1 Results from Prior NSF Support

PI Tandy Warnow: DBI 1062335 (PI: Warnow), 2010-2015 (no-cost extension to June, 2015). Collaborative Research: Novel Methodologies for Genome-scale Evolutionary Analysis of Multi-locus data. Total award $349,999. Publications: [119, 12, 13, 16, 79, 78, 120, 121, 63, 116, 80, 123, 15]. *Intellectual merit:* The most significant contribution were new methods (ASTRAL [80] and statistical binning [79]) for species tree estimation in the presence of incomplete lineage sorting.

Broader Impacts: Freely available open-source software including ASTRAL [80] and statistical binning [79], that were used in biological dataset analyses [116, 63]. The project held summer software schools at UT-Austin, NESCENT, and the Evolution 2014 meeting, and online bioinformatics course materials were also provided.

Human resource development: Siavash Mirarab and Md. Shamsuzzoha Bayzid (doctoral students of Warnow) are expected to graduate with PhDs in Spring 2015.

Co-PI Chandra Chekuri: NSF CCF-1016684 (PI: Chekuri): "AF: Small: Approximation Algorithms for Graph and Combinatorial Optimization Problems." 09/01/2010 to 8/31/2014
Intellectual Merit: Developed approximation algorithms for submodular optimization, both maximization and minimization via continuous extensions and mathematical programming framework. Developed algorithms for several node-weighted network design problems and for maximum disjoint paths in node-capacited graphs and in bounded treewidth graphs. Ten publications by the PI were supported by this grant [25, 34, 35, 29, 28, 32, 27, 30, 31, 33], and approximately another ten were written independently by the students of the PI.

Broader Impacts: The work on flow-cut gaps for polymatroidal networks was motivated by and found applications in network information theory, and the work on submodular optimization, in particular the multilinear relaxation approach for maximization problems, has found applications in machine learning.

Human Resource Development: Three PhD students and an MS student graduated under the PI's supervision with support from the grant. All three PhD students are faculty members in Computer Science departments, two in the United States and one in UK.

Satish Rao. NSF 0963904 (PI: Rao), "Medium: Collaborative Research: Geometric Network Analysis Tools: Algorithmic Methods for Identifying Structure in Large Informatics Graphs", from August 27, 2009 to September 1, 2013. Total award: $418,011.

Intellectual Merit: The project used traditional and recently-developed approximation algorithms for the graph partitioning problem as "experimental probes" of large informatics graphs in order to characterize in a more robust and scalable manner the structural and dynamic properties of very large informatics graphs. We developed tools that have sufficient algorithmic and statistical flexibility to characterize the local and global structures of large networks in a rich and robust way. This project extended recent theoretical and algorithmic developments to derive improved algorithms based on geometrical algorithms, and applied them to real-world problems.

Broader Impacts: The project enhanced interdisciplinary education at Berkeley and Stanford, and more generally. The project developed an empirical testbed for testing algorithms for topic modeling. It includes implementations of many methods, as well as data repository, data generators, and several evaluation techniques; software developed by the project is publicly available on github.

Human Resource Development: This project supported a postdoctoral fellow, two graduate students, and two undergraduate students. Dr. Virginia Vassilevska William was supported as a postdoctoral fellow, and a portion of her work on improving the running time for matrix multiplication as well as her work on improving the complexity for a variety of shortest path problems in graphs; she is now a professor at Stanford.

2 Background

The centrality of evolution in biology is expressed by the famous statement of Dobhzansky, who said "Nothing in biology makes sense except in the light of evolution" [42]. At its simplest, this implies that the *Tree of Life* will provide a comparative framework for understanding biology [36]. However, the evolutionary trees of individual parts of the genome ("gene trees") can differ from evolution at the species level, and both are of interest. For example, understanding the evolution of genomic regions that code for proteins is useful for understanding protein function and structure, protein-protein interactions, and other biological questions (this is how Jonathan Eisen defined "phylogenomics") [44, 97, 96]. Comparisons of gene trees and species trees together enable a deeper understanding of how species respond to their environments. Accurate phylogenies are useful for detecting selection, for estimating dates of speciation events, etc. Finally, taxonomic classifications of microbiome data (which are generally very short sequences, often with substantial sequencing error) depends on having highly accurate species trees as references [83, 77]. Thus, phylogeny estimation, whether of gene trees or species trees, is a basis for much biological research.

