HHS Public Access

Author manuscript

J Math Biol. Author manuscript; available in PMC 2022 June 13.

Published in final edited form as:

J Math Biol. ; 84(5): 35. doi:10.1007/s00285-022-01734-2.

Identifiability of species network topologies from genomic sequences using the logDet distance

Elizabeth S. Allman,

Department of Mathematics and Statistics, University of Alaska Fairbanks, Fairbanks, AK, 99775, USA

Hector Baños,

Department of Biochemistry & Molecular Biology, Faculty of Medicine, Dalhousie University, Halifax, Nova Scotia, CANADA; Department of Mathematics & Statistics, Faculty of Science, Dalhousie University, Halifax, Nova Scotia, CANADA

John A. Rhodes

Department of Mathematics and Statistics, University of Alaska Fairbanks, Fairbanks, AK, 99775, USA

Abstract

Inference of network-like evolutionary relationships between species from genomic data must address the interwoven signals from both gene flow and incomplete lineage sorting. The heavy computational demands of standard approaches to this problem severely limit the size of datasets that may be analyzed, in both the number of species and the number of genetic loci. Here we provide a theoretical pointer to more efficient methods, by showing that logDet distances computed from genomic-scale sequences retain sufficient information to recover network relationships in the level-1 ultrametric case. This result is obtained under the Network Multispecies Coalescent model combined with a mixture of General Time-Reversible sequence evolution models across individual gene trees. It applies to both unlinked site data, such as for SNPs, and to sequence data in which many contiguous sites may have evolved on a common tree, such as concatenated gene sequences. Thus under standard stochastic models statistically justifiable inference of network relationships from sequences can be accomplished without consideration of individual genes or gene trees.

Keywords

species network; identifiability; logDet; phylogenetic inference

1 Introduction

As genomic-scale sequencing has become increasingly common, attention in phylogenetics has shifted from inferring trees of evolutionary relationships for individual genetic loci from a set of species to inferring relationships between the species themselves. A substantial

e.allman@alaska.edu .

complication is that population genetic processes within species, as modeled by the *Multispecies Coalescent* (MSC) model can lead to individual gene trees having quite different topological structures than the tree relating the species overall. If the evolutionary history of the species also involved hybridization or other forms of horizontal gene flow, so that a species network is a more suitable depiction of relationships, the relationships of gene trees to the network, as modeled by the *Network Multispecies Coalescent* (NMSC) model, is even more complex.

Inference of species networks, through a combined NMSC and sequence substitution model, can be performed in a Bayesian framework [Zhang et al., 2017, Wen and Nakhleh, 2018] but computational demands severely limit both the number of taxa and the number of genetic loci considered. Other methods take a faster two-stage approach, first inferring gene trees which are treated as "data" for a second inference of a species network. Approaches include maximum pseudolikelihood using either rooted triples (PhyloNet) or quartets (SNaQ) displayed on the gene trees [Yu and Nakhleh, 2015, Solís-Lemus and Ané, 2016], or the faster, distance-based analysis built on gene quartets of NANUQ [Allman et al., 2019a]. Still, the first stage of these approaches, the inference of individual gene trees, can be a major computational burden. Avoiding such gene tree inference, and passing more directly from sequences to an inferred network, could substantially reduce total computational time in data analysis pipelines.

The goal of this paper is to show that most topological features of a level-1 species network can be identified from logDet intertaxon distances computed from aligned genomic-scale sequences. In particular this can be done without partitioning the sequences by genes, under a combined model of the NMSC and a mixture of general time-reversible (GTR) substitution processes on gene trees. While the main result, that the logDet distances retain enough information to recover most of the species network, despite having lost information on individual genes, is a theoretical one, it points the way toward faster algorithms for practical inference. In particular, since the computation of logDet distances requires little effort, it suggests that a distance-based approach similar to NANUQ's, but avoiding individual gene tree inference, may offer substantially faster analyses than current methods.

The model of sequence evolution underlying our result accounts not only for base substitutions along each gene tree, but also for variation in gene trees due to their formation under a coalescent process combined with hybridization or similar gene transfer. Our model extends to networks the mixture of coalescent mixtures model on species trees of Allman et al. [2019b], which itself extended the coalescent mixture introduced by Chifman and Kubatko [2015]. More specifically, for a fixed species network, gene trees are formed under the Network Multispecies Coalescent model [Meng and Kubatko, 2009, Yu et al., 2011, Zhu et al., 2016] for each site independently. GTR substitution parameters for base evolution on each site's tree are then independently chosen from some distribution, leading to a site pattern distribution. These site distributions are finally combined to give a site pattern distribution for genomic sequences. (As discussed in Section 2, this distribution also applies to a more realistic model in which multisite genes with a single substitution process have lengths chosen independently from some distribution.) While this pattern frequency

distribution thus reflects the substitution processes on all the gene trees, information about pattern frequencies arising on any individual gene tree is hidden.

The logDet distance was first introduced in the context of a single class general Markov model of sequence evolution on a single gene tree [Steel, 1994, Lockhart et al., 1994], and has been used both to obtain both gene tree identifiability results and for inference of individual gene trees. Considering genomic sequences, Liu and Edwards [2009], and independently Dasarathy et al. [2015], showed that for a Jukes-Cantor substitution model and an ultrametric species tree, the Jukes-Cantor distances obtained under the coalescent mixture model still allowed for consistent inference of topological species trees. By passing to the logDet distance, Allman et al. [2019b] extended this result to the more realistic mixture of coalescent mixtures model, showing that the logDet distance allowed for consistent inference of a topological species tree, assuming it is ultrametric in generations. This study builds on all these works on gene and species tree models, but considers level-1 species networks on which all extant species are equidistant from the root.

Passing from species trees to networks is a substantial step, however, and our approach is strongly motivated by the approach taken by Baños [2019] in studying identifiability of features of unrooted level-1 topological species networks from gene tree quartet concordance factors (probabilities of the different quartet topologies displayed on gene trees). In the ultrametric setting of this work, we show that logDet distances computed from genomic sequences suffice to determine 4-cycles on undirected rooted triple networks, and then that this 4-cycle information for different rooted triples can be combined to determine all cycles of size 4 or more, and even all hybrid nodes in those cycles of size 5 or more. We do not obtain information on 2- or 3- cycles, so our results closely parallel those of Baños [2019], despite the rather different source of information.

There are a number of other theoretical works in the literature on determining phylogenetic networks from limited information. For instance, Jansson and Sung [2006] investigate determining a level-1 network from the rooted triple trees it displays, Huber et al. [2017, 2018] discuss how knowledge of trinets (induced 3-taxon directed rooted networks) and quarnets (induced 4-taxon undirected unrooted networks) determine larger networks, and van Iersel et al. [2020] explore determination of networks from distances. However, the question of how, or whether, these results can be applied to biological data is not addressed, and the setting of these works is not directly applicable to obtaining our results.

Other works [Gross and Long, 2018, Gross et al., 2020, Hollering and Sullivant, 2021] use algebraic approaches to show that certain types of level-1 networks can be identified from joint pattern frequency arrays under group-based models of sequence evolution such as the Jukes-Cantor and Kimura models. In addition to their restriction on sequence evolution models, these works do not incorporate a coalescent process. That is, all sequence sites are assumed to have evolved on one of the finitely-many trees displayed on the network. Since the absence of a coalescent process is a limiting case of our coalescent-based model, our results allowing for mixtures of more general sequence evolution models extend those results in the ultrametric case. Algebraic study of a network model combined with the

general Markov model, again with no coalescent process, was also conducted by Casanellas and Fernández-Sánchez [2020].

This paper proceeds as follows. Section 2 defines the networks and models under consideration, as well as the logDet distance. For most of the paper we restrict to a model of unlinked sites, only later passing to a model allowing concatenated genes whose sites evolve on the same gene tree. Section 3 uses combinatorial arguments to show how information on undirected rooted triple networks can be used to determine features of a larger directed network from which they are induced. Expected frequencies of site patterns for sequences produced by the mixture of coalescent mixtures model are studied in Section 4, and shown to be expressible as convex combinations of pattern frequencies from simpler networks. In Section 5 we show that the ordering by magnitude of logDet distances for triples of taxa tells us about the induced rooted triple species network, and by combining this with the result of Section 3 we obtain our main identifiability result, Theorem 1. Section 6 discusses two variations on our main result that are implied by it. The first is to a model with genes of linked sites that evolve on a common tree. The second is to a non-coalescent model, in which all gene trees must be displayed on the species network. Section 7 further studies the logDet distances from a rooted triple network, in order to better understand what triples of distances can arise under the mixture of coalescent mixtures model. We conclude in Section 8 with an outline of how these results can be developed into a practical inference algorithm.

2 Networks and models

2.1 Phylogenetic Networks

Although there are many variations on the notion of a phylogenetic network in the literature, we adopt ones appropriate to the Network Multispecies Coalescent (NMSC) model. This model, which describes the formation of trees of gene lineages in the presence of both incomplete lineage sorting and hybridization, will be further developed in the next subsection. First, we focus on setting forth combinatorial aspects of the networks.

Definition 1 [Solís-Lemus and Ané, 2016, Baños, 2019] A *topological binary rooted phylogenetic network* $\mathcal{N}^+$ on taxon set X is a connected directed acyclic graph with vertices $V = V(\mathcal{N}^+)$ and edges $E = E(\mathcal{N}^+)$, where V is a disjoint union $V = \{r\} \sqcup V_L \sqcup V_H \sqcup V_T$ and E is a disjoint union $E = E_H \sqcup E_T$, 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* $e = (v, w) \in E_H$ is an edge whose child node w is hybrid.
6. A *tree edge* $e = (v, w) \in E_T$ is an edge whose child node w is either a tree node or a leaf.

When $|X| = 3$ or 4, we refer to $\mathcal{N}^+$ as a *rooted triple network* or a *rooted quartet network*, respectively.

The vertices, and edges, of $\mathcal{N}^+$ are partially ordered by the directedness of the graph. For instance, a node u is *below* a node v, and v is *above* u, if there exists a non-empty directed path in $\mathcal{N}^+$ from v to u. The root is thus above all other nodes.

A metric notion of the network above incorporates some of the parameters of the NMSC model. This introduces edge lengths, measured in generations throughout this article, as well as probabilities that a gene lineage at a hybrid node follows one or the other hybrid edge as it traces back in time toward the network root. Since we focus on binary networks, only hybrid edges are allowed to have length 0, to model possibly instantaneous jumping of a lineage from one population to another.

Definition 2 A *metric binary rooted phylogenetic network* $(\mathcal{N}^+, \{\ell_e\}_{e \in E}, \{\gamma_e\}_{e \in E_H})$ is a topological binary rooted phylogenetic network together with an assignment of weights or *lengths* ℓ_e to all edges and *hybridization parameters* γ_e to all hybrid edges, subject to the following restrictions:

1. The length ℓ_e of a tree edge $e \in E_T$ is positive.
2. The length ℓ_e of a hybrid edge $e \in E_H$ is non-negative.
3. The hybridization parameters γ_e and $\gamma_{e'}$ for a pair of hybrid edges $e, e' \in E_H$ with the same child hybrid node are positive and sum to 1.

A metric network of this sort is said to be *ultrametric* if every directed path from the root to a leaf has the same total length. This is equivalent to requiring the ultrametricity of all trees displayed on the network. An example of a simple ultrametric network is shown in Figure 1 (Right).

On directed networks there are several analogs [Steel, 2016] of the most recent common ancestor of a set of taxa on a tree. The following is the most useful in this work.

Definition 3 [Steel, 2016] Let $\mathcal{N}^+$ be a (metric or topological) binary rooted phylogenetic network on a set of taxa X and let $Z \subseteq X$. Let D be the set of nodes which lie on every directed path from the root r of $\mathcal{N}^+$ to any $z \in Z$. Then the *lowest stable ancestor of Z on $\mathcal{N}^+$*, denoted $\mathrm{LSA}(Z, \mathcal{N}^+)$, is the unique node $v \in D$ such that v is below all $u \in D$ with $u \neq v$. The *lowest stable ancestor (LSA)* of a network on X is $\mathrm{LSA}(X)$.

Phylogenetic networks as defined here have no cycles in the usual sense for a directed graph. The term *cycle* will thus be used to refer to a collection of edges that form a cycle when all edges are undirected. A cycle must contain at least two hybrid edges sharing a hybrid node, and may contain any non-negative number of tree edges. The class of networks we focus on is those in which cycles are separated, in the following sense.

Definition 4 A rooted binary phylogenetic network $\mathcal{N}^+$ is said to be *level-1* if no two distinct cycles in $\mathcal{N}^+$ share an edge.

Although this is not the standard definition of level-1 [Rosselló and Valiente, 2009], in the setting of binary networks it is equivalent.

Each cycle on a level-1 phylogenetic network contains exactly one hybrid node and two hybrid edges with that node as a child. Thus there is a one-to-one correspondence between cycles and the hybrid nodes they contain. A cycle composed of n edges, 2 of which are hybrid, is called an *n-cycle*. If the cycle's hybrid node has k leaf descendants, it is an n_k*-cycle*.

Passing from a large network to one on a subset of the taxa is similar to the process for trees.

Definition 5 *Suppressing a node* with both in- and out-degree 1 in a directed phylogenetic network means replacing it and its two incident edges with a single edge from its parent to its child. For a metric network, the new edge is assigned a length equal to the sum of lengths of the two replaced. If the outedge was hybrid, the new edge is also hybrid and retains the hybridization parameter.

Similarly, suppressing a node of degree 2 between two undirected edges means replacing it and its two incident edges with a single undirected edge.

Definition 6 Let $\mathcal{N}^+$ be a (metric or topological) binary rooted phylogenetic network on X and let $Y \subset X$. The *induced rooted network* $\mathcal{N}^+_Y$ on Y is the network obtained from $\mathcal{N}^+$ by retaining nodes and edges in every path from the root r on $\mathcal{N}^+$ to any $y \in Y$, and then suppressing all nodes with in- and out-degree 1. We then say $\mathcal{N}^+$ *displays* $\mathcal{N}^+_Y$.

We also need the notion of a *rooted undirected network*, in which all edges have been undirected but the root retained. Note that if a rooted network is a tree, knowledge of the root alone is enough to recover the direction of every edge, so this notion is not useful in that setting. If cycles are present, knowledge of the root determines only the direction of every cut edge (an edge whose deletion results in a graph with two connected components), and edges directly descended from cut edges. Knowing the root and all hybrid nodes in an undirected level-1 network does, however, determine the full directed network.

Several other notions of networks induced from a directed one are needed.

Definition 7 Let $\mathcal{N}^+$ be a (metric or topological) binary rooted phylogenetic network on X.

1. [Baños, 2019] The *LSA network* $\mathcal{N}^{\oplus}$ induced from $\mathcal{N}^+$ is the network on X obtained by deleting all edges and nodes above $\mathrm{LSA}(X, \mathcal{N}^+)$, and designating $\mathrm{LSA}(X, \mathcal{N}^+)$ as the root node.
2. The *undirected LSA network* $\mathcal{N}^{\ominus}$ is the rooted network obtained from the LSA network $\mathcal{N}^{\oplus}$ by undirecting all edges.

3. [Baños, 2019] The *unrooted semidirected network* $\mathcal{N}^{-}$ is the unrooted network obtained from the LSA network $\mathcal{N}^{\oplus}$ by undirecting all tree edges and suppressing the root, but retaining directions of hybrid edges.

For a binary level-1 network $\mathcal{N}^{+}$, the only possible structure above the LSA has the form of a (possibly empty) chain of 2-cycles [Baños, 2019], an example of which is shown in Figure 2. The LSA network $\mathcal{N}^{\oplus}$ is obtained by simply deleting that chain.

