On the robustness to gene tree rooting (or lack thereof) of triplet-based species tree estimation methods

Tanjeem Azwad Zaman[1,†], Rabib Jahin[1,†], and Md. Shamsuzzoha Bayzid[1,*]

[1]Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology
Dhaka-1205, Bangladesh
[†]These authors contributed equally to this work
[*]To whom correspondence should be addressed.

Abstract

Species tree estimation is frequently based on phylogenomic approaches that use multiple genes from throughout the genome. This process becomes particularly challenging due to gene tree heterogeneity (discordance), often resulting from Incomplete Lineage Sorting (ILS). Triplet- and quartet-based approaches for species tree estimation have gained substantial attention as they are provably statistically consistent in the presence of ILS. However, unlike quartet-based methods, the limitation of rooted triplet-based methods in handling unrooted gene trees has restricted their adoption in the systematics community. Furthermore, since the induced triplet distribution in a gene tree depends on the placement of the root, the accuracy of triplet-based methods depends on the accuracy of gene tree rooting. Despite progress in developing methods for rooting unrooted gene trees, greatly understudied is the choice of rooting technique and downstream effects on species tree inference under realistic model conditions. This study involves rigorous empirical testing with different gene tree rooting approaches to establish a nuanced understanding of the impact of rooting on species tree accuracy. Moreover, we aim to investigate the conditions under which triplet-based methods provide more accurate species tree estimations than the widely-used quartet-based methods such as ASTRAL.

Keywords: Species tree, gene tree, gene tree heterogeneity, incomplete lineage sorting (ILS), triplets, and quartets.

1 Introduction

The estimation of species trees using multiple loci has become increasingly common. However, combining multi-locus data is difficult, especially in the presence of gene tree

discordance, where different genes may have different evolutionary histories. A traditional approach for species tree estimation from multiple genes is called concatenation (also called 'combined analysis'), where alignments of the genes are concatenated into a supermatrix, which is then used to estimate the species tree. Although it is a widely used technique, concatenation can be problematic as it is agnostic to the topological differences among the gene trees, can be statistically inconsistent (1), and can return incorrect trees with high confidence (2; 3; 4; 5). As a result, "summary methods", which operate by computing gene trees from different loci and then combining the inferred gene trees into a species tree, are becoming increasingly popular (6).

Recent modeling and computational advances have produced summary methods that explicitly take gene tree discordance into account while estimating species trees from multi-locus data. While several biological processes, such as gene duplication and loss, incomplete lineage sorting (ILS), and horizontal gene transfer, can result in this gene tree conflict, ILS (modeled by the multi-species coalescent (7)), is potentially the most prevalent cause of gene tree heterogeneity. Therefore, recent literature has primarily focused on estimating species trees in the presence of ILS. Quartet and triplet-based summary methods (8; 9; 10; 11; 12; 13; 14; 15; 16) have gained significant attention as quartets (4-leaf unrooted trees) and triplets (3-leaf rooted trees) do not contain the "anomaly zone" (17; 18; 19), a condition where the most probable gene tree topology may differ from the species tree topology. ASTRAL (20), the most widely used summary method, is a quartet-based method that uses a dynamic programming (DP) approach to find a species tree that is consistent with the largest number of quartets induced by the set of gene trees. STELAR (16), on the other hand, is a triplet-based method which seeks a species tree by maximizing the number of consistent triplets with respect to the input gene trees. Both ASTRAL and STELAR are statistically consistent, have comparable accuracy and scalable to large dataset. However, while ASTRAL has been a widely adopted choice, STELAR is not as popular as ASTRAL, possibly due to its limitation in analyzing unrooted gene trees. Gene trees are usually estimated using time reversible mutation models which makes the root of the tree non-identifiable (19).

Unrooted gene trees are commonly transformed into rooted trees by incorporating an outgroup that places the root between the outgroup and the remaining taxa in the tree. Another approach involves introducing the assumption of a molecular clock. However, identifying a suitable outgroup proves to be challenging, and the use of a pre-specified outgroup may lead to biased root placement (21; 22). Therefore, in the absence of a molecular clock or a reliable outgroup, alternative techniques for rooting phylogenetic trees have been developed (23; 24; 25; 26; 27; 28). Despite the long history of thinking about tree rooting, there has been a lack of investigation into their impact in the context of estimating species trees using methods that rely on rooted gene trees.

In this study, we focus on triplet-based species tree estimating methods and investigated the robustness (or lack thereof) of these methods to variations in gene tree rooting. We report, on an extensive experimental study using a collection of simulated as well as empirical datasets, the performance of STELAR with gene trees rooted by different techniques. Furthermore, we identified different model conditions where STELAR with gene trees rooted using specific rooting techniques outperforms ASTRAL. These results indicate the potential for finding appropriate roots for the gene trees that result in better species tree estimations than ASTRAL, and thus we believe the underlying causalities

need further research.

2 Background and preliminaries

We now give a brief overview of various rooting techniques analyzed in this study, namely Mid Point (MP), Minimum Variance (MV), Minimal Ancestor Deviation (MAD), Root-digger (RD), outgroup (OG), and random rooting.

2.1 Rooting Methods

Phylogenetic trees can be categorized as rooted or unrooted, each serving distinct purposes in evolutionary research. Rooted trees identify the last common ancestor, providing crucial insights into the directionality of evolution, whereas unrooted trees focus on the relationships among taxa without specifying evolutionary paths. Many methods, each with their own strengths and drawbacks, have been established for rooting unrooted phylogenetic trees.

The outgroup (OG) rooting method is the most commonly used technique for rooting phylogenetic trees. While this method generally provides better rooting accuracy when compared to other rooting methods (29), the main challenge lies in selecting an appropriate outgroup (30; 31; 32; 33; 34; 35). In cases where an appropriate outgroup is unknown, inferring a root is possible using a molecular clock (29; 36). A strict molecular clock assumes a constant rate of substitution across all lineages under consideration; a rather problematic assumption in cases where the ingroup taxa may be distantly related and, consequently, have varying rates of molecular evolution. Relaxed molecular clocks, while comparatively more robust (36), still struggle with accuracy when deviations from a clock-like rate are substantial (37).

Furthermore, there are rooting methods that take into account the distribution of branch lengths (and/or branch supports). Midpoint Rooting (38), Minimum Variance Rooting (39), and Minimal Ancestor Deviation (37) fall under this category. These methods also assume a clock-like behavior and, though they exhibit less dependence, their accuracy also decreases with deviations from this assumption.

Other methods that do not root using an outgroup include gene duplication-based rooting (23; 40; 41), indel-based rooting (24; 42; 43), rooting species trees using distributions of unrooted gene trees (25), probabilistic co-estimation of gene and species trees (26), rooting using a non-reversible Markov model with the help of Multiple Sequence Alignments (27), and rooting species trees using the distribution of quintets induced by gene trees (28).

In this study, we explore the effect of the following methods on gene tree rooting: Outgroup (OG) rooting, Mid-Point (MP) rooting, Minimum Variance (MV) rooting, Minimum Ancestor Deviation (MAD) rooting, Root-Digger (RD), and Random (RAND) rooting. These methods were not restricted to only species trees and thus could be utilized to root gene trees. We discuss in details about these rooting methods in the subsequent sections.

2.1.1 Outgroup Rooting Method (OG)

The outgroup method is a popular rooting approach in phylogenetics (29; 44; 45; 46), often used to determine the evolutionary starting point of a tree. This method assumes that one or more taxa (the outgroup) are divergent from the main group being studied (the ingroup). The branch connecting the outgroup and ingroup sets the evolutionary baseline for the tree (47; 48). The outgroup method not only aids in understanding ingroup evolution but also helps in identifying unique features within ingroup sequences. However, selecting an appropriate outgroup is challenging (30; 31; 32; 33; 34; 35), particularly for datasets with a large number of taxa where a consensus about the outgroup is lacking (49). If an outgroup is too distantly related to the ingroups, it might adversely impact the rooting accuracy owing to a substantially different molecular evolution. Conversely, outgroups that are too closely related with the ingroups may fail to perform its functions as an appropriate outgroup.

Incorrect outgroup selection can also lead to long branch attraction (LBA), a phenomenon where distant outgroup taxa erroneously influence the tree due to their large divergence time or rapid evolution (46). This results in artifactual or random rooting (50; 51) . To mitigate LBA, several criteria, such as low substitution rate and phylogenetic proximity, are suggested for choosing outgroups, particularly in complex cases like arthropod classes (52). Additionally, (22) recommend using multiple outgroup samples within the sister group to minimize LBA and enhance the robustness of the outgroup rooting method. Gatesy et al. (53) demonstrated that introducing just one additional taxon can significantly alter the resulting tree topology, impacting even the existing ingroup taxa. Holland's simulations further confirm that outgroups that influence ingroup topologies are common (54).

OGs can also be used as "true" roots, if the original root is known beforehand (as in the case in simulated datasets).

2.1.2 Mid-point Rooting Method (MP)

Midpoint rooting in phylogenetics is a method where the root is placed at the midpoint between the two furthest tips of the tree. (38; 55) This method works effectively if the tree exhibits constant rates of evolution, relying on the assumption of a molecular clock and homogenous evolution rates across branches (54). Midpoint rooting is particularly suitable for balanced trees but has limitations, especially if the tree data is not clock-like or the topology is unbalanced. The midpoint rooting method usually displays an impressively high success rate, which is especially remarkable in the multiple outgroup, consistently rooted datasets(MOC). (45)

This technique is frequently used in studies where outgroups are not available, such as in viral genetics. An example is (56), who applied midpoint rooting to analyze the evolutionary relationships of SARS coronaviruses, focusing on genes encoding envelope matrix and nucleocapsid proteins. In (57), phylogenetic trees were rooted using Midpoint-rooting method which helped in genome-based phylogeny and classification of prokaryotic viruses. Moreover, phylogenetic reconstuction was done in (58) using midpoint rooting technique. The appropriateness of midpoint rooting is supported by the results of Tajima's relative rate test (59), indicating no rate heterogeneity among the coronavirus groups (56). It is advised to not use MP as the default method and to only use it

if the outgroup method is not applicable e.g. due to a lack of priori knowledge regarding the outgroup or LBA (45).

2.1.3 Minimum Variance Rooting Method (MV)

This section discusses the rooting method introduced in (39), which focuses on reducing the variance of root-to-tip distances in the phylogenetic tree. It is effective in cases where deviations from the strict molecular clock are random and can be implemented in a linear time algorithm, similar to the traditional midpoint rerooting method.

The effectiveness of this method has been explored through extensive simulations, considering factors like gene tree estimation error, divergence from the clock, and outgroup distance to ingroups. (39) showed that MV rooting performs better or at least as well as midpoint rooting across various conditions, especially when the divergence from the clock is less and outgroup distance is smaller. In (60), this method was used to root phylogenies under Outgroup-free rooting.

2.1.4 Root-Digger Rooting Method (RD)

RootDigger, introduced by (27), is a tool that takes an unrooted phylogenetic tree and its corresponding Multiple Sequence Alignment (MSA), and outputs a rooted tree. It uses a non-reversible Markov model to determine the most likely root location on the tree and infer confidence values for each potential root placement. It attempts to addresses the limitations of existing methods like molecular clock analysis (including midpoint rooting) and outgroup rooting. RootDigger circumvents the computationally demanding task of inferring a tree with a non-reversible model by using a reversible model for fast tree inference and then applying a non-reversible model solely for rooting the tree in a final step. RootDigger operates in its default configuration by running in search mode with early stopping enabled. In this mode, RootDigger seeks out the most probable root and terminates the search as soon as it identifies the same root position consecutively, rather than waiting for the likelihood value to repeat. We used search mode in our experiment. RootDigger was used in (61) for Phylogenetic analysis of bestrhodopsins. Here traditional outgroup method failed due to significant evolutionary distance. Alternative approaches such as RootDigger, which uses non-reversible Markov models, was applied to improve rooting accuracy.

RootDigger operates in two modes: Search and Exhaustive. The Search mode quickly identifies the most likely root using heuristics, while the Exhaustive mode thoroughly evaluates the likelihood of placing the root in every branch of the tree, reporting the Likelihood Weight Ratio for each branch. For computational efficiency, we employed the Search mode in this study. We acknowledge that the Exhaustive mode may potentially enhance species tree estimation accuracy by offering a more comprehensive analysis.

