Fast and Accurate Species Trees from Weighted Internode Distances

Baqiao Liu
Department of Computer Science, University of Illinois Urbana-Champaign, IL, USA

Tandy Warnow[1]
Department of Computer Science, University of Illinois Urbana-Champaign, IL, USA

Abstract

Species tree estimation is a basic step in many biological research projects, but is complicated by the fact that gene trees can differ from the species tree due to processes such as incomplete lineage sorting (ILS), gene duplication and loss (GDL), and horizontal gene transfer (HGT), which can cause different regions within the genome to have different evolutionary histories (i.e., "gene tree heterogeneity"). One approach to estimating species trees in the presence of gene tree heterogeneity resulting from ILS operates by computing trees on each genomic region (i.e., computing "gene trees") and then using these gene trees to define a matrix of average internode distances, where the internode distance in a tree T between two species x and y is the number of nodes in T between the leaves corresponding to x and y. Given such a matrix, a tree can then be computed using methods such as neighbor joining. Methods such as ASTRID and NJst (which use this basic approach) are provably statistically consistent, very fast (low degree polynomial time) and have had high accuracy under many conditions that makes them competitive with other popular species tree estimation methods. In this study, inspired by the very recent work of weighted ASTRAL, we present weighted ASTRID, a variant of ASTRID that takes the branch uncertainty on the gene trees into account in the internode distance. Our experimental study evaluating weighted ASTRID shows improvements in accuracy compared to the original (unweighted) ASTRID while remaining fast. Moreover, weighted ASTRID shows competitive accuracy against weighted ASTRAL, the state of the art. Thus, this study provides a new and very fast method for species tree estimation that improves upon ASTRID and has comparable accuracy with the state of the art while remaining much faster. Weighted ASTRID is available at `https://github.com/RuneBlaze/internode`.

2012 ACM Subject Classification Applied computing → Molecular evolution

Keywords and phrases Species tree estimation, ASTRID, ASTRAL, multi-species coalescent, incomplete lineage sorting

Digital Object Identifier 10.4230/LIPIcs.WABI.2022.8

Related Version *Previous Version*: `https://www.biorxiv.org/content/10.1101/2022.05.24.493312v1`

Supplementary Material *Software (Source Code)*: `https://github.com/RuneBlaze/internode` archived at `swh:1:dir:e8b673193c6c6e4a1d11e821fbc9a8ff2e6b74b9`

Funding *Baqiao Liu*: This work was supported in part by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research under the Secure Biosystems Design Initiative and by the Laboratory Directed Research and Development (LDRD) program of Sandia National Laboratories, which is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.
Tandy Warnow: NSF grant 2006069

Acknowledgements The authors thank the members of the Warnow lab for insightful comments.

[1] corresponding author

22nd International Workshop on Algorithms in Bioinformatics (WABI 2022).
Editors: Christina Boucher and Sven Rahmann; Article No. 8; pp. 8:1–8:24
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

Species tree estimation is a common task in phylogenomics and is a prior step in many downstream analyses (e.g., estimating divergence, understanding adaptation). Despite the recent increase in the availability of genome-scale data, species tree estimation remains challenging due to gene tree heterogeneity, where gene trees (the evolutionary history of genes) differ from species trees [16]. Among common factors for gene tree heterogeneity, incomplete lineage sorting (ILS), a population-level process modeled statistically by the multi-species coalescent (MSC) [47, 24], is extremely common and well-studied.

A standard approach to species-tree reconstruction under the presence of ILS is to concatenate the alignments of the individual genes and running a maximum likelihood (ML) heuristic on the combined alignment. This simple approach, however, has been established to be statistically inconsistent under the MSC, and can even be positively misleading, converging to the wrong topology with probability 1 as the number of genes increases [43, 42]. Empirically, concatenation can also suffer from degraded accuracy under higher levels of ILS, and can be affected by scalability issues under large data [29, 34]. In response, many accurate ILS-aware methods have since been developed. Those that are most commonly used in practice fall into a class of so-called summary methods, where gene trees are first independently estimated from each genomic region; the inferred gene trees are then used as input to the summary method to "summarize" the input gene trees into a species tree.

In recent years, many accurate summary methods statistically consistent under the MSC have been developed, such as MP-EST [22], NJst [40], ASTRAL [29], ASTRID [49], FASTRAL [9], and wQFM [25]. Many of these methods are scalable to genomic-scale data, and under sufficient gene signal and ILS tend to be more accurate than concatenation [34]. Among these methods, ASTRAL is the most commonly used, compares favorably to other methods in accuracy, and has been successfully applied to many large scale data [33, 55]. However, it is well known that summary methods suffer under inaccurate gene trees [34, 52] as a result of summary methods not explicitly taking gene tree uncertainty into account. Under inaccurate gene trees, it might still be preferable to use concatenation even under substantial ILS [34]. Although co-estimating the gene trees and species trees such as using StarBeast [36] or directly inferring quartets from the alignment and then combining them using SVDquartets+PAUP* [6] circumvents this problem, neither can scale to large data.

This sensitivity of summary methods to gene tree error has motivated approaches preprocessing the gene trees to improve the quality of the signal. Although throwing out inaccurate gene trees generally does not help [34], statistical binning [31, 4] and contracting low-support branches [55] improved accuracy for ASTRAL on many conditions. Nonetheless, these approaches require setting arbitrary thresholds: statistical binning requires a threshold to determine which branches are trustworthy, and contracting low-support branches also requires such a threshold. Suboptimal parameter selection in either case can lead to little accuracy improvement, or even worse, degraded accuracy compared to simply running on the original input [4, 55]. Thus, pragmatically, accurately applying such methods faces the difficulty of parameter selection.

Very recently, Zhang and Mirarab introduced weighted ASTRAL [54]. By directly incorporating gene tree uncertainty into the ASTRAL optimization problem, weighted ASTRAL improved ASTRAL in accuracy under all of their tested conditions. Notably, under conditions where concatenation proved more accurate than unweighted ASTRAL, weighted ASTRAL achieved the largest improvement, shrinking substantially the long known gap [32, 33, 37] between summary methods and concatenation under low gene signal. More specifically, (unweighted) ASTRAL heuristically searches for a species tree that maximizes the amount of quartet trees (unrooted four-taxon tree) shared with the input gene trees. By using branch support and lengths to weigh the reliability of gene tree quartets, weighted ASTRAL instead heuristically maximizes the weighted agreement with respect to the input gene trees, effectively discounting the contribution of unreliable quartets. Weighted ASTRAL is threshold-free, was shown to be more accurate than running ASTRAL on contracted gene trees [54], and in fact might be the most accurate summary method under ILS that can scale to large datasets.

Here, inspired by weighted ASTRAL, we introduce weighted ASTRID, incorporating gene tree uncertainty into ASTRID. ASTRID, a fast and more accurate variant of NJst, is based on the internode distance, defined by ASTRID as the number of edges between two taxa in a gene tree. We explore variations of this internode distance where branch uncertainty is considered. Notably, ASTRID is shown to have competitive accuracy against ASTRAL [49, 34] while having a much faster running time [49, 9], both of which we hope to generalize to weighted ASTRID when compared against weighted ASTRAL, obtaining a fast alternative to a very accurate method.

The rest of the study is organized as follows. In Section 2 we describe weighted ASTRID and introduce the two ways of weighting the internode distance matrix before providing some theoretical running time bounds. In Section 3 we describe our experimental study, choosing parameters for weighted ASTRID and comparing it to other methods. In Section 4, we present our experimental results showing that support-weighted ASTRID is very fast, is more accurate than ASTRID, has comparable accuracy with weighted ASTRAL, and provides an alternative accurate species-tree inference method robust to low gene signal, whereas branch-length weighted ASTRID has mixed accuracy and is less accurate than support-weighted ASTRID. In Section 5 we summarize our main conclusions and discuss future work.

2 Materials and methods

2.1 Basic definitions

Let n denote the number of taxa and let k denote the number of genes assuming a set of gene trees. Given an unrooted phylogenetic tree T, we denote its leafset by $\mathcal{L}(T)$ and its edge-set $E(T)$. For each edge e in T, deleting e from T partitions the leaves into two sets defined by the two connected components separated by e. Let this bipartition be denoted by π_e for some $e \in E(T)$. We denote the set of bipartitions of T by $C(T) = \{\pi_e \,|\, e \in E(T)\}$, and $C(T)$ uniquely defines the (unrooted) topology of T. A bipartition π_e is said to be trivial if e is incident to a leaf, since in such case $\pi_e \in C(T)$ for any T on the same leafset. Two trees are said to be compatible if the union of their bipartitions can coexist in a single tree. The Robinson-Foulds distance (RF distance) [41] between two trees T and T' on the same leafset is the size of the symmetric difference between the bipartitions of T and T', i.e., $|C(T) \triangle C(T')|$. Given two binary trees T and T', we define the nRF (error) rate as their RF distance normalized by $2n - 6$ (the number of non-trivial bipartitions), obtaining a value that is between 0 and 1.

For taxa $u, v \in \mathcal{L}(T)$, let $P_T(u, v)$ denote the set of edges on the unique path connecting u and v in T. Given an estimated gene tree G, we assume that each internal branch e is associated with a branch support value $s(e)$, some measure of confidence that this branch is correctly estimated, where $s(e) \in [0, 1]$. Note that we say a branch is correctly estimated (in topology) if the bipartition associated with that edge is present in the true gene tree. We extend this notion of branch support also to pendant edges, where we simply define $s(e) = 1$ if e is incident to a leaf. Similarly, for a true or estimated gene tree G, let $l(e)$ denote the length of a branch, where for estimated gene trees is usually given in substitution units inferred by some maximum likelihood tree inference method.