Phylogenetic estimation has a solid foundation in statistical inference [49, 50], and phylogenetic trees are generally computed using statistical inference methods (maximum likelihood and Bayesian MCMC estimation), based on stochastic models of evolution. These models range from fairly simple ones that model sequence evolution using substitutions of individual nucleotides or amino acids (see discussion of different models in [57]), to ones that model sequence evolution using insertions and deletions (indels) in addition to substitutions (e.g., [111, 112, 73, 95, 21]). There is also great interest in the inference of evolution at the genome-scale (i.e., the other meaning of "phylogenomics"). For example, research in recent years has focused on how different parts of the genome can have different phylogenies due to processes such as incomplete lineage sorting [75], gene duplication and loss [48, 58, 74], and horizontal gene transfer [56]. Therefore, phylogenomics projects that examine multiple parts of the genome ("loci") and estimate species trees under statistical models that address gene tree heterogeneity are increasingly common [63, 116].

However, most statistical estimation methods are computationally expensive. For example, maximum likelihood phylogeny estimation is NP-hard [89], and heuristics based on hill-climbing and randomization techniques are used to explore the exponentially-sized treespace. Maximum likelihood heuristics are in common use (e.g., RAxML [101], a popular program for maximum likelihood phylogeny estimation, has been cited more than 6000 times), but are too expensive to use on datasets that contain many thousands of species and loci (genomic regions, sometimes comprised of genes). Even some datasets with small numbers of species but large numbers of sites can be challenging to analyze; for example, a RAxML maximum likelihood analysis of an avian dataset with approximately 50 species and several million sites took more than 200 CPU years and 1TB of memory [63].

In addition, model complexity increases the running time for likelihood-based methods. As a result, the models underlying statistical methods are typically simplified in order to make them usable on larger datasets. For example, most sequence evolution models assume homogeneity of the evolutionary process – an assumption that is known to be violated on biological datasets, and especially on biological datasets spanning large evolutionary distances. Therefore, maximum likelihood phylogenetic estimation software under non-homogeneous models of sequence evolution has been developed [22]. However, maximum likelihood methods based on non-homogeneous models are too slow to use on even moderately large datasets, and so simpler models that assume homogeneity are used instead.

The combination of these two issues – simplifications of model complexity and limitations of statistical methods to small datasets – means that large-scale phylogenetic estimation is based on methods that may not be able to provide adequate accuracy on biological datasets spanning large evolutionary distances.

We propose to develop novel discrete optimization techniques and graph-theoretic methods, including graph-based divide-and-conquer strategies, to develop improved phylogeny estimation methods. We specifically target three related computational problems in phylogenomics, where the combination of theoretical computer science and statistical inference has the potential to result in new methods with better accuracy and scalability, and their use may result in more accurate biological discoveries:

- Aim 1: Algorithms to assemble a species tree from estimated species trees on subsets of the taxa (the "Supertree Problem"),
- Aim 2: Algorithms to assemble a species tree from estimated gene trees, taking biological causes for discord between gene trees and species trees into account ("Species Tree Estimation"), and
- Aim 3: Algorithms that enable computationally expensive statistical methods to analyze large datasets using divide-and-conquer ("Scaling Statistical Estimation Methods").

For each of these scientific aims, we seek to develop algorithms that are *statistically consistent* under realistic models of evolution, that run in *polynomial time* and produce results on ultra-large datasets within reasonable timeframes, and that are *highly accurate on benchmark datasets*. This research program therefore requires a combination of expertise - discrete algorithms, graph theory, probability theory, statistical inference, and rigorous evaluation of methods using both biological and simulated datasets. Finally, we are actively engaged in collaborations with biologists in major phylogenomics projects around the world (e.g., the Avian Phylogenomics project [63]) and the Thousand Plant Transcriptome project [116]), and so our methods will be used to analyze datasets in these collaborations.

Below, we describe the problems in greater detail, the biological and statistical framework for each problem, and the computational and mathematical approaches we will use to achieve the goals.

