\mathcal{N}}$ is induced from a rooted network. Thus, the node labeled x in Fig. 23 cannot be the hybrid node. This illustrates that although we cannot always identify the hybrid node on 4-cycles, sometimes the structure of the resulting network $\widetilde{\mathcal{N}}$ restricts the possible nodes for its placement.

9 Further Results on 3_2-Cycles

Under some special circumstances, for example, when a set of taxa satisfy the Cycle property but not the BC property, it is possible to detect further information about the

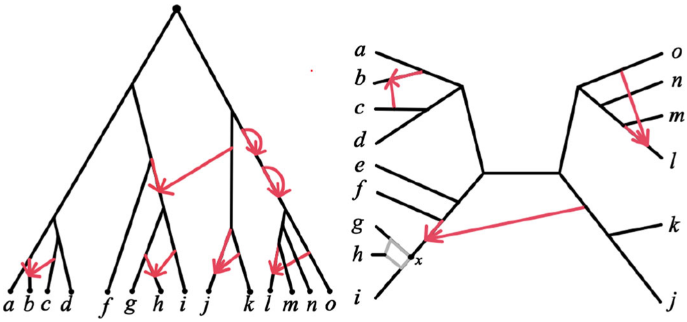


Fig. 23 A rooted metric phylogenetic network $\mathcal{N}^+$ (left) and the network structure $\widetilde{\mathcal{N}}$ (right) that can be identified by Theorem 4. The 4-cycle on the network in the right, colored gray, has 3 different candidates for the hybrid node

topology of the network than that given in Theorem 4. For instance, some 3-cycles are identifiable under such hypothesis. In this section, we discuss these extensions briefly, as it is difficult to formulate general statements on identifiability.

Recall that a 3_2-cycle may lead to concordance factors satisfying the Cycle property, but it need not, as shown in Proposition 7. There is a full-dimensional subset of parameters space on which concordance factors indicate a 3_2-cycle and another in which it fails to. Nonetheless, the following gives a positive, but limited, identifiability result.

Proposition 12 *Let $\mathcal{N}^+$ be a metric rooted level-1 network on X and suppose $\{a, b, c, d\} \subset X$ satisfies the Cycle property but not the BC property. Then under the NMSC model, for generic parameters, if there is no taxon $e \in X$ such that $\{i, j, k, e\}$ satisfies the BC property for any distinct $i, j, k \in \{a, b, c, d\}$ then $\mathcal{N}^-$ contains a 3-cycle with at least two descendants of the hybrid node.*

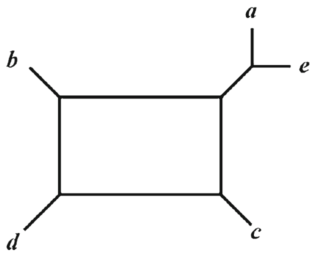
Proof Since $\{a, b, c, d\} \subset X$ satisfy the Cycle property but not the BC property, by Proposition 8, there is a 3_2-cycle in $\mathcal{Q}^-_{abcd}$. Thus, three taxa of a, b, c, d are in distinct v-blocks in $\mathcal{Q}^-_{abcd}$. This implies that there exists a cycle C_v in $\mathcal{N}^-$ where three taxa of a, b, c, d are in distinct v-blocks. Since $\{i, j, k, e\}$ does not satisfy the BC property for any distinct $i, j, k \in \{a, b, c, d\}$, this implies C_v is not a k-cycle for $k \geq 4$. Thus, by Proposition 7, C_v has size 3 and at least two of a, b, c, d descend from v. □

Let $\mathcal{Q}^-_{abcd}$ be an unrooted level-1 quartet network where $\{a, b, c, d\}$ satisfies the Cycle property but not the BC property. It can be shown that if, for example, the smallest entry in CF_{abcd} is the one corresponding to the quartet $AB|CD$, then either a, b or c, d are in the v-hybrid block. This proof is very similar to that of Proposition 11.

