SPECIES NETWORK INFERENCE UNDER THE MULTISPECIES COALESCENT MODEL

By

Hector Daniel Baños Cervantes

A Dissertation Submitted in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

in

Mathematics

University of Alaska Fairbanks

May 2019

APPROVED:

Dr. Elizabeth S. Allman, Committee Co-Chair
Dr. John A. Rhodes, Committee Co-Chair
Dr. Ronald Barry, Committee Member
Dr. Jill Faudree, Committee Member
Dr. Anthony Rickard, Chair
Department of Mathematics and Statistics
Dr. Leah Berman, Dean
College of Natural Sciences and Mathematics
Dr. Michael Castellini, *Dean of the Graduate School*

Abstract

Species network inference is a challenging problem in phylogenetics. In this work, we present two results on this. The first shows that many topological features of a level-1 network are identifiable under the network multispecies coalescent model (NMSC). Specifically, we show that one can identify from gene tree frequencies the unrooted semidirected species network, after suppressing all cycles of size less than 4. The second presents the theory behind a new, statistically consistent, practical method for the inference of level-1 networks under the NMSC. The input for this algorithm is a collection of unrooted topological gene trees, and the output is an unrooted semidirected species network.

Table of Contents

Title Page i

Abstract iii

Table of Contents v

List of Figures vii

Acknowledgments xv

Chapter 1: Introduction 1

Chapter 2: The Network Multispecies Coalescent model 6
- 1 The coalescent model 6
- 2 The network multispecies coalescent model (NMSC) 10

Chapter 3: Identifying Species Network Features from Gene Tree Quartets Under the Coalescent Model 13
- 1 Introduction 14
- 2 Phylogenetic Networks 16
- 3 Structure of level-1 Networks 22
- 4 The Network Multispecies Coalescent model and Quartet Concordance Factors 25
- 5 Computing Quartet Concordance Factors 27
- 6 The Cycle property 38
- 7 The Big Cycle property 41
- 8 Identifying cycles in networks 45
- 9 Further results in 3_2-cycles 47
- 10 Discussion 49
- 11 Appendix 50

Chapter 4: NANUQ: A method for inferring species networks from gene trees under the coalescent model 55
- 1 Introduction 57
- 2 Phylogenetic Networks 60
- 3 The Network Multispecies Coalescent model and Quartet Concordance Factors 62
- 4 Network split systems and distances 68
- 5 Quartet distance for level-1 networks 70
- 6 Split networks from the network quartet distance 74
- 7 The NANUQ algorithm for inference of phylogenetic networks 80

8 Examples ... 87

Chapter 5: Conclusions and future work ... 95

References ... 96

List of Figures

Chapter 1

Figure 1 A species network relating species a, b, c and d. 2

Figure 2 A non level-1 species network relating species a, b, c and d. 2

Figure 3 (Left) A gene tree within the species tree. In this case one lineage was sampled from each species. (Right) The same metric gene tree as the gene tree in the left. 3

Chapter 2

Figure 1 (Left) A single population with 4 lineages sampled. Each horizontal line of dots represents a generation, and each dot represents an individual. (Right) The resulting gene tree observed from the coalescent process in the population on the left. 7

Figure 2 A population y with length δ_y in cu, where 3 lineages enter y and 2 leave it. Time $u^{y}{}_{3}$ has elapsed between the lineages entering y and the coalescent event that reduces the number of lineages from 3 to 2. 9

Figure 3 (Left) A metric gene tree $\mathcal{G}$. (Right) The two different embeddings of $\mathcal{G}$ in $\mathcal{N}$. 11

Chapter 3

Figure 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. 18

Figure 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. 18

Figure 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. 20

Figure 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.......... 21

Figure 5 In a level-1 network on X, the structure between the root and m =LSA(X) is a chain of two cycles. The number of two cycles in the chain could be zero... 23

Figure 6 (Left) A level-1 unrooted network $\mathcal{N}^-$ and (Right) the tree of cycles of $\mathcal{N}^-$. 25

Figure 7 Two gene trees within a species network with one hybrid node. 26

Figure 8 Cases 1-4 (Left-Right) of Example 3-1, of how lineages may behave under the NMSC model on the network of Figure 3-7. 26

Figure 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.......... 28

Figure 10 A graph with two 3_2 cycles. Each dashed edge represents a chain of 2-cycles with, possibly, other cycles. 29

Figure 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. 29

Figure 12 A level-1 rooted network where the root differs from the LSA(a, b, c, d). 30

Figure 13 The two trees T_d and T_c in the proof of Lemma 3-9, obtained when $K = 2$, $K_y = 2$ and the lineages c and d trace different hybrid edges.......... 34

Figure 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. 35

Figure 15 An unrooted quartet with a single 2_2-cycle. 37

Figure 16 An unrooted quartet with a single 3_1-cycle. 37

Figure 17 An unrooted quartet with a single 4_1-cycle. 38

Figure 18 An unrooted quartet with a single 3_2-cycle. 38

Figure 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 3-17.......... 40

Figure 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. 43

Figure 21 Four unrooted metric level-1 quartet networks with the same concordance factors. 44

Figure 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. 45

Figure 23 A rooted metric phylogenetic network $\mathcal{N}^+$ (left) and the network structure $\widetilde{\mathcal{N}}$ (right) that can be identified by Theorem 3-4. The 4-cycle on the network in the right, colored gray, has 3 different candidates for the hybrid node. 48

Figure 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. 49

Figure 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. 51

Figure 26 The treks in case 1 (left), case 2 (center), and case 3 (right). 52

Figure 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. 52

Chapter 4

Figure 1 (Left) A rooted phylogenetic network N^+ with root r and lowest common ancestor m. (Right) The unrooted network N^- induced from N^+. 61

Figure 2 The quartet networks Q_{abdf} and Q_{bcef} respectively induced from the network on the right of Figure 4-1. 62

Figure 3 Cycles in a level-1 quartet network are classified as type n_k if they have n edges and k descendants of the hybrid node. The only cycles possible in a level-1 quartet network are (Left) of type 2_1, 2_2, and 2_3; (Center) of type 3_1 and 3_2; and (Right) of type 4_1. The dashed lines represent subgraphs that may contain other cycles. 63

Figure 4 On the left a planar projection of the simplex Δ_2, where the gray lines represent concordance factors that arise from quartet networks with no 3_2-cycle and no 4-cycle (including quartet trees). In the center, the gray segments in Δ_2 represent all the concordance factors arising from quartet networks with a 3_2-cycle. On the right, the gray shaded area represents concordance factors arising from quartet networks with a 4-cycle with generic numerical parameters. In all figures, the quartets labeling line segments, and the quartet networks labeling regions, represent the quartet network after contraction of all 2- and 3-cycles. 65

Figure 5 An unrooted quartet with a single 3_2-cycle. Internal edge lengths are denoted t_i, and hybridization parameters γ and $1-\gamma$. 66

Figure 6 The shaded area in the x_1x_3 unit square indicates solutions of $x_1 \leq 4/(5-x_3)$, which lead to treelike CFs on the network of Figure 4-5. Here $x_i = \exp(-t_i)$, with t_i the branch length in coalescent units. 67

Figure 7 The shaded area represents the solutions of $M \leq \frac{2}{x_1} - \frac{3}{2}$ with the x_1-axis horizontal and the M axis vertical. With $M = \max(x_2, x_4)$, points in this region lead to tree-like CFs on the network of Figure 4-5. 68

Figure 8 The two trees in the grove $\mathcal{G}$, where N^- is the unrooted network of Figure 4-1. 70

Figure 9 On the left, two quartet networks Q_{xyzw}, and on the right, the networks $\widetilde{Q}_{xyzw}$ obtained by contracting all 2- and 3-cycles and suppressing degree 2 nodes. For the top network $\rho_{xy}(Q_{xyzw}) = 0$ since x, y are not separated in the quartet. For the bottom $\rho_{xy}(Q_{xyzw}) = 1/2$, since x, y are separated on one of the quartet trees obtained by deleting a hybrid edge from $\widetilde{Q}_{xyzw}$ but not on the other. 72

Figure 10 A cycle in a level-1 network (L) and the two simpler networks (R) produced from it by deleting one hybrid edge. The cycle edges in these networks that arise from the original cycle are shown in blue. If N^- has a single cycle, then the networks on the right are the two trees in $\mathcal{G}$. 78

Figure 11 A m-dart, for $m = 5, 6, 7$ respectively. The frontier edges, shown in bold outline, are as characterized in the text. The outer vertices labelled by the X_i are the corners. The point of the dart is the unique corner which is $m - 3$ frontier edges away from both of its closest corners. This figure was obtained by applying the Circular Network Algorithm implemented in SplitsTree4 *Huson and Bryant* [2006] to appropriate distance matrices. 79

Figure 12 An unrooted level-1 network N^- (top L), the network obtained from it by contracting 2- and 3-cycles and undirecting 4-cycles (top R), and the split graph (bottom) obtained from it by Proposition 4-33. Note the split graph has a 4-cycle, a 5-dart, and a 6-dart, arising from the 4-, 5-, and 6-cycles of N^-. As described in Section 4-8, the split graph was obtained by applying the NANUQ algorithm to a large simulated data set of gene trees, and drawn by SplitsTree4 *Huson and Bryant* [2006]. 81

Figure 13 The species network N^+ of Example 4-36. See Newick notation in text for branch lengths and hybridization parameters........ 88

Figure 14 Representative simplex plots from empirical QCCFs, with hypothesis testing results........ 88

Figure 15 Simplex plots for hypothesis test results on the yeast data set, with two choices of significance levels $\alpha = 10^{-2}$ and 10^{-4} with $\beta = 10^{-1}$. The choice of β here is largely irrelevant, as no plotted empirical CFs are near the center. Smaller α results in fewer empirical CFs being called as supporting 4-cycles. .. 89

Figure 16 Split networks inferred for yeast dataset of Example 4-37, with $\alpha = 10^{-2}$ (L) or 10^{-4} (R) and $\beta = 10^{-1}$.. 90

Figure 17 Simplex plots showing hypothesis test results for Heliconius data set of Example 4-38.. 90

Figure 18 Split graph for Heliconius dataset of Example 4-38, with $\alpha = 10^{-40}$, $\beta = 10^{-6}$.. 91

Figure 19 Inferred network structure from Heliconius dataset of Example 4-38, from the split graph of Figure 4-18.. 92

Acknowledgments

I would like to express my deepest appreciation to my advisers Dr. Elizabeth Allman and Dr. John A. Rhodes for letting me work with them, for their support, patience, and contributions to this work. Being mentored by such remarkable people has been an honor.

I would also like to express appreciation to my committee, Dr. Jill Faudree and Dr. Ron Barry, for taking the time to answer questions, review and provide suggestions for this work.

In addition, I would like to thank Dr. Leah Berman, Dr. Gordon Williams, Dr. Latrice Bowman, Dr. David Maxwell, Dr. Alexei Rybkin, Dr. John Gimbel, Dr. Julie McIntyre, Dr. Scott Goddard, and specially Dr. Ed Bueler, Dr. Margaret Short and Kellie Dolan. At last, I would like to thank Cheto and Flor.

Chapter 1: Introduction

Phylogenetics is a branch of evolutionary biology whose main objective is to infer evolutionary relationships between species. Recently, more evidence has appeared showing that hybridization has played an important role in such relationships [*Nakhleh*, 2011]. In particular, there is strong evidence showing hybridization as an crucial factor in the evolution of some groups of plants, fish, and frogs [*Rieseberg et al.*, 2000; *Ellstrand et al.*, 1996]. However, methods for inferring hybridization in a statistical framework are only beginning to be developed.

One tool used to depict the relation between species in the presence of hybridization is a *species network* [*Steel*, 2016]. A species network N is a branching diagram, such as that in Figure 1, composed of edges, depicted by "tubes," representing populations. When two edges are incident this represents either a speciation event or a hybridization event. Two edges that are incident representing a hybridization event are called *hybrid edges* (hybrid populations), edges that are not hybrid are called *tree edges*. In the species network of Figure 1 the hybrid edges are colored in red and the tree edges in white. In such a figure, the *leaves*, the ends of terminal edges labeled as a, b, c, and d, represent the current species being related. The rest of the edges represent ancestral populations. The *root*, the edge on the top of the diagram, is the most recent common ancestor of all the species that are being related. Even if not explicitly marked, the edges in the diagram are directed away from the root. This is interpreted as time flowing from the root (the past), to the leaves (the present). A hybridization event is depicted by the heads of two edges being incident, while a speciation event is depicted by the tails of two edges being incident.

Phylogenetics is also concerned with quantifying how distant the relationships are between species. We can assign to each population in a species network an *edge length*, which

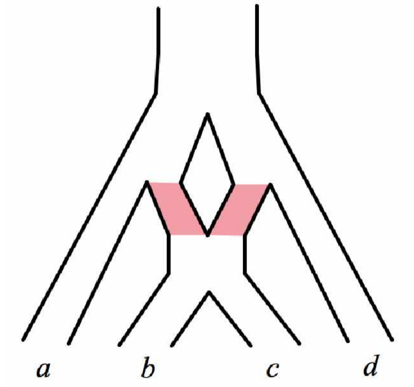


Figure 1: A species network relating species a, b, c and d.

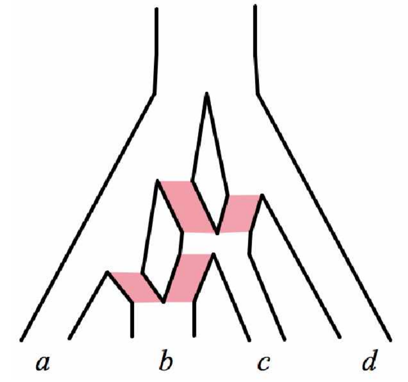


Figure 2: A non level-1 species network relating species a, b, c and d.

represents how much time, in number of generations, has passed between speciation or hybridization events. We also assign a number $\gamma(e) \in (0, 1)$, called *hybridization parameter*, to any hybrid edge e. This number represents the probability that a lineage that is below e has an ancestral lineage in e. We formalize this in Chapter 2.

Due to the complexity of an arbitrary species network, it is common to work with networks with a simpler structure. In this work we restrict to *level-1* networks. Informally, a level-1 network is a network whose cycles have no edges in common.

A common approach of inference methods for a species network is to use as data *gene trees*, see for example [*Degnan and Salter*, 2005; *Carstens et al.*, 2007; *Chifman and Kubatko*,

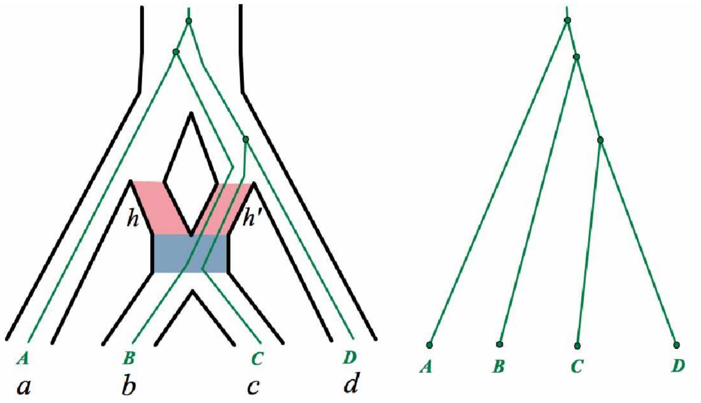


Figure 3: (Left) A gene tree within the species tree. In this case one lineage was sampled from each species. (Right) The same metric gene tree as the gene tree in the left.

2015; *Solís-Lemus and Ané*, 2016]. Gene trees are acyclic, connected graphs [*Semple and Steel*, 2005], that show the relation between genes of individuals sampled from the different species of interest. They show direct ancestral relationships from child to parent and are inferred from DNA sequences. There exist various methods of gene tree inference [*Allman and Rhodes*, 2005], but in this work we do not focus on any of these methods, but assume gene trees are somehow known. A gene tree could also contain metric information. We denote by (G, t) a *metric gene tree*, where G contains the topological information (the shape of the tree) and t the edge length information of G. On the right in Figure 3 a gene tree relating lineages A, B, C, and D is depicted.

There is much evidence showing that for a given set of species, trees for different genes relate the species differently. This is known as *gene tree incongruence.* There are several sources of gene tree incongruence, for example, the presence of hybridization [*Degnan*, 2010], and *incomplete lineage sorting* [*Syring et al.*, 2005; *Pollard et al.*, 2006; *Carstens et al.*, 2007]. Hybridization occurs when, going forwards in time, two distinct populations merge genetically to produce a new population. Incomplete lineage sorting occurs when, going backwards in time, two lineages that enter the same population do not coalesce after until they leave that population, so they may coalesce with lineages from more distantly related organisms first. For example, in Figure 3 (Left), we see that lineages B and C enter the population colored in blue, but they do not coalesce until the root population. There are other possible sources of gene tree incongruence, for example, gene loss and gene duplication [*Olson*, 1999], but here we only consider gene tree incongruence due to hybridization and incomplete lineage sorting. That already is a challenge for inference methods.

Incomplete lineage sorting together with hybridization is modeled by the network multispecies coalescent model (NMSC) [*Meng and Kubatko*, 2009], which is a generalization of the coalescent model for lineages in a single population [*Kingman*, 1988]). We describe both models in Chapter 2.

In using a model-based inference method, a goal is typically to estimate the parameters underlying the data. Specifically, for model-based statistical inference to have a solid basis, we need the parameters of the model to be *identifiable*. That is, the probability distribution for the model must uniquely determine the parameters (i.e the model parametrization map is injective).

In this work we undertake two main investigations, both based under the network multispecies coalescent model. The first one is presented in the paper "Identifying species network features from gene tree quartets under the coalescent model," published in the Bulletin of Mathematical Biology [*Baños*, 2019]. Its main result is to solve a model identifiability problem: given information on frequencies of gene trees, under the NMSC, for some unknown network N, what can be determined about the topology of N? This work consists of 3 stages: 1) Provide rigorous arguments for gaps left unaddressed in previous literature; 2) describe gene tree frequencies for subtrees with four species using algebraic statistics; and 3) use the frequencies of the subtrees with four species to reconstruct most of the network features using graph theoretical methods.

The second main result is an algorithm, that we refer to by the acronym "NANUQ." NANUQ is a new practical method of species network inference. It takes as input a set of gene trees and produces an unrooted network with certain properties. The algorithm steps are to 1) determine the gene tree frequencies for subtrees with four species; 2) use a statistical test to determine whether any subset of four species display hybridization; 3) construct a distance between species using information on tree and networks for subsets of four species; 4) use SplitsTree4 [*Huson and Bryant*, 2006] to obtain a circular splits graph; and 5) give an interpretation of the circular splits graphs produced by SplitsTree4 to reconstruct the network. This work is presented in the preprint "NANUQ: A method for inferring species networks from gene trees under the coalescent model," which is a joint work with J. A. Rhodes and E. S. Allman [*Allman et al.*, 2019].

For NANUQ, Step 1 is based on the insights of [*Solís-Lemus and Ané*, 2016; *Baños*, 2019]; Step 2 is based on the statistical test developed in [*Allman et al.*, 2018a]; Step 3 is done by generalizing the distance method for trees in [*Rhodes*, 2017]; Step 4 uses NeighborNet [*Bryant and Moulton*, 2004], and an algorithm that constructs circular splits graphs [*Dress and Huson*, 2004], both implemented in SplitsTree4; Step 5 required developing new theory on splits graphs for networks.

An outline of this thesis is as follows: Chapter 2 gives a detailed explanation of the coalescent model and the network multispecies coalescent model. Chapter 3 contains the paper "Identifying species network features from gene tree quartets under the coalescent model." Chapter 4 contains the paper "NANUQ: A method for inferring species networks from gene trees under the coalescent model." Finally, in Chapter 5 we present conclusions and outline future work.

Chapter 2: The Network Multispecies Coalescent model

Before introducing the coalescent model, we formalize the idea of edge lengths and hybridization parameters mentioned in Chapter 1. We associate to a topological network N a function $\tau : E(N) \to [0, \infty)$, where $E(N)$ is the set of edges (i.e populations) on N not including the root. The function τ gives the number of generations spanned by a population or the edge length. This represents how much time, in number of generations, has passed between speciation or hybridization events. For the root r, we set $\tau(r) = \infty$. We also assign to N a function $\gamma : E_h(N) \to (0, 1)$, where $E_h(N)$ is the set of hybrid edges of N, and require for any two incident hybrid edges e and $\hat{e} \in E_h(N)$, that $\gamma(e) = 1 - \gamma(\hat{e})$. The number $\gamma(e)$ is called the hybridization parameter of e.

We assign additionally to each edge e a function $N_e(t) : [0, \tau(e)) \to \mathbb{R}_{>0}$, where $\tau(e)$ is the edge length of e. The function $N_e(t)$, called the population function of e, represents the population size in number of generations in edge e at time t. We denote the set of population functions of a network N by $\{N_e\}$. The quadruplet $(N, \tau, \gamma, \{N_e\})$ is called a *metric species network* [*Steel*, 2016]. A metric species network with no hybridization events is also known as a *metric species tree.*

1 The coalescent model

The simplest form of coalescent theory models the formation of gene trees within one population. The coalescent model traces, backwards in time, the ancestries of a finite set of individual copies of a gene as the lineages *coalesce* to form ancestral lineages. The model has as parameters a population size function and gives a distribution of metric gene trees. For now, we assume that any lineage from the population can trace back in time infinitely,

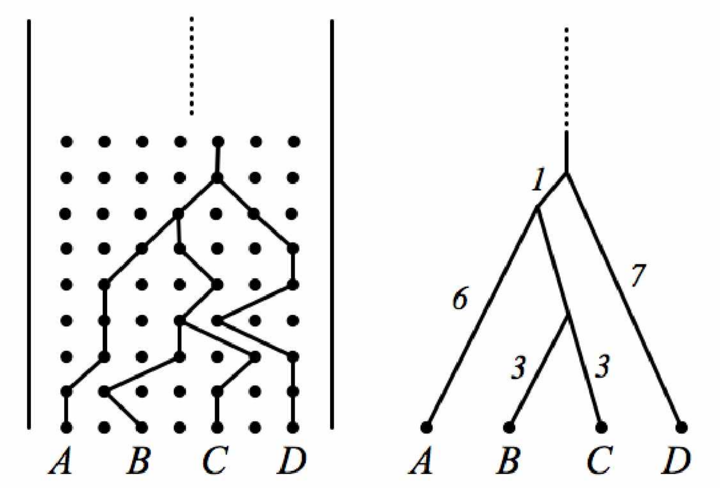


Figure 1: (Left) A single population with 4 lineages sampled. Each horizontal line of dots represents a generation, and each dot represents an individual. (Right) The resulting gene tree observed from the coalescent process in the population on the left.

that is, the population has a infinite number of past generations. On the left of Figure 1 we observe lineages A, B, C, and D sampled from a population, each horizontal sequence of dots represents a generation, and each dot represents an individual. Each lineage traces backwards in time from an individual to an individual in the previous generation. The first coalescent event in Figure 1 occurs at the 3th generation, and it involves lineages B and C. The second coalescent event occurs at the 6th generation, and the last coalescent event occurs at the 7th generation before the present. On the right of Figure 1 we see the resulting gene tree of this process. This depiction of the coalescent model with discrete time and population size is known as the Wright-Fisher model. We focus on the continuous version of the coalescent model, that is, considering continuous notions of both the number of generations and population sizes. This model is known as the Kingman coalescent model. We refer the reader to [*Wakeley*, 2008] for a detailed explanation, history and survey of both models. For now on, we refer to the Kingman coalescent model as the coalescent model.

For simplicity in presentation of the model, we measure time u in *coalescent units* (cu). Coalescent units are obtained by scaling time t in number of generations, inversely by population size ($\Delta u = \frac{\Delta t}{N(t)}$).