3 Algorithms for Aim 1: Supertree Construction

3.1 Background

We begin with a classic problem in computational biology – constructing a tree from a set of smaller trees. This is the "Supertree" problem, which has a large literature (an entry into this literature can be obtained from [19, 104, 3, 76]). Moreover, given the many millions of living species, methods that construct large trees by combining smaller trees have long been considered a necessary complementary approach for estimating the Tree of Life [20].

We address the classic "supertree" problem, which models the most common usage of this technique. Since nearly all phylogeny estimation methods produce unrooted trees, the input to the supertree problem is a set $\mathcal{T}$ of unrooted leaf-labelled trees, which are called "source trees". The leaves of the source trees are taken from a set S (the set of species, or "taxa"), and every element in S appears as a leaf in at least one tree in $\mathcal{T}$. The scientific objective is to combine these trees into an estimated species tree on the full set S, that comes close to the true species tree on S.

When all the source trees can be combined into a tree on the full set of species, without any conflict, then the source trees are said to be "compatible", and the tree that agrees with all the source trees is called a "compatibility tree". Thus, one desired objective in supertree estimation is to determine if the source trees are compatible, and to return the compatibility tree if it exists. However, the compatibility problem is NP-complete [103]. Furthermore, biological source trees nearly always have some estimation error. Therefore, supertree methods are typically described as optimization problems, so that the input is a set $\mathcal{T}$ of source trees, and the output is a tree T on the full set of species that optimizes some criterion. Examples of optimization criteria in supertree estimation include:

- **Matrix Representation with Parsimony [11]:** The set $\mathcal{T}$ is represented with a binary matrix called the MRP matrix, and the objective is to find a tree with the best maximum parsimony [52] score (where maximum parsimony is the Hamming distance Steiner Tree problem).
- **Minimum Robinson-Foulds Supertree [8]:** The objective is a tree T that minimizes the sum

of the Robinson-Foulds (RF) distances to the trees in $\mathcal{T}$. The RF distance between two trees T, T' on the same leafset is $|C(T) \triangle C(T')|$, where $C(t)$ is the set of bipartitions of the tree t induced by deleting single edges of tree t. The set $C(t)$ is called the "bipartition set" of t, and the tree t is uniquely defined by $C(t)$. The definition of the *RF* distance is extended to trees on different leafsets by restricting the two trees to the intersection of their leafsets.

- **The Minimum Quartet Distance Supertree (MQDS) Problem:** The objective is a tree T that minimizes the minimum total *quartet distance* to set $\mathcal{T}$, where the quartet distance between two trees is the number of four-species subsets of S that induce different quartet trees in the two trees (a four-species subset that is not fully contained in one of the trees does not contribute to the distance).

It is not hard to see that when the input source trees are compatible, then an optimal solution to any of these problems will be a compatibility supertree (i.e., a tree that induces each of the input source trees). However, each of these optimization problems is NP-hard, and approaches to solving these problems are heuristics without performance guarantees.

MRP heuristics are the most popular techniques for supertree estimation, and they have good accuracy [109], as explored on simulated and biological dataset analyses. MQDS heuristics also have very good accuracy (better in some cases than MRP [108]), and MinRF heuristics also show promise. However, supertree methods are nearly always based on techniques that explore an exponentially sized treespace using hill-climbing and randomization techniques, and so are not scalable to large datasets. One of the exceptions to this paradigm is the Quartets MaxCut (QMC) [98, 99] method developed by PI Satish Rao and colleague Sagi Snir). QMC runs in polynomial time, but because it examines all $\Theta(n^4)$ four-leaf subsets, it is too slow for large datasets.

PI Warnow and her students tested the use of QMC as a supertree method [110, 108]. In this setting, all quartet trees are computed for all source trees, and then QMC is applied to the multi-set containing all these quartet trees. This approach had high accuracy – better in many cases than MRP – but was computationally intensive since it had to examine $\Theta(n^4)$ quartet trees. Using a sparser sampling of the quartet trees for each source tree improved the speed but also tended to reduce accuracy.