Let $\mathcal{N}^+$ be a network such that $\widetilde{\mathcal{N}}$ (in the network obtained from $\mathcal{N}^+$ in Theorem 4) is as shown in Fig. 24. Observe that $\{a, c, d\}$ satisfies the BC property by Theorem 3. If $\{a, e, b, d\}$ satisfies the Cycle property, then the following Proposition indicates the hybrid node in the network shown in Fig. 24 can be determined.

Proposition 13 *Let $\mathcal{N}^+$ be a metric rooted level-1 network on X and let C_v be a 4-cycle in $\mathcal{N}^-$. Let $a, b, c, d \in X$ be in different v-blocks in $\mathcal{N}^-$. Suppose under the*

Fig. 24 A network $\widetilde{\mathcal{N}}$ with a four cycle such that if $\{a, b, c, e\}$ satisfies the Cycle property, the hybrid block can be detected



NMSC model, for generic parameters, for distinct $i, j, k \in \{a, b, c, d\}$, there exists a taxon $e \in X$ such that $\{i, j, k, e\}$ satisfies the Cycle property but not the BC property. Then the v-block containing e is the v-hybrid block.

Proof Without loss of generality suppose that $i = a$, $j = b$ and $k = c$. Note that e is not in the same v-block as d, otherwise $\{a, b, c, e\}$ would satisfy the BC property. Thus, e is the same v-block as a, b or c. Without loss of generality suppose that is in the same v-block as a. Thus, $\{e, b, c, d\}$ satisfies the BC property and by Theorem 4 the order of the cycle can be determined. Without loss of generality suppose that the order is the one as in Fig. 24. By Lemma 13, $\{a, b, c, d\}$ and $\{e, b, c, d\}$ determine the same hybrid node v. Since $\{a, b, c, e\}$ satisfies the Cycle property, Corollary 4 shows $\mathcal{Q}^-_{abce}$ has a 3_2-cycle. The 4-cycle in $\mathcal{Q}^-_{abcd}$ and the 3-cycle in $\mathcal{Q}^-_{abce}$ have to have the same hybrid edges, otherwise the level-1 condition would be violated. Observe that the only possibility for $\mathcal{Q}^-_{abce}$ having a 3_2-cycle is if e and a are in the hybrid block. □

In Solís-Lemus and Ané (2016) it is stated that one could identify the hybrid node in a 4-cycle when the number of taxa in the network is greater than 4 by using multiple concordance factors at once.

10 Discussion

In this work, we show that for generic numerical parameters, under the network multispecies coalescent model the collection of orderings of quartet concordance factors identifies the unrooted semidirected topological network obtained from $\mathcal{N}^-$ by contracting all 2- and 3-cycles, and ignoring the directions of hybrid edges in 4-cycles, while retaining directions of hybrid edges in larger cycles.

As mentioned in the introduction, the proof of this result suggests combinatorial methods for constructing the network under noiseless data, but the question remains open in the presence of noise. There are two challenges when noise is introduced. The first one consists of detecting whether a quartet network contains a 4-cycle or not. We would never expect the empirical concordance factors to be exactly treelike. For this challenge, one could develop a statistical test to determine when concordance factors are sufficiently close to treelike to doubt the presence of a 4-cycle. The second challenge arises after determining such test. Since the test will not be accurate all the time, some quartets will not be inferred correctly and thus we need a method to reconstruct the network with some erroneous quartets. We leave this for future work.

Acknowledgements The author deeply thanks John A. Rhodes and Elizabeth S. Allman for their technical assistance and suggestions during the development of this work, and the reviewers for their valuable suggestions and observations.

Appendix

Here, Proposition 1 of Section 2 is proved. The argument uses the following.

