Molecular Biology and Evolution, 2025, **42**, 1–7
https://doi.org/10.1093/molbev/msaf049
Advance access publication 7 March 2025
Brief Communications

MBE OXFORD

Misspecification Strikes: ASTRAL can Mislead in the Presence of Hybridization, even for Nonanomalous Scenarios

Vu Dinh [1,*] **Hector Baños** [2]

[1]Department of Mathematical Sciences, University of Delaware, Newark, DE 197111, USA
[2]Department of Mathematics, California State University, San Bernardino, CA 92407, USA

***Corresponding author**: E-mail: vucdinh@udel.edu.
Associate editor: Daniel Falush

Abstract

ASTRAL is a powerful and widely used tool for species tree inference, known for its computational speed and robustness under incomplete lineage sorting. The method has often been used as an initial step in species network inference to provide a backbone tree structure upon which hybridization events are later added to such a tree via other methods. However, we show empirically and theoretically, that this methodology can yield flawed results. Specifically, we demonstrate that under the network multispecies coalescent model—including nonanomalous scenarios—ASTRAL can produce a tree that does not correspond to any topology displayed by the true underlying network. This finding highlights the need for caution when using ASTRAL-based inferences in suspected hybridization cases.

Keywords: ASTRAL, model misspecification, phylogenetic networks, network multispecies coalescent

Introduction

Inferring species networks remains a challenging problem in phylogenetics. An initial difficulty in determining species relationships, even without hybridization, is incomplete lineage sorting (ILS), which leads to gene tree discordance. The complexity of gene tree discordance increases when hybridization or other lateral gene transfer events are considered. The network multispecies coalescent model (NMSC) (Meng and Salter Kubatko 2009) is a standard probabilistic model for describing gene tree formation within species networks in the presence of ILS and hybridization. Although a variety of coalescent-based network inference methods exist, most are either restricted to a rather simple family of networks (known as level-1) (Solís-Lemus and Ané 2016; Allman et al. 2019; Kong et al. 2024; Allman et al. 2024), or lack scalability (Yu and Nakhleh 2015; Zhang et al. 2017; Wen and Nakhleh 2018).

A reasonable approach for inferring species networks, avoiding either oversimplification of the network or scalability issues, is first to infer a *"displayed"* (or *"backbone"*) tree, representing underlying tree-like relationships among species, and then inferring hybridization events on top of it using different techniques, notably, through Dsuite (Malinsky et al. 2021). Due to the absence of an established method for inferring a displayed tree, many researchers have used ASTRAL (Mirarab et al. 2014), a leading method for inferring species trees. This network inference methodology has been used in many works, for example (Owens et al. 2023; Zhou et al. 2022; Singh et al. 2022; Jensen et al. 2023; Sanderson et al. 2023; Ciezarek et al. 2024; Scherz et al. 2022; Yang et al. 2023; Lopes et al. 2023; Feng et al. 2022; Bernhardt et al. 2020; DeRaad et al. 2022; Herrig et al. 2024; Zhang et al. 2023; Zhou et al. 2023), among others. Such a pipeline relies on the assumption that ASTRAL accurately infers a tree displayed in the network, as adding hybridization edges to a tree not displayed in the network cannot yield the true network.

In this work, we demonstrate that under the NMSC, such a pipeline can be erroneous, as ASTRAL may fail to produce a reliable displayed tree. This failure results from model misspecification (given ASTRAL does not account for hybridization) and not from a flaw in ASTRAL's approach to species tree inference. We demonstrate that even for relatively simple networks under the NMSC, such as one with a 5-cycle (see Fig. 1(d) for an example of such network), at least ∼6% of the parameter space can generate data that cause ASTRAL to infer a nondisplayed tree.

Challenges associated with using ASTRAL in the context of hybridization have been previously explored. Solís-Lemus et al. (2016) demonstrated that *anomalous networks* pose significant challenges for ASTRAL. Anomalous networks are such that a gene tree not displayed in the network occurs more frequently than one that is (Ané et al. 2024). Inferring anomalous networks presents challenges even for dedicated species network inference methods. Here, we show that ASTRAL's behavior under the NMSC applies even to networks that are far from anomalous. In a different context, Long and Kubatko (2018) showed that ASTRAL can be misleading in the setting of a three-taxon isolation-with-migration model (with a similar structure to an anomalous network). However, their analysis incorporates continuous gene flow, which is not addressed by the standard NMSC. Lastly, in Pang and Zhang (2022), the authors primarily focus on cases

Received: November 14, 2024. **Revised:** February 24, 2025. **Accepted:** February 25, 2025

where ASTRAL fails to recover the "major" tree, that is, the "predominant" tree structure within the network. However, the pipeline for network inference using ASTRAL, as described above, is not compromised in these instances, as the ASTRAL tree inferred may be displayed by the network. In such work, the authors describe these cases as "anomalous;" however, as mentioned before, here we adopt the term anomalous as defined in Ané et al. (2024).