In the coalescent model, coalescent events between any two lineages occur independently by a Poisson process with the instantaneous rate of coalescence 1. That is, coalescent events between any two lineages occur rarely, with equal probability in any small interval of a fixed size. The probability of simultaneous coalescence of more than two lineages is zero. After a coalescent event occurs this process re-starts but with one fewer lineage. Because the rate of coalescence for any two pairs of lineages is the same, all lineages behave the same in the coalescent process. This important observation is expressed by saying lineages are *exchangeable.*

Let $h(u_k)$ be the probability that no two lineages out of k lineages present have coalesced by time u_k within a population. There are $\binom{k}{2}$ possible pairs of lineages, and each of the $\binom{k}{2}$ possible coalescent events occur are independently. Also, since coalescent events follow a Poisson process, $h(u_k)$ should decrease and the instantaneous rate of the possible $\binom{k}{2}$ independent events should be added, to obtain the rate for the first coalescent event. Thus the instantaneous rate $h'(u_k)$ is negative and equal to

$$h'(u_k) = -\binom{k}{2} h(u_k).$$

Since $h(0) = 1$, we find that

$$h(u_k) = e^{-\binom{k}{2}u_k}. \tag{1}$$

Also, given k gene lineages in a population, the probability density of the time u_k until the first pair of lineages coalesce is given by

$$f(u_k) = \frac{d}{du_k}(1 - h(u_k)) = \binom{k}{2} e^{-\binom{k}{2}u_k}. \tag{2}$$

Now we review the coalescent process in a population that has a finite number of genera-

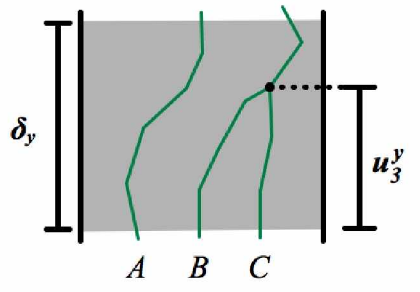


Figure 2: A population y with length δ_y in cu, where 3 lineages enter y and 2 leave it. Time u_3^y has elapsed between the lineages entering y and the coalescent event that reduces the number of lineages from 3 to 2.

tions, which will be needed in the next section. Let y be a population of length $\delta_y \in (0, \infty)$ in cu, p be the number of lineages entering y and q be the number of lineages leaving it. There are then $p - q$ coalescent events in y. In Figure 2 we see that in the population shaded in gray $p = 3$, and $q = 2$. Let u_k^y denote the time from the moment lineages entered population y to the time of the coalescent event that reduces the number of lineages in this branch from k to $k - 1$. Figure 2 depicts u_3^y, that is, u_3^y is the time from the moment lineages A, B, and C entered y, to the coalescent event that reduces the number of lineages from 3 to 2. By the exchangeability property, when there are j lineages present at a certain point in time, all pairs are equally likely to coalesce, so the density for a coalescent event has to be weighted by $1/\binom{j}{2}$. Therefore we can write the joint density of coalescent times $u_p^y, u_{p-1}^y, ..., u_{q+1}^y$ within population y as

$$f_y(u_u^y, u_{p-1}^y, ..., u_{q+1}^y) = \prod_{j=q+1}^{p} \left[\exp\left(-\binom{j}{2} (u_j^y - u_{j+1}^y) \right) \right] \exp\left(-\binom{v}{2} (\delta_y - u_{q+1}^y) \right), \quad (3)$$

where we define $u_{p+1}^y = 0$. It follows from Equation 1, that the probability that there are no coalescent events in y, i.e, $q = p$, is given by

$$P(\text{no coalescent among } p \text{ lineages in population } y) = \exp\left(-\binom{p}{2} \delta_y \right). \quad (4)$$

Equations (3) and (4) describe the coalescent process in a single population y of length δ_y.

2 The network multispecies coalescent model (NMSC)

The *network multispecies coalescent model* (NMSC) generalizes the coalescent model. It applies the coalescent model to multiple populations connected by a species network. The parameters of the model are a metric species network $(N, \tau, \gamma, \{N_e\})$, and it gives a distribution of metric gene trees. When there is no hybridization in the species network, i.e. the network is a tree, the NMSC is simply referred to as the multispecies coalescent model (MSC) [*Pamilo and Nei*, 1988]. The NSMC is used to obtain the probability of a metric gene tree in the presence of incomplete lineage sorting. By marginalization over branch lengths in N, the NMSC gives a distribution of topological gene trees. We briefly explain how to obtain the density of a metric gene tree.

When a lineage L is in a population y that is below two hybrid populations h and h', like population y depicted in blue in the species network of Figure 3, the probability that L traces backwards in time from an individual in y to an individual in h is given by $\gamma(h)$. Analogously, the probability that L trace backwards in time from an individual in y to and individual in h' is $\gamma(h') = 1 - \gamma(h)$. The random variables determining which hybrid edge is entered are independent of any other lineages leaving y. For example, in the species network of Figure 3, the population shaded in blue is below the hybrid populations h and h', so any lineage sampled from b has probability $\gamma(h')$ of tracing its ancestry back through h'. Moreover, the probability that, for example, both lineages B and C, conditioned on not having coalesced below the hybrid populations, trace through to h is $\gamma(h)^2$.

Let $\mathcal{G} = (G, t)$ be a metric gene tree and let $\mathcal{N} = (N, \tau, \gamma, \{N_e\})$ be a metric species network. To find the density of the metric gene tree $\mathcal{G}$ conditional on the species network $\mathcal{N}$

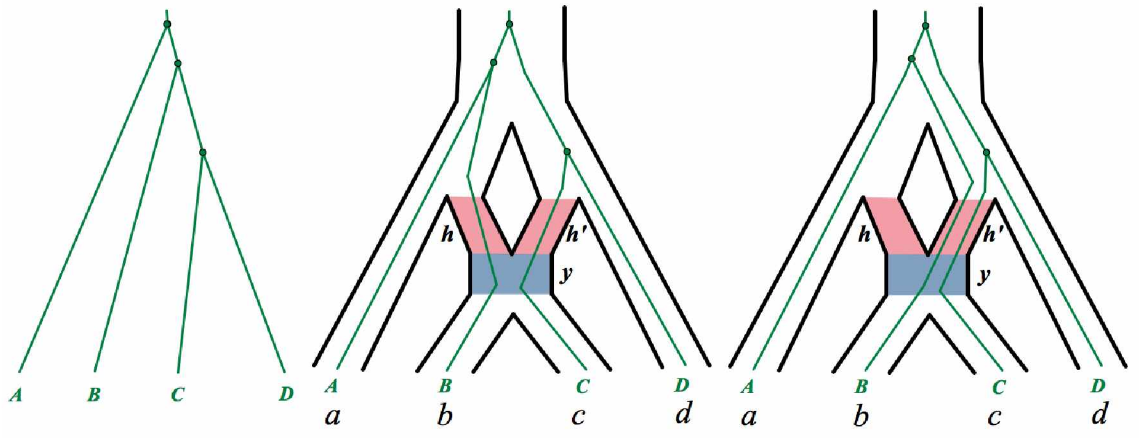


Figure 3: (Left) A metric gene tree $\mathcal{G}$. (Right) The two different embeddings of $\mathcal{G}$ in $\mathcal{N}$.

we introduce the following. An *embedding* K of $\mathcal{G}$ in $\mathcal{N}$ (also denoted by $K(\mathcal{G}|\mathcal{N})$), encodes a coalescent process of observing $\mathcal{G}$ in $\mathcal{N}$. We say two embeddings are distinct if there is at least one lineage in $\mathcal{G}$ that traced through different populations of $\mathcal{N}$ in each embedding. For example, the left of Figure 3 shows a metric gene tree and the right of the figure shows the two distinct embeddings of it in a species network. For any metric gene tree the number of embeddings is finite, and if there are no hybridization events in the species network, there is a unique embedding.

For a given embedding $\mathcal{K} = K(\mathcal{G}|\mathcal{N})$, the number of lineages entering and leaving any given population are specified. Since the coalescent processes within different populations are conditionally independent, the probability densities for individual populations (as in equation (3)) of such embedding can be multiplied to obtain the probability density for the embedding $\mathcal{K}$. This is

$$g(\mathcal{K}) = \left(\prod_{e\in T} f_e(u^e_{p_e}, u^e_{p_e-1}, ..., u^e_{q_e+1})\right) \cdot \left(\prod_{h\in H} \gamma(h)^{p_h} \cdot f_h(u^h_{p_h}, u^h_{p_h-1}, ..., u^h_{q_h+1})\right) \qquad (5)$$

where T is the set of tree edges of N, H is the set of hybrid edges of N, p_y is the number of lineages entering population y, and q_y is the number of lineages leaving y. Thus the density of the metric gene tree $\mathcal{G}$ conditional on the species tree $\mathcal{N}$ is given by

$$f(\mathcal{G}|\mathcal{N}) = \sum_{i} g(\mathcal{K}_i) \tag{6}$$

where the index i of the finite sum is over different embeddings of $\mathcal{G}$ in $\mathcal{N}$.

Note that when there is no hybridization in a species network, computing the probability of observing a metric gene tree is challenging, but straightforward algorithmically [*Pamilo and Nei*, 1988]. The presence of hybridization complicates this probability computation due to the different embeddings of $\mathcal{G}$ in $\mathcal{N}$. However, the probability of a topological gene tree is even more difficult to compute on $\mathcal{N}$, since we need to consider all possible metric structures of gene trees that can be embedded in the species network, in addition to all possible embeddings.

Inference of species networks for gene trees using "standard" approaches (such as Maximum likelihood or Bayesian methods) needs the computations of gene tree densities. Also, one might prefer to work with topological gene trees because the metric information that most gene tree inference methods produce may not be reliable and could lead to a wrong conclusion in the species network inference. The next chapters indicate a way around these challenges.

Chapter 3: Identifying Species Network Features from Gene Tree Quartets Under the Coalescent Model

In this chapter we present the paper "Identifying Species Network Features from Gene Tree Quartets Under the Coalescent Model." As mentioned in Chapter 1 this paper consists of 3 parts.

The first part (Sections 1 through 5) give rigorous arguments for gaps left unaddressed in previous literature. In [*Solís-Lemus and Ané*, 2016], the authors present a statistical method, based on the NMSC and a pseudo-likelihood function, to infer level-1 species networks. While their method is novel, some of the results presented on identifiability did not contain complete statements, nor proofs. To build on their work, it was necessary to revise some of the claims and formally provide introductory definitions and arguments to fully address both the claims and the extensions of them.

The second part (Sections 6 and 7) consists of describing gene tree frequencies, under the NMSC, for subtrees with four taxa. This is the main idea used in the detection of "big cycles" in the species network. That is, we can determine whether there is hybridization among a set of four species by looking at the gene tree frequencies for those species. These ideas are also used in NANUQ, and this will be fully addressed in Chapter 4.

The last part (Sections 8 and 9) describe how to use frequencies of subtrees with four taxa to reconstruct most topological features of the network. The main tools used in these sections are combinatorial and graph theoretical.

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

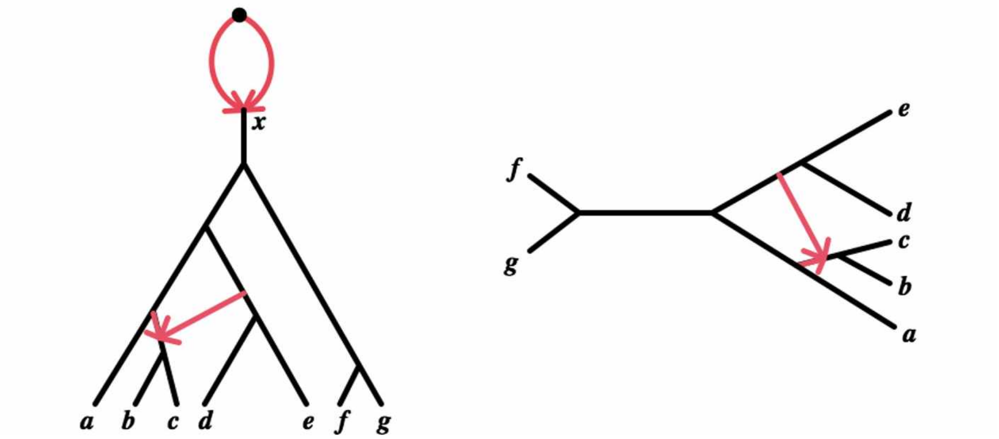
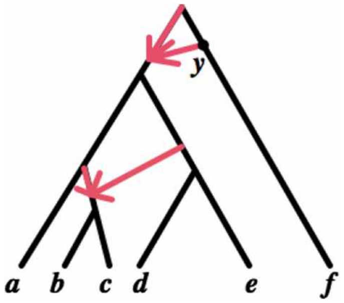


Fig. 1 (Left) A binary rooted phylogenetic network on X, with $\mathrm{LSA}(X)$ the node labeled x, and (Right) its induced unrooted semidirected network. In a depiction of a rooted network, all edges are directed downward, from the root, but arrowheads are shown only on hybrid edges. For the unrooted 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 $\mathrm{LSA}(X)$. $\mathrm{LSA}(X)$ 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 $\mathrm{LSA}(Z)$ for $\mathrm{LSA}(Z, \mathcal{N}^+)$. To see that $\mathrm{LSA}(Z)$ 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, $\mathrm{LSA}(X)$ need not to be the root of the network, as Fig. 1 (left) shows. Furthermore, there can be nodes below $\mathrm{LSA}(X)$ 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 $\mathrm{LSA}(Z)$ is two. Then the outdegree would be one, and the child of $\mathrm{LSA}(Z)$ would be in any path from the root to any taxa in Z, contradicting the definition of $\mathrm{LSA}(Z)$.

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 $Z \subseteq Y$, $\mathrm{LSA}(Y)$ must be equal or above $\mathrm{LSA}(Z)$ 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 $m = \mathrm{LSA}(Z)$, 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 $\mathcal{N}^+$.*

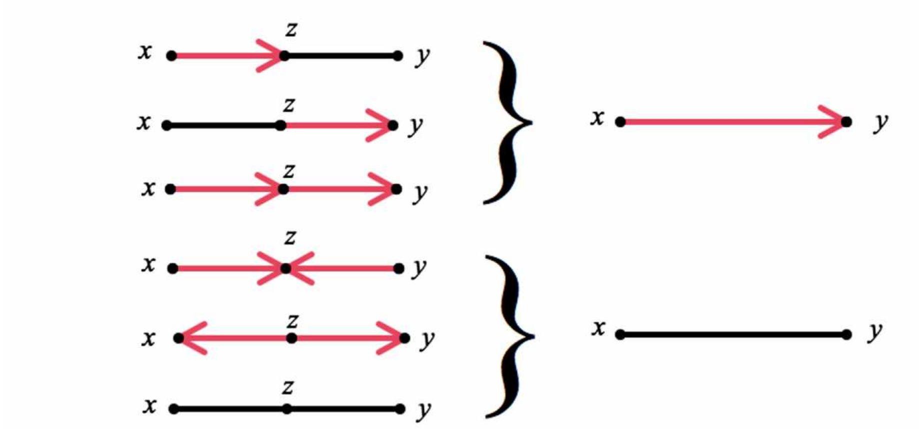


Fig. 3 On the left are all the semidirected graphs, up to isomorphism, on a degree two node z and its adjacent vertices x and y. On the right are the corresponding graphs obtained by suppressing z

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.

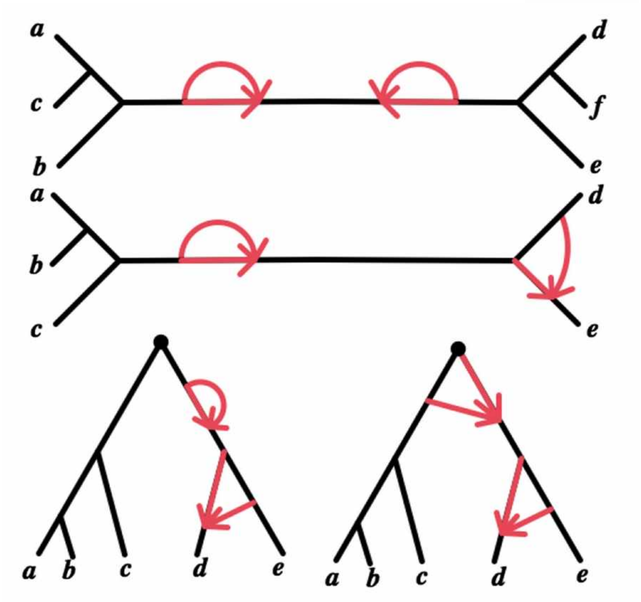


Fig. 4 The top graph is not a topological unrooted semidirected phylogenetic network, since its directed edges cannot be obtained by suppressing the root of any 6-taxon topological binary rooted phylogenetic network. The middle graph is the induced topological unrooted network from either of the bottom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

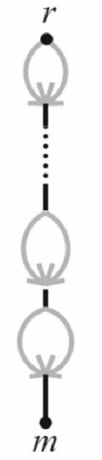


Fig. 5 In a level-1 network on X, the structure between the root and $m = \mathrm{LSA}(X)$ is a chain of two cycles. The number of two cycles in the chain could be zero

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 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.

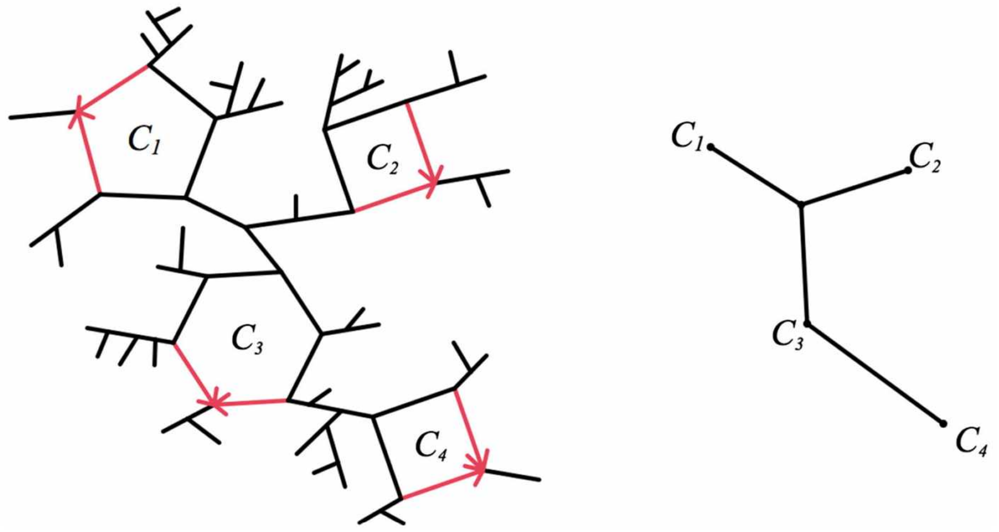


Fig. 6 (Left) A level-1 unrooted network $\mathcal{N}^-$ and (Right) the tree of cycles of $\mathcal{N}^-$

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 $e^{-z}(\gamma^2 P_1 + (1-\gamma)^2 P_2 + \gamma(1-\gamma)P_3 + \gamma(1-\gamma)P_4)$.

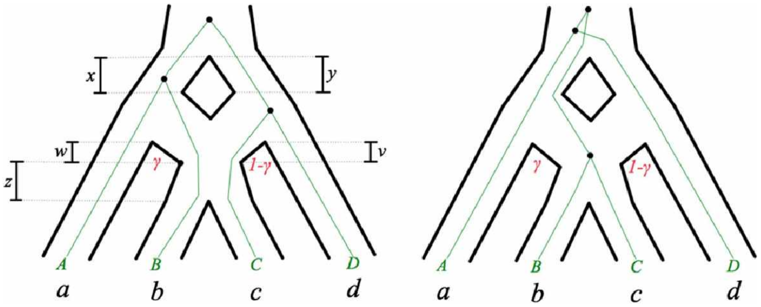


Fig. 7 Two gene trees within a species network with one hybrid node

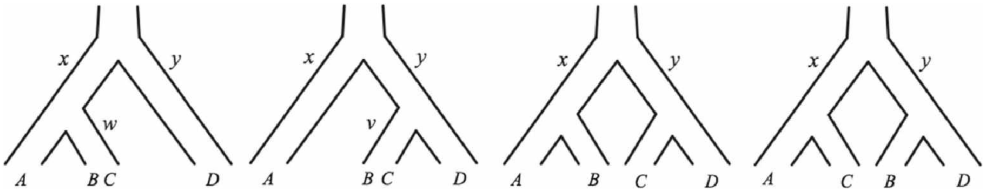


Fig. 8 Cases 1-4 (Left-Right) of Example 1, of how lineages may behave under the NMSC model on the network of Fig. 7

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.

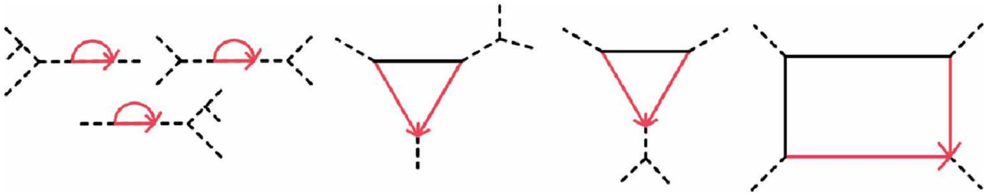


Fig. 9 (Left) The three types of 2-cycles in an unrooted quartet network (2_1-,2_2- and a 2_3-cycle); (Center) The two types of 3-cycles in the unrooted quartet network (3_1- and a 3_2-cycle). (Right) The only type of 4-cycle in an unrooted quartet network (a 4_1-cycle). The dashed lines represent subgraphs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

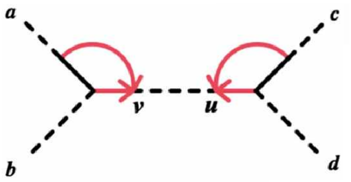


Fig. 10 A graph with two 3_2 cycles. Each dashed edge represents a chain of 2-cycles with, possibly, other cycles

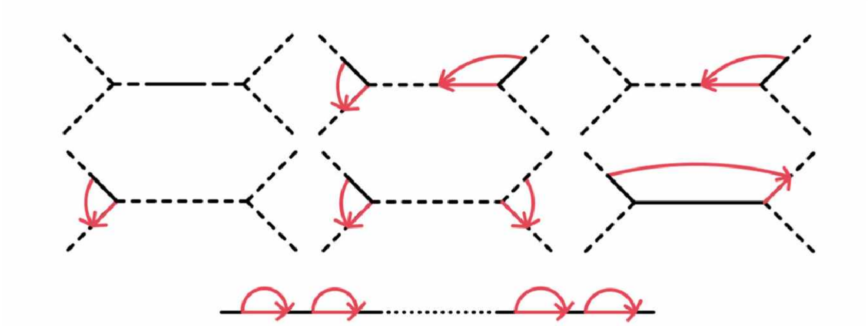


Fig. 11 Possible structures for unrooted quartet networks. Every dashed arrow represents a chain of an arbitrary number of 2-cycles, as the one in the bottom of the figure. The direction of these 2-cycles must be such that the obtained graph is induced from a rooted network

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

Proof Suppose $Q = Q^{-}_{abcd}$ 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.*

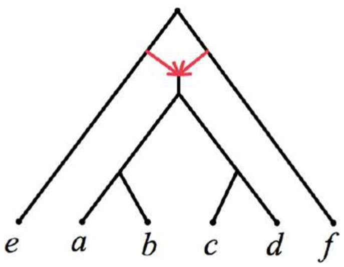


Fig. 12 A level-1 rooted network where the root differs from the LSA (a, b, c, d)

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 $Q^{\oplus}_{abcd}$.*

Proof In any realization of the coalescent process if there are fewer than 4 lineages at the LSA(a, b, c, d) in $\mathcal{N}^+_{abcd} = Q^+_{abcd}$, 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, $CF_{abcd}(Q^{\oplus}_{abcd}) = CF_{abcd}(Q^+_{abcd})$. □

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 $Q^{\oplus} = Q^{\oplus}_{abcd}$ 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:

- $Q^{\oplus}_1$ denotes the rooted quartet network obtained from $Q^{\oplus}$ by removing h_2.
- $Q^{\oplus}_2$ denotes the rooted quartet network obtained from $Q^{\oplus}$ by removing h_1.
- $Q^{\oplus}_0$ 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 $Q^{\oplus}_0$.