Lemma 17 *Let $\mathcal{N}^+$ be a (metric or topological) rooted network on X and let $Z \subset X$. For any edge e below LSA(Z), with a descendant in Z, there are $x, y \in Z$ such that e is in a simple trek in $\mathcal{N}^+$ from x to y whose edges are below LSA(Z).*

Proof Let $x \in Z$ be below e. By Lemma 2 there exists $y \in Z$ with LSA(x, y) above e.

Suppose y is not below e. Let P_x be a path from LSA(x, y) to x containing e and let P_y be a path from LSA(x, y) to y. Let u be the minimal node in the intersection of P_x and P_y. Since y is not below e, u cannot be below e. Then the subpath of P_x from u to x, which contains e, and the subpath of P_y from f to y form a simple trek containing e.

Now assume y is below e. Since e is below LSA(x, y), there exists a path from LSA(x, y) to one of y or x that does not pass through the child of e. Without loss of generality suppose such a path P_y goes from LSA(x, y) to y. Let P_x be a path from LSA(x, y) to x that passes through e. Let $A = A(P_x, P_y)$ be the set of nodes above e, common to P_y and P_x. Let $a \in A$ be the minimal node in A.

Let $B(P_y, P_x)$ be the set of nodes below e, common to P_y and P_x. We may assume that we choose P_x and P_y such that $B = B(P_y, P_x)$ has minimal cardinality. If $B = \emptyset$ then the desired trek is easily constructed, with top a. So suppose $B \neq \emptyset$ has minimal element b^- and maximal element b^+. We are going to contradict the minimality of B. Note that b^+ must be the hybrid node of a cycle containing e (see Fig. 25 for a graphical reference).

Since b^- is not LSA(x, y), there exists a path P^* from LSA(x, y) to one of x or y that does not pass through b^-. Note that P^* has to intersect at least one of P_y or P_x at an internal node below b^-. Let C_1 be the set of nodes below b^-, common to P^* and P_y and let C_2 be the set of nodes below b^-, common to P^* and P_y. Let c be the maximal node in $C_1 \cup C_2$. We can assume, without loss of generality, that c is in P_y. This is because if instead, c were in P_x, we can construct paths P'_x and P'_y where P'_i contains all the edges in P_i above b^- and all edges of P_j below b^- for $i, j \in \{x, y\}$, $i \neq j$. Note that P'_x passes through e and does not contains c, while P'_y does not pass through e, contains c, and $B = B(P'_y, P'_x)$.

Denote by W the set of nodes in $(P^* \cap P_y) \cup (P^* \cap P_x)$ and let w be the minimal node of W above b^-. Since $\mathcal{N}^+$ is binary, w cannot be a or b^+ (see Fig. 25 for a graphical reference). There are 5 different cases of the location of w in the network composed by the paths P_y and P_x. These are

1. w is in P_y, above b^+ but below a.
2. w is in P_x, above b^+ but below e.
3. w is in P_x, above e but below a.

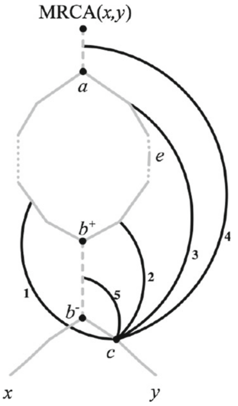


Fig. 25 In gray we see the subgraph composed by P and P', the dashed edges represent that P and P' could intersect, the dotted segments represent just a succession of edges. In black we see the different cases of the possible edges in P^* above b but below a

4. w is in one or more of P_x or P_y, above a.
5. w is in one or more of P_x or P_y, above b^- but below b^+.

Figure 25 depicts in gray the graph composed by the paths P_y and P_x, and in black we see the possible subpaths of P^* from w to c. In any of case 1, 2 or 3 we can find a simple trek containing e as depicted in Fig. 26 by choosing the appropriate edges, and thus, B was not minimal. For case 4 and 5 there are two possibilities; (i) w is in both P_y and P_x; (ii) w is only in one of P_y or P_x. For case 4 (i), the situation is simple, and we can find a simple trek as depicted on the left in Fig. 27. For case 4 (ii), we first find the node in A that is right above w. Then as depicted on the left of Fig. 27 we can find a simple trek.