Background

The NMSC

The NMSC is a probabilistic model describing the formation of gene trees, within a species network, in the presence of ILS and hybridization. In this model, gene lineages trace independently their ancestry backward in time through ancestral populations (the network edges). The probability that two lineages present in a population do not coalesce over time t, where t is measured in coalescent units, is given by e^{-t}. The coalescence rate depends on the number of lineages present, decreasing with each coalescent event. Lineages below a hybridization event trace their ancestry independently back to a particular population with probability γ, known as the hybridization parameter, which is specific to that population. This coalescent process continues indefinitely until all lineages have coalesced, resulting in the formation of a rooted gene tree.

Marginalizing over all taxa other than $\{A, B, C, D\}$, root location, and edge lengths, yields the probability that a gene tree displays each of the resolved topologies $AB \mid CD$, $AC \mid BD$, $AD \mid BC$. Quartet gene tree probabilities are pivotal in species tree and network inference; for instance, ASTRAL's statistical consistency relies on the property that, under the standard multispecies coalescent model (MSC), the most probable quartet gene tree shares the same topology as the quartet species tree (Allman et al. 2011).

As shown in Ané et al. (2024), for an arbitrary network, quartet gene tree probabilities are independent of root placement. Therefore, in this work, we generally depict networks without specifying the root, with only hybrid edges directed and nonhybrid edges left undirected, commonly known as *semidirected* networks.

For a network on n taxa with a single n-cycle—meaning a cycle with n edges when considered undirected—the distribution of gene trees under the NSMC model is a mixture of two gene tree distributions derived from the MSC (Rannala and Yang 2003). These distributions correspond to the displayed trees, which are obtained by removing one of the hybrid edges. See Fig. 1(a–f) for an example of a 4- and 5-cycle with their displayed trees.

As an illustrative example, and one that will be useful in a later section, the quartet probabilities for the 4-cycle network in Fig. 1(a) are given by:

$$P(AB \mid CD) = (1-\gamma)\left(1 - \frac{2}{3}e^{-x}\right) + \gamma\left(\frac{1}{3}e^{-y}\right), \quad (1)$$

$$P(AD \mid BC) = (1-\gamma)\left(\frac{1}{3}e^{-x}\right) + \gamma\left(1 - \frac{2}{3}e^{-y}\right), \quad (2)$$

$$P(AC \mid BD) = (1-\gamma)\left(\frac{1}{3}e^{-x}\right) + \gamma\left(\frac{1}{3}e^{-y}\right). \quad (3)$$

In this particular case, $P(AC \mid BD) < P(AB \mid CD), P(AD \mid BD)$, meaning that the two displayed trees are more frequent than the one not displayed by the network. Specifically,

$$P(AB \mid CD) > P(AD \mid BC) \text{ if and only if } (1-\gamma)(1-e^{-x}) > \gamma(1-e^{-y}). \quad (4)$$

One of the key insights underlying our main results involves examining the differences in the probability between the most frequent quartet tree and its runner-up tree. For this scenario, this difference is given by

$$P(AB \mid CD) - P(AD \mid BC) = (1-\gamma)(1-e^{-x}) - \gamma(1-e^{-y}). \quad (5)$$

Particularly, when $\gamma = 0$, i.e. when there is no hybridization

$$P(AB \mid CD) - P(AD \mid BC) = 1 - e^{-x}.$$

ASTRAL

ASTRAL is a quartet-based method for estimating species trees from multiple genes. The input of ASTRAL is a set of unrooted gene trees $\mathcal{T}_m = \{T_1, T_2, \ldots, T_m\}$, assumed to have arisen from the MSC on a tree. ASTRAL aims to find the (unrooted) species tree that agrees with the largest number of quartet trees induced by the set of gene trees. In principle, ASTRAL searches for a species tree $\mathbb{T}$ such that

$$\frac{1}{m}\sum_{q \in Q(\mathbb{T})} w_m(q, \mathcal{T}_m)$$

is maximized, where $Q(\mathbb{T})$ is the set of quartet trees induced by $\mathbb{T}$ and $w_m(q, \mathcal{T}_m)$ is the number of the trees in $\mathcal{T}_m$ that induce quartet topology q.

In ASTRAL, gene trees $\mathcal{T}_m = \{T_1, T_2, \ldots, T_m\}$ are assumed to be independently and identically distributed samples from the MSC on an unknown species tree $\mathbb{T}$. Under this assumption, in the limit of large numbers of gene trees, the proportion of the trees in $\mathcal{T}_m$ that induce a given quartet topology q converges to its quartet gene tree probabilities. Thus, if $\mathbb{T}'$ is the true species tree, there is

$$\frac{1}{m}\sum_{q \in Q(\mathbb{T}')} w_m(q, \mathcal{T}_m) \approx \sum_{q \in Q(\mathbb{T}')} w(q, \mathbb{T}'),$$

where $w(q, \mathbb{T}')$ is probability of observing the quartet topology q under the MSC model from a tree $\mathbb{T}'$. The quantity of the right-hand side, which we will refer to as the expected ASTRAL score of the tree $\mathbb{T}'$ and denote by $A(\mathbb{T}')$, can be used as a proxy to study the behavior of ASTRAL on a large collection of error-free gene trees.

ASTRAL Under Model Misspecification

In this section, we demonstrate theoretically and empirically that the ASTRAL tree inferred from data generated under the NMSC model for a 5-cycle network is not displayed by the network.

