Recent Progress on Methods for Estimating and Updating Large Phylogenies

Paul Zaharias*and Tandy Warnow†
University of Illinois Urbana-Champaign

September 2021

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.

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., microbial phylogenies of 100,000 or more sequences or 10,000 whole genomes) is very difficult using standard approaches, except perhaps when supercomputers are available.

In this paper we present a variety of techniques that use divide-and-conquer in order to scale computationally intensive but highly accurate methods to large and even ultra-large datasets. An example where divide-and-conquer is used for maximum likelihood tree estimation is provided in Park et al. (2021): the input sequence dataset is divided into disjoint subsets, maximum likelihood trees are estimated on the subsets, and then these subset trees are merged together. By design, the recent divide-and-conquer methods have fast techniques for the initial stage (dividing into subsets) and the final stage (merging disjoint trees), so that this approach has both high accuracy, low computational effort, and excellent scalability to large datasets.

*ORCID: 0000-0003-3550-2636
†ORCID: 0000-0001-7717-3514

This study presents advances in four main topics: divide-and-conquer methods for 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. Thus, while this survey is specifically 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.

Due to space constraints, this survey is perforce limited, and some promising approaches will be omitted or not covered in adequate depth.

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. 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 alignment accurate methods to large datasets. These methods (e.g., Smith et al. (2009); Liu et al. (2009); Mirarab et al. (2015); Smirnov and Warnow (2021); 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).

The most accurate of these divide-and-conquer strategies is MAGUS, which substantially improves on the previous most accurate method (PASTA) through the use of a new technique, the Graph Clustering Merger, for merging a set of disjoint alignments; 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). Figure 1 from Smirnov (2021) demonstrates how three variants of MAGUS produce more accurate alignments than leading alignment methods on the HomFam benchmark, with up to 98,681 sequences.

3 Recent Advances in Maximum Likelihood Tree Estimation

Maximum likelihood 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., 2015), 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 be comparable to RAxML with respect to topological accuracy (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 a tree in about 5 days using 32 CPUs.