2.2 Intertaxon-distance based summary method

Under the context of summary methods, given an input of (unrooted) gene trees, our task is to infer an unrooted species tree topology from these input gene trees. A class of summary methods first introduced by Liu and collaborators [23, 21] infers the species tree by first metricizing the gene trees by defining a specific intertaxon distance between leaf nodes on the gene trees, and then for each pair of taxa averages all such distances across the input gene trees. The final averaged distance, represented as a distance matrix, is then fed as input to a distance-based tree inference method, such as UPGMA [28] or neighbor-joining [44], which infers a tree topology that will be the output species tree of this summary method. Here we focus on ASTRID (a faster and more accurate variant of NJst) and specifically describe its most accurate setting under no missing data.

ASTRID metricizes the gene trees by the **internode distance** $d_G(u, v)$ where $u, v \in \mathcal{L}(G)$ is defined to simply be the number of edges in between u, v in G, i.e., $d_G(u, v) = |P_G(u, v)|$. Although this definition from ASTRID differs from the original internode distance from NJst [21], where it was defined as the number of nodes (instead of edges) in between two taxa, this change has essentially no impact for our purposes [2] and generalizes to our later modifications better. Given this measure of internode distance, and assuming that each pair of taxa appears together in some gene tree (no missing data), ASTRID proceeds as follows with the input set of gene trees $\mathcal{G}$:

1. Initialize an $n \times n$ matrix D (n the number of taxa).
2. For each pair of taxa u, v, set $D[u, v]$ to be the empirical mean of $d_G(u, v)$ where G ranges in the input gene trees $\mathcal{G}$ where both u, v are present, i.e., set $D[u, v] = \sum_{G \in \mathcal{G}_{u,v}} d_G(u, v)/|\mathcal{G}_{u,v}|$, where $\mathcal{G}_{u,v}$ is $\{G \mid G \in \mathcal{G} \wedge u, v \in \mathcal{L}(G)\}$. This empirical mean is well defined as we assumed $|\mathcal{G}_{u,v}| > 0$.
3. Run FastME's heuristic for balanced minimum evolution [17] (referred to as FastME from here on) on D, outputting an unrooted species tree.

ASTRID is statistically consistent under the MSC when given true gene trees [2], and is in practice very fast while having competitive accuracy [49, 34].

2.3 Weighted ASTRID

In the definition of the internode distance, each branch contributes equally to the internode distance for each pair of taxa it separates. Intuitively, under the realistic assumption that gene trees are estimated with a non-trivial amount of error, some branches will be more reliably estimated than the others. As such it makes sense to assign weights to branches as some confidence of them correctly contributing to the internode distance. For example, because zero-support branches very likely do not contribute correctly to the internode distance, assigning a weight of 0 to such branches likely discounts incorrect contributions to the internode distance. The branch lengths could also be used as such a proxy, because short branches are empirically hard to estimate. Thus, our problem becomes to choose appropriate weighting schemes for the edges based on information already annotated in the gene trees, that is, the branch support and branch lengths. Because branch support is already designed as some statistical confidence of the correctness of some branch, it seems natural to naively assign the support directly as the weight for each branch. We alternatively explore simply assigning the branch length as the weight. The details are presented as follows.

2.3.1 Distance defined by branch support

We now formally introduce **wASTRID-s** (weighted ASTRID by support), analogous to the naming of weighted ASTRAL by support. As mentioned above, if a branch has low support, it is less likely that this branch contributes correctly to the internode distance, and thus branches with low support should contribute less. Here we try one simple approach, defining each branch's contribution to the internode distance as its support instead of 1, which gives rise to the following definition of $d_G(u,v)$, the new support-weighted intertaxon distance replacing the internode distance from step 2 of ASTRID:

$$d_G(u,v) = \sum_{e \in P_G(u,v)} s(e)$$

In reality, several different measures of support exist with different running time and accuracy trade-offs [3]. Weighted ASTRAL discovered that the approximate Bayesian support [3] of IQ-TREE led to the most accurate species tree reconstruction, although other measures of support also led to accuracy improvements over unweighted ASTRAL. We leave this choice of support as a parameter to be decided later for wASTRID-s.

While ASTRID is statistically consistent under the MSC when given true gene trees [2], the statistical consistency of wASTRID-s only makes sense under *estimated* gene trees because only estimated gene trees can have meaningful branch support. We conjecture that similar to wASTRAL-s (support weighted ASTRAL), wASTRID-s is statistically consistent under the MSC under some probabilistic interpretation of branch support when given estimated gene trees.

2.3.2 Distance defined by branch lengths

Unlike branch support, which is designed to be a measure of statistical confidence on the correctness of a branch, branch lengths can only serve as proxies to such information, where shorter branches are empirically harder to estimate likely as a result of shorter branches containing less information (fewer substitutions) [50]. We do not attempt a complex conversion here, and simply just assign the branch length as the confidence similar to how we use the support values:

$$d_G(u,v) = \sum_{e \in P_G(u,v)} l(e)$$

Notably, this definition of $d_G(u,v)$ coincides with STEAC [23] (motivated by a different perspective of the coalescence time between genes), and potentially under a more accurate setting when paired with the FastME step of ASTRID. In addition, we also explore whether and how to normalize the input branch lengths of the gene trees for the weighting. We name this final algorithm **wASTRID-pl** (weighted ASTRID by path-lengths).

2.4 Running time and fast distance calculation

Clearly, both wASTRID-s and wASTRID-pl retain the original theoretical running time of ASTRID. Recall that n is the number of species and k is the number of gene trees given in the input. For each gene tree, our metricization simply assigns already-computed values as lengths to each edge; thus calculating the intertaxon distance across all pairs of taxa per gene tree takes $O(n^2)$ time (for normalizing branch lengths in wASTRID-pl, we only explore ways to normalize that does not affect this asymptotic running time). The averaged intertaxon distance thus can be calculated in $O(kn^2)$ time. The running time of FastME (using the balanced minimum evolution heuristic) is $O(n^2 \times \text{diam}(T))$ [8], where $\text{diam}(T)$ is the topological diameter of the output species tree. Assuming the common Yule-Harding distribution [53, 12] or the uniform distribution [10, 1, 27] on the output species tree, the expected diameter is either $\log n$ [10] or $\sqrt{n}$ [8, 10], respectively. Therefore we re-derive the original running time result:

▶ **Theorem 1.** *The averaged intertaxon distance of wASTRID-s and wASTRID-pl can be obtained in $O(kn^2)$ time. The final species tree can be obtained using an additional $O(n^2 \lg n)$ time by FastME assuming the output species tree is under the Yule-Harding distribution (or under the uniform distribution, $O(n^2\sqrt{n})$ time), where n is the number of species and k is the number of genes.*

While the algorithm for the $O(kn^2)$ step can be theoretically uninteresting, we implemented another algorithm for the distance calculation in hope of better in-practice speed. The asymptotically optimal algorithm is easy to devise because the naive algorithm, which given a gene tree, starts a BFS at each leaf to obtain the all-pairs intertaxon distance, is already quadratic time per gene and also asymptotically optimal due to each gene tree having $\binom{n}{2}$ distances. The original ASTRID implementation, in this vein, uses an algorithm which implicitly performs multiple traversals in the tree. For weighted ASTRID, we instead implemented an intertaxon-distance algorithm from TreeSwift [35] based on post-order traversal, through which we hope to achieve better empirical performance due to better cache locality in its simultaneous maintenance of multiple distances from the leaves in an array.

3 Experimental Study

3.1 Overview

We conduct three experiments. In Experiment 1, we explore parameter choices (choice of branch support for wASTRID-s and normalization scheme for wASTRID-pl) for weighted ASTRID. In Experiment 2, we compare the accuracy and running time of weighted ASTRID against other methods on a diverse set of simulated conditions. In Experiment 3, we compare ASTRID, wASTRAL-h, and the best variant of wASTRID (as determined by previous experiments) on the Jarvis et al. [14] avian biological dataset, comparing the quality of the reconstructed species trees.

3.2 Datasets

We assembled a set of diverse data from prior studies (see Table 1), consisting of various simulated conditions with estimated gene trees and one biological dataset ("avian biological") from the avian phylogenomics project [14]. We use the nomenclature of the original ASTRID study and refer to the SimPhy-simulated datasets from the ASTRAL-II study by an "MC" name. The ILS levels of the datasets are measured in average discordance (AD), defined as the average nRF rate between the true species tree and the true gene trees. Same as the original ASTRID study [49], we classify the ILS levels of the datasets into four categories according to their AD percentages, where below 25% is classified as low ILS (L), between 26% and 39% medium ILS (M), between 40% and 59% high ILS (H), and higher AD considered

Table 1 Dataset statistics. The ILS levels of the datasets are categorized according to their AD percentages, where below 25% is low ILS (L), between 26% and 39% mid ILS (M), between 40% and 59% high ILS (H), and higher AD very high ILS (VH). SH-like denotes FastTree default support; BS denotes standard bootstrap support using FastTree or RAxML; aBayes denotes IQ-TREE approximate Bayesian support. (*): training dataset: only 10/50 replicates were used for training.

Dataset	# taxa	# genes	# reps	ILS (AD %)	Branch Support
ASTRAL-II MC2* [33]	201	1000	10	33 (M)	SH-like, BS, aBayes
ASTRAL-III S100 [55]	101	1000	50	46 (H)	aBayes
ASTRAL-II MC3 [33]	201	1000	50	21 (L)	aBayes
ASTRAL-II MC5 [33]	201	1000	50	34 (M)	aBayes
ASTRAL-II MC1 [33]	201	1000	50	69 (VH)	aBayes
ASTRAL-II MC6H [20]	201	1000	50	9 (L)	aBayes
ASTRAL-II MC11H [20]	1001	1000	50	35 (M)	aBayes
Avian 2x [31]	48	1000	20	29 (M)	aBayes
Avian 1x [31]	48	1000	20	47 (H)	aBayes
Avian 0.5x [31]	48	1000	20	60 (VH)	aBayes
Mammalian 2x [31]	37	200	20	21 (L)	aBayes
Mammalian 1x [31]	37	200	20	29 (M)	aBayes
Mammalian 0.5x [31]	37	200	20	50 (H)	aBayes
Jarvis et al. avian [14]	48	14446	1	not known	BS

very high ILS (VH). For the simulated conditions, we subsample the gene trees to size $k = 50, 200, 1000$ except for the mammalian simulation ($k = 50, 100, 200$ instead as only 200 gene trees were provided).