The ASTRAL Tree for Perfect Data under the NMSC Model

Consider the 5 taxon network N in Fig. 1(d). The network N has a 5-cycle, hybrid parameter γ, internal nonhybrid edge lengths e_2,e_3, e_4 in coalescent units, displayed tree $\mathbb{T}_1$ (Fig. 1(e)), and displayed tree $\mathbb{T}_2$ (Fig. 1(f)). Note that, regardless of the placement of the root, each leaf is represented by only a single sample from the associated population. As a result, no coalescent events can occur along these edges, meaning their lengths do not affect the quartet gene tree topology

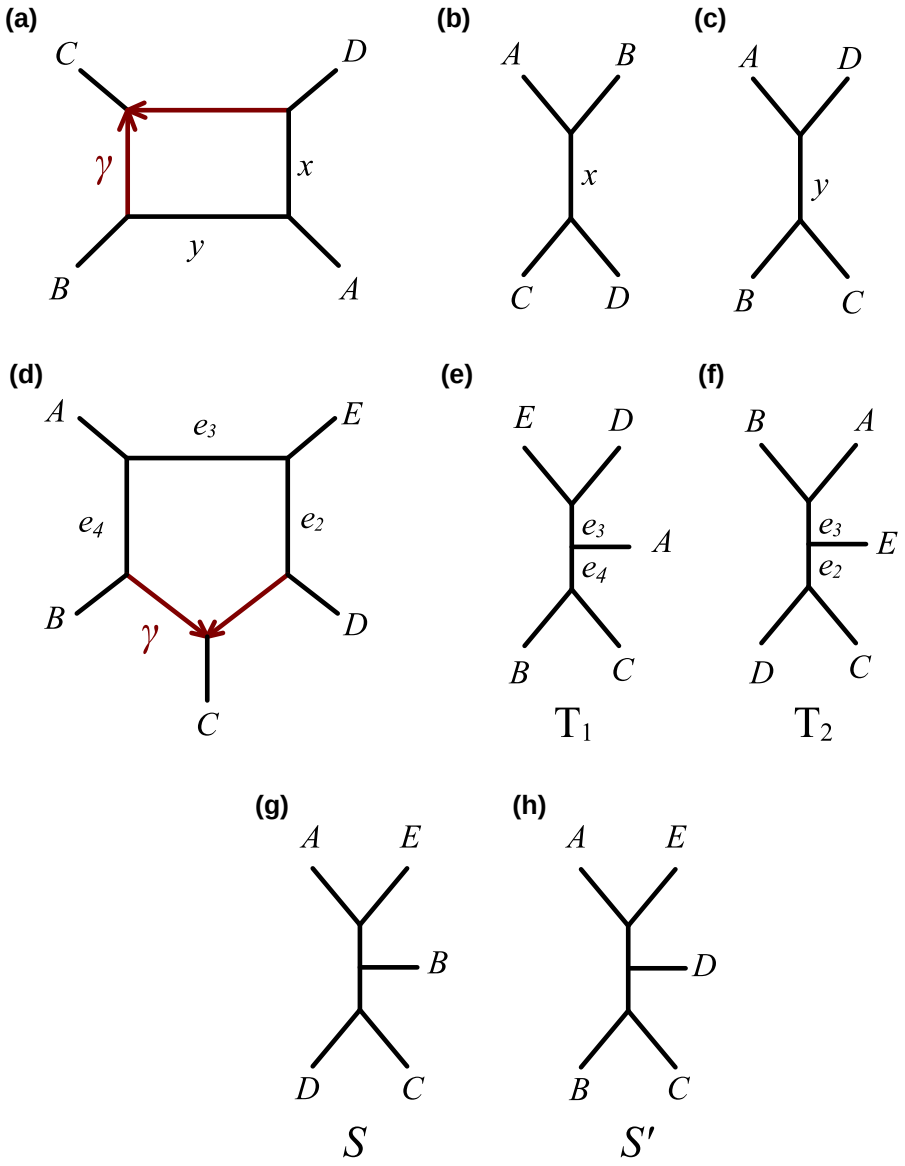

Fig. 1. a) A semidirected network with a 4-cycle, hybrid parameter γ, and edge lengths x and y in coalescent units. b) The displayed tree obtained by removing the edge with probability γ. c) The displayed tree obtained by removing the edge with probability $1-\gamma$. d) Similarly to above, a semidirected network N with a 5-cycle, hybrid parameter γ, and edge lengths e_2,e_3, e_4 in coalescent units. e) The displayed tree $\mathbb{T}_1$ obtained from removing the edge with probability $1-\gamma$. f) The displayed tree $\mathbb{T}_2$ obtained from removing the edge with probability γ. g, h) The trees S and S' from Theorem 1 and Corollary 2, respectively.

probabilities. For this network, there are five distinct 4-taxon sets. Note that, as shown in Baños (2019), 5-cycle networks are not *quartet-anomalous*, meaning no quartet gene tree not displayed in the network occurs more frequently than one that is. We now show the main result of this work.