Note that the terminology of "n_k-cycles" can be applied to LSA networks $\mathcal{N}^{\oplus}$, as hybrid edges retain their direction. On undirected LSA networks $\mathcal{N}^{\ominus}$, however, "n-cycle" can still be applied, but "n_k-cycle" generally cannot.

Definition 8 By *suppressing a cycle* C in a topological level-1 network we mean deleting all edges in C, identifying all nodes in C, and if the resulting node is of degree 2 suppressing it. If the network is rooted and this results in the root becoming a degree 1-node, then the resulting edge below the root is also deleted, with its child becoming the root.

Suppressing an n-cycle in a binary level-1 network results in a non-binary network when $n \geq 4$. However if only 2- and 3-cycles are suppressed, the result is binary.

2.2 Coalescent Model on Networks

The formation of gene trees within a species network, as ancestral lineages of sampled loci from extant taxa join together moving backwards in time, is given a mechanistic description by the Network Multispecies Coalescent Model (NMSC) [Meng and Kubatko, 2009, Yu et al., 2011, Zhu et al., 2016].

Parameters of the NMSC for a set of taxa X include a metric rooted binary phylogenetic network $(\mathcal{N}^{+}, \{\ell_e\}, \{\gamma_e\})$ on X, with edge lengths ℓ_e in generations. In addition, for each edge $e = (u, v)$ fix a function $N_e : [0, \ell_e) \to \mathbb{R}^{>0}$ giving the (haploid) population size along the edge, where $N_e(0)$ is the population size at the child node v and $N_e(t)$ is the population at time t units above it. Finally, let $N_r : [0, \infty) \to \mathbb{R}^{>0}$ be an additional population size function for an infinite length 'edge' ancestral to the root r of the network. The N_e need not be constant nor equal, although those are common assumptions in other works. As did Allman et al. [2019b], we make the biologically-plausible technical assumptions that the functions N_e are bounded, and that all $1/N_e(t)$ are integrable over finite intervals.

Figure 1 (Left) depicts an example species network that is ultrametric in generations, with hybrid edges h and h', and population functions N_e on each edge depicted by time-varying widths of the network edges. The edge lengths ℓ_e are measured on the t-axis between the horizontal lines indicating speciation and hybridization events. Figure 1 (Right) gives a schematic of the same species tree, without a depiction of population functions.

The standard Kingman coalescent models the formation of gene trees, with edge lengths in generations, within a single population edge e, with pairs of lineages coalescing independently as they trace backward in time, at instantaneous rate $1/N_e(t)$. The multispecies

coalescent model (MSC) extends this to a tree of populations, by using the standard coalescent on each edge, as well as an infinite length edge above the root, allowing multiple gene lineages to enter a population from its descendant ones at a tree node. The NMSC extends this further, so that lineages reaching hybrid nodes randomly enter one or the other hybrid edge above them, with the choice determined independently according to the hybridization parameter probabilities. Thus the NMSC parameters $(\mathcal{N}^+, \{\ell_e\}, \{\gamma_e\})$ and $\{N_e\}$ determine a distribution of rooted metric gene trees. The structure of the NMSC also ensures that the distributions of gene trees obtained by marginalization to a subset Y of taxa are the same as the distributions obtained from the NMSC on the displayed network $\mathcal{N}_Y^+$.

2.3 Sequence substitution models on gene trees

The k-state *general time-reversible model* (GTR) for sequence evolution is a continuous-time Markov process on a metric gene tree. Gene tree edge lengths are in substitution units, and sequences are composed of k possible states, or bases. Model parameters are a $k \times k$ instantaneous rate matrix Q together with a k-state distribution $\boldsymbol{\pi}$, with non-negative entries summing to 1, satisfying the following:

1. off-diagonals entries of Q are positive,
2. row sums of Q are 0,
3. trace $Q = -1$,
4. $\boldsymbol{\pi}Q = 0$,
5. $\mathrm{diag}(\boldsymbol{\pi})Q$ is symmetric.

In the ultrametric framework for our species networks, we introduce an additional time-dependent but lineage-independent rate scalar $\mu(t)$ for Q, where t is measured in generations from leaves to the root and beyond, and $\mu(t)$ has units of substitutions/generation. We assume μ is piecewise-continuous, $\mu(t) > 0$ for all $t \geq 0$ so that the mutations process never stops, and $\int_0^\infty \mu(t)dt = \infty$ so that the total amount of possible mutation is unbounded. Following Allman et al. [2019b], this substitution model is denoted by GTR+μ.

For any node u on a gene tree, let t_u denote the distance, in generations, to that node from its descendant leaves. The states at a single site in sequences at the taxa at the leaves on the gene tree are then determined as follows: A state is randomly chosen at the root of the tree from the distribution $\boldsymbol{\pi}$. For each edge $e = (u, v)$ descendant from a node u the site undergoes random state changes with rates $\mu(t)Q$ for times $t \in [t_v, t_u]$ to obtain states at the child nodes. The full substitution process on the edge is thus described by the Markov matrix

$$M_e = \exp\left(\int_{t_v}^{t_u} \mu(t)\, dt\, Q\right).$$

A similar process is then repeated for those nodes' children, and so on, until states at the taxa have been determined.

2.4 Mixture of coalescent mixtures

The model we focus on is the m-class *mixture of coalescent mixtures* [Allman et al., 2019b] extended from a tree to an ultrametric network. This model has as parameters an ultrametric species network $(\mathcal{N}^+, \{\ell_e\}, \{\gamma_e\})$, population size functions $\{N_e\}$, a finite collection $\{(Q_i, \pi_i; \mu_i)\}_{i=1}^m$ of GTR+μ parameters for the m classes, and a vector λ of m positive class size parameters summing to 1.

Sequence data is generated as follows: For each site:

1. a gene tree T is sampled according to the NMSC model on $(\mathcal{N}^+, \{\ell_e\}, \{\gamma_e\})$ with population sizes $\{N_e\}$,
2. class i is sampled from the distribution λ to determine parameters (Q_i, π_i, μ_i),
3. the bases for each $x \in X$ are sampled under the GTR+μ process on T with parameters (Q_i, π_i, μ_i).

This model is denoted by $\mathcal{M} = \mathcal{M}(\theta)$ where

$$\theta = ((\mathcal{N}^+, \{\ell_e\}, \{\gamma_e\}), \{N_e\}, \lambda, \{(Q_i, \pi_i; \mu_i)\}).$$

Sampling n independent sites from this model produces k-state aligned sequences of n unlinked sites. As usual in phylogenetics, these are summarized through counts of site patterns across the sequences in an $|X|$-dimensional $k \times k \times \cdots \times k$ array. Marginalizations of this array to 2-dimensions give pairwise $k \times k$ site pattern count matrices that compare only the sequences for two taxa in X.

In the tree context, two extensions of this model were discussed by Allman et al. [2019b]. For the first, the model assumption of one independently drawn gene tree for each site is modified to a more realistic one for genomic sequences in which all sites for a genetic locus share a gene tree. If the lengths (in number of sites) of the loci are independent identically distributed draws from some distribution, then the expected site pattern distribution for such a model is unchanged from that determined by $\mathcal{M}$. Only the rate of convergence, as the number of sampled genes grows, of frequencies of sampled site patterns to the asymptotic distribution will be slowed. This model is considered in Section 6, as its analysis follows easily from that for unlinked sites.

Another extension in the tree setting of Allman et al. [2019b] allowed for relaxing the ultrametric condition while retaining strong results on identifiability from the logDet distances. In that extension, the scalar rate function was allowed to be edge dependent as long as a certain symmetry condition on mixture components resulted in ultrametricity in substitution units “on average” across gene trees. While a similar model extension in the network setting seems likely to lead to similar results, it is not explored here, as the technical complications are greater than in the tree case.

2.5 LogDet distance

The fundamental tool we use to study relationships of taxa under the mixture of coalescent mixtures model $\mathscr{M}$ is the logDet distance between a pair of aligned sequences. It is computed as follows: For taxa $a, b \in X$, let $\hat{F}^{ab}$ be a $k \times k$ matrix of empirical relative site-pattern frequencies, obtained by normalizing the site pattern count matrix for a and b, so that its entries sum to 1. Thus the ij entry of $\hat{F}^{ab}$ is the proportion of sites in the sequences exhibiting base i for a and base j for b. With $\hat{f}_a$ and $\hat{f}_b$ the vectors of row and column sums of $\hat{F}^{ab}$, which give the proportions of various bases in the sequences for a and b, let $\hat{g}_a$ and $\hat{g}_b$ the products of the entries of $\hat{f}_a$, $\hat{f}_b$, respectively. Then the empirical logDet distance is

$$\hat{d}_{LD}(a,b) = -\frac{1}{k}\left(\ln\det\left(\hat{F}^{ab}\right) - \frac{1}{2}\ln(\hat{g}_a\hat{g}_b)\right) \tag{1}$$

Under most phylogenetic models, including the mixture of coalescent mixtures model, individual site patterns in sequences are assumed to be independent and identically distributed. By the weak law of large numbers, $\hat{F}^{ab}$ computed from a sample will converge in probability to its expected value F^{ab} as the sequence length goes to ∞. By the continuous function theorem, e.g. [van der Vaart, 1998], the empirical logDet distance thus converges in probability to the logDet distance computed by the same formula from the expected F^{ab}, a quantity we refer to as the *theoretical logDet distance* and denote by $d_{LD}(a, b)$.

3 Rooted Networks from Undirected Rooted Triple Networks

The goal of this section is to establish Proposition 1, a combinatorial result indicating features of a topological level-1 rooted n-taxon network that can be recovered from its induced undirected rooted triple networks with 2- and 3-cycles suppressed. This is a rooted analog of a key result of Baños [2019] relating unrooted semidirected networks and their induced undirected quartet networks. Later sections of this paper focus on identifying these rooted triple networks under the model $\mathscr{M}$.

There are several possible routes to Proposition 1. One approach would be to follow the argument of the quartet analog, with modifications throughout due to the rooted setting. Another would be to imitate the alternate proof of the quartet result given by Allman et al. [2019a], based on an extension of the intertaxon quartet distance of Rhodes [2019], but instead using the rooted triple distance also introduced in that work. The argument presented here is shorter than these approaches, as it leverages information about undirected rooted triple networks to obtain information about undirected quartet networks, and then applies the theory of Baños [2019].

The following result, extracted from the proof of Theorem 4 of Baños [2019], will be used. In it, and throughout this work, by a network *modulo 2- and 3-cycles* we mean the network obtained by suppressing all 2- and 3-cycles. Similarly, *modulo directions of edges in 4-cycles* means that all edges in 4-cycles are undirected. As a result, which of the edges in a 4-cycle are hybrid, and therefore which node is hybrid, is not indicated.

Lemma 1 ([Baños, 2019]) *Let $\mathcal{N}^+$ be a level-1 rooted binary topological phylogenetic network on X. Let Q be the set of undirected quartet networks obtained from those displayed on $\mathcal{N}^+$ by unrooting, suppressing all cycles of size 2 and 3, and undirecting all edges. Then modulo 2- and 3-cycles and directions of edges in 4-cycles, the semidirected unrooted network $\mathcal{N}^-$ is determined by Q.*

In order to apply this to rooted triples, we first recall some combinatorial properties of rooted triple and quartet networks.

Lemma 2 ([Baños, 2019]) *Let $\mathcal{Q}^-$ be a level-1 unrooted semidirected binary quartet network. Then $\mathcal{Q}^-$ has no k-cycles for $k \geq 5$, and at most one 4-cycle. If $\mathcal{Q}^-$ has a 4-cycle, then it has neither 3- nor 2_2-cycles. If there is no 4-cycle, then there are at most two 3-cycles, with at most one of these a 3_2-cycle.*

Lemma 2 can be used to characterize possible cycles in a rooted triple network, by attaching an outgroup at the root. More specifically, by *attaching an outgroup o to the root* of an *n*-taxon network on taxa X with $o \notin X$ we mean identifying the root r of the network with the node r on an edge (r, o) and undirecting all tree edges. This gives a $(n + 1)$-taxon unrooted semidirected network. The rooted triple networks displayed on the original network are then in one-to-one correspondence with induced semidirected quartet networks containing o on the new network. This construction yields the following.

Corollary 1 *Let $\mathcal{N}^+$ be a level-1 binary rooted triple network. Then $\mathcal{N}^+$ has no k-cycles for $k \geq 5$, and at most one 4-cycle in which case there are no 3- or 2_2-cycles. If there is no 4-cycle, then there are at most two 3-cycles, with at most one of these a 3_2-cycle.*

Considering a rooted quartet network $\mathcal{Q}^+$, and the impact of passing to its associated unrooted semidirected quartet network $\mathcal{Q}^-$, Lemma 2 also immediately yields the following.

Corollary 2 *Let $\mathcal{Q}^+$ be a level-1 rooted binary quartet network. Then $\mathcal{Q}^+$ has no k-cycles for $k \geq 6$, and has at most a one 5-cycle or 4-cycle, but not both.*

We now catalog the rooted quartet networks with 4- or 5-cycles, modulo smaller cycles.

Lemma 3 *Let $\mathcal{Q}^+$ be a level-1 binary rooted quartet network with one 4-cycle or one 5-cycle. Then modulo 2- and 3- cycles and up to taxon relabelling, the LSA network $\mathcal{Q}^{\oplus}$ is one of those shown in* Figure 3. *Thus $\mathcal{Q}^+$ displays either 1, 2, or 3 rooted triples with a 4-cycle.*

Proof Let $\mathcal{Q}^+$ be a rooted level-1 network on $\{a, b, c, d\}$ with a cycle C of size 4 or 5. By Corollary 2, C is the only cycle of size greater than 3. Figure 3 shows the topologies, up to taxon relabeling, of all the rooted quartet networks with a 4- or 5-cycle and no 2- or 3-cycles, as determined by enumerating all possible locations for adding hybrid edges to a rooted 4-taxon tree. The top row of Figure 3 shows the quartet networks with exactly one displayed rooted triple, on $\{a, b, c\}$, having a 4-cycle. The middle row shows the networks with exactly two displayed rooted triples, on $\{a, b, c\}$ and $\{a, b, d\}$, having a 4-cycle. The

bottom row shows those with exactly three displayed rooted triples, on $\{a, b, c\}$, $\{a, b, d\}$, and $\{a, c, d\}$, having a 4-cycle.

Now we proceed to the main result of this section.

Proposition 1 *Let $\mathcal{N}^+$ be a level-1 rooted binary topological phylogenetic network on X. Let S be the set of undirected rooted triple networks obtained from those displayed on $\mathcal{N}^+$ by suppressing all cycles of size 2 and 3 and undirecting all edges. Then modulo 2- and 3-cycles and directions of edges in 4-cycles, the LSA network $\mathcal{N}^{\oplus}$ is determined by S.*

Proof We first build a set of rooted quartet networks from S. Let $\{a, b, c, d\} \in X$ and let $S_{abcd} \subseteq S$ be the set of undirected rooted triple networks on any three elements of $\{a, b, c, d\}$, so $|S_{abcd}| = 4$. By Corollary 2 and Lemma 3, there are $k = 0$, 1, 2, or 3 elements of S_{abcd} with a 4-cycle. We consider each possibility in turn, showing that we can determine the undirected rooted quartet network $N^{\ominus}_{abcd}$ modulo 2- and 3-cycles.

If $k = 0$, all rooted triples in S_{abcd} are trees and since $\mathcal{N}^{+}_{abcd}$ has no 4- or 5-cycles by Lemma 3, the undirected LSA network $\mathcal{N}^{\ominus}_{abcd}$ modulo 2-and 3-cycles is a tree. By a well-known result for trees [Semple and Steel, 2005], S_{abcd} determines $\mathcal{N}^{\ominus}_{abcd}$ modulo 2- and 3-cycles.