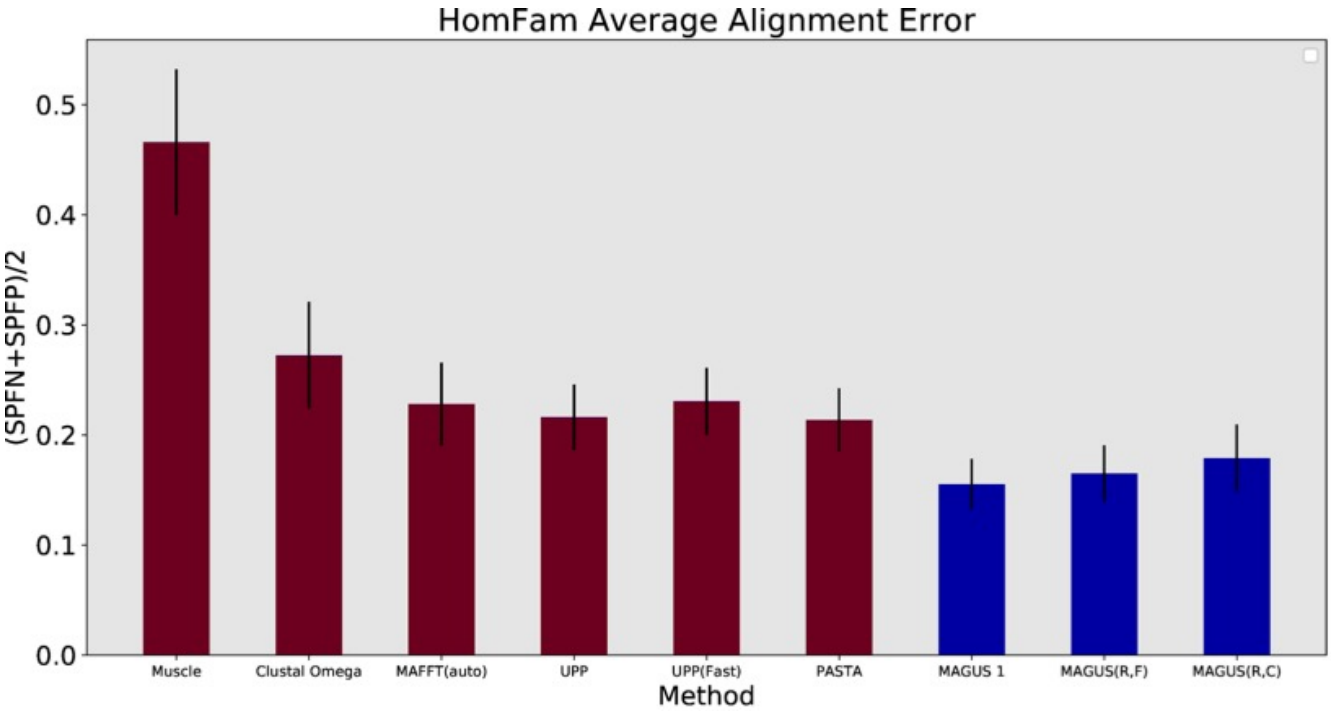

Figure 1: **Average alignment error on HomFam datasets with up to 93,681 sequences.** Alignment error rate SPFN is the fraction of the reference pairwise homologies missing in the estimated alignment and SPFP is the fraction of the inferred pairwise homologies that are not in the reference alignment.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

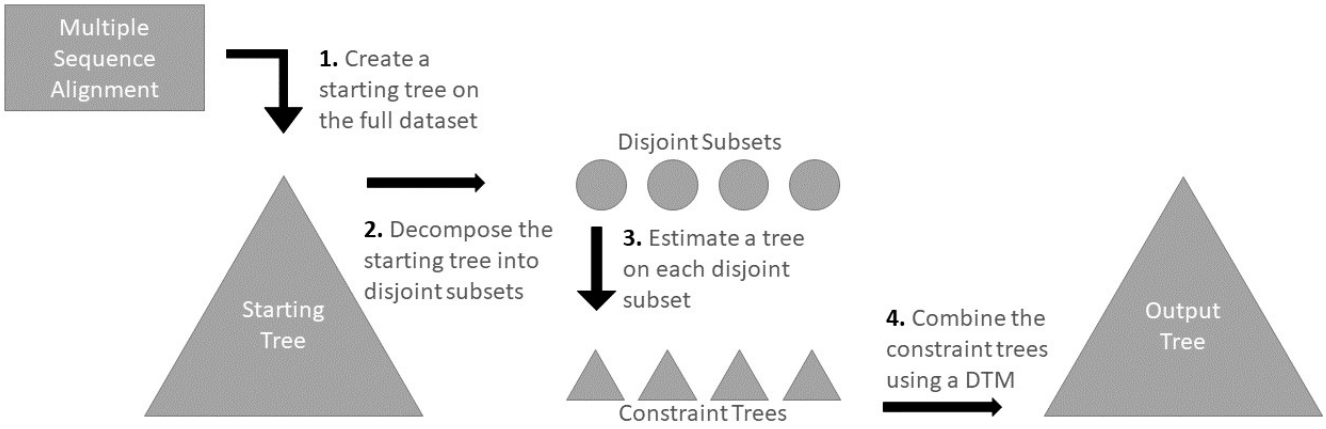

Figure 2: **DTM Pipeline for constructing a tree from an input sequence alignment using maximum likelihood**. (1) A starting tree is computed (e.g., using FastTree 2 or IQ-TREE 2). (2) Edges are deleted from the starting tree to produce small subsets. (3) Trees are estimated on the subsets using a selected 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. Figure from Park et al. (2021)

Although FastTree 2 clearly dominates RAxML for speed and memory usage and can be comparable in topological accuracy, recent research has shown that FastTree 2 can have reduced topological accuracy when the input alignment contains many fragmentary sequences (Park et al., 2021) or is otherwise very gappy (Sayyari et al., 2017) and when the sequences have evolved under heterotachy (Park et al., 2021); in contrast, RAxML seems more robust to those conditions (see Figure 3).