Note that $Q^{\oplus}_i$, 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 $Q^{\oplus}_i$ 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 $Q^\oplus = Q^\oplus_{abcd}$ 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 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*

$$CF(Q^{\oplus}) = \gamma^2 CF\left(Q_1^{\oplus}\right) + (1-\gamma)^2 CF\left(Q_2^{\oplus}\right) + 2\gamma(1-\gamma) CF\left(Q_0^{\oplus}\right).$$

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

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

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, $CF(Q_i^{\oplus} \mid K=1) = CF(Q^{\oplus} \mid K=1)$ for $i \in \{0, 1, 2\}$ and

$$\begin{aligned} CF(Q^{\oplus} \mid K=1) = \gamma^2 CF\left(Q_1^{\oplus} \mid K=1\right) + (1-\gamma)^2 CF\left(Q_2^{\oplus} \mid K=1\right) \\ + 2\gamma(1-\gamma) CF\left(Q_0^{\oplus} \mid K=1\right); \end{aligned} \tag{4}$$

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 $CF(Q_i^{\oplus} \mid K=1 \text{ or } K_y=1)$ are equal for $i \in \{0, 1, 2\}$, and

$$\begin{aligned} CF\left(Q^{\oplus} \mid K=1 \text{ or } K_y=1\right) = \gamma^2 CF(Q_1^{\oplus} \mid K=1 \text{ or } K_y=1) \\ + (1-\gamma)^2 CF\left(Q_2^{\oplus} \mid K=1 \text{ or } K_y=1\right) \\ + 2\gamma(1-\gamma) CF\left(Q_0^{\oplus} \mid K=1 \text{ or } K_y=1\right) \end{aligned} \tag{5}$$

Even though Eq. (5) is equal to $CF(Q_0^{\oplus} \mid K=1 \text{ or } K_y=1)$, 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,

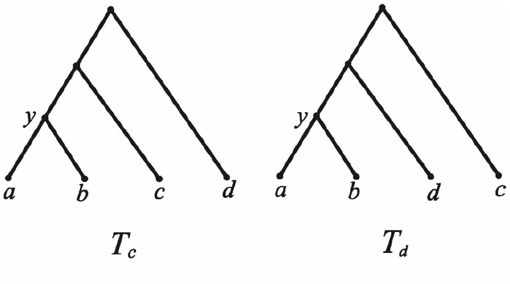


Fig. 13 The two trees T_d and T_c in the proof of Lemma 9, obtained when $K = 2$, $K_y = 2$ and the lineages c and d trace different hybrid edges

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

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,

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

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

Lemma 10 *Let $Q^{\oplus} = Q_{abcd}^{\oplus}$ 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 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

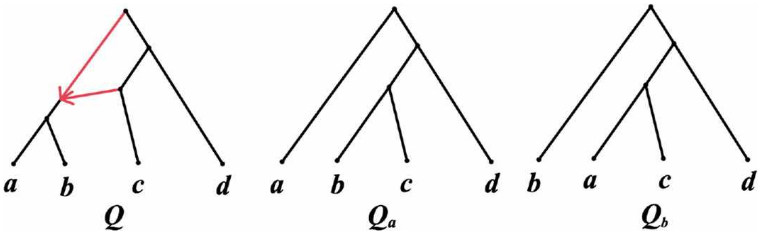


Fig. 14 A LSA quartet $Q^{\oplus}$ with a cycle C that induces a 3_2-cycle in the unrooted quartet and the graphs obtained by deleting everything below the hybrid node, disjointing, and labeling the leaves

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

Let $Q^{\oplus} = Q^{\oplus}_{abcd}$ be a metric level-1 LSA quartet network, and C_v a cycle in $Q^{\oplus}$, that induces a 3_2-cycle in 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 $Q^{\oplus} = Q^{\oplus}_{abcd}$ be a metric level-1 LSA quartet network, C_v be a cycle in $Q^{\oplus}$, that induces a 3_2-cycle in 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.

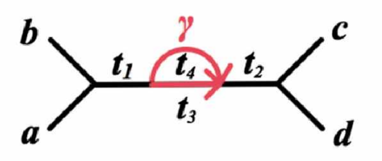


Fig. 15 An unrooted quartet with a single 2_2-cycle

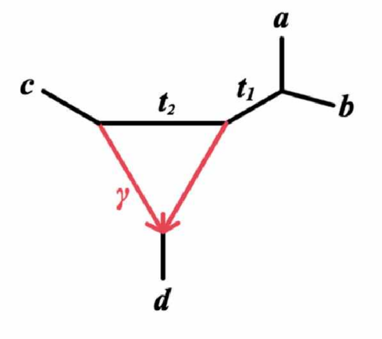


Fig. 16 An unrooted quartet with a single 3_1-cycle

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:

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

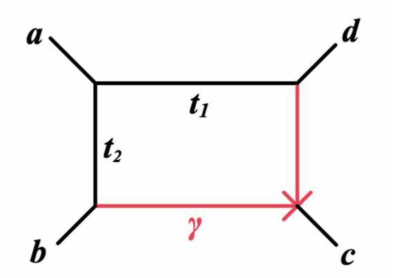


Fig. 17 An unrooted quartet with a single 4_1-cycle

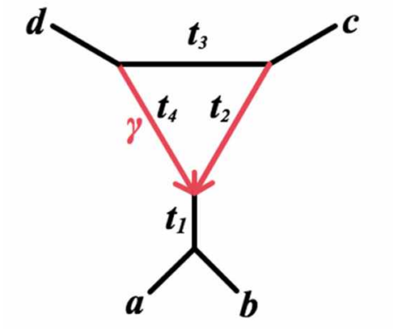


Fig. 18 An unrooted quartet with a single 3_2-cycle

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:

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

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.

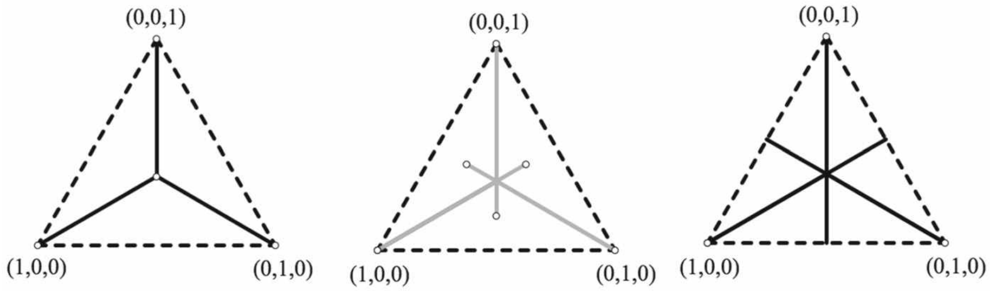


Fig. 19 On the left a planar projection of the simplex Δ_2, where the black lines represent concordance factors that are treelike. In the center, the gray segments in Δ_2 represent all the concordance factors arising from unrooted quartet networks with a 3_2-cycle. On the right, the black lines represent 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 $Q = Q^-_{abcd}$ 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 Q^-_s contains only a 3_2-cycle, as in Fig. 18. Then $CF(Q^-_s)$ 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 $CF(Q^-_s) \to \left(\frac{2}{12}, \frac{5}{12}, \frac{5}{12}\right)$. To minimize $CF_{AD|BC}$ it is enough to let $t_1 \to \infty$, so $CF(Q^-_s) \to (1, 0, 0)$.

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 $CF(Q^-_s)$ 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 $CF(Q^-_s)$ 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 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 $CF(Q^-_s)$. By Lemmas 9 and 10, we may suppress 2_2- and 3_1-cycles by expressing $CF(Q^-_s)$ as a convex sum of networks with a 3_2-cycle, but one fewer cycle. Thus, $CF(Q^-_s)$ 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.

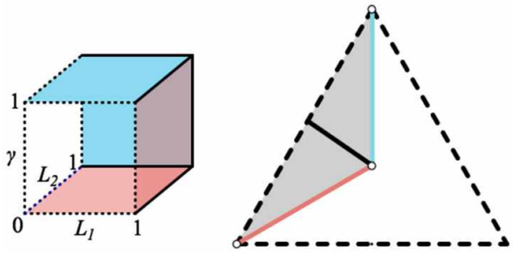


Fig. 20 The function f maps the cube χ_s (left) to Δ_2 (right). The blue facets (rear and top) of the cube are mapped by f to the blue (vertical) segment and the red facets (bottom and right) to the red (skewed) segment. The full cube is mapped onto the shaded triangle with all the concordance factor 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

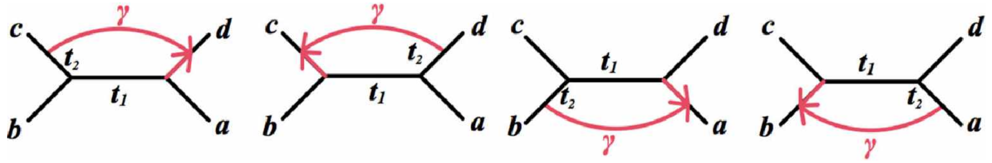


Fig. 21 Four unrooted metric level-1 quartet networks with the same concordance factors

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:

$$
\begin{aligned}
CF_{AB|CD} &= (1-\gamma)\left(1 - \frac{2}{3}e^{-t_1}\right) + \gamma\left(\frac{1}{3}e^{-t_2}\right),\\
CF_{AC|BD} &= (1-\gamma)\left(\frac{1}{3}e^{-t_1}\right) + \gamma\left(\frac{1}{3}e^{-t_2}\right),\\
CF_{AD|BC} &= (1-\gamma)\left(\frac{1}{3}e^{-t_1}\right) + \gamma\left(1 - \frac{2}{3}e^{-t_2}\right).
\end{aligned}
$$

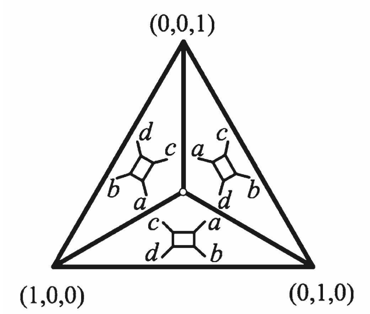


Fig. 22 Each section of the simplex is depicted with an unrooted quartet network topology whose image under the concordance factor map fills that region, independent of the placement of the hybrid node

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

$$Y_a = \underset{\{s\in S|a\in s\}}{\qquad} s\setminus\{a\}$$

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 $X\setminus Y_x$ 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 $q^{BC}_{a_1a_jxy} \neq a_1a_j|xy$ 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 $\{q^{BC}_{a_1a_pa_ja_m}\}$ 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 $q^{BC}_{a_1a_ja_px} \neq a_1a_j|xa_p$ 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 $q \in X \setminus P_u$. 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 $\widetilde{\mathcal{N}} \setminus \{u\}$ 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

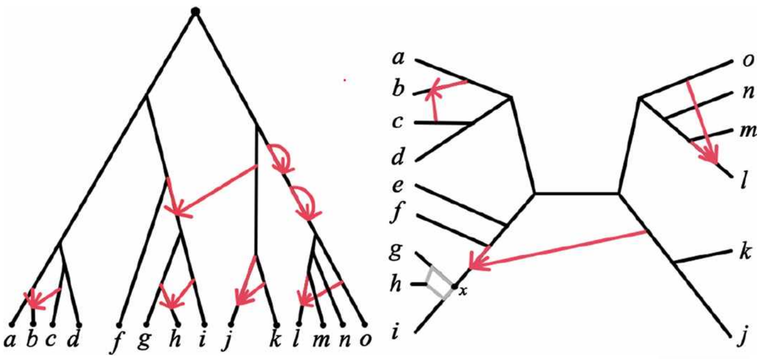


Fig. 23 A rooted metric phylogenetic network $\mathcal{N}^+$ (left) and the network structure $\widetilde{\mathcal{N}}$ (right) that can be identified by Theorem 4. The 4-cycle on the network in the right, colored gray, has 3 different candidates for the hybrid node

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 $\{a, b, c, d\}$ 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*

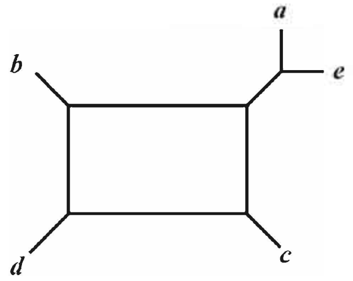


Fig. 24 A network $\tilde{\mathcal{N}}$ with a four cycle such that if $\{a, b, c, e\}$ satisfies the Cycle property, the hybrid block can be detected

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.

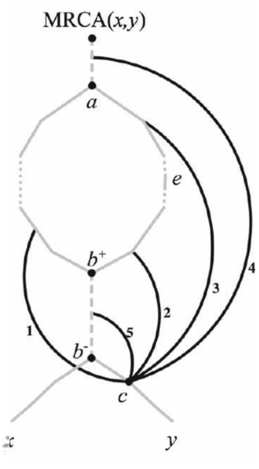


Fig. 25 In gray we see the subgraph composed by P and P', the dashed edges represent that P and P' could intersect, the dotted segments represent just a succession of edges. In black we see the different cases of the possible edges in P^* above b but below a

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

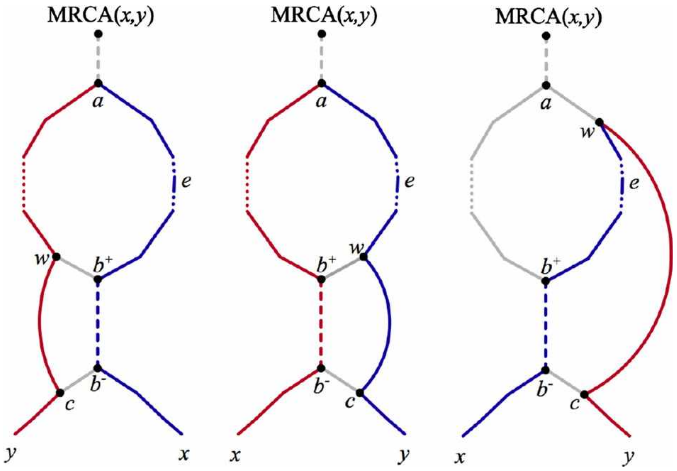


Fig. 26 The treks in case 1 (left), case 2 (center), and case 3 (right)

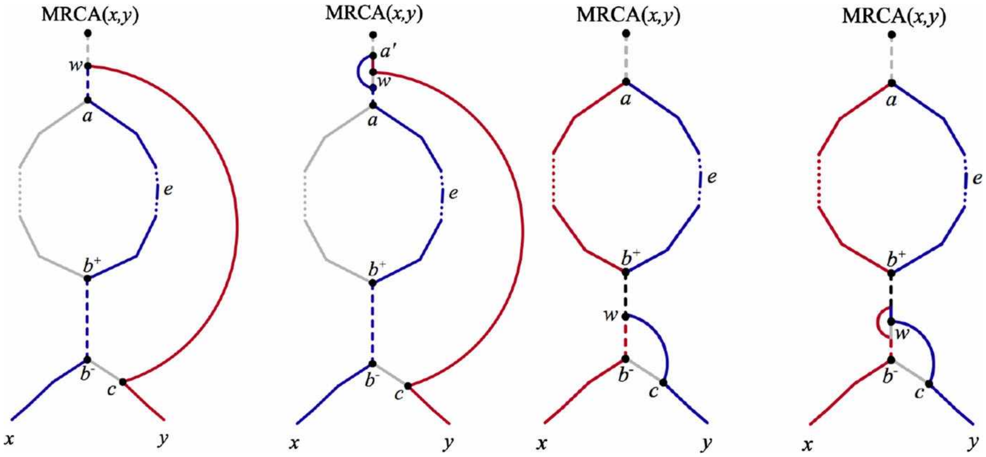


Fig. 27 (Left) The treks in the two possibilities of case 4. (Right) The two possibilities of case 5, where the black segments represent possible edges red and blue at the same time

LSA$(Z, \mathcal{N}^+)$≠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

Chapter 4: NANUQ: A method for inferring species networks from gene trees under the coalescent model

The proof of Theorem 4 in [*Baños*, 2019], presented in Chapter 3, suggests a method to infer level-1 species networks under noiseless data, but empirical data contains noise. The sources of noise include both stochastic noise from having a finite sample and additional noise from having obtained that "sample" not directly from the model but from an earlier inference of gene trees. In this chapter we present the paper "NANUQ: A method for inferring species networks from gene trees under the coalescent model," an inference method under the NMSC for a topological species network from gene tree data.

This method not only obtains a statistically consistent estimator but it is also robust under missing taxa on gene trees. Another strength of NANUQ is that it provides signals of when the NMSC and level-1 assumptions are not satisfied. This is a unique and novel feature of it, not present in any existing network inference schemes, and can help identify data sets that arise from other processes. NANUQ is designed to be an aid for biologists to analyze data in the suspected presence of hybridization.

We have implemented the first three steps of NANUQ in R [*R Core Team*, 2013], and use existing software, SplitsTree4 [*Huson and Bryant*, 2006], for the next two steps. The last step is an "interpretation" that is to be performed by the user.

Even though in this paper we mainly present the theory behind NANUQ, we also present the analysis of two empirical data sets of yeast and butterflies. We also analyze a data set that was simulated using Hybrid-Lambda [*Zhu et al.*, 2015].

My main contributions for this paper include the following:

- Initial conjecture explorations and coding with simulated data,
- first draft of some sections ("Phylogenetic networks", "Quartet concordance factors", "3_2-cycles", "Network split systems and distances", part of "Split networks from the network quartet distance", and "Examples"),
- contributions to the formulation and proofs of some theorems of the split graph section, and
- editing throughout.

NANUQ: A METHOD FOR INFERRING SPECIES NETWORKS FROM GENE TREES UNDER THE COALESCENT MODEL

HECTOR BAÑOS, ELIZABETH S. ALLMAN, AND JOHN A. RHODES

University of Alaska Fairbanks

Abstract. Species networks generalize the notion of species trees to allow for hybridization or other lateral gene transfer. Under the Network Multispecies Coalescent Model, individual gene trees arising from a network can have any topology, but arise with frequencies dependent on the network structure and numerical parameters. We propose a new algorithm for statistical inference of a level-1 species network under this model, from data consisting of gene tree topologies, and provide the theoretical justification for it. The algorithm is based in an analysis of quartets displayed on gene trees, combining several statistical hypothesis tests with combinatorial ideas such as a quartet-based intertaxon distance appropriate to networks, the NeighborNet algorithm for circular split systems, and the Circular Network algorithm for constructing a split graph.

1. Introduction

In this paper we provide the theory supporting a new, statistically consistent method of inferring most topological features of a level-1 hybridization network under the network multispecies coalescent (NMSC) model. The method uses as data a collection of unrooted topological gene trees, which may themselves have been inferred from sequences.

Unlike psuedo-likelihood methods [23, 28], it does not require an assumed limit on the number of hybridization events in the network, nor does it involve a time-intensive search over the space of possible networks. Instead, it computes a certain distance between taxa which, under ideal circumstances, corresponds to a circular split system. When this expected distance is processed through particular algorithms to produce a split graph, interpretation rules allow one to read off network information. The total theoretical running time of the algorithm is $\mathcal{O}(n^4 m)$ for an input of m binary gene trees on n taxa, making it computationally feasible when n has moderate size.

While we illustrate the method's utility through several examples with simulated and empirical data, our focus in this work is on providing its theoretical basis. This draws on a number of independent research works, but also requires new results on the nature of the split graphs that are produced under ideal circumstances.

We call this new method the <u>N</u>etwork inference <u>A</u>lgorithm via <u>N</u>eighbourNet <u>U</u>sing <u>Q</u>uartet distance, or by the acronym NANUQ[1]. It involves the following steps, applied to a collection of unrooted gene tree topologies assumed to have arisen under the NMSC on an unknown binary level-1 network:

Date: March 28, 2019.

[1] The word for "polar bear" in Inupiaq and other Inuit languages, pronounced and sometimes written as 'Nanook'.

(a) For each subset of 4 taxa, determine the empirical quartet count concordance factors from the gene trees, which will reflect possible cycles on the network, as shown in [3, 23].
(b) Apply a statistical hypothesis test to these counts, as in [2], to judge evidence as to whether or not the quartet species network displays a 4-cycle.
(c) Use the test results on quartets to construct a network quartet distance between taxa, extending the ideas of [19].
(d) Apply the NeighborNet [6] and Circular Algorithms [9] to construct a split graph from the quartet distance.
(e) Interpret the abstract network produced in the previous step by certain rules developed in Section 6 of this paper to infer most features of the unknown network.

All but the last step have been fully automated, in R for the steps (a–c), and SplitsTree4 [12] for step (d). While it is conceivable the last step could be as well, there are advantages to not doing so until more experience with the method has accumulated. For instance, some data sets may not support a hypothesis of evolution on a level-1 hybridization network, and a human interpretation of both the hypothesis test results and the SplitsTree4 output may suggest this, while simply showing a hybridization network most in accord with the output might be misleading.

While we show NANUQ is statistically consistent as an inference tool on its own, we suspect the greatest usefulness of this method is likely to be in an initial stage of data analysis, in which an empiricist seeks to explore the possibility and extent of hybridization among some taxa. Indeed, NANUQ offers several important advantages over other network inference methods we know of, in that it can indicate poor model fit to the level-1 NMSC and, in the case of reasonable fit, indicate the number of hybridization events. In contrast, pseudo-likelihood methods, which can be used for network inference [23, 28], are known broadly to be poor for judging model fit, though often perform well for inference. Moreover, while NANUQ gives information only on network topology, psuedo-likelihood can be used to obtain metric information as well. We thus view NANUQ as a complementary tool to the quartet-based pseudo-likelihood approach of SNAQ [23], and suggest using the two in tandem.

Several recent works [25, 30] have taken a Bayesian approach to inference of species networks from genetic sequence data, to obtain a joint posterior on both species networks and gene trees. As attractive as one might find this as a conceptual approach, it produces a formidable computational challenge for data sets with many taxa or gene trees. Indeed, the largest analysis in these works are quite small, involving only 7 taxa and 106 gene trees from a yeast dataset we also analyze. The alternative approaches offered by NANUQ and the pseudo-likelihood algorithms easily handle much larger datasets, with thousands of genes, as have already been assembled by researchers.

We note that this is the first instance, to our knowledge, of a split graph being given a firm interpretation as supporting a biological process underlying a data set. Split graphs are generally viewed as exploratory devices for judging the extent to which a data set is "tree-like," and authors often warn against interpreting them as supporting any particular biological mechanism. We fully agree with this general statement; only in the framework of our multi-step algorithm do we claim that an interpretation of support for a hybridization network is fully justified by theory. While an earlier step in this direction was taken by [13], that work assumed no incomplete lineage sorting was involved in the formation of gene trees, and provided less detailed description of the form of the split graph than does our work here.