If $k = 1$, then modulo 2- and 3-cycles and relabelling of taxa, $\mathcal{N}^{+}_{abcd}$ is isomorphic to one of the networks in the top row of Figure 3. But for these networks if a, b, c are the taxa in the rooted triple network with a 4-cycle, then the rooted 4-taxon network is obtained by attaching d as an outgroup to it. Thus $\mathcal{N}^{\ominus}_{abcd}$ is determined modulo 2- and 3-cycles.

If $k = 2$, $\mathcal{N}^{+}_{abcd}$ is isomorphic, modulo 2- and 3-cycles and relabeling, to one of the networks in the middle row of Figure 3. Note that for all those rooted quartet networks, the displayed rooted triple networks with 4-cycles are on $\{a, b, c\}$ and $\{a, b, d\}$, and the 4-taxon network can be obtained from either of these by replacing c or d with a cherry on $\{c, d\}$, thus determining $\mathcal{N}^{\ominus}_{abcd}$ modulo 2- and 3-cycles.

If $k = 3$, $\mathcal{N}^{+}_{abcd}$ is isomorphic, modulo 2-, and 3-cycles and relabeling, to one of the networks in the bottom row of Figure 3. In both of these, there is exactly one taxon, a, that is in all three rooted triple networks with 4-cycles, and there is exactly one taxon, c, that has graph-theoretic distance 3 from a in exactly one of the two rooted triples with 4-cycles it appears in. Thus we can determine which taxon is a, and which is c. For the remaining pair b, d, if there is a taxon that is at distance 4 from a in both 4-cycle rooted triple networks it appears in, then the 4-taxon network is the one shown on the left, and that taxon is d. Otherwise, the network is the one shown on the right. In this case there is exactly one rooted triple network on a and c which has its third taxon at distance 2 from the root, and this determines b. Thus we obtain the rooted 4-taxon network $\mathcal{N}^{\oplus}_{abcd}$ modulo 2- and 3-cycles, and hence $\mathcal{N}^{\ominus}_{abcd}$ modulo 2- and 3-cycles

With all rooted 4-taxon networks $\mathcal{N}^{\ominus}_{abcd}$ modulo 2- and 3-cycles determined, we attach an outgroup o to all, giving the collection of all 5-taxon unrooted networks including o, modulo 2- and 3-cycles, induced from the unrooted network $\mathcal{N}'$ formed by attaching o to the root of $\mathcal{N}^{+}$. But the unrooted 4-taxon networks displayed on these 5-taxon ones form the collection of all 4-taxon undirected networks (possibly including o) modulo 2- and 3-cycles displayed on $\mathcal{N}'$.

Lemma 1 now determines $\mathcal{N}'$ modulo 2- and 3-cycles, with directions of cut edges and edges in cycles of size ≥ 5, though not in 4-cycles. Rooting N' by the outgroup o we recover the topology of $\mathcal{N}^{\oplus}$ modulo 2- and 3-cycles and directions of edges in 4-cycles.

4 Expected pattern frequencies as convex sums

The theoretical logDet distance between taxa depends on the matrix of expected relative site-pattern frequencies F^{xy} in aligned sequences for taxa x, y, under the mixture of coalescent mixtures model $\mathcal{M}(\theta)$. The goal of this section is to show that F^{xy} on a level-1 ultrametric rooted triple network can be expressed as a convex combination of frequency matrices for networks with no cycles below the LSA of the taxa. In this way, we reduce the computation of F^{xy} to its computation on simpler networks. This is complicated somewhat by the fact that the convex combination may have terms which are expected pattern frequencies conditioned on a pair of lineages coalescing below a certain node in a network.

The lemmas that follow often involve modifying a network $\mathcal{N}^{+}$ by removing a hybrid edge, to obtain a new network $\mathcal{N}^{+}_{i}$. If one hybrid edge in a cycle is removed, the hybrid node is then suppressed as the other hybrid edge is joined to the descendant tree edge and given the induced length and population size. We retain all other edge lengths and population sizes, as well as hybrid parameters for unaffected cycles. The parameters for the substitution process describing sequence evolution on gene trees are also retained. If θ denotes the full set of parameters associated to $\mathcal{N}^{+}$, then θ_i denotes the full set of parameters associated to $\mathcal{N}^{+}_{i}$ in this way. Notation such as $F^{xy}(\theta)$ or $F^{xy}(\theta_i)$ denotes the dependence of F^{xy} on the parameters θ or θ_i, which include the network $\mathcal{N}^{+}$ or $\mathcal{N}^{+}_{i}$.

The most straightforward network simplifications occur when the hybrid node of a cycle has a single descendant leaf, as depicted by the example 2_1-, 3_1- and 4_1-cycles in Figure 4.

Lemma 4 *(Removing 2_1-cycles) Let $\mathcal{N}^{+}$ be a binary level-1 ultrametric rooted triple network on $\{a, b, c\}$ and let C be a 2_1-cycle in $\mathcal{N}^{+}$ with hybrid edges h_1, h_2. Let $\mathcal{N}^{+}_{1}$ be the network obtained from $\mathcal{N}^{+}$ by removing h_2. Then, under the model $\mathcal{M}$ for any $x, y \in \{a, b, c\}$,*

$$F^{xy}(\theta) = F^{xy}(\theta_1).$$

Proof Since the hybrid node of C has only one descendant, the combined coalescent and substitution process on $\mathcal{N}^+$ can be expressed as a linear combination of those processes on $\mathcal{N}_1^+$, $\mathcal{N}_2^+$, weighted by $\gamma_1 = \gamma(h_1)$, $\gamma_2 = \gamma(h_2)$. That is, for any $x, y \in \{a, b, c\}$,

$$F^{xy}(\theta) = \gamma_1 F^{xy}(\theta_1) + \gamma_2 F^{xy}(\theta_2).$$

But $\mathcal{N}_1^+$ and $\mathcal{N}_2^+$ only differ by h_1 and h_2 which have the same length, though possibly different population sizes. However, since only one lineage can be present in the population for those edges, those population sizes have no impact in model $\mathcal{M}$, so $F^{xy}(\theta_2) = F^{xy}(\theta_1)$. Since $\gamma_1 + \gamma_2 = 1$, the claim follows.

If a network $\mathcal{N}^+$ has multiple 2_1-cycles, then applying Lemma 4 repeatedly gives $F^{xy}(\theta) = F^{xy}(\tilde{\theta})$ where $\widetilde{\mathcal{N}}^+$ is a rooted network with no 2_1-cycles obtained from $\mathcal{N}^+$ by deleting one hybrid edge in each of the 2_1-cycles on $\mathcal{N}^+$.

Lemma 5 *(Decomposing* 3_1- and 4_1-*cycles) Let* $\mathcal{N}^+$ *be a binary level-1 ultrametric rooted triple network on* $\{a, b, c\}$ *and let C be either a* 3_1- *or a* 4_1-*cycle on* $\mathcal{N}^+$. *Let* h_1, h_2 *be the hybrid edges of C with* $\gamma_i = \gamma(h_i)$. *Let* $\mathcal{N}_i^+$ *be the network obtained from* $\mathcal{N}^+$ *by removing* h_j, $j \neq i$. *Then, under the model* $\mathcal{M}$ *for any* $x, y \in \{a, b, c\}$,

$$F^{xy}(\theta) = \gamma_1 F^{xy}(\theta_1) + \gamma_2 F^{xy}(\theta_2).$$

Proof Since the hybrid node of C has only one descendant, we can express the combined coalescent and substitution process on $\mathcal{N}^+$ as a linear combination of the processes of the $\mathcal{N}_i$ with coefficients γ_i, $i = 1, 2$.

A level-1 rooted triple network may have one 4_1-cycle, one 3_1-cycle, or two 3_1-cycles. In the last case, Lemma 5 may be applied twice, to express the pattern frequency matrix under the model as a convex combination of four such matrices for networks with no 3_1-cycles.

With Lemma 4 this shows that computation of the matrix of relative site-pattern frequencies of a level-1 ultrametric rooted triple network $\mathcal{N}^+$ reduces to cases where there are no 2_1-, 3_1-, or 4_1-cycles. The effects of 2_2- and 3_2-cycles are more complicated, however, as a coalescent event may or may not occur below the hybrid nodes of such cycles.

The following definition facilitates studying the impact of such cycles. In it a node p may be either an existing node or a new node introduced along an edge of a network, with appropriate division of the original edge length and population function. Although strictly speaking this second case passes out of the class of binary networks, we allow this only to simplify reference to intermediate states of the coalescent process.

Definition 9 Let $K_p(\theta)$ be the random variable giving the number of lineages at node $p \in V(\mathcal{N}^+)$ under the NMSC. With $X_p \subseteq X$ denoting the set of taxa below p, $K_p(\theta)$ has sample space $\{1, 2, \ldots, |X_p|\}$.

When θ is clear from context we write $K_p = K_p(\theta)$. We also use the notation $F^{xy}_{|K_p=m}(\theta)$ to denote the joint distribution of site patterns conditioned on $K_p = m$ under the model $\mathcal{M}$ with parameters θ.

Lemma 6 *(Decomposing 2_2-cycles) Let $\mathcal{N}^+$ be a binary level-1 ultrametric rooted triple network on $\{a, b, c\}$ without 2_1- or 3_1-cycles. Suppose, as depicted in* Figure 5, *C is a 2_2-cycle on $\mathcal{N}^+$, with edges h_1, h_2 from node q to hybrid node p, hybridization parameters $\gamma_i = \gamma(h_i)$, leaf descendants a, b of p, and no cycles below p. Denote by $\mathcal{N}_i^+$, $i = 1, 2$ the network obtained from $\mathcal{N}^+$ by removing h_j, $j \neq i$ and by $\mathcal{N}_0^+$ the network obtained from $\mathcal{N}^+$ by deleting all edges and nodes below q and attaching edges (q, a) and (q, b) of appropriate length so that $\mathcal{N}_0^+$ is ultrametric. Then, under the model $\mathcal{M}$ for any $x, y \in \{a, b, c\}$,*

$$F^{xy}(\theta) = \gamma_1^2 F^{xy}(\theta_1) + \gamma_2^2 F^{xy}(\theta_2) + P(K_p = 2)2\gamma_1\gamma_2 F^{xy}(\theta_0) + P(K_p = 1)2\gamma_1\gamma_2 F^{xy}_{|K_p=1}(\theta_1).$$

Proof Since the structure of the model for $\mathcal{N}^+$, $\mathcal{N}_1^+$, and $\mathcal{N}_2^+$ is identical below p, we may also use K_p to denote $K_p(\theta_1)$ and $K_p(\theta_2)$. Thus

$$\begin{aligned} F^{xy}(\theta) &= P(K_p = 2)F^{xy}_{|K_p=2}(\theta) + P(K_p = 1)F^{xy}_{|K_p=1}(\theta) \\ &= P(K_p = 2)\left[\gamma_1^2 F^{xy}_{|K_p=2}(\theta_1) + \gamma_2^2 F^{xy}_{|K_p=2}(\theta_2) + 2\gamma_1\gamma_2 F^{xy}(\theta_0)\right] + P(K_p = 1)F^{xy}_{|K_p=1}(\theta). \end{aligned} \quad (2)$$

But since $F^{xy}_{|K_p=1}(\theta) = F^{xy}_{|K_p=1}(\theta_i)$ for $i = 1, 2$ by the argument used for Lemma 4, and the identity $1 = \gamma_1^2 + \gamma_2^2 + 2\gamma_1\gamma_2$,

$$F^{xy}_{|K_p=1}(\theta) = \gamma_1^2 F^{xy}_{|K_p=1}(\theta_1) + \gamma_2^2 F^{xy}_{|K_p=1}(\theta_2) + 2\gamma_1\gamma_2 F^{xy}_{|K_p=1}(\theta_1).$$

Substituting this into equation (2) and using $P(K_p = 1) + P(K_p = 2) = 1$ yields the claim.

Note that while $\mathcal{N}_1^+$ and $\mathcal{N}_2^+$ of Lemma 6 have the same topology and edge lengths, the hybrid edges h_1, h_2 may have different population sizes. Thus $F^{xy}(\theta_1) \neq F^{xy}(\theta_2)$ is possible. This is in contrast to the argument on removing 2_1-cycles in Lemma 4, in which hybrid edge population sizes did not play a role.

Since a level-1 3-taxon rooted network cannot have a 2_2-cycle above a 3_2-cycle, Lemma 6 can be applied recursively to the $\mathcal{N}_i^+$, $i \in \{1, 2\}$ to eliminate all 2_2-cycles. Thus the

remaining complication to producing an expression for $F^{xy}(\theta)$ as a convex combination of such matrices for networks without 2_1-, 3_1-, or 2_2-cycles is the presence of terms of the form $F^{xy}_{|K_p=1}(\theta')$ where $\mathcal{N}'^{+}$ has cherry $\{a, b\}$ and neither 2_1- nor 3_1-cycles. Such terms are handled with the following.

Lemma 7 *(Decomposing 2_2- and 3_2-cycles conditioned on coalescence) Let $\mathcal{N}^+$ be a binary level-1 ultrametric rooted triple network on $\{a, b, c\}$ on which $\{a, b\}$ form a cherry, with no 2_1-, 3_1-, or 4_1-cycles, and at least one 2_2- or 3_2-cycle. (See* Figure 6*.) Let p be the hybrid node parental to the common parent of a, b. Let $\widetilde{\mathcal{N}}^+$ be the network obtained from $\mathcal{N}^+$ by removing one hybrid edge from each 2_2-cycle.*

If $\mathcal{N}^+$ has no 3_2-cycle, then

$$F^{xy}_{|K_p=1}(\theta) = F^{xy}_{|K_p=1}(\widetilde{\theta}).$$

If $\mathcal{N}^+$ has a 3_2-cycle, with hybrid edges h_1, h_2 and hybridization parameters $\gamma_i = \gamma(h_i)$, then let $\widetilde{\mathcal{N}}_i^+$ be the network obtained from $\widetilde{\mathcal{N}}^+$ by removing h_j, $j \neq i$. Then

$$F^{xy}_{|K_p=1}(\theta) = \gamma_1 F^{xy}_{|K_p=1}(\widetilde{\theta}_1) + \gamma_2 F^{xy}_{|K_p=1}(\widetilde{\theta}_2).$$

Proof Conditioned on $K_p = 1$, there is only one lineage in any population above p and below the hybrid node of a 3_2-cycle, if such a cycle is present, or the LSA otherwise. Thus, as in the proof of Lemma 4, no 2_2-cycle will have any effect on the joint distribution. If there is no 3_2-cycle on $\mathcal{N}^+$ this yields the claim. If there is a 3_2-cycle, since only one lineage reaches the hybrid node of the 3_2-cycle, we obtain the claim as in the proof of Lemma 5.

