 doi:10.20944/preprints202110.0258.v2

Recent Progress on Methods for Estimating and Updating Large Phylogenies

Paul Zaharias*and Tandy Warnow†

Department of Computer Science
University of Illinois Urbana-Champaign
Urbana, IL 61801

Abstract

With the increased availability of sequence data and even of fully sequenced and assembled genomes, phylogeny estimation of very large trees (even of hundreds of thousands of sequences) is now a goal for some biologists. Yet, the construction of these phylogenies is a complex pipeline presenting analytical and computational challenges, especially when the number of sequences is very large. In the last few years, new methods have been developed that aim to enable highly accurate phylogeny estimations on these large datasets, including divide-and-conquer techniques for multiple sequence alignment and/or tree estimation, methods that can estimate species trees from multi-locus datasets while addressing heterogeneity due to biological processes (e.g., incomplete lineage sorting and gene duplication and loss), and methods to add sequences into large gene trees or species trees. Here we present some of these recent advances and discuss opportunities for future improvements.

Keywords— phylogeney estimation, multiple sequence alignment, phylogenetic placement, phylogenomics, taxon identification, maximum likelihood

1 Introduction

Large-scale phylogeny estimation presents substantial computational and statistical challenges: the most accurate methods are often likelihood-based methods (Maximum Likelihood or Bayesian Inference) that can use substantial time and memory to produce reliable trees. Multiple sequence alignment (a precursor to phylogeny estimation) is also challenging, especially on large datasets that have high rates of evolution. Furthermore, species tree estimation presents additional challenges due to heterogeneity in phylogenetic trees between different loci, which can result from processes such as incomplete lineage sorting (ILS), gene duplication and loss (GDL), and horizontal gene transfer (HGT) (Maddison, 1997). Yet because dense taxonomic sampling has been seen to improve phylogenetic accuracy (Nabhan and Sarkar, 2012), the interest in statistically rigorous methods for large-scale phylogeny estimation (whether of gene trees or species trees) has not abated.

The last decade has produced methods for alignment and phylogeny estimation that have excellent accuracy on small to moderate-sized datasets, but only a few of these methods can analyze even moderately large datasets (1,000 sequences). Some of the methods with the best scalability are distance-based (e.g., FastME (Lefort et al., 2015)). However, studies (e.g., Lees et al. (2018)) comparing methods based on maximum likelihood to distance-based approaches have observed that maximum likelihood methods tend to be more accurate on large datasets.

Because maximum likelihood methods can be computationally intensive (both for time and memory), substantial effort has been made to improve the running time through careful implementation of the numerical calculations and use of parallelism (see recent surveys in Bader and Madduri (2019); Guindon and Gascuel (2019); Stamatakis (2019)). Despite the advances in the last decade, the construction of very large maximum likelihood phylogenies (e.g., gene phylogenies of 100,000 or more sequences or 10,000 whole genomes) is very difficult using standard approaches, except perhaps when supercomputers are available.

Divide-and-conquer is a natural technique to speed up computationally intensive analyses: for example, rather than estimating a tree on a dataset with 100,000 sequences, the input could be divided into many smaller datasets (perhaps 100 datasets with approximately 1000 sequences each), trees could be estimated on each subset, and then combined into a tree on the entire dataset. An obvious divide-and-conquer technique would use taxonomic information to define the

*zaharias@illinois.edu
†warnow@illinois.edu

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2022 doi:10.20944/preprints202110.0258.v2

subsets; however, using taxonomies presents potentially significant challenges. For example, when estimating gene trees, discordance between gene trees and species trees (resulting from various biological processes) can mean that taxonomically-derived decompositions do not form connected subtrees in the true gene trees. An additional complication that impacts all estimation problems is that taxonomies can have mistakes; as a result, techniques that use taxonomic information are often combined with opportunities for the user to correct potential mistakes. Finally, taxonomies may not include all the sequences in the input. Despite the challenges in using taxonomies, they can be very useful in constraining the search space, and so result in reduced running time. PyPHLAWD (Smith and Walker, 2019) and PhyLoTA (Sanderson et al., 2008) are two such techniques, and strategies like these have been used in phylogenomic analyses (e.g., Asnicar et al. (2020); Janssens et al. (2020)).

In this paper we present new divide-and-conquer techniques to scale computationally intensive but highly accurate methods to large and even ultra-large datasets, without using taxonomic information. We show how divide-and-conquer can improve large-scale multiple sequence alignment (a precursor to phylogeny estimation), maximum likelihood tree estimation, species tree estimation without requiring orthology detection, and phylogenetic placement methods (e.g., adding new sequences or species to a given phylogeny) that can be used to update a large phylogeny or taxonomically characterize new sequences (e.g., in a microbiome analysis). Thus, while this survey is relevant to microbial phylogenetics and biodiversity assessment, all large-scale systematics research presents similar challenges. These techniques reduce the computational effort compared to traditional methods, and so reduce the need for supercomputers or high-performance computing environments while providing very high accuracy.

2 Recent Advances in Multiple Sequence Alignment

Multiple sequence alignment is a precursor to phylogeny estimation as well as to other bioinformatics problems, such as sequence classification and protein function prediction. When the input is a set of sequences for a group of closely related individuals, then techniques that operate by inferring pairwise alignments to a single reference sequence can have good accuracy; however, the estimation of multiple sequence alignments for more distantly related sequences requires other techniques. There are many well established methods (surveyed in Katoh (2021)), but only some of these provide good accuracy on large sequence datasets, especially when they have evolved under high rates of evolution.

Divide-and-conquer techniques have been very powerful tools in scaling the most accurate alignment methods to large datasets. These methods (e.g., Smith et al. (2009); Liu et al. (2009); Mirarab et al. (2015); Smirnov and Warnow (2021a); Smirnov (2021)) divide the input sequence dataset into disjoint subsets, produce alignments on each subset using a selected "base method" and then merge the subset alignments together. Two of these methods, PASTA (Mirarab et al., 2015) and recursive MAGUS (Smirnov, 2021), can be used to produce highly accurate alignments of datasets with up to 1,000,000 sequences. When combined with iteration (so that each iteration uses the previous iteration's alignment to compute a new tree and then decomposes the dataset using the tree), the methods can produce highly accurate alignments and trees, typically in just a few iterations. MAFFT (Katoh and Standley, 2013) is the default method for subset alignment for many of these pipelines, but these pipelines have been studied with other methods and found that they improved accuracy and/or reduced running time when analyzing large datasets. For example, using BAli-Phy (Redelings and Suchard, 2005) (a Bayesian method for co-estimation of alignments and trees) within PASTA has been able to produce highly accurate alignments on datasets with 1,000 sequences (Nute and Warnow, 2016).

A new and promising divide-and-conquer strategy is used in MAGUS (Smirnov and Warnow, 2021a), a recently developed MSA method that is closely related to PASTA. Specifically, whereas PASTA merges a set of disjoint alignments by merging selected pairs of alignments and then using transitivity to complete the merger, MAGUS achieves the merger by first computing a graph where the vertices represent the sites in the alignments, and then clustering the sites together to define the merged alignment. This clustering step, performed using the Graph Clustering Merger (described in Smirnov and Warnow (2021a)) is the key to the improved accuracy that MAGUS has over PASTA, as all other algorithmic differences between MAGUS and PASTA are very minor. As demonstrated in Zaharias et al. (2021), the Graph Clustering Merger is an effective strategy for solving the Maximum Weight Trace problem (Kececioglu, 1993) in the context of merging alignments. The recursive version of MAGUS (Smirnov, 2021) is able to align very large datasets with high accuracy (up to 1,000,000 sequences so far). As shown in Smirnov (2021), MAGUS and its recursive version are more accurate than leading alignment methods on large biological benchmark datasets and simulated datasets (up to 1,000,000 sequences). Here, alignment error is based on pairwise homology statements for each alignment, where two letters that are in the same column of an alignment are considered homologous according to that alignment. The fraction of the pairwise homologies (defined by the reference alignment) that are not in the estimated alignment is the sum-of-pairs false negative (SPFN) error rate, and the fraction of the pairwise homologies in the estimated alignment that are not in the reference alignment is the sum-of-pairs false positives (SPFP) error rate. Figure 1 from Smirnov (2021) demonstrates that each of the three variants of MAGUS produces more accurate alignments than leading alignment methods on HomFam benchmark datasets with 10,099 to 98,681 sequences.

Potential limitations of MAGUS. Previous studies (Smirnov and Warnow, 2021a; Smirnov, 2021) have established that MAGUS produces highly accurate alignments in comparison to other methods, such as PASTA, MAFFT, Clustal-Omega (Sievers et al., 2011), Muscle (Edgar, 2004), and UPP (Nguyen et al., 2015b), on both biological and simulated datasets. Although the explored model conditions have varied in terms of type of data (i.e. nucleotides and proteins), rate of evolution,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2022 doi:10.20944/preprints202110.0258.v2