Theorem 1 Let N be the semidirected network on 5 taxa with displayed trees $\mathbb{T}_1$ and $\mathbb{T}_2$, as depicted in Fig. 1(d–f). For any branch length of N, let $y_i = \exp(-e_i)$. If

$$(1-y_2) > \frac{\gamma}{1-\gamma}(1-y_3y_4) \tag{6}$$

and

$$(1-y_3) < \min\left(\frac{\gamma}{2-\gamma}(1-y_4),\ \frac{1-\gamma}{1+\gamma}(1-y_2)\right), \tag{7}$$

then for data generated from N under the NMSC, the tree $\mathcal{S} = (((A, E), B), C, D)$; (depicted in Fig. 1(g)), which is not displayed by N, has a higher expected ASTRAL score than both $\mathbb{T}_1$ and $\mathbb{T}_2$.

Proof. We begin by identifying, for each quartet network, the gene tree quartet with the highest and second-highest probability.

The subnetwork of N on A, B, C, D, is depiced in Fig. 1(a), where $x = e_2 + e_3$ and $y = e_4$. Thus, the probabilities of the three gene tree quartets on this set of taxa are given by Equations (1)–(3). Note that, for e_2,e_3,$e_4 > 0$, then $(1-\gamma)(1-y_2y_3) > (1-\gamma)(1-y_2)$ and $\gamma(1-y_3y_4) > \gamma(1-y_4)$. Together with Condition (6), this implies

$$(1-\gamma)(1-y_2y_3) > \gamma(1-y_4).$$

Thus, by Equation (4), $P(AB\,|\,CD) > P(AD\,|\,BC) > P(AC\,|\,BD)$.

Using a similar reasoning, we observe that $\gamma(1-y_4y_3) > \gamma(1-y_3)$, which together with Condition (6), gives $(1-\gamma)(1-y_2) > \gamma(1-y_3)$. Consequently, for the subnetwork on A, C, D, E, this yields $P(AE\,|\,CD) > P(AC\,|\,DE) > P(AD\,|\,CE)$.

Directly from Condition (6), we have $(1-\gamma)(1-y_2) > \gamma(1-y_3y_4)$, which, analogously to the previous cases, gives $P(BE\,|\,CD) > P(BC\,|\,DE) > P(BD\,|\,CE)$.

Following this approach, Condition (7) yields $\gamma(1-y_4) > (1-\gamma)(1-y_3)$, implying $P(AE|BC) > P(AB|CE) > P(AC|BE)$.

Finally, for the quartet on A, B, E, D, we note that the corresponding subnetwork is the tree $AB\,|\,DE$ with internal branch e_3, giving gene tree probabilities $P(AB\,|\,DE) > P(AE\,|\,BD) = P(AE\,|\,BD)$.

From this, we observe that the expected ASTRAL score of any tree cannot exceed

$$\alpha^* = P(AB\,|\,CD) + P(AE\,|\,CD) + P(BE\,|\,CD) + P(AE\,|\,BC) + P(AB\,|\,DE).$$

The expected ASTRAL score for $\mathbb{T}_1$, $\mathbb{T}_2$, and $\mathcal{S}$, which depends on their displayed quartets, is given by

$$\begin{aligned} A(\mathbb{T}_1) &= P(AD\,|\,BC) + P(AC\,|\,DE) + P(BC\,|\,DE) \\ &\quad + P(AE\,|\,BC) + P(AB\,|\,DE), \\ A(\mathbb{T}_2) &= P(AB\,|\,CD) + P(AE\,|\,CD) + P(BE\,|\,CD) \\ &\quad + P(AB\,|\,CE) + P(AB\,|\,DE), \\ A(\mathcal{S}) &= P(AB\,|\,CD) + P(AE\,|\,CD) + P(BE\,|\,CD) \\ &\quad + P(AE\,|\,BC) + P(AE\,|\,BD). \end{aligned}$$

The tree $\mathbb{T}_1$ has several nonoptimal gene tree quartets. Particularly one on $\{A, C, D, E\}$. Thus, by Equation (5),

$$\begin{aligned} \alpha^* - A(\mathbb{T}_1) &\geq P(AE\,|\,CD) - P(AC\,|\,DE) \\ &= (1-\gamma)(1-y_2) - \gamma(1-y_3). \end{aligned}$$

The tree $\mathbb{T}_2$ displays 4 of the optimal quartets, and only has a single nonoptimal topology, the one on $\{A, B, C, E\}$. Thus, by Equation (5),

$$\begin{aligned} \alpha^* - A(\mathbb{T}_2) &= P(AE\,|\,BC) - P(AB\,|\,CE) \\ &= \gamma(1-y_4) - (1-\gamma)(1-y_3). \end{aligned}$$

Note that $\mathcal{S}$ also displays 4 optimal quartets, and only gives a nonoptimal topology on $\{A, B, E, D\}$. Thus, by Equation (5),

$$\alpha^* - A(\mathcal{S}) = P(AB\,|\,DE) - P(AE\,|\,BD) = 1 - y_3$$

By Condition (7), we have

$$\begin{aligned} 1-y_3 &< \gamma(1-y_4) - (1-\gamma)(1-y_3) \quad \text{and} \\ 1-y_3 &< (1-\gamma)(1-y_2) - \gamma(1-y_3). \end{aligned}$$

Therefore, $A(\mathcal{S}) > A(\mathbb{T}_1), A(\mathbb{T}_2)$.