Lemma 8 *(Decomposing 3_2-cycles) Let $\mathcal{N}^+$ be a binary level-1 ultrametric rooted triple network on $\{a, b, c\}$ with no cycles below its LSA except a 3_2-cycle C. Let p denote the hybrid node of C, and h_1, h_2 the hybrid edges with hybridization parameters $\gamma_i = \gamma(h_i)$ and lengths y, z, as depicted at the top of* Figure 7*. Let $\mathcal{N}_1^+$, $\mathcal{N}_2^+$, $\mathcal{N}_3^+$, and $\mathcal{N}_4^+$ be the networks derived from $\mathcal{N}^+$ shown at the bottom of* Figure 7*. Then, under the model $\mathcal{M}$, for any $x, y \in \{a, b, c\}$, with $K_p = K_p(\theta)$,*

$$\begin{aligned} F^{xy}(\theta) = \gamma_1^2 F^{xy}(\theta_1) + \gamma_2^2 F^{xy}(\theta_2) + P(K_p = 2)\gamma_1\gamma_2(F^{xy}(\theta_3) + F^{xy}(\theta_4)) \\ + P(K_p = 1)\gamma_1\gamma_2\left(F^{xy}_{|K_p=1}(\theta_1) + F^{xy}_{|K_p=1}(\theta_2)\right). \end{aligned}$$

Proof Observe that

$$\begin{aligned} F^{xy}(\theta) &= P(K_p = 2)F^{xy}_{|K_p=2}(\theta) + P(K_p = 1)F^{xy}_{|K_p=1}(\theta) \\ &= P(K_p = 2)\Big[\gamma_1^2 F^{xy}_{|K_p=2}(\theta_1) + \gamma_2^2 F^{xy}_{|K_p=2}(\theta_2) + \gamma_1\gamma_2 F^{xy}(\theta_3) + \gamma_1\gamma_2 F^{xy}(\theta_4)\Big] \\ &\quad + P(K_p = 1)F^{xy}_{|K_p=1}(\theta)\,. \end{aligned} \tag{3}$$

Since $F^{xy}_{|K_p=1}(\theta) = \gamma_1 F^{xy}_{|K_p=1}(\theta_1) + \gamma_2 F^{xy}_{|K_p=1}(\theta_2)$ and $\gamma_1 + \gamma_2 = 1$,

$$F^{xy}_{|K_p=1}(\theta) = \gamma_1^2 F^{xy}_{|K_p=1}(\theta_1) + \gamma_2^2 F^{xy}_{|K_p=1}(\theta_2) + \gamma_1\gamma_2\Big(F^{xy}_{|K_p=1}(\theta_1) + F^{xy}_{|K_p=1}(\theta_2)\Big).$$

Using this and $P(K_p = 1) + P(K_p = 2) = 1$ in equation (3) yields the claim.

5 Theoretical logDet distances

In this section, we show that, under the mixture of coalescent mixtures model $\mathscr{M}$ on an ultrametric level-1 rooted triple network, the theoretical logDet distances between taxa determine most topological features of the network. The previous section established that the pattern frequency matrices for the model on such networks can be expressed as convex combinations of those on simpler networks (possibly subject to conditioning), whose only cycles are 2_3-cycles located above LSA(a, b, c), such as depicted in Figure 2. The following algebraic lemma is key to drawing conclusions about the determinants of such linear combinations of matrices.

Lemma 9 ([Allman et al., 2019b], **Lemma 3.1**) *Suppose for each i, F_i and G_i are $\kappa \times \kappa$ symmetric positive definite matrices such that $y^T F_i y \geq y^T G_i y$ for every $y \in \mathbb{R}^\kappa$ with the inequality strict for some y and some i. For $\alpha_i \geq 0$, let*

$$F = \sum_{i=1}^{m} \alpha_i F_i, \quad G = \sum_{i=1}^{m} \alpha_i G_i\,.$$

Then

$$\det F > \det G\,.$$

Analyzing the pattern frequency matrix for networks with 2_3-cycles above LSA(a, b, c) requires a detailed look at the coalescent process in such a chain of 2-cycles. For a simple case, aspsume lineages x and y enter the single cycle chain depicted in Figure 8. Population functions N_1, N_2, N_3, and N_4 are fixed for each edge, where for convenience, we shift domains from the convention in Section 2.2 so that N_1 is defined on $[0, t_0)$, N_2, N_3 on $[t_0, t_1)$, and N_4 on $[t_1, \infty)$.

The probability density $c(t)$ for time to coalescence of the lineages x, y entering at the bottom node ($t = 0$) can be calculated piecewise as follows: For $t \in [0, t_0)$,

$$c(t) = \frac{1}{N_1(t)} \exp\left(-\int_0^t \frac{1}{N_1(\tau)}\, d\tau\right), \tag{4}$$

as given by Allman et al. [2019b].

For $t \in [t_0, t_1)$,

$$c(t) = p_0\left(\gamma^2 c_2(t) + (1-\gamma)^2 c_3(t)\right)$$

where $p_0 = 1 - \int_0^{t0} c(t)\, dt$ is the probability of no coalescence before t_0, and for $i = 2, 3$

$$c_i(t) = \frac{1}{N_i(t)} \exp\left(-\int_{t_0}^t \frac{1}{N_i(\tau)} d\tau\right).$$

Finally, for $t \in [t_1, \infty)$, with $p_1 = 1 - \int_0^{t1} c(t)\, dt$ the probability of no coalescence before t_1,

$$c(t) = p_1 \frac{1}{N_4(t)} \exp\left(-\int_{t_1}^t \frac{1}{N_4(\tau)}\, d\tau\right).$$

It is straightforward to extend this analysis of $c(t)$ to a chain with an arbitrary number of 2-cycles. Since we will not need an explicit formula for the distribution of coalescent times for two lineages entering such a chain of 2-cycles, we omit a complete derivation, and only state the properties of it that we use.

Formally, a *chain of 2-cycles* is a species network with leaf a_0, internal vertices b_1, a_1, b_2, $a_2, \ldots, a_n$, with root $r = a_n$, tree edges $e_i = (b_i, a_{i-1})$, and hybrid edges $e_i' = (a_i, b_i)$, $e_i'' = (a_i, b_i)$, together with edge lengths, piecewise-continuous population size functions on each edge, including above the root, and hybrid parameters γ_i', $\gamma_i'' = 1 - \gamma_i'$ for each pair of hybrid edges e_i', e_i''.

Using the technical assumptions given in Subsection 2.2, it is straightforward to deduce the following.

Lemma 10 *Consider a fixed chain of 2-cycles with leaf a_0. Let $c : [0, \infty) \to \mathbb{R}^{\geq 0}$ denote the probability density function under the NMSC for the time T of coalescence of two lineages entering the chain at a_0. Then $c(t)$ is piecewise continuous, and $c(t) > 0$ for all $t \in [0, \infty)$.*

The next three technical lemmas generalize Lemmas 4.1, 4.4, and 4.5 of Allman et al. [2019b] from a tree to a network setting. These culminate in Proposition 2 below, which justifies the application of Lemma 9.

Lemma 11 *Let* $c:[0,\infty)\to\mathbb{R}^{\geq 0}$ *be the probability density function under the NMSC for the time T of coalescence of two lineages entering a chain of 2-cycles, and for times* $t_2 > t_1 \geq 0$ *let* c_i *be the conditional density given* $T \geq t_i$. *Then the cumulative distribution functions for* c_1 *and* c_2 *satisfy*

$$C_1(t) \geq C_2(t),$$

with the inequality strict on some interval.

Proof Since $0 = c_2(t) \leq c_1(t)$ for all $t \leq t_2$, the inequality is immediate for $t \leq t_2$. Since using Lemma 10 we have $c_1(t) > c_2(t) = 0$ for $t \in (t_1, t_2)$, the inequality is strict on a subinterval.

For $t \geq t_2$, let $J = \int_{t_1}^{t_2} c_1(t)\,dt$ and $I(t) = \int_{t_2}^{t} c_1(s)\,ds$, so

$$\begin{aligned} C_1(t) - C_2(t) &= J + I(t) - \frac{I(t)}{1-J} \\ &= J - \frac{J}{1-J} I(t). \end{aligned}$$

Differentiating and using Lemma 10 shows $C_1(t) - C_2(t)$ is decreasing for $t > t_2$. Since $C_1(t) - C_2(t) \to 0$ as $t \to \infty$, this implies $C_1(t) - C_2(t) \geq 0$, as claimed.

Lemma 12 *Let* c_1, c_2 *be probability density functions on* $[0,\infty)$, *with cumulative distribution functions* C_1, C_2, *such that* $C_1(t) \geq C_2(t)$ *for all t, with the inequality strict on some interval. Let* $s(t) = \int_0^t \mu(x)\,dx$ *for a positive, piecewise-continuous* μ *on* $[0,\infty)$ *such that* $s(\infty) = \infty$. *For* $\lambda \leq 0$ *let*

$$f(\lambda, \mu, C_i) = \int_0^\infty \exp(2\lambda s(t)) c_i(t)\,dt.$$

Then if $\lambda = 0$,

$$f(0, \mu, C_1) = f(0, \mu, C_2) = 1.$$

while for $\lambda < 0$

$$f(\lambda, \mu, C_1) > f(\lambda, \mu, C_2).$$

Proof For $\lambda = 0$ we find $f(0, \mu, C_i) = \int_0^\infty c_i(t)\,dt = 1$.

If $\lambda < 0$, integrating by parts yields

$$f(\lambda, \mu, C_i) = \exp(2\lambda s(t))C_i(t) \mid_{t=0}^{\infty} - 2\lambda \int_0^{\infty} \mu(t) \exp(2\lambda s(t))C_i(t)dt$$
$$= -2\lambda \int_0^{\infty} \mu(t) \exp(2\lambda s(t))C_i(t)dt .$$

Thus

$$f(\lambda, \mu, C_1) - f(\lambda, \mu, C_2) = -2\lambda \int_0^{\infty} \mu(t) \exp(2\lambda s(t))(C_1(t) - C_2(t))dt .$$

As the integrand is non-negative, and positive on some interval, the claim for $\lambda < 0$ follows.

Lemma 13 *Consider a GTR substitution model with rate matrix* $Q \neq 0$, *a scalar-valued rate function* $\mu(t)$ *satisfying the assumptions of* Subsection 2.3, *and a cumulative distribution function* $C(t)$ *for the time* T *to coalescence of 2 lineages in a population.*

Let $F(x) = F(Q, \mu, C, x)$ *be the expected site-pattern frequency array for two lineages that enter a population at time 0 and undergo substitutions at rate* $\mu(t)Q$ *conditioned on* $T \geq x$. *For* $x < x_1$ *let* $\widetilde{F}(x, x_1) = \widetilde{F}(Q, \mu, C, x, x_1)$ *be the expected site-pattern frequency array for two lineages that enter a population at time 0 and undergo substitutions at rate* $\mu(t)Q$ conditioned, on $x < T < x_1$.

Then for all $0 \neq y \in \mathbb{R}^k$ *the functions* $y^T F(x)y$ *and* $y^T \widetilde{F}(x, x_1)y$ *are positive-valued and decreasing in x. Moreover there exists a y for which both are strictly decreasing, and for which if* $x_0 < x_1 \leq x_2$

$$y^T \widetilde{F}(x_0, x_1)y > y^T F(x_2)y .$$

Proof Let $c_x(t)$ denote the conditional probability density function for the coalescent time T given $T > x$. With $s(t) = \int_0^t \mu(\tau)\, d\tau$, the Markov matrix describing the substitution process on a single lineage from time 0 to time t is

$$M(\mu, Q, t) = \exp(s(t)Q) .$$

Thus using time-reversibility of the substitution process, with $\boldsymbol{\pi}$ the stationary distribution for Q,

$$F(x) = \mathrm{diag}(\pi) \int_0^{\infty} (M(\mu, Q, t))^2 c_x(t)dt .$$

Here the square of the Markov matrix accounts for substitutions in the two lineages before coalescence.

Now $S^{-1}QS$ is diagonal for a matrix $S = \mathrm{diag}(\boldsymbol{\pi})^{-1/2}U$ with U orthogonal, and Q's eigenvalues satisfy $0 = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$ with at least one $\lambda_i < 0$ (Lemma 2.2 of Allman et al. [2019b]). Thus diagonalizing the Markov matrix yields

$$U^T \mathrm{diag}(\pi)^{-1/2} F(x) \mathrm{diag}(\pi)^{-1/2} U = \int_0^\infty \Lambda_{M(\mu,Q,t)} c_c(t)\, dt$$

where $\Lambda_M(\mu,Q,t)$ is diagonal with entries $\exp(2s(t)\lambda_i)$. The diagonal entries of this integral are thus

$$\int_0^\infty \exp(2s(t)\lambda_i) c_x(t)\, dt\,.$$

But Lemmas 11 and 12 show this is positive, decreasing in x, and strictly decreasing for some i. This establishes the claims about F, by choosing y to be any eigenvector of Q whose eigenvalue is negative to obtain a strictly decreasing function.

The corresponding claims about $\widetilde{F}$ are given by the same argument with the cumulative distribution function C replaced by the conditional distribution function given the coalescent time $T < x_1$, that is, with

$$\widetilde{C}_{x_1}(t) = \begin{cases} C(t) / C(x_1) & \text{if } t \leq x_1 \\ 1 & \text{if } t > x_1 \end{cases}\,.$$

Finally, since for every t the function $\widetilde{C}_{x_1}(t)$ is decreasing in x_1, then for any y and x_0, a similar diagonalization argument and again using Lemma 12 shows the function $y^T \widetilde{F}(x_0, x_1) y$ is decreasing in x_1. Thus if $x_0 < x_1 \leq x_2$, then

$$y^T \widetilde{F}(x_0, x_1) y \geq \lim_{x_1 \to \infty} y^T \widetilde{F}(x_0, x_1) y = y^T F(x_0) y \geq y^T F(x_2) y\,.$$

Moreover, if y is an eigenvector of Q whose eigenvalue is negative, then strict inequality holds.

Proposition 2 *Let $\mathcal{N}^+$ be a binary level-1 ultrametric rooted triple network on $\{a, b, c\}$ whose LSA network has topology $((a, b), c)$, but above $\mathrm{LSA}(\{a, b, c\}, \mathcal{N}^+)$ there is possibly a chain of 2-cycles. Then, under a coalescent mixture model on $\mathcal{N}^+$ with fixed parameters $\mu(t)$, $\{N_e\}$, Q, $\boldsymbol{\pi}$, the relative site-pattern frequency matrices F^{ab}, F^{bc}, and F^{ac} are symmetric positive definite, with $F^{ac} = F^{bc}$, and satisfy*

$$y^T F^{ab} y \geq y^T F^{ac} y$$

for every $y \in \mathbb{R}^k$, *with the inequality strict for some y. Moreover, the same statements hold when the arrays* F^{xy} *are replaced by* $F^{xy}_{|K_p=1}$ *with p a node placed above the parent of a, b and below the parent of c.*

Proof Let x_1 be the length of the pendant edges to a and b, and x_2 the length of the pendant edge to c, so $x_2 > x_1$. Then applying Lemma 13 for an appropriately chosen distribution $C(t)$ of coalescent times so

$$F^{ab} = F(x_1), \quad F^{ac} = F^{bc} = F(x_2),$$

the result is immediate.

Let x_p denote the distance from a or b to p, so $x_1 < x_p < x_2$. Then conditioning on $K_p = 1$, in the notation of Lemma 13 we have

$$F^{ab}_{|K_p=1} = \tilde{F}(x_1, x_p), \quad F^{ac}_{|K_p=1} = F^{bc}_{|K_p=1} = F^{bc} = F(x_2),$$

so again Lemma 13 yields the claim.

We now turn from considering a coalescent mixture model, with a single substitution model class, to the mixture of coalescent mixtures $\mathcal{M}$.

Lemma 14 *Let* $\mathcal{N}^+$ *be a level-1 ultrametric rooted triple network on* $\{a, b, c\}$ *with no* 4*-cycle. Suppose* $\{a, b\}$ *form a cherry in the tree topology obtained from suppressing all cycles of* $\mathcal{N}^+$. *Then, under the mixture of coalescent mixtures model* $\mathcal{M}$ *on* $\mathcal{N}^+$, $F^{ac}(\theta) = F^{bc}(\theta)$.