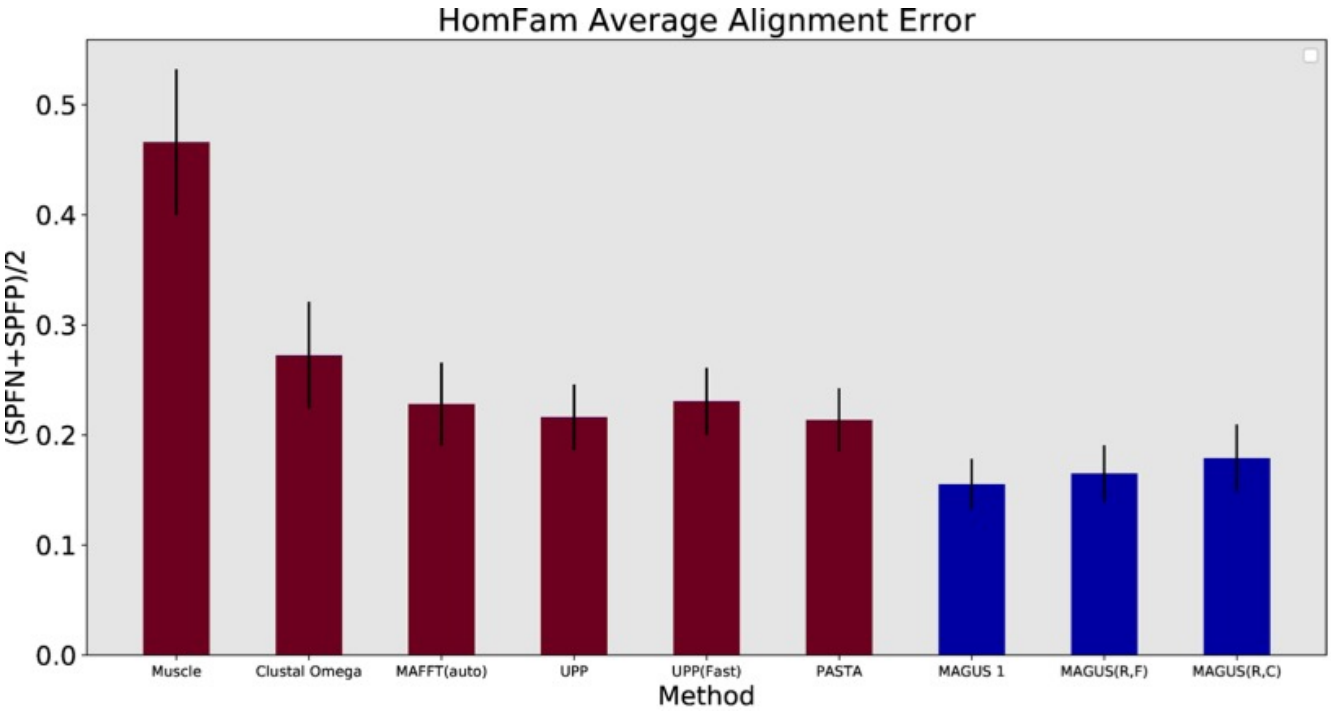

Figure 1: **Average alignment error on 19 HomFam datasets with 10,099 to 93,681 sequences.** Alignment error rate SPFN is the fraction of the reference pairwise homologies missing in the estimated alignment and SPFP is the fraction of the inferred pairwise homologies that are not in the reference alignment. Results are averaged over the datasets where all methods completed (Muscle segfaulted on two). Error bars show standard error. Reproduced from Smirnov (2021) under the Creative Commons Attribution License.

and dataset size, they have not fully explored conditions with substantial sequence length heterogeneity (i.e., where the input sequences vary substantially in length). One type of sequence length heterogeneity is when most of the input sequences are full-length but the rest of the sequences are fragmentary, a condition that arises when the input sequences include reads or other incompletely assembled sequences. UPP is an MSA method that is specifically designed for datasets with this type of sequence length heterogeneity, and it has produced highly accurate alignments on large datasets with up to 1,000,000 sequences (Nguyen et al., 2015b). A recent study (Shen et al., 2021) showed that MAGUS is less accurate than UPP when there are fragmentary sequences, but the combination of MAGUS and UPP is more accurate than either MAGUS or UPP on such datasets. What is not yet known is how well MAGUS, UPP, or their combination perform when aligning datasets that exhibit other patterns of sequence length heterogeneity (e.g., where there are some very long sequences, or where the dataset evolved with very large insertions or deletions), and whether there are other methods already developed that provide better accuracy under such conditions. In general, multiple sequence alignment of datasets with sequence length heterogeneity is not well studied, and so this issue impacts multiple sequence alignment method development more generally.

3 Recent Advances in Maximum Likelihood Tree Estimation

Maximum likelihood (ML) gene tree estimation is one of the core problems in phylogeny estimation. Finding the optimal maximum likelihood tree is NP-hard (Roch, 2006) and so the best heuristics, such as RAxML (Stamatakis, 2014) and IQ-TREE (Nguyen et al., 2015a), use many different strategies to search for the tree optimizing the likelihood score. FastTree 2 (Price et al., 2010) is a very fast heuristic that does not make a very substantial attempt to optimize likelihood (and hence does not find very good maximum likelihood scores), but can (in some cases) be comparable to RAxML with respect to topological accuracy on simulated datasets (Liu et al., 2011).

RAxML has been modified over the years to improve scalability to large datasets, and the current version, RAxML-ng (Kozlov et al., 2019), is able to analyze very large datasets. However, Park et al. (2021) showed that RAxML-ng, using 16 CPUs, did not converge on a 10,000-sequence dataset even after a week. In contrast, Price et al. (2010) showed that FastTree 2 was able to estimate an ML tree with 237,882 distinct sequences in 22 hours. Smirnov (2021) benchmarked FastTree 2 on a million-sequence dataset, and showed that FastTree 2 produced an ML tree in about 5 days using 32 CPUs.

Although FastTree 2 clearly dominates RAxML for speed and memory usage and can be comparable in topological accuracy, several studies have shown that FastTree 2 can have reduced topological accuracy when the input alignment

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2022 doi:10.20944/preprints202110.0258.v2

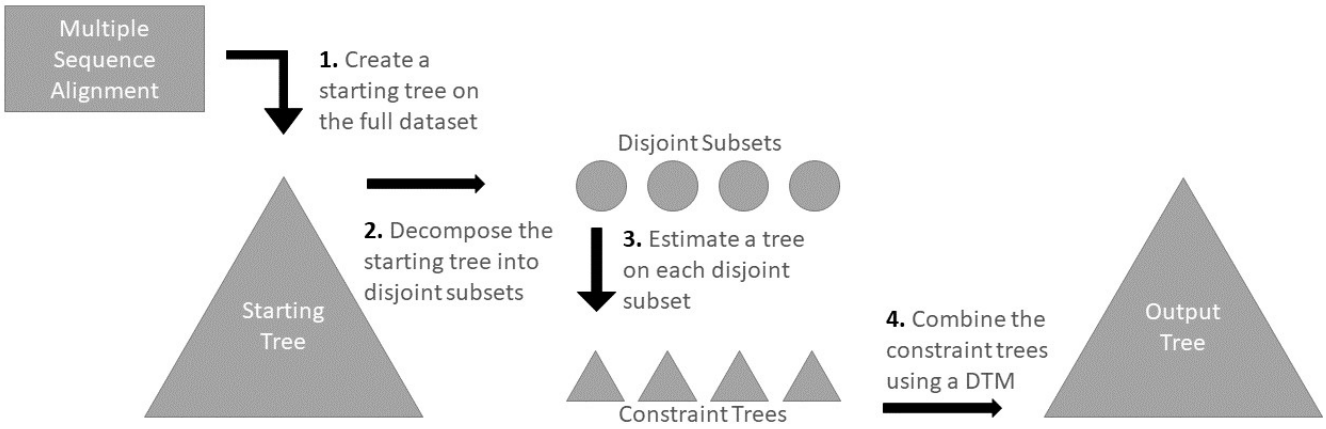

Figure 2: **DTM Pipeline for constructing a tree from an input sequence alignment using maximum likelihood**. (1) A starting tree is computed (e.g., using FastTree 2 or IQ-TREE 2). (2) Edges are deleted from the starting tree to produce small subsets. (3) Trees are estimated on the subsets using a selected maximum likelihood method (e.g., IQ-TREE 2 or RAxML-ng). (4) The selected DTM method merges the disjoint trees into a tree on the full dataset. DTM pipelines that operate from multi-locus inputs and compute species trees have also been developed, with suitable adjustments to the algorithmic steps. Reproduced from Park et al. (2021) under the Creative Commons Attribution License.

contains many fragmentary sequences (Park et al., 2021; Smirnov and Warnow, 2021b) or is otherwise very gappy (Sayyari et al., 2017); further, a recent study showed reduced accuracy for FastTree 2 when the sequences have evolved under heterotachy (Park et al., 2021). In contrast, RAxML and to a somewhat lesser extent also IQ-TREE 2 seem more robust to those conditions (Park et al., 2021). These recent studies indicate the importance of simulating sequences under models that are more complex than the reconstruction model. Further research is needed to assess the degree to which maximum likelihood under standard models, such as the Generalized Time Reversible (GTR) model (Tavaré, 1986), is robust to model violations, such as heterogeneity of rates across sites (Yang, 1994), heterogeneity of substitution processes across sites (Lartillot and Philippe, 2004), compositional heterogeneity (Foster and Hickey, 1999), heterotachy (Lopez et al., 2002), heteropecilly (Roure and Philippe, 2011), selection, and other processes where the evolutionary processes are non-i.i.d.

Several strategies have been developed to overcome the burden of computationally intensive maximum likelihood analyses. Some of these (e.g., DACTAL (Nelesen et al., 2012)) operate by dividing the input set into overlapping subsets, constructing trees on the subsets, and then using supertree methods to merge the subset trees into a tree on the full dataset. This is a natural approach to large-scale tree estimation (Bininda-Emonds, 2004), but the choice of decomposition strategy can impact the final accuracy, and random decompositions in particular can produce poor supertrees (Roshan et al., 2004). Furthermore, the requirement to use supertree methods (which are not yet very fast) constrains the scalability of these approaches (Warnow, 2019).

To overcome these limitations, a new type of divide-and-conquer approach has been developed. In this approach (see Figure 2), an initial tree is computed on the input. Then edges are deleted from the tree until each subset is small enough (below a user-provided threshold). Then trees are estimated on each subset, and finally merged into a tree on the full dataset. This four-stage approach divides the input dataset into disjoint rather than overlapping sets, and hence requires additional information, such as a distance matrix or a guide tree, in order to merge the subset trees into a full tree.

The problem of merging a set of leaf-disjoint trees into a single tree is called "Disjoint Tree Merging", and methods that perform that operation are referred to as "Disjoint Tree Mergers" (DTMs). Pipelines that use DTMs can be used to estimate both gene trees (in which case the subset trees can be computed using maximum likelihood) and species trees from multi-locus datasets. Provided that the DTM is designed carefully, these pipelines can be proven to be statistically consistent for both types of analyses.