For case 5 we do not find a simple trek directly, instead we construct two paths P_1 and P_2 from LSA(x, y) to x, y, respectively, only one of which contains e with at least one less node in $B(P_1, P_2)$ than B. For case 5 (i), we just take P_1 to be the same as P_x and for P_2 we consider the same edges that are in P_y above w, the edges below c, and the edges in P^* between w and c. For case 5 (ii), we assume without loss of generality that w is in P_x. Let b be the node in B right above w. Let P_1 be the path containing the edges in P_x that are above b, the edges in P_y that are below b but above the node $b' \in B$ right below w, and at last the edges in P_x below b'. Let P_2 the path containing the edges in P_y that are above b, the edges in P_x that are above a but below b, the edges in P^* that are above c but below w and at last the edges in P_y that are below c. Figure 27 (right) depicts P_1 (red) and P_2 (blue) for (i) and (ii). Since $B(P_1, P_2)$ has at least one less node that B and we assumed B, the minimality of B is contradicted. □

Proof (of Proposition 1) Let $M^+ = \mathcal{N}_Z^{\oplus}$. Let M^- be the graph obtained from M^+ by ignoring the direction of all tree edges and then suppressing the LSA$(Z, \mathcal{N}^+)$, that is, the induced unrooted network from M^+. Denote by M' the graph obtained by ignoring all directions of the tree edges in M^+, so that by suppressing degree two nodes of either M^- or M' gives $(\mathcal{N}_Z^+)^-$. Let K be the graph obtained by considering all the edges in simple treks in $\mathcal{N}^-$ from x to y for all $x, y \in Z$, so that suppressing degree two nodes in K gives $(\mathcal{N}^-)_Z$. Showing either $M' = K$ or $M^- = K$, will prove the claim.

First we show that if LSA$(Z, \mathcal{N}^+) \neq$LSA$(X, \mathcal{N}^+)$ then $M' = K$, by arguing that M' and K have the same edges. Let e be an edge of M'. Since

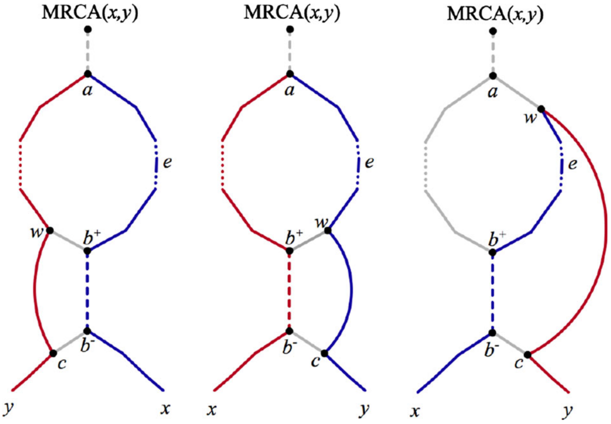


Fig. 26 The treks in case 1 (left), case 2 (center), and case 3 (right)