Proof By Lemmas 4 and 5, we may assume $\mathcal{N}^+$ has neither a 2_1- nor a 3_1-cycle, so there are no cycles below the parent of *a, b*. By the ultrametricity of the network, a and b are exchangeable under the combined coalescent and substitution model for each substitution model class, and therefore for the model $\mathcal{M}$.

This result is used to show that logDet distances from rooted triple networks with only 2- and 3_1-cycles satisfy the same equality and inequality relationships as those from trees.

Proposition 3 *(No* 4_1*-cycles or* 3_2*-cycles) Let* $\mathcal{N}^+$ *be a level-1 ultrametric rooted triple network on* $\{a, b, c\}$ *with neither a 4-cycle nor a* 3_2*-cycle. Let* $\mathcal{T} = ((a, b), c)$ *be the tree topology obtained after suppressing all cycles in* $\mathcal{N}^+$. *Under the mixture of coalescent mixtures model* $\mathcal{M}$ *on* $\mathcal{N}^+$ *the theoretical logDet distances satisfy*

$$d_{LD}(a, c) = d_{LD}(b, c) > d_{LD}(a, b).$$

Proof Under the model $\mathcal{M}$, the frequencies of bases at any taxon are identical, given by the same convex combination of the base frequency vectors $\boldsymbol{\pi}_i$ for substitution classes i. Thus

the value of $\ln(g_{ua}g_{va})$ in the definition of the logDet distance, equation (1), is identical for every pair of distinct taxa $x, y \in \{a, b, c\}$. It thus suffices to show

$$\det F^{ab}(\theta) \geq \det F^{ac}(\theta) = \det F^{bc}(\theta).$$

Lemma 14 gives the equality. By Lemmas 4, 5, and 6, we can express $F^{xy}(\theta)$ as a convex combination of relative site-pattern frequency matrices, possibly conditioned on $K_p = 1$, of networks of the form of the tree $\mathcal{T}$ joined to a (possibly empty) chain of 2-cycles above $\mathcal{T}$'s root, such as depicted in Figure 2. By Proposition 2 each of those matrices for coalescent mixture models satisfy the hypotheses of Lemma 9. Lemma 9 thus yields the claim for mixtures of coalescent mixtures by considering a convex combination across both the networks and substitution model classes.

A weaker result, without the inequality, applies to networks with 3_2-cycles.

Proposition 4 *(3_2-cycle) Let $\mathcal{N}^+$ be a level-1 ultrametric rooted triple network on $\{a, b, c\}$ with a 3_2-cycle. Let $\mathcal{T} = ((a, b), c)$ be the tree topology obtained after suppressing all cycles in $\mathcal{N}^+$. Then under the mixture of coalescent mixtures model $\mathcal{M}$ on $\mathcal{N}^+$, the theoretical logDet distances satisfy*

$$d_{LD}(a, c) = d_{LD}(b, c).$$

Proof From Lemma 14, $F^{ac}(\theta) = F^{bc}(\theta)$, so the result follows as in the previous proof.

Proposition 3, and the arguments leading to it, show that the equality and inequality relationships of logDet distances between only 3 taxa carry no signal of either 2- or 3_1-cycles. Proposition 4, however, leaves open the possibility that for a network with a 3_2-cycle the smallest distance may not necessarily correspond to the taxa which are neighbors after 2- and 3- cycles are suppressed. This suggests that the presence of a 3_2-cycle might be detectable, at least under some circumstances. In Section 7 we return to this issue, providing a more in-depth analysis of triples of logDet distances.

Proposition 5 *(4_1-cycle) Let $\mathcal{N}^+$ be a level-1 ultrametric rooted triple network on $\{a, b, c\}$ with a 4-cycle, such that contracting all cycles except the 4-cycle and then deleting one of its hybrid edges gives the trees $((a, b), c)$ and $((a, c), b)$. (See* Figure 9*.) Then under the mixture of coalescent mixtures model $\mathcal{M}$ on $\mathcal{N}^+$, the theoretical logDet distances satisfy*

$$d_{LD}(b, c) > d_{LD}(a, b) \text{ and } d_{LD}(b, c) > d_{LD}(a, c).$$

Moreover, if all other parameters are fixed, then for generic values of the hybridization parameters,

$$d_{LD}(a, b) \neq d_{LD}(a, c).$$

Proof As in Proposition 3, to establish these inequalities for the logDet distance, it is enough to show

$$\det F^{bc}(\theta) < \det F^{ab}(\theta) \text{ and } \det F^{bc}(\theta) < \det F^{ac}(\theta)\,. \tag{5}$$

From Lemmas 4 and 5, for $x, y \in \{a, b, c\}$

$$F^{xy}(\theta) = \gamma_1 F^{xy}(\theta_1) + \gamma_2 F^{xy}(\theta_2)$$

where $\mathcal{N}_1^+$ and $\mathcal{N}_2^+$ have the structure of the trees $((a, b), c)$ and $((a, c), b)$ with chains of 2-cycles possibly attached above their roots. Proposition 2 implies that for each GTR substitution model class

$$y^T F^{ab}(\theta_1) y \geq y^T F^{bc}(\theta_1) y = y^T F^{ac}(\theta_1) y \quad \text{and} \quad y^T F^{ac}(\theta_2) y \geq y^T F^{ab}(\theta_2) y = y^T F^{bc}(\theta_2) y,$$

for every $y \in \mathbb{R}^k$, with the inequalities strict for some choices of y. From this and Lemma 9 we obtain the in equalities (5).

To see $d_{LD}(a, b) \neq d_{LD}(a, c)$ for generic hybridization parameters, first observe that these distances extend to analytic functions of the γ on all of $\mathbb{C}$. To show the inequality for generic γ, it is enough to show there exists one specific choice of $\gamma \in \mathbb{C}$ for which they are not equal. First consider a choice on the boundary of the parameter space, by letting $\gamma_e = 1$, $\gamma_{e'} = 0$ for every pair e, e' of hybrid edges with a common child so that the model reduces to one on the tree $((a, c), b)$. In this case Theorem 1 of Allman et al. [2019b] establishes the inequality. Continuity implies that there are then choices of $0 < \gamma_e < 1$, where the model does not degenerate to one on a tree, for which these distances are also not equal.

Assuming generic parameter values, Proposition 5 combined with earlier results implies that the presence of a 4-cycle is indicated by three distinct logDet distances computed from expected pattern frequencies. However, the three networks at the top of Figure 9 all satisfy the hypothesis of Proposition 5, but using equalities and inequalities of logDet distances we cannot distinguish them. We can only identify their undirected version as depicted in the bottom of Figure 9.

Nonetheless, the combinatorial result of Proposition 1 yields information on larger cycles and their hybrid nodes by first using logDet distances to determine undirected rooted triple networks. This gives our main result.

Theorem 1 *Let $\mathcal{N}^+$ be a binary level-1 ultrametric network on X with a $|X| \geq 3$. Let $\widetilde{N}$ denote the topological LSA network $\mathcal{N}^{\oplus}$ modulo 2- and 3-cycles and directions of edges in 4-cycles. Then for generic hybridization parameters under the mixture of coalescent mixtures model $\mathcal{M}$ on $\mathcal{N}^+$, $\widetilde{N}$ is identifiable from the theoretical logDet distances for pairs of taxa.*

Proof Propositions 3, 4, and 5 imply that for generic parameters the three logDet distances for any choice of 3 taxa are distinct if, and only if, the induced rooted triple network has a 4-cycle. Moreover, the unrooted topology of the 4-cycle is determined by the largest of the three distances. Thus the set S of Proposition 1 is determined, yielding the result.

An example of a rooted level-1 network and the structure that we have shown to be identifiable from logDet distances under the model $\mathscr{M}$ is given in Figure 10. On the left is a level-1 rooted phylogenetic network with cycles of various sizes, and on the right the partially directed network that could be inferred from it for generic parameters.

6 Modifying the model

In this section we show how our results apply to two variants of the model used throughout earlier sections. In the first, we no longer require that sites be independent, allowing instead finite subsets of sites (e.g., modeling individual genes) evolving on common gene trees. In the second, we consider a limiting case of the model, in which gene lineages entering a population have an immediate common ancestor, without any delay from a coalescent process. Other variants, such as one combining the features of the two considered here, could be treated similarly.

6.1 Variant 1: A model for unlinked genes

The first model variation allows for unlinked genetic loci, each composed of linked sites evolving on a common gene tree. This is a relaxation of the model assumption in Section 2 that sites be unlinked. The original model only properly applies to unlinked SNP data, while this variant allows for concatenated gene sequences. We require only that the length of each locus be a random draw from some length distribution with finite mean, independent of the topology of the gene tree.

To formalize this, let g be a probability mass function supported on $\mathbb{N}$, with mean $m = \sum_{n=1}^{\infty} g(n)n < \infty$. The model description in Section 2 is modified so that sequence data is generated as follows: For each gene,

1. a gene tree T is sampled according to the NMSC model on $(\mathcal{N}^+, \{\ell_e\}, \{\gamma_e\})$ with population sizes $\{N_e\}$,
2. class i is sampled from the distribution λ to determine parameters $(Q_i, \boldsymbol{\pi}_i, \mu_i)$, and gene length n is sampled according to g, and
3. for n independent sites, the bases for each extant taxon $x \in X$ are sampled under the GTR+μ process on T with parameters $(Q_i, \boldsymbol{\pi}_i, \mu_i)$.

All sites are then summarized by a site pattern frequency array, so that information as to which sites evolved on the same gene tree is lost.

To show that Theorem 1 applies to this model, we need only show that the expected pattern frequency array for two taxa, $\widetilde{F}^{ab}$, under this model, is the same as the expectation, F^{ab}, under the model of Section 2. Let $\widetilde{F}^{ab}_{|T}$ and $F^{ab}_{|T}$ denote expected pattern frequencies

conditioned on a particular gene tree T. Then with dT denoting the probability measure for gene trees under the NMSC with the given parameters,

$$\begin{aligned}\widetilde{F}^{ab} &= \int_T \widetilde{F}^{ab}_{\,|\,T}\, dT \\ &= \int_T \left(\frac{1}{m}\sum_{n=1}^{\infty} g(n) n F^{ab}_{\,|\,T}\right) dT \\ &= \int_T \left(\frac{1}{m}\sum_{n=1}^{\infty} g(n) n\right) F^{ab}_{\,|\,T}\, dT \\ &= \int_T F^{ab}_{\,|\,T}\, dT \\ &= F^{ab}.\end{aligned}$$

Note that in applications of the theory developed here, empirical frequency arrays produced from gene sequences are likely to converge more slowly to their expected values than for those produced from SNP data, due to the linkage of sites. The argument above suggests that enough genes are needed so that the variation in gene length averages out over each possible gene tree.

6.2 Variant 2: A non-coalescent model

The second model variation we consider is a non-coalescent model for an ultrametric level-1 species network, in which gene trees must be displayed on the species network. One can think of this as simply requiring immediate coalescence of gene lineages when they enter a common population. Population size parameters are thus no longer relevant, but all other features of the model of Section 2 are retained.

This model is similar to the non-coalescent model considered by Gross et al. [2020], who used algebraic and combinatorial arguments to obtain an identifiability result for most features of a level-1 species network topology assuming generic numerical parameters. However, we impose one more restrictive assumption, namely that the network be ultrametric. On the other hand, we considerably relax their assumptions on the sequence substitution model, from a requirement of a single Jukes-Cantor or Kimura process to the mixture of GTR processes used throughout this paper.

Informally, to produce immediate coalescence of gene lineages in a coalescent model, one can simply take a limit as the population sizes approach 0. Small population size produces bottlenecks, which encourage rapid coalescence of lineages. In general, results obtained under the coalescent model will still apply under a non-coalescent model, provided the arguments respect taking such a limit.

To sketch how this applies in our arguments, first fix all population sizes N_e on edges in a species network to have a common value N. Note that population size plays no role in any of our arguments before those of Section 5, except through probabilities such as $P(K_p = 1)$ and $P(K_p = 2)$ which appear in formulas in Section 4 but are not computed there. Thus all results through Section 4 remain valid.

As $N \to 0^+$, the density function $c(t)$ of equation (4) for the time to coalescence of two lineages in a population is easily seen to approach δ_0, a point mass at $t = 0$. Thus with probability 1 lineages coalesce immediately upon entering a common population. While this observation can be traced through the remaining lemmas of Section 5 (making some modifications to their presentation), it is simpler to give a direct proof of the following analog of Proposition 2.

Proposition 6 *Let $\mathcal{N}^+$ be a binary level-1 ultrametric rooted triple network on $\{a, b, c\}$ whose LSA network has topology $((a, b), c)$, but above $\mathrm{LSA}(\{a, b, c\}, \mathcal{N}^+)$ there is possibly a chain of 2-cycles. Then, under a non-coalescent model on $\mathcal{N}^+$ with fixed parameters $\mu(t)$, Q, π, the relative site-pattern frequency matrices F^{ab}, F^{bc}, and F^{ac} are symmetric positive definite, with $F^{ac} = F^{bc}$, and satisfy*

$$y^T F^{ab} y \geq y^T F^{ac} y$$

for every $y \in \mathbb{R}^k$, with the inequality strict for some y.

Proof Let x_1 be the length of the pendant edges to a, b and x_2 the length of the pendant edge to c. With $s(t) = \int_0^t \mu(\tau)\, d\tau$, the Markov matrix describing the substitution process on a single lineage from time 0 to time t is

$$M(\mu, Q, t) = \exp(s(t)Q).$$

Thus using time-reversibility of the substitution process

$$F^{ab} = \mathrm{diag}(\pi) M(\mu, Q, x_1)^2 = \mathrm{diag}(\pi) \exp(2s(x_1)Q)$$

$$F^{ac} = F^{bc} = \mathrm{diag}(\pi) M(\mu, Q, x_2)^2 = \mathrm{diag}(\pi) \exp(2s(x_2)Q)\ .$$

Since Q is a GTR rate matrix, the result follows by diagonalization, as in Lemma 13.

The remainder of the arguments of Section 5 apply unchanged, to yield an analog of Theorem 1. Note that while population sizes are no longer model parameters, all other parameters are unchanged in the limit.

Remark 1 In general, results under the MSC and NMSC models yield results for simpler non-coalescent models in the limit as population sizes decrease to 0. For instance, without considering a site substitution process Baños [2019] and Allman et al. [2019a] show that most features of a level-1 network can be identified from the frequencies of displayed gene quartet trees under the NMSC. Letting all population sizes $\to 0$ then gives that most features of a level-1 network can be identified from the frequencies of its displayed quartet trees.

7 Normalized triples of logDet distances.

In the previous section, we obtained linear equalities and inequalities that the logDet distances between three taxa must satisfy if they are related by various level-1 rooted networks. Combined with the combinatorial result of Section 3 these are sufficient for proving the identifiability claim that is the main focus of this work. However, it is worthwhile to seek a more complete characterization of what distances are achievable by various network topologies. In particular, with an eye toward practical application, any tighter characterizations would enable stronger testing for network topology from the empirical distances.

Here we conduct a partial investigation, characterizing not the triple of theoretical logDet distances that may be produced on rooted 3-taxon networks, but rather the *normalized triple* obtained by dividing the distances by their sum. The triple of distances forms a point in the non-negative octant $(\mathbb{R}_{\geq 0})^3$, while the normalized triple gives a point in the 2-dimensional simplex. Thus plots can be made with the normalized distances that are analogous to the simplex plots for visualizing gene quartet concordance factors [Baños, 2019, Mitchell et al., 2019, Allman et al., 2021]. Just as simplex plots of concordance factors aid in understanding genomic data sets, we anticipate that the 2-simplex visualization of the normalized logDet distance triples will be similarly useful.