Several DTMs have been developed, including NJMerge (Molloy and Warnow, 2019a), TreeMerge (Molloy and Warnow, 2019b), Constrained-INC (Le et al., 2020; Zhang et al., 2020a), and the Guide Tree Merger (Smirnov and Warnow, 2020). Of these, the Guide Tree Merger (GTM) has been shown to be very fast and generally as accurate as the other DTMs. The auxiliary information for GTM is a guide tree T, which it uses to merge the disjoint subset trees. Specifically, GTM adds edges to link up the disjoint subset trees so as to produce a merged tree T^*, and does this while guaranteeing that T^* has the minimum topological distance to the guide tree T; thus, T^* has the largest number of shared edges to T among all such merged trees. GTM performs this merger exactly in polynomial time. Hence, when the initial tree and the subset trees are all estimated using statistically consistent methods, then pipelines using GTM are provably statistically consistent.

Figure 3 shows results from Park et al. (2021), comparing a DTM pipeline using GTM to two leading maximum likelihood methods (RAxML-ng and IQ-TREE 2). For topological error, we report the false negative error rate, which indicates the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2022 doi:10.20944/preprints202110.0258.v2

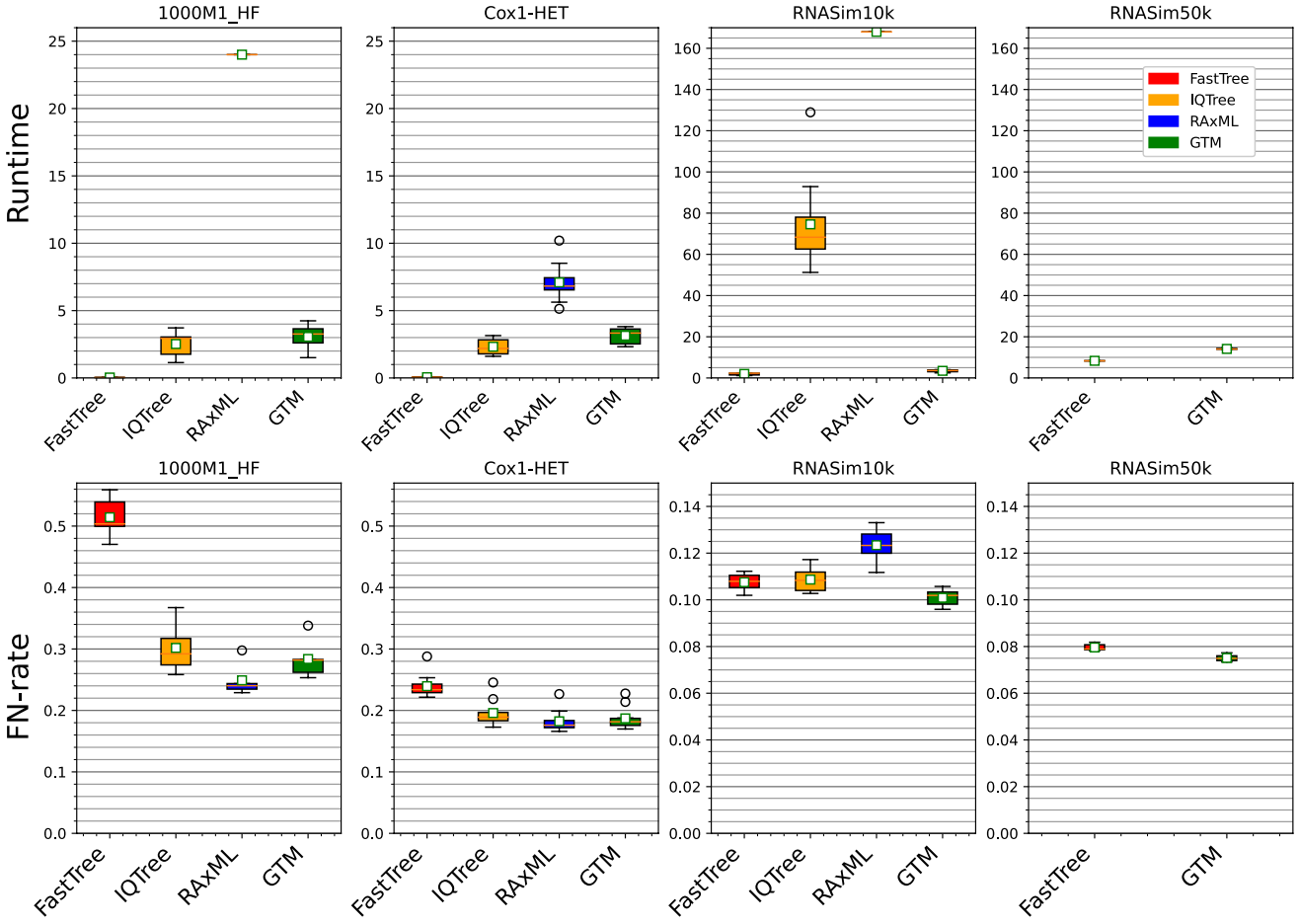

Figure 3: We compare standard maximum likelihood methods (RAxML-ng, IQ-TREE 2, and FastTree 2) to a divide-and-conquer pipeline using the Guide Tree Merger (GTM) on four simulated datasets with 1,000 to 50,000 sequences. 1000M1-HF datasets each have 1,000 sequences that evolved under a GTRGAMMA+indel model and includes fragmentary sequences, Cox1-HET datasets each have 2341 sequences that evolved with heterotachy, and the RNASim (Mirarab et al., 2015) datasets have 10,000 to 50,000 sequences each and evolved under a non-i.i.d. model with indels and under selective pressures to maintain the RNA secondary structure. Top: running time (hrs), bottom: missing branch (FN) error rates across 10 replicates per model condition. Results not shown for IQ-TREE 2 and RAxML on the RNASim 50K dataset are because IQ-TREE 2 failed to return a tree within the allowed time (24 hrs for the two smaller datasets and 168 hrs for the two larger datasets) and RAxML-ng produced trees with at least 99.96% FN error. Adapted from Park et al. (2021) under the Creative Commons Attribution License.

proportion of the non-trivial splits in the true tree that are not produced in the estimated tree. The GTM pipeline matches or improves on the topological accuracy of both IQ-TREE 2 and FastTree 2 and is competitive with RAxML-ng, while being much faster than RAxML-ng. A comparison on the largest dataset with 50,000 sequences, limited to 168 hours (1 week) of analysis, shows that only the GTM pipeline and FastTree 2 are acceptable: RAxML-ng has at least 99.96% false negative error on that model condition while IQ-TREE 2 fails to return a tree at all due to memory issues.

It is also worth noting that in this figure, only the 1000M1-HF model condition (which has fragmentary sequences) has substitutions modelled by a GTRGAMMA model, which is an assumption made by all the maximum likelihood methods. Thus, this figure demonstrates that RAxML-ng and IQ-TREE 2 exhibit a fair amount of robustness to model violations; this robustness is also true for FastTree 2, but to a lesser extent. This trend is also consistent with a study comparing deep neural network methods (DNNs) to standard protein maximum likelihood methods on datasets that evolve with very high levels of heterogeneity, and which showed that standard maximum likelihood methods, even under simple protein evolution models, were also very accurate, and typically more accurate than the DNNs (Zaharias et al., 2022).

Potential limitations of DTM pipelines. The results shown in Figure 3 show that the GTM pipeline had some advantages over RAxML-ng and IQ-TREE 2, two leading maximum likelihood heuristics, on the largest datasets (with 10,000 or 50,000

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2022 doi:10.20944/preprints202110.0258.v2

sequences), and was comparable to these methods on the other datasets. Even so, it is premature to generalize and assume that GTM pipelines will continue to provide competitive performance under other conditions. Moreover, the main advantage of the GTM pipeline over IQ-TREE 2 and RAxML-ng occurs on the two largest datasets, and most notably on the largest dataset (with 50,000 sequences) where IQ-TREE 2 fails due to memory issues, indicating that the memory available (64 Gb) was insufficient, and RAxML-ng produces trees that have at least 99.96% topological error. To understand this performance, we note that Park et al. (2021) used RAxML-ng in default mode (10 random starting trees and 10 random addition parsimony trees), and the result shown is consistent with RAxML-ng completing only a few rounds of heuristic search and so returning a tree that is close to the starting tree, as large random trees are expected to have very few bipartitions in common with the true tree (i.e., the probability of a shared bipartition converges to 0 with the number of leaves (Steel and Penny, 1993)). It is possible that RAxML-ng might have been able to produce a good tree on this dataset using a different starting tree (e.g., using FastTree 2). Thus, while Park et al. (2021) does show advantages to using a GTM-pipeline for large-scale maximum likelihood compared to both IQ-TREE 2 and RAxML-ng, future work is needed to better understand how to use RAxML-ng and IQ-TREE 2 to estimate ultra-large trees without requiring very large amounts of memory or time.

A limitation that is specific to the use of GTM (rather than other DTMs) to merge constraint trees is that GTM by design combines subset trees by adding edges that connect the subset trees; this means that the subset trees are not allowed to "blend", a restriction that inherently can limit the accuracy of the final tree, making it very dependent on the initial tree used to define the dataset decomposition. Thus, GTM pipelines need to be evaluated under conditions where the initial trees that are computed and then used for decomposition have poor accuracy. In such cases, other DTM pipelines that allow blending (e.g., Constrained-INC) may provide better accuracy and offer advantages over the best current ML heuristics, but this needs to be evaluated.

4 Recent Advances in Species Tree Estimation

A traditional approach to multi-locus species tree estimation concatenates the individual gene sequence alignments into a "supermatrix" and estimates a tree on the supermatrix, often using maximum likelihood. These "concatenation analyses" are appealing but can be very computationally expensive: the maximum likelihood analysis of the 48 bird genomes in Jarvis et al. (2014) took 250 CPU years, and the maximum likelihood concatenation pipeline of Zhu et al. (2019b) took ∼33,000 CPU hours (about 3.8 CPU years) to build a tree on 10,575 genomes. In addition, because different genomic regions can have different evolutionary histories due to processes such as incomplete lineage sorting (ILS) and gene duplication and loss (GDL), the use of concatenation (which assumes that all the sites evolve down a single tree topology) has been significantly criticized (Jiang et al., 2020; Kubatko and Degnan, 2007). As a result, new approaches based on statistical models for gene evolution within species trees have been developed and are now increasingly used, and some of these approaches are very scalable. Here we present recent advances for species tree estimation that provide high accuracy and scalability.