For the measure of support for the datasets, the weighted ASTRAL study provided gene trees reannotated with aBayes support and IQ-TREE inferred branch lengths for the ASTRAL-II and ASTRAL-III datasets. We also reannotated the avian and mammalian simulation with aBayes support and IQ-TREE branch lengths because aBayes was determined as the best measure of support also for wASTRID-s. We use the MC2 condition as the training data for both wASTRID-s and wASTRID-pl, where to explore the choice of branch support for wASTRID-s, we also took the original FastTree [38] inferred trees annotated with the default FastTree SH-like support and also computed a version of the gene trees with standard bootstrap support [11] using 100 FastTree trees.

3.3 Methods

For testing, we compare against ASTRID, ASTRAL, and weighted ASTRAL, with the following settings:

- **ASTRID** (v2.2.1), the base method of weighted ASTRID. We use the original implementation, turning off missing data imputation, as our data has no missing entries in the final averaged matrix.
- **ASTRAL(-III)** (v5.7.8). We do not contract very low support edges in the gene trees because a good threshold can be hard to determine across datasets. In addition, no contraction allows us the fully explore the impact of weighting.
- **wASTRAL-h** (hybrid weighted ASTRAL, v.1.4.2.3). This was the most accurate weighted ASTRAL from the original study, using both branch lengths and support to weigh gene tree quartets. wASTRAL-h supports parallelization, so we run wASTRAL-h with 16 threads.

Many of the datasets have FastTree-inferred gene trees that were reannotated with IQ-TREE approximate Bayesian support. FastTree-inferred trees have polytomies for identical sequences, but these polytomies will be resolved when the trees get reannotated by IQ-TREE, adding false positive edges which may adversely affect the accuracy of the unweighted methods. In these cases, we run the unweighted methods on the original FastTree gene trees.

3.4 Evaluation criterion

For simulated datasets, we compare the topological error rate of the reconstructed species trees using the normalized Robinson-Foulds error (nRF error) with respect to the true species trees. Because all the inferred and true species trees are binary, the nRF error rate is equivalent to the missing branch rate (ratio of branches in the reference tree missing from the reconstructed tree).

On the avian biological dataset, as the true tree is not known, we compare the estimated species trees against prior topologies (wASTRAL tree and published trees). We also compute the local posterior-probability (localPP) branch support [45] for the reconstructed species trees obtained using wASTARL-h to assess the reliability of the branches.

On all datasets, we keep track of the wall-clock running time of the methods, the time taken from consuming the input gene trees (that may have been preprocessed with new branch support values) until outputting the species tree.

3.5 Experimental Environment

All experiments were conducted on the Illinois Campus Cluster, a heterogeneous cluster that has a four-hour running time limit. The heterogeneity of the hardware makes the wall-clock running times not directly comparable across runs, but can still be used to gather obvious running time trends.

4 Results & Discussion

4.1 Experiment 1: Parameter selection

In Figure 1, we explore the choice of branch support among the default FastTree SH-like support, IQ-TREE approximate Bayesian (aBayes) support (normalized to the $[0, 1]$ range), and bootstrap support (100 FastTree replicates) on the training datasets. The best accuracy of wASTRID-s is obtained by using the normalized aBayes support on gene trees. All measures of support, however, improved the species tree estimation error in general. The superiority of (normalized) aBayes support is consistent with the support chosen for weighted ASTRAL, where it was also found superior to SH-like support and bootstrap support. This advantage is even more pronounced when considering that aBayes support can be obtained much faster than bootstrap support [3].

On this training dataset, wASTRID-pl attained the highest accuracy when normalizing the branch lengths in each gene tree by the maximum path length in that gene tree (better than no normalization). Interestingly, while worse than ASTRID with fewer genes $k \in \{50, 200\}$, wASTRID-pl attained higher accuracy than ASTRID when $k = 1000$. However, when comparing wASTRID-s and wASTRID-pl, wASTRID-pl was always less accurate.

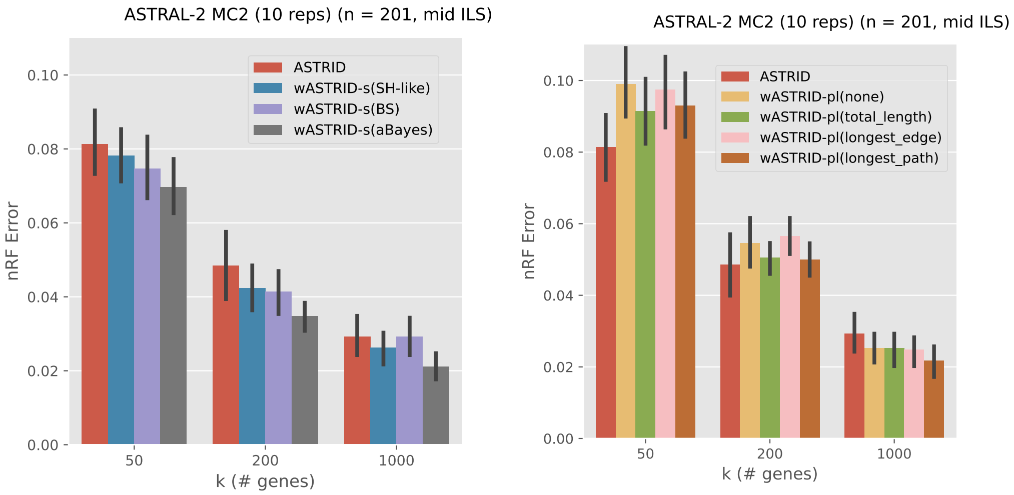


Figure 1 Comparison of the choice of branch support and the choice of branch length normalization strategy for wASTRID-s and wASTRID-pl respectively on the training data, showing the species tree topological error rates (nRF error). "SH-like" is FastTree SH-like support. "aBayes" is IQ-TREE approximate Bayesian support. "BS" is bootstrap support using 100 FastTree trees. The x-axis varies in the number of genes k in the input. Results are shown averaged across ten replicates. Error bars show standard error.

4.2 Experiment 2: Results on simulated datasets

In this experiment, we show four-way comparisons among ASTRID, ASTRAL, weighted ASTRAL (wASTRAL-h), and weighted ASTRID (wASTRID-s). We omit showing the branch-length weighted wASTRID-pl, as it was discovered to be on all datasets less accurate than wASTRID-s. We put an emphasis on the accuracy (nRF error), while later revisiting the problem of running time.

4.2.1 ASTRAL-III S100

This 101-species dataset contains four conditions that varied in the gene tree estimation error (GTEE, measured by the average nRF error between the estimated gene trees and their corresponding true gene trees) by varying the sequence lengths. We show the results of three of the four conditions in Figure 2. In all such figures, we show the unweighted methods (ASTRID, ASTRAL) in dotted lines. On S100, aside from the obvious trend that summary methods become more accurate as k (the number of genes) increases, all methods also improve in accuracy when given more accurate estimated gene trees. These two trends are unsurprising and well-documented across studies for summary methods in general.

More interestingly, the weighted methods (wASTRID-s, wASTRAL-h) are clearly more accurate than their unweighted counterparts, especially at higher levels of GTEE (GTEE $= 0.55, 0.42$). The improvement in accuracy from the weighted methods does not seem to depend on the number of genes, suggesting that the noise brought by low-quality gene trees is not simply resolved by having ample data. This advantage of the weighted methods, however, is smaller as more accurate gene trees are used (GTEE $= 0.31$), as expected.

wASTRID-s clearly improves upon ASTRID on this dataset, with an obvious advantage in the nRF error across all conditions and all numbers of genes. wASTRID-s notably almost matches the accuracy of wASTRAL-h. With $k \in \{200, 1000\}$, no clear benefit exists for using wASTRAL-h on the shown conditions.

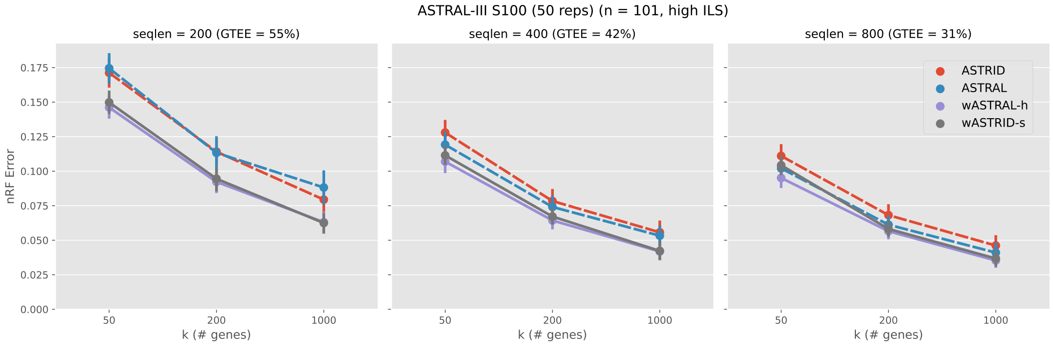


Figure 2 Topological error of species tree across methods on the ASTRAL-III S100 dataset ($n = 101, \mathrm{AD} = 46\%$). Subfigures vary the sequence lengths, affecting the gene tree estimation error (measured in GTEE, the average distance between estimated gene trees and true gene trees). The x-axis varies in the number of genes. Results are shown averaged across 50 replicates with standard error bars. All weighted methods (wASTRID, wASTRAL) ran on gene trees reannotated with IQ-TREE aBayes support branch support and lengths. All methods achieve better accuracy when given more (larger k) or better (lower GTEE) data. Weighted methods are more accurate than unweighted ones. wASTRID-s and wASTRAL-h have almost the same accuracy.