Analogously, by symmetry, the following corollary holds.

Corollary 2 Let N be the semidirected network as in Theorem 1. For any branch length of N, let $y_i = \exp(-e_i)$. If

$$(1-y_4) > \frac{1-\gamma}{\gamma}(1-y_3y_2)$$

and

$$(1-y_3) < \min\left(\frac{\gamma}{2-\gamma}(1-y_4), \frac{1-\gamma}{1+\gamma}(1-y_2)\right)$$

then for data generated from N under the NMSC, then the tree $\mathcal{S}' = (((A, E), D), C, B)$; (depicted in Fig. 1(h)), which is not displayed in N, has a higher expected ASTRAL score than both $\mathbb{T}_1$ and $\mathbb{T}_2$.

While these theoretical results correspond to 5-taxon networks with a single 5-cycle, they can be extended to any arbitrary n-taxon network that has a 5-cycle as a subnetwork. Furthermore, in the Supplementary materials, we generalize Theorem 1 to n-cycle networks, with $n \geq 5$, demonstrating that these results also extend to networks containing a k-cycle as a subnetwork, where $k \geq 5$.

Simulations

In the results above, the term data refers to a "perfect" sample under the NMSC. In this section, we present simulations that empirically validate our theoretical results. As a proof of concept, we first simulated a sample $\mathcal{T}_{100K}$ of 100,000 gene trees from the rooted network N', with extended Newick notation:

```
((((C:1)#H1:1::0.5, B:1)G:1.1, A:1)F:0.1,
((#H1:1::0.5, D:1)J:2, E:1)K:0.1)r;
```

using `PhyloCoalSimulations` (Fogg et al. 2023) with default parameters. When unrooting N', the resulting network is that in Fig. 1(d), where $e_2 = 2$, $e_3 = 0.2$, $e_4 = 1.1$, $\gamma = 0.5$, with all other edges set to 1, thus satisfying the conditions of Theorem 1.

Using default parameters, we then ran ASTRAL-III (Zhang et al. 2018) v.5.7.8 on $\mathcal{T}_{100K}$. As expected, the resulting ASTRAL tree is $\mathcal{S}$. We primarily used ASTRAL-III in this work, as it has been the most widely adopted tool for the network inference pipeline described in the introduction. However, for completeness, we also input $\mathcal{T}_{100K}$ through ASTRAL-IV v1.19.4.6 (Zhang and Mirarab 2022), and ASTRAL-PRO v1.19.3.6 (Zhang et al. 2020), with default parameters. For all methods, the tree recovered is $\mathcal{S}$. In $\mathcal{T}_{100K}$, the two most frequent unrooted gene tree topologies agree with the displayed trees $\mathbb{T}_1$ and $\mathbb{T}_2$ with frequencies ~0.174 and ~0.214, respectively. The third and fourth most frequent topologies are $((B, E), A), C, D)$; and $\mathcal{S}$, with frequencies ~0.131 and ~0.130, respectively, indicating that N' is not anomalous (nor quartet-anomalous as mentioned before).

To estimate the size of the parameter space of a 5-cycle network N whose data would yield an ASTRAL tree not displayed by N, we uniformly sampled 10^6 points from $[0, 1]^4$. The first three entries correspond to transformed branch lengths $\log(e_2)$, $\log(e_3)$, $\log(e_4)$, corresponding to sampling edge lengths from an exponential distribution with mean 1, and the remaining one to the hybridization parameter γ. We denote by Θ the set of parameters that satisfy either condition of Theorem 1 or Corollary 2. We found that ~0.06 of the samples are in Θ.

Figure 2 illustrates the wide range of parameters in Θ. Each histogram in the figure represents the distribution of a specific parameter, derived from the points in Θ from the 10^6 sampled parameters. To demonstrate that these parameters are indeed empirically problematic for ASTRAL, we simulated 1,000 gene trees for 100 randomly sampled parameter sets in Θ. We then ran ASTRAL-III with default parameters and found that in 86 out of the 100 trials, the ASTRAL tree was not displayed in the network. In the 14 cases where the tree was displayed, we simulated 50,000 gene trees. In all 14 instances,