The theory we present is based thoroughly on consideration of the quartets displayed on a collection of gene trees, but it differs in important ways from the more purely combinatorial work of [10] on undirected networks of level-1 and higher. First, we crucially focus on unrooted phylogenetic networks in the sense of [3], which retain the direction of hybrid edges from the rooted species network underlying the biological model, rather than fully undirected networks of [10]. This leads to a different notion of the trees and quartets displayed on the network, and the set of splits we associate to a network. Second, unlike most purely combinatorial studies, our algorithm takes into account that due to the coalescent process some gene trees will display quartets inconsistent with the species network. It nonetheless provides a means of determining, up to statistical inference error, which quartets are displayed on the network. Third, if these quartets are known exactly, we are able to recover not only the undirected version of the network (modulo contraction of 2- and 3-cycles) but even directions of hybrid edges in cycles of size 5 or larger.

This paper proceeds as follows: Sections 2 through 6 outline and develop theory behind our algorithm in a purely theoretical setting with no discussion of data. Section 7 more carefully outlines the algorithm for data analysis. Section 8 concludes with a few examples of network inference.

In more detail, Section 2 formally defines the type of rooted directed networks which underly our model, as well as unrooted semidirected networks induced from them. While this notion of unrooted network appeared in [3], it is not standard to the literature, yet it is essential to our work. Section 3 briefly recalls the network multispecies coalsecent model (NMSC) and the notion of a quartet concordance factor (CF). It summarizes results of [23] and [3] indicating how these concordance factors reflect quartet network topology, and provides a new analysis indicating the extent to which one can avoid the one important case of ambiguity in interpreting CFs. In Section 4 after recalling terminology for split systems, we define a split system associated to a unrooted semidirected level-1 network. In Section 5, we define a new quartet intertaxon distance for a level-1 topological network, and explore its relationship to the network. Section 6 investigates the form of a split graph computed from the quartet distance of a binary level-1 network. This requires establishing some new theoretical results which enable us to directly relate the form of a level-1 hybridization network to the form of the split graph of our network quartet distance.

Section 7 presents our algorithm in full, making use of all the theory above, as well as hypothesis testing using CFs as developed in [2], as well as the NeighborNet algorithm [6] and Circular Network Algorithm [9] as implemented in SplitsTree4 [12]. We give a running time analysis for NANUQ and establish its statistical consistency. As our primary goal in this paper is to provide the theoretical background to our algorithm, and not to extensively investigate its performance on simulated data or demonstrate its applicability to biological data, we offer only a minimial set of example analyses in Section 8. A later work, directed more at empiricists, will focus further on issues arising in data analysis.

2. Phylogenetic Networks

2.1. **Rooted and unrooted phylogenetic networks.** We begin by establishing terminology for phylogenetic networks. Throughout, $X = \{x_1, x_2, \ldots, x_n\}$ denotes a fixed set of taxa.

Our focus is on a explicit network [14], that can be interpreted as providing an evolutionary history of species relationships, including hybridization or other forms of lateral gene transfer that occur at discrete moments in time.

Definition 1 ([3, 24])**.** *A* topological binary rooted phylogenetic network 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* $E = E_H \sqcup E_T$*, together with 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* N^+ *be a topological binary rooted phylogenetic network with* $|E| = m$ *and* $|E_H| = 2h$*. A* metric for N^+ *is a pair* (λ, γ)*, where* $\lambda : E \to \mathbb{R}^{\geq 0}$ *assigns* edge lengths *and* $\gamma : E_H \to (0, 1)$ *assigns* hybridization parameters *satisfiing*

1. $\lambda(e) > 0$ *for* $e \in E_T$*,*
2. $\gamma(e_1) + \gamma(e_2) = 1$ *whenever* $e_1, e_2 \in E_H$ *have the same hybrid-node child.*

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

While the idea of unrooting a tree is simple, unrooting a network is more subtle. For example, it may not be clear how to proceed when the two edges incident to the root have the same child. We follow [3] in elucidating this concept.

In a directed 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 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).

Definition 3 ([24])**.** *Let* 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 directed path from the root* r *of* N^+ *to any* $z \in Z$*. Then the* lowest stable ancestor of Z of N^+*, denoted* $\mathrm{LSA}(Z)$*, is the unique node* $v \in D$ *such that* v *is below all* $u \in D$*,* $u \neq v$*.*

The lowest stable ancestor is a generalization (though not the only one) on a network of the concept of most recent common ancestor on a tree.

If z is a degree two node on a semidirected graph, with nodes x and y adjacent to z, then by *suppressing* z we mean deleting z and its incident edges, and introducing a new edge from x to y. If the deleted edges formed a semidirected path, we direct this new edge consistently with that path; otherwise the new edge is undirected.

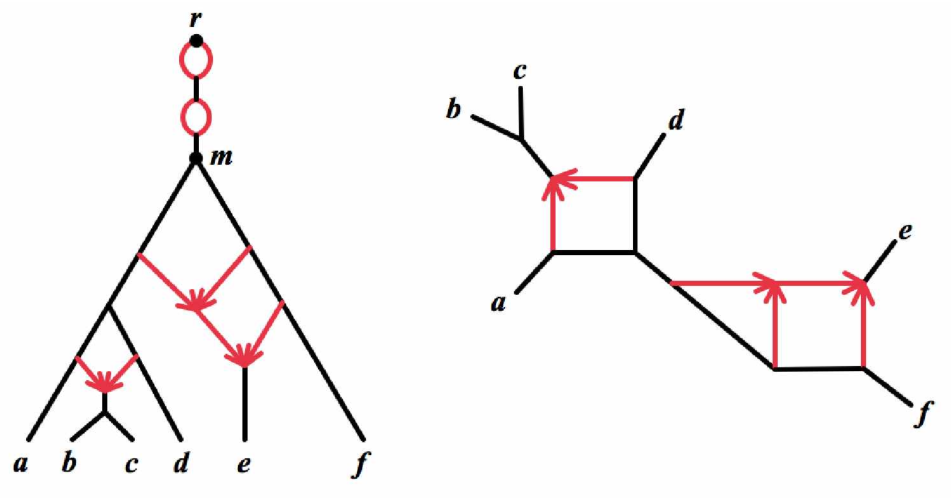


FIGURE 1. (Left) A rooted phylogenetic network N^+ with root r and lowest common ancestor m. (Right) The unrooted network N^- induced from N^+.

Definition 4. *Let N^+ be a binary topological rooted phylogenetic network on a set of taxa X. Then N^-, the* topological unrooted phylogenetic network induced from N^+, *is the semidirected network obtained by*

(1) *deleting all edges and nodes above* LSA(X),
(2) *undirecting all tree edges, and*
(3) *suppressing* LSA(X).

If N^+ has a metric structure, then N^- inherits one in an obvious way. Edge lengths on N^- are the sum of conjoined edge lengths in N^+, and hybridization parameters are the same as those on N^+.

Note that in some other phylogenetic works the term "unrooted network" is used for a fully undirected network. An unrooted network in our sense retains directions on hybrid edges, and thus encodes some information about possible root locations on N^+. Figure 1 depicts a topological binary rooted phylogenetic network on the left and its induced topological unrooted network on the right.

For simplicity, when we refer to an *unrooted network* N^- later in this paper, we always mean a semidirected network induced from a rooted binary phylogenetic network N^+ by this definition. That is, we implicitly assume the existence of N^+, even though there exist unrooted networks by the standard graph theory definition of that term which are not so induced.

Given that a unrooted network still retains some directed edges, a useful definition of induced quartet network is more elaborate than the analog for a tree. Recall a *trek* between vertices x, y on a network is the union of semidirected paths from some vertex v to x and from v to y. A trek is *simple* if the two paths intersect only at v.

Definition 5. *Let N^- be a unrooted network on X, and let $a, b, c, d \in X$. The* induced quartet network Q_{abcd} *is the unrooted network obtained by*

(1) *keeping only the edges in simple treks between elements of $\{a, b, c, d\}$, and*

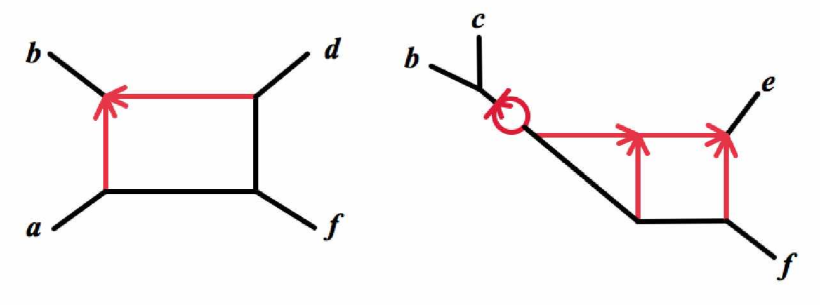


FIGURE 2. The quartet networks Q_{abdf} and Q_{bcef} respectively induced from the network on the right of Figure 1.

(2) *then suppressing all degree two nodes.*

If N^- is a metric network, Q_{abcd} inherits a metric structure as well. Edge lengths are simply sums of lengths of conjoined edges. Since a hybrid edge e in Q_{abcd} arises from conjoining a single hybrid edge $\tilde{e}$ of N^- with tree edges, the hybridization parameters for e is set equal to that for $\tilde{e}$.

Figure 2 shows several quartet networks induced from the unrooted network in Figure 1.

Finally, most of our results will be established only for a subclass of phylogenetic networks exhibiting a *level-1* structure. Although the definition we give is not the standard one for level-1 (e.g., [24]), it is equivalent for binary directed networks [22]. We also use it in the context of unrooted networks, which preserve the notion of hybrid nodes from any rooted version of them.

Definition 6. *Let N be a (rooted or unrooted) binary topological network. If no two cycles in the undirected graph of N share a vertex, then N is said to be* level-1.

3. The Network Multispecies Coalescent Model and Quartet Concordance Factors

The *multi-species coalescent model* (MSC) [18, 15] is the standard probabilistic model of incomplete lineage sorting, by which gene trees, showing direct ancestral relationships, form within species trees composed of mulit-individual populations. It traces, backwards in time, the lineages of a finite set of individual copies of a gene, sampled from different extant species, as they *coalesce* at common ancestors.

The *network multi-species coalescent model* (NMSC) [17, 27, 31] is a generalization of the MSC, which allows a finite number of hybridization events, or other discrete horizontal gene transfer events, between populations. Its parameters are captured by a metric, rooted phylogenetic network, which we assume to be binary, as defined in Section 2. At a hybrid node in the network, a gene lineage may pass into either of two ancestral populations, with probabilities given by the hybridization parameters γ for that node. This differs from other generalizations of the MSC, such a those built on a structured coalescent, where genes may switch populations continuously over an interval in time.

3.1. Quartet concordance factors.

The NSMC model is often used to obtain the probability of observing a specific gene tree (metric or topological, rooted or unrooted) in a species network. Our algorithm focuses on the probability that a species network produces various gene tree quartets (unrooted topological gene trees on 4-taxa) under the NMSC. The study of these probabilities, and their use for network inference, was pioneered in [23], with further work in [3]. A key concept is that of a quartet concordance factor, whose definition we recall.

A binary unrooted topological tree on four taxa a, b, c, d is called a *quartet*, and can be denoted as $ab|cd$ if deletion of its internal edge gives a connected component containing a and b. We say a large tree *displays* a quartet $ab|cd$ if the induced unrooted tree on the four taxa is $ab|cd$.

Definition 7. *Let N^+ be a metric rooted network on a taxon set X. Let A, B, C, D be genes sampled from species $a, b, c, d \in X$ respectively. Given a gene quartet $AB|CD$, the* concordance factor $CF_{AB|CD}$ *is the probability under the NMSC on N^+ that a gene tree displays the quartet $AB|CD$. The* concordance factor CF_{abcd} *is the ordered triple*

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

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

In a modeling context, we generally have a fixed rooted metric network N^+ in mind, but if there could be some ambiguity we denote the concordance factor by $CF_{abcd}(N^+)$. When a, b, c, d are clear from context (e.g., if N^+ has only four taxa), we write CF for CF_{abcd}. Also, while the language of 'concordance factor' is sometimes used for both theoretical values and empirical estimates, in this work we use it only for the first, being careful to refer to estimators of CFs, or empirical CFs, when they are found from gene tree data.

As established in [23, 3], the concordance factors for a level-1 network N^+ actually depend only on the unrooted N^-, and, more precisely, CF_{abcd} only depends on the quartet network Q_{abcd} induced from N^-. Finally, these concordance factors carry information about what cycles might be on that quartet network. To elucidate this, we review some of the results of those works.

In a level-1 network, each cycle has exactly one hybrid node. An n-cycle with exactly k taxa descended from its hybrid node is referred to as a *n_k-cycle.* In a level-1 quartet network there are 6 types of cycles that may appear: 2_1- ,2_2-, 2_3-, 3_1-, 3_2-, and 4_1-cycles as depicted in Figure 3. There are also several restrictions in the number and types of cycles of certain sizes that may simultaneously appear. For example, Q_{abcd} can have at most one of a 4_1-cycle or a 3_2-cycle.

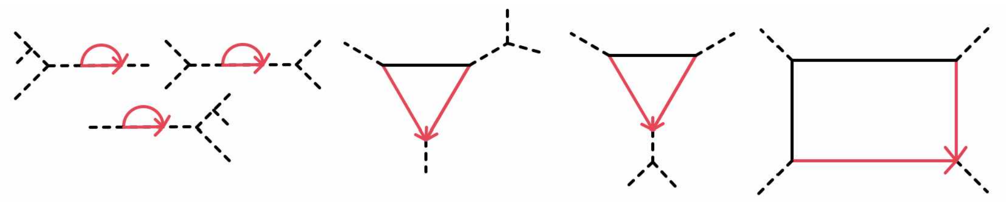


FIGURE 3. Cycles in a level-1 quartet network are classified as type n_k if they have n edges and k descendants of the hybrid node. The only cycles possible in a level-1 quartet network are (Left) of type 2_1, 2_2, and 2_3; (Center) of type 3_1 and 3_2; and (Right) of type 4_1. The dashed lines represent subgraphs that may contain other cycles.

Definition 8. *If the two smallest entries of CF_{abcd} are equal, then we say the concordance factor is* tree-like*. If a tree-like CF has a unique largest entry, $CF_{XY|ZW}$, we say it* supports *the quartet $xy|zw$. If all 3 entries are equal we say it supports all three quartets.*

This terminology arises from the fact that if the CF arises from a species *tree*, then it is tree-like, and its largest entry indicates the quartet species tree topology [1]. However, as was first shown in [23], certain types of non-tree networks also produce tree-like CFs under the NMSC.

Viewing a CF as a point in the probability simplex $\Delta^2 = \{(x_1, x_2, x_3)) \mid x_i \geq 0, \sum x_i = 1\}$, as in Figure 4, the tree-like CFs form the 3 line segments radiating from the central point $(1/3, 1/3, 1/3)$ to the vertices. With the ordering

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

the diagonal segment to $(1, 0, 0)$ is those CFs supporting $ab|cd$, the diagonal segment to $(0, 1, 0)$ is those supporting $ac|bd$, and the vertical segment to $(0, 0, 1)$ is those supporting $ad|bc$.

The following summarizes several results from [3]. By the *contraction* of a cycle, we mean removal of its edges together with identification of all vertices in it.

Proposition 9. *Let N^+ be a level-1 binary quartet network and $\widetilde{N}^-$ the network obtained from N^- by contracting all 2- and 3-cycles and then suppressing degree 2 nodes.*

If N^- has no cycles of type 4_1 or 3_2, then its CF is tree-like, and supports the quartet $\widetilde{N}^-$. That is if $\widetilde{N}^- = ab|cd$ then

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

If N^- has a 3_2-cycle, its CF may or may not be tree-like. If it is tree-like, it supports the quartet $\widetilde{N}^-$. If it is not tree like, the CF is on the line containing the tree-like CFs supporting the quartet $\widetilde{N}^-$. More precisely, if $\widetilde{N}^- = ab|cd$ then

$$CF_{AB|CD} \geq 1/6, \text{ and } CF_{AC|BD} = CF_{AD|BC}.$$

If N^- has a 4_1-cycle, then its CF is non-tree-like, and if $\widetilde{N}^-$ displays a 4-cycle joined to the taxa in circular order a, b, c, d, then

$$CF_{AB|CD} > CF_{AC|BD} \text{ and } CF_{AD|BC} > CF_{AC|BD}.$$

This proposition is perhaps most easily understood through Figure 4. Note that the CFs for binary quartet networks with no 3_2 cycles can be used to unambiguously determine whether that network had a 4-cycle or not. Applied to large networks through their induced quartet networks, this shows CFs carry some information about even large network topology.

But an issue that must be addressed is that, as Figure 4 shows, different quartet networks may give the same theoretical CFs. Most of these possibilities are harmless; for instance, as Proposition 9 indicates, the presence of any 2-cycle, or of a 3_1 cycle has no impact on what quartet is supported in the tree-like case, or what network(s) in the non-tree-like case. This ultimately leads to the non-identifiability of such cycles on a network by the method laid out in [3], and will similarly prevent NANUQ from detecting them. Since 2- and 3_1-cycles on a large network model 'hybridization' between the most closely related populations (two that split and then rejoin, or hybridization between two populations which have just split from a common one) inability to infer that such hybridization events occurred by our method may

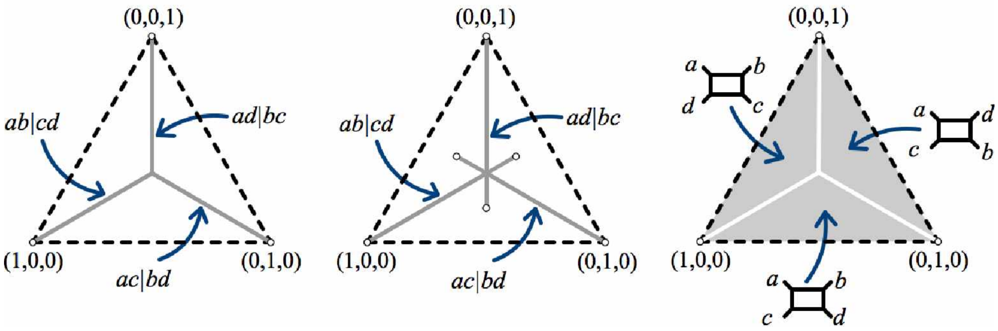


FIGURE 4. On the left a planar projection of the simplex Δ_2, where the gray lines represent concordance factors that arise from quartet networks with no 3_2-cycle and no 4-cycle (including quartet trees). In the center, the gray segments in Δ_2 represent all the concordance factors arising from quartet networks with a 3_2-cycle. On the right, the gray shaded area represents concordance factors arising from quartet networks with a 4-cycle with generic numerical parameters. In all figures, the quartets labeling line segments, and the quartet networks labeling regions, represent the quartet network after contraction of all 2- and 3-cycles.

not be too surprising. (Note that the SNaQ algorithm [23] is likewise unable to detect these, as it is based on CFs using a pseudo-likelihood framework.)

However, the CFs possibly arising from a 3_2 cycle must be carefully considered. If they are tree-like, then as with the 3_1-cycle we simply are unable to detect them via concordance factors. On the other hand, if they are non-tree-like, the same CF could have arisen from a 4_1-cycle. In the theoretical world, we understand that exact CFs from a network with a 4_1- cycle cannot match those from a 3_2-cycle for generic numerical parameters, since only a measure-zero subset of 4-cycle numerical parameters will place the CF where a 3_2-cycle CF may lie. Thus after excluding a negligible set of metric parameters from consideration, we can unambiguously determine those non-tree-like theoretical CFs arising from 3_2 cycles. We can then interpret these as supporting the quartet obtained by contracting all cycles in the quartet network, and identify the network with all 2- and 3-cycles contracted, as in [3].

An alternative to this is to argue that non-tree-like CFs from 3_2-cycles are 'rare,' and unlikely to occur in practice. This can be made more precise by determining what metric structure on a 3_2-cycle can produce a non-tree-like CF, and making a formal assumption that rules out that possibility. Thus we are motivated to study CFs from 3_2-cycles in more detail.

3.2. 3_2-cycles.

Let N^- be the unrooted quartet network shown in Figure 5, with parameters t_i in coalescent units and γ as shown. Let $x_i = e^{-t_i}$ for $i = 1, 2, 3, 4$. As shown in [3, 23], the quartet concordance factors of N^- are given by:

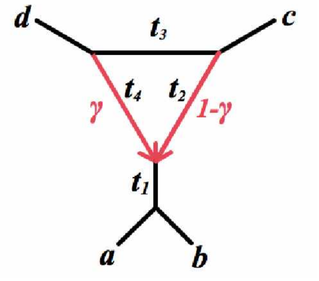


FIGURE 5. An unrooted quartet with a single 3_2-cycle. Internal edge lengths are denoted t_i, and hybridization parameters γ and $1-\gamma$.

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

We say that a choice of parameters $\{t_1, t_2, t_3, t_4, \gamma\}$, or their transformed versions x_i, is *tree-like* if the CF for the network is tree-like when given that metric structure.

The set of tree-like parameters for N^- is a region in the 5-dimensional cube,

$$0 \le \gamma, x_1, x_2, x_3, x_4 \le 1,$$

defined by the polynomial inequality

$$CF_{AB|CD} \ge CF_{AC|BD}.$$

To get a sense of the size of the tree-like region, we uniformly at random sampled 10^{10} points in $[0,1]^5$, and computed the proportion of these sampled parameters that are tree-like. We found that 99.468% of the sampled points were tree-like. Thus in some sense, non-tree-like CFs due to 3_2-cycles are rare. More precisely, since a uniform distribution for $x_i \in [0,1]$ corresponds to an exponential distribution with mean 1 for $t_i \in [0,\infty)$, if one assumes edge lengths are exponentially distributed random variables, and the hybridization parameter is uniformly distributed, the chance of a non-tree-like CF is estimated as ≈ 0.00532. While this number is quite small, we caution that we have no real justification for these priors on parameters (although exponential priors on branch lengths are often used in Bayesian approaches to tree inference).

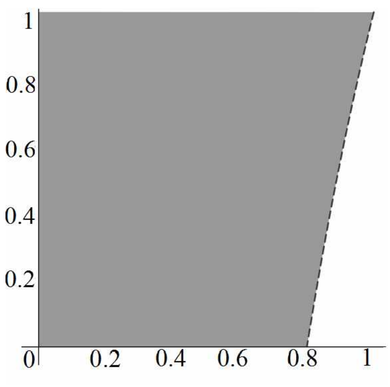


FIGURE 6. The shaded area in the x_1x_3 unit square indicates solutions of $x_1 \leq 4/(5 - x_3)$, which lead to treelike CFs on the network of Figure 5. Here $x_i = \exp(-t_i)$, with t_i the branch length in coalescent units.

For additional insight on tree-like parameters on N^-, we find

$$\begin{aligned}
(1)\qquad CF_{AB|CD} - CF_{AC|BD} &= 1 + (1-\gamma)^2(-x_1x_2) + \gamma(1-\gamma)(-3x_1 + x_1x_3) + \gamma^2(-x_1x_4) \\
&\geq 1 + (1-\gamma)^2(-x_1) + \gamma(1-\gamma)(-3x_1 + x_1x_3) + \gamma^2(-x_1) \\
&= 1 - x_1 + \gamma(1-\gamma)(-x_1 + x_1x_3) \\
&= 1 - x_1 - \gamma(1-\gamma)x_1(1-x_3) \\
&\geq 1 - x_1 - \frac{1}{4}x_1(1-x_3).
\end{aligned}$$

This last quantity is positive, and hence parameters are tree-like, when $x_1 \leq 4/(5 - x_3)$, a region shown in Figure 6. This region represents approximately 89% of the area of $[0,1]^2$. More crudely, if $x_1 \leq 4/5$ (that is, $t_1 \geq -\log(4/5) \approx 0.2231$) then a tree-like CF results regardless of all other parameter values. Thus non-tree-like parameters require that t_1 be fairly short, causing substantial incomplete lineage sorting. (For comparison, if the internal branch on a rooted 3-taxon species tree has length t and $e^t > 4/5$, then fewer than half of the gene trees will have the same rooted topology as the species tree under the MSC.)