2.1.5 Minimal Ancestor Deviation Rooting Method (MAD)

The Minimal Ancestor Deviation (MAD) rooting method is a phylogenetic technique designed to identify the root of an unrooted tree by minimizing deviations from the strict molecular clock hypothesis (37). While strict ultrametricity rarely holds in practice, the midpoint criterion asserts that the middle of the path between two operational taxonomic

units (OTUs) should coincide with their last common ancestor (LCA). The MAD algorithm evaluates this criterion by considering each branch of the tree as a potential root and then calculating the mean relative deviation from the molecular clock expectation for all OTU pairs. The branch that minimizes the deviation is considered the best candidate for the root. The deviation between the observed and expected distances from a putative ancestor node to any two OTUs is quantified by comparing the distance of each OTU to the ancestor with half the distance between the two OTUs. This method tries to ensure that the root placement is the one that best aligns with the expected molecular clock-based distances.

3 Experimental studies

3.1 Datasets

We studied a collection of previously used simulated and biological datasets to evaluate the impact of various rooting techniques on the triplet-based summary method STELAR. We also compared STELAR (with various rooting techniques) with ASTRAL, the leading coalescent-based summary methods, which maximizes quartet-consistency and thus does not require rooted gene trees.

3.1.1 Simulated dataset

We used two biologically-based simulated datasets from (62), which were generated based on species trees estimated by MP-EST (13) on avian and mammalian datasets from (63) and (64), respectively. For the avian simulations, gene sequence lengths were varied (250bp, 500bp, 1000bp, and 1500bp), while in the mammalian simulations, we examined sequence lengths of 250bp, 500bp, and 1000bp to assess the impact of phylogenetic signal. In both simulations, we modeled three levels of incomplete lineage sorting (ILS) by scaling internal branch lengths in the species tree and also explored varying gene counts.

We also used a 15-taxon dataset from (65) with a caterpillar-like (pectinate or ladder-like) model species tree, featuring 12 consecutive short internal branches (0.1 coalescent units) – creating conditions for high levels of ILS. Ultrametric gene trees were simulated along this tree under a multi-species coalescent model, adhering to a strict molecular clock without branch length transformations. Sequence data were generated for each gene tree, and four model conditions were constructed by using gene sequence lengths of 100 or 1000 sites and using either 100 or 1000 genes.

For 11 taxa, we specifically analyzed the regions exhibiting strong incomplete lineage sorting (ILS). The number of genes analyzed varied incrementally, with counts set at 5, 15, 25, 50, and 100 genes.

The simulated datasets we studied varied in many respects (number of genes, sequence length per gene, whether the sequence evolution is ultrametric or not, and the ILS level). Thus, they represent a wide range of model conditions on which we evaluated the impact of various rooting techniques on the performance of triplet-based species tree estimation method STELAR.

We also analyzed two more relatively large datasets containing 200 and 500 taxa. For both 200-taxa and 500 taxa datasets, we analyzed tree lengths 2 M generations with spe-

Table 1: Properties of the simulated datasets.The level of ILS is represented by the average topological distance between the true gene trees and the true species tree.

Dataset	ILS level	# genes	# sites	# Replicates	Ref.
11-taxon	85%	5 - 100	2000	20	(67)
15-taxon	82%	100 - 1000	100 - 1000	10	(65)
37-taxon	18%, 32%, 54%	200 - 800	500 -1000	20	(62)
48-taxon	35%, 47%, 59%	25 - 1000	250 -1000	20	(62)

ciation rates 1e-6 per generation. SimPhy was used to simulate species trees (66) based on the Yule process, defined by parameters such as the number of taxa, the maximum tree length, and the speciation rate, which together specify a model condition. The tree length impacts the amount of ILS, with lower length resulting in shorter branches, and therefore higher levels of ILS. We used medium tree length (2 M) that indicates moderate level ILS. 10 replicates of each dataset (200 taxa, 500 taxa) were analyzed, with each replicate comprising 1,000 gene trees. These datasets were simulated according to the multi-species coalescent model with the population size fixed to 200000 (66).

3.1.2 Empirical dataset

Angiosperm dataset We used the angiosperm dataset of 310 nuclear genes analyzed in (68), (66). The nuclear gene taxon sampling included 42 species representing all major angiosperm clades (35 families and 28 orders (69)). Three gymnosperms (*Picea glauca* [Moench] Voss, *Pinus taeda* L., and *Zamia vazquezii* D.W. Stev., Sabato & De Luca) and one lycophyte (*Selaginella moellendorffii* Hieron.) were included as outgroups. These three gymnosperms span the crown node of extant gymnosperms (70).

Amniota dataset We used the Amniota dataset from (71) containing 16 species and 248 genes. We analyzed the nucleotide (nt) part of the dataset where *Xenopus tropicalis* and/or *Protopterus annectens* used as outgroups.

3.2 Methods compared

We investigated the performance of different rooting techniques paired with STELAR, a statistically consistent triplet-based species tree estimation method. Thus our experimental pipeline contains the following steps: 1) take a set of unrooted gene trees, 2) root them using an appropriate rooting method of our choice, and finally, 3) infer the rooted species tree using a species tree rooting method (STELAR(16)). We used the following rooting methods: Outgroup Rooting (OG)(72), Midpoint Rooting (MP)(72), Minimum Variance rooting (MV) (39), RootDigger(RD) (27), Random rooting (RAND). We refer to STELAR paired with a particular rooting techniques using the convention STELAR–⟨quartet-generation-technique⟩ (e.g, STELAR-OG, STELAR-MP, etc.). We also compared STELAR paired with different rooting techniques with ASTRAL, the leading species tree estimation methods which maximizes quartet consistency.

3.3 Evaluation Metrics

We compared the estimated trees (on simulated datasets) with the model species tree using normalized Robinson-Foulds (RF) distance (73), which is a widely used metric to measure the tree error. The RF distance between two trees is the sum of the bipartitions (splits) induced by one tree but not by the other, and vice versa.

We also compared the triplet and quartet scores of candidate species trees. Triplet score of a rooted species tree T with respect to a particular set $\mathcal{G}$ of rooted gene trees is defined as the number of triplets induced by the gene trees that a the candidate tree T is consistent with. Maximizing triplet score is a statistically consistent criterion for estimating species trees (16). We use the term true triplet score (TTS) when we compute the triplet score with respect to true gene trees. Similarly, the quartet score is defined as the number of quartets in the gene trees that are consistent with the species tree.

To assess the correctness of the root placements inferred by various techniques, we computed the topological distance between the inferred root and the true root (27). This is defined as the number of nodes in the path from the inferred/estimated root to a predefined true root. This distance quantifies the error in root placement without any scaling or normalization (27) based on tree size. This metric was introduced to quantify errors in root placement, addressing the limitations of the binary "percentage of correct rooting" measure, which failed to capture such nuances.

4 Results and discussion

4.1 Simulated Datasets:

4.1.1 37 Taxa

Figure 1 shows the performance of STELAR with gene trees rooted by different methods and ASTRAL on the 37-taxon dataset under various model conditions. One of the key observations from the 37-taxon dataset is the robustness of the triplet-based species tree estimation method, STELAR, to gene trees rooted by various techniques. Most rooting methods, including MP, MV, and MAD, resulted in competitive species trees with no statistically significant differences. Surprisingly, while random rooting (RAND) performed worse in many instances, it still remained competitive under certain model conditions, such as 2X ILS (Figure 1c), 1000 bp, and true gene trees (Figure 1b). Despite yielding comparable accuracy in the 2x ILS model condition (Figure 1c), STELAR performed substantially worse in all other model conditions when working with gene trees rooted by RD.

Interestingly, STELAR sometimes yielded more accurate species trees with rootings other than the known/true OG. There were model conditions where gene trees rooted using MP, MV, and MAD resulted in superior species trees compared to OG and even ASTRAL, albeit these differences were not statistically significant. Such cases include MAD for 0.5X and 1X ILS, and MP, MV for the 1X-200gene-500bp and 2X-200gene-500bp model conditions.

The general patterns were consistent with our expectations: for all methods, the species tree estimation accuracy was improved by increasing the number of genes and sequence lengths (i.e., decreasing the gene tree estimation error), but was reduced by

increasing the amount of gene tree discordance (i.e., the amount of ILS). In general, the choice of rooting techniques becomes less impactful in "easier" model conditions (e.g., higher gene count and more base pairs, with STELAR with MP, MV, OG, and MAD all producing to the same or simiar tree accuracy (e.g., under conditions with 800 genes and true gene trees).

Next, we compared the relative performance of different rooting techniques in terms of species tree accuracy. MP and MV perform notably well on this dataset, especially in "unfavorable" model conditions with fewer genes, shorter gene sequences, and higher ILS. MP matched or outperformed most other methods, including ASTRAL, across various model conditions. Similarly, MV performed on par with other methods, including ASTRAL, and showed significant improvements under low ILS (2X) conditions. Although MAD achieved the best species tree accuracy at high ILS (0.5x), its performance remained largely unaffected by varying levels of ILS. This unresponsiveness of MAD was also observed toward varying gene tree estimation errors (controlled by sequence lengths). However, like other methods, MAD's performance improved with an increasing number of genes.

To further investigate why MP, MV, and occasionally MAD yielded better RF scores than OG, and why RD performed poorly, we conducted a series of experiments examining triplet and quartet scores, as well as the correctness of the roots of the gene trees and species trees inferred by various techniques. First, we compared the triplet scores (Table 2) and quartet scores (Table 4) of the estimated species trees. As expected, due to the statistical consistency property of the triplet score, in almost all cases involving STELAR, a lower RF score corresponds to a higher triplet score. For instance, under high ILS conditions (0.5X), MAD achieved the highest triplet score and the lowest RF score. Interestingly, OG, which is expected to yield the highest triplet score when the outgroup is known for simulated datasets, did not perform as expected under certain model conditions (e.g., 1X-200gt-1000bp and 1X-400gt-500bp). In these scenarios, STELAR using gene trees rooted by OG resulted in marginally lower triplet scores and higher RF scores compared to STELAR with gene trees rooted by MP and MV. There were a few exceptions to this trend, such as in the 1X-200gt-500bp and 2X-200gt-500bp conditions, where OG achieved higher triplet scores despite MP and MV performing better in terms of RF scores.

We also looked at the triplet score of ASTRAL-estimated trees. Though the ASTRAL output species tree is meant to be unrooted, we use the root implied from its Newick representation. Expectedly, ASTRAL–which maximizes quartet scores–consistently achieved substantially lower triplet scores than STELAR despite its higher accuracy. ASTRAL often failed to achieve the best RF scores, even though it always had the highest quartet scores (see 0.5x-200gt-500bp, 1x-200gt-500bp, 2x-200gt-500bp model conditions). Similar to triplet-scores, when STELAR-MP and STELAR-MV outperformed STELAR-OG in species tree accuracy (e.g., 1X-200gt-1000bp and 1X-400gt-500bp), they tended to achieve higher quartet scores. However, as with triplet scores, there were exceptions to this trend in some model conditions such as 0.5x-200gt-500bp and 2X-200gt-500bp.

Next, we examined the percentage of correctly rooted gene trees among the inputs to STELAR for each rooting method. This simulated dataset is rooted using the outgroup (Chicken, Turkey). Aside from the obvious 100% accuracy from OG, we see MV consistently reaching the second best, closely followed by MP and then MAD (Table 6). To

further evaluate gene tree rooting accuracy for the rooting methods, we calculated the average topological distance between the true root and inferred root for the gene trees, in Table 8. This analysis supports the trend seen above: by definition, OG scores 0, indicating perfect accuracy, while MV, MP, and MAD demonstrate progressively increasing topological distances, reflecting their respective accuracies in inferring the root position.

Despite the fact that MP and MV did not achieve the same level of rooting accuracy as OG, as previously noted, the gene trees rooted by MP and MV produced species trees that were equally good or even better than those inferred using OG. This indicates the robustness of triplet-based methods in species tree estimation from gene trees rooted using different techniques under certain model conditions. RAND, as expected, failed to recover the correct root in most cases (around 98% of the times). Surprisingly, RD was even worse than random rooting (RAND), almost never recovering the correct root. This poor rooting accuracy of the gene trees rooted by RD likely explains the poor performance of STELAR-RD.