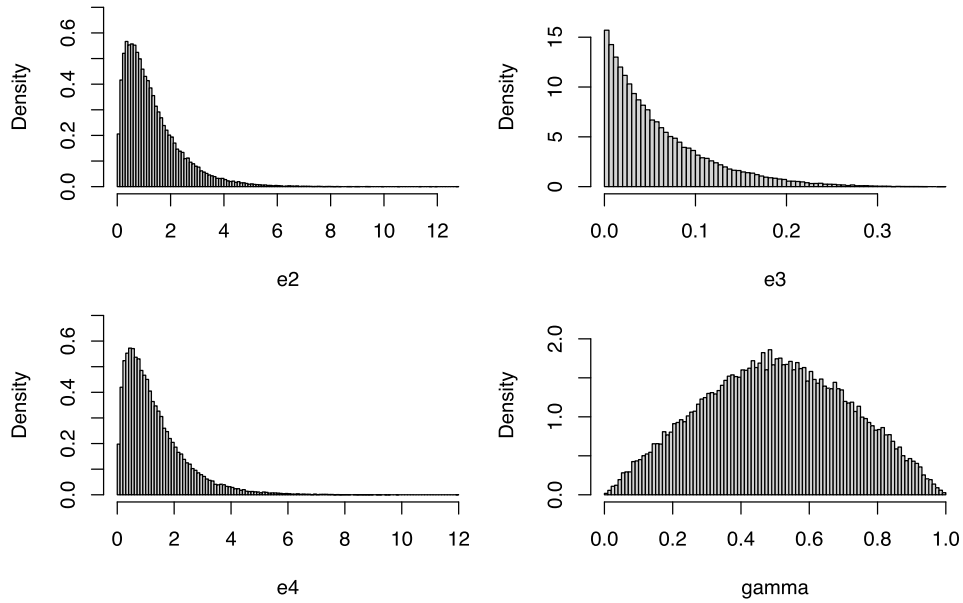

Fig. 2. Histograms illustrating the range of parameters in Θ, i.e. those satisfying the hypothesis of Theorem 1 or Corollary 2. In each histogram, the *y*-axis represents the density, while the *x*-axis corresponds to a specific parameter. The histograms are organized from left to right and top to bottom, corresponding to e_2 (with mean value ∼1.3), e_3 (mean ∼0.06), e_4 (mean ∼1.3), and γ (mean ∼0.5). Note that all edge lengths (e_2, e_3, and e_4) are expressed in coalescent units.

Table 1. The proportion of parameters in Θ (i.e. satisfying the conditions of Theorem 1 or Corollary 2).

Range for γ	**(0,1)**	**(0.2,0.8)**	**(0.4,0.6)**	**(0,0.1)∪(0.9,1)**
Proportion of parameters in Θ	0.06	0.08	0.10	0.01

Each proportion is calculated from values of γ within Θ, based on 10^6 parameter samples. In each sample, γ is uniformly selected within the specified interval, while e_2, e_3, and e_4 are sample edges from an exponential distribution with mean 1.

ASTRAL failed to recover a displayed tree in the larger data sets, suggesting the initial successes were likely due to finite sample error.

Note that the average values for each parameter (as shown in the figure captions) represent biologically reasonable estimates. For edge lengths, there are some cases where the parameters may be unrealistic, for example, edge lengths greater than 5 coalescent units. Nonetheless, the majority of edge lengths fall below this threshold. For the edge e_3, there are many instances where the edge length is very short, this may be the result of rapid divergences or populations with large effective sizes. Regarding the hybridization parameter, the results suggest that most issues arise from events with a strong hybridization signal (γ close to 0.5). However, even in cases with minimal hybridization (e.g. when γ is close to 0 or 1), erroneous inference can still occur.

To further investigate the parameter space, we restricted γ to different hybridization levels. Table 1 displays the proportion of parameters in Θ across different γ ranges. As γ approaches 0.5 (indicative of a strong hybridization signal) more edge length sets are satisfying the conditions of our results. For example, Table 1 shows that when γ is between 0.4 and 0.6, approximately, 10% of the parameter space is in Θ. Finally, we showed that even when γ values are close to 0 or 1, indicating a low hybridization signal, parameters still exist that can lead to erroneous inferences by ASTRAL (last column of Table 1).

Discussion

This study highlights the downsides of species network inference using a pipeline in which a tree is first inferred with ASTRAL. Our results indicate that for data generated under the NMSC, the inferred ASTRAL tree can differ from the one displayed in the network. We emphasize that the issues with ASTRAL arise from model misspecification rather than limitations inherent to ASTRAL's theoretical framework. Nonetheless, our results show the need for a tool that can reliably infer a displayed tree from a network. In principle, it is possible to quantify the fit between data and a fixed network, for example, using tools such as those developed by Cai and Ané (2020). However, no existing method allows testing, whether a specific tree is displayed in a network. If there were a reliable method for inferring a displayed tree, the pipeline discussed would not, in principle, present fundamental issues. A methodology to infer a displayed tree was introduced in Pyron et al. (2024), where the authors proposed a heuristic approach to infer displayed trees. While promising, this methodology is not easily scalable and has theoretical limitations, highlighting a need for further development.

While here we showed that the ASTRAL tree can differ from a displayed tree by a single nearest neighbor interchange (NNI) move, additional simulations in the Supplementary Material, show that this discrepancy can become more pronounced in networks with larger cycles as well as with more hybridization events. We also anticipate similar issues occur in other quartet-based methods under model misspecification, impacting not only tree inference but also network inference methods that assume a fixed network level, particularly those using a pseudolikelihood approach (Solís-Lemus and Ané 2016; Kong et al. 2024). Addressing these limitations requires an alternative framework, which we will explore in future work.

Supplementary Material

Supplementary material is available at *Molecular Biology and Evolution* online.

Acknowledgments

The authors would like to thank the Institute for Computational and Experimental Research in Mathematics (ICERM) for the resources during the "Theory, Methods, and Applications of Quantitative Phylogenomics" semester program. The authors would also like to thank John A. Rhodes for all the insightful discussions and suggestions. Both authors contributed equally to this work, and the author's order was decided uniformly at random.