On this dataset, wASTRID-pl (as seen in full results in Appendix Figure 8), similar to trends seen in the training dataset, attains better accuracy than ASTRID when $k = 1000$, but is almost always worse than wASTRID-s. While this better accuracy than ASTRID suggests potential for wASTRID-pl and STEAC-like methods in general, this accuracy disadvantage to wASTRID-s led to us dropping wASTRID-pl from future experiments.

4.2.2 ASTRAL-II SimPhy

This dataset was generated varying the speciation rates, ILS level, and number of taxa, with the "H"-suffixed conditions regenerated from a previous study [20] halving sequence lengths to increase GTEE. We show the results in Figure 3 (only three out of five of the conditions visualized for brevity; see Appendix Figure 9 for the full results).

Across all conditions in this dataset, the weighted methods are more accurate than the unweighted methods. This advantage does not seem to depend on the level of ILS or the number of species. Even under the easiest condition (MC3), wASTRAL-h and wASTRID-s still consistently achieved better accuracy. All methods also performed worse in accuracy as ILS increased, as expected.

While wASTRID-s still consistently improved upon ASTRID in accuracy on this dataset, we also see datasets where wASTRID-s is worse than wASTRAL-h (MC1 as shown here and MC6H as shown in Appendix Figure 9). The relative performance of wASTRID-s and wASTRAL-h seems related to the relative performance of the base methods: MC1 and MC6H are the two conditions that ASTRAL is in general more accurate than ASTRID, but the relative performance of the base methods does not explain the whole picture – for MC1 going to $k = 1000$, ASTRID became more accurate than ASTRAL yet wASTRID-s is still worse than wASTRAL-h. More positively, on the other conditions of this dataset, wASTRAL-h has nearly the same accuracy as wASTRID-s, although wASTRAL-h is marginally more accurate, which might be due to the hybrid weighting of wASTRAL-h also using the branch lengths.

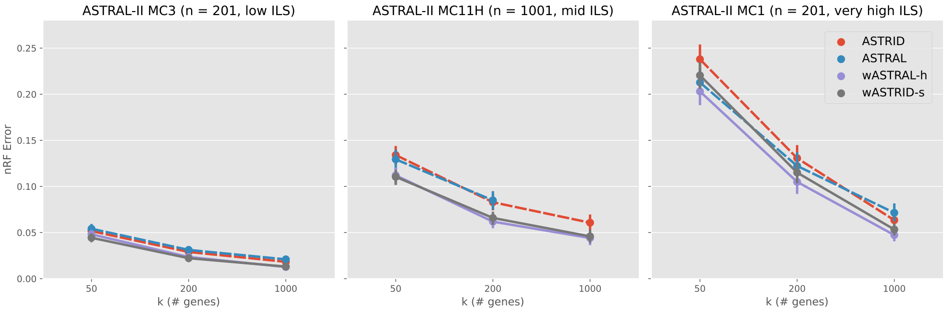


Figure 3 Topological error (nRF error rate) of species tree across methods on selected conditions on the ASTRAL-II SimPhy conditions ($n = 201, 1001, 201$, AD $= 21, 35, 69\%$ respectively). Each subfigure depicts a different model condition. The x-axis varies in the number of genes. Results are shown averaged across 50 replicates with standard error bars. ASTRAL did not finish 24 out of the 50 replicates within four hours for $k = 1000$ on MC11H and thus the data point was omitted. All weighted methods (wASTRID, wASTRAL) were run on gene trees reannotated with IQ-TREE aBayes support branch support and lengths. Weighted methods are more accurate than unweighted ones. wASTRID-s on MC1 was less accurate than wASTRAL-h and otherwise has the same accuracy.

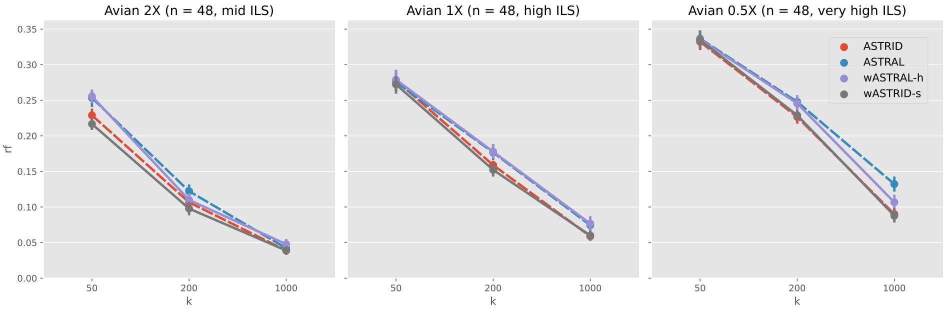


Figure 4 Topological error (nRF error rate) of species tree across methods on the avian simulation ($n = 48$). Each subfigure depicts a different model condition. The x-axis varies in the number of genes. Results are shown averaged across 20 replicates with standard error bars. All weighted methods (wASTRID, wASTRAL) ran on gene trees reannotated with IQ-TREE aBayes support branch support and lengths. ASTRID and wASTRID-s are more accurate than ASTRAL and wASTRAL-h, with a slight accuracy advantage to the weighted methods over the unweighted ones.

On the largest input of these conditions (MC11H, $k = 1000$), ASTRAL did not finish under our four-hour time limit on around half of the replicates (see Appendix Section A.7 for more details), but wASTRAL-h did. We comment on this scalability advantage of wASTRAL-h over ASTRAL later.

4.2.3 Avian and mammalian simulation

These two datasets were generated based on model trees inferred on biological datasets. Both datasets have three conditions with varying ILS by scaling the model tree branch lengths by 2X, 1X, or 0.5X, with shorter branch lengths leading to higher degrees of ILS. Notably, prior results from the ASTRID study [49] showed that ASTRID outperformed ASTRAL on the avian simulation, while on the mammalian simulation ASTRAL was more accurate. Also, the mammalian simulation only has 200 genes available, so we vary k among $50, 100, 200$ unlike the other datasets.

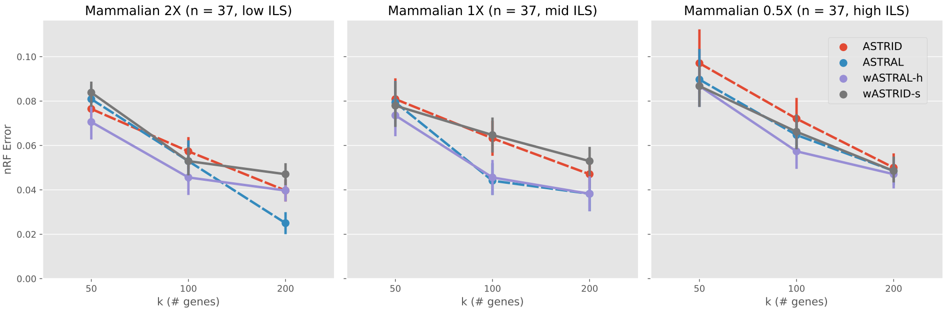


Figure 5 Topological error (nRF error rate) of species tree across methods on the mammalian simulation ($n = 37$). Each subfigure depicts a different model condition. The x-axis varies in the number of genes. Results are shown averaged across 20 replicates with standard error bars. All weighted methods (wASTRID, wASTRAL) ran on gene trees reannotated with IQ-TREE aBayes support branch support and lengths. ASTRAL and wASTRAL-h are more accurate than ASTRID and wASTRID-s. The weighted methods have mixed accuracy compared to the unweighted ones.

On the avian simulation (Figure 4), aside from obvious trends (ILS increases difficulty; more genes leads to more accurate reconstruction), same as the original study, ASTRID is consistently more accurate than ASTRAL. Strangely, although the weighted methods inherit the relative performance of their base methods, in a few cases the weighted methods do not help in accuracy, but they do not erode accuracy either. Even on conditions where the weighted methods improved accuracy, the improvement was small. For example, wASTRAL-h, even though improving upon ASTRAL, is even less accurate than ASTRID, whereas on previously shown data wASTRAL-h was consistently the best in accuracy. This avian simulation does carry substantial GTEE ($> 50\%$), so it is not clear what led to the weighted methods underperforming.

The results for the mammalian simulation (Figure 5) paint a more perplexing picture. On the 2X condition, surprisingly, the weighted methods are less accurate than their unweighted counterparts in general. This trend continues with the 1X condition, where wASTRAL-h only mostly matches ASTRAL in accuracy, and wASTRID-s is worse than ASTRID in accuracy. Only on the 0.5X condition, both weighted methods clearly help in accuracy. wASTRAL-h is clearly better than wASTRID-s on this dataset, but this difference can be explained by the accuracy advantage of ASTRAL on ASTRID. While it is again unclear why the weighted methods underperformed, this dataset is relatively easy compared to the previously shown datasets, with all methods achieving around 0.05 nRF error rate even with $k = 200$, so despite the puzzling relative performance, the difference in accuracy among methods is very small.