Finally, we examined the root of the estimated species trees. Unlike the quartet-based species tree estimation method ASTRAL, triplet-based methods work with rooted gene trees and infer a rooted species tree that maximizes the triplet score, making the rooting implied by the output of STELAR meaningful. We therefore analyzed the percentage of correct rooting in the inferred species trees as well as the average topological distance between the inferred species tree roots and the true root, as shown in Tables 7 and 9 respectively. Interestingly, despite the previously noted differences in RF scores, triplet scores, and quartet scores, all triplet-based methods except for STELAR-RD and STELAR-RAND correctly predicted the rooting of the species tree across all model conditions. Moreover, the topological distances between the roots inferred by STELAR-RAND and STELAR-RD and the true roots were not substantially large. These findings highlight the robustness of the triplet-based species tree estimation method STELAR in accurately inferring the root of the estimated species trees, even when the roots of the input gene trees differ.

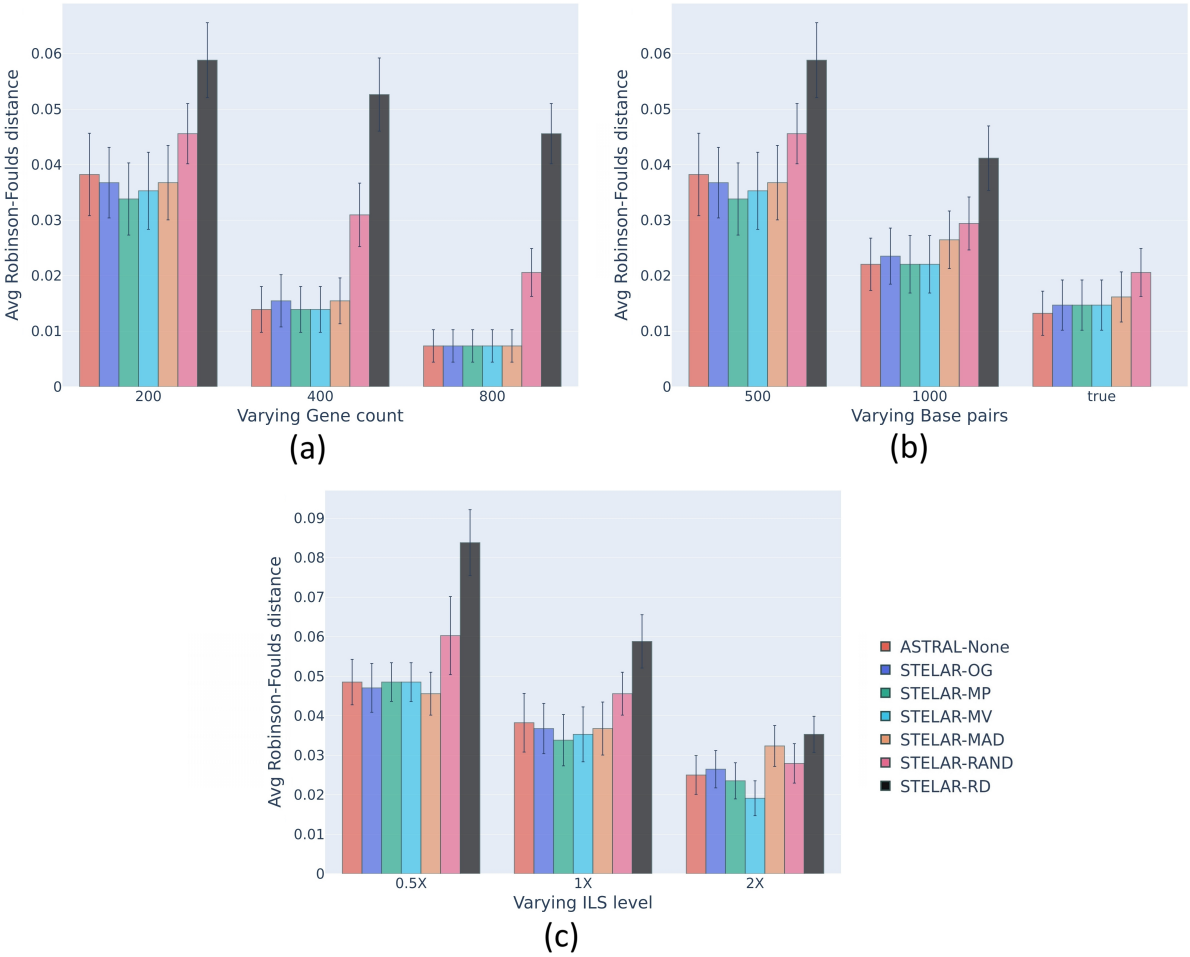


Figure 1: Comparison of Average Robinson-Foulds distance with standard error bars over 20 replicates for ASTRAL and STELAR with different rooting methods for the 37-taxa mammalian dataset. (a) We fixed the ILS at moderate (1X) and the base pairs at 500bp, and varied gene counts between 200, 400 and 800. RD could not be run on 400 genes, since the gene sequences corresponding to the gene trees were not available. (b) We varied the sequence length, and consequently the gene tree estimation error, between 500, 1000 and true-gt while keeping moderate ILS (1X) and 200 gene trees. (c) We set the gene trees at 200, the sequence length at 500bp, and varied the ILS between high (0.5X), moderate (1X) and low (2X)

4.1.2 15-taxa

A general trend similar to that in the 37-taxa dataset is also observed in the analysis of the 15-taxa dataset: all methods, including STELAR with different gene tree rootings, yield better performances with "easier" model conditions. However, a primary point of interest in this dataset is the significantly better performance of STELAR-MAD in terms of RF scores (Figure 2), even outperforming ASTRAL in all model conditions [1]. Despite displaying lower Triplet Scores compared to the other STELAR variants (Table 2), and lower Quartet scores compared to ASTRAL (Table 4), STELAR-MAD showed better robustness to gene tree estimation errors as reflected in its mostly unmatched True Triplet Scores (Table 3) and True Quartet Scores (Table 5). Subsequent discussions on this dataset focus more on a comparative analysis involving methods other than the evidently best-performing STELAR-MAD.

For this dataset in general, STELAR exhibits the same robustness to gene tree rooting as seen in the 37-taxa dataset. In terms of species tree accuracy, comparing the relative performance of STELAR with different rooting techniques shows that STELAR-MP and STELAR-MV performed on par with STELAR-OG, except in the 100gt-true case where STELAR-MV performed better than STELAR-MP, but worse than STELAR-OG. Although STELAR-RD performed better than the other rooting methods on the "unfavorable" model conditions with small numbers of genes and short sequence lengths (100gt - 100bp), it gradually fell off with easier model conditions. STELAR with random rooting yielded competitive performances in some cases (100bp), despite worse results in most other conditions. This drop in performance for STELAR-RAND was more significant than that in the 37-taxa dataset, indicating a possible correlation of worse performance with increasing number of taxa.

A comparison of the performance of ASTRAL with the STELAR variants on this dataset shows that, as previously stated, STELAR-MAD yields significantly better performances in all model conditions. Considering the other rooting methods, STELAR yielded the same species trees as ASTRAL under conditions of 1000gt-1000bp and 1000gt-true, and showed competitive performances in 100gt-1000bp and 100gt-true. However, unlike similar 37-taxa conditions, lower basepairs corresponded to worse performances in general by STELAR; with notable differences from ASTRAL in 100gt-100bp and 1000gt-100bp.

When trying to correlate the triplet scores to performances in terms of RF scores, we notice a trend different from that observed in the 37-taxa dataset. STELAR-MAD yielded the lowest triplet scores in most cases (Table 2). However, we see that triplet scores are not the best indicator of species tree accuracy in this dataset, since STELAR-MAD had the best performance in terms of RF scores. STELAR-MP, STELAR-MV and STELAR-OG yielded the same triplet score in all cases except the true-gt cases where STELAR-MP came in second after STELAR-OG, followed by STELAR-MV in last place. Similar to its performance in the previous dataset, ASTRAL had the highest quartet scores in all cases, whereas STELAR-MAD yielded the best quartet scores among the STELAR variants (Table 4). Comparing the higher quartet scores of of STELAR-MAD with its RF-scores shows that Quartet scores were better correlated to species tree accuracy in this dataset.

[1] except for 100gt-true and 1000gt-true, since the branch lengths were absent in these cases and the MAD rooting method cannot not be run on such gene trees

Robustness to gene tree rooting of Triplet-based Species-tree Estimation methods

Analysis of the percentage of correctly rooted gene trees (Table 6), and the average topological distance between true root and inferred roots (Table 8) yields fascinating insights into the relationship between gene tree rooting and the robustness of STELAR to gene tree estimation error, which were not evident from our analysis of the 37-taxa dataset results. All rooting methods failed to correctly root the gene trees in the true-gt cases. Though MP and MV correctly predicted the gene tree rootings in all other cases, MAD actually rooted the gene trees at points other than the outgroup quite frequently (getting the "correct" root around 20% and 80% of the time in the 100bp and 1000bp cases respectively, from Table 6). A similar trend is reflected in the average topological distances in Table 8 as well. However, this seemingly "inaccurate" rooting by MAD actually helped STELAR in overcoming gene tree estimation errors, resulting in the best performances in this dataset as evident from prior discussions.

Looking at the corresponding metrics for species tree rooting accuracy from Tables 7 and 9 reveals that STELAR, unsurprisingly, always predicted the correct species tree rooting when the gene trees were correctly rooted (e.g. STELAR-MP and STELAR-MV in cases other than true-gt). However, interestingly, despite the inaccuracies in gene tree rooting by MAD in the 1000bp cases, STELAR always correctly predicted the species tree roots. Furthermore, STELAR-MAD achieved accuracies of 50% and 80% in species tree root prediction despite having around 20% accuracy in gene tree rooting in the 100gt-100bp and 1000gt-100bp cases respectively. Moreover, the average topological distances between the predicted species tree root and the true root were only around 1 and 0.2 in these cases, respectively. This attests to the robustness of STELAR to gene tree rooting in the case of species tree root prediction, as also seen from prior discussion on the 37-taxa dataset.

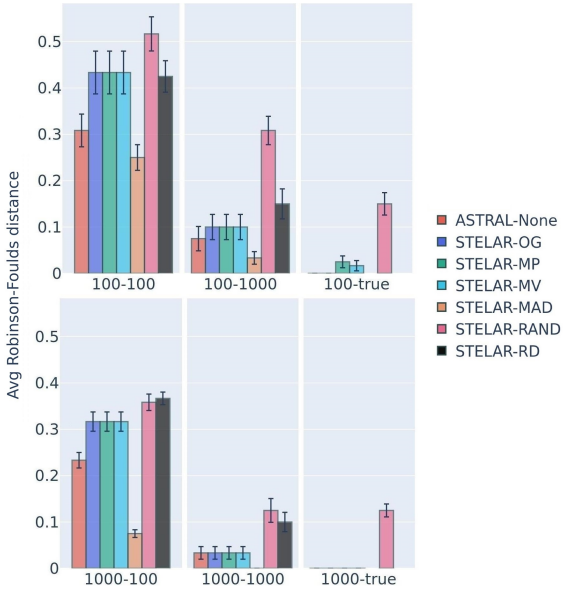


Figure 2: Comparison of Robinson-Foulds distance for ASTRAL and STELAR with different rooting methods on the 15-taxa dataset. Model Condition X-Y denotes X gene trees, Y base pairs. We varied the number of estimated gene trees (100genes - 1000genes) as well as the sequence length (100bp - 1000bp). We also analyzed model conditions with true gene trees; however, due to absense of branch lengths, we could not run MAD on these cases (100-true and 1000-true).

4.1.3 11 Taxa

With varying genecount, the RF scores (figure 3) depict a general trend similar to that of the 37-taxa dataset: increasing genecounts results in more accurate species trees for all methods, leading to lower RF scores. All variants of STELAR (except STELAR-RAND) yielded almost identical performances in terms of RF scores when compared to ASTRAL, with STELAR-MP and STELAR-MV yielding the same species trees in all cases.

The minor differences in RF scores could not be captured completely by the triplet scores, for example - despite having a higher triplet score than ASTRAL in the 50 and 100 gene conditions (Table 2), STELAR-OG had a slightly worse RF score (figure 3). Upon inspecting the True Triplet Scores (TTS) from Table 3, we find that a worse RF score actually corresponded to a worse TTS for STELAR-OG. Other model conditions also mirrored this trend, where the TTS (and not the Triplet Scores) were a good indicator of the RF scores.