Funding

H.B. was supported by the National Science Foundation (NSF) grant DMS-2331660, and V.D. by a startup fund from the University of Delaware, a University of Delaware Research Foundation's Strategic Initiatives Grant, and NSF grant DMS-1951474. This research was conceived and performed while the authors were in residence at ICERM in Rhode Island, which is supported by NSF grant DMS-1929284.

Conflict of Interests

None declared.

Data Availability

The data underlying this article were entirely simulated and can be easily replicated using the methods described in the manuscript.

References

Allman ES, Baños H, Rhodes JA. NANUQ: a method for inferring species networks from gene trees under the coalescent model. *Algorithms Mol Biol*. 2019:14(1):24. https://doi.org/10.1186/s13015-019-0159-2.

Allman ES, Baños H, Rhodes JA, Wicke K. NANUQ+: a divide-and-conquer approach to network estimation. bioRxiv. https://doi.org/10.1101/2024.10.30.621146. 2024, preprint: not peer reviewed. https://www.biorxiv.org/content/early/2024/11/03/2024.10.30.621146.

Allman ES, Degnan JH, Rhodes JA. Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. *J Math Biol*. 2011:62(6):833–862. https://doi.org/10.1007/s00285-010-0355-7.

Ané C, Fogg J, Allman ES, Baños H, Rhodes JA. Anomalous networks under the multispecies coalescent: theory and prevalence. *J Math Biol*. 2024:88(3):29. https://doi.org/10.1007/s00285-024-02050-7.

Baños H. Identifying species network features from gene tree quartets. *Bull Math Biol*. 2019:81(2):494–534. https://doi.org/10.1007/s11538-018-0485-4.

Bernhardt N, Brassac J, Dong X, Willing E-M, Poskar CH, Kilian B, Blattner FR. Genome-wide sequence information reveals recurrent hybridization among diploid wheat wild relatives. *Plant J*. 2020:102(3):493–506. https://doi.org/10.1111/tpj.v102.3.

Cai R, Ané C. Assessing the fit of the multi-species network coalescent to multi-locus data. *Bioinformatics*. 2020:37(5):634–641. https://doi.org/10.1093/bioinformatics/btaa863.

Ciezarek AG, Mehta TK, Man A, Ford AGP, Kavembe GD, Kasozi N, Ngatunga BP, Shechonge AH, Tamatamah R, Nyingi DW, *et al.* Ancient and recent hybridization in the Oreochromis cichlid fishes. *Mol Biol Evol*. 2024:41(7):msae116. https://doi.org/10.1093/molbev/msae116.

DeRaad DA, McCormack JE, Chen N, Peterson AT, Moyle RG. Combining species delimitation, species trees, and tests for gene flow clarifies complex speciation in scrub-jays. *Syst Biol*. 2022:71(6):1453–1470. https://doi.org/10.1093/sysbio/syac034.

Feng X, Merilä J, Löytynoja A. Complex population history affects admixture analyses in nine-spined sticklebacks. *Mol Ecol*. 2022:31(20):5386–5401. https://doi.org/10.1111/mec.v31.20.

Fogg J, Allman ES, Ané C. PhyloCoalSimulations: a simulator for network multispecies coalescent models, including a new extension for the inheritance of gene flow. *Syst Biol*. 2023:72(5):1171–1179. https://doi.org/10.1093/sysbio/syad030.

Herrig DK, Ridenbaugh RD, Vertacnik KL, Everson KM, Sim SB, Geib SM, Weisrock DW, Linnen CR. Whole genomes reveal evolutionary relationships and mechanisms underlying gene-tree discordance in neodiprion sawflies. *Syst Biol*. 2024:73(5):839–860. https://doi.org/10.1093/sysbio/syae036.

Jensen A, Swift F, de Vries D, Beck RMD, Kuderna LFK, Knauf S, Chuma IS, Keyyu JD, Kitchener AC, Farh K, *et al.* Complex evolutionary history with extensive ancestral gene flow in an African primate radiation. *Mol Biol Evol*. 2023:40(12):msad247. https://doi.org/10.1093/molbev/msad247.

Kong S, Swofford DL, Kubatko LS. Inference of phylogenetic networks from sequence data using composite likelihood. *Syst Biol*. 2024:74(1):53–69. https://doi.org/10.1093/sysbio/syae054.

Long C, Kubatko L. The effect of gene flow on coalescent-based species-tree inference. *Syst Biol*. 2018:67(5):770–785. https://doi.org/10.1093/sysbio/syy020.

Lopes F, Oliveira LR, Beux Y, Kessler A, Cárdenas-Alayza S, Majluf P, Páez-Rosas D, Chaves J, Crespo E, Brownell Jr RL, *et al.* Genomic evidence for homoploid hybrid speciation in a marine mammal apex predator. *Sci Adv*. 2023:9(18):eadf6601. https://doi.org/10.1126/sciadv.adf6601.

Malinsky M, Matschiner M, Svardal H. Dsuite - fast D-statistics and related admixture evidence from VCF files. *Mol Ecol Resour*. 2021:21(2):584–595. https://doi.org/10.1111/1755-0998.13265.

Meng C, Kubatko LS. Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. *Theor Popul Biol*. 2009:75(1):35–45. https://doi.org/10.1016/j.tpb.2008.10.004.

Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*. 2014:30(17):i541–i548. https://doi.org/10.1093/bioinformatics/btu462.

Owens GL, Huang K, Todesco M, Rieseberg LH. Re-evaluating homoploid reticulate evolution in helianthus sunflowers. *Mol Biol Evol*. 2023:40(2):msad013. https://doi.org/10.1093/molbev/msad013.

Pang X-X, Zhang D-Y. Impact of ghost introgression on coalescent-based species tree inference and estimation of divergence time. *Syst Biol*. 2022:72(1):35–49. https://doi.org/10.1093/sysbio/syac047.

Pyron RA, O'Connell KA, Myers EA, Beamer DA, Baños H. Complex hybridization in a clade of polytypic salamanders (Plethodontidae: Desmognathus) uncovered by estimating higher-level phylogenetic networks. *Syst Biol*. 2024:74(1):124–140. https://doi.org/10.1093/sysbio/syae060.

Rannala B, Yang Z. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*. 2003:164(4):1645–1656. https://doi.org/10.1093/genetics/164.4.1645.

Sanderson BJ, Gambhir D, Feng G, Hu N, Cronk QC, Percy DM, Freaner FM, Johnson MG, Smart LB, Keefover-Ring K, *et al.* Phylogenomics reveals patterns of ancient hybridization and differential diversification that contribute to phylogenetic conflict in willows, poplars, and close relatives. *Syst Biol*. 2023:72(6):1220–1232. https://doi.org/10.1093/sysbio/syad042.

Scherz MD, Masonick P, Meyer A, Hulsey CD. Between a rock and a hard polytomy: phylogenomics of the rock-dwelling mbuna cichlids of Lake Malaŵi. *Syst Biol*. 2022:71(3):741–757. https://doi.org/10.1093/sysbio/syac006.

Singh P, Irisarri I, Torres-Dowdall J, Thallinger GG, Svardal H, Lemmon EM, Lemmon AR, Koblmüller S, Meyer A, Sturmbauer C. Phylogenomics of trophically diverse cichlids disentangles processes driving adaptive radiation and repeated trophic transitions. *Ecol Evol*. 2022:12(7):e9077. https://doi.org/10.1002/ece3.v12.7.

Solís-Lemus C, Ané C. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS Genet*. 2016:12(3):e1005896. https://doi.org/10.1371/journal.pgen.1005896.

Solís-Lemus C, Yang M, Ané C. Inconsistency of species tree methods under gene flow. *Syst Biol*. 2016:65(5):843–851. https://doi.org/10.1093/sysbio/syw030.

Wen D, Nakhleh L. Coestimating reticulate phylogenies and gene trees from multilocus sequence data. *Syst Biol*. 2018:67(3):439–457. https://doi.org/10.1093/sysbio/syx085.

Yang L-H, Shi X-Z, Wen F, Kang M. Phylogenomics reveals widespread hybridization and polyploidization in Henckelia (Gesneriaceae). *Ann Bot*. 2023:131(6):953–966. https://doi.org/10.1093/aob/mcad047.

Yu Y, Nakhleh L. A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genomics*. 2015:16(1):1–10. https://doi.org/10.1186/1471-2164-16-1.

Zhang B-L, Chen W, Wang Z, Pang W, Luo M-T, Wang S, Shao Y, He W-Q, Deng Y, Zhou L, *et al*. Comparative genomics reveals the hybrid origin of a macaque group. *Sci Adv*. 2023:9(22):eadd3580. https://doi.org/10.1126/sciadv.add3580.

Zhang C, Mirarab S. Weighting by gene tree uncertainty improves accuracy of quartet-based species trees. *Mol Biol Evol*. 2022:39(12):msac215. https://doi.org/10.1093/molbev/msac215.

Zhang C, Ogilvie HA, Drummond AJ, Stadler T. Bayesian inference of species networks from multilocus sequence data. *Mol Biol Evol*. 2017:35(2):504–517. https://doi.org/10.1093/molbev/msx307.

Zhang C, Rabiee M, Sayyari E, Mirarab S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*. 2018:19(S6):153. https://doi.org/10.1186/s12859-018-2129-y.

Zhang C, Scornavacca C, Molloy EK, Mirarab S. ASTRAL-Pro: quartet-based species-tree inference despite paralogy. *Mol Biol Evol*. 2020:37(11):3292–3307. https://doi.org/10.1093/molbev/msaa139.

Zhou B-F, Yuan S, Crowl AA, Liang Y-Y, Shi Y, Chen X-Y, An Q-Q, Kang M, Manos PS, Wang B. Phylogenomic analyses highlight innovation and introgression in the continental radiations of fagaceae across the northern hemisphere. *Nat Commun*. 2022:13(1):1320. https://doi.org/10.1038/s41467-022-28917-1.

Zhou W, Furey NM, Soisook P, Thong VD, Lim BK, Rossiter SJ, Mao X. Diversification and introgression in four chromosomal taxa of the Pearson's horseshoe bat (*Rhinolophus pearsoni*) group. *Mol Phylogenet Evol*. 2023:183:107784. https://doi.org/10.1016/j.ympev.2023.107784.