We begin with the logDet triples from 3-taxon trees.

Proposition 7 *Let* $\ell = (\ell_{ab}, \ell_{ac}, \ell_{bc})$ *with* $0 < \ell_{ab} \leq \ell_{ac} = \ell_{bc}$ *be a triple of positive numbers summing to 1. Then there exists an ultrametric rooted tree with topology* $((a, b), c)$ *and GTR substitution model parameters such that the normalized theoretical logDet distances of sequences generated under the coalescent mixture model are* ℓ.

Proof Consider the metric species tree $((a{:}0, b{:}0){:}x/2, c{:}x/2)$, and constant population sizes $\epsilon > 0$ on all edges. Fix a single substitution model, say the Jukes-Cantor, for sequence generation. Since small population sizes ϵ result in rapid coalescence with arbitrarily high probability, by taking ϵ sufficiently small one can show the expected frequency array can be made arbitrarily close to that which would arise if all gene trees exactly matched the species tree. Thus the theoretical logDet distances can be made arbitrarily close to $d_{LD}(a, b) = 0$ and $d_{LD}(a, c) = d_{LD}(b, c) = x$, which normalizes to (0, 1/2, 1/2).

The unresolved species tree $(a{:}x/2, b{:}x/2, c{:}x/2)$, regardless of choice of population functions on the edges yields, by exchangeability of the taxa, a triple of equal logDet distances, which normalizes to (1/3, 1/3, 1/3).

While the two trees above have 0-length edges and hence are non-binary, perturbations to binary trees with positive length edges can produce normalized logDet distances that are arbitrarily close.

Since the normalized logDet distances are continuous functions of parameters, the parameter space is connected, and the image of the normalized distances lies in a line segment by Proposition 3, the claim follows.

We turn now to networks with a single cycle.

Proposition 8 *Let* $\ell = (\ell_{ab}, \ell_{ac}, \ell_{bc})$ *with* $0 < \ell_{ab} \leq \ell_{ac} < \ell_{bc}$ *be a triple of positive numbers summing to 1. Then there exists a binary ultrametric rooted network on taxa a, b, c with a single 4-cycle and GTR substitution model parameters such that the normalized theoretical logDet distances of sequences generated, under a single-class coalescent mixture model are* ℓ

Proof The 4-cycle network we construct is shown in Figure 11, with t_0, t_1 measured in generations, and the hybrid edges of length 0. Consider a single constant population size $N > 0$ for all populations over the tree and above the root, and a Jukes-Cantor substitution process with constant rate $\mu > 0$. We will choose values for t_0, $t_1 > 0$, $\gamma \in [1/2, 1)$ so that the normalized distances for the coalescent mixture model with this single substitution process are given by ℓ

Recall that if $M(t)$ denotes the Jukes-Cantor Markov matrix for a substitution process over time t with rate 1, then the common value of all its off-diagonal entries is

$$f(t) = \frac{1}{4}\left(1 - e^{-\frac{4}{3}t}\right).$$

With $D = \text{diag}(1/4, 1/4, 1/4, 1/4)$, the Jukes-Cantor pattern frequency array is $DM(t)$, and the logDet distance (equal to Jukes-Cantor distance) is

$$t = f^{-1}(f(t)) = -\frac{3}{4}\log(1 - 4f(t)).$$

Note that f is an increasing function.

From equation 4.1 of Allman et al. [2019b], for a coalescent mixture Jukes-Cantor model on an ultrametric tree with uniform population size N and mutation rate μ, sequences for two taxa x, y whose MRCA is at time t before the present has expected pattern frequency array

$$F(t) = DM(2t\mu)\widetilde{M}(\mu, N),$$

where $\widetilde{M}(\mu, N)$ is a Markov matrix of Jukes-Cantor form describing the expected additional substitutions due to the coalescent model delaying lineages merging until some time above the MRCA. The logDet distance between x, y is then the same as the Jukes-Cantor distance, which is computed to be

$$d_{LD}(x, y) = 2t\mu + \beta$$

where $\beta = \beta(\mu, N) > 0$ can be explicitly computed from $\widetilde{M}(\mu, N)$, though we will not do so here. Since β is continuous and $\beta(\mu, N) \to 0$ as $N \to 0$ and $\beta(\mu, N) \to \infty$ as $N \to \infty$, it follows that β takes on all positive values.

Now by Lemma 5 on the 4-cycle network of Figure 11 the expected pattern frequency array for a, b is

$$\gamma F(t_0) + (1-\gamma)F(t_1) = DM_{ab}\widetilde{M}(\mu, N)$$

where

$$M_{ab} = \gamma M(2t_0\mu) + (1-\gamma)M(2t_1\mu)$$

has the usual Jukes-Cantor form, with off-diagonal entries

$$f_{ab} = \gamma f(2t_0\mu) + (1-\gamma)f(2t_1\mu)\,.$$

This shows

$$d_{LD}(a,b) = f^{-1}(f_{ab}) + \beta\,.$$

A similar calculation shows

$$d_{LD}(a,c) = f^{-1}(f_{ac}) + \beta,$$

where

$$f_{ac} = \gamma f(2t_1\mu) + (1-\gamma)f(2t_0\mu)\,.$$

The expected pattern frequencies for b, c sequences is $F(t_1)$, so

$$d_{LD}(b,c) = f^{-1}(f_{bc}) + \beta$$

where

$$f_{bc} = f(2t_1\mu)\,.$$

We now determine parameters which produce the normalized triple of distances ℓ. Fixing values of μ, N determines a fixed value of $\beta > 0$. Next, choose some value m so that

$$f(m\ell_{ab} - \beta) > \frac{1}{8},$$

which can be done since $f : \mathbb{R}^{>0} \to (0, 1/4)$ is surjective and increasing. Then, with $x_{ij} = f(m\ell_{ij} - \beta)$, because $\ell_{ab} \le \ell_{ac} < \ell_{bc}$ we have

$$\frac{1}{8} < x_{ab} \leq x_{ac} < x_{bc} < \frac{1}{4}.$$

Let $x_0 = x_{ab} + x_{ac} - x_{bc}$, so $0 < x_0 < \frac{1}{4}$. Determine t_0 by $f(2t_0\mu) = x_0$, and $\gamma \in [1/2, 1)$ by

$$\gamma = \frac{x_{bc} - x_{ab}}{2x_{bc} - x_{ab} - x_{ac}}, \text{ so } 1 - \gamma = \frac{x_{bc} - x_{ac}}{2x_{bc} - x_{ab} - x_{ac}}.$$

Then choose t_1 by $f(2t_1\mu) = x_{bc}$.

To verify that these parameter choices give the desired normalized triple of distances, the expected distance between a, b is

$$\begin{aligned} d_{LD}(a,b) &= f^{-1}(\gamma f(2t_0\mu) + (1-\gamma) f(2t_1\mu)) + \beta \\ &= f^{-1}(\gamma x_0 + (1-\gamma) x_{bc}) + \beta \\ &= f^{-1}(x_{ab}) + \beta \\ &= m\ell_{ab}. \end{aligned}$$

Similarly, we see $d_{LD}(a, c) = m\ell_{ac}$. Finally we have

$$d_{LD}(b,c) = f^{-1}(f(2t_1\mu)) + \beta = f^{-1}(x_{bc}) + \beta = m\ell_{bc}.$$

Note that even if $\ell_{ac} = \ell_{bc}$, the argument of Proposition 8 can be modified slightly by taking $\gamma = 1$ in the analytic continuation of the parameterization. However, that choice of the hybridization parameter essentially means that in place of a 4-cycle network parameter we have a tree.

Finally, we consider a network with a 3_2-cycle. While Proposition 4 shows the normalized triples of theoretical logDet distances lie on the same line as those for a tree, we establish they need not be restricted to the same line segment of tree-like distances. However, we do not completely characterize the extent of the segment they fill out.

Proposition 9 *Let $\ell = (\ell_{ab}, \ell_{ac}, \ell_{bc})$ with $\ell_{ac} = \ell_{bc}$ be a triple of positive numbers summing to 1 with $0 < \ell_{ab} < \frac{1}{2}$. Then there exists a binary ultrametric rooted network on taxa $\{a, b, c\}$ with a single 3_2-cycle whose leaf-descendants are a, b and GTR substitution model parameters such that the normalized theoretical LogDet distances of sequences generated under the coalescent mixture model are ℓ*

Proof We construct several 3_2-cycle species networks of the form shown in Figure 12, with edge lengths $t_i = \ell(e_i)$. In making choices of numerical parameters, since the network is ultrametric we view t_1, t_3, t_5, t_7 as independent, determining t_2, t_4, t_6. The population size on edge e_i for $3 \leq i \leq 8$ are constants N_i, with the sizes on terminal edges irrelevant. The hybridization parameters are $1 - \gamma$ and γ on edges e_4 and e_5 respectively. We also fix a single Jukes-Cantor substitution process with any constant rate $\mu > 0$.

By Proposition 4, for any choices of the t_i, N_i, γ, the theoretical LogDet distances will satisfy $d_{LD}(a, c) = d_{LD}(b, c)$ so the normalized theoretical LogDet distance triple lies on a line. Since the parameter space is connected, it is enough to show that

$$\frac{d_{LD}(a,b)}{2d_{LD}(a,c)+d_{LD}(a,b)} \tag{6}$$

is arbitrarily close to 0 for some choice of the parameters, and arbitrarily close to 1/2 for others, to conclude that the rescaled expected distances give all the described triples.

To make expression (6) near 0, we choose parameters with t_1 and N_3 sufficiently small so that with high probability the a, b lineages coalesce quickly. Specifically, let $t_3 = 1$, and fix any positive values for t_5, t_7 and N_i for $i \neq 3$. Now for any $\epsilon > 0$, as $N_3 \to 0^+$, the probability of lineages from a, b coalescing on e_3 within ϵ of entering it approaches 1. Using this, it is straightforward to show that as $N_3 \to 0^+$ the expected pattern frequency array for a, b approaches that for the JC model on a 2-taxon tree of total length $2t_1$. This then implies that $d_{LD}(a, b) \to 2\mu t_1$ as $N_3 \to 0^+$. On the other hand, for all values of $N_3 > 0$ one can show $d_{LD}(a, c) > 2\mu(t_1 + 2)$. Thus for a sufficiently small choices of t_1 and N_3, we can make $d_{LD}(a, b)/(2d(a, c) + d(a, b))$ as close to 0 as desired.

To produce a value of expression (6) near 1/2 is more subtle. We choose parameters so that a, b lineages are likely to enter e_5, but if they both do they are then unlikely to coalesce in it, and coalescence of any pair of lineages in e_7 is likely to occur quickly. First set $t_5 = 0$, $t_7 = 1$ and N_8 arbitrary. For any t_1, t_3 and γ, by choosing $N_3 = N_4 = N_5$ sufficiently large, the probability that the a, b lineages coalesce on e_3, e_4, or e_5 can be made arbitrarily small, so that if they coalesce below the root with (conditional) probability approaching 1 they must do so on e_7. This requires that both the a, b lineages follow e_5, which occurs with probability γ^2. If lineages a, c coalesce below the root, they must do so on e_7, requiring the a lineage to follow e_5, which occurs with probability γ. By picking N_7 sufficiently small, the probability that two lineages in edge e_7 coalesce near the lower end can be made close to 1. All this shows that once t_1, t_3 and γ are chosen, by appropriate choices of the N_i we can ensure the expected frequency arrays for a, b and a, c are arbitrarily close to

$$\gamma^2 F(t_1 + t_3) + (1 - \gamma^2)G(t_1 + t_3 + 1, N_8)$$

and

$$\gamma F(t_1 + t_3) + (1 - \gamma)G(t_1 + t_3 + 1, N_8),$$

respectively, where $F(t)$ is the expected pattern frequency array for two samples at distance $2t$ and $G(t, N)$ is the expected array under the coalescent for 2 lineages which enter a common population of size N at time t. Further picking sufficiently small values for t_1, t_3, the pattern frequency arrays for a, b and a, c can be made arbitrarily close to

$$\gamma^2 \frac{1}{4} I + (1 - \gamma^2)G(1, N_8)$$

and

$$\gamma \frac{1}{4} I + (1-\gamma) G(1, N_8),$$

respectively. Thus for any γ the theoretical distance can be made arbitrarily close to the distance computed from the above arrays. Using the formulas defined in the proof of Proposition 8, we find these distances are

$$d_{LD}(a,b) = f^{-1}((1-\gamma^2)\delta)$$

and

$$d_{LD}(a,c) = f^{-1}((1-\gamma)\delta)$$

where $\delta > 0$ is the off-diagonal entry of $G(1, N_8)$. Thus once γ is specified, by choosing t_1, t_3, $N_3 = N_4 = N_5$, N_7 we can ensure expression (6) is arbitrarily close to

$$\frac{\log(1-4\delta(1-\gamma^2))}{2\log(1-4\delta(1-\gamma)) + \log(1-4\delta(1-\gamma^2))}. \tag{7}$$

Applying L'Hopital's rule shows the limit of expression (7) as $\gamma \to 1$ is $\frac{1}{2}$. Thus for any $\epsilon > 0$, by first choosing γ near 1 so that the expression (7) is within $\epsilon/2$ of 1/2, and then choosing $t_1 = t_3$, $N_3 = N_4 = N_5$, N_7 so that expression (6) is within $\epsilon/2$ of expression (7), we obtain the desired result.

The results of this section, combined with those of Section 5 are summarized by Figure 13, which indicates the various regions of the simplex which normalized logDet triples fill, according to whether the network has a 4-cycle, a 3_2-cycle, or neither.

Note that the possibility that a 3_2-cycle (as depicted in the center of Figure 13) leads to a triple of normalized logDet distances lying on an extension of the corresponding line segment for the tree topology displayed on the networks (as depicted to the right of the figure) echoes a number of similar results arising in studies of network inference under the coalescent from gene tree data. For unrooted quartets, these include the works of Solís-Lemus et al. [2016], Baños [2019] and Allman et al. [2019a], and for rooted triples Long and Kubatko [2018] and Jiao and Yang [2020]. In essence, all these results indicate that the coalescent can lead to anomalous gene trees, in the sense that the most frequent gene tree topology may not match that of the trees displayed on the species network, even though all such displayed trees have the same topology.

8 Conclusion

Theorem 1 states that most topological features of an ultrametric level-1 network can be identified from theoretical logDet distances under a fairly general model of sequence

evolution with incomplete lineage sorting. It more generally implies network identifiability from pattern frequency arrays, since logDet distances are functions of these. In particular, individual gene trees, or even sequences partitioned into genes, are not required for network identifiability.

While identifiability is a theoretical question about the model, it has important implications for data analysis. Indeed, it is a key requirement for a statistically consistent inference procedure to exist. While our method of proof of identifiability, using the logDet distance, suggests using that distance as a basis for an inference procedure, others might be developed as well.

In subsequent work, we will explore using the logDet distance in a procedure for level-1 network inference following the framework of NANUQ [Allman et al., 2019a]. In outline, for each triple of taxa, the location of the normalized triple of logDet distances in simplex plots such as those of Figure 13 can indicate whether the rooted triple has a 4-cycle or not. A triple near the lines through the centroid can, through some statistical test, be judged unlikely to have arisen from a 4-cycle, while those farther away are judged to have arisen from a 4-cycle. Then, modifying the rooted triple distance of Rhodes [2019] to a network setting, similarly to how NANUQ modified the quartet distance, an intertaxon distance can be computed from the results of these statistical tests. Rules for relating a splits graph for the expected rooted triple distance to the original network will be developed. When applied to the splits graph constructed by NeighborNet from the empirically-derived distance, this should lead to consistent network inference. Since individual gene trees are never inferred, this will potentially give a much faster data analysis pipeline than the current version of NANUQ, which is built on quartet concordance factors across gene trees.