An alternative approach to the MQDS problem constrains the allowed search space by providing a set X of *allowed bipartitions*, and seeks an optimal solution to MQDS subject to the constraint that the output tree have all its bipartitions in X. In other words, we pose the following problem:

MQDS with constrained search space (MQDS-C):

- Input: Set $\mathcal{T}$ of source trees on species set S, and set X of bipartitions on S
- Output: Tree T on S minimizing the quartet distance to $\mathcal{T}$, subject to $C(T) \subseteq X$ where $C(T)$ denotes the set of edge-induced bipartitions of T.

Unlike the MQDS problem, which is NP-hard, the MQDS-C problem can be solved in polynomial time using dynamic programming; a DP algorithm for MQDS-C is implemented in the publicly available ASTRAL software [80]. ASTRAL also has automated ways of computing the set X of allowed bipartitions from the input set. However, how X is set will impact the accuracy and running time of this approach: if X is too small or is not selected well, then solutions to this problem will not provide good accuracy (i.e., the supertree that is computed will not reflect the species tree well), and if X is too large then the running time will be excessively high. Because ASTRAL was not developed to be a supertree method (but rather to address Aim 2's challenge of estimating the species tree in the presence of gene tree heterogeneity due to a biological process referred to as incomplete lineage sorting), we do not know how well ASTRAL will perform as a supertree method.

SuperFine: improving the scalability of supertree methods. SuperFine [107] is another approach to supertree estimation, but operates as a "booster" for a given supertree method. Thus, instead of being a supertree method, it is used *with* some other base supertree method, and seeks to improve the scalability (and accuracy) of the method. Thus, SuperFine+MRP refers to SuperFine used with MRP heuristics, SuperFine+QMC refers to SuperFine used with QMC, etc. SuperFine operates as follows: First, SuperFine computes an unrooted constraint tree from the input source trees using the "strict consensus merger", and this constraint tree will only contain edges that are shared by all the source trees. Computing the strict consensus merger is very fast, and takes low degree polynomial time. In general, the constraint tree will not be fully resolved (i.e., it will have nodes of degree greater than three) - although the degree of resolution will depend on properties of the source trees, including the overlap pattern and how compatible they are [93]. The constraint tree is then refined into a binary (fully resolved) tree using the base supertree method (e.g., QMC or MRP heuristics) in a computationally efficient way: each node of degree d in the constraint tree defines a new set of source trees, each on d leaves, and the supertree method is applied to this new set of source trees. When the maximum degree D in the constraint tree is relatively small compared to the number of number of taxa in S, then this approach provides a substantial speed-up for the supertree method. However, in some biological datasets, the constraint tree has nodes of very high degree, which means that using SuperFine may not provide much speed-up at all [82]. See [107, 84, 108] for SuperFine used with various different supertree methods, and [82] for a study of a parallel implementation of SuperFine.

Summary of current limitations for supertree methods: Despite the great interest in supertree methods, nearly all supertree methods are limited to at best moderately large datasets (perhaps several hundred species) because of their algorithmic designs, which use heuristic search strategies that combine hill-climbing and randomization to explore tree space. SuperFine can provides a large speed-up for some biological datasets, but the speed-up depends on how well resolved the initial constraint tree is, and our research demonstrated that for some biological datasets, there was little or no speed-up provided through SuperFine. More importantly, we conjecture that on supertree datasets with large numbers of species and/or source trees, the constraint tree produced by SuperFine will provide little resolution, and hence SuperFine will not provide much benefit. Taken together, these observations suggest that *highly accurate and fast supertree estimation on datasets containing thousands or species and hundreds of source trees is not likely to be acheived using existing supertree methods.*

3.2 Proposed Research

Our overall goal for Aim 1 is to develop fast supertree methods that can analyze the large biological datasets that are being generated, where individual source trees might only contain a few thousand species each, but which together contain many thousands to millions of species. Speed is not the only criterion, however; these methods need to also be highly accurate, and able to provide substantially well resolved supertrees even in the presence of estimation error in the source trees.

Modelling the source tree distribution. There are two aspects of the source tree distribution that are we will study: the *taxon sampling strategy*, which determines the set of species (or "taxa") in each subset tree, and the *source tree estimation error*. The first aspect must reflect how biologists select their taxa, and thus involves the choice of clades (species sets below a node within the species tree) or backbone sets (randomly selected sets of taxa from throughout the species tree), and also the shape of the species tree. We will build on mathematical models for tree shape (e.g., birth-death models, but see also [4]), and explore published biological datasets to develop appropriate taxon-sampling models.