4.1 Species tree estimation in the presence of ILS

The problem of species tree estimation in the presence of ILS is very well studied. Although species trees have traditionally been estimated using maximum likelihood and other methods on a concatenation of the individual gene sequence alignments, this approach has been shown to be statistically inconsistent when there is gene tree heterogeneity due to incomplete lineage sorting (Roch and Steel, 2015).

One of the statistically consistent approaches for species tree estimation when ILS is present operates by estimating gene trees for each gene and then combining the gene trees. These "summary methods" are generally faster than concatenation (especially on large datasets). Two of the best known methods are MP-EST (Liu et al., 2010) and ASTRAL (Mirarab et al., 2014b), but ASTRAL is generally faster on large datasets. ASTRID (Vachaspati and Warnow, 2015) and DISTIQUE are two other fast and scalable summary methods that are often comparable in accuracy to ASTRAL (Sayyari and Mirarab, 2016), but ASTRAL is more frequently used than ASTRID.

ASTRAL constructs an unrooted species tree from a set of unrooted gene trees by solving the "Maximum Quartet Support Supertree" problem (i.e., finding a species tree that agrees with as many quartet trees induced by the input gene trees as possible). Since this is an NP-hard problem, the default setting for ASTRAL solves the problem within a constrained search space that is computed from the input gene trees. Specifically, ASTRAL only considers those candidate species trees that draw their bipartitions from a constraint set that contains the input gene tree bipartitions and potentially some additional bipartitions. ASTRAL uses dynamic programming to solve this constrained search problem exactly, which allows it to be polynomial time on every input. Although it is polynomial time, the worst-case runtime is nearly quadratic in the number of distinct bipartitions found in the constraint set. Since this constraint set can be quite large when there is substantial heterogeneity between gene trees and large numbers of genes, ASTRAL can sometimes take a long time to complete (i.e., days).

In addition to parallelism (Yin et al., 2019), two high-level techniques have been developed to improve ASTRAL's speed. The first is the use of Disjoint Tree Merger pipelines, which greatly reduce the running time for ASTRAL on large taxon sets (Molloy and Warnow, 2019b; Smirnov and Warnow, 2020). The second technique operates by replacing the constraint set that ASTRAL computes from the input with a smaller constraint set. One such approach uses "external constraints", for example partial information about the species tree, in order to reduce the constraint set size (Rabiee and Mirarab, 2020a), an

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2022 doi:10.20944/preprints202110.0258.v2

approach we refer to as "ASTRAL-J" to reflect the flag used in ASTRAL for this case. Another approach runs ASTRID on a collection of subsamples of the gene trees, so that each ASTRID analysis of each subsample produces a candidate species tree. The bipartitions from those estimated trees are then used as the constraint set for ASTRAL. This approach, which is called "FASTRAL" (Dibaeinia et al., 2021), is provably statistically consistent under the multi-species coalescent model, and comparisons on simulated and biological datasets reported in Dibaeinia et al. (2021) show FASTRAL generally is similar in accuracy to ASTRAL while being much faster when the number of species and/or genes is large enough. Finally, FASTRAL-J, a combination of FASTRAL and ASTRAL-J, has been developed that provides runtime advantages over ASTRAL-J and comparable accuracy (Liu and Warnow, 2021).

Potential limitations of ASTRAL and its variants. ASTRAL has been in wide use by biologists, potentially due to its ease of use, generally fast speed (for most datasets), and concordance in many cases with concatenation analyses. Yet there are other summary methods that may be competitive with ASTRAL for accuracy, and some of these are also fast. For example, ASTRID (which is much faster than ASTRAL) is sometimes more accurate than ASTRAL and sometimes less accurate, and it is not clear what the conditions are where each should be used (e.g., see Yan et al. (2021)). Another method that is promising is wQFM (Mahbub et al., 2021), which uses a different technique to combine quartet trees and may provide better accuracy than ASTRAL (albeit at a computational cost). More generally, there are many summary methods available, and while ASTRAL is perhaps the most commonly used summary method for ILS-based species tree estimation, the development of new summary methods could lead to improved accuracy and comparable or better scalability and speed.

One important limitation for ASTRAL and any summary method, for that matter, is the reliance on estimated gene trees. As shown in Molloy and Warnow (2018), ASTRAL and other summary methods are often less accurate than concatenation analyses when gene tree estimation error is high enough, and this inferior accuracy can even be present when there is a high level of ILS. Further, when used with estimated gene trees, there are no guarantees for statistical consistency (Roch et al., 2019), so that using estimated gene trees has both theoretical and empirical consequences. Moreover, phylogenomic datasets can have both properties–high levels of ILS, resulting from rapid speciation, as well as low gene tree accuracy (e.g., see discussion about gene tree resolution in Jarvis et al. (2014)). Thus, the estimation of species trees under conditions with generally poor gene trees can be challenging, if summary methods are used. This is one reason that statistical binning (Mirarab et al., 2014a) and its weighted variant (Bayzid et al., 2015) (methods to re-estimate gene trees by binning gene sequence alignments together using branch support, in order to produce a better input to summary methods) were developed, but these binning techniques are unlikely to provide good results if the gene trees have very low support. More generally, even using statistical binning, there are realistic conditions where all summary methods, are inferior to concatenation analyses and other methods.

Alternative approaches have been developed that avoid the problems with summary methods and also provide statistical guarantees in the presence of ILS. One such example is SVDquartets (Chifman and Kubatko, 2014), a method that uses properties of the multi-species coalescent model to estimate quartet trees and then combines the quartet trees into a tree on the full set of species. SVDquartets (and its variants, e.g. Vachaspati and Warnow (2018)) can provide superior accuracy compared to summary methods under conditions with high gene tree estimation error (Molloy and Warnow, 2018), but more study is needed in order to understand the empirical conditions under which they are more reliable than standard maximum likelihood concatenation analyses. Finally, co-estimation of gene trees and species trees, are also more robust to gene tree estimation error, and methods such as StarBEAST 2 (Ogilvie et al., 2017), can provide outstanding accuracy when they can run. However, these methods are so far limited to small numbers of species and loci (though see Zimmermann et al. (2014).

The variants of ASTRAL we have described here—FASTRAL, ASTRAL-J, and FASTRAL-J–were developed in order to improve the speed for the ASTRAL analyses, not to improve accuracy. Each of these operates by constraining the search space that ASTRAL explores, and so by design will reduce running time but could either improve or reduce accuracy. FASTRAL constrains the search space using bipartitions from ASTRID trees on different subsets of the genes, and demonstrably (and by design) reduces this space and so reduces the runtime. However, FASTRAL has the potential to reduce accuracy if the ASTRID trees constrain the search space too much, removing true bipartitions compared to ASTRAL's default search space. While substantial reductions in accuracy were not observed in Dibaeinia et al. (2021), clearly more extensive explorations are needed, in order to understand the conditions under which FASTRAL can be safely used, without degrading accuracy. A similar statement is true for FASTRAL-J, which also uses FASTRAL to constrain the search space.

We now turn to ASTRAL-J, an approach that uses partial knowledge of the true species tree; ASTRAL-J in other respects uses the same technique as ASTRAL to find a good solution to its optimization problem. Note that by design, this approach cannot reduce accuracy provided that the constraint tree is valid (i.e., has no false bipartitions). The challenge in using ASTRAL-J, therefore, is obtaining this partial knowledge. However, there are many biological datasets where some information is clearly available, and so ASTRAL-J (and hence also FASTRAL-J) may provide benefits to large-scale species tree estimation.

4.2 Species tree estimation in the presence of GDL

Genes can evolve with duplication and loss (GDL), in which case a given organism can have multiple copies of a given gene. As a consequence, the phylogeny for that gene (called a "gene family tree") can have multiple copies of one or more species, and so is called a "MUL-tree" to distinguish it from a single-copy tree.

When estimating a species tree, it is common practice to eliminate those genes that have multiple copies of species (and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2022 doi:10.20944/preprints202110.0258.v2

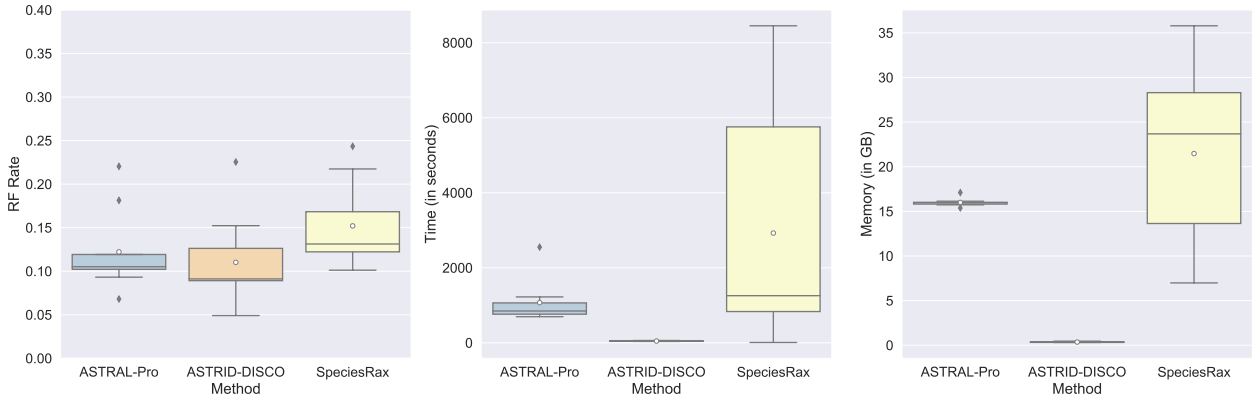

Figure 4: Species tree error (Robinson-Foulds error rates), wall clock running time (s), and peak memory usage of ASTRAL-Pro, SpeciesRax, and ASTRID-DISCO on simulated datasets (evolved under GDL and ILS) of 1001 species and 50 estimated gene trees. All estimated and model trees are fully resolved, so the RF error rate is the fraction of bipartitions defined by internal edges of the model tree that are not in the estimated tree. Reproduced from Willson et al. (2021) under the Creative Commons Attribution Non-Commercial License.

so evolve with GDL) and restrict instead to those genes that are single copy. This practice reduces available data, and so raises the concern that accuracy could be reduced. Alternatively, methods to detect orthology are used, so that the multi-copy family can be reduced to single-copy genes. However, orthology detection is still not reliably solved well (Glover et al., 2019), and so this approach also has some problems. Phyldog (Boussau et al., 2013) is a statistically rigorous approach that co-estimates gene trees and species trees from multi-locus datasets that have evolved under GDL, but is computationally very intensive. Finally, methods that can construct species trees from MUL-trees can be used.