Acknowledgements

This work was supported by the National Institutes of Health [R01 GM117590], under the Joint DMS/NIGMS Initiative to Support Research at the Interface of the Biological Mathematical Sciences, and [2P20GM103395], an NIGMS Institutional Development Award (IDeA), and by the National Science Foundation [2051760]. H.B. was also partially supported by the Moore-Simons Project on the Origin of the Eukaryotic Cell, Simons Foundation grant 735923LPI (DOI: https://doi.org/10.46714/735923LPI) awarded to Andrew J. Roger and Edward Susko.

References

Allman ES, Baños H, and Rhodes JA. NANUQ: A method for inferring species networks from gene trees under the coalescent model. Algorithms Mol. Biol, 14(24):1–25, 2019a. [PubMed: 30839948]

Allman ES, Long C, and Rhodes JA. Species tree inference from genomic sequences using the log-det distance. SIAM J. Appl. Algebra Geometry, 3:107–127, 2019b.

Allman ES, Mitchell JD, and Rhodes JA. Gene tree discord, simplex plots, and statistical tests under the coalescent. Syst. Biol, 2021. Advance article 10.1093/sysbio/syab008.

Baños H. Identifying species network features from gene tree quartets. Bulletin of Mathematical Biology, 81:494–534, 2019. [PubMed: 30094772]

Casanellas M and Fernández-Sánchez J. Rank conditions on phylogenetic networks. arXiv:2004.12988, to appear in Research Perspectives CRM Barcelona, Spring 2019, vol. 10, in Trends in Mathematics Springer-Birkhauser, 2020.

Chifman J and Kubatko L. 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. [PubMed: 25791286]

Dasarathy G, Nowak R, and Roch S. Data requirement for phylogenetic inference from multiple loci: A new distance method. IEEE/ACM Trans. Comput. Biol. and Bioinf, 12(2):422–432, 2015.

Gross E and Long C. Distinguishing phylogenetic networks. SIAM Journal on Applied Algebra and Geometry, 2(1):72–93, 2018. doi: 10.1137/17M1134238.

Gross E, van Iersel L, Janssen R, Jones M, Long C, and Murakami Y. Distinguishing level-1 phylogenetic networks on the basis of data generated by Markov processes. arXiv:2007.08782, 2020.

Hollering B and Sullivant S. Identifiability in phylogenetics using algebraic matroids. Journal of Symbolic Computation, 104:142–158, 2021. ISSN 0747-7171. doi: 10.1016/j.jsc.2020.04.012.

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

Huber KT, Moulton V, Semple C, and Wu T. Quarnet inference rules for level-1 networks. Bulletin of Mathematical Biology, 80(8):2137–2153, 2018. doi: 10.1007/s11538-018-0450-2. URL 10.1007/s11538-018-0450-2. [PubMed: 29869043]

Jansson J and Sung W-K. Inferring a level-1 phylogenetic network from a dense set of rooted triplets. Theoretical Computer Science, 363(1):60–68, 2006. ISSN 0304-3975. doi: 10.1016/j.tcs.2006.06.022. Computing and Combinatorics.

Jiao X and Yang Z. Defining species when there is gene flow. Syst. Biol, 70(1):108–119, 07 2020. ISSN 1063-5157. doi: 10.1093/sysbio/syaa052. URL 10.1093/sysbio/syaa052.

Liu L and Edwards SV. Phylogenetic analysis in the anomaly zone. Systematic Biology, 58(4):452–460, 2009. [PubMed: 20525599]

Lockhart PJ, Steel MA, Hendy MD, and Penny D. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol, 11:605–612, 1994. [PubMed: 19391266]

Long C and Kubatko L. The effect of gene flow on coalescent-based species-tree inference. Syst. Biol, 67(5):770–785, 03 2018. ISSN 1063-5157. doi: 10.1093/sysbio/syy020. URL 10.1093/sysbio/syy020. [PubMed: 29566212]

Meng C and Kubatko LS. Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: A model. Theoretical Population Biology, 75(1):35–45, 2009. ISSN 00405809. doi: 10.1016/j.tpb.2008.10.004. [PubMed: 19038278]

Mitchell JD, Allman ES, and Rhodes JA. Hypothesis testing near singularities and boundaries. Electron. J. Statist, 13(1):2150–2193, 2019.

Rhodes JA. Topological metrizations of trees, and new quartet methods of tree inference. IEEE/ACM Trans. Comput. Biol. Bioinform, early access, 2019. doi: 10.1109/TCBB.2019.2917204.

Rosselló F and Valiente G. All that glisters is not galled. Mathematical Biosciences, 221(1):54–59, 2009. ISSN 00255564. doi: 10.1016/j.mbs.2009.06.007. [PubMed: 19576908]

Semple C and Steel M. Phylogenetics. Oxford University Press, 2005. ISBN 0 19 850942 1.

Solís-Lemus C and Ané C. Inferring Phylogenetic Networks with Maximum Pseudolikelihood under Incomplete Lineage Sorting. PLoS Genetics, 12(3), 2016. ISSN 15537404. doi: 10.1371/journal.pgen.1005896.

Solís-Lemus C, Yang M, and Ané C. Inconsistency of species tree methods under gene flow. Syst. Biol, 65(5):843–851, 05 2016. ISSN 1063-5157. doi: 10.1093/sysbio/syw030. URL 10.1093/sysbio/syw030. [PubMed: 27151419]

Steel M. Phylogeny: Discrete and Random Processes in Evolution. SIAM, Philadelphia, 2016. ISBN 9781611974478.

Steel MA. Recovering a tree from the leaf colourations it generates under a Markov model. Applied Mathematics Letters, 7(2):19–24, 1994.

van der Vaart AW. Asymptotic Statistics. Cambridge University Press, 1998.

van Iersel L, Moulton V, and Murakami Y. Reconstructibility of unrooted level-k phylogenetic networks from distances. Advances in Applied Mathematics, 120:102075, 2020.

Wen D and Nakhleh L. Coestimating reticulate phylogenies and gene trees from multilocus sequence data. Systematic Biology, 67(3):439–457, 2018. [PubMed: 29088409]

Yu Y and Nakhleh L. A maximum pseudo-likelihood approach for phylogenetic networks. BMC Genomics, 16(Suppl 10):S10, 2015. ISSN 1471-2164. doi: 10.1186/1471-2164-16-S10-S10.

Yu Y, Than C, Degnan JH, and Nakhleh L. Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. Systematic Biology, 60(2):138–149, 2011. [PubMed: 21248369]

Zhang C, Ogilvie HA, Drummond AJ, and Stadler T. Bayesian inference of species networks from multilocus sequence data. Molecular Biology and Evolution, 35(2):504–517, 12 2017.

Zhu J, Yu Y, and Nakhleh L. In the light of deep coalescence: revisiting trees within networks. BMC Bioinformatics, 5:271–282, 2016.

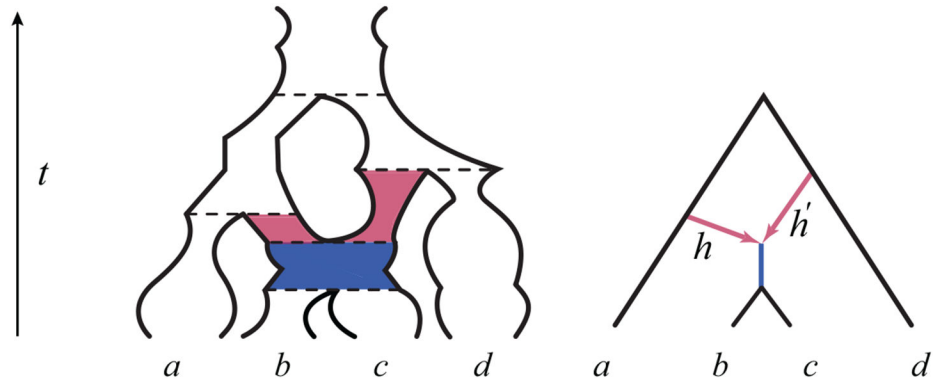

Fig. 1.

(Left) An ultrametric species network $\mathcal{N}^+$ with time t in generations before the present, hybrid edges h and h' shown in red, and population functions $N_e(t)$ on each edge depicted by widths of "tubes." The edge lengths τ are measured on the t-axis between the dashed lines indicating speciation and hybridization events. The dashed red/blue boundary represents a hybrid node, the top dashed line the root of the network, and other dashed lines tree nodes. (Right) A schematic of the same species tree, which does not show population sizes. Hybridization parameters γ and γ' are omitted from both drawings.

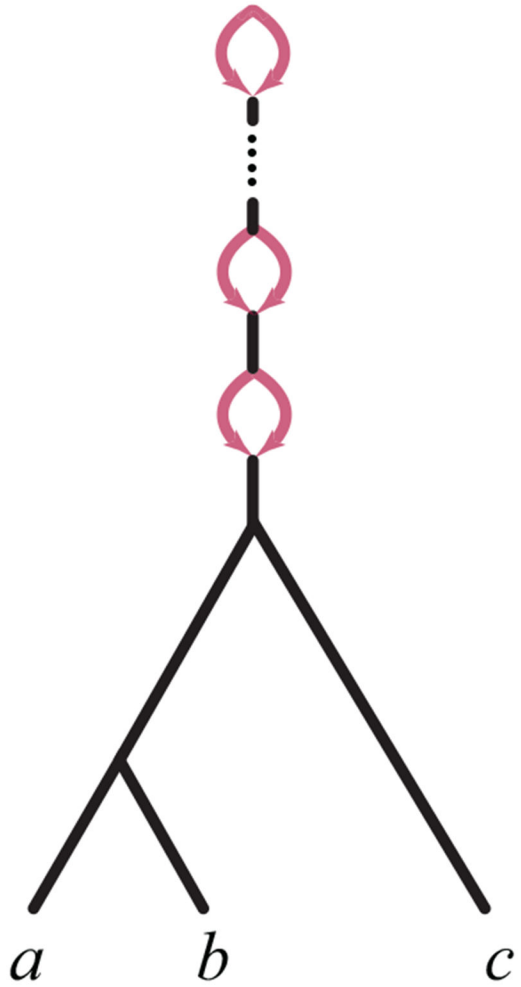

Fig. 2.

A rooted network $\mathcal{N}^{+}$ whose LSA network $\mathcal{N}^{\oplus}$ is the rooted tree ((*a, b*), *c*), but which has a chain of 2-cycles above LSA(*a, b, c*).

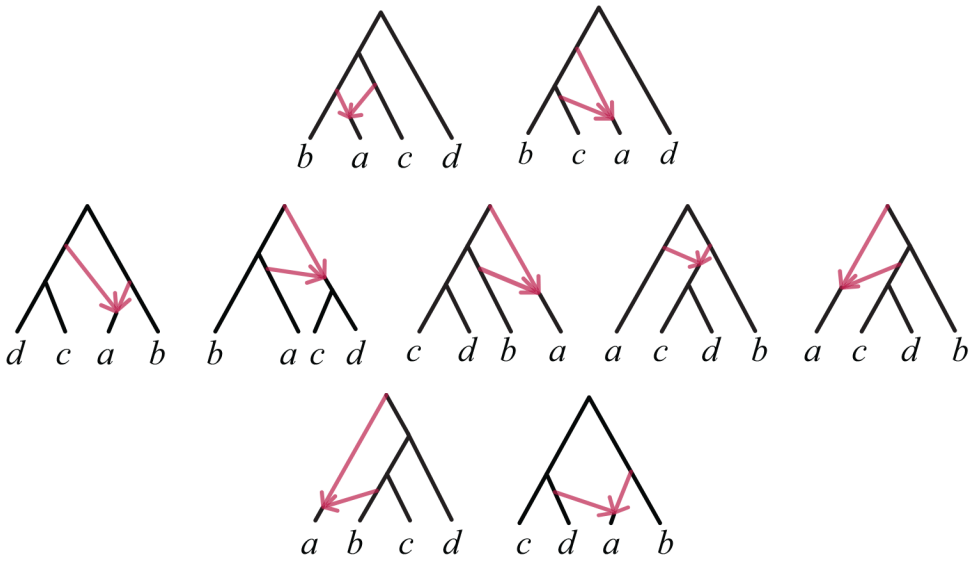

Fig. 3.

All rooted directed topological quartet networks with a single 4- or 5-cycle, and no other cycles, up to relabeling of taxa. Networks in the top row display exactly one rooted triple with a 4-cycle, those in the middle row display two, and those in the bottom row display three.

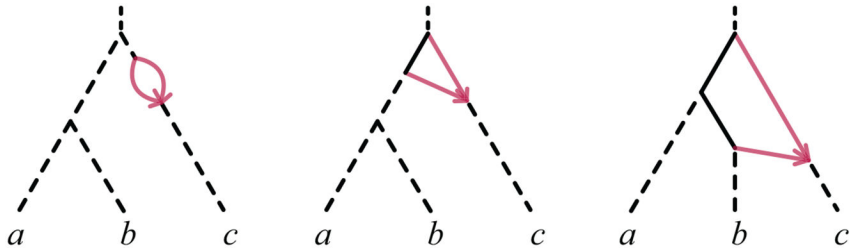

Fig. 4.

Examples of level-1 rooted triple networks with 2_1-, 3_1-, and 4_1-cycles. While multiple 2_1-cycles may be present along any pendant edge shown here in dashes, there can be at most two 3_1-cycles, whose hybrid nodes are located on a dashed pendant edge. At most one 4_1-cycle can be present. Site-pattern frequency matrices from the model $\mathscr{M}$ on rooted triple networks with these types of cycles are convex combinations of such matrices for 1, 2, or 4 networks without those cycles, as shown by Lemmas 4 and 5.

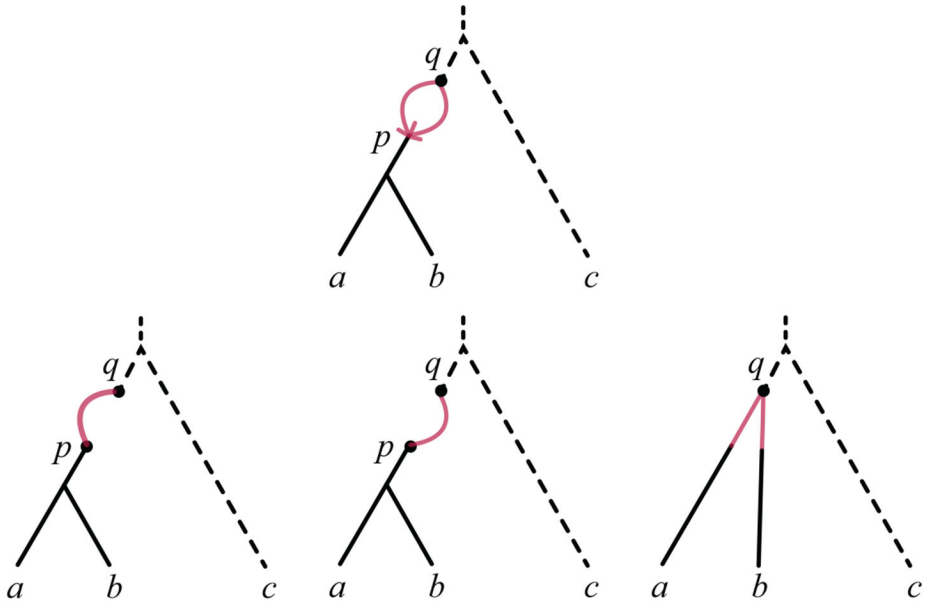

Fig. 5.