4.2.4 Running time

We show the wall-clock running time of the four methods under three representative conditions ($n = 101, 201, 1001$) in Table 2, with a direct comparison of the two most accurate methods visualized in Figure 6. While the heterogeneity of the hardware dilutes the comparability of the running times, clearly wASTRID-s and ASTRID are much faster than wASTRAL-h and ASTRAL, with wASTRID-s on average taking less than 12 seconds even on the largest input, whereas on the same input wASTRAL-h on average takes roughly two hours. In general, wASTRID-s is around two orders of magnitude faster than wASTRAL-h. Although we note that the default flags of both ASTRAL and wASTRAL-h (that we used in the experiments)

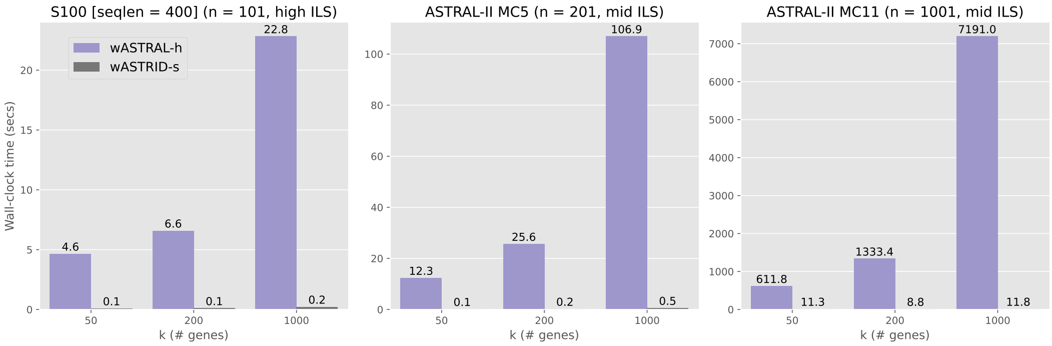


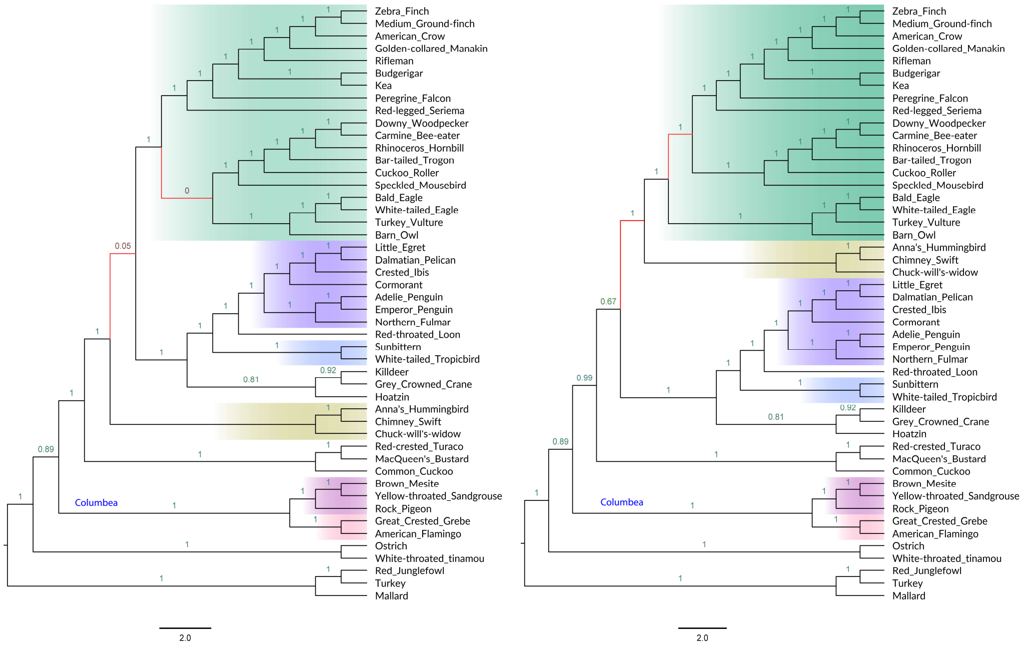
Figure 6 Wall-clock running time (sec) comparison of wASTRID-s and wASTRAL-h on selected representative simulated conditions on $n = 101, 201, 1001$ for k ranging in $50, 200, 1000$. Bars and labels show averages across 50 replicates. wASTRID-s is dramatically faster than wASTRAL-h.

Table 2 Wall-clock running time (sec) across methods on selected representative simulated conditions on $n = 101, 201, 1001$ for k ranging in $50, 200, 1000$. Data points show averages across 50 replicates. ASTRAL did not finish 24 out of the 50 replicates within four hours for $k = 1000$ on MC11H and thus the data point was omitted. The methods sorted by the fastest to the slowest are almost always wASTRID-s, ASTRID, wASTRAL-h, and ASTRAL across all shown conditions. ASTRID and wASTRID-s are much faster than ASTRAL and wASTRAL-h.

Running time (s)		ASTRAL	ASTRID	wASTRAL-h	wASTRID-s
	k (# genes)				
S100 ($n = 101$)	50	12.1	0.1	4.6	0.1
	200	29.6	0.2	6.6	0.1
	1000	300.7	1.4	22.8	0.2
MC5 ($n = 201$)	50	20.3	0.2	12.3	0.1
	200	57.5	0.6	25.6	0.2
	1000	600.7	1.8	106.9	0.5
MC11H ($n = 1001$)	50	552.9	18.0	611.8	11.3
	200	2059.6	14.9	1333.4	8.8
	1000	–	27.5	7191.0	11.8

also calculate support and lengths for reconstructed species tree, in practice, this is a fast step relative to the species tree reconstruction, and does not affect our running time analysis in any substantial way. On MC11H, wASTRID-s and ASTRID took less time going from $k = 50$ to 200, likely due to the $k = 50$ trees having larger diameters, negatively impacting the FastME step running time which has a linear dependency on the diameter of the output tree.

The weighted methods are faster than their unweighted counterparts. For example on S100 (seqlen = 400), wASTRAL-h under $k = 1000$ is more than ten times faster than ASTRAL, and ASTRAL did not finish around half of the datasets for the largest input (MC11H, $k = 1000$). Aside from the benefit of parallelization (we ran wASTRAL-h using 16 threads, but off-the-shelf ASTRAL does not support parallelization), this speed advantage under a large number of genes of wASTRAL-h over ASTRAL can also be attributed to the algorithmic change implemented in wASTRAL-h. The new weighted ASTRAL algorithm removes the in-practice quadratic dependency of ASTRAL's search algorithm on the number



(a) wASTRID-s avian tree.

(b) wASTRAL-h avian tree.

Figure 7 Results on the avian biological dataset ($n = 48, k = 14446$). In (a) and (b), we show the reconstructed species tree topology of wASTRID-s and wASTRAL-h, annotated with local posterior-probability branch support (localPP) computed using wASTRAL-h. The red branches show where the two trees differ. The six well corroborated clades according to Braun and Kimball [5] are displayed by both trees and highlighted in gradients. For wASTRID-s, the red branches also coincide with the only two low support branches. Contracting the very low support branches for wASTRID-s arrives at a topology compatible with wASTRAL-h. wASTRAL-h took around 294 seconds to infer its tree. wASTRID-s was very fast and took 1 second to infer its topology.

of genes [30] and instead has a linear dependency on k for the running time. wASTRID-s is consistently faster (at least two times faster on most conditions) than ASTRID, showing that the new distance calculation algorithm implemented is more efficient than the original one.

4.3 Experiment 3: Results on the avian biological dataset

Jarvis et al. studied the phylogeny of birds using a dataset on 48 taxa using 14446 genes [14]. The original gene trees were annotated with RAxML [46] bootstrap support, which we directly use in our wASTRID-s analysis. This dataset is known to have very low gene resolution, with the average support only 32% [31]. ASTRAL on the original set of gene trees reconstructed a species tree that contained obvious inaccuracies, but contracting low support branches enabled ASTRAL to construct a very plausible topology in agreement with established results [55]. Zhang and Mirarab reanalyzed the original gene trees using wASTRAL-h, and reconstructed the same topology as ASTRAL running on contracted gene trees [54]. In addition to directly comparing against their wASTRAL-h tree, we also compute the number of differing branches between the inferred trees and the two published trees of the Jarvis et al. study: the ExaML-based concatenation tree, called the TENT ("total evidence nucleotide tree"), and the coalescent-based published tree based on binned MP-EST

(MP-EST* tree). We show the results in Figure 7. In addition to showing the topology and the localPP branch support, we also highlight in gradients six clades that are thought to be strongly corroborated [5] for avian phylogeny.

The ASTRID tree (shown in Appendix Figure 10) differed in eight or nine edges with each of the wASTRAL-h, TENT, and MP-EST* trees. In addition, the ASTRID tree did not recover the Columbea clade, which is a clade that has seen strong support in various analyses of this data [14, 55, 39].

Using wASTRID-s, we recovered a topology that is much more in agreement with the other trees, differing in two branches (4.4% of the branches) with the wASTRAL-h tree. It is also in much higher agreement with the published trees (differing in three branches with the TENT, two branches with the MP-EST* tree). Looking closer, the two branches where it differs from the wASTRAL-h tree coincide with the only two very low support branches (no more than 5% in support), and thus contracting the very low support branches arrives at a topology compatible with the wASTRAL-h tree. Both wASTRID-s and wASTRAL-h recover the Columbea-Passerea split, a major conclusion in the original analysis of this data, and even agree on the placement of the hard-to-place Hoatzin.

For running time, both ASTRID and wASTRID-s finished quickly. ASTRID completed in 6.4 seconds, and wASTRID-s completed in 1.1 seconds. Our rerun of wASTRAL-h on this data finished in 294.2 seconds, showcasing the much better scalability of the new weighted ASTRAL optimization algorithm in the number of genes, whereas ASTRAL on the same input took 32 hours in the ASTRAL-III study. In summary, on this dataset, wASTRID-s inferred a more accurate tree compared to ASTRID, is much faster than wASTRAL-h, and is compatible with wASTRAL-h after contraction of two very low support branches.

4.4 Additional remarks

Impact of number of genes, ILS, and GTEE

In all results shown, all methods achieved better accuracy when given more genes, and all methods have decreased accuracy as the degree of ILS or GTEE increases. These trends are well known and quite expected for all summary methods. The number of species does not seem to influence the accuracy of the methods. The relative performance of the methods does shift across different numbers of genes, but there seems to be no reliable predictor of this change across the conditions. All methods seem equally adversely impacted by ILS in accuracy, but for GTEE, the weighted methods are seen to be more robust to the low signal, as by design the weighted methods can extract better signals from the gene trees under high GTEE.