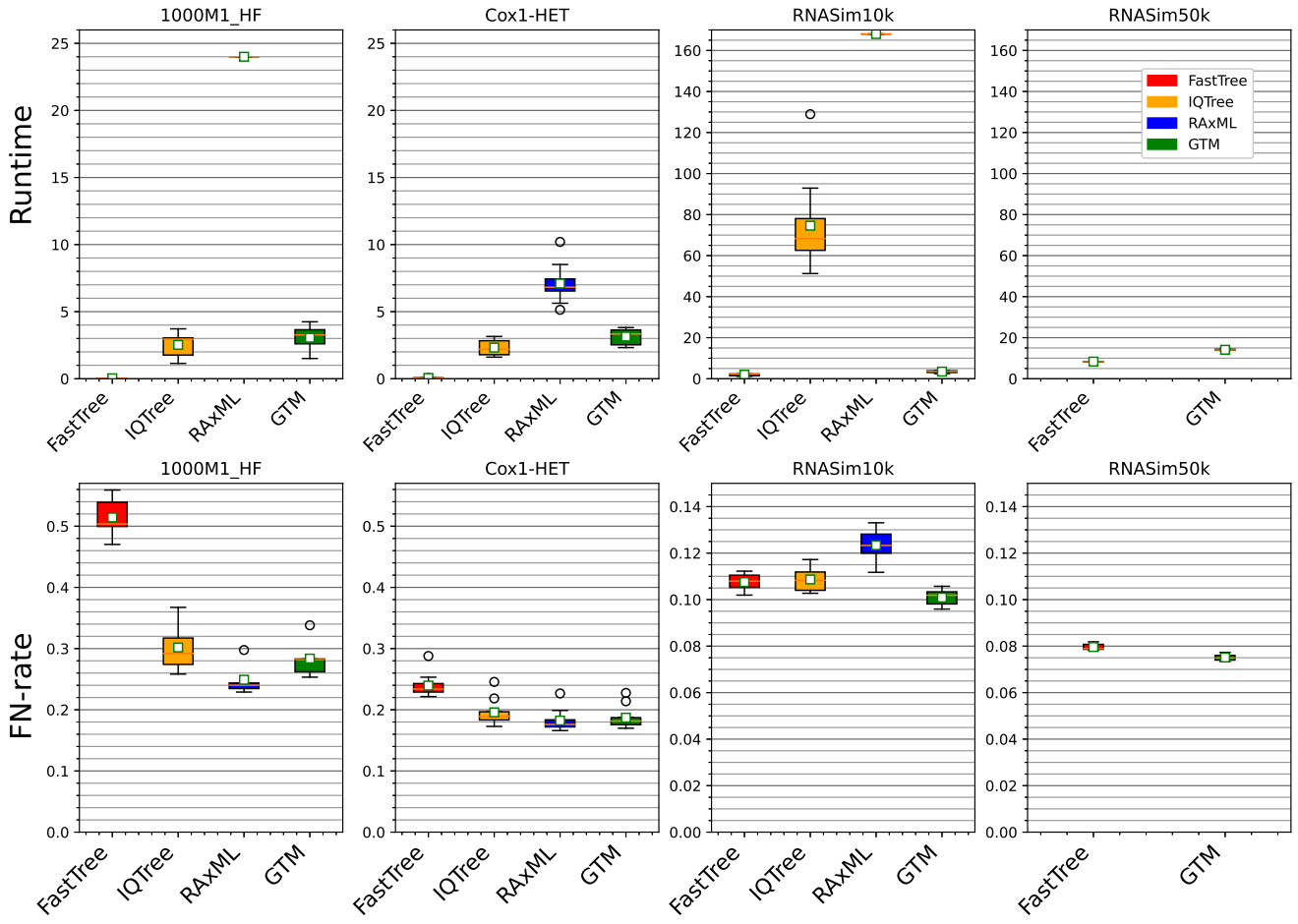

Figure 3: We compare standard maximum likelihood methods (RAxML-ng, IQ-TREE 2, and FastTree 2) to a divide-and-conquer pipeline using the Guide Tree Merger (GTM) on four simulated datasets with 1,000 to 50,000 sequences. 1000M1-HF is a model condition with 1,000 sequences that includes fragmentary sequences, Cox-HET is a model condition with 2341 sequences with heterotachy, RNASim model conditions evolve under selection and have 10,000 or 50,000 sequences. 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 produced a tree with 100% topological error. Figure adapted from Park et al. (2021)