While this argument assumed the network of Figure 5, a general level-1 quartet network with a 3_2 cycle may have additional 2_1-, 2_2-, and 3_1-cycles. One can show, though, that the same result applies as long as the t_1 is then taken as the length of the edge descended from the 3_2 hybrid node.

In particular, we have the following.

Proposition 10. *If on a level-1 network N^+ all branches descending from hybrid nodes have length $\geq -\log(4/5)$, then under the NMSC model all CFs will be tree-like except for those for which the associated quartet network has a 4-cycle.*

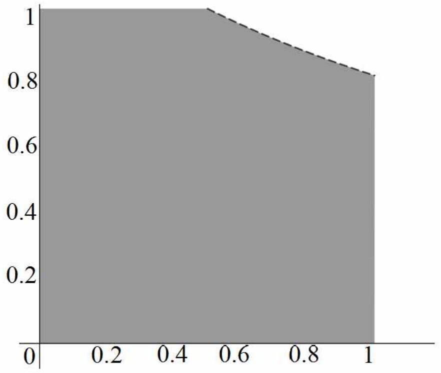


FIGURE 7. The shaded area represents the solutions of $M \leq \frac{2}{x_1} - \frac{3}{2}$ with the x_1-axis vertical and the M axis horizontal. With $M = \max(x_2, x_4)$, points in this region lead to tree-like CFs on the network of Figure 5.

Letting $M = \max(x_2, x_4)$, from equation (1) we also find

$$\begin{aligned} CF_{AB|CD} - CF_{AC|BD} &\geq 1 + (1-\gamma)^2(-x_1 M) + \gamma(1-\gamma)(-3x_1) + \gamma^2(-x_1 M) \\ &= 1 - x_1 M - \gamma(1-\gamma)x_1(3 - 2M) \\ &\geq 1 - x_1 M - \frac{1}{4}x_1(3-2M) \\ &= 1 - x_1\left(\frac{3+2M}{4}\right) \end{aligned}$$

Thus parameters will be tree-like if $M \leq \frac{2}{x_1} - \frac{3}{2}$. In particular, this is satisfied if $M \leq \frac{1}{2}$ ($\min\{t_2, t_4\} \geq -\log(1/2) \approx 0.694$), independent of all other parameters. Of course these are lengths of hybrid edges, and one might not want to make *a priori* modeling assumption that they even have positive length. Figure 7 shows the region where the inequality $M \leq \frac{2}{x_1} - \frac{3}{2}$ is satisfied. This region represents approximately 95% of the area of $[0,1]^2$.

While a large network may induce quartet networks with 3_2-cycles, which may be responsible for non-tree-like CFs, our goal here has been to suggest that one might make reasonable assumptions on branch lengths to rule out this possibility. Note that assuming there are no 3_2 cycles in a large network is not sufficient, since larger cycles, such as 4_2-cycles, will lead to 3_2-cycles on some of its induced quartet networks.

4. NETWORK SPLIT SYSTEMS AND DISTANCES

The ability to use quartet CFs to determine if a 4-cycle is on a quartet network, as discussed in the last section, will ultimately allow us to define a distance between taxa from which we can infer features of a larger network. But before we do that, we review the concept of a weighted circular split system for a set of taxa, and the distance associated to it. We also define the split system of an unrooted network in the sense of this paper, which differs from the standard definition due to hybrid edges being directed.

4.1. **Split systems.** We adopt standard terminology concerning splits [7]. A *split* $A|B = B|A$ of taxa X is a bipartition $X = A \sqcup B$ with A, B non-empty. The set of all splits of X is denoted by (X). A subset of (X) is often called a *split system* on X.

Definition 11. *A split system $\mathcal{S} \subseteq (X)$ is said to be* circular *if there exists a linear ordering $x_1 < ... < x_n$ of the elements of X such that each split in S has the form $A|B$ with*

$$A = \{x_p, x_{p+1}, ..., x_{q-1}, x_q\}$$

for appropriately chosen $1 \leq p < q < n$. The ordering of the x_i is called a circular ordering *for $\mathcal{S}$.*

A circular ordering for $\mathcal{S}$ can always be modified by cyclic permutations (e.g., replaced with $x_2 < x_3, < \cdots < x_n < x_1$) or inversion (replaced with $x_n < x_{n-1} < \cdots < x_1$) and will remain a circular ordering for $\mathcal{S}$. We treat such variants as the same, without further comment.

Given a tree T on X, deleting an edge defines a split according to the connected components of the resulting graph. The set of all such displayed splits is denoted $\mathcal{S}(T)$, and it is easy to see from a planar depiction of a tree that $\mathcal{S}(T)$ is circular.

For a tree, the correspondence between edges and displayed splits allows edge weights to be viewed as split weights, by setting weights of non-displayed splits to 0. This is a special case of a *weighted split system on X*, a map

$$\omega : (X) \to \mathbb{R}^{\geq 0}.$$

A weighted split system ω on X induces a distance function d_ω on X by

$$d_\omega(x, y) = \sum_{s \in S_{xy}} \omega(s),$$

where $S_{xy} \subseteq (X)$ is the set of splits separating x and y, i.e., splits $A|B$, with $x \in A$ and $y \in B$. Clearly d_ω is non-negative valued, with $d_\omega(x, x) = 0$, $d_\omega(x, y) = d_\omega(y, x)$.

Recall the support of a weighted split, denoted $\operatorname{supp}(\omega)$, is the set of splits on which ω is non-zero.

Definition 12. *A weighted split system ω on X is said to be* circular *if* $\operatorname{supp}(\omega)$ *is circular. A distance function d on X is said to be* circular *if $d = d_\omega$ for some circular weighted split system ω.*

As pointed out in [7], it follows from [4] that a circular distance function d uniquely determines the weighted split system ω such that $d = d_\omega$.

4.2. **Splits from unrooted networks.** Our notion of splits associated to a network, and sone related terminology, will not be standard, but is essential to this work. In particular, we focus only on phylogenetic unrooted networks as in Definition 4. We remind the reader that these networks are always assumed to be induced from rooted phylogenetic networks, and thus have additional features, including some directed edges, not implied by term "unrooted" as normally used in graph theory.

Definition 13. *Let N^- be a unrooted network on X. An unrooted tree T on X is said to be* displayed *on N^- if it can be obtained from N^- by deleting some edges, including at least one hybrid edge from each pair, undirecting remaining hybrid edges, and suppressing degree 2 nodes. The set of all unrooted topological trees on X that are displayed on N^- is called the* grove *of N^-, denoted $\mathcal{G}(N^-)$.*

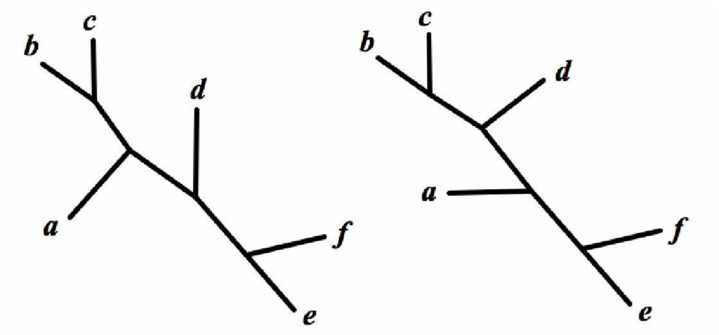


FIGURE 8. The two trees in the grove $\mathcal{G}(N^-)$, where N^- is the unrooted network of Figure 1.

The notion of displayed tree given here for unrooted networks extends that for an unrooted network given in [10], since in that work unrooted networks have no directed edges, hence there is no concept of a hybrid edge.

If N^- has either 2-cycles or a 3-cycles, then different choices of which hybrid edge in those cycles is deleted will yield trees with the same topology, and hence give the same element of $\mathcal{G}(N^-)$. For larger cycles, the choice of hybrid edge to delete does affect the tree topology. For a level-1 network N^- with k cycles of size ≥ 4, $|\mathcal{G}(N^-)| = 2^k$. This is, each tree in $\mathcal{G}(N^-)$ is determined by a unique choice of one hybrid edge in each cycle of size ≥ 4. This is not true more generally, as shown by the non-level-1 network of Figure 1 and the two trees in its grove displayed in Figure 8.

Definition 14. *For an unrooted network N^-, the set of splits*

$$\mathcal{S}(N^-) = \cup_{T \in \mathcal{G}(N^-)} \mathcal{S}(T)$$

is called the (unweighted) split system for N^-. *A* weighted split system for N^- *is any weighted split system with support $\mathcal{S}(N^-)$.*

From the study in [10] of undirected networks, we have the following.

Theorem 15 ([10])**.** *Let S be a split system on a set X and for any undirected network N on X let $\mathcal{S}(N)$ be the set of splits of all trees displayed on N. Then S is circular if and only if there exists an undirected level-1 network N such that $S \subset \mathcal{S}(N)$.*

Since the split system $\mathcal{S}(N^-)$ we associate to a semidirected unrooted network is a subset of the splits of all trees displayed on an completely undirected version of the network, we obtain the following.

Corollary 16. *If N^- is a level-1 unrooted network, then $\mathcal{S}(N^-)$ is circular.*

5. Quartet Distance for level-1 networks

As was shown in [19], topological trees have a natural metrization tied to the quartets displayed on a tree. Moreover, intertaxon distances from this metrization can be computed from the collection of displayed quartets, without having to know the full tree, giving a means of inferring the tree topology. After briefly reviewing this in the tree setting, we generalize it to the setting of level-1 networks.

5.1. **Quartet distance on a tree.** Recall a *quartet* means a binary unrooted topological 4-taxon tree, and is denoted by $ab|cd$ if taxa a, b and c, d are in different connected components when the internal edge is removed.

For an unrooted binary topological phylogenetic tree T on X, any edge e induces a partition of X into 4 non-empty blocks, X_1, X_2, X_3 and X_4, where the split associated to e is $s_e = X_1 \cup X_2 | X_3 \cup X_4$, and the splits associated to the 4 adjacent edges have an X_i as one split set. Similarly, a pendant edge e to taxon a induces a partition into 3 blocks X_1, X_2 and $\{a\}$, where $s_e = \{a\}|X_1 \cup X_2$, and the splits associated to the 2 edges adjacent to e have an X_i as one split set. The quartet weight function $w_T : (X) \to \mathbb{R}$ is defined as

$$w_T(s) = \begin{cases} |X_1||X_2| + |X_3||X_4| & \text{if } s = s_e \text{ for an internal edge } e, \\ |X_1||X_2| & \text{if } s = s_e \text{ for a pendant edge } e, \\ 0 & \text{if } s \text{ is not displayed on } T \end{cases}$$

This split weight function then induces d_{w_T}, the quartet distance function on X. The following theorem shows the distance function can be computed another way, from the set of quartets displayed on T.

Proposition 17. [19] *For any quartet q on taxa in X with $|X| = n$, let $\rho_{xy}(q) = 1$ if $q = xz|yw$ separates x, y, and 0 otherwise. Then for an unrooted binary tree T on X, and any $x, y \in X$,*

$$d_{w_T}(x, y) = 2 \sum_{q \text{ on } T} \rho_{xy}(q) + 2n - 4. \tag{2}$$

5.2. **Quartet distance on a network.** To generalize Proposition 17 to a binary unrooted network N^- on X, we begin with the following definition.

Definition 18. *The* quartet weight function ω_{N^-} *of a unrooted network N^- is defined by*

$$\omega_{N^-}(s) = \sum_{T \in \mathcal{G}(N^-)} w_T(s),$$

where $w_T(s)$ is the quartet weight function on T.

Note that since $\operatorname{supp}(w_T) = \mathcal{S}(T)$ for each T, $\operatorname{supp}(\omega_{N^-}) = S(N^-)$. Thus by Corollary 16, the quartet weight function ω_{N^-} is a weighted circular split system for N^-.

Lemma 19. *Let N^- be a level-1 unrooted network on X. Then*

$$d_{\omega_{N^-}} = \sum_{T \in \mathcal{G}(N^-)} d_{w_T}.$$

Proof. For $x, y \in X$, let $S_{xy} \subset (X)$ be the set of splits separating x and y. Then

$$\begin{aligned} d_{\omega_{N^-}}(x, y) &= \sum_{s \in S_{xy}} \omega_{N^-}(s) = \sum_{s \in S_{xy}} \sum_{T \in \mathcal{G}(N^-)} w_T(s) \\ &= \sum_{T \in \mathcal{G}(N^-)} \sum_{s \in S_{xy}} w_T(s) = \sum_{T \in \mathcal{G}(N^-)} d_{w_T}(x, y). \end{aligned}$$

□

To state a network analog of Proposition 17, we first extend the indicator function ρ_{xy} used in it to quartet networks.

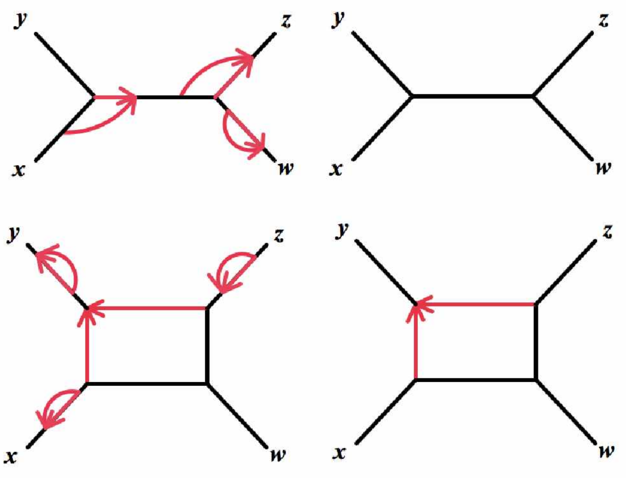


FIGURE 9. On the left, two quartet networks Q_{xyzw}, and on the right, the networks $\widetilde{Q}_{xyzw}$ obtained by contracting all 2- and 3-cycles and suppressing degree 2 nodes. For the top network $\rho_{xy}(Q_{xyzw}) = 0$ since x, y are not separated in the quartet. For the bottom $\rho_{xy}(Q_{xyzw}) = 1/2$, since x, y are separated on one of the quartet trees obtained by deleting a hybrid edge from $\widetilde{Q}_{xyzw}$ but not on the other.

Definition 20. *Let Q_{xyzw} be an unrooted level-1 4-taxon network on 4 distinct taxa $x, y, z, w \in X$. After contracting all 2- and 3-cycles, and suppressing degree 2 nodes, we obtain a network $\widetilde{Q}_{xyzw}$ that is either a tree or has a single 4-cycle. Let*

$$\rho_{xy}(Q_{xyzw}) = \begin{cases} 0 & \text{if } \widetilde{Q}_{xyzw} \text{ is a tree of the form } xy|zw, \\ 1/2 & \text{if } \widetilde{Q}_{xyzw} \text{ has a 4-cycle with } x, y \text{ adjacent in its circular ordering,} \\ 1 & \text{otherwise.} \end{cases}$$

This definition agrees with that in Proposition 17 in the case Q_{xyzw} is a tree. An intuitive way of viewing the extension to networks is that only 4-cycles change the definition, and in that case we take the average of the values we would get for the two trees obtained by dropping one or the other hybrid edge in $\widetilde{Q}_{xyzw}$. See Figure 9.

Lemma 21. *For a unrooted level-1 network N^-, with k cycles of size ≥ 4, and $x, y, z, w \in X$, let Q_{xyzw} be an induced unrooted 4-taxon network on x, y, z, w. Then*

$$\rho_{xy}(Q_{xyzw}) = \frac{1}{2^k} \sum_{T \in \mathcal{G}(N^-)} \rho_{xy}(T_{xyzw}).$$

Proof. If $\rho_{xy}(Q_{xyzw}) = 0$, then for no $T \in \mathcal{G}(N^-)$ will T_{xyzw} be a quartet separating x, y so the equation holds.

If $\rho_{xy}(Q_{xyzw}) = 1/2$, then $\widetilde{Q}_{xyzw}$ has the two hybrid edges, which are induced from hybrid edges of N^-. Each of these is deleted in exactly half of the 2^k trees in $\mathcal{G}(N^-)$, so 2^{k-1} of the $T \in \mathcal{G}(N^-)$ have T_{xyzw} displaying a quartet separating x, y. Thus the equation holds.

If $\rho_{xy}(Q_{xyzw}) = 1$, so $\widetilde{Q}_{xyzw}$ is either a quartet tree separating x, y, or has a 4-cycle with x, y opposite in its circular ordering, then for all $T \in \mathcal{G}(N^-)$, T_{xyzw} will display a quartet separating x, y, so the equation holds. □

We now define a distance function in terms of quartet networks displayed on the network.

Definition 22. *Let N^- be an unrooted level-1 network on X. Then the* quartet distance *d_{Q,N^-} is*

$$d_{Q,N^-}(x,y) = 2 \sum_{z,w \neq x,y} \rho_{xy}(Q_{xyzw}) + 2n - 4.$$

Note that if $N^- = T$ is a tree, $d_{Q,N^-}(x,y)$ reduces to the right hand side of equation (2).

The following is a network analog of Proposition 17.

Proposition 23. *Let N^- be an unrooted level-1 network on X, with k cycles of size ≥ 4. Then*

$$d_{\omega_{N^-}} = 2^k d_{Q,N^-}.$$

Proof. Using Lemma 19, Proposition 17, and Lemma 21, for $x \neq y \in X$,

$$\begin{aligned}
d_{\omega_{N^-}}(x,y) &= \sum_{T \in \mathcal{G}(N^-)} d_{w_T}(x,y) \\
&= \sum_{T \in \mathcal{G}(N^-)} \left(2 \sum_{q \text{ on } T} \rho_{xy}(q) + 2n - 4 \right) \\
&= 2 \sum_{T \in \mathcal{G}(N^-)} \sum_{q \text{ on } T} \rho_{xy}(q) + 2^k(2n-4) \\
&= 2 \sum_{T \in \mathcal{G}(N^-)} \sum_{z,w \neq x,y} \rho_{xy}(T_{xyzw}) + 2^k(2n-4) \\
&= 2 \sum_{z,w \neq x,y} \sum_{T \in \mathcal{G}(N^-)} \rho_{xy}(T_{xyzw}) + 2^k(2n-4) \\
&= 2 \sum_{z,w \neq x,y} 2^k \rho_{xy}(Q_{xyzw}) + 2^k(2n-4) \\
&= 2^k d_{Q,N^-}(x,y).
\end{aligned}$$

□

The import of this proposition is that from the induced quartet networks on N^- we can compute the distance d_{Q,N^-}, which is, up to scaling, $d_{\omega_{N^-}}$, the distance from a weighted split system. This contrasts with computing $d_{\omega_{N^-}}$ directly from its definition, which requires knowing $\mathcal{G}(N^-)$, the collection of trees on X displayed on N^-. This is at the heart of our algorithm for network inference under the network multispecies coalescent model, as we can obtain information about induced quartet networks from biological data relatively easily, using empirical concordance factors, while information about trees displayed on the species network does not seem to be directly obtainable.

Furthermore, since by Corollary 16 the underlying quartet weighted split system is circular, we have the following.

Corollary 24. *Let N^- be an unrooted level-1 network. Then the distance d_{Q,N^-} arises from a weighted circular split system, with support $\mathcal{S}(N^-)$.*

Thus given sufficient information on induced quartet networks to compute d_{Q,N^-} even approximately, methods for analyzing distances from weighted circular split systems, such as the NeighborNet algorithm, can be productively applied, as we show in the next section.

6. Splits networks from the network quartet distance

The last several sections have shown a path toward obtaining, under the NMSC model, the distance associated to the weighted circular split system ω_{N^-}. But for this to have value, we need to be able to extract from this distance information about features of N^-. While there is a well developed theory of split graphs associated to distances from such split systems, and split graphs are networks, one can not hope that such a split graph give N^- directly. In particular split graphs have no directed edges, and are generally not level-1.

Our goal in this section is to investigate the relationship between a level-1 network and the split graphs obtainable from the quartet distance for that network. We develop precise rules by which one can interpret features in a split graph for ω_{N^-} to obtain much information on the topological features of N^-. While there is some overlap between the results in this section and those of [13], we give a complete presentation as is necessary for our more detailed results.

In a level-1 unrooted network N^-, it is convenient to give terminology for several types of edges, in addition to the already defined tree and hybrid edges. A *cycle edge* is an undirected edge in a cycle. A *cut edge* is an undirected edge that is not a cycle edge. Thus any edge is either a cut edge, a cycle edge, or a hybrid edge. The cut and cycle edges together comprise the tree edges. A k-cycle is composed of $k-2$ cycle edges along with 2 hybrid edges.

For any $T \in \mathcal{G}(N^-)$, the edges of T arise from those of N^- in one of the following ways:

(1) An edge e of T is obtained directly from an edge of N^-. Then e is called a cycle or cut edge of T according to its classification in N^-.
(2) An edge e of T is obtained from several edges of N^- by suppressing internal nodes of degree 2. Since N^- is level-1, at least one of these conjoined edges of N^- is a cut edge, so we refer to e as a cut edge of T.

Note that edges in 2-cycles and 3-cycles on N^- induce *only* cut edges on any $T \in \mathcal{G}(N^-)$. For $k \geq 4$, a k-cycle on N^- will induce $k-3$ cycle edges on any $T \in \mathcal{G}(N^-)$, since one hybrid edge is deleted, one hybrid edge is conjoined with its descendent cut edge, and one cycle edge is conjoined with a cut edge.

A split $s \in \mathcal{S}(N^-)$ is called a *cycle split* (respectively, a *cut split*) if $s = s_e$ for a cycle edge (respectively, a cut edge) e on some $T \in \mathcal{G}(N^-)$. Note that the cut splits are precisely those splits obtained from N^- by deletion of a cut edge, and that these two classes of splits form a partition of $\mathcal{S}(N^-)$.

The following lemma indicates that the splits from a network N^- will show no sign of any 2- or 3-cycle on it.

Lemma 25. *Let $\widetilde{N}^-$ be the graph obtained from a level-1 binary network N^- by contracting each 2- and 3-cycle to a vertex and then suppressing degree 2 nodes. Then $\omega_{\widetilde{N}^-} = \omega_{N^-}$.*

Proof. If one or the other hybrid edge in a 2- or 3-cycle on N^- is deleted, the resulting network has the same topology as contracting the cycle. Thus N^- and $\widetilde{N}^-$ display the same topological trees. □

Lemma 26. *Let $s \in \mathcal{S}(N^-)$ for a level-1 binary network N^-. Then the following are equivalent:*

(1) *$s \in \mathcal{S}(T)$ for all $T \in \mathcal{G}(N^-)$,*
(2) *On every $T \in \mathcal{G}(N^-)$ there is a cut edge e such that $s = s_e$,*
(3) *s is compatible with every $s' \in \mathcal{S}(N^-)$.*

Proof. Clearly (2) implies (1). To see (1) implies (2), suppose on some tree $T \in \mathcal{G}(N^-)$ there is a cycle edge e with $s = s_e$. Then e arises from a cycle edge in N^- and that cycle has hybrid edges e_1 and e_2, where e_1 was deleted to form T. Then no tree $T' \in \mathcal{G}(N^-)$ which is formed by deleting e_2 will display s. This contradicts (1).

That (1) implies (3) is immediate. For the converse, observe that since N^- is binary, so is each $T \in \mathcal{G}(N^-)$. But the set of splits on a binary tree is maximal with respect to compatibility, so (3) implies (1). □

While Lemma 26 already implies that a split from a cycle edge in some $T \in \mathcal{G}(N^-)$ will be incompatible with some split from a cycle edge on another tree in $\mathcal{G}(N^-)$, this observation is refined by the following.