Referring to the Quartet Scores in Table 4, we see that ASTRAL has marginally higher quartet scores in each case despite not always being the best in terms of RF scores as previously seen. Similar to the TTS, we find that True Quartet Scores (TQS) from Table 5 better capture the subtle changes in RF scores.

Referring to Table 8, we see that MV displayed a slightly higher inaccuracy in rooting the gene trees, when compared to MP. Despite this, STELAR was robust in estimating the rooting of the final species tree in both cases (Table 9), with the estimated root differing from the true root by one node on average.

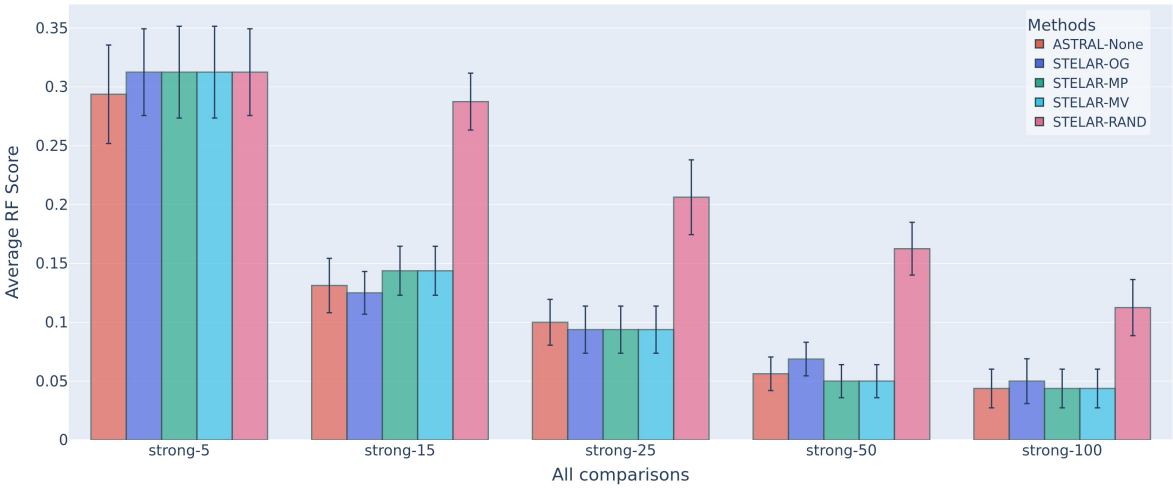


Figure 3: [11-taxa] All gene trees in 1 plot

4.1.4 48 Taxa

All methods followed the trend of better RF scores with "easier" model conditions (figure 4). STELAR yielded better results with outgroup rooting. In this particular dataset, despite yielding competitive performances in "harder" conditions, STELAR-MP, STELAR-MV, STELAR-MAD failed to keep pace with STELAR-OG in higher gene counts (figure 5) and lower ILS (figure 6). Furthermore, STELAR proved to be more robust to random

rooting. We also see that STELAR-RD yielded rather comparable results in most model conditions.

As with previous observations, the Triplet and Quartet Scores followed the trend of increasing with lower RF scores and vice versa (Tables 2 and 4). Exceptions were observed in the cases of STELAR-RAND, and STELAR-RD for triplet scores and STELAR-RD for quartet scores. Under high ILS conditions (0.5X), STELAR-OG achieved the highest triplet score and the lowest RF score among the STELAR variations. Though ASTRAL achieved the lowest rf score, it didn't achieve the higher triplet score compared to STELAR-OG which proves STELAR maximizes triplet score to estimate the species tree. This is true for other ILS conditions as well. Similarly, in the case of quartet Score,

In order to explain the cause for the decline in performance for STELAR-MP, STELAR-MV and STELAR-MAD, we inspect the quality of gene tree rooting depicted in Table 8. Compared to the precious datasets, we find gross inaccuracies in gene tree rooting for MV and MAD, as well as for MP in a slight;y less severe sense. Since the rooting methods MP, MV and MAD are closely dependent on the branch lengths and branch supports of the input gene trees, we suspect irregularities in these values for this 48-taxa dataset, leading to worse gene tree rootings. This irregularity in gene tree rooting led to worse performance for the corresponding STELAR variants.

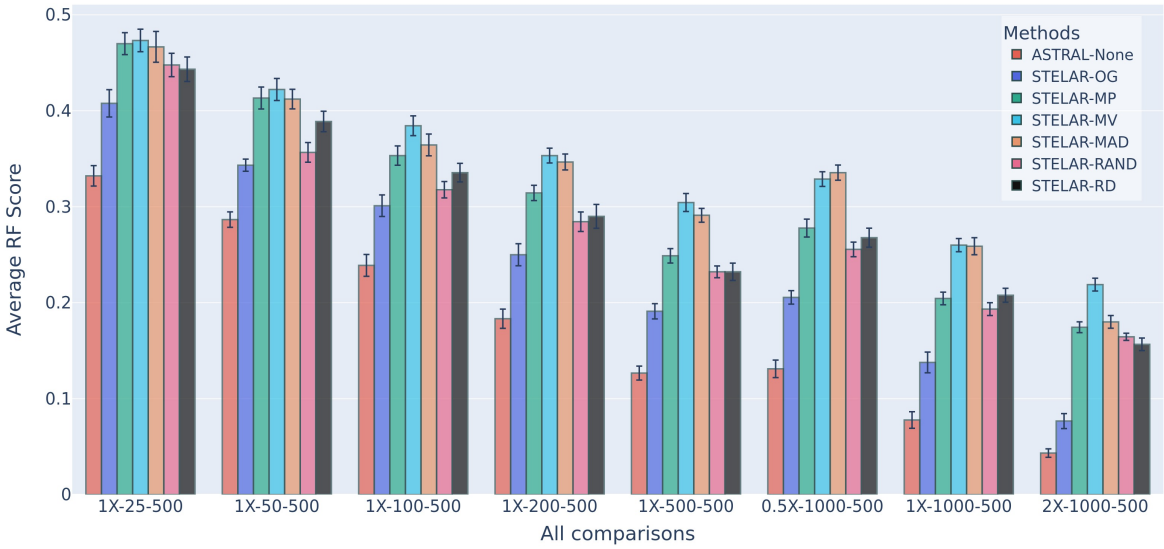


Figure 4: [48-taxa] All gene trees in 1 plot

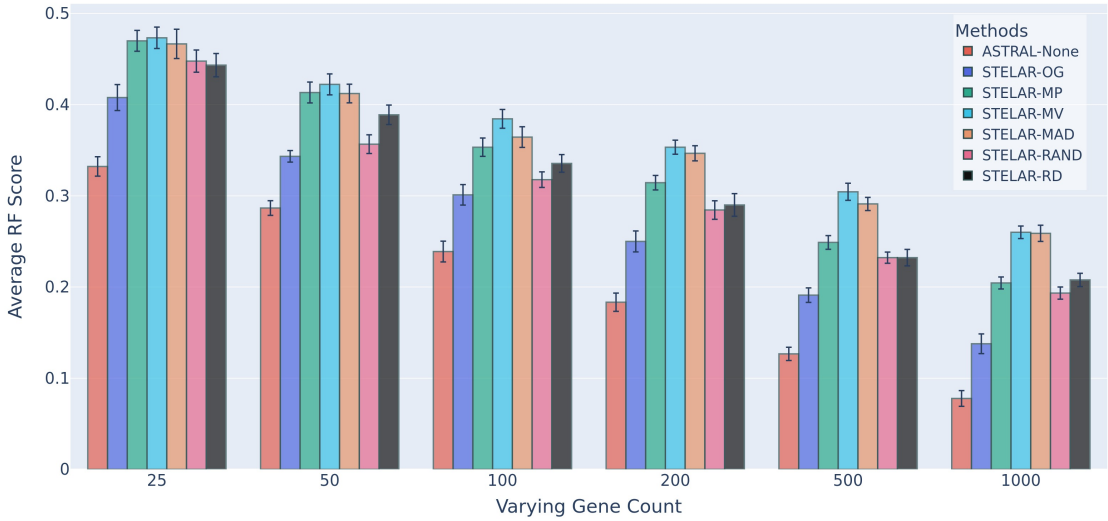


Figure 5: [48-taxa] ILS - 1X, Base Pairs - 500

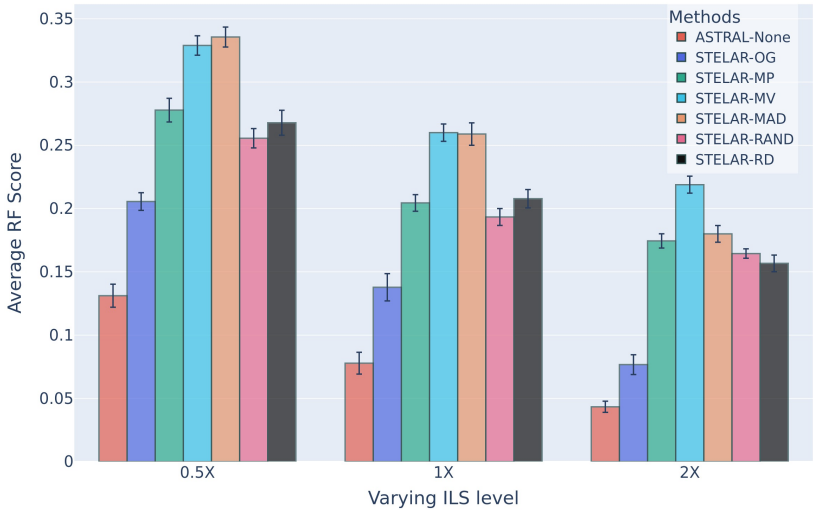


Figure 6: [48-taxa] Gene Count - 1000, Base Pairs - 500

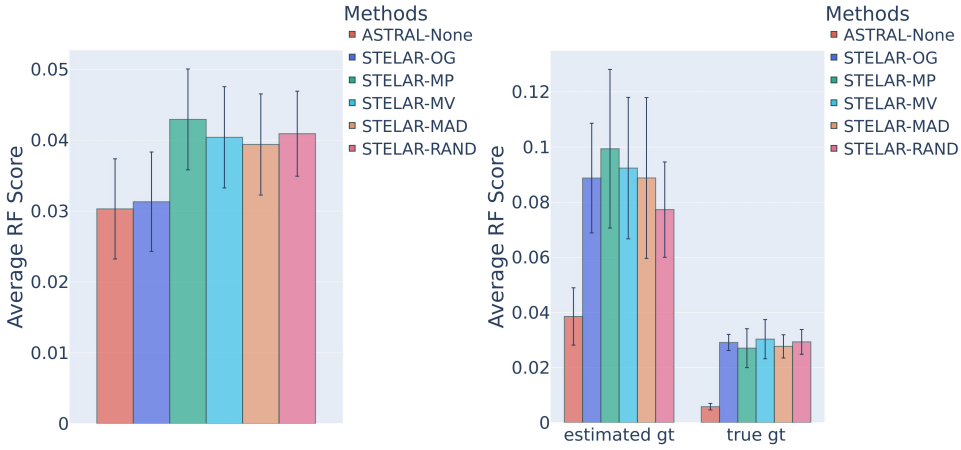
4.1.5 Results on relatively higher number of taxa

On the 200 taxa dataset, STELAR-OG outperformed all other variants of STELAR in terms of RF scores by a significant margin, as seen from Figure 7a. Strikingly, RAND showed a performance comparable to that of MP, MV, MAD. This also translated to the triplet scores, where RAND yielded the highest triplet score among all methods. Though ASTRAL yielded slightly better performance when compared to STELAR-OG. this was not statistically significant.

A similar robustness to gene tree rooting for STELAR is noticed in the 500 taxa dataset depicted in Figure 7b. In the simulated case, STELAR-RAND came in second, outperforming even STELAR-OG in terms of RF scores. STELAR-RAND also showed

a performance comparable to the other STELAR variants in the true gene tree case. However, ASTRAL displayed statistically significant improvement in terms of RF scores when compared to the STELAR variants.