(Top) A rooted level-1 ultrametric network on $\{a, b, c\}$, with the 2_2-cycle closest to LSA(a, b) shown. (Bottom) The networks $\mathcal{N}_1^+$, $\mathcal{N}_2^+$, and $\mathcal{N}_0^+$ obtained from $\mathcal{N}^+$, respectively, as described in Lemma 6. Note that there may be additional cycles along the dashed lines, with hybrid nodes above node q and taxon c.

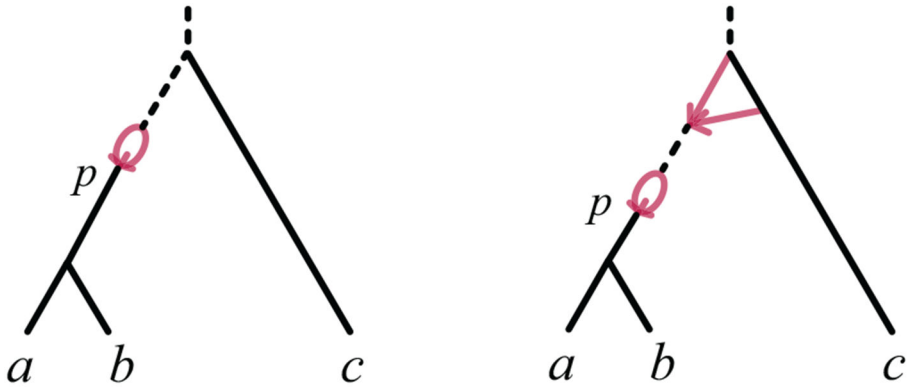

Fig. 6.

Networks $\mathcal{N}^+$ meeting the hypothesis of Lemma 7, with at least one 2_2- or 3_2-cycle, and possibly 2_3-cycles. In both figures the dashed internal edge represents a possible chain of 2_2-cycles, and the dashed edge above the LSA a possible chain of 2_3-cycles. Note that a network with a 3_2-cycle may also have no 2_2-cycles (not shown), in which case p would be the 3_2-cycle's hybrid node.

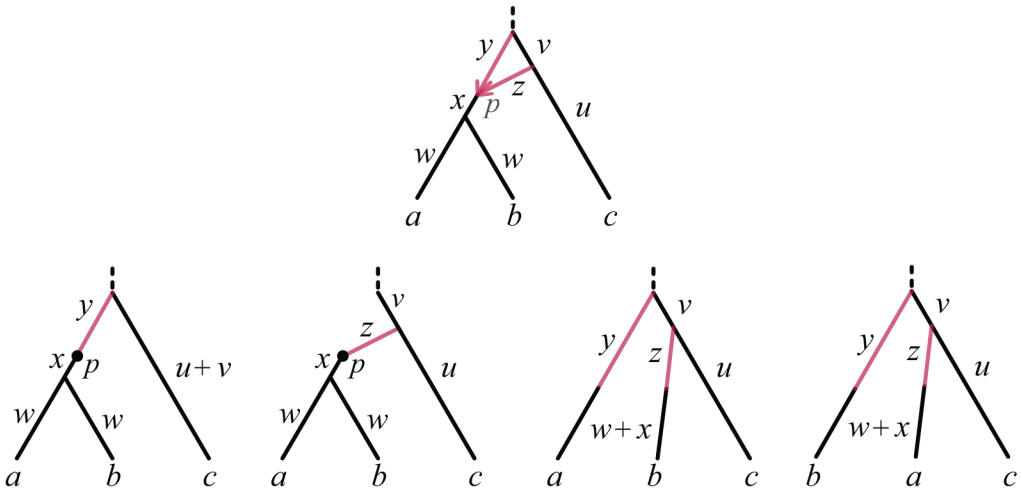

Fig. 7.

(Top) A rooted level-1 ultrametric network with a 3_2-cycle, and (Bottom) the networks $\mathcal{N}_1^+$, $\mathcal{N}_2^+$, $\mathcal{N}_3^+$, and $\mathcal{N}_4^+$ used in Lemma 8. Although only topology and branch lengths are shown, population size parameters for each edge of $\mathcal{N}_i^+$ are obtained from the corresponding ones of $\mathcal{N}^+$.

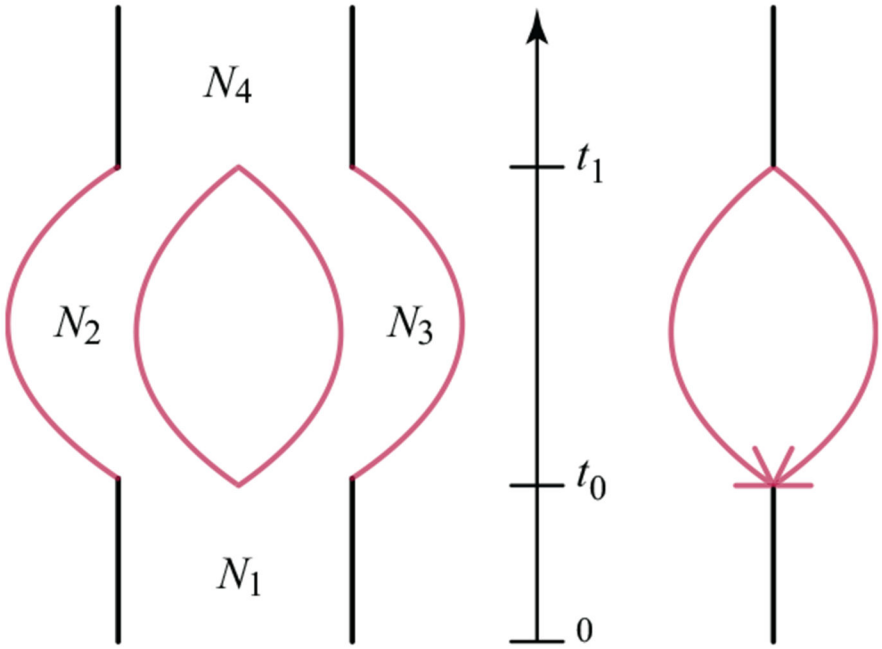

Fig. 8.

A 2-cycle and adjacent tree edges in a species network, depicted (Left) with pipes whose width represent population sizes, and (Right) as a schematic.

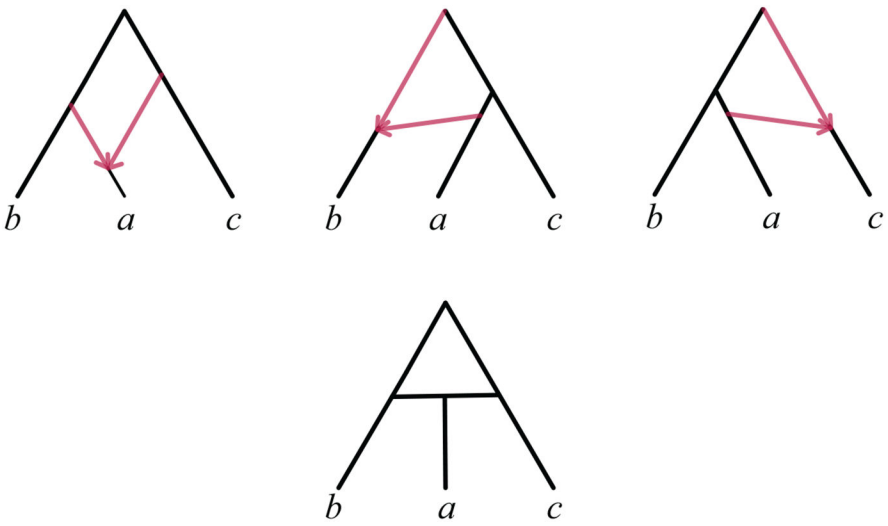

Fig. 9.

(Top) Three topologically-distinct rooted triple networks with a 4-cycle displaying the trees ((*a, b*), *c*) and ((*a, c*), *b*). (Bottom) The undirected rooted topology shared by them.

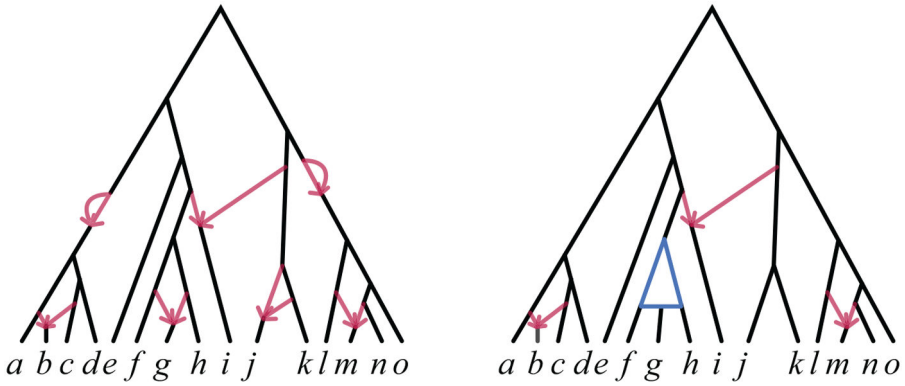

Fig. 10.

(Left) A rooted binary level-1 network and (Right) that part of its structure that Theorem 1 identifies from logDet distances under the model $\mathscr{M}$ for generic parameters. Both 2- and 3-cycles are lost, as are the directions of 4-cycle edges, and hence knowledge of the hybrid nodes in 4-cycles. Directed edges in cycles of size greater than 4 are identifiable.

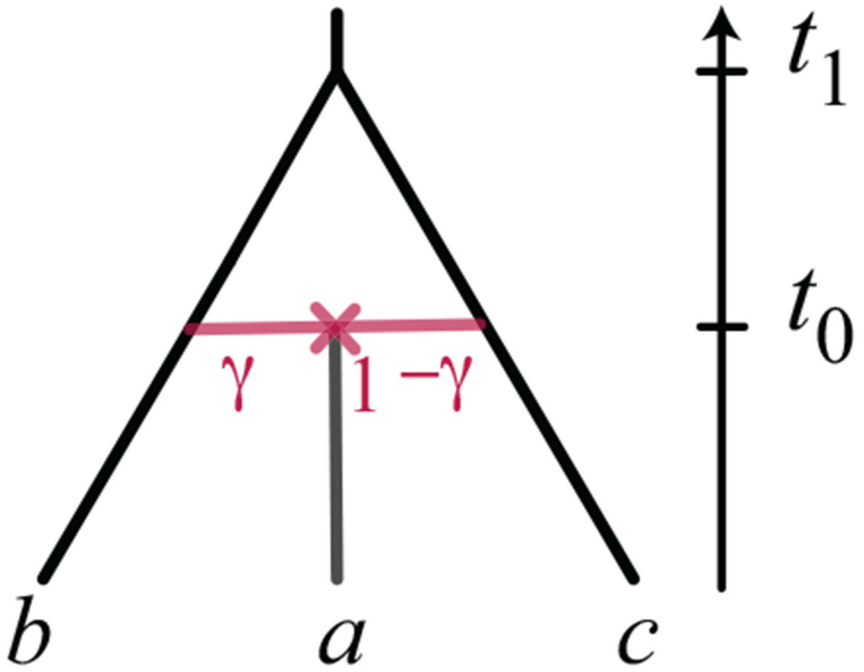

Fig. 11.

The 4-cycle network, with times in generations, constructed in Proposition 8. Hybridization parameters are γ, $1-\gamma$, and hybrid edges have length 0.

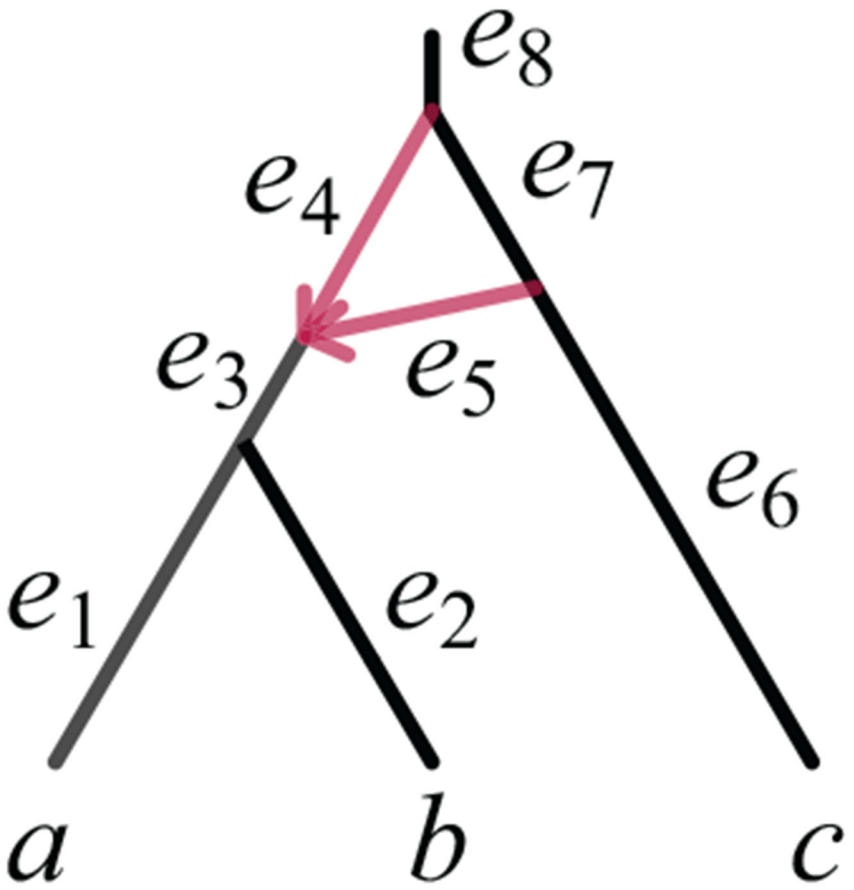

Fig. 12.

A 3_2-network, with numbered edges, as used in Proposition 9. The hybridization parameter on edge e_5 is γ, and on e_4 is $1 - \gamma$.

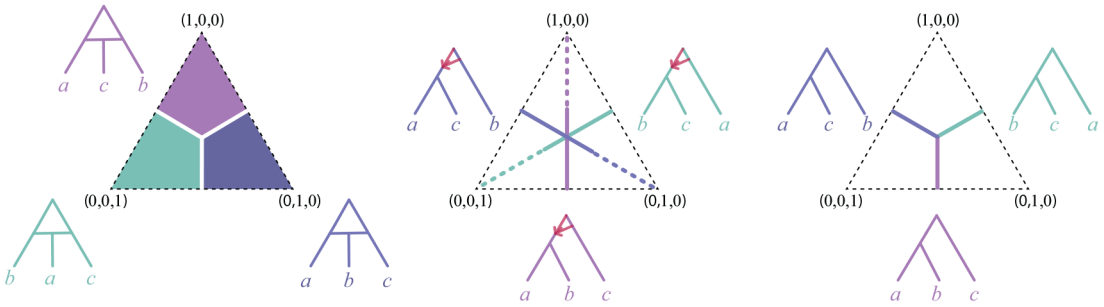

Fig. 13.

The regions of the simplex filled by normalized triples of logDet distances under the model $\mathscr{M}$ on a 3-taxon network. The networks shown are those obtained by suppressing all cycles other than 4- and 3_2-cycles, and then undirecting the 4-cycle edges. Normalized logDet distances are ordered as $(\ell_{ab}, \ell_{ac}, \ell_{bc})$. Networks with 3_2-cycles fill the solid line segments in the center simplex, but it is unknown whether they may also produce points in the dashed line segments.