A recent theoretical advance is the proof that ASTRAL-multi and ASTRAL-one, two modifications of ASTRAL to enable them to estimate species trees from MUL-trees, are statistically consistent under statistical models of gene evolution that allow for GDL (Legried et al., 2021; Markin and Eulenstein, 2021). However, these statistically consistent methods are not as accurate as ASTRAL-Pro (Zhang et al., 2020b), a variant of ASTRAL recently developed specifically to address GDL (Zhang et al., 2020b). Other methods that can estimate species trees from a set of MUL-trees have been developed, with gene tree parsimony the most well known (e.g., DupTree (Wehe et al., 2008)), but also including MixTrEm-DLRS (Ullah et al., 2015), MulRF (Chaudhary et al., 2015), FastMulRFS (Molloy and Warnow, 2020), and SpeciesRax (Morel et al., 2021). While not all of them have been compared to ASTRAL-Pro, those that have been evaluated have not been shown to be as reliably accurate as ASTRAL-Pro (Yan et al., 2021).

Tree-decomposition represents an alternative approach to methods like ASTRAL-Pro that combine MUL-trees to estimate the species tree. In a tree-decomposition approach, each gene family tree is decomposed into a set of single-copy trees, and then the resultant set of single-copy trees is given to a selected summary method, such as ASTRAL or ASTRID. There are several such tree-decomposition methods, with DISCO (Willson et al., 2021) a recent and promising technique. As seen in Figure 4, using DISCO with ASTRID on a dataset with 1,000 species produces a tree that is more accurate than ASTRAL-Pro and SpeciesRax, while being much faster and having lower memory requirements than both methods.

Potential limitations of ASTRAL-Pro and DISCO-based methods. The results shown in Figure 4, which were taken from Willson et al. (2021), show that ASTRAL-Pro and ASTRID-DISCO both have very good accuracy, better than SpeciesRax, and that ASTRID-DISCO is the fastest of the three methods. The runtime advantage of ASTRID-DISCO over both ASTRAL-Pro and SpeciesRax is essentially guaranteed by its design, as it uses ASTRID (which is very fast) to estimate species trees from single-copy gene trees produced by DISCO, and the other methods are not as fast. However, the relative accuracy will depend on the model condition, with dataset size (number of species and genes), gene tree estimation error rate, and cause(s) for discordance between gene trees and species trees relevant factors. Zhang et al. (2020b) explored a wide range of model conditions with varying levels of ILS, GDL, and gene tree estimation error, and ASTRAL-Pro matched or improved on all other methods they explored; however, they did not explore SpeciesRax or ASTRID-DISCO, as neither had been developed by that time. The conditions explored by Willson et al. (2021) in comparing ASTRAL-Pro, ASTRID-DISCO, and SpeciesRax also included varying levels of ILS, GDL, and gene tree estimation error, but did not include HGT. Finally, Morel et al. (2021) explored variants of the model conditions explored in Zhang et al. (2020b), modified to include HGT (so that they had mainly DLS and HGT, but also some conditions with ILS). It is interesting to note that the SpeciesRax study (Morel et al., 2021) reports that SpeciesRax is more accurate than ASTRAL-Pro, which is not what was reported in Willson et al. (2021). Willson et al. (2021) notes that the conditions in Morel et al. (2021) where there is an advantage for SpeciesRax over ASTRAL-Pro are for small numbers of genes and species and high levels of HGT; in other conditions, the differences between the two methods are minor or favor ASTRAL-Pro.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2022 doi:10.20944/preprints202110.0258.v2

Table 1: Average delta error (Δe) for phylogenetic placement methods in backbone trees of size n. Analyses were limited to 64Gb of memory.

	$n = 5000$	$n = 10,000$	$n = 50,000$	$n = 100,000$	$n = 200,000$
			Δe		
pplacer-XR	0.150	0.132	0.085	0.084	0.075
EPA-ng	0.239	0.219	X	X	X
APPLES	0.366	0.330	0.239	0.247	0.250

We now turn to limitations of the DISCO decomposition strategy, and to ASTRID-DISCO, which runs ASTRID on the DISCO trees. Because ASTRID-DISCO depends on ASTRID specifically, the relative accuracy of ASTRID-DISCO and other methods (e.g., ASTRAL-DISCO and ASTRAL-Pro) will depend on the extent to which ASTRID provides good accuracy. This is a question that we do not yet have an answer, since we have seen that ASTRAL and ASTRID vary in their relative accuracy, and the conditions under which each provides an advantage are not known. Thus, future work is needed to explore ASTRID-DISCO in comparison to ASTRAL-Pro and other methods under a larger number and range of model conditions.

5 Recent Advances in Updating Large Trees

Once a large tree is estimated, if new sequence data become available, then starting all over is undesirable (especially since the first tree may have already required a great deal of computational effort and time). Hence, the problem of updating a tree by adding newly found sequences into the tree becomes relevant. We consider this in two contexts: adding leaves to gene trees and to species trees.

The methods described in this section are also relevant to understanding microbial diversity: given a sequence, placing it into a taxonomy makes it possible to taxonomically characterize the sequence, and so also enables an assessment of microbial diversity in a population (Nguyen et al., 2014; Segata et al., 2013; Czech et al., 2020; Shah et al., 2021). This approach is particularly relevant for characterizing novel sequences (i.e., sequences that are not in public databases) and the accuracy of the taxonomic assignment improves on larger trees (Shah et al., 2021). Therefore, methods for placing sequences into large trees also have utility for assessment of microbial diversity.

Phylogenetic placement is also useful when the input sequence dataset exhibits sequence length heterogeneity: for example, FastTree can have poor topological accuracy on datasets with fragmentary sequences (see Figure 3 and also Sayyari et al. (2017); Smirnov and Warnow (2021b)), with the consequence that in some conditions constructing trees on the full length sequences and then using phylogenetic placement to add the remaining sequences can be more accurate than FastTree on a good alignment (Smirnov and Warnow, 2021b).

5.1 Adding sequences to gene trees

One of the earliest methods for phylogenetic placement is pplacer (Matsen et al., 2010), which assumed that the input is a binary tree with sequences at the leaves in an alignment, and a set of query sequences that need to be added into the tree. The approach used in pplacer is likelihood-based, with maximum likelihood or Bayesian options both available; here we describe the maximum likelihood version. For a given query sequence q, pplacer would find the best location in the tree to add q (i.e., the best edge in the tree to subdivide and then make q a leaf adjacent to the new node) in order to optimize the maximum likelihood score. Because pplacer is likelihood-based, this approach can be computationally intensive (Balaban et al., 2020).

Other phylogenetic placement methods have been developed that seek to improve scalability to larger trees or reduce running time (e.g., UShER (Turakhia et al., 2021), RAPPAS (Linard et al., 2019), EPA-ng (Barbera et al., 2019), APPLES (Balaban et al., 2020), and APPLES-2 (Balaban et al., 2021)). EPA-ng is likelihood-based and has been optimized for "batch processing" of query sequences (so that the cost of performing phylogenetic placement of a large number of query sequences is much less than the cost of placing them one-by-one). EPA-ng has slightly reduced accuracy compared to pplacer. APPLES is a very fast distance-based method and uses dynamic programming to place each query sequence into the tree so as to minimize the weighted least squares error. APPLES-2 is an improvement on APPLES with respect to accuracy and running time, and also scales to at least 200,000 sequences. Recent studies (Balaban et al., 2020, 2021; Wedell et al., 2021) show that APPLES and APPLES-2 can run on trees with 200,000 leaves and are much faster than both pplacer and EPA-ng; however, even APPLES-2 does not match the accuracy of pplacer. UShER is parsimony-based and very fast, but has not been compared to pplacer, APPLES, or APPLES-2, while RAPPAS, which is based on k-mers, is very fast but not as accurate as EPA-ng or pplacer (Linard et al., 2019)). Thus, the highest accuracy in phylogenetic placement is obtained using likelihood-based methods, but these tend to be relatively computationally intensive compared to other approaches, especially distance-based or k-mer based methods.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 January 2022 doi:10.20944/preprints202110.0258.v2

Recently, two divide-and-conquer methods, pplacer-XR (pplacer-eXtended Range) (Wedell et al., 2021) and pplacer-DC(pplacer-Divide-and-Conquer) (Koning et al., 2021), were developed in order to improve accuracy for phylogenetic placement when inserting into trees that are too large for pplacer. Here we describe the pplacer-XR approach, as a comparison of pplacer-XR to pplacer-DC on the RNASim VS datasets reported in Wedell et al. (2021) and Koning et al. (2021) shows that pplacer-XR is faster, uses less memory, and is more accurate than pplacerDC. In addition, pplacer-XR is able to scale to trees with 200,000 leaves whereas pplacer-DC scales only to 100,000 sequences (Koning et al., 2021; Wedell et al., 2021).

The pplacer-XR pipeline uses four stages to insert a query sequence q into a tree T. First, a leaf that has the greatest similarity to q is found (where similarity is based on percent ID). In the second stage, a contiguous subtree t is extracted from T that includes the nearest leaf and up to $N-1$ additional leaves (where $N = 2000$ when the XR framework is used with pplacer). In the third stage, pplacer is used to insert the query sequence into the subtree t (i.e., an edge e in the subtree t is identified); since N was set to be only 2,000, pplacer can complete on this dataset. Finally, in the fourth stage, an edge e' in the tree T is found corresponding to the edge e, and the query sequence is placed into edge e'. By design, this four-stage approach can be modified to suit a different phylogenetic placement method, so that methods that can run on larger trees can have larger values for N. For example, when using the XR framework with EPA-ng, N is set to 10,000. Every stage of this pipeline, other than the third stage (which runs pplacer), is very fast and uses little memory.