The second aspect, modelling source tree estimation error, is also challenging. Since we are interested in the use of supertree methods on biological datasets, we will model source tree estimation error by considering phylogenetic estimation given sequence alignments generated under standard Markov models of evolution.

Theoretical understanding of MQDS-C: The MQDS-C problem seeks an optimal solution to the MQDS problem, subject to the constraint that the solution take its bipartitions from a set X provided in the input. The question we will address here is how X affects the accuracy of exact solutions to MQDS-C; in other words, what can we say about the difference in the trees that are obtained by an optimal solution to MQDS and an optimal solution to MQDS-C? To explore this question, we will assume that the source trees are obtained from a distribution defined by the species tree (see previous paragraph), with or without source tree estimation error.

Modifying MQDS for sparse subsets of quartets. The MQDS problem is stated in terms of quartet distances, where all subsets of four leaves contribute to the distance; here we consider the case where the problem formulation is changed so that only some four-leaf subsets count towards the distance. The motivation for this reformulation is the observation that QMC (and its extension to weighted Quartets MaxCut [7]) are too expensive to use on large datasets, since they begin by encoding every source tree by its set of quartet trees. However, it may be possible to find a sparse encoding of each source tree as a set of quartet trees, while ensuring that QMC, Weighted QMC, and similar methods have high accuracy.

We will explore this question theoretically, seeking to characterize the impact on tree accuracy based on how the set of four-species subsets is selected. For example, we will consider quartet-tree representations of source trees that are guaranteed to include the "short quartets" (the quartets formed by selecting the nearest leaf in each of the four subtrees around an edge in the source tree), as these representations produced better results than random samplings [110, 108]). We will explore this question by assuming a probability distribution on the set of source trees.

Related to this issue, we note that a polynomial time approximation scheme (PTAS) for the MQDS problem has been developed [64], and that approximation algorithms for the variant of MQDS based on a random sample of quartet trees have also been developed [100]. However, results shown in [110, 108] suggest that QMC analyses based on sparse random subsets of quartets for each source tree are not likely to produce highly accurate species trees. We will explore approximation algorithms for variants of MQDS where we use more suitable representation of the source trees by quartet trees.

Distance-based supertree methods. Source trees for the supertree problem are based on phylogenetic trees that are typically estimated from molecular sequence data, and so come with edge lengths reflecting the model parameters (expected number of substitutions on the edge for a random site). These edge-weighted trees can be represented as *additive* matrices, which are matrices of path-distances in the tree. In some cases, these additive matrices are *ultrametric*, which means that the tree can be rooted so that the distance from the root to each leaf is the same; this is a natural outcome when edge lengths represent the amount of elapsed time. Finally, the input may not only be the set of source trees, but also associated multiple sequence alignments (MSAs) for each source tree along with a "distance" matrix computed on the MSA. Because of how phylogenetic distances are computed on molecular sequences, these distances may not satisfy the triangle inequality; therefore, estimated distance matrices are normally referred to as "dissimilarity matrices" rather than "distance matrices", to avoid an abuse of the term. As Willson [117] observed, the use of these "distance matrices" is a valuable source of information in supertree construction, since they naturally enable the estimation of branch lengths in the supertree, something that is not possible when just using the source trees alone. Willson also presented the "Build-with-Distance Supertree method", which was shown to have excellent accuracy on a collection of simulated datasets in [24].

We approach this problem by generalizing the MRD (Matrix Representation with Distances) supertree problem from [68]. The input is a set of dissimilarity matrices (which may be additive or ultrametric), and where each matrix defines "distances" (loosely defined, since they may not obey the triangle inequality) between sequences in a set of taxa. However, since the source trees can have proper subsets of the species set, the matrices too may only be on subsets of the species set. Our objective is an additive matrix on the full species set that *minimizes the total distance to the input matrices*; in other words, we seek a median tree with respect to some way of measuring distances between matrices. Common ways of measuring distances use the L_∞, L_1, and L_2 norms, but other norms can also be used. Note also that we require that the output be an additive matrix, so that we can use it to define the supertree topology and branch lengths. If we wish to compute a rooted supertree with branch lengths reflecting elapsed time, then we will require that the output be an ultrametric matrix.