Impact of weighting

Weighting improved accuracy on all the datasets (simulated and biological) except on the mammalian simulation, and the improvement of accuracy tends to be the same degree for both wASTRAL-h and wASTRID-s, with a slight advantage to wASTRAL-h, likely due to the hybrid weighting using more information than support-weighted ASTRID. Even with ample genes or lower levels of GTEE, the weighted methods still in general improved upon the unweighted variants in accuracy. It is not clear how the mammalian simulation (and to some degree, the avian simulation) differs from the other datasets, where the weighted methods could in cases be worse than the original methods in accuracy. The mammalian simulation has a non-trivial amount of GTEE, but has the smallest number of species among the tested datasets and was generated from a different protocol than the ASTRAL simulation datasets, and all these factors could potentially affect the accuracy. Although the weighted methods are in general faster than their unweighted counterparts, this speed advantage is orthogonal to the weighting and entirely due to faster algorithmic design or implementation.

wASTRAL-h vs. wASTRID-s

ASTRAL and ASTRID are known to be among the most accurate summary methods under ILS, and their relative accuracy is dataset dependent, as also shown in our results. The relative accuracy of their weighted counterparts is clearly influenced by the accuracy of their base methods, as seen in the performance differences in the avian and mammalian simulation, where either wASTRID-s and wASTRAL-h can be the more accurate method. On the ASTRAL simulated datasets (ASTRAL-III S100, ASTRAL-II SimPhy), wASTRID-s and wASTRAL-h have comparable accuracy, with a small advantage to wASTRAL-h, which might be due to wASTRAL-h's more effective hybrid weighting. Somewhat unsurprisingly, on all conditions except the mammalian simulation, wASTRID-s is more accurate than unweighted ASTRAL, even on conditions where ASTRAL is more accurate than ASTRID.

As for running time, wASTRID-s is clearly the winner. Despite wASTRAL-h's improved algorithm and parallelism, wASTRID-s is still two orders of magnitude faster than wASTRAL-h, and hence can provide better scalability on large-scale data. We do not extend this running time discussion to a highly parallelized setting (e.g., 64 cores, common for large-scale data analysis) since wASTRAL-h has not been efficiently parallelized yet (as noted in their future work), so any current discussion likely does not reflect the future parallel efficiency of wASTRAL-h. The wASTRID-s algorithm is trivially parallelizable in its distance calculation (albeit very fast already), but parallelizing the FastME step can be difficult both in design and in implementation, which can be a bottleneck under large n. Intriguingly, wASTRAL-h theoretically can scale better in the number of species than wASTRID-s as wASTRAL-h has a running time that is subquadratic in the number of species, but our results show that at $n = 1001$ this point is far from being reached. Thus wASTRID-s is much faster than wASTRAL-h, while having comparable accuracy.

5 Conclusions

While the estimation of species trees using summary methods, such as ASTRAL, ASTRID, MP-EST, and others, is now commonplace, it is known that gene tree estimation error reduces the accuracy of the estimated species tree. We presented support weighted ASTRID (wASTRID-s), an improvement over ASTRID that incorporates uncertainty in gene tree branches into its estimation of the averaged internode distance. Our work is largely inspired by the recent work of weighted ASTRAL (wASTRAL-h), which improved upon ASTRAL, and we showed that wASTRID-s obtained similar accuracy improvements over ASTRID. The advantage provided by wASTRID-s over ASTRID is most noteworthy under higher degrees of gene tree estimation error. wASTRID-s has very close accuracy to wASTRAL-h and is sometimes more accurate, but overall wASTRAL-h has a small advantage in accuracy. The improvement of wASTRAL-h over wASTRID-s may be due to its weighting also incorporating branch lengths. A branch-length weighted version of ASTRID (wASTRID-pl) has mixed accuracy compared to ASTRID, and does not compete against the support-weighted wASTRID-s in accuracy. In general, wASTRID-s and wASTRAL-h serve as accurate species tree inference methods under ILS and are more robust to GTEE than ASTRID and ASTRAL, two of the most accurate summary methods. Both can scale very well, with wASTRID-s much faster. However, the differences in accuracy are dataset dependent, just as the comparison between ASTRAL and ASTRID for accuracy seems dataset dependent.

This study was limited to datasets where the only cause for gene tree discordance with the species tree was ILS and gene tree estimation error. When considering real world datasets that may have other sources of gene tree heterogeneity, such as GDL or HGT, it seems likely that ASTRAL and other quartet-based methods may have an advantage over ASTRID and wASTRID-s, due to the theoretical proofs of statistical consistency for quartet-based methods for those conditions (and the lack of such proofs for distance-based species tree estimation methods) [18, 26, 7, 13].

Although wASTRID-s is highly accurate and very fast, we recommend using wASTRID-s in conjunction with wASTRAL-h and other species tree estimation methods. Due to its speed, the inclusion of wASTRID-s adds little computational burden, and having multiple different approaches for estimating the species tree, each based on a very different (but statistically consistent) technique, can provide insights into what parts of the species tree are most reliably recovered, and which parts may need further data in order for full resolution.

For future work, finding a way to incorporate branch lengths into the branch certainty scoring for wASTRID-s, a hybrid weighting, will improve accuracy and might close the accuracy gap between wASTRID-s and wASTRAL-h under some conditions. Missing data in the final averaged distance matrix has not been addressed, and generalizing the approach from ASTRID [48] for handling missing entries to weighted ASTRID will be important. ASTRID has been used to speed-up and sometimes improve analyses for ASTRAL by assisting in constraining the search space explored by ASTRAL (e.g., FASTRAL [9]), suggesting that a weighted version of ASTRID might provide even better accuracy improvements in FASTRAL. Another direction for future work is species tree estimation in the presence of gene duplication and loss, where gene family trees have multiple copies of species and so are called MUL-trees. The combination of DISCO [51], a method for decomposing the MUL-trees into single-copy gene trees, with ASTRID produced very good accuracy and scalability [51], suggesting that combining DISCO with wASTRID might be even more accurate.

References

1 David J. Aldous. Stochastic Models and Descriptive Statistics for Phylogenetic Trees, from Yule to Today. *Statistical Science*, 16(1):23–34, 2001. URL: `https://www.jstor.org/stable/2676778`.

2 Elizabeth S. Allman, James H. Degnan, and John A. Rhodes. Species tree inference from gene splits by unrooted star methods. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 15(1):337–342, 2018. `doi:10.1109/TCBB.2016.2604812`.

3 Maria Anisimova, Manuel Gil, Jean-François Dufayard, Christophe Dessimoz, and Olivier Gascuel. Survey of Branch Support Methods Demonstrates Accuracy, Power, and Robustness of Fast Likelihood-based Approximation Schemes. *Systematic Biology*, 60(5):685–699, October 2011. `doi:10.1093/sysbio/syr041`.

4 Md Shamsuzzoha Bayzid, Siavash Mirarab, Bastien Boussau, and Tandy Warnow. Weighted Statistical Binning: Enabling Statistically Consistent Genome-Scale Phylogenetic Analyses. *PLOS ONE*, 10(6):e0129183, June 2015. `doi:10.1371/journal.pone.0129183`.

5 Edward L. Braun and Rebecca T. Kimball. Data types and the phylogeny of neoaves. *Birds*, 2(1):1–22, 2021. `doi:10.3390/birds2010001`.

6 Julia Chifman and Laura Kubatko. Quartet Inference from SNP Data Under the Coalescent Model. *Bioinformatics*, 30(23):3317–3324, December 2014. `doi:10.1093/bioinformatics/btu530`.

7 Constantinos Daskalakis and Sébastien Roch. Species trees from gene trees despite a high rate of lateral genetic transfer: A tight bound. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1621–1630. SIAM, 2016.

8 Richard Desper and Olivier Gascuel. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *Journal of Computational Biology*, 9(5):687–705, 2002. PMID: 12487758. doi:10.1089/106652702761034136.

9 Payam Dibaeinia, Shayan Tabe-Bordbar, and Tandy Warnow. FASTRAL: improving scalability of phylogenomic analysis. *Bioinformatics*, 37(16):2317–2324, August 2021. doi:10.1093/bioinformatics/btab093.

10 Péter L. Erdös, Michael A. Steel, László A. Székely, and Tandy J. Warnow. A few logs suffice to build (almost) all trees: Part II. *Theoretical Computer Science*, 221(1):77–118, June 1999. doi:10.1016/S0304-3975(99)00028-6.

11 Joseph Felsenstein. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution*, 39(4):783–791, 1985. doi:10.1111/j.1558-5646.1985.tb00420.x.

12 E. F. Harding. The probabilities of rooted tree-shapes generated by random bifurcation. *Advances in Applied Probability*, 3(1):44–77, 1971. doi:10.2307/1426329.

13 Max Hill, Brandon Legried, and Sébastien Roch. Species tree estimation under joint modeling of coalescence and duplication: sample complexity of quartet methods. *arXiv preprint*, 2020. arXiv:2007.06697.