Several strategies have been developed to overcome the burden of computationally intensive maximum likelihood analyses. One such approach uses taxonomic information about the input sequences to constrain the search space; these approaches are discussed in the species tree section as they are not as relevant to gene tree estimation due to the potential discordance between gene trees and species trees. Other divide-and-conquer techniques have been developed that do not use external taxonomic information. 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 approaches to large-scale tree estimation (Bininda-Emonds, 2004), but the requirement to use supertree methods (which are not yet very fast) constrains the scalability of these approaches (Warnow, 2019).

To overcome this limitation, a new type of divide-and-conquer approach has been developed that divides the input dataset into disjoint rather than overlapping sets, estimates trees on these subsets, and then merges the trees together using information obtained in the input. This approach, referred to as "Disjoint Tree Mergers" (DTMs) (see Figure 2), can be used to estimate both gene trees (in which case the subset trees are computed using maximum likelihood) and species trees from multi-locus datasets, and are statistically consistent for both types of analyses.

Figure 3 shows results from Park et al. (2021), comparing a DTM pipeline (using the Guide Tree Merger

(Smirnov and Warnow, 2020)) to two leading maximum likelihood methods (RAxML-ng and IQ-TREE 2). The GTM pipeline matches or improves on the topological accuracy compared to 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 100% error on that model condition and the IQ-TREE 2 analysis fails to return a tree.

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. (2019) 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). 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., 2014), but ASTRAL is generally faster on large datasets. ASTRID (Vachaspati and Warnow, 2015) is another fast and scalable summary method that is often comparable in accuracy to ASTRAL, 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, 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, 2019; Smirnov and Warnow, 2020). The third 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). 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 much faster than ASTRAL. Interestingly, it is also competitive with ASTRAL for accuracy – and sometimes more accurate!

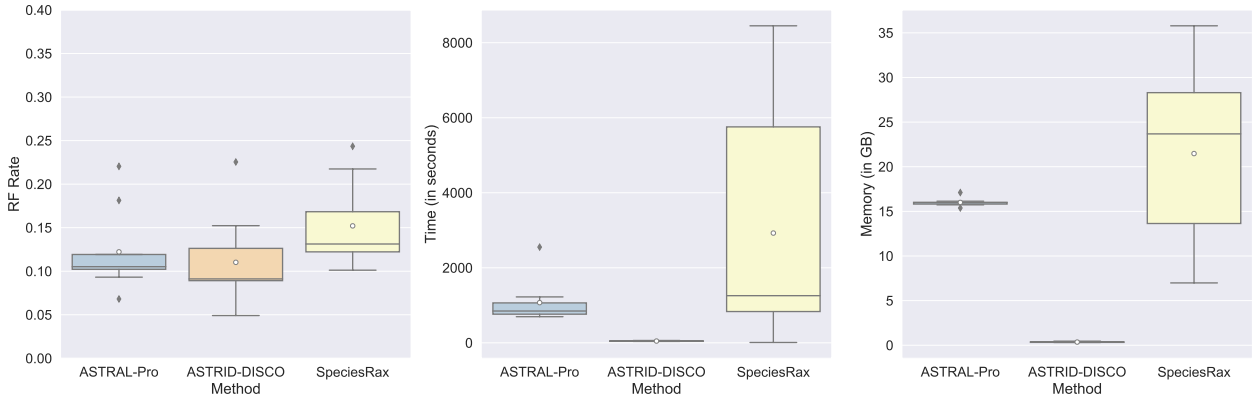

Figure 4: Species tree error (Robinson-Foulds error rates), wall clock running time (s), and peak memory usage of ASTRAL-Pro, SpeciesRax, and ASTRID-DISCO on a simulated set (under GDL and ILS) of 1001 species and 50 estimated gene trees. (Figure taken from Willson et al. (2021).)