While distance-based supertree estimation is not new (e.g., [117, 37, 69, 68]), most methods average the distance matrices to produce a new matrix, from which they estimate the supertree. The approach in [68], however, addresses one of the cases we identify, which is where the input matrices are additive and the objective is the additive matrix that is closest to the input under the L_2 norm. None of these distance-based methods reliably match the accuracy of MRP, but the Build-with-Distances method of [117] came close to MRP and in some cases was more accurate [24], leading the authors to conclude that distance-based supertree estimation might have potential to replace MRP.

To our knowledge the computational complexity of the optimization problems we pose have not been explicitly studied. There is a substantial literature on related problems, where the input is a single dissimilarity matrix M and the objective is an additive matrix or an ultrametric matrix that is optimally close to M, under some metric between distance matrices. Some of these problems are solvable in polynomial time, others are NP-hard but can be approximated, and some are hard to approximate; see [2, 47, 1] for an entry to this literature. Thus, there is a rich literature about related problems where the input contains only a single matrix.

Nearly every optimization problem is NP-hard for the case where there is a single input matrix, and so will be NP-hard when we allow multiple input matrices. But some of these problems are solvable in polynomial time when the input has a single matrix, and it is possible that these problems will remain polynomial time for small numbers of input matrices, or that the problems will be fixed parameter tractable. We will establish the computational complexity of these problems, and develop approximation algorithms for those that are NP-hard.

4 Algorithms for Aim 2: Combining gene trees into a species tree

4.1 Introduction

In this aim, we consider the case where the input set of source trees is obtained by estimating phylogenetic trees for different parts of the genome, and we consider differences between these estimated trees that result from biological processes such as incomplete lineage sorting [75], gene duplication and loss [122, 43, 58], and horizontal gene transfer [17, 18, 55, 56, 39]. In keeping with the scientific literature, we refer to the phylogenetic trees on the different loci (genomic regions) as “gene trees”, even though they may not be based on genic regions. Because these processes can result in gene trees that are different from the species tree, genome-scale phylogeny estimation must address the case where the input set of gene trees is highly heterogeneous. On the face of it, Aim 2 seems very similar to Aim 1, and perhaps even identical. However, each biological cause of gene tree discord creates a distribution of gene trees with properties that depend on the biological process, and these properties should be considered in the algorithm design. Importantly, for some of these biological processes, the species tree can be identified from these distributions, and so inference methods, if based on appropriate mathematical models of evolution and designed correctly, can be statistically consistent and enable highly accurate estimates of the species tree – even in the presence of massive gene tree heterogeneity.

We focus on two specific challenges in species tree estimation – incomplete lineage sorting (ILS) and horizontal gene transfer (HGT). In the case of ILS, the species history is correctly represented by a tree, and there is a rich mathematical literature on how to identify the species tree from the distribution on the gene trees. In the case of HGT, the species history is properly represented by a phylogenetic network consisting of a species tree with additional "non-tree" branches representing the HGT events. However, for many biologically realistic cases, the species tree contained in the phylogenetic network is also identifiable from the gene tree distribution. Our research focuses on developing methods to estimate the species tree in the presence of ILS and/or HGT.

4.2 When gene trees differ from species trees due to incomplete lineage sorting (ILS)

4.2.1 Background

Incomplete lineage sorting (ILS) refers to a population-level process that can result in gene trees having different topologies than the species trees. This population-level process is modelled by the multi-species coalescent [66, 65], which makes statistical inference under this model possible. ILS is considered a major confounding problem in species tree estimation [40] from multiple loci, because it is mathematically proven to happen with high probability for at least some genes. ILS is believed to occur with substantial frequency in many phylogenetic analyses of taxonomic groups, including birds [63] and plants [116].

Most of the methods for estimating species trees in the presence of ILS operate by combining gene trees, and so are called "summary methods." Many of these summary methods have been proven to be statistically consistent under the multi-species coalescent model, which means that for any model species tree (T,Θ) (where T is the rooted tree and Θ denotes the branch lengths in coalescent units) and for any $\varepsilon > 0$, there is some $K > 0$ so that given $k > K$ randomly selected true gene trees, the probability that the method returns the unrooted version of T is at least $1-\varepsilon$.