14 Erich D. Jarvis, Siavash Mirarab, Andre J. Aberer, Bo Li, Peter Houde, Cai Li, Simon Y. W. Ho, Brant C. Faircloth, Benoit Nabholz, Jason T. Howard, Alexander Suh, Claudia C. Weber, Rute R. da Fonseca, Jianwen Li, Fang Zhang, Hui Li, Long Zhou, Nitish Narula, Liang Liu, Ganesh Ganapathy, Bastien Boussau, Md. Shamsuzzoha Bayzid, Volodymyr Zavidovych, Sankar Subramanian, Toni Gabaldón, Salvador Capella-Gutiérrez, Jaime Huerta-Cepas, Bhanu Rekepalli, Kasper Munch, Mikkel Schierup, Bent Lindow, Wesley C. Warren, David Ray, Richard E. Green, Michael W. Bruford, Xiangjiang Zhan, Andrew Dixon, Shengbin Li, Ning Li, Yinhua Huang, Elizabeth P. Derryberry, Mads Frost Bertelsen, Frederick H. Sheldon, Robb T. Brumfield, Claudio V. Mello, Peter V. Lovell, Morgan Wirthlin, Maria Paula Cruz Schneider, Francisco Prosdocimi, José Alfredo Samaniego, Amhed Missael Vargas Velazquez, Alonzo Alfaro-Núñez, Paula F. Campos, Bent Petersen, Thomas Sicheritz-Ponten, An Pas, Tom Bailey, Paul Scofield, Michael Bunce, David M. Lambert, Qi Zhou, Polina Perelman, Amy C. Driskell, Beth Shapiro, Zijun Xiong, Yongli Zeng, Shiping Liu, Zhenyu Li, Binghang Liu, Kui Wu, Jin Xiao, Xiong Yinqi, Qiuemei Zheng, Yong Zhang, Huanming Yang, Jian Wang, Linnea Smeds, Frank E. Rheindt, Michael Braun, Jon Fjeldsa, Ludovic Orlando, F. Keith Barker, Knud Andreas Jønsson, Warren Johnson, Klaus-Peter Koepfli, Stephen O'Brien, David Haussler, Oliver A. Ryder, Carsten Rahbek, Eske Willerslev, Gary R. Graves, Travis C. Glenn, John McCormack, Dave Burt, Hans Ellegren, Per Alström, Scott V. Edwards, Alexandros Stamatakis, David P. Mindell, Joel Cracraft, Edward L. Braun, Tandy Warnow, Wang Jun, M. Thomas P. Gilbert, and Guojie Zhang. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, 346(6215):1320–1331, December 2014. doi:10.1126/science.1253451.

15 Alexey M Kozlov, Diego Darriba, Tomáš Flouri, Benoit Morel, and Alexandros Stamatakis. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, 35(21):4453–4455, May 2019. doi:10.1093/bioinformatics/btz305.

16 Laura Salter Kubatko and James H. Degnan. Inconsistency of Phylogenetic Estimates from Concatenated Data under Coalescence. *Systematic Biology*, 56(1):17–24, February 2007. doi:10.1080/10635150601146041.

17 Vincent Lefort, Richard Desper, and Olivier Gascuel. FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program. *Molecular Biology and Evolution*, 32(10):2798–2800, October 2015. doi:10.1093/molbev/msv150.

18 Brandon Legried, Erin K Molloy, Tandy Warnow, and Sébastien Roch. Polynomial-time statistical estimation of species trees under gene duplication and loss. *Journal of Computational Biology*, 28(5):452–468, 2021.

19 Frédéric Lemoine and Olivier Gascuel. Gotree/Goalign: toolkit and Go API to facilitate the development of phylogenetic workflows. *NAR Genomics and Bioinformatics*, 3(3), August 2021. doi:10.1093/nargab/lqab075.

20 Baqiao Liu and Tandy Warnow. Data from scalable species tree inference with external constraints, 2021. University of Illinois at Urbana-Champaign. doi:10.13012/B2IDB-2566000_V1.

21 Liang Liu and Lili Yu. Estimating Species Trees from Unrooted Gene Trees. *Systematic Biology*, 60(5):661–667, October 2011. doi:10.1093/sysbio/syr027.

22 Liang Liu, Lili Yu, and Scott V. Edwards. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolutionary Biology*, 10(1):302, October 2010. doi:10.1186/1471-2148-10-302.

23 Liang Liu, Lili Yu, Dennis K. Pearl, and Scott V. Edwards. Estimating Species Phylogenies Using Coalescence Times among Sequences. *Systematic Biology*, 58(5):468–477, October 2009. doi:10.1093/sysbio/syp031.

24 Wayne P. Maddison. Gene Trees in Species Trees. *Systematic Biology*, 46(3):523–536, September 1997. doi:10.1093/sysbio/46.3.523.

25 Mahim Mahbub, Zahin Wahab, Rezwana Reaz, M Saifur Rahman, and Md Shamsuzzoha Bayzid. wQFM: highly accurate genome-scale species tree estimation from weighted quartets. *Bioinformatics*, 37(21):3734–3743, November 2021. doi:10.1093/bioinformatics/btab428.

26 Alexey Markin and Oliver Eulenstein. Quartet-based inference is statistically consistent under the unified duplication-loss-coalescence model. *Bioinformatics*, 37(22):4064–4074, 2021.

27 Andy McKenzie and Mike Steel. Distributions of cherries for two models of trees. *Mathematical Biosciences*, 164(1):81–92, March 2000. doi:10.1016/S0025-5564(99)00060-7.

28 Charles D Michener and Robert R Sokal. A quantitative approach to a problem in classification. *Evolution*, 11(2):130–162, 1957.

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

30 Siavash Mirarab. Species Tree Estimation Using ASTRAL: Practical Considerations. *arXiv:1904.03826 [q-bio]*, October 2019. arXiv:1904.03826.

31 Siavash Mirarab, Md. Shamsuzzoha Bayzid, Bastien Boussau, and Tandy Warnow. Statistical binning enables an accurate coalescent-based estimation of the avian tree. *Science*, 346(6215):1250463, December 2014. doi:10.1126/science.1250463.

32 Siavash Mirarab, Md Shamsuzzoha Bayzid, and Tandy Warnow. Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. *Systematic Biology*, 65(3):366–380, 2016.

33 Siavash Mirarab and Tandy Warnow. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics*, 31(12):i44–i52, June 2015. doi:10.1093/bioinformatics/btv234.

34 Erin K Molloy and Tandy Warnow. To Include or Not to Include: The Impact of Gene Filtering on Species Tree Estimation Methods. *Systematic Biology*, 67(2):285–303, March 2018. doi:10.1093/sysbio/syx077.

35 Naima Moshiri. TreeSwift: A massively scalable Python tree package. *SoftwareX*, 11:100436, January 2020. doi:10.1016/j.softx.2020.100436.

36 Huw A. Ogilvie, Remco R. Bouckaert, and Alexei J. Drummond. StarBEAST2 Brings Faster Species Tree Inference and Accurate Estimates of Substitution Rates. *Molecular Biology and Evolution*, 34(8):2101–2114, August 2017. doi:10.1093/molbev/msx126.

37 Swati Patel, Rebecca T Kimball, and Edward L Braun. Error in phylogenetic estimation for bushes in the tree of life. *J. Phylogenet. Evol. Biol*, 1(2):1–10, 2013.

38 Morgan N. Price, Paramvir S. Dehal, and Adam P. Arkin. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLOS ONE*, 5(3):e9490, March 2010. doi:10.1371/journal.pone.0009490.

39 Maryam Rabiee and Siavash Mirarab. Forcing external constraints on tree inference using astral. *BMC genomics*, 21(2):1–13, 2020.

40 John A. Rhodes, Michael G. Nute, and Tandy Warnow. NJst and ASTRID are not statistically consistent under a random model of missing data, 2020. doi:10.48550/ARXIV.2001.07844.

41 D. F. Robinson and Leslie R. Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1):131–147, February 1981. doi:10.1016/0025-5564(81)90043-2.

42 Sébastien Roch, Michael Nute, and Tandy Warnow. Long-branch attraction in species tree estimation: inconsistency of partitioned likelihood and topology-based summary methods. *Systematic Biology*, 68(2):281–297, 2019.

43 Sébastien Roch and Mike Steel. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. *Theoretical Population Biology*, 100:56–62, March 2015. doi:10.1016/j.tpb.2014.12.005.

44 Naruya Saitou and Masatoshi Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, July 1987. doi:10.1093/oxfordjournals.molbev.a040454.

45 Erfan Sayyari and Siavash Mirarab. Fast Coalescent-Based Computation of Local Branch Support from Quartet Frequencies. *Molecular Biology and Evolution*, 33(7):1654–1668, July 2016. doi:10.1093/molbev/msw079.

46 Alexandros Stamatakis. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, January 2014. doi:10.1093/bioinformatics/btu033.

47 Naoyuki Takahata. Gene genealogy in three related populations: consistency probability between gene and population trees. *Genetics*, 122(4):957–966, August 1989. doi:10.1093/genetics/122.4.957.

48 Pranjal Vachaspati. *Large scale phylogenomic estimation.* PhD thesis, University of Illinois at Urbana-Champaign, 2019.

49 Pranjal Vachaspati and Tandy Warnow. ASTRID: Accurate Species TRees from Internode Distances. *BMC Genomics*, 16(10):S3, October 2015. doi:10.1186/1471-2164-16-S10-S3.

50 John J Wiens, Caitlin A Kuczynski, Sarah A Smith, Daniel G Mulcahy, Jack W Sites Jr, Ted M Townsend, and Tod W Reeder. Branch lengths, support, and congruence: testing the phylogenomic approach with 20 nuclear loci in snakes. *Systematic Biology*, 57(3):420–431, 2008.

51 James Willson, Mrinmoy Saha Roddur, Baqiao Liu, Paul Zaharias, and Tandy Warnow. DISCO: Species Tree Inference using Multicopy Gene Family Tree Decomposition. *Systematic Biology*, 71(3):610–629, May 2022. doi:10.1093/sysbio/syab070.

52 Zhenxiang Xi, Liang Liu, and Charles C. Davis. Genes with minimal phylogenetic information are problematic for coalescent analyses when gene tree estimation is biased. *Molecular Phylogenetics and Evolution*, 92:63–71, November 2015. doi:10.1016/j.ympev.2015.06.009.

53 George Udny Yule. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis. *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character*, 213(402-410):21–87, January 1925. doi:10.1098/rstb.1925.0002.

54 Chao Zhang and Siavash Mirarab. Weighting by gene tree uncertainty improves accuracy of quartet-based species trees. *bioRxiv*, 2022. doi:10.1101/2022.02.19.481132.

55 Chao Zhang, Maryam Rabiee, Erfan Sayyari, and Siavash Mirarab. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, 19(6):153, May 2018. doi:10.1186/s12859-018-2129-y.