The number of species trees in the search space that differs from the "best" tree by k edges increases as the number of taxa increase. As such STELAR has more options to arrive at a species tree with a better/same RF score, despite working on randomly rooted gene trees. Furthermore a single mismatched bipartition has a more severe impact on the RF score when the number of Taxa is lower. The better performance of STELAR-RAND in terms of RF score for large datasets can quite possibly attributed to these facts. Thus, STELAR grows increasingly robust to gene tree rooting as the number of taxa increases.



(a) 200-taxa **(b)** 500-taxa

Figure 7: Higher-order taxa datasets

4.2 Biological Datasets

4.2.1 Angiosperm Dataset:

Analyses on this dataset aim to address long-standing questions in the phylogeny of Angiosperms, with the key challenges lying in the phylogenetic placement of *Amborella trichopoda* and its relative branching order with other lineages like Nymphaeales (i.e. *Nuphar* or waterlilies)

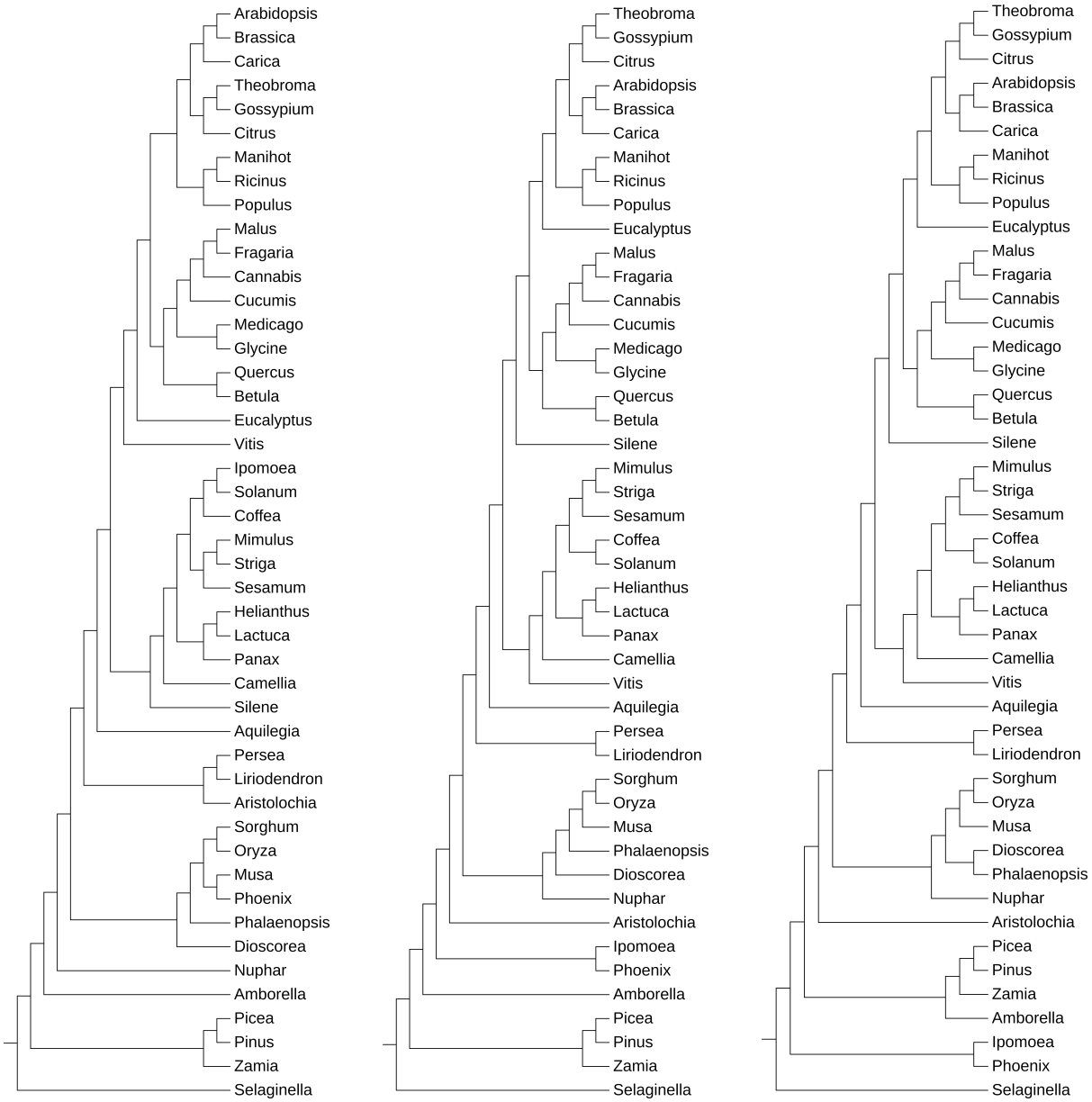
STELAR-MAD correctly placed *Amborella* as a sister to the outgroup comprised of (Selaginella, (Zamia, (Pinus and Picea))) (Fig. 8b). As such, the tree inferred by this method correctly places Amborella as the sister to the other Angiosperms. But, this method could not predict Nuphar as the next sister to the angiosperms (as inferred in (68), nor does it predict the clade of (Amborella, Nuphar) as done in (66). These relationships were not inferred by the other STELAR methods as well.

Furthermore, unlike STELAR-MAD, STELAR-OG, STELAR-MP, and STELAR-MV misplaced *Selaginella*, causing *Amborella* to be placed as a sister to the clade comprised of the 3 other outgroup taxa (Zamia, (Pinus, Picea)) (Fig. 8c).

All STELAR variants had issues with the following placements: swapped positions of Vitis and Silene, misplaced Eucalyptus, Phoenix, Ipomoea and Aristolochia.

The species tree from ASTRAL is the same as stated in (70), and does not misplace the aforementioned taxa (Fig. 8a). It predicts the clade of (Amborella, Nuphar) as sister to the other Angiosperms.

It is worth noting that none of the methods originally rooted the estimated species trees at the outgroup. The STELAR methods predicted the root at (Phoenix, Ipomoea) and ASTRAL originally rooted at *Camellia*. For ease of discussion, the species trees were all rerooted at *Selaginella*.



(a) ASTRAL **(b)** STELAR-MAD **(c)** STELAR-OG, MP, MV

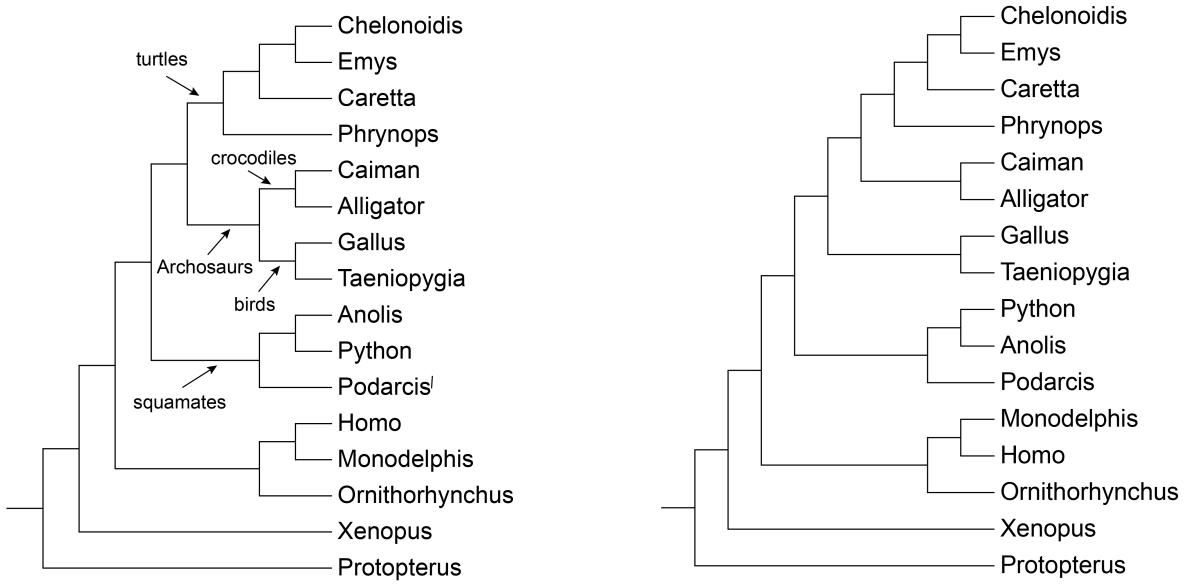
Figure 8: Species trees for Angiosperm dataset

4.2.2 Amniota dataset:

We analyzed the nuleotide (DNA) variant of the dataset from (71), containing 16 amniota taxa. The key challenge lies in resolving the position of turtles relative to crocodiles and birds. Past studies like (12) place turtles as sisters to the archosaurs clade (comprised of birds and crocodiles).

Both the MP and MV variants of STELAR correctly predicted the archosaurs clade, similar to ASTRAL (fig. 9a). However, STELAR-OG failed to predict the Archosaurs clade as seen from fig 9b. All methods correctly predicted the other significant clades like turtles, crocodiles, birds and squamates.

It is worth noting that only STELAR-OG could correctly predict the rooting at Protopterus; ASTRAL and the other STELAR variants could not, and thus yielded lower triplet scores (Table 2) despite their correct clade predictions. The rooting inferred from ASTRAL is *Ornithorhynchus*, whereas STELAR-MV placed the root at the clade ((Protopterus, Xenopus), (Ornithorhyncus, (Homo, Monodelphis)))



(a) ASTRAL, STELAR-MP & MV **(b)** STELAR-OG

Figure 9: Species trees for Amniota (Nucleotide) dataset

In the amino acid variant of the Amniota dataset, ASTRAL correctly inferred all important clades. All variants of STELAR failed to place alligators correctly.

4.2.3 Mammalian Dataset

We analyzed the Mammalian Dataset from (62), containing 37 taxa and 424 genes. Our aim was to study the impact of different rooting methodologies applied to the input gene trees on the species tree inferred by STELAR.

We found that ASTRAL and all the variants of STELAR inferred same species trees from which STELAR-MV, STELAR-MAD inferred the same rooting and other variations inferred different rooting in the species tree. To investigate more, we calculated the quartet score. All the variations of STELAR and ASTRAL yielded same quartet score which supports the inference of same species trees by all the methods.

In conclusion, STELAR demonstrated robustness in inferring species trees regardless of the rooting techniques applied. Even when random rooting was used on the input gene trees, STELAR was able to infer the same species tree. This indicates that, for the Mammalian dataset, the species tree inference by STELAR was unaffected by the rooting of the input gene trees which proves its robustness.