Lemma 27. *Let $s, s' \in \mathcal{S}(N^-)$ for a level-1 binary network N^-. Then s, s' are incompatible if and only if there are cycle edges e, e' (not necessarily distinct) on N^- in the same cycle C, and $T, T' \in \mathcal{G}(N^-)$ such that e, e' induces cycle edges $\bar{e}, \bar{e}'$ on T, T' with $s = s_{\bar{e}}, s' = s_{\bar{e}'}$ and T, T' were obtained by deleting different hybrid edges from C.*

Proof. Consider incompatible $s, s' \in \mathcal{S}(N^-)$. Then by Lemma 26, there exist $T, T' \in \mathcal{G}(N^-)$ with cycle edges $\bar{e}, \bar{e}'$ where $s = s_{\bar{e}}, s' = s_{\bar{e}'}$. The edges $\bar{e}, \bar{e}'$ are induced from unique cycle edges e, e' in N^-.

Suppose e, e' are in cycles $C \neq C'$. Now T determines a hybrid edge of C whose removal from N^-, along with the removal of e, determines the split s, and T' similarly determines a hybrid edge of C'. Removing these two hybrid edges, together with one hybrid edge from every other cycle on N^- determines a tree $T'' \in \mathcal{G}(N^-)$. But T'' has both s, s' as displayed splits, which implies they are compatible. Thus e, e' must be in the same cycle on N^-.

Moreover, T, T' must be obtained by deleting different hybrid edges in the cycle containing e, e', since if the same hybrid edge were deleted, the splits s, s' would be displayed on a common tree, and hence be compatible.

For the converse, suppose e, e' are cycle edges in cycle C of N^-, which induce cycle edges in trees $T, T' \in \mathcal{G}(N^-)$, where T, T' are obtained by deleting different hybrid edges in C. Let $X = X_0 \sqcup X_1 \sqcup X_2 \sqcup \cdots \sqcup X_n$ be the partition of X obtained from the connected components of the graph resulting from removing all edges of C from N^-. Suppose further that the ordering of these sets reflecting the ordering around the cycle, so that X_0 is descendants of the the hybrid node, and X_1, X_n are its neighbors, etc. Then, without loss of generality, we may assume that split s_e displayed on T is $X_0 \cup \cdots \cup X_k | X_{k+1} \cup \cdots \cup X_n$ with $1 \leq k \leq n-2$, while the split $s_{e'}$ displayed on T' is $X_0 \cup X_n \cup \cdots \cup X_{\ell+1} | X_\ell \cup \cdots \cup X_1$ with $2 \leq \ell \leq n-1$. These splits are incompatible as claimed. □

Split networks [14] provide a valuable visual tool for interpreting split systems. In these, each edge is colored by exactly one of the splits, with each split possibly coloring multiple edges. Deleting all edges with a common color leaves two connected components, with taxon labels on the components giving the split sets. Unfortunately split networks are generally not uniquely determined by split systems. However, since the split systems of interest here arise from level-1 networks N^-, and are thus circular by Corollary 16, we can impose a few additional requirements. The Circular Network Algorithm of [9] is the key to both showing split networks with additional properties exist in this case, and producing them in specific instances.

Recall that the *frontier* of a planar graph is the subset of edges adjacent to the unbounded component of its complement in the plane (more informally, the "outside" edges of the graph). A graph is *outer-labelled* if the labelled vertices are in the frontier. Also, a *blob* on a network is a maximal set of edges in undirected edge-intersecting cycles.

Lemma 28. *Let $S = S_c \sqcup S_i$ be a circular split system, with S_c the subset of splits compatible with all others in S, and S_i those incompatible with at least one other. Then the Circular Network Algorithm of* [9] *produces an outer-labelled planar split network N_S such that*

(1) *If $s \in S_c$, then s colors only one edge in the frontier of N_S, which does not lie in any blob.*
(2) *If $s \in S_i$ it colors precisely 2 edges in the frontier (and possibly additional edges not in the frontier) which lie in the same blob.*
(3) *If $s, s' \in S_i$ are incompatible, then they color frontier edges in the same blob.*

Proof. The Circular Network Algorithm works iteratively, by adding new vertices and edges as each split is considered in some order, to produce an outer-labelled split graph [9].

We may assume the trivial splits are in the system. The algorithm begins with these splits represented by a star tree, and the stated properties hold. Each time an additional split is considered, the algorithm 'duplicates' parts of the frontier, composed of edges labelled by splits incompatible with the new one, joining the duplicated section to the old part by 'ladder' edges colored by the new split.

In the case of the new split $s \in S_i$, this makes the frontier grow by 2 edges colored by the new split, and ensures that any previous splits incompatible with it that only colored one frontier edge will color two. Any two edges colored by the same split lie in the same blob, as do frontier edges coloring incompatible splits.

If the new split $s \in S_c$, then the duplication is of a single vertex, and only one edge is introduced and colored by that split. Moreover, this edge is not in a blob. □

The features of the split graphs produced by the Circular Network Algorithm above can be characterized in a less algorithm-dependent way. We call an outer-labelled planar split graph *frontier-minimal* if it contains the minimal number of frontier edges of all outer-labelled planar split graphs for the split system it depicts.

Proposition 29. *Any frontier-minimal split graph for a circular split system S has the properties listed in the previous lemma. The Circular Network Algorithm produces a frontier-minimal split graph.*

Proof. First, observe every split in S must label at least one frontier edge, else deletion of edges labelled by s would not disconnect the graph.

Recall the operation of contraction of a split s in a split network for $\mathcal{S}$, which identifies the two endpoints of each edge labelled by s and deletes the edge, yields a split network for $\mathcal{S} \setminus \{s\}$ (Lemma 5.10.1 of [14]). Note that frontier edges resulting from contraction must arise from frontier edges in the original split network. If $s, s' \in S_i$ are incompatible splits in a split network for $\mathcal{S}$, then contracting all other splits we obtain a split network depicting only these two. Now if only one frontier edge in this split graph were labelled by s, deletion of that edge must separate the graph. But then, since s' is incompatible, s' must label edges whose deletion disconnects each of the components obtained by deleting the s edge. But this implies that deleting only the s' edges separates the graph into at least 3 components, which contradicts that it is a split graph. Thus s labels at least 2 frontier edges.

Thus any split graph has at least $|S_c| + 2|S_i|$ frontier edges, and since this minimal count is achieved by the split graph output by the Circular Network Algorithm, a frontier-minimal split graph has $|S_c| + 2|S_i|$ frontier edges.

Since each element of S_i colors at least two frontier edges, and each element of S_c at least one, given the total count of frontier edges in a frontier-minimal split graph, it must be that elements of S_i color precisely two frontier edges, and elements of S_c precisely one.

Next, the one frontier edge labelled by an element of S_c cannot lie in a blob, else deleting it would not disconnect the graph.

If $s \in S_i$, then for any $s' \in S_i$ incompatible with s, contracting all splits but s, s' in a frontier-minimal split graph must give a split graph with 4 frontier edges. By considering all possible such graphs, these edges must form a 4-cycle (with edges labelled in order s, s', s, s'). Since these four edges are in the same blob on this graph, they must be in the same blob in the original graph. □

In [9] it is shown that the Circular Network Algorithm produces a split graph minimal in a different sense: it has the smallest number of edges among all split graphs whose bounded faces are parallelograms (i.e., quadrilaterals with opposite sides sharing colors). This addresses internal structure of the blobs, which our notion of frontier-minimal ignores. We have not investigated whether the two notions of minimality are equivalent, nor to what extent a frontier-minimal split graph for a circular split system is unique.

The *tree of blobs* of a graph is the graph obtained by contracting edges and vertices in each blob to a single vertex.

Corollary 30. *The tree of blobs of a level-1 network N^- is isomorphic to the tree of blobs of a frontier-minimal split network for $\mathcal{S}(N^-)$.*

Proof. The tree of blobs of N^- displays precisely those splits associated to cut edges of N^-. By Lemma 26, these are precisely the splits compatible with all others in $\mathcal{S}(N^-)$. But by Proposition 29, the tree of blobs of a frontier-minimal split graph displays the same set. □

To go further, we investigate how the structure of a blob (a cycle) in N^- corresponds to a related structure of a blob (*not* generally a cycle) in a frontier-minimal split network such as $\mathcal{S}(N^-)$. The following, which characterizes splits associated to a cycle in N^-, follows straightforwardly from definitions, so a formal proof is omitted. The argument is readily supplied by considering Figure 10, which depicts a single cycle in N^-, and the two networks obtained from it by deleting one hybrid edge.

Lemma 31. *Suppose a level-1 unrooted network N^- has k cycles of size ≥ 4. Let C be an m-cycle on N^-, $m \geq 4$,and $X = X_0 \sqcup X_1 \sqcup X_2 \sqcup \cdots \sqcup X_{m-1}$ the partition of X obtained from the connected components of the graph resulting from removing all edges of C from N^-. Suppose further that the ordering of these sets reflects the ordering around the cycle, so that X_0 is descendants of the the hybrid node, and X_1, X_{m-1} are its neighbors, etc. (see Figure*

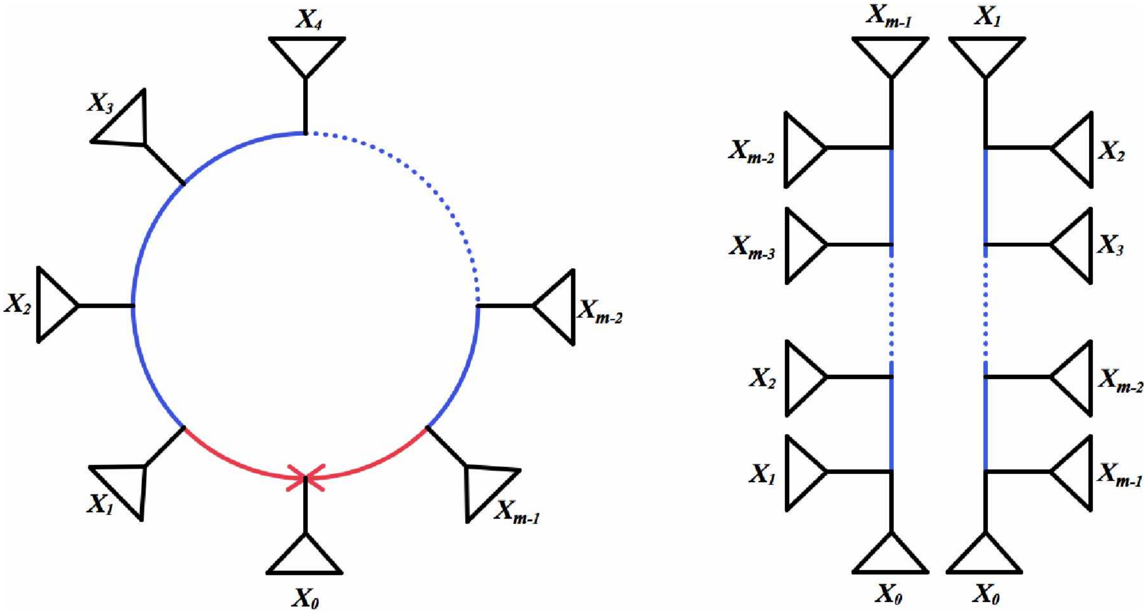


Figure 10. A cycle in a level-1 network (L) and the two simpler networks (R) produced from it by deleting one hybrid edge. The cycle edges in these networks that arise from the original cycle are shown in blue. If N^- has a single cycle, then the networks on the right are the two trees in $\mathcal{G}(N^-)$.

10). Then the cycle splits in $\mathcal{S}(N^-)$ arising from edges in C are

$$X_0 \cup X_1 \cup \cdots \cup X_i | X_{i+1} \cup \cdots \cup X_{m-1}, \quad 1 \leq i \leq m-3, \tag{3}$$

$$X_0 \cup X_{m-1} \cup \cdots \cup X_{j+1} | X_j \cup \cdots \cup X_1, \quad 2 \leq j \leq m-2, \tag{4}$$

all with $\omega_{N^-}(s) = 2^{k-1}$. Those splits of the form (3) (respectively (4)) are compatible with all others of that form. Spits of the form (3) are incompatible with those of the form (4). Splits of the form (3) or (4) are compatible with all other elements of $\mathcal{S}(N^-)$.

Moreover, $(X_0, X_1, X_2, \ldots, X_{m-1})$ is the only circular ordering of the X_i consistent with these splits, and with $X_m = X_0$ the number of cycle splits arising from C that separate X_i from X_{i+1} is

$$\begin{cases} m-3 & \text{if } i = 0, m-1, \\ 1 & \text{if } i = 1, m-2, \\ 2 & \text{otherwise.} \end{cases}$$

Lemma 32. *A frontier-minimal split graph for the cycle splits $\mathcal{S}(C)$ arising from a single cycle C of size $m \geq 4$ in N^- as in Lemma 31 forms a single blob whose frontier is a cycle of size $4(m-3)$. Moreover, there are distinct vertices labelled in circular order by $X_0, X_1, \ldots, X_{m-1}$ along the frontier, with the number of edges between labels X_i, X_{i+1} equal to the number of splits in $S(C)$ that separate X_i, X_{i+1}.*

Proof. Consider two splits associated to the cycle. By Lemma 31, they are either incompatible, or they are both incompatible with a third split from the same cycle. By Lemma 28, they therefore color edges in the same blob. Thus there is only one blob in the split graph.

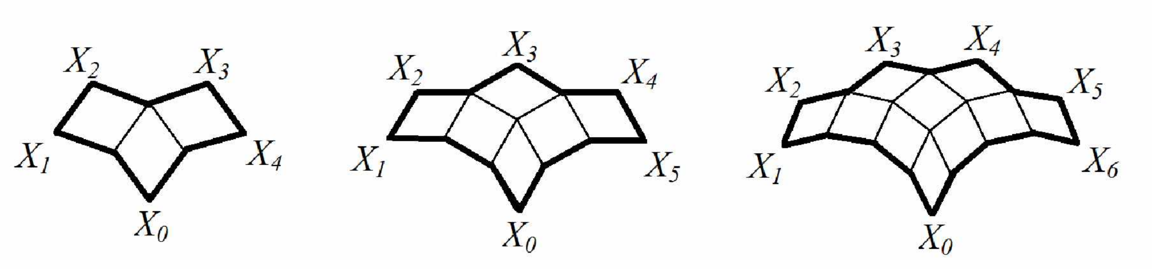


FIGURE 11. A m-dart, for $m = 5, 6, 7$ respectively. The frontier edges, shown in bold outline, are as characterized in the text. The outer vertices labelled by the X_i are the corners. The point of the dart is the unique corner which is $m-3$ frontier edges away from both of its closest corners. This figure was obtained by applying the Circular Network Algorithm implemented in SplitsTree4 [12] to appropriate distance matrices.

Since by Lemma 31 there are $2(m-3)$ splits associated to the cycle, by Lemma 28 the blob has $4(m-3)$ edges in its frontier.

Since from Lemma 31 there are splits separating any X_i, X_j, $i \neq j$, the X_i must label distinct vertices. Since any split separating X_i and X_{i+1} must label at least one edge in any frontier path between them, the number of edges in a minimal frontier path between X_i and X_{i+1} is at least the number of splits separating them. This then implies that the X_i must be in order along the frontier, at the distances claimed. □

Suppose C is an m-cycle in N^-. If $m = 4$ this lemma indicates a frontier-minimal split graph for the splits associated to C will also be a 4-cycle, that is, the undirected version of the cycle. However, if $m \geq 5$, the split graph is more complicated, having frontier as those depicted in the examples of Figure 11. We refer to such blobs as *m-darts*. The *corners* of the m-dart are the vertices on the frontier of the dart that are labeled by sets of taxa X_i. The *point* of the m-dart is the unique corner that is $m-3$ frontier edges away from the two closest corners, which is labelled by X_0. Thus, in a closed walk around the frontier of the dart, starting at the point, the number of edges between consecutive corners is

$$m-3, 1, 2, 2, \dots, 2, 2, 1, m-3.$$

Putting all this together, we have the following.

Proposition 33. *Given a level-1 unrooted network N^-, the frontier of any frontier-minimal split graphs for $\mathcal{S}(N^-)$ can be obtained by the following steps:*

1. *Contract any 2- and 3-cycles to vertices,*
2. *Undirect the hybrid edges in any 4-cycle,*
3. *Replace any m-cycle, $m \geq 5$ with an m-dart pointed at the hybrid node, with the m cut edges incident to the cycle connected to the corners of the dart in the same circular ordering as in the cycle.*

Proof. By Lemma 25, we may assume N^- has no 2- or 3-cycles. Let k denote the number of cycles of size ≥ 4 on N^-, and G a frontier-minimal split graph for $\mathcal{S}(N^-)$.

By Corollary 30, the tree of blobs of N^- and the tree of blobs of G are isomorphic, so we identify them. Moreover, since cycles in N^- are vertex-disjoint, each cycle of size

$m \geq 4$ on N^- gives rise to a node of degree m in the tree of blobs, so the tree of blobs has k multifurcations. This implies G has at least k blobs. Note however, though it could conceivably have more than k blobs, since if two shared a vertex they would be collapsed to a single node in the tree of blobs.

By Proposition 29 property (3) frontier edges colored by splits associated with a single cycle of N^- all lie in a single blob of G, since Lemma 31 shows two such splits are either incompatible, or both incompatible with a third. Thus G has at most k blobs, hence exactly k, and none share a vertex. Moreover, each blob has only splits associated to a single cycle coloring its frontier edges. Thus there is a one-to-one correspondence between cycles in N^- and blobs in G, according to the coloring of frontier edges.

Fixing a cycle C on N^-, and contracting all edges of G not labeled by splits associated to C preserves the frontier of the blob of G corresponding to C. By Lemmas 32, this frontier is either a 4-cycle (if $m = 4$) or an m-dart (if $m \geq 5$). Moreover, the partition of X according to the connected components of N^- with C deleted is the same is the same as that from the labeled corners of the 4-cycle or m-dart, with the same circular ordering, and in the case $m \geq 5$ the descendants of the hybrid node of C label the dart's point. Thus both C in N^- and the blob of G associated to C must map to the same multifurcation in the tree of blobs, and G must have the form described. □

Figure 12 illustrates this proposition for a particular network N^-. In that figure the frontier-minimal split graph produced by the Circular Network Algorithm, as implemented in SplitsTree4, is shown.

Importantly for applications, one can apply Proposition 33 "in reverse" to obtain information about the original network N^+ from the frontier-minimal split graph for $\mathcal{S}(N^-)$. Note that the correspondence between level-1 networks N^- and frontier-minimal split graphs as described in Proposition 33 is not one-to-one. In particular, one loses all information about 2- and 3-cycles, as well as an indication of which node in a 4-cycle is the hybrid one. However, for cycles of size $m \geq 5$, the form of an m-dart allows one to infer both the existence and ordering of an m-cycle in N^- and which node on it was hybrid. In conjunction with previous sections of this paper, this recovers the main result of [3]:

Corollary 34. *From the gene tree quartet frequencies of the NMSC model on a level-1 network N^+ with generic numerical parameters, one can identify the network obtained from N^- by suppressing 2- and 3-cycles and undirecting 4-cycles.*

Beyond providing a different argument for this corollary, Proposition 33 provides theoretical underpinnings to a practical algorithm for (partial) network inference from a sample of gene trees, as outlined in the next section.

7. The NANUQ algorithm for inference of phylogenetic networks

Here we revisit and formalize the NANUQ algorithm sketched in the introduction.

Algorithm (NANUQ).

Input: A collection of unrooted topological gene trees on subsets of an taxon set X, such that each 4-element subset of X appears on at least one tree; and two hypothesis testing levels $0 < \alpha, \beta < 1$.

(1) *For each subset of 4 taxa, determine the empirical quartet counts across the gene trees for each of the 3 resolved topologies. If all four taxa are not on a gene tree,*

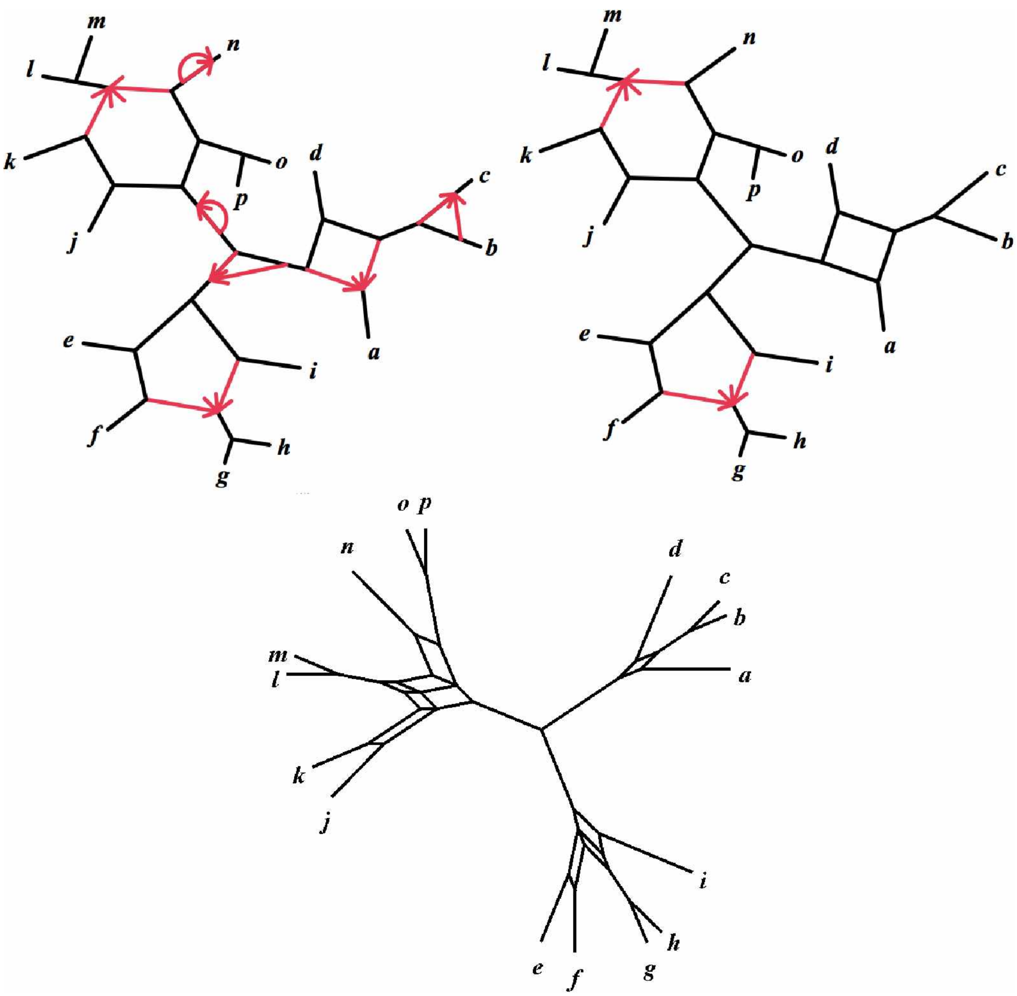


FIGURE 12. An unrooted level-1 network N^- (top L), the network obtained from it by contracting 2- and 3-cycles and undirecting 4-cycles (top R), and the split graph (bottom) obtained from it by Proposition 33. Note the split graph has a 4-cycle, a 5-dart, and a 6-dart, arising from the 4-, 5-, and 6-cycles of N^-. As described in Section 8, the split graph was obtained by applying the NANUQ algorithm to a large simulated data set of gene trees, and drawn by SplitsTree4 [12].

that tree does not affect the counts. These 3 counts form an empirical vector quartet count concordance factor *(QCCF) for the 4 taxa.*

(2) *For each set of 4 taxa, apply two statistical hypothesis tests to its QCCF, with levels α, β, as described below in subsection 7.1, to determine whether to view the QCCF as supporting (1) a star tree, (2) a resolved tree, or (3) a 4-cycle network on the taxa. In cases (2) and (3), use the maximum likelihood estimate of the topology from the QCCF to determine which tree or network is supported.*

(3) *Use the quartet networks/trees from the previous step to construct a network quartet distance between taxa, as in Definition 22, with the modification described below in subsection 7.2 for unresolved quartets.*
(4) *Use the NeighborNet Algorithm* [6] *to determine a weighted circular split system approximating the quartet distance.*
(5) *Use the Circular Network Algorithm of* [9] *to determine a split graph for the circular system.*

Output: A split graph to interpret via Proposition 33 for features of N^+.