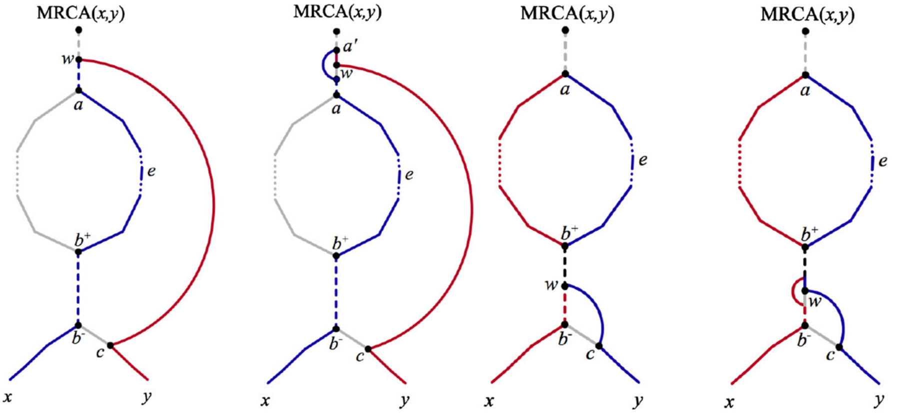


Fig. 27 (Left) The treks in the two possibilities of case 4. (Right) The two possibilities of case 5, where the black segments represent possible edges red and blue at the same time

LSA$(Z, \mathcal{N}^+)\neq$LSA$(X, \mathcal{N}^+)$, M' is a subgraph of $\mathcal{N}^-$ and e is directed in M^+. By Lemma 17, e is in a simple trek in M^+ from x to y, for some $x, y \in Z$. This trek induces a simple trek in M' from x to y, and therefore a simple trek in $\mathcal{N}^-$ from x to y. Thus, e is in K.

Now let e be an edge of K. Then there exists a simple trek $(\overline{P_1}, \overline{P_2})$ in $\mathcal{N}^-$ from x to y, for some $x, y \in Z$ containing e. Let v =top$(\overline{P_1}, \overline{P_2})$ and let T be the sequence of incident edges in $\mathcal{N}^+$ from x to v conformed of edges inducing those in $\overline{P_1}$ and $\overline{P_2}$. Since $(\overline{P_1}, \overline{P_2})$ is simple, T does not have repeated edges. Following T in $\mathcal{N}^+$ from x to y, edges are first transversed "uphill" (in reverse direction) until there is a first "downhill" edge (u, w). The next edge in T cannot be uphill, as otherwise it would be hybrid and $(\overline{P_1}, \overline{P_2})$ would have not been a trek in $\mathcal{N}^-$. This argument applies for all consecutive edges in T until we end at y. Thus, there is a simple trek $(\overline{P_1}, \overline{P_2})$ from x to y in $\mathcal{N}^+$ with top u. Note that u must be below or equal to LSA$(Z, \mathcal{N}^+)$ since otherwise the trek would not be simple. Moreover, P_1 and P_2 contain only edges in

M^+ and thus in M' after the directions of the tree edges is omitted. Thus, e is in M', so $K = M'$.

If LSA$(Z, \mathcal{N}^+)$=LSA$(X, \mathcal{N}^+)$ then $M^- = K$ follows from a straight forward modification of the previous argument to account for the suppression of LSA$(z, \mathcal{N}^+)$ in both M^- and K. □

References

Allman ES, Degnan JH, Rhodes JA (2011) Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. J Math Biol 62(6):833–862

Ané C, Larget B, Baum DA, Smith SD, Rokas A (2007) Bayesian estimation of concordance among gene trees. Mol Biol Evolut 24(2):412–426

Arnold ML (1997) Natural hybridization and evolution, vol 53. Oxford University Press, Oxford

Bapteste E, van Iersel L, Janke A, Kelchner S, Kelk S, McInerney JO, Morrison DA, Nakhleh L, Steel M, Stougie L, Whitfield J (2013) Networks: expanding evolutionary thinking. Trends Genet 29(8):439–441

Carstens BC, Knowles LL, Tim C (2007) Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: an example from melanoplus grasshoppers. Syst Biol 56(3):400–411

Degnan JH (2010) Probabilities of gene trees with intraspecific sampling given a species tree. In: Knowles LL, Kubatko LS (eds) Estimating species trees: practical and theoretical aspects. Wiley-Blackwell, pp 53–78. ISBN 0470526858

Ellstrand NC, Whitkus R, Rieseberg LH (1996) Distribution of spontaneous plant hybrids. Proc Nat Acad Sci U S A 93(10):5090–5093