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 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. Finally, methods that can construct species trees from MUL-trees can be used.

A recent theoretical advance is the proof that modifications of ASTRAL to deal with multi-copy gene family trees are statistically consistent under statistical models of GDL evolution (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. 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), SpeciesRax (Morel et al., 2021). However, of these methods have been shown to be as accurate as ASTRAL-Pro.

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 species tree estimation 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.

4.3 General techniques to reduce runtime

Taxonomic information can be used to constrain the search space, and hence improve running time. For example, the different species in the input could be organized into clades that are consistent with an external taxonomy and trees on these clades could be estimated and then rooted through inclusion of outgroups. By design, this approach produces a rooted tree that is compatible with the selected taxonomy. PyPHLAWD (Smith and Walker, 2019) and PhyLoTA (Sanderson et al., 2008) are examples of pipelines using such strategies, and this type of approach has been used in several phylogenomic analyses (e.g., Asnicar et al. (2020); Janssens et al. (2020)). However, if the taxonomy has errors or if the sequences in the input are incorrectly taxonomically labeled, accuracy can be reduced; hence, these techniques are often combined with opportunities for the user to correct potential mistakes. In addition, rooting subset trees using outgroups can be unreliable (Tian and Kubatko, 2017). This type of approach

is therefore useful but presents challenges.

Disjoint Tree Merger (DTM) pipelines have been used with ASTRAL and RAxML concatenation analysis for multi-locus species tree estimation, where they have reduced computational effort and maintained or improved topological accuracy (Molloy and Warnow, 2019; Smirnov and Warnow, 2020). DTM pipelines do not use taxonomic information, which provides both advantages and disadvantages compared to the methods that use taxonomic information.

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.

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), 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; recent studies (Balaban et al., 2020, 2021; Wedell et al., 2021) showed that APPLES can run on trees with 200,000 leaves and is much faster than both pplacer and EPA-ng. APPLES-2 is an improvement on APPLES with respect to accuracy and running time, and also scales to at least 200,000 sequences. However, even APPLES-2 does not match the accuracy of pplacer. Finally, UShER is parsimony-based and very fast, but has not been compared to pplacer, APPLES, or APPLES-2.

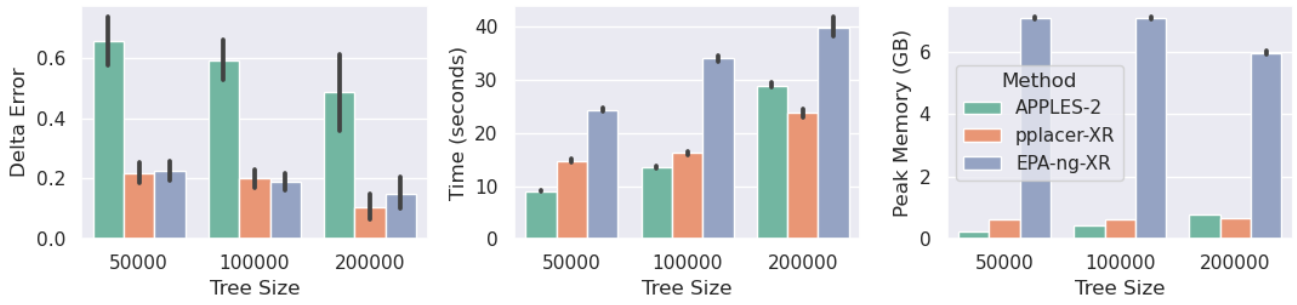

Figure 5: Phylogenetic placement of fragmentary sequences into large backbone trees. Results shown are averaged across 1,000 queries on the 50,000- and 100,000-taxon backbone trees and 200 queries on the 200,000-taxon backbone trees. Left: placement delta error. Central: running time. Right: peak memory usage. EPA-ng-XR and pplacer-XR both use the XR framework and have the best accuracy.