To analyze the running time for this algorithm, suppose $|X| = n$ and the input set contains m trees. First note that tallying displayed quartets in Step (1) can be done in time $\mathcal{O}(n^4 m)$, as was shown in [19]. The hypothesis tests for Step (2) are performed in constant time for each set of 4 taxa, for a total of $\mathcal{O}(n^4)$. Step (3) in which the distance is computed requires running through the inferred quartet trees and networks for an additional time of $\mathcal{O}(n^4)$. The NeighborNet algorithm in Step (4) takes time $\mathcal{O}(n^3)$ [6]. Since NeighborNet might produce positive weights for all $\mathcal{O}(n^2)$ splits consistent with some circular ordering of the taxa, the time for the Circular Network Algorithm in Step (5) is $\mathcal{O}(n^4)$ [9]. Thus the total time for NANUQ is $\mathcal{O}(n^4 m)$.

We implemented Steps (1-3) of the NANUQ algorithm in an R package `MSCquartets`, with a function accepting an input file of gene trees, and producing an output file of the distances. When this file is opened by SplitsTree4 [12], steps (4) and (5) are performed. With these implementations, we have found step (1) by far dominates computational time, as is consistent with this analysis. However, the use of R probably slows computations considerably over what could be achieved.

The package `MSCQuartets` is currently available on request from the authors, and will be made publicly downloadable after further refinement.

7.1. **Testing Empirical Quartet Counts.** The statistical tests of step (2) of the NANUQ algorithm require more explanation.

We use a hypothesis testing framework, in which two tests are performed. One test is used to decide whether the topological signal in a QCCF is weak enough that rather than having any belief in a particular resolved network or tree, we should view the quartet as unresolved. The second test is used to decide whether the QCCF supports a 4-cycle network or a tree. The particular network or tree is then chosen via maximum likelihood.

These tests are performed for each set of four taxa as if they are independent. Of course they are not independent, since the quartet trees they consider are drawn from the same gene trees for all sets of 4 taxa, and the gene trees themselves are presumed to have formed in the same species tree or network. However their lack of independence depends in part upon the species tree or network, which is still unknown.

Suppose for a set of 4 taxa, one has tabulated the counts of the quartets displayed on the gene trees (ignoring those for which some of the taxa are missing) to obtain the QCCF. Assuming the NMSC model, these counts can be viewed as a multinomial sample from the distribution given by the theoretical CF. Normalizing by the total count, we obtain an empirical CF which estimates the theoretical one. Although in the tree-like case this is unlikely to lie exactly on the set where theoretical ones must, an appropriate statistical test can be used for deciding whether the QCCF supports a quartet tree or a network.

For a specific QCCF we first perform a hypothesis test for a star tree. More formally with null hypothesis

$$H_0: \text{ The QCCF arises from a 4-taxon star tree.}$$

the alternative is that it does not, which means it may have arisen from either a resolved tree or a network (or a more complicated model). As the star tree has theoretical CF $(1/3, 1/3, 1/3)$, we perform this test by computing the likelihood ratio statistic from the three quartet counts, using a χ^2 distribution with 2 degrees of freedom to compute a p-value.

With a chosen level β for the test, we reject the star tree hypothesis for p-values smaller than β. (Note that β is used here as the probability of a type I error, *not* the probability of a type II error.) For larger p-values, we view the QCCF as supporting the star tree.

Under the NMSC on a binary network, with enough data (sufficiently many gene trees), we should always reject star trees. However, with finite and noisy data, this test can be important to prevent interpreting a QCCF that is nearly uniform from indicating support for a particular tree or network topology. Assuming the data was produced by the NMSC on a binary level-1 network, performing this test has no effect on the asymptotic behavior of the algorithm as the amount of data increases. Nonetheless, performing it can suppress weak and possibly erroneous signals in finite datasets.

Next we perform a test for support for a tree. We formulate a null hypothesis of

$$H_0: \text{ The QCCF is tree-like.}$$

with alternative that it is not tree-like. However, assuming the NMSC model underlies our data, and we have restricted our presumed parameter space to avoid non-tree-like CFs from 3_2-cycles (as discussed in subsection 3.2), the alternative can be interpreted as the QCCF arises from a quartet network with a 4-cycle.

Geometrically, the model for the null-hypothesis is the 3 line segments in the left simplex of Figure 4. The alternative model is the remainder of the simplex, the complement of the 3 line segments. For the test, we compute the likelihood ratio statistic for this hypotheses. Using a standard χ^2 distribution with 1 degree of freedom (the asymptotic distribution for a resolved tree) to judge its value would be a standard approach. However, since the model has a singularity at the center of the simplex, and justification for the χ^2 depends on the model be approximated well by its tangent line, for finite sample sizes this may behave poorly in the vicinity of the singularity. While the region on which the asymptotic distribution behaves poorly shrinks as the sample size grows, it is present for any finite size. However, this particular model has been studied in [2], and an alternative approximate distribution, dependent on the sample size, has been developed to address this behavior near the singularity. We therefore adopt the test of that work for the likelihood ratio statistic, which returns a p-value.

For the algorithm with a chosen level α, we interpret a p-value greater than α as support for a tree. The particular tree topology is then chosen as the maximum likelihood estimate from the QCCF. This is simply the quartet topology with the largest count in the QCCF. A p-value less than α is interpreted as support for a 4-cycle network. The particular 4-cycle topology is taken as the maximum likelihood estimate from the QCCF, which is determined by which of the 3 triangular regions in the simplex it lies, as in Figure 4.

With two tests being performed in this way, it is possible that for a particular set of 4 taxa we find we fail to reject the first hypothesis (that the QCCF arises a star tree) but reject the second (that it arises from a tree). This can be forced to occur by taking β quite small while α is large, but it may occur for less extreme values. In a such a situation one must give priority to one test over the other. We choose to prioritize the first test, so that in such a case we view the tests as supporting a star tree, on the principal that evidence for hybridization should be judged by the strictest standards.

The output of NANUQ will depend on the choices of α and β, with smaller values of α requiring stronger evidence for 4-cycles, and smaller values of β requiring stronger evidence for any resolution of the network. Since "empirical" gene tree collections are likely to be noisy from the error introduced by inferring them from gene sequences, it is reasonable to set α quite small, which imposes a high standard for evidence of hybridization. There is no reason that α and β should be chosen to have equal values, and we believe appropriate choices of both will depend upon the level of noise in the data. In particular, *a priori* choices of conventional values such as 0.05 may be poor choices. Investigating the impact of a variety of choices of α and β on the final split graph is a necessary part of the analysis. We will briefly discuss this issue in Section 8 for a few simulated and empirical data sets, but defer more detailed comments to a future paper directed at empiricists.

The testing framework described here treats any QCCF judged non-tree-like as supporting a 4-cycle. As shown in subsection 3.2, the presence of a 3_2-cycle on a quartet network can, however, lead to a non-tree-like CF under some circumstances. By an assumption of sufficiently long edges descended from all hybrid nodes, one can rule such behavior out, using Proposition 10. Nonetheless, an empiricist may prefer not to make such an assumption. While in a future version of NANUQ we intend to offer a choice of using an additional statistical test for 3_2-cycle networks, such a test will also be nonstandard, due to the model being composed of three crossing line segments, and thus requires additional theoretical development. Moreover, the fundamental non-identifiability problem that some CFs may arise either from a 4-cycle or a 3_2-cycle means that in some circumstances a quartet network could still be miscalled.

7.2. **Quartet distance with unresolved quartets.**

The quartet distance defined for a binary network in Section 5 required that all quartet networks, after contraction of 2- and 3-cycles be binary, with positive lengths for all tree edges. However, in step 2 of the algorithm we include a hypothesis test for a star tree to reduce the possibility of calling a particular resolved tree or 4-cycle when the QCCF is nearly uniform and thus give weak evidence as to what the resolved topology should be. Thus we must explain how we modify the quartet distance computed in step 3 to deal with unresolved quartets.

We seek to modify the distance defined in Definition 22. To do this, we need only extend the Definition 20 of $\rho_{xy}(Q_{xyzw})$. Guided by the results in [19] on quartet distances for non-binary trees, we set

$$\rho_{xy}(Q_{xyzw}) = 1 \text{ if } \widetilde{Q}_{xyzw} \text{ is a star tree.}$$

In particular, this means a star tree is viewed as separating any two taxa on it.

Under our assumptions of a binary network, this modification has no impact on the asymptotic behavior of the algorithm under the NMSC model, as the chance of calling any star trees through the hypothesis test goes to 0 as the size of the data set grows.

7.3. **Statistical consistency.**

An inference procedure is statistically consistent for a particular model if the probability of inferring the correct result from a data set of size m produced in accord with the model approaches 1 as m approaches ∞. Since the NANUQ algorithm depends upon choices of two significance levels, α and β, these choices must be taken into account in formulating an appropriate notion of consistency for it.

Assuming the unknown network is binary, the value of $0 < \beta < 1$ will not matter in the limit as the data set grows. Informally, this is a consequence of the fact that the probability of rejecting the null hypothesis of a star tree will approach 1 for each choice of 4 taxa.

However, even for large data sets we expect to reject the null hypothesis of a tree even when the quartet network is tree-like with probability α in the limit. In other words, we will incorrectly call a positive proportion of the tree-like networks as 4-cycle networks. Even if we use a small value of α, this will put some error in the quartet distance, which the NeighborNet algorithm may not fully remove in projecting to a circular split system. Although in practice this error may be 'clear' to the human eye in viewing the split network, and easily removed by filtering out splits with small support, analyzing this error theoretically would be rather involved as it depends in understanding both the error introduced in the quartet distance and the impact of NeighborNet on it.

One solution to this problem of understanding the asymptotic behavior of the algorithm is to choose a sequence of values of α_m dependent on the sample size m so that as m increases the probability of calling correct tree-like networks (avoiding type 1 errors) goes to 1 while the probability of calling correct 4-cycle networks (avoiding type 2 errors) also goes to 1. Taking this approach, we formulate the following.

Proposition 35. *Under the NMSC model on a binary level-1 metric phylogenetic network N^+, for generic numerical parameters in which all induced quartet networks with 3_2-cycles are tree-like, there exists a sequence $\alpha_1, \alpha_2, \dots$, with $0 < \alpha_m < 1$ and $\alpha_m \to 0$ such that for any $0 < \beta < 1$ the NANUQ algorithm with significance levels α_m and β on a data set of m gene trees will, with probability approaching 1 as $m \to \infty$, infer the binary unrooted phylogenetic network associated to N^+ by Proposition 33.*

Proof. It is enough to show that the α_m can be chosen so that with probability approaching 1 the quartet distance computed in the NANUQ algorithm exactly agrees with the theoretical one based on the true network. This will follow from showing that as $m \to \infty$ with probability approaching 1 the hypothesis tests performed will all reject a star tree and either fail to reject a tree-like quartet network when the true one is tree-like, or reject a tree-like quartet network when the true one is non-tree-like.

Consider first the hypothesis test for a star tree for a particular choice of 4 taxa. While the result we need for this test is essentially a standard one, that the test is consistent, we give a full argument as an orientation to what will follow for the second test.

Since the network is binary, the expected CF is $(p_1, p_2, p_3) \neq (1/3, 1/3, 1/3)$. Then, since the likelihood ratio statistic λ for a sample involves the supremum of the log likelihood over the alternative hypothesis, it is easy to see

$$\lambda \geq \lambda_1$$

where λ_1 is the likelihood ratio statistics for a null hypothesis of $(1/3, 1/3, 1/3)$ vs. an alternative of (p_1, p_2, p_3). We view a sample of size m with QCCF (m_1, m_2, m_3) as arising from a multinomial sample with parameters (p_1, p_2, p_3). Then

$$\begin{aligned}\lambda_1 &= 2(m_1 \log p_1 + m_2 \log p_2 + m_3 \log p_3 - m \log(1/3))\\ &= 2m\left(\frac{m_1}{m}\log p_1 + \frac{m_2}{m}\log p_2 + \frac{m_3}{m}\log p_3 - \log(1/3)\right)\\ &= mX.\end{aligned}$$

where X is a random variable. By the law of large numbers X converges in probability to

$$c = 2(p_1 \log p_1 + p_2 \log p_2 + p_3 \log p_3 - \log(1/3)) > 0.$$

Thus for any $\epsilon > 0$ there exits an M such that $m > M$ implies $\mathbb{P}(X > c/2) > 1-\epsilon$, and thus that $\mathbb{P}(\lambda > mc/2) > 1-\epsilon$. With any significance level $0 < \beta < 1$ for the χ^2_2 distribution, we thus have that for any $\epsilon > 0$ there exists an M such that if $m > M$ then the probability of rejecting the null hypothesis is $> 1-\epsilon$. Thus as $m \to \infty$ the probability of rejecting the null hypothesis goes to 1. As there are only finitely many 4-taxon sets, the probability of rejecting the null hypothesis for all also goes to 1.

Turning now to the hypothesis test for a tree-like quartet network on 4 specific taxa, suppose first the true CF is tree-like. The likelihood ratio statistic is judged according to the distribution of the random variable $W_m = W$ described in Theorem 3.1 of [2]. Since the true network is binary, from results in that paper W_m has a limiting distribution as $m \to \infty$, which is χ^2_1. To ensure that the probability of failing to reject the null hypothesis approaches 1 as $m \to \infty$, it is enough to choose any sequence of significance levels $\alpha_1, \alpha_2, \dots$ with $\alpha_m \to 0$.

If the true CF is non-tree-like, however, we must pick significance levels more carefully. Without loss of generality, suppose the true CF is (p_1, p_2, p_3) with $p_1 \geq p_2 > p_3$. Then the likelihood ratio statistic satisfies

$$\lambda \geq \lambda_1$$

where λ_1 is the likelihood ratio statistics for a null hypothesis of tree-like-ness vs. an alternative of (p_1, p_2, p_3). If for a sample of size m the QCCF is (m_1, m_2, m_3), then

$$\begin{aligned}\lambda_1 &= 2(m_1 \log p_1 + m_2 \log p_2 + m_3 \log p_3 - m_1 \log(m_1/m) - (m_2+m_3)\log((m_2+m_3)/2m)\\ &= 2m\Big(\frac{m_1}{m}\log p_1 + \frac{m_2}{m}\log p_2 + \frac{m_3}{m}\log p_3\\ &\qquad\qquad - \frac{m_1}{m}\log(m_1/m) - \frac{(m_2+m_3)}{m}\log((m_2+m_3)/2m)\Big)\\ &= mY.\end{aligned}$$

where Y is a random variable. By the law of large numbers Y converges in probability to

$$d = 2(p_2 \log p_2 + p_3 \log p_3 - (p_2+p_3)\log((p_2+p_3)/2) > 0.$$

Thus for any $\epsilon > 0$ there exits an M such that $m > M$ implies $\mathbb{P}(Y > d/2) > 1-\epsilon$, and thus that $\mathbb{P}(\lambda > md/2) > 1-\epsilon$. Let $\alpha'_m = \mathbb{P}(W_m > md/2)$. Then we have that for any $\epsilon > 0$ there exists an M such that for $m > M$ the probability of rejecting the null hypothesis is $> 1-\epsilon$. Thus as $m \to \infty$ the probability of rejecting the null hypothesis goes to 1. As the W_m converge in distribution to a χ^2_1, one also sees that $\alpha'_m \to 0$.

Since there are a finite number of non-tree-like subsets of 4 taxa, we choose α_m to be the minimum of the α'_m for these subsets, to ensure the probability of rejecting the null hypothesis for all of them goes to 1 as $m \to \infty$. As $\alpha_m \to 0$, this sequence has all the desired properties. □

Note that the assumption of the above proposition that all 3_2-cyles are tree-like can be justified by, for example, Proposition 10 by requiring that no edges descending from hybrid nodes have length less than $-\log(4/5)$.

Although we do not give a formal proof here, NANUQ remains statistically consistent even in the absence of incomplete lineage sorting. Informally, one can "turn off" ILS in the multispecies coalescent model by shrinking all population sizes on the species network. Equivalently, if the species network's branch lengths, measured in coalescent units, go to ∞, then the distribution of rooted topological gene trees approaches that of a hybridization model with no ILS. One can thus establish consistency either by taking appropriate limits in the argument above, or by analyzing quartet concordance factors for the pure hybridization model directly.

7.4. Sources of Error.

While NANUQ is a statistically consistent (in the precise sense of Proposition 35) method of inferring certain network features from a collection of gene trees produced by the NMSC model, in practice it must be applied to a finite set of inferred gene trees. Possible sources of errors of conclusions drawn from NANUQ include:

1. Input of incorrect gene trees, due to their inference from sequence data,
2. Small sample size (number of gene trees),
3. Miscalls of evidence for/against hybridization in individual quartets, in step 2,
4. The NeighborNet Algorithm's projection of the split system onto a circular one, in step 4,
5. The presence of non-tree-like 3_2-cycles on some induced quartet networks,
6. NMSC model misspecification due to any of:
   (a) a non-level-1 network,
   (b) structure within populations,
   (c) continuous gene flow between populations.

Thus one should not expect empirical data to necessarily lead to a split graph exactly conforming to form described by Proposition 33.

Note that the algorithm of [20] offers an alternative to NeighborNet that might reduce the error arising in passing to a circular split system from the quartet distance. However, this has not been implemented in general purpose softwares yet, so we were unable to test its performance.

We have chosen not to suggest any automatic interpretation of the output of NANUQ, such as a mechanism for producing the closest split network (by some measure) that conforms exactly to the form described by Proposition 33. Thus the user must visually consider the output, which shows some of the error. In particular, SplitsTree offers a capability of removing splits with small weight from a split graph, and this can be useful for removing some of the noise remaining after projecting onto a circular split system.

8. Examples

In this section we present three examples of data analysis with NANUQ. The first uses a simulated data set of gene trees (without any gene tree inference error), the second the well-known and well-studied yeast data set of [21], and the third a butterfly data set of [16]. For the empirical data sets, we use gene trees previously inferred from genetic sequences.

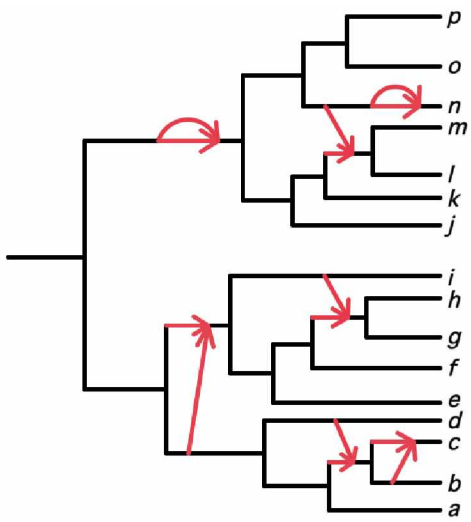


FIGURE 13. The species network N^+ of Example 36. See Newick notation in text for branch lengths and hybridization parameters.

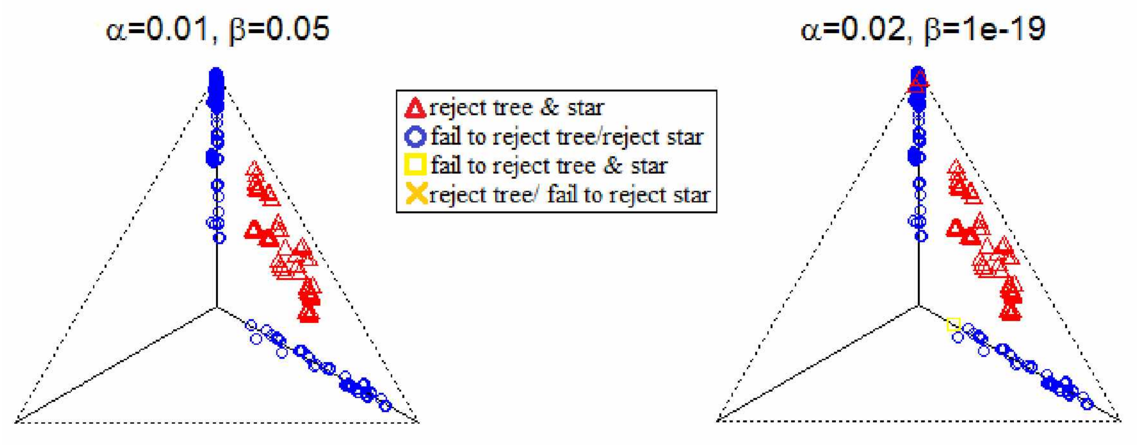


FIGURE 14. Representative simplex plots from empirical QCCFs, with hypothesis testing results.

Example 36. *We first generated a data set of 1000 gene trees using Hybrid-Lambda* [32] *with the species network N^+ shown in Figure 13, with branch lengths in coalescent units and hybridization parameters given in extended Newick format with internal node labels by*

$$
\begin{gathered}
(((((a{:}1.5,(((b{:}.8,h1\#.5{:}.1)x1{:}.2,(c{:}.7)h1\#.5{:}.3)x2{:}.3)h2\#.5{:}.2)x3{:}1.5,(h2\#.5{:}.2,d{:}1.5)\\
x4{:}1.5)x5{:}2,h3\#.5{:}1.5)x6{:}0.5,(((e{:}2,(f{:}1,((g{:}.25,h{:}.25)x7{:}.25)h4\#.5{:}.5)x8{:}1)x9{:}1,\\
(h4\#.5{:}.5,i{:}1)x10{:}2)x11{:}0.5)h3\#.5{:}2)x12{:}1,((((j{:}4.5,(k{:}3.5,((l{:}2.75,m{:}2.75)x13{:}.25)\\
h5\#.3{:}.5)x14{:}1)x15{:}1,((((n{:}1)h6\#.5{:}2,h6\#.5{:}2)x16{:}.5,h5\#.3{:}.5)x17{:}1,(o{:}3.5,p{:}3.5)x18{:}1)\\
x19{:}1)x20{:}.25)h7\#.5{:}.5,h7\#.5{:}.5)x21{:}.25)r;
\end{gathered}
$$

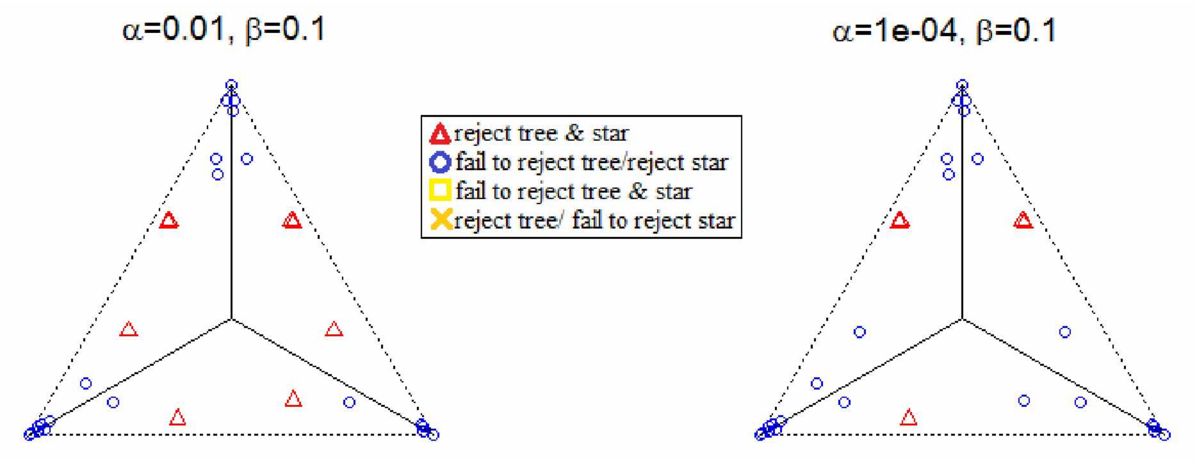


FIGURE 15. Simplex plots for hypothesis test results on the yeast data set, with two choices of significance levels $\alpha = 10^{-2}$ and 10^{-4} with $\beta = 10^{-1}$. The choice of β here is largely irrelevant, as no plotted empirical CFs are near the center. Smaller α results in fewer empirical CFs being called as supporting 4-cycles.