4.3 Metrics

Table 2: Average Triplet Score (TS)

Dataset	Model Conditions	ASTRAL	STELAR					
			MAD	MP	MV	OG	RAND	RD
11-taxon	strong-5	761.65		641.6	641.6	761.65	533.947	
	strong-15	2290.65		1930.45	1930.45	2291.1	1851.1	
	strong-25	3798.6		3198.6	3198.6	3798.6	3149.2	
	strong-50	7589.55		6389.6	6389.6	7589.8	6352.3	
	strong-100	15164.4		12764.4	12764.4	15165.05	12733.9	
15-taxon	100-100	15184.2	17235.2	17695.4	17695.4	17695.4	16457.4	17675.7
	100-1000	17272.8	20110.1	20290.5	20290.5	20290.5	18536.9	20340.2
	100-true	25885.8		23959.9	22207.4	32822.3	19874.4	
	1000-100	149866.4	171608.1	175925.6	175925.6	175925.6	163764.7	175895.2
	1000-1000	170197.6	198719.1	199613.3	199613.3	199613.3	182300.7	201799.8
	1000-true	258798.3		241057.2	218618.5	327645.8	203841.5	
37-taxon	0.5X-200-500	928498.4	989474.5	989467.45	989148.2	989344.1	874149.65	843120.75
	1X-200-1000	1051124.35	1158433.25	1158536.85	1158536.85	1158536.3	961302.25	940074.8
	1X-200-500	981359.05	1101439.55	1101722.65	1101557.0	1102434.0	952731.8	932451.85
	1X-200-true	1305923.7	1410709.85	1410709.85	1410709.85	1410709.85	998847.45	
	1X-400-1000	2218385.85	2314597.85	2314598.0	2314598.0	2314598.0	1926274.9	1925994.8
	1X-400-500	2001280.158	2202684.211	2202711.0	2202711.0	2202709.421	1905627.263	1905137.263
	1X-400-true	2729554.05	2820623.75	2820623.75	2820623.75	2820623.75	1998009.45	
	1X-800-1000	4285710.65	4624759.05	4624759.05	4624759.05	4624759.05	3855959.8	3855226.2
	1X-800-500	4004590.9	4415965.4	4415965.4	4415965.4	4415965.4	3810705.75	3809699.85
	1X-800-true	5461561.3	5479549.1	5643918.0	5643918.0	5643918.0	3997968.1	
	2X-200-500	1006160.45	1152943.65	1154764.9	1153979.25	1154896.6	988994.0	967021.4
48-taxon	0.5X-1000-500	9020299.15	9500062.25	10357359.35	9574800.2	10608691.5	7930725.65	7518799.15
	1X-100-500	1009324.85	720083.6	1068425.35	884224.4	1177023.25	852615.85	833157.05
	1X-1000-500	9855638.35	7606195.8	11379455.1	8782297.3	11723809.35	8499165.85	8171234.05
	1X-200-500	1991144.1	1500354.9	2213331.85	1745214.8	2352178.85	1712047.25	1649962.45
	1X-25-500	254762.85	196975.8	270455.75	219774.65	298140.6	215021.0	212741.2
	1X-50-500	507800.25	365375.65	543951.15	453327.45	591600.8	427917.35	420172.9
	1X-500-500	4920514.05	3797144.3	5693187.55	4361695.9	5862315.2	4311531.4	4099042.55
	2X-1000-500	10634954.65	6965546.05	11458453.75	7945845.2	11915250.75	8947757.9	8574869.6
200-taxon	inner200	717159608.4	755349429.111	765279555.556	749798311.556	759675755.889	785804351.333	
Amniota	Amino Acid	25132		30807	25073	38136	29755	
	Nucleotide	31162		34600	28448	42448	28447	
Angiopsperm	Nuclear	912705	1214602	1200657	1200657	1200657	867932	

Table 3: True Triplet Scores

Taxa	Model Conditions	ASTRAL	STELAR					
			MAD	MP	MV	OG	RAND	RD
11-taxon	strong-5	781.2		660.35	660.35	780.4	558.4	
	strong-15	2404.05		2044.9	2044.9	2407.3	1936.35	
	strong-25	4024.65		3425.05	3425.05	4025.05	3329.9	
	strong-50	8057.45		6859.25	6859.25	8050.7	6748.0	
	strong-100	16124.4		13726.8	13726.8	16116.95	13601.95	
15-taxon	100-100	19589.4	30030.818	30461.9	30461.9	30461.9	20156.0	28224.7
	100-1000	21719.3	32766.9	32521.4	32521.4	32521.4	20864.3	26324.6
	1000-100	187521.0	321010.5	309185.1	309185.1	309185.1	198069.8	289730.0
	1000-1000	231557.5	327645.8	326824.3	326824.3	326824.3	199191.7	262838.1
37-taxon	0.5X-200-500	1105674.15	1248086.15	1248052.1	1248244.05	1248109.45	920812.15	920358.95
	1X-200-1000	1201271.65	1410385.05	1410652.65	1410652.65	1410650.6	998723.7	832956.55
	1X-200-500	1145167.2	1410381.45	1410467.05	1410405.45	1409539.75	998074.45	869667.85
	1X-400-1000	2591563.25	2820598.0	2820594.95	2820594.95	2820594.95	1997631.55	
	1X-400-500	2355862.105	2820118.789	2820230.579	2820230.579	2820224.316	1996615.684	
	1X-800-1000	4774830.25	5479549.1	5479549.1	5479549.1	5479549.1	3881310.75	3147458.75
	1X-800-500	4600609.9	5479505.2	5479505.2	5479505.2	5479505.2	3881010.3	3371166.75
	2X-200-500	1174273.3	1497359.75	1497454.1	1498962.2	1496526.2	1034799.55	1034311.6
48-taxon	0.5X-1000-500	9636236.65	10051174.3	11095538.9	10139002.25	11530857.55	8288721.45	7824656.45
	1X-100-500	1065149.15	732817.85	1127397.95	918274.75	1251100.15	897977.55	877085.75
	1X-1000-500	10462831.35	7840046.25	12066355.55	9109073.35	12591978.75	8972138.9	8617815.35
	1X-200-500	2100679.5	1538912.75	2338739.1	1812038.7	2506970.35	1803167.7	1736660.8
	1X-50-500	528423.65	372201.95	571365.9	469727.05	624282.25	449897.6	441395.05
	1X-500-500	5208654.0	3909483.9	6035359.7	4527019.1	6280826.55	4547023.75	4318498.2
	2X-1000-500	11741419.55	7213265.45	12494254.9	8357037.45	13361902.55	9542505.2	9084815.15

Table 4: Average Quartet Scores

Datasets	Model Conditions	ASTRAL	STELAR					
			MAD	MP	MV	OG	RAND	RD
11-taxon	strong-100	26705.4		26705.4	26705.4	26702.7	26535.85	
	strong-15	4068.25		4066.6	4066.6	4064.6	3970.7	
	strong-25	6695.4		6695.15	6695.15	6695.15	6596.05	
	strong-5	1357.4		1356.9	1356.9	1355.35	1318.9	
	strong-50	13383.1		13383.05	13383.05	13380.1	13190.25	
15-taxon	100-100	69933.8	69686.3	69276.5	69276.5	69276.5	68780.0	69545.4
	100-1000	82166.6	82105.2	82013.2	82013.2	82013.2	80610.5	81830.5
	100-true	84634.9		84516.7	84570.0	84634.9	83301.7	
	1000-100	693656.1	691299.8	684572.4	684572.4	684572.4	682891.75	685192.0
	1000-1000	818022.2	817937.3	817532.111	817532.111	817532.111	811553.333	815358.222
	1000-true	844184.3		844184.3	844184.3	844184.3	833455.6	
37-taxon	0.5X-200-500	10007425.55	10007255.25	10007260.7	10007291.35	10007287.65	9999226.7	9995322.45
	1X-200-1000	11586641.45	11586613.85	11586550.25	11586550.25	11586540.25	11581925.35	11578626.2
	1X-200-500	11270657.55	11270311.8	11270455.75	11270303.45	11270534.45	11262629.65	11257710.1
	1X-200-true	11746203.5	11746194.95	11746198.85	11746198.85	11746198.85	11741292.0	
	1X-400-1000	23159900.85	23159890.65	23159900.85	23159900.85	23159900.85	23148064.55	23144744.4
	1X-400-500	22530090.421	22529758.158	22530050.053	22530050.053	22529985.526	22518243.421	22502535.105
	1X-400-true	23487558.65	23487558.65	23487558.65	23487558.65	23487558.65	23481670.65	
	1X-800-1000	46342464.15	46342464.15	46342464.15	46342464.15	46342464.15	46329619.95	46307948.85
	1X-800-500	45096874.45	45096874.45	45096874.45	45096874.45	45096874.45	45083925.95	45052620.65
	1X-800-true	45632640.45	45632640.45	45632640.45	45632640.45	45632640.45	45621015.4	
	2X-200-500	11947266.5	11945264.3	11947184.75	11947089.25	11947252.75	11943092.85	11939246.25
48-taxon	0.5X-1000-500	110583148.55	103668933.6	109362112.0	104025102.85	110383433.15	109578470.35	109734374.2
	1X-100-500	12383876.5	12025621.5	12138032.25	11964202.2	12318052.45	12207533.7	12179035.15
	1X-1000-500	123024287.2	119274515.2	121790825.05	114929114.55	122916649.8	121998800.7	122241813.9
	1X-200-500	24713063.7	23962846.1	24378714.6	23863357.3	24642898.2	24418292.65	24372977.5
	1X-25-500	3150240.45	3052890.9	3051834.85	3043066.7	3115783.05	3067542.55	3054728.5
	1X-50-500	6234610.35	6065409.95	6058419.0	6010449.1	6182000.55	6122005.95	6084794.85
	1X-500-500	61574240.65	59710450.05	60903919.2	58540593.6	61477542.2	60979019.25	61104007.6
	2X-1000-500	131631393.9	129156563.95	129905861.95	126831800.25	131546355.5	130155966.9	130740197.05
200-taxon	inner200	47622712046.3	47536451073.2	47446554046.6	47519602308.4	47615850756.2	47390448761.7	
Amniota	Amino Acid	83604		82250	82250	82250	82077	
	Nucleotide	97890		97890	97890	97744	97881	
Angiosperm	Nuclear	11553053	10963619	10936281	10936281	10936281	11210579	

Table 5: True Quartet Scores

Dataset	Model Conditions	ASTRAL	STELAR					
			MAD	MP	MV	OG	RAND	RD
11-taxon	strong-100	30708.3		30708.3	30708.3	30663.25	30039.25	
	strong-15	4513.6		4517.15	4517.15	4533.65	4273.95	
	strong-25	7622.3		7625.5	7625.5	7625.5	7283.15	
	strong-5	1391.85		1386.95	1386.95	1387.5	1371.4	
	strong-50	15322.95		15327.75	15327.75	15282.75	14743.55	
15-taxon	100-100	82158.6	83019.8	80525.5	80525.5	80525.5	78400.5	80608.9
	100-1000	84345.1	84545.7	84127.9	84127.9	84127.9	82597.5	83759.7
	1000-100	827970.1	8421420	816941.9	816941.9	816941.9	815221.4	808329.0
	1000-1000	843362.8	844184.3	843362.8	843362.8	843362.8	837052.9	839741.3
37-taxon	0.5X-200-500	10311630.05	10312053.55	10311354.45	10311638.25	10311920.2	10304742.0	10296194.95
	1X-200-1000	11743809.5	11743714.8	11744886.3	11744886.3	11744820.7	11739393.55	11739601.85
	1X-200-500	11739721.0	11740608.15	11740822.35	11740575.95	11740228.2	11734417.9	11729777.7
	1X-400-1000	23486728.75	23486827.8	23486728.75	23486728.75	23486728.75	23470981.1	
	1X-400-500	23482457.0	23479373.158	23481661.053	23481661.053	23481461.053	23468128.368	
	1X-800-1000	45632640.45	45632640.45	45632640.45	45632640.45	45632640.45	45619021.2	45608973.15
	1X-800-500	45631302.8	45631302.8	45631302.8	45631302.8	45631302.8	45610570.45	45596005.65
	2X-200-500	12565281.6	12560555.4	12567292.7	12568800.8	12565175.0	12559268.9	12547261.1
48-taxon	0.5X-1000-500	119513731.6	110530209.8	117731361.0	110982521.8	119135686.95	118129439.45	118354949.2
	1X-100-500	13567423.65	13051253.55	13248847.0	12954507.65	13454157.6	13359236.1	13315353.65
	1X-1000-500	135985986.4	129816269.5	133918071.6	124794923.4	135761706.2	134327102.0	134657697.5
	1X-200-500	27170167.45	25974395.8	26665882.2	25855260.35	27003290.0	26767338.2	26695842.15
	1X-50-500	6776224.95	6549690.9	6565844.35	6482442.4	6708723.9	6668536.7	6611222.05
	1X-500-500	67891776.05	64911557.35	66810392.75	63519122.5	67684969.8	66998284.4	67178572.9
	2X-1000-500	148393885.35	143157302.05	144655982.55	139243603.85	148183540.35	145197013.0	146399056.5