A Commands

A.1 Weighted ASTRID

The commands differ depending on the type of support annotated on the gene trees. For FastTree SH-like, bootstrap support, and aBayes support, the commands for wASTRID-s are respectively:

```
wastrid -i $genes -o $output # FastTree SH-like
wastrid -x 100 -i $genes -o $output # bootstrap
wastrid -n 0.333 -i $genes -o $output # aBayes
```

For the final setting of wASTRID-pl (distance defined by path lengths, with branch lengths normalized by the maximum path-length in the tree), the above commands need to be appended with an additional flag: `-m n-length` (distance using "normalized branch length" instead of the default `-m support`).

A.2 Weighted ASTRAL

We ran hybrid-weighted ASTRAL (v1.4.2.3) using the following command on trees annotated with aBayes support:

```
astral-hybrid -x 1 -n 0.333 -i $genes -o $output
```

A.3 ASTRID

We ran ASTRID (v2.2.1) using the following command:

```
ASTRID -s -i $genes -o $output
```

The `-s` flag is specified to skip the missing data imputation step of ASTRID, as all tested data have no missing data in the averaged internode distance matrix (i.e., each pair of taxa appears together at least once in some gene tree).

A.4 ASTRAL

We ran ASTRAL (v5.7.8) using the following command:

```
java -jar astral.5.7.8.jar -i $genes -o $output
```

A.5 Approximate Bayesian support

We computed IQ-TREE (v2.1.2) aBayes support using the following command:

```
iqtree2 -s $aln -te $gtree -m GTR+G -abayes
```

where `$aln` is the alignment file for the gene tree, and `$gtree` is the path to the gene tree topology.

A.6 Bootstrap support

We computed bootstrap support (on the training dataset) using FastTree (v2.1.11) on bootstrap replicates generated by Goalign (v0.3.5) [19].

The following command was used to generate bootstrap replicates (`$aln` is the original gene alignment):

```
goalign build seqboot -i $aln -S -n 100
```

The FastTree command used for inferring a tree from one bootstrap replicate is (`$bs_aln` is one bootstrap alignment replicate generated by Goalign):

```
FastTree -nt -gtr -nosupport $bs_aln > $output
```

The FastTree bootstrap trees are then randomly resolved to eliminate polytomies, and then mapped by RAxML-NG [15] to the original gene trees using the following command:

```
raxml-ng --support --tree $gtree --bs-trees $bs_trees --bs-metric fbp
```

where `$gtree` is the path to the original gene tree, and `$bs_trees` contains new-line separated newick trees inferred on the bootstrap replicates.

A.7 Failures to complete

ASTRAL failed to complete 24/50 replicates (replicates 2, 4, 6, 8, 9, 11, 13, 15, 16, 17, 20, 23, 27, 28, 29, 30, 32, 36, 38, 39, 40, 41, 48, and 49) on the MC11H condition ($n = 1001$) under $k = 1000$. The 24 log files indicate that ASTRAL has indeed timed out on each of the replicates. An example of the last three lines of the log files is attached:

```
Calculated 700000 weights; time (seconds): 1549
Calculated 800000 weights; time (seconds): 1544
slurmstepd: error: *** JOB 5328647 ON golub041 CANCELLED AT 2022-05-08T05:24:48 DUE TO TIME LIMIT ***
```

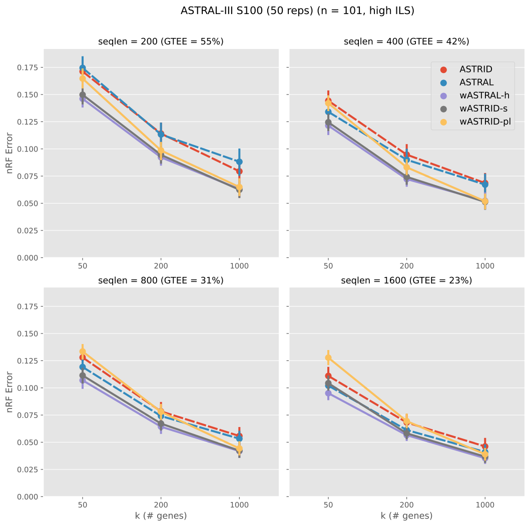


Figure 8 Topological error of species tree across methods on the ASTRAL-III S100 dataset ($n = 101, \mathrm{AD} = 46$). Results are shown averaged across 50 replicates, with error bars showing the standard error. All weighted methods (wASTRID-s, wASTRID-pl, wASTRAL) ran on gene trees reannotated with IQ-TREE aBayes support branch support and lengths.

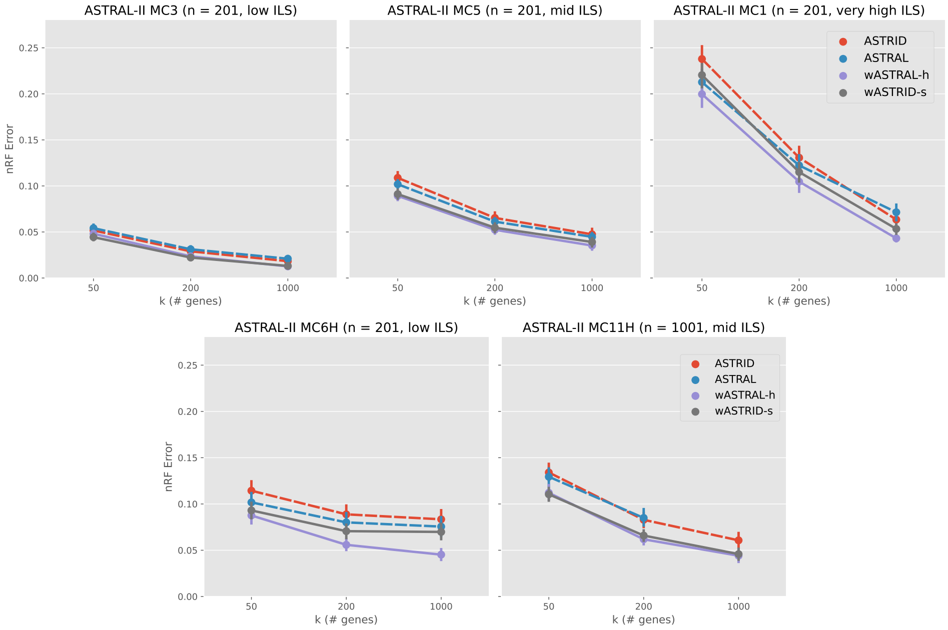


Figure 9 Topological error (nRF error rate) of species tree across methods on selected conditions on the ASTRAL-II SimPhy conditions. Results are shown averaged across 50 replicates with standard error bars. ASTRAL did not finish 24 out of the 50 replicates within four hours for $k = 1000$ on MC11H and thus the data point was omitted. All weighted methods (wASTRID, wASTRAL) were run on gene trees reannotated with IQ-TREE aBayes support branch support and lengths.

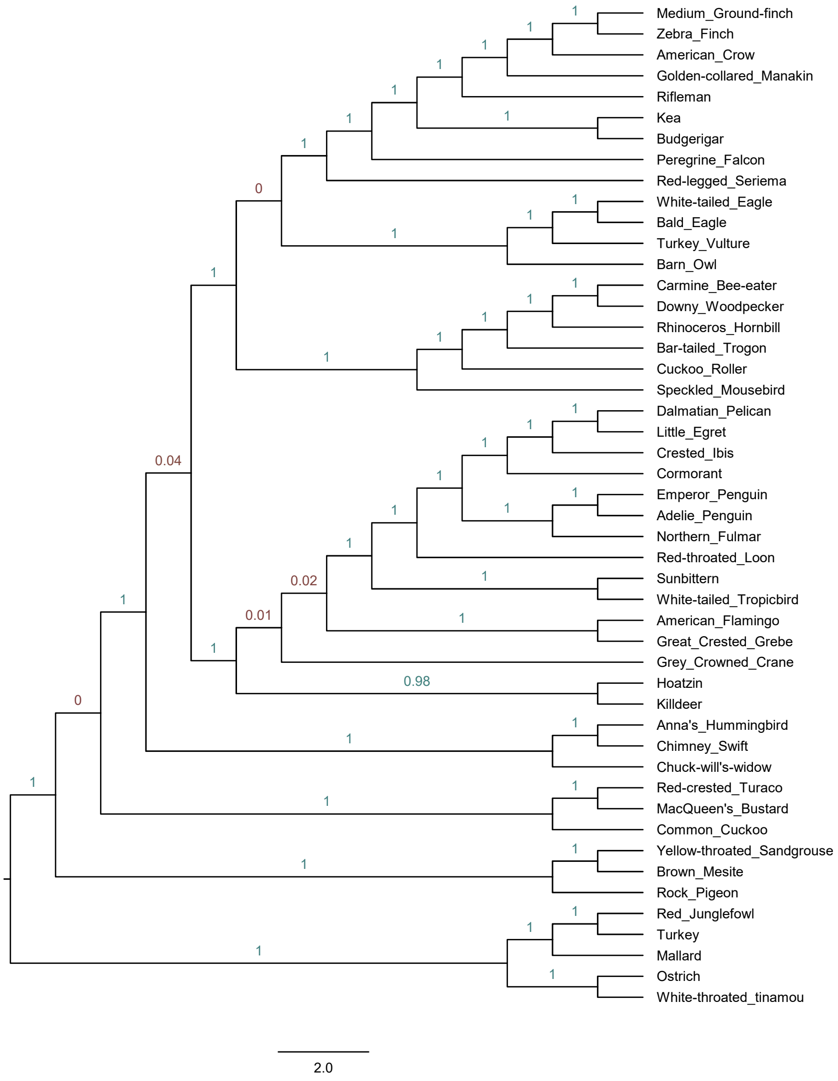


Figure 10 Reconstructed species tree on the Jarvis et al. avian biological data using ASTRID. Branch support values are in localPP values calculated by wASTRAL-h.