Table 1 compares pplacer-XR (i.e., pplacer used within the XR framework) to APPLES and EPA-ng with respect to delta-error (a measure for the increase in topological error in the tree produced by the phylogenetic placement method, see Balaban et al. (2020); Wedell et al. (2021) for the definition). The placement methods are given full-length sequences in the true alignment and place these sequences in a leave-one-out strategy into the model tree on the remaining sequences, with trees varying from 5000 to 200,000 sequences. EPA-ng fails to be able to place into the largest trees due to memory requirements, but APPLES and pplacer-XR succeed on all trees. Note that pplacer-XR has the lowest placement error of all methods.

5.2 Adding species to species trees

While the methods above focused on adding sequences into gene trees, adding species (represented by genome-scale data) into species trees is another kind of phylogenetic placement problem. One such method is MGPlacer (Kay et al., 2015), which uses reads from across a genome to place a genome into a species tree. Other approaches, such as INSTRAL (Rabiee and Mirarab, 2020b), have been developed that consider heterogeneity across the genome due to processes such as ILS. Given an existing species tree T, INSTRAL will add the new species into the existing tree to optimize the quartet tree support for the extended species tree (i.e., INSTRAL extends the theoretical approach in ASTRAL). Another new method is DEPP (Jiang et al., 2021), which computes distances using a deep neural network (DNN) and then runs APPLES to place the new species into the tree. By training the DNN appropriately, these distances can be appropriate to this problem of adding species into species trees.

5.3 Potential limitations of phylogenetic placement methods

Clearly, pplacer-XR is a very accurate phylogenetic placement method for adding sequences into large gene trees and DEPP and INSTRAL provide advantages for phylogenetic placement of sequences or genomes into species trees. Since less is known yet about phylogenetic placement in the context of species trees, we focus the discussion of limitations to the problem of adding sequences into gene trees.

The first limitation worth noting is that pplacer-XR has not been compared to APPLES-2, and given that APPLES-2 is more accurate than APPLES, a comparison between pplacer-XR and APPLES-2 is merited. It is also worth noting that the scalability of pplacer-XR was only evaluated on the RNASim VS (Variable Size) datasets. Further, while the RNASim simulation itself is complex (see description of the simulation protocol in Mirarab et al. (2015)), as shown in Mirarab et al. (2015); Smirnov and Warnow (2021a), the dataset itself is "easy" to analyze in some ways (e.g., the average pairwise distance between two sequences is not very large, making multiple sequence alignment and tree estimation relatively easy compared to other datasets with higher rates of evolution. Furthermore, while pplacer-XR is reasonably fast, it is not designed to efficiently place a large number of query sequences into a tree, which is a natural use of phylogenetic placement when performing taxon identification of microbiome samples. This, on the other hand, is a strong benefit provided by EPA-ng (Barbera et al., 2019), which is specifically designed for the batch placement problem. Thus, while pplacer-XR provides specific advantages over other phylogenetic placement methods, it does not provides the full benefits needed for all the applications of phylogenetic placement.

6 Concluding Remarks

This review has shown the significant innovations over the last few years in the development of methods that provide high accuracy on very large datasets (even up to 1,000,000 sequences). As our survey notes, for each of the problems we addressed, whether it be multiple sequence alignment, gene tree estimation, or species tree estimation, there is often a choice between

 doi:10.20944/preprints202110.0258.v2

the most accurate methods and the ones that are computationally feasible. Here we have focused our attention on techniques for scaling excellent but computationally intensive methods to large datasets.

However, we did not discuss all the relevant problems for large-scale tree estimation, including how to efficiently and accurately estimate numeric parameters (e.g., branch lengths) or evaluate branch support in a large tree. There is active work on these problems (e.g., see Sharma and Kumar (2021); Lemoine et al. (2018); Guindon and Gascuel (2019)), but each of these problems is likely to remain an important direction for research. We also did not address Bayesian inference, which is an important class of phylogenetic methods (Chen et al., 2014; Czech et al., 2020; Holder and Lewis, 2003). Bayesian methods, such as MrBayes (Ronquist and Huelsenbeck, 2003), are well established in the research community and have been shown to provide highly accurate point estimates of alignments, gene trees, and species trees; however, most Bayesian methods use MCMC (Markov Chain Monte Carlo) and are computationally intensive on large datasets since convergence to the stationary distribution is required for high confidence in an accurate result. Some progress has been made on improving the scalability of these point estimations using Bayesian methods, e.g., by using divide-and-conquer to break a large dataset into subsets or constraining the search space (e.g., Zimmermann et al. (2014); Nute and Warnow (2016); Wang et al. (2020); Gupta et al. (2021)). However, Bayesian methods produce distributions from which point estimates can be obtained, and these distributions have significant additional value since they enable uncertainty quantification. Scaling Bayesian methods to large datasets so that a good estimate of the distribution can be obtained is of great interest, but is generally not enabled through the techniques that focus on scaling the point estimates. Here we note that Zhang et al. (2020a) has made some progress in scaling MrBayes, suggesting that additional effort in this direction is merited. In general, fully scaling Bayesian methods requires additional techniques beyond the ones explored in this survey.

We also did not discuss in full how different causes for gene tree discord can affect species tree estimation. Quartet-based methods, such as ASTRAL, ASTRAL-Pro, and wQFM, can have very good accuracy on simulated and biological datasets when ILS and/or GDL are the cause for discordance between gene trees and species trees and there is even some evidence that quartet-based methods can provide good accuracy when HGT is also present Davidson et al. (2015); this empirical performance is consistent with the strong theoretical properties for quartet-based methods that have been established under the relevant models Legried et al. (2021); Markin and Eulenstein (2021); Roch and Snir (2013); Daskalakis and Roch (2016). However, gene flow can also create discordance between gene trees and species trees, and quartet-based methods such as ASTRAL have been proven to be statistically inconsistent under some conditions when gene flow is present (Solís-Lemus et al., 2016). A simulation study in Solís-Lemus et al. (2016) showed that using PhyloNet (Wen et al., 2018) to construct a maximum likelihood phylogenetic network with hybridization edges and then suppressing the "minor" hybrid edge produced the most accurate results, followed by ASTRAL, NJst (Liu and Yu, 2011), and finally concatenation. Thus, while ASTRAL provided superior accuracy compared to the other tree inference methods, when gene flow was present an explicit phylogenetic network approach was key to obtaining high accuracy (see Morrison (2011) for the difference between explicit phylogenetic networks and other types, which he refers to as "data-display networks"). Thus, Solís-Lemus et al. (2016) identifies a basic limitation for tree-based methods that do not consider a wide range of causes for gene tree discordance, and argues for the use of explicit phylogenetic network methods, with recent surveys in Kong et al. (2021); Blair and Ané (2020). Unfortunately, methods for constructing phylogenetic networks are enormously complex and sufficiently computationally intensive so that even the most scalable methods are limited to a few tens of species (Elworth et al., 2019; Lutteropp et al., 2021; Rabier et al., 2021; Mirarab et al., 2021).

Therefore, method development for phylogenetic network estimation is also needed, or–at a minimum–methods for estimating trees that are reasonable models of the evolutionary processes underway. It is helpful that despite the need for further research and method development that can address the full range of biological processes that create heterogeneity across the genome, some of the recently developed methods for large-scale multiple sequence alignment and tree estimation (both gene trees and species trees) are much more accurate than earlier methods. At the same time, because so many of the methods discussed in this review are extremely new, additional studies are needed to explore and understand the conditions under which these methods are reliably more accurate than alternative methods, and our review has suggested some potential directions where such study is needed.

This study did not discuss all the recent advances in large-scale alignment and tree estimation, and some of these may provide even better scalability and accuracy. For example, there are new methods for large-scale maximum likelihood tree estimation (e.g., Very Fast Tree (Piñeiro et al., 2020)), new techniques to speed up co-estimation of gene trees and species trees (Wang and Nakhleh, 2018; Wang et al., 2020), and even divide-and-conquer approaches to phylogenetic network estimation (Zhu et al., 2019a). This continued effort to develop methods that are highly accurate and scalable leads us to the optimistic prediction that the next 5 to 10 years will result in new scalable methods to estimate accurate alignments, trees, and even phylogenetic networks, and that these methods will enable biologists to make discoveries on the large and ultra-large phylogenomic datasets that they assemble.

Funding Statement

This research was supported in part by the US National Science Foundation grant 2006069 to TW and by the Grainger Foundation support to TW.

Acknowledgments

The authors thank Siavash Mirarab and Luay Nakhleh for comments on the manuscript and suggestions for additional papers to discuss and cite, which led to improvements in the manuscript.

References

Asnicar, F., Thomas, A. M., Beghini, F., Mengoni, C., Manara, S., Manghi, P., Zhu, Q., Bolzan, M., Cumbo, F., May, U., et al. (2020). Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. *Nature Communications*, 11(1):1–10.

Bader, D. A. and Madduri, K. (2019). High-performance phylogenetic inference. In *Bioinformatics and Phylogenetics*, pages 39–46. Springer.

Balaban, M., Jiang, Y., Roush, D., Zhu, Q., and Mirarab, S. (2021). APPLES-2: Faster and more accurate distance-based phylogenetic placement using divide and conquer. *Molecular Ecology Resources*. In press, https://doi.org/10.1111/1755-0998.13527.

Balaban, M., Sarmashghi, S., and Mirarab, S. (2020). APPLES: scalable distance-based phylogenetic placement with or without alignments. *Systematic Biology*, 69(3):566–578.

Barbera, P., Kozlov, A. M., Czech, L., Morel, B., Darriba, D., Flouri, T., and Stamatakis, A. (2019). EPA-ng: massively parallel evolutionary placement of genetic sequences. *Systematic Biology*, 68(2):365–369.

Bayzid, M. S., Mirarab, S., Boussau, B., and Warnow, T. (2015). Weighted statistical binning: enabling statistically consistent genome-scale phylogenetic analyses. *PloS One*, 10(6):e0129183.

Bininda-Emonds, O. R. (2004). The evolution of supertrees. *Trends in Ecology & Evolution*, 19(6):315–322.