Table 6: Percentage of correctly rooted gene trees

Taxa	Model Conditions	STELAR					
		MAD	MP	MV	OG	RAND	RD
11-taxon	strong-5		0.0	0.0	100.0	11.0	0.0
	strong-15		0.0	0.0	100.0	8.9	0.0
	strong-25		0.0	0.0	100.0	9.0	0.0
	strong-50		0.0	0.0	100.0	9.9	0.0
	strong-100		0.0	0.0	100.0	9.15	0.0
15-taxon	100-100	19.5	100.0	100.0	100.0	7.0	5.9
	100-1000	85.4	100.0	100.0	100.0	7.7	0.0
	100-true	0.0	0.0	0.0	100.0	6.5	
	1000-100	19.6	100.0	100.0	100.0	5.7	0.5
	1000-1000	83.8	100.0	100.0	100.0	5.9	0.0
	1000-true	0.0	0.0	0.0	100.0	6.1	
37-taxon	0.5X-200-500	60.35	70.75	79.15	100.0	2.4	0.1
	1X-200-1000	78.75	81.35	86.9	100.0	2.4	0.1
	1X-200-500	66.85	77.6	85.35	100.0	2.5	0.0
	1X-400-1000	79.25	81.3	87.0	100.0	2.35	0.0
	1X-400-500	64.6	74.25	81.1	100.0	2.25	0.0
	1X-800-1000	78.5	80.9	86.55	100.0	2.45	0.0
	1X-800-500	67.45	78.2	85.25	100.0	2.3	0.0
	2X-200-500	71.6	80.8	87.1	100.0	2.8	0.1
48-taxon	0.5X-1000-500	7.0	0.6	6.55	100.0	1.7	0.0
	1X-100-500	6.2	1.0	7.25	100.0	1.8	0.0
	1X-1000-500	5.15	0.75	6.1	100.0	1.55	0.0
	1X-200-500	5.7	0.9	6.85	100.0	1.6	0.0
	1X-25-500	7.0	0.8	8.4	100.0	2.2	0.0
	1X-50-500	6.3	1.2	7.6	100.0	2.2	0.0
	1X-500-500	5.25	0.7	6.2	100.0	1.75	0.0
	2X-1000-500	3.6	0.8	5.0	100.0	1.7	0.0
200-taxon	estimated	6.6	9.3	11.2	100.0	0.0	
500-taxon	true	25.3	0.0	0.0	100.0	0.0	
	estimated	7.4	12.5	16.7	100.0	0.0	

Table 7: Percentage of correctly rooted species trees

Taxa	Model Conditions	STELAR					
		MAD	MP	MV	OG	RAND	RD
11-taxon	strong-5		0.0	0.0	100.0	0.0	0.0
	strong-15		0.0	0.0	100.0	0.0	0.0
	strong-25		0.0	0.0	100.0	0.0	0.0
	strong-50		0.0	0.0	100.0	0.0	0.0
	strong-100		0.0	0.0	100.0	0.0	0.0
15-taxon	100-100	50.0	100.0	100.0	100.0	0.0	60.0
	100-1000	100.0	100.0	100.0	100.0	0.0	0.0
	100-true	0.0	0.0	0.0	100.0	0.0	
	1000-100	80.0	100.0	100.0	100.0	0.0	90.0
	1000-1000	100.0	100.0	100.0	100.0	0.0	0.0
	1000-true	0.0	0.0	0.0	100.0	0.0	
37-taxon	0.5X-200-500	100.0	100.0	100.0	100.0	0.0	0.0
	1X-200-1000	100.0	100.0	100.0	100.0	0.0	0.0
	1X-200-500	100.0	100.0	100.0	100.0	0.0	0.0
	1X-400-1000	100.0	100.0	100.0	100.0	0.0	0.0
	1X-400-500	100.0	100.0	100.0	100.0	0.0	0.0
	1X-800-1000	100.0	100.0	100.0	100.0	0.0	0.0
	1X-800-500	100.0	100.0	100.0	100.0	0.0	0.0
	2X-200-500	100.0	100.0	100.0	100.0	0.0	0.0
48-taxon	0.5X-1000-500	0.0	0.0	0.0	100.0	0.0	0.0
	1X-100-500	0.0	0.0	0.0	100.0	0.0	0.0
	1X-1000-500	0.0	0.0	0.0	100.0	0.0	0.0
	1X-200-500	0.0	0.0	0.0	100.0	0.0	0.0
	1X-25-500	0.0	0.0	0.0	100.0	0.0	0.0
	1X-50-500	0.0	0.0	0.0	100.0	0.0	0.0
	1X-500-500	0.0	0.0	0.0	100.0	0.0	0.0
	2X-1000-500	0.0	0.0	0.0	100.0	0.0	0.0
200-taxon	estimated	30.0	40.0	40.0	100.0	0.0	
500-taxon	true	40.0	0.0	0.0	100.0	0.0	
	estimated	10.0	20.0	30.0	100.0	0.0	

Table 8: Average Topological Distance between true and estimated roots of gene trees

Taxa	Model Conditions	MAD	MP	MV	OG	RAND	RD
11-taxon	strong-5		1.18	1.33	0.0	3.34	
	strong-15		1.2	1.353	0.0	3.47	
	strong-25		1.182	1.344	0.0	3.42	
	strong-50		1.192	1.339	0.0	3.392	
	strong-100		1.199	1.347	0.0	3.475	
15-taxon	100-100	3.536	0.0	0.0	0.0	5.253	3.685
	100-1000	0.763	0.0	0.0	0.0	5.031	
	100-true		2.948	3.159	0.0	5.048	
	1000-100	3.516	0.001	0.007	0.0	5.329	
	1000-1000	0.82	0.0	0.0	0.0	5.13	
	1000-true		2.916	3.105	0.0	4.958	
37-taxon	0.5X-200-500	1.423	0.634	0.547	0.0	7.606	8.085
	1X-200-1000	0.604	0.387	0.282	0.0	7.65	8.072
	1X-200-500	1.374	0.557	0.44	0.0	7.745	8.149
	1X-400-1000	0.602	0.391	0.282	0.0	7.616	8.094
	1X-400-500	1.341	0.551	0.444	0.0	7.649	8.168
	1X-800-1000	0.614	0.405	0.297	0.0	7.574	8.108
	1X-800-500	1.348	0.563	0.454	0.0	7.702	8.172
	2X-200-500	1.298	0.555	0.443	0.0	7.689	8.107
48-taxon	0.5X-1000-500	5.555	5.075	5.677	0.0	9.026	9.324
	1X-100-500	7.554	5.7	6.815	0.0	9.267	9.514
	1X-1000-500	7.523	5.65	6.861	0.0	9.247	9.538
	1X-200-500	7.497	5.637	6.809	0.0	9.187	9.472
	1X-25-500	7.034	5.386	6.402	0.0	9.158	9.304
	1X-50-500	7.301	5.449	6.643	0.0	9.116	9.391
	1X-500-500	7.495	5.614	6.832	0.0	9.121	9.545
	2X-1000-500	8.599	6.047	7.496	0.0	9.364	9.83
200-taxon	estimated	4.684	4.74	4.76	0.0	11.091	
500-taxon	true	2.632	2.51	2.326	0.0	11.739	
	estimated	5.612	5.917	5.875	0.0	14.15	

Table 9: Average Topological Distance between true and estimated roots of species tree

Taxa	Model Conditions	MAD	MP	MV	OG	RAND	RD
11-taxon	strong-5		1.0	1.0	0.0	2.0	
	strong-15		1.0	1.0	0.0	1.15	
	strong-25		1.0	1.0	0.0	1.05	
	strong-50		1.0	1.0	0.0	1.0	
	strong-100		1.0	1.0	0.0	1.0	
15-taxon	100-100	1.1	0.0	0.0	0.0	2.8	0.4
	100-1000	0.0	0.0	0.0	0.0	4.4	2.5
	100-true		4.9	5.6	0.0	6.0	
	1000-100	0.2	0.0	0.0	0.0	3.0	0.1
	1000-1000	0.0	0.0	0.0	0.0	5.5	3.0
	1000-true		5.0	5.9	0.0	5.8	
37-taxon	0.5X-200-500	0.0	0.0	0.0	0.0	4.1	4.45
	1X-200-1000	0.0	0.0	0.0	0.0	4.05	4.1
	1X-200-500	0.0	0.0	0.0	0.0	4.05	4.15
	1X-400-1000	0.0	0.0	0.0	0.0	4.0	4.05
	1X-400-500	0.0	0.0	0.0	0.0	4.053	4.053
	1X-800-1000	0.0	0.0	0.0	0.0	4.0	4.0
	1X-800-500	0.0	0.0	0.0	0.0	4.0	4.0
	2X-200-500	0.0	0.0	0.0	0.0	3.95	4.1
48-taxon	0.5X-1000-500	2.0	2.0	2.0	0.0	4.7	8.6
	1X-100-500	5.65	2.8	4.0	0.0	6.0	8.0
	1X-1000-500	4.05	2.0	3.2	0.0	5.0	8.6
	1X-200-500	4.7	2.35	3.8	0.0	5.5	8.4
	1X-25-500	6.2	2.7	4.6	0.0	6.75	7.0
	1X-50-500	6.05	2.55	4.0	0.0	5.95	7.25
	1X-500-500	4.15	2.0	3.35	0.0	4.8	8.35
	2X-1000-500	5.0	2.05	3.9	0.0	4.7	8.85
200-taxon	estimated	1.0	1.0	1.1	0.0	3.5	
500-taxon	true	3.1	2.2	2.5	0.0	6.4	
	estimated	3.444	4.9	3.4	0.0	7.0	

5 Conclusions

Both quartets and triplets avoid the "anomaly zone" – a scenario where the most likely gene tree topology may differ from the species tree topology (17; 18; 19). Thus, statistically consistent methods have been developed by maximizing quartet and triplet consistency. While quartet-based summary methods like ASTRAL are in wide use, triplet-based methods like STELAR has not gained notable attention from the community. Moreover, triplet-based methods require rooted gene trees which is often difficult to obtain in the absence of or molecular clock or reliable outgroups. Consequently, various techniques have been developed to root a given set of unrooted gene trees. However, little is known about the impact of these rooting techniques on species tree inference.

Several root inference methods have been described in the literature, differing in the type of data that can be analysed, the assumptions of the evolutionary dynamics of the data, and their scalability or general applicability. In this study, we considered a broad range of rooting approaches and their effects on species tree inference by maximizing triplet consistency. We found that while STELAR was robust to the choice of rooting in some simulated model conditions as well as on the real biological dataset (e.g., biological mammalian dataset), there are conditions where we noticed notable differences in the species tree accuracy based on different gene tree rooting techniques. More interestingly, under some model conditions (like lower ILS for 37 taxa and almost all cases for 15 taxa), STELAR produced more accurate species trees with rootings other than the known outgroup. We further investigated the performance of different methods in terms of the triplet and quartet scores. Both of them are statistically significant criteria, but under practical model conditions with limited numbers of genes with gene tree estimation errors, quartet scores tend to be more correlated with the RF score than triplet scores.

Given the mixed performance of various rooting techniques, relative performance on finite data clearly depends on the model conditions. Hence, we do not make any general recommendation in favor of one type of rooting method or another. These observations highlight the complexity and interplay between different methods and conditions in species tree estimation. The choice of rooting technique and the approach to summarizing gene tree information significantly influence species tree estimation accuracy, underscoring the importance of methodological considerations in phylogenetic studies.

Competing interests

The authors declare that they have no competing interests.

References

[1] Sebastien 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, 2015.

[2] L S Kubatko and J H Degnan. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst Biol*, 56:17, 2007.

[3] Scott V Edwards, Liang Liu, and Dennis K Pearl. High-resolution species trees without concatenation. *Proceedings of the National Academy of Sciences*, 104(14):5936–5941, 2007.

[4] A D Leaché and B Rannala. The accuracy of species tree estimation under simulation: a comparison of methods. *Syst Biol*, 60(2):126–137, 2011.

[5] Michael DeGiorgio and James H Degnan. Fast and consistent estimation of species trees using supermatrix rooted triples. *Molecular biology and evolution*, 27(3):552–569, 2009.

[6] Md Shamsuzzoha Bayzid and Tandy Warnow. Naive binning improves phylogenomic analyses. *Bioinformatics*, 29(18):2277–2284, 2013.

[7] J F C Kingman. The coalescent. *Stoch Proc Appl*, 13:235–248, 1982.

[8] J Heled and A J Drummond. Bayesian inference of species trees from multilocus data. *Mol Biol Evol*, 27:570–580, 2010.

[9] L S Kubatko, B C Carstens, and L L Knowles. Stem: Species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinf*, 25:971–973, 2009.

[10] B Larget, S K Kotha, C N Dewey, and C Ané. BUCKy: Gene tree/species tree reconciliation with the Bayesian concordance analysis. *Bioinf*, 26(22):2910–2911, 2010.

[11] Rezwana Reaz, Md Shamsuzzoha Bayzid, and M Sohel Rahman. Accurate phylogenetic tree reconstruction from quartets: A heuristic approach. *PLoS One*, 9(8):e104008, 2014.

[12] 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.

[13] Liang Liu, Lili Yu, and Scott V Edwards. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolutinary Biology*, 10:302, 2010.