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 it is faster, uses less memory, and is more accurate than pplacerDC; in addition, it has also been able to scale to trees with 200,000 leaves whereas pplacer-DC scales only to 100,000 sequences.

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, we find an edge e' in the tree T corresponding to the edge e, and we place the query sequence into that 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.

Phylogenetic placement is also useful when the input sequence dataset exhibits sequence length heterogeneity, as tree estimation methods can have poor accuracy under such conditions (e.g., see FastTree's poor topological accuracy for datasets with fragmentary sequences in Sayyari et al. (2017)). Furthermore, phylogenetic placement is useful for adding reads or just partially assembled sequences into trees, which in turn is useful for taxon identification when the tree is a taxonomy.

Figure 5 compares pplacer-XR (i.e., pplacer used within the XR framework) and EPA-ng-XR (i.e., EPA-ng used within the XR framework) to APPLES and APPLES-2 in placing fragmentary sequences into trees of increasing size, from 50,000 to 200,000 sequences. These are trees that are potentially too large for pplacer or EPA-ng to run on without substantial computational resources, but using the XR framework allows both to easily run to completion. Delta error measures the additional topological error produced for each phylogenetic placement. This figure shows that APPLES-2 is a substantial improvement over APPES: it has much lower delta-error than APPLES and is also faster and uses much less memory. EPA-ng-XR and pplacer-XR, are nearly equal in accuracy to each other and both are much more accurate than APPLES-2 and APPLES. We also see that pplacer-XR is faster and has much lower memory requirements than EPA-ng-XR. Finally, APPLES-2 is the fastest of the four methods on trees with 50,000 and 100,000 leaves, but pplacer-XR is only slightly slower than APPLES-2 on those datasets and is faster (by a small amount) than APPLES-2 on the 200,000-leaf trees. Thus, APPLES-2 and pplacer-XR are the best performing of these methods, each providing an advantage over the other in a different part of the parameter space.

5.2 Adding species to species trees

To add a species into an existing species tree, it can help to consider heterogeneity across the genome due to processes such as ILS. INSTRAL (Rabiee and Mirarab, 2020b) is a recent example of such a method. 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.

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), highlighting the techniques for scaling excellent but computationally intensive methods to large datasets.

However, due to space constraints, we did not discuss all the relevant problems for large-scale tree estimation, including how to efficiently and accurately estimate the 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.

In closing, we note that the currently available methods for large-scale analysis are already much more accurate than earlier methods, and the development of new pipelines that integrate techniques (such as disjoint tree mergers) for scaling strong methods to large datasets is likely to provide further improvements. Thus, we predict that new methods will be developed that will advance the capability of biologists to estimate accurate alignments and trees and then use these in biological discovery.

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.

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, available on bioRxiv, doi=10.1101/2021.02.14.431150.

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.

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

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.

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

Dibaeinia, P., Tabe-Bordbar, S., and Warnow, T. (2021). FASTRAL: improving scalability of phylogenomic analysis. *Bioinformatics*, 37:2317–2324. `https://doi.org/10.1093/bioinformatics/btab093`.

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.

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

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.

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.

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., 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.

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

Markin, A. and Eulenstein, O. (2021). Quartet-based inference is statistically consistent under the unified duplication-loss-coalescence model. *Bioinformatics*. doi: 10.1093/bioinformatics/btab414.

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., 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. (2014). ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*, 30(17):i541–i548.

Molloy, E. K. and Warnow, T. (2019). 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*.

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. (2015). 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.

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

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

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.

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. 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.

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., 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*.

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.

Smirnov, V. (2021). Recursive MAGUS: scalable and accurate multiple sequence alignment. *PLOS Computational Biology*. In press, preprint available in bioRxiv 2021.

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. (2021). MAGUS: multiple sequence alignment using graph clustering. *Bioinformatics*, 37(12):1666–1672.

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.

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.

Tian, Y. and Kubatko, L. (2017). Rooting phylogenetic trees under the coalescent model using site pattern probabilities. *BMC evolutionary biology*, 17(1):1–11.

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.

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. *Bioinformatics and Phylogenetics*, pages 121–150.

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.

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`.

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, 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. (2019). 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.