This network has unrooted version N^- as shown in Figure 12.

Running NANUQ on this dataset, our implementation of steps 1-3 in R required about 76 s of computation time on a desktop computer. We considered several values of α and β for our hypothesis tests. To visualize outcomes of the hypothesis tests, we produced simplex plots such as those shown in Figure 14, which plot empirical CFs (i.e., QCCFs normalized to sum to 1) for each set of 4 taxa, color coded to indicate test outcomes. Due to the rather clean separation of empirical CFs into those close to the 3 tree-like lines and those farther from them we found that for $10^{-17} \leq \alpha \leq .01$ we drew the same conclusions as to which QCCFs supported a 4-cycle. The close clustering around the lines of points not rejected as tree-like also suggest little error in them, so that a rather large value of β might be sufficient to test for star trees. Using $\alpha = .01$ and $\beta = .05$, from SplitsTree4 we obtain the split graph on left of Figure 12. Under the rules of Proposition 33, this leads to the correct inference of all features of N^- inferable by the method.

Reducing the sample size to 300 gene trees, using the same value of α and β, we obtained the same correct inference result.

Example 37. *For the second example we use a subset of the yeast data set of* [21], *which has been analyzed by multiple investigators* [5, 11, 25, 26, 29, 30]. *The data set consists of 106 gene trees, each with a single allele sampled from seven Saccharomyces species: S. cerevisiae (Scer), S. paradoxus (Spar), S. mikatae (Smik), S. kudriavzevii (Skud), S. bayanus (Sbay), S. castellii (Scas), S. kluyveri (Sklu), and the outgroup fungus Candida albicans (Calb). Running time for NANUQ steps 1-3 was about .5 s.*

Figure 15 shows the results of hypothesis tests for several choice of α and β. As all of the QCCFs are far from the central point of the simplex, only a quite large β would lead to failing to reject the star tree for any set of 4 taxa. Thus for this data set, we simply set $\beta = 0.1$ and call quartet networks as either resolved trees or 4-cycle networks. (We also see no empirical CFs plotted where those from 3_2-cycles that were not tree-like would

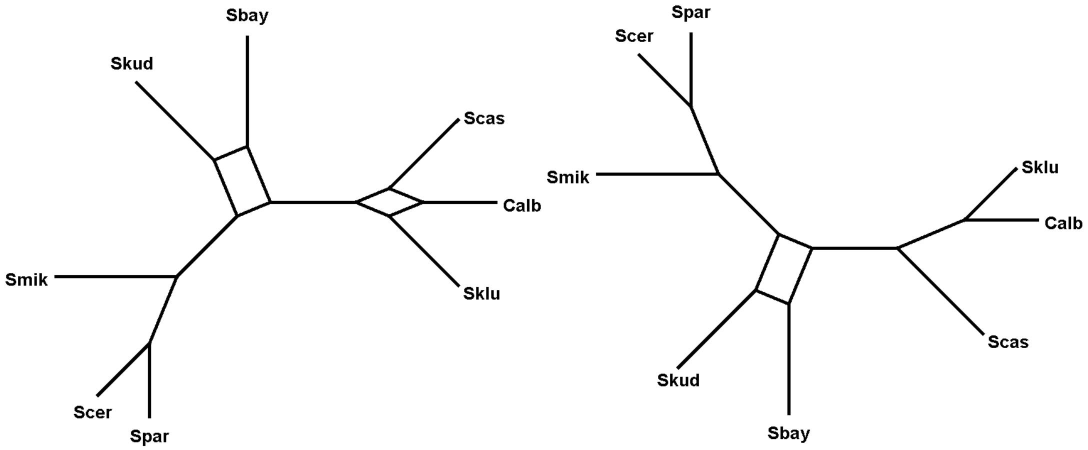


FIGURE 16. Split networks inferred for yeast dataset of Example 37, with $\alpha = 10^{-2}$ (L) or 10^{-4} (R) and $\beta = 10^{-1}$.

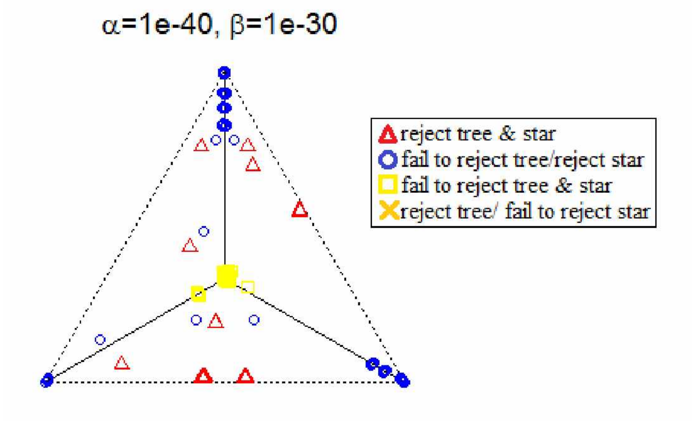


FIGURE 17. Simplex plots showing hypothesis test results for Heliconius data set of Example 38.

lie, giving us some confidence in NANUQ's assumption that there are none.) We choose values of $\alpha = 10^{-4}$ and 10^{-2} as the first of these results in only the most extreme (far from the tree-like lines) empirical CFs being interpreted as supporting 4-cycle networks, while the second brings in the remaining ones that appear far from the lines. Further increasing α to $> .08$ would result in additional calls of 4-cycle networks, but we choose to interpret their deviations from the tree-like lines as being stochastic (or other) noise.

For each of the choices of α, β, the split graphs produced in NANUQ's use of SplitsTree are shown in Figure 16. Since these show only 4-cycles, they can be directly interpreted as indicating the undirected version of the true level-1 network topology relating the taxa, with all 2- and 3-cycles contracted. We obtain no information on root location since no cycles have size larger than 4.

Example 38. *For the third example we use a butterflies data set* [16], *which was also analyzed in* [8]. *This data set consists of 2909 loci, each with alleles sampled from seven Heliconius*

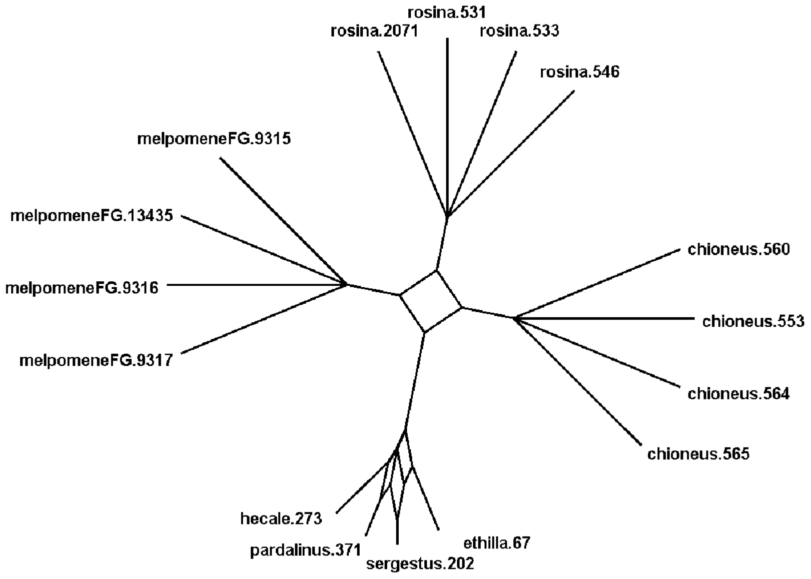


FIGURE 18. Split graph for Heliconius dataset of Example 38, with $\alpha = 10^{-40}$, $\beta = 10^{-6}$

species: H. rosina, H. melpomene, H. cydno (labelled chioneus), H. ethilia, H. hecale, H. p. sergestus, and H. pardalinus. For three of these taxa, rosina, melpomene, and chioneus, four individuals were sampled, with one sampled for the other four taxa. Running time for steps 1-3 of the algorithm was about 218 s. Figure 17 shows results of hypothesis tests for one choice of α and β, with Figure 18 the resulting split graph.

While difficult to see in the SplitsTree4 output, the split graph depicts a split with very small weight separating ethilia, sergestus, and pardalinus from the rest of the taxa. However, that software allows such small weight splits to be filtered out, and doing so leaves a 5-dart. Since the 5-dart is pointed at sergestus, the network structure that is inferred is as shown in 19. Note that with empirical CFs plotted so close to the central point in the simplex of Figure 17, the choice of β led to some of these being treated as unresolved quartets, giving the multifurcations in the split graph for the multisampled taxa.

For this data set the 4-cycle separating the groups rosina, melpomene, chioneus and all the other taxa remains stable under a wide range of choices of α and β. However, cycles appear in the single taxon groups if β is made larger, so that star trees are rejected more often. Varying α to 10^{-17} or larger, which lowers the standard for inference of hybridization, changes the 5-dart to a different 5-dart pointed at ethilla. Thus while the central 4-cycle is well supported, one should probably not draw firm conclusions on other hybridizations from this data set.

REFERENCES

[1] E.S. Allman, J.H. Degnan, and J.A. Rhodes. Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. *Journal of Mathematical Biology*, 62(6):833–862, 2011.

[2] E.S. Allman, J.D. Mitchell, and J.A. Rhodes. Hypothesis testing near singularities and boundaries. *https://arxiv.org/abs/1806.08458*, 2018.

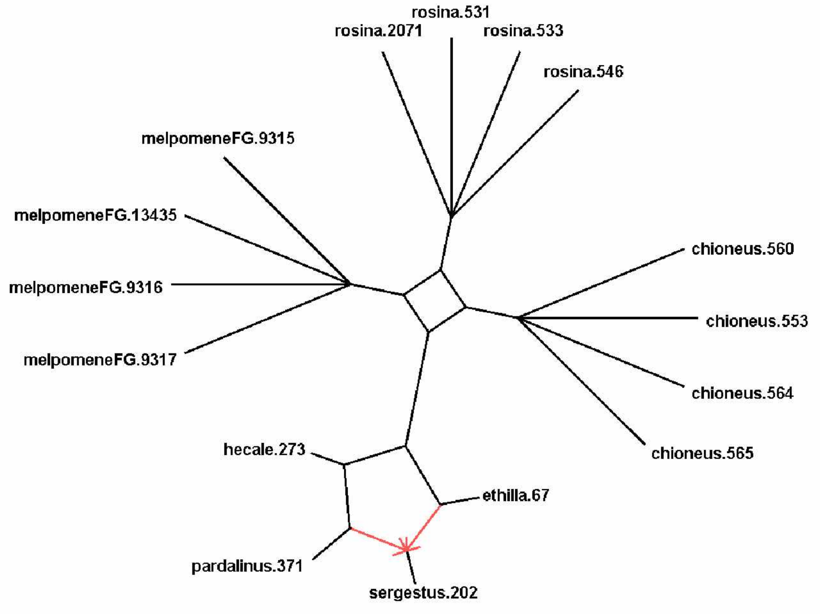


FIGURE 19. Inferred network structure from Heliconius dataset of Example 38, from the split graph of Figure 18.

[3] H. Baños. Identifying species network features from gene tree quartets. *Bulletin of Mathematical Biology*, 81:494–534, 2019.

[4] H. Bandelt and A. Dress. A canonical decomposition theory for metrics on a finite set. *Advances in Mathematics*, 92:47–105, 1992.

[5] D. Bloomquist and M. Suchard. Unifying vertical and nonvertical evolution: A stochastic arg-based framework. *Systematic Biology*, 59:27–41, 2010.

[6] D. Bryant and V. Moulton. Neighbor-net: An agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution*, 21:255–265, 2004.

[7] D. Bryant, V. Moulton, and A. Spillner. Consistency of the neighbor-net algorithm. *Algorithms for Molecular Biology*, 2:8, 2007.

[8] J. Chifman and L. Kubatko. Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes, site-specific rate variation, and invariable sites. *Journal of Theoretical Biology*, 374:35–47, 2015.

[9] A.W.M. Dress and D.H. Huson. Constructing splits graphs. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(3):109–115, July 2004.

[10] P. Gambette, V. Berry, and C. Paul. Quartets and unrooted phylogenetic networks. *Journal of Bioinformatics and Computational Biology*, 10(4):1250004, 2012.

[11] B.R. Holland, K.T. Huber, V. Moulton, and P.J. Lockhart. Using consensus networks to visualize contradictory evidence for species phylogeny. *Molecular Biology and Evolution*, 21(7):1459–1461, 2004.

[12] D.H. Huson and D. Bryant. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, 23(2):254–267, 2005.

[13] D.H. Huson, T. Klöpper, P.J. Lockhart, and M.A. Steel. Reconstruction of reticulate networks from gene trees. In S. Miyano, J. Mesirov, S. Kasif, S. Istrail, P.A. Pevzner, and M. Waterman, editors, *Research in Computational Molecular Biology. RECOMB 2005.*, volume 3500 of *Lecture Notes in Computer Science*, Berlin, Heidelberg, 2005. Springer.

[14] D.H. Huson, R. Rupp, and C. Scornavacca. *Phylogenetic Networks*. Cambridge University Press, Cambridge, 2010.

[15] L. Liu, L. Yu, L. Kubatko, D.K. Pearl, and S.V. Edwards. Coalescent methods for estimating phylogenetic trees. *Mol. Phylogenet. Evol.*, 53(1):320–328, 2009.

[16] S.H. Martin, Dasmahapatra K.K., N.J. Nadeau, C. Salazar, J.R. Walters, F. Simpson, M. Blaxter, A. Manica, J. Mallet, and C.D. Jiggins. Genome-wide evidence for speciation with gene flow in heliconius butterflies. *Genome Res*, 23:1817–1828, 2013.

[17] C. Meng and L.S. Kubatko. Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: A model. *Theoretical Population Biology*, 75(1):35–45, 2009.

[18] P. Pamilo and M. Nei. Relationships between gene trees and species trees. *Mol. Biol. Evol.*, 5(5):568–583, 1988.

[19] J.A. Rhodes. Topological metrizations of trees, and new quartet methods of tree inference. *https://arxiv.org/abs/1704.02004*, 2017.

[20] S. Roch and K.-C. Wang. Circular networks from distorted metrics. In B. Raphael, editor, *Research in Computational Molecular Biology, RECOMB 2018*, volume 10812 of *Lecture Notes in Computer Science*. Springer, 2018.

[21] A. Rokas, B. Williams, and S. Carrol. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, 425:798–804, 10 2003.

[22] Francesco Rosselló and Gabriel Valiente. All that glisters is not galled. *Mathematical Biosciences*, 221(1):54–59, 2009.

[23] C. Solís-Lemus and C. Ané. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS Genetics*, 12(3), 2016.

[24] M. Steel. *Phylogeny: Discrete and Random Processes in Evolution*. SIAM, Philadelphia, 2016.

[25] D. Wen and L. Nakhleh. Coestimating reticulate phylogenies and gene trees from multilocus sequence data. *Systematic Biology*, 67(3):439–457, 2018.

[26] Q. Wu, S. James, I. Roberts, V. Moulton, and K. Huber. Exploring contradictory phylogenetic relationships in yeasts. *FEMS Yeast Research*, 8:641–650, 2008.

[27] Y. Yu, J.H. Degnan, and L. Nakhleh. The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. *PLoS Genetics*, 8:e1002660, 2012.

[28] Y. Yu and L. Nakhleh. A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genomics*, 16:S10, 2015.

[29] Y. Yu, C. Than, J.H. Degnan, and L. Nakhleh. Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. *Systematic Biology*, 60(2):138–149, 2011.

[30] C. Zhang, H.A. Ogilvie, A.J. Drummond, and T. Stadler. Bayesian inference of species networks from multilocus sequence data. *Mol. Biol. Evol.*, 35(2):504–517, 2018.

[31] J. Zhu, Y. Yu, and L. Nakhleh. In the light of deep coalescence: Revisiting trees within networks. *BMC Bioinformatics*, 17:415, 2016.

[32] S. Zhu, J.H. Degnan, S. Goldstien, and B. Eldon. Hybrid-Lambda: Simulation of multiple merger and Kingman gene genealogies in species networks and species trees. *BMC Bioinformatics*, 16(1):292, Sep 2015.

Chapter 5: Conclusions and future work

In this work we have shown that, under the NSMC, one can recover from expected gene tree frequencies most of the topology of a level-1 species network. Also, we have presented an algorithm for the inference of level-1 networks under this model.

The main result in Chapter 3 considers only level-1 networks, and though it is a common assumption in the literature [*Rosselló and Valiente*, 2009; *Huson et al.*, 2010; *Huber et al.*, 2017; *Solís-Lemus and Ané*, 2016; *Steel*, 2016], we have to move beyond this eventually. Also, this result focuses only in identifying the topology of the species network and left unaddressed that question for inferring metric parameters of the species network. These are very interesting questions to pursue. One could try to approach these questions by looking at several quartet networks at the same time, or maybe move beyond them and consider networks on 5 taxa (quintet networks) instead. We leave this for future work.

Also, these identifiability results use as data gene trees, but the question remains open as to whether the species network topology is identifiable from whole genome DNA sequences. We have been working on a generalization of an species tree inference method developed in [*Allman et al.*, 2018b], that is based under a well known DNA sequence metric, the *log-det* distance [*Steel*, 2016].

As discussed in Chapter 4, there exist 3_2-cycles that pose challenges to the NANUQ algorithm. For future work, one could also investigate more on this to find a way around it.

References

Allman, E., and J. Rhodes (2005), Lecture notes: The mathematics of phylogenetics.

Allman, E., J. Mitchell, and J. Rhodes (2018a), Hypothesis testing near singularities and boundaries, *https://arxiv.org/abs/1806.08458.*

Allman, E., C. Long, and J. Rhodes (2018b), Identifiability of species tree topologies from genomic sequences using the logdet distance, *https://arxiv.org/abs/1806.04974.*

Allman, E., H. Baños, and J. Rhodes (2019), NANUQ: A method for inferring species networks from gene trees under the coalescent model, *In preparation.*

Baños, H. (2019), Identifying species network features from gene tree quartets, *Bulletin of Mathematical Biology*, *81*, 494534.

Bryant, D., and V. Moulton (2004), Neighbor-net: An agglomerative method for the construction of phylogenetic networks, *Molecular Biology and Evolution*, *21*, 255–265.

Carstens, B. C., L. L. Knowles, and T. Collins (2007), Estimating Species Phylogeny from Gene-Tree Probabilities Despite Incomplete Lineage Sorting: An Example from Melanoplus Grasshoppers, *Systematic Biology*, *56*(3), 400–411, doi:10.1080/10635150701405560.

Chifman, J., and L. Kubatko (2015), Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes, site-specific rate variation, and invariable sites, *Journal of Theoretical Biology*, *374*, 35–47.

Degnan, J. H. (2010), Probabilities of gene tree topoligies with intraspecific sampling given a species tree, in *Estimating Species Trees: Practical and Theoretical Aspects*, edited by L. L. Knowles and L. S. Kubatko, chap. 4, pp. 53–78, Wiley-Blackwell, Hoboken, NJ.

Degnan, J. H., and L. Salter (2005), Gene tree distributions under the coalescent process., *Evolution; International Journal of Organic Evolution*, *59*(1), 24–37.

Dress, A. W. M., and D. H. Huson (2004), Constructing splits graphs, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *1*(3), 109–115, doi:10.1109/TCBB.2004.27.

Ellstrand, N. C., R. Whitkus, and L. H. Rieseberg (1996), Distribution of spontaneous plant hybrids, *Proceedings of the National Academy of Sciences of the United States of America*, *93*(10), 5090–5093, doi:10.1073/pnas.93.10.5090.

Huber, K. T., L. van Iersel, V. Moulton, C. Scornavacca, and T. Wu (2017), Reconstructing phylogenetic level-1 networks from nondense binet and trinet sets, *Algorithmica*, *77*(1), 173–200, doi:10.1007/s00453-015-0069-8.

Huson, D. H., and D. Bryant (2006), Application of phylogenetic networks in evolutionary studies.

Huson, D. H., R. Rupp, and C. Scornavacca (2010), *Phylogenetic Networks*, Cambridge University Press, Cambridge.

Kingman, J. F. C. (1988), On the genealogy of large populations, *Journal of Applied Probability*, *19*, 2743.

Meng, C., and L. S. Kubatko (2009), Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: A model, *Theoretical Population Biology*, *75*(1), 35–45.

Nakhleh, L. (2011), Evolutionary phylogenetic networks: Models and issues, in *Problem Solving Handbook in Computational Biology and Bioinformatics*, edited by L. S. Heath and N. Ramakrishnan, chap. 7, pp. 125–158, Springer US, Boston, MA.

Olson, M. V. (1999), When less is more: gene loss as an engine of evolutionary change, *American Journal of Human Genetics*, *64*, 18–23.

Pamilo, P., and M. Nei (1988), Relationships between gene trees and species trees., *Molecular Biology and Evolution*, *5*(5), 568–583, doi:10.1093/oxfordjournals.molbev.a040517.

Pollard, D. A., V. N. Iyer, A. M. Moses, and M. B. Eisen (2006), Widespread discordance of gene trees with species tree in *Drosophila*: Evidence for incomplete lineage sorting, *PLoS Genetics*, *2*(10), 1634–1647, doi:10.1371/journal.pgen.0020173.

R Core Team (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Rhodes, J. A. (2017), Topological metrizations of trees, and new quartet methods of tree inference, *https://arxiv.org/abs/1704.02004*.

Rieseberg, L. H., S. J. Baird, and K. A. Gardner (2000), Hybridization, introgression, and linkage evolution, *Plant Molecular Biology*, *42*(1), 205–224, doi:10.1023/A:1006340407546.

Rosselló, F., and G. Valiente (2009), All that glisters is not galled, *Mathematical Biosciences*, *221*(1), 54–59, doi:10.1016/j.mbs.2009.06.007.

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

Solís-Lemus, C., and C. Ané (2016), Inferring Phylogenetic Networks with Maximum Pseudolikelihood under Incomplete Lineage Sorting, *PLoS Genetics*, *12*(3), doi:10.1371/journal.pgen.1005896.

Steel, M. (2016), *Phylogeny: Discrete and Random Processes in Evolution*, SIAM, Philadelphia.

Syring, J., A. Willyard, R. Cronn, and A. Liston (2005), Evolutionary relationships among Pinus (Pinaceae) subsections inferred from multiple low-copy nuclear loci, *American Journal of Botany*, *92*(12), 2086–2100, doi:10.3732/ajb.92.12.2086.

Wakeley, J. (2008), *Coalescent Theory: An Introduction*, Roberts and Company Publishers.

Zhu, S., J. H. Degnan, S. J. Goldstien, and B. Eldon (2015), Hybrid-Lambda: Simulation of multiple merger and Kingman gene genealogies in species networks and species trees, *BMC Bioinformatics*, *16*(1), 292, doi:10.1186/s12859-015-0721-y.