[14] Julia Chifman and Laura Kubatko. Quartet from snp data under the coalescent model. *Bioinformatics*, 30(23):3317–3324, 2014.

[15] Pranjal Vachaspati and Tandy Warnow. Astrid: accurate species trees from internode distances. *BMC genomics*, 16(10):S3, 2015.

[16] Mazharul Islam, Kowshika Sarker, Trisha Das, Rezwana Reaz, and Md Shamsuzzoha Bayzid. Stelar: A statistically consistent coalescent-based species tree estimation method by maximizing triplet consistency. *BMC genomics*, 21(1):1–13, 2020.

[17] J H Degnan and N A Rosenberg. Discordance of species trees with their most likely gene trees. *PLoS Genetics*, 2:762 – 768, 2006.

[18] J H Degnan and N A Rosenberg. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecology Evolution*, 26(6), 2009.

[19] James H Degnan. Anomalous unrooted gene trees. *Systematic biology*, 62(4):574–590, 2013.

[20] Siavash Mirarab, Rezwana Reaz, Md S Bayzid, Théo Zimmermann, M Shel Swenson, and Tandy Warnow. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*, 30(17):i541–i548, 2014.

[21] H. Philippe and P. Forterre. The rooting of the universal tree of life is not reliable. *Journal of Molecular Evolution*, 49(4):509–523, 1999.

[22] S. W. Graham, R. G. Olmstead, and S. CH Barrett. Rooting phylogenetic trees with distant outgroups: a case study from the commelinoid monocots. *Molecular Biology and Evolution*, 19(10):1769–1781, 2002.

[23] MO Dayhoff and RM Schwartz. Prokaryote evolution and the symbiotic origin of eukaryotes. *Endocytobiology: endosymbiosis and cell biology: a synthesis of recent research*, 1:63–84, 1980.

[24] Maria C Rivera and James A Lake. Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. *Science*, 257(5066):74–76, 1992.

[25] Elizabeth S Allman, James H Degnan, and John A Rhodes. Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. *Journal of mathematical biology*, 62:833–862, 2011.

[26] B. Boussau, G. J. Szöllősi, L. Duret, M. Gouy, E. Tannier, and V. Daubin. Genome-scale coestimation of species and gene trees. *Genome Research*, 23(2):323–330, 2013.

[27] Stamatakis A. Bettisworth B. Root digger: a root placement program for phylogenetic trees. *BMC Bioinformatics*, 22(225), 2021.

[28] Yasamin Tabatabaee, Kowshika Sarker, and Tandy Warnow. Quintet rooting: rooting species trees under the multi-species coalescent model. *Bioinformatics*, 38(Suppl 1):i109–i117, 2022. Supplementary data available at Bioinformatics online.

[29] J. P. Huelsenbeck, J. P. Bollback, and A. M. Levine. Inferring the root of a phylogenetic tree. *Systematic Biology*, 51(1):32–43, 2002.

[30] Larry E Watrous and Quentin D Wheeler. The out-group comparison method of character analysis. *Systematic Biology*, 30(1):1–11, 1981.

[31] Wayne P Maddison, Michael J Donoghue, and David R Maddison. Outgroup analysis and parsimony. *Systematic biology*, 33(1):83–103, 1984.

[32] Andrew B Smith. Rooting molecular trees: problems and strategies. *Biological Journal of the Linnean Society*, 51(3):279–292, 1994.

[33] D. L. Swofford, G. J. Olson, P. J. Waddell, and D. M. Hillis. *Phylogenetic Inference*, pages 407–425. Sinauer Associates, Sunderland, Massachusetts, 2nd edition edition, 1996.

[34] Michel C Milinkovitch and James Lyons-Weiler. Finding optimal ingroup topologies and convexities when the choice of outgroups is not obvious. *Molecular Phylogenetics and Evolution*, 9(3):348–357, 1998.

[35] James Lyons-Weiler, Guy A Hoelzer, and Robin J Tausch. Optimal outgroup analysis. *Biological Journal of the Linnean Society*, 64(4):493–511, 1998.

[36] Alexei J Drummond, Simon Y W Ho, Matthew J Phillips, and Andrew Rambaut. Relaxed phylogenetics and dating with confidence. *PLoS biology*, 4(5):e88, 2006.

[37] Fernando Domingues Kümmel Tria, Giddy Landan, and Tal Dagan. Phylogenetic rooting using minimal ancestor deviation. *Nature Ecology Evolution*, 1(7):0193, 2017.

[38] James S. Farris. Estimating phylogenetic trees from distance matrices. *The American Naturalist*, 106(951):645–668, 1972.

[39] Uyen Mai, Erfan Sayyari, and Siavash Mirarab. Minimum variance rooting of phylogenetic trees and implications for species tree reconstruction. *PLOS ONE*, 12(8):e0182238, 2017.

[40] J. P. Gogarten, H. Kibak, P. Dittrich, L. Taiz, E. J. Bowman, B. J. Bowman, M. F. Manolson, R. J. Poole, T. Date, T. Oshima, J. Konishi, K. Denda, and M. Yoshida. Evolution of the vacuolar h^+-atpase: implications for the origin of eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, 86(17):6661–6665, Sep 1989.

[41] Naoyuki Iwabe, Kei-ichi Kuma, Masami Hasegawa, Syozo Osawa, and Takashi Miyata. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proceedings of the National Academy of Sciences*, 86(23):9355–9359, 1989.

[42] Sandra L Baldauf and Jeffrey D Palmer. Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins. *Proceedings of the National Academy of Sciences*, 90(24):11558–11562, 1993.

[43] James A Lake, Craig W Herbold, Maria C Rivera, Jacqueline A Servin, and Ryan G Skophammer. Rooting the tree of life using nonubiquitous genes. *Molecular biology and evolution*, 24(1):130–136, 2007.

[44] Ward C. Wheeler. Nucleic acid sequence phylogeny and random outgroups. *Cladistics*, 6(4):363–367, 1990.

[45] PABLO N. HESS and CLAUDIA A. DE MORAES RUSSO. An empirical test of the midpoint rooting method. *Biological Journal of the Linnean Society*, 92(4):669–674, 12 2007.

[46] Rosa Tarrıo, Francisco Rodrıguez-Trelles, and Francisco J. Ayala. Tree rooting with outgroups when they differ in their nucleotide composition from the ingroup: The drosophila saltans and willistoni groups, a case study. *Molecular Phylogenetics and Evolution*, 16(3):344–349, 2000.

[47] Tom A Williams, Peter G Foster, Cymon J Cox, and T Martin Embley. Rooting the eukaryotic tree of life. *Current Biology*, 24(4):R151–R152, 2014.

[48] Laura M Boykin, Charles D Bell, Gregory Evans, Ian Small, and Paul J De Barro. Is agriculture driving the diversification of the bemisia tabaci species complex (hemiptera: Sternorrhyncha: Aleyrodidae)?: Dating, diversification and biogeographic evidence revealed. *BMC Evolutionary Biology*, 13:228, 2013.

[49] Yin-Long Qiu, Jungho Lee, Barbara A. Whitlock, Fabiana Bernasconi-Quadroni, and Olena Dombrovska. Was the ANITA Rooting of the Angiosperm Phylogeny Affected by Long-Branch Attraction? *Molecular Biology and Evolution*, 18(9):1745–1753, 09 2001.

[50] W. C. Wheeler and D. S. Gladstein. MALIGN: A Multiple Sequence Alignment Program. *J Hered*, 85(5):417–418, 1994.

[51] Michael D. Hendy and David Penny. A Framework for the Quantitative Study of Evolutionary Trees. *Systematic Biology*, 38(4):297–309, 12 1989.

[52] Omar Rota-Stabelli and Maximilian J. Telford. A multi criterion approach for the selection of optimal outgroups in phylogeny: Recovering some support for mandibulata over myriochelata using mitogenomics. *Molecular Phylogenetics and Evolution*, 48(1):103–111, 2008.

[53] John Gatesy, Rob DeSalle, and Niklas Wahlberg. How many genes should a systematist sample? conflicting insights from a phylogenomic matrix characterized by replicated incongruence. *Systematic Biology*, 56(2):355–363, 2007.

[54] BR Holland, D Penny, and MD Hendy. Outgroup misplacement and phylogenetic inaccuracy under a molecular clock—a simulation study. *Systematic biology*, 52(2):229–238, 2003.

[55] SWOFFORD D. L. Phylogenetic inference. *Molecular Systematics, 2nd ed.*, pages 407–514, 1996.

[56] John Stavrinides and David S Guttman. Mosaic evolution of the severe acute respiratory syndrome coronavirus. *Journal of virology*, 78(1):76–82, 2004.

[57] Jan P Meier-Kolthoff and Markus Göker. VICTOR: genome-based phylogeny and classification of prokaryotic viruses. *Bioinformatics*, 33(21):3396–3404, 07 2017.

[58] Nicky Wybouw, Wannes Dermauw, Luc Tirry, Christian Stevens, Miodrag Grbić, René Feyereisen, and Thomas Van Leeuwen. A gene horizontally transferred from bacteria protects arthropods from host plant cyanide poisoning. *eLife*, 3:e02365, apr 2014.

[59] Fumio Tajima. Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics*, 135(2):599–607, 1993.

[60] Panagiotis S Adam, George E Kolyfetis, Till LV Bornemann, Constantinos E Vorgias, and Alexander J Probst. Genomic remnants of ancestral methanogenesis and hydrogenotrophy in archaea drive anaerobic carbon cycling. *Science advances*, 8(44):eabm9651, 2022.

[61] Andrey Rozenberg, Igor Kaczmarczyk, Donna Matzov, Johannes Vierock, Takashi Nagata, Masahiro Sugiura, Kota Katayama, Yuma Kawasaki, Masae Konno, Yujiro Nagasaka, et al. Rhodopsin-bestrophin fusion proteins from unicellular algae form gigantic pentameric ion channels. *Nature structural & molecular biology*, 29(6):592–603, 2022.

[62] 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, 2014.

[63] Erich D Jarvis, Siavash Mirarab, Andre J Aberer, Bo Li, Peter Houde, Cai Li, Simon YW Ho, Brant C Faircloth, Benoit Nabholz, Jason T Howard, et al. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, 346(6215):1320–1331, 2014.

[64] Sen Song, Liang Liu, Scott V Edwards, and Shaoyuan Wu. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proceedings of the National Academy of Sciences*, 109(37):14942–14947, 2012.

[65] Md Shamsuzzoha Bayzid, Siavash Mirarab, Bastien Boussau, and Tandy Warnow. Weighted statistical binning: enabling statistically consistent genome-scale phylogenetic analyses. *PLoS One*, 10(6), 2015.

[66] 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, 2015.

[67] Y Chung and C Ané. Comparing two Bayesian methods for gene tree/species tree reconstruction: A simulation with incomplete lineage sorting and horizontal gene transfer. *Syst Biol*, 60(3):261–275, 2011.

[68] Zhenxiang Xi, Liang Liu, Joshua S Rest, and Charles C Davis. Coalescent versus concatenation methods and the placement of amborella as sister to water lilies. *Systematic biology*, 63(6):919–932, 2014.

[69] The Angiosperm Phylogeny Group, M. W. Chase, M. J. M. Christenhusz, M. F. Fay, J. W. Byng, W. S. Judd, D. E. Soltis, D. J. Mabberley, A. N. Sennikov, P. S. Soltis, and P. F. Stevens. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Botanical Journal of the Linnean Society*, 181(1):1–20, 04 2016.

[70] Zhenxiang Xi, Joshua S. Rest, and Charles C. Davis. Phylogenomics and coalescent analyses resolve extant seed plant relationships. *PLoS One*, 8(11):e80870, 2013.

[71] Ylenia Chiari, Vincent Cahais, Nicolas Galtier, and Frédéric Delsuc. Phylogenomic analyses support the position of turtles as the sister group of birds and crocodiles (archosauria). *Bmc Biology*, 10(1):65, 2012.

[72] T Kinene, J Wainaina, S Maina, and LM Boykin. Rooting Trees, Methods for. *Encyclopedia of Evolutionary Biology*, pages 489–93, 2016.

[73] D.F. Robinson and L.R. Foulds. Comparison of phylogenetic trees. *Math. Biosci.*, 53:131–147, 1981.