Blair, C. and Ané, C. (2020). Phylogenetic trees and networks can serve as powerful and complementary approaches for analysis of genomic data. *Systematic Biology*, 69(3):593–601.

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

Chaudhary, R., Fernández-Baca, D., and Burleigh, J. G. (2015). MulRF: a software package for phylogenetic analysis using multi-copy gene trees. *Bioinformatics*, 31(3):432–433.

Chen, M.-H., Kuo, L., and Lewis, P. O. (2014). *Bayesian phylogenetics: methods, algorithms, and applications*. CRC Press.

Chifman, J. and Kubatko, L. (2014). Quartet inference from SNP data under the coalescent model. *Bioinformatics*, 30(23):3317–3324.

Czech, L., Barbera, P., and Stamatakis, A. (2020). Genesis and Gappa: processing, analyzing and visualizing phylogenetic (placement) data. *Bioinformatics*, 36(10):3263–3265.

Daskalakis, C. and Roch, S. (2016). 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.

Davidson, R., Vachaspati, P., Mirarab, S., and Warnow, T. (2015). Phylogenomic species tree estimation in the presence of incomplete lineage sorting and horizontal gene transfer. *BMC genomics*, 16(10):1–12.

Dibaeinia, P., Tabe-Bordbar, S., and Warnow, T. (2021). FASTRAL: improving scalability of phylogenomic analysis. *Bioinformatics*, 37:2317–2324.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797.

Elworth, R. L., Ogilvie, H. A., Zhu, J., and Nakhleh, L. (2019). Advances in computational methods for phylogenetic networks in the presence of hybridization. In *Bioinformatics and Phylogenetics*, pages 317–360. Springer.

Foster, P. G. and Hickey, D. A. (1999). Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. *Journal of Molecular Evolution*, 48(3):284–290.

Glover, N., Dessimoz, C., Ebersberger, I., Forslund, S. K., Gabaldón, T., Huerta-Cepas, J., Martin, M.-J., Muffato, M., Patricio, M., Pereira, C., et al. (2019). Advances and applications in the quest for orthologs. *Molecular Biology and Evolution*, 36(10):2157–2164.

Guindon, S. and Gascuel, O. (2019). Numerical optimization techniques in maximum likelihood tree inference. In *Bioinformatics and Phylogenetics*, pages 21–38. Springer.

Gupta, M., Zaharias, P., and Warnow, T. (2021). Accurate large-scale phylogeny-aware alignment using BAli-Phy. *Bioinformatics*. `https://doi.org/10.1093/bioinformatics/btab555`.

Holder, M. and Lewis, P. O. (2003). Phylogeny estimation: traditional and Bayesian approaches. *Nature Reviews Genetics*, 4(4):275–284.

Janssens, S. B., Couvreur, T. L., Mertens, A., Dauby, G., Dagallier, L.-P. M., Abeele, S. V., Vandelook, F., Mascarello, M., Beeckman, H., Sosef, M., et al. (2020). A large-scale species level dated angiosperm phylogeny for evolutionary and ecological analyses. *Biodiversity Data Journal*, 8:e39677.

Jarvis, E. D., Mirarab, S., Aberer, A. J., Li, B., Houde, P., Li, C., Ho, S. Y., Faircloth, B. C., Nabholz, B., Howard, J. T., et al. (2014). Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, 346(6215):1320–1331.

Jiang, X., Edwards, S. V., and Liu, L. (2020). The multispecies coalescent model outperforms concatenation across diverse phylogenomic data sets. *Systematic Biology*, 69(4):795–812.

Jiang, Y., Balaban, M., Zhu, Q., and Mirarab, S. (2021). DEPP: deep learning enables extending species trees using single genes. *bioRxiv*. `https://doi.org/10.1101/2021.01.22.427808`.

Katoh, K., editor (2021). *Multiple Sequence Alignment: Methods and Protocols*. Springer.

Katoh, K. and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30(4):772–780.

Kay, G. L., Sergeant, M. J., Zhou, Z., Chan, J. Z.-M., Millard, A., Quick, J., Szikossy, I., Pap, I., Spigelman, M., Loman, N. J., et al. (2015). Eighteenth-century genomes show that mixed infections were common at time of peak tuberculosis in Europe. *Nature Communications*, 6(1):1–9.

Kececioglu, J. (1993). The maximum weight trace problem in multiple sequence alignment. In *Annual Symposium on Combinatorial Pattern Matching*, pages 106–119. Springer.

Kong, S., Pons, J. C., Kubatko, L., and Wicke, K. (2021). Classes of explicit phylogenetic networks and their biological and mathematical significance. *arXiv preprint arXiv:2109.10251*.

Koning, E., Phillips, M., and Warnow, T. (2021). pplacerDC: a new scalable phylogenetic placement method. In *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 1–9.

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., and Stamatakis, A. (2019). RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, 35(21):4453–4455.

Kubatko, L. S. and Degnan, J. H. (2007). Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic Biology*, 56(1):17–24.

Lartillot, N. and Philippe, H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution*, 21(6):1095–1109.

Le, T., Sy, A., Molloy, E. K., Zhang, Q., Rao, S., and Warnow, T. (2020). Using Constrained-INC for large-scale gene tree and species tree estimation. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 18(1):2–15.

Lees, J. A., Kendall, M., Parkhill, J., Colijn, C., Bentley, S. D., and Harris, S. R. (2018). Evaluation of phylogenetic reconstruction methods using bacterial whole genomes: a simulation based study. *Wellcome Open Research*, 3.

Lefort, V., Desper, R., and Gascuel, O. (2015). FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. *Molecular Biology and Evolution*, 32(10):2798–2800.

Legried, B., Molloy, E. K., Warnow, T., and Roch, S. (2021). Polynomial-time statistical estimation of species trees under gene duplication and loss. *Journal of Computational Biology*, 28(5):452–468.

Lemoine, F., Entfellner, J.-B. D., Wilkinson, E., Correia, D., Felipe, M. D., De Oliveira, T., and Gascuel, O. (2018). Renewing Felsenstein's phylogenetic bootstrap in the era of big data. *Nature*, 556(7702):452–456.

Linard, B., Swenson, K., and Pardi, F. (2019). Rapid alignment-free phylogenetic identification of metagenomic sequences. *Bioinformatics*, 35(18):3303–3312.

Liu, B. and Warnow, T. (2021). Scalable species tree inference with external constraints. *bioRxiv*. https://doi.org/10.1101/2021.11.05.467436, Accepted to the Journal for Computational Biology.

Liu, K., Linder, C. R., and Warnow, T. (2011). RAxML and FastTree: comparing two methods for large-scale maximum likelihood phylogeny estimation. *PloS one*, 6(11):e27731.

Liu, K., Raghavan, S., Nelesen, S., Linder, C. R., and Warnow, T. (2009). Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science*, 324(5934):1561–1564.

Liu, L. and Yu, L. (2011). Estimating species trees from unrooted gene trees. *Systematic Biology*, 60(5):661–667.

Liu, L., Yu, L., and Edwards, S. V. (2010). A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolutionary Biology*, 10(1):1–18.

Lopez, P., Casane, D., and Philippe, H. (2002). Heterotachy, an important process of protein evolution. *Molecular Biology and Evolution*, 19(1):1–7.

Lutteropp, S., Scornavacca, C., Kozlov, A. M., Morel, B., and Stamatakis, A. M. (2021). NetRAX: accurate and fast maximum likelihood phylogenetic network inference. *bioRxiv*. https://doi.org/10.1101/2021.08.30.458194.

Maddison, W. P. (1997). Gene trees in species trees. *Systematic Biology*, 46(3):523–536.

Mahbub, M., Wahab, Z., Reaz, R., Rahman, M. S., and Bayzid, M. (2021). wQFM: highly accurate genome-scale species tree estimation from weighted quartets. *Bioinformatics*, 37:3734–3743.

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

Matsen, F. A., Kodner, R. B., and Armbrust, E. V. (2010). pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, 11(1):1–16.

Mirarab, S., Bayzid, M. S., Boussau, B., and Warnow, T. (2014a). Statistical binning enables an accurate coalescent-based estimation of the avian tree. *Science*, 346(6215).

Mirarab, S., Nakhleh, L., and Warnow, T. (2021). Multispecies coalescent: theory and applications in phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, 52:247–268.

Mirarab, S., Nguyen, N., Guo, S., Wang, L.-S., Kim, J., and Warnow, T. (2015). PASTA: ultra-large multiple sequence alignment for nucleotide and amino-acid sequences. *Journal of Computational Biology*, 22(5):377–386.

Mirarab, S., Reaz, R., Bayzid, M. S., Zimmermann, T., Swenson, M. S., and Warnow, T. (2014b). ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*, 30(17):i541–i548.

Molloy, E. K. and Warnow, T. (2018). To include or not to include: the impact of gene filtering on species tree estimation methods. *Systematic Biology*, 67(2):285–303.

Molloy, E. K. and Warnow, T. (2019a). Statistically consistent divide-and-conquer pipelines for phylogeny estimation using NJMerge. *Algorithms for Molecular Biology*, 14(1):1–17.

Molloy, E. K. and Warnow, T. (2019b). TreeMerge: a new method for improving the scalability of species tree estimation methods. *Bioinformatics*, 35(14):i417–i426.

Molloy, E. K. and Warnow, T. (2020). FastMulRFS: fast and accurate species tree estimation under generic gene duplication and loss models. *Bioinformatics*, 36(Supplement_1):i57–i65.

Morel, B., Schade, P., Lutteropp, S., Williams, T. A., Szöllösi, G. J., and Stamatakis, A. (2021). SpeciesRax: a tool for maximum likelihood species tree inference from gene family trees under duplication, transfer, and loss. *bioRxiv*. https://doi.org/10.1101/2021.03.29.437460.

Morrison, D., editor (2011). *Introduction to Phylogenetic Networks*. RJR Productions, Uppsala. ISBN 978-91-980099-0-3.

Nabhan, A. R. and Sarkar, I. N. (2012). The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy. *Briefings in Bioinformatics*, 13(1):122–134.