Gusfield D, Bansal V, Bafna V, Song YS (2007) A decomposition theory for phylogenetic networks and incompatible characters. J Comput Biol 14(10):1247–1272

Huber KT, van Iersel L, Moulton V, Scornavacca C, Wu T (2017) Reconstructing phylogenetic level-1networks from nondense binet and trinet sets. Algorithmica 77(1):173–200

Huber KT, Moulton V, Semple C, Wu T (2017) Quarnet inference rules for level-1 networks. https://arxiv.org/pdf/1711.06720.pdf

Keijsper JCM, Pendavingh RA (2014) Reconstructing a phylogenetic Level-1 network from quartets. Bull Math Biol 76(10):2517–2541

Linder CR, Rieseberg LH (2004) Reconstructing patterns of reticulate evolution in plants. Am J Bot 91(10):1700–1708

Liu Liang Yu, Scott Lili Edwards, V. (2010) A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evolut Biol 10(1):302

Mallet J (2005) Hybridization as an invasion of the genome. Trends Ecol Evolut 20(5):229 – 237. **Special issue: invasions, guest edited by Michael E. Hochberg and Nicholas J. Gotelli**

Meng C, Kubatko LS (2009) Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. Theor Popul Biol 75(1):35–45

Nakhleh L (2010) Evolutionary phylogenetic networks: models and issues. In: Heath L, Ramakrishnan N (eds) Problem solving handbook in computational biology and bioinformatics. Springer, Boston, pp 125–158

Noor MA, Feder JL (2006) Speciation genetics: evolving approaches. Nat Rev Genet 7(11):851–861

Pamilo P, Nei M (1988) Relationships between gene trees and species trees. Mol Biol Evolut 5:568583

Pollard DA, Iyer VN, Moses AM, Eisen MB (2006) Widespread discordance of gene trees with species tree in drosophila: evidence for incomplete lineage sorting. PLoS Genet 2(10):1634–1647

Rieseberg LH, Baird SJ, Gardner KA (2000) Hybridization, introgression, and linkage evolution. Plant Mol Biol 42(1):205–224

Rosselló F, Valiente G (2009) All that glisters is not galled. Math Biosci 221(1):54–59

Semple C, Steel M (2005) Phylogenetics. Oxford University Press, Oxford

Solís-Lemus C, Ané C (2016) Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. PLoS Genet 12(3):e1005896

Solís-Lemus C, Ané C, Yang M (2016) Inconsistency of species tree methods under gene flow. Syst Biol 65(5):843–851

Steel M (2016) Phylogeny discrete and random processes in evolution. SIAM, Philadelphia

Sullivant S, Talaska K, Draisma J (2010) Trek separation for gaussian graphical models. Ann Statist 38(3):1665–1685

Syring J, Willyard A, Cronn R, Liston A (2005) Evolutionary relationships among Pinus (Pinaceae) subsections inferred from multiple low-copy nuclear loci. Am J Bot 92(12):2086–2100

John Wakeley (2008) Coalescent theory: an introduction, vol 58. Roberts and Company Publishers, Englewood

Yu Y, Degnan JH, Nakhleh L (2014) Maximum likelihood inference of reticulate evolutionary histories. PNAS 111(296–305):11

Yu Y, Degnan JH, Nakhleh L (2012) The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. PLoS Genet 8:e1002660

Yu Y, Than C, Degnan JH, Nakhleh L (2011) Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. Syst Biol 60(2):138–149

Zhang C, Ogilvie HW, Drummond AJ, Stadler T (2018) Bayesian inference of species networks from multilocus sequence data. Mol Biol Evolut 35(504–517):02

Zhu J, Yu Y, Nakhleh L (2016) In the light of deep coalescence: revisiting trees within networks. BMC Bioinform 17:415

Zhu S, Degnan J (2017) Displayed trees do not determine distinguishability under the network multispecies coalescent. Syst Biol 66:283298