Nelesen, S., Liu, K., Wang, L.-S., Linder, C. R., and Warnow, T. (2012). DACTAL: divide-and-conquer trees (almost) without alignments. *Bioinformatics*, 28(12):i274–i282.

Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A., and Minh, B. Q. (2015a). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32(1):268–274.

Nguyen, N.-p., Mirarab, S., Liu, B., Pop, M., and Warnow, T. (2014). TIPP: taxonomic identification and phylogenetic profiling. *Bioinformatics*, 30(24):3548–3555.

Nguyen, N.-p. D., Mirarab, S., Kumar, K., and Warnow, T. (2015b). Ultra-large alignments using phylogeny-aware profiles. *Genome Biology*, 16(1):1–15.

Nute, M. and Warnow, T. (2016). Scaling statistical multiple sequence alignment to large datasets. *BMC Genomics*, 17(10):135–144.

Ogilvie, H. A., Bouckaert, R. R., and Drummond, A. J. (2017). StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. *Molecular Biology and Evolution*, 34(8):2101–2114.

Park, M., Zaharias, P., and Warnow, T. (2021). Disjoint tree mergers for large-scale maximum likelihood tree estimation. *Algorithms*, 14(5):148.

Piñeiro, C., Abuín, J. M., and Pichel, J. C. (2020). Very Fast Tree: speeding up the estimation of phylogenies for large alignments through parallelization and vectorization strategies. *Bioinformatics*, 36(17):4658–4659.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PloS One*, 5(3):e9490.

Rabiee, M. and Mirarab, S. (2020a). Forcing external constraints on tree inference using ASTRAL. *BMC Genomics*, 21(2):1–13.

Rabiee, M. and Mirarab, S. (2020b). INSTRAL: discordance-aware phylogenetic placement using quartet scores. *Systematic Biology*, 69(2):384–391.

Rabier, C.-E., Berry, V., Stoltz, M., Santos, J. D., Wang, W., Glaszmann, J.-C., Pardi, F., and Scornavacca, C. (2021). On the inference of complex phylogenetic networks by Markov Chain Monte-Carlo. *PLoS Computational Biology*, 17(9):e1008380.

Redelings, B. D. and Suchard, M. A. (2005). Joint Bayesian estimation of alignment and phylogeny. *Systematic Biology*, 54(3):401–418.

Roch, S. (2006). A short proof that phylogenetic tree reconstruction by maximum likelihood is hard. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 3(1):92–94.

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

Roch, S. and Snir, S. (2013). Recovering the treelike trend of evolution despite extensive lateral genetic transfer: a probabilistic analysis. *Journal of Computational Biology*, 20(2):93–112.

Roch, S. and Steel, M. (2015). Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. *Theoretical Population Biology*, 100:56–62.

Ronquist, F. and Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572–1574.

Roshan, U., Moret, B. M., Williams, T. L., and Warnow, T. (2004). Performance of supertree methods on various data set decompositions. In *Phylogenetic Supertrees*, pages 301–328. Springer.

Roure, B. and Philippe, H. (2011). Site-specific time heterogeneity of the substitution process and its impact on phylogenetic inference. *BMC Evolutionary Biology*, 11(1):1–14.

Sanderson, M. J., Boss, D., Chen, D., Cranston, K. A., and Wehe, A. (2008). The PhyLoTA Browser: processing GenBank for molecular phylogenetics research. *Systematic Biology*, 57(3):335–346.

Sayyari, E. and Mirarab, S. (2016). Anchoring quartet-based phylogenetic distances and applications to species tree reconstruction. *BMC Genomics*, 17(10):101–113.

Sayyari, E., Whitfield, J. B., and Mirarab, S. (2017). Fragmentary gene sequences negatively impact gene tree and species tree reconstruction. *Molecular Biology and Evolution*, 34(12):3279–3291.

Segata, N., Börnigen, D., Morgan, X. C., and Huttenhower, C. (2013). PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nature Communications*, 4(1):1–11.

Shah, N., Molloy, E. K., Pop, M., and Warnow, T. (2021). TIPP2: metagenomic taxonomic profiling using phylogenetic markers. *Bioinformatics*, 37:1839–1845.

Sharma, S. and Kumar, S. (2021). Fast and accurate bootstrap confidence limits on genome-scale phylogenies using little bootstraps. *Nature Computational Science*, 1(9):573–577.

Shen, C., Zaharias, P., and Warnow, T. (2021). MAGUS+eHMMs: improved multiple sequence alignment accuracy for fragmentary sequences. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btab788.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 7(1):539.

Smirnov, V. (2021). Recursive MAGUS: Scalable and accurate multiple sequence alignment. *PLOS Computational Biology*, 17(10):1–17. Publisher: Public Library of Science.

Smirnov, V. and Warnow, T. (2020). Unblended disjoint tree merging using GTM improves species tree estimation. *BMC Genomics*, 21(2):1–17.

Smirnov, V. and Warnow, T. (2021a). MAGUS: multiple sequence alignment using graph clustering. *Bioinformatics*, 37(12):1666–1672.

Smirnov, V. and Warnow, T. (2021b). Phylogeny estimation given sequence length heterogeneity. *Systematic Biology*, 70(2):268–282.

Smith, S. A., Beaulieu, J. M., and Donoghue, M. J. (2009). Mega-phylogeny approach for comparative biology: an alternative to supertree and supermatrix approaches. *BMC Evolutionary Biology*, 9(1):1–12.

Smith, S. A. and Walker, J. F. (2019). PyPHLAWD: A python tool for phylogenetic dataset construction. *Methods in Ecology and Evolution*, 10(1):104–108.

Solís-Lemus, C., Yang, M., and Ané, C. (2016). Inconsistency of species tree methods under gene flow. *Systematic Biology*, 65(5):843–851.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313.

Stamatakis, A. (2019). A review of approaches for optimizing phylogenetic likelihood calculations. In *Bioinformatics and Phylogenetics*, pages 1–19. Springer.

Steel, M. A. and Penny, D. (1993). Distributions of tree comparison metrics—some new results. *Systematic Biology*, 42(2):126–141.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, 17(2):57–86.

Turakhia, Y., Thornlow, B., Hinrichs, A. S., De Maio, N., Gozashti, L., Lanfear, R., Haussler, D., and Corbett-Detig, R. (2021). Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nature Genetics*, 53(6):809–816.

Ullah, I., Parviainen, P., and Lagergren, J. (2015). Species tree inference using a mixture model. *Molecular Biology and Evolution*, 32(9):2469–2482.

Vachaspati, P. and Warnow, T. (2015). ASTRID: accurate species trees from internode distances. *BMC Genomics*, 16(10):1–13.

Vachaspati, P. and Warnow, T. (2018). SVDquest: Improving SVDquartets species tree estimation using exact optimization within a constrained search space. *Molecular Phylogenetics and Evolution*, 124:122–136.

Wang, Y. and Nakhleh, L. (2018). Towards an accurate and efficient heuristic for species/gene tree co-estimation. *Bioinformatics*, 34(17):i697–i705.

Wang, Y., Ogilvie, H. A., and Nakhleh, L. (2020). Practical speedup of Bayesian inference of species phylogenies by restricting the space of gene trees. *Molecular Biology and Evolution*, 37(6):1809–1818.

Warnow, T. (2019). Divide-and-conquer tree estimation: Opportunities and challenges. In *Bioinformatics and Phylogenetics*, pages 121–150. Springer.

Wedell, E., Cai, Y., and Warnow, T. (2021). Scalable and accurate phylogenetic placement using pplacer-XR. In *International Conference on Algorithms for Computational Biology*, pages 94–105. Springer.

Wehe, A., Bansal, M. S., Burleigh, J. G., and Eulenstein, O. (2008). DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony. *Bioinformatics*, 24(13):1540–1541.

Wen, D., Yu, Y., Zhu, J., and Nakhleh, L. (2018). Inferring phylogenetic networks using PhyloNet. *Systematic Biology*, 67(4):735–740.

Willson, J., Roddur, M. S., Liu, B., Zaharias, P., and Warnow, T. (2021). DISCO: species tree inference using multi-copy gene family tree decomposition. *Systematic Biology*. `https://doi.org/10.1093/sysbio/syab070`.

Yan, Z., Smith, M. L., Du, P., Hahn, M. W., and Nakhleh, L. (2021). Species tree inference methods intended to deal with incomplete lineage sorting are robust to the presence of paralogs. *Systematic Biology*. https://doi.org/10.1093/sysbio/syab056.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from dna sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution*, 39(3):306–314.

Yin, J., Zhang, C., and Mirarab, S. (2019). ASTRAL-MP: scaling ASTRAL to very large datasets using randomization and parallelization. *Bioinformatics*, 35(20):3961–3969.

Zaharias, P., Grosshauser, M., and Warnow, T. (2022). Re-evaluating deep neural networks for phylogeny estimation: The issue of taxon sampling. *Journal of Computational Biology*. Accepted, special issue for RECOMB 2021.

Zaharias, P., Smirnov, V., and Warnow, T. (2021). The maximum weight trace alignment merging problem. In *International Conference on Algorithms for Computational Biology*, pages 159–171. Springer.

Zhang, C., Huelsenbeck, J. P., and Ronquist, F. (2020a). Using parsimony-guided tree proposals to accelerate convergence in Bayesian phylogenetic inference. *Systematic biology*, 69(5):1016–1032.

Zhang, C., Scornavacca, C., Molloy, E. K., and Mirarab, S. (2020b). ASTRAL-Pro: quartet-based species-tree inference despite paralogy. *Molecular Biology and Evolution*, 37(11):3292–3307.

Zhu, J., Liu, X., Ogilvie, H. A., and Nakhleh, L. K. (2019a). A divide-and-conquer method for scalable phylogenetic network inference from multilocus data. *Bioinformatics*, 35(14):i370–i378.

Zhu, Q., Mai, U., Pfeiffer, W., Janssen, S., Asnicar, F., Sanders, J. G., Belda-Ferre, P., Al-Ghalith, G. A., Kopylova, E., McDonald, D., et al. (2019b). Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nature Communications*, 10(1):1–14.

Zimmermann, T., Mirarab, S., and Warnow, T. (2014). BBCA: Improving the scalability of *BEAST using random binning. *BMC Genomics*, 15(6):1–9.