The justification for using summary methods is that when ILS can occur, standard methods for species tree estimation can be statistically inconsistent (and so may not converge to the true species tree as the number of genes increases). Worse, some standard methods are *positively misleading*, which means they may converge to the wrong tree with high probability, as the number of genes increase. For example, the standard maximum likelihood concatenation approach, which concatenates alignments for individual genes into a large supermatrix and then estimates the tree directly from the supermatrix using maximum likelihood, can be positively misleading [91, 92]. Furthermore, under sufficient levels of ILS, the most probable gene tree will not be the true species tree [41]; species trees for which this occurs are said to be in the "anomaly zone", and the gene tree that is more probable than the gene tree matching the species tree topology is called an "anomalous gene tree".

The most popular of the "coalescent-based" species tree estimation methods use likelihood calculations to find the species tree – maximum likelihood (as in STELLS [118]), maximum pseudo-likelihood (as in MP-EST [72]), or Bayesian MCMC (as in BEST [71] or *BEAST [59]); however, these likelihood-based methods are too computationally intensive to use on large datasets. Of these likelihood-based methods, only the summary methods (such as MP-EST) are able to analyze reasonably large datasets in reasonable time frames.

However, *none of these summary methods have any theoretical guarantees in the presence of gene tree estimation error*. There is also increasing empirical and experimental evidence that summary methods have impaired accuracy in the presence of sufficient gene tree estimation error [85, 16, 79, 14]. Furthermore, many summary methods are too slow to use on large datasets; even MP-EST [72], which is one of the fastest of the summary methods, is too slow to use on datasets with 100 or more species [15]. MP-EST is also limited by the requirement that the input gene trees all be rooted, a requirement that is difficult for many biological datasets.

To motivate our proposed research, we present a very simple and yet provably statistically consistent summary method. As shown in [5], under the multi-species coalescent model there are no anomalous unrooted four-leaf gene trees, which means that the most probable unrooted gene tree on any four species is topologically identical to the unrooted species tree on the four species. Hence, as the number of gene trees increases, the most frequently observed quartet tree on any set A of four leaves will, with probability converging to 1, be the unrooted species tree for A. This means that a very simple method of estimating the species tree is statistically consistent under the multi-species coalescent model: (1) examine all the gene trees, and determine the most frequent quartet tree for each set A of four species (these are called the "dominant quartets"), and (2) construct the tree that is compatible with all the dominant quartets, if it exists. This is a polynomial time method (determining compatibility of the set of dominant quartet trees is very easy). The proof of statistical consistency follows directly from the result in [5] that there are no anomalous quartet trees; hence, for all $\varepsilon > 0$ there is a large enough number of gene trees so that the dominant quartets will all be identical to the species tree (restricted to the subset of four species) with probability at least $1 - \varepsilon$.

This reasoning is also the basis for some of the well-known coalescent-based methods, such as the population tree in BUCKy [70]. However, BUCKy and other similar methods combine dominant quartet trees even when they are not compatible (by contrast, this simple method we described does not return any tree if the dominant quartet trees are not compatible).

A limitation for BUCKy and other methods that use dominant quartets is that they do not take into account the strength of the probability that the dominant quartet tree is the true species tree. In other words, this general approach does not distinguish between dominant quartet trees that are overwhelmingly likely to be the true species tree and those that are only marginally more likely to be the true species tree.

The quartet-based method, ASTRAL [80], which we described earlier, addresses this issue, and is easily seen to be a statistically consistent method under the multi-species coalescent model. However, instead of using only the dominant quartet trees, ASTRAL considers all quartet trees and the frequency in which they appear in the input set of gene trees. This consideration enables ASTRAL to have excellent accuracy on both biological and simulated data, with substantial improvements over MP-EST and other coalescent-based summary methods on biological datasets and biologically realistic simulated datasets [80].

ASTRAL solves MQDS-C, the constrained version of the Minimum Quartet Distance Supertree problem, in $O(n^2k|X|^2)$ time, where X is the set of allowed bipartitions, k is the number of genes, and n is the number of species. ASTRAL also has automated ways to define the set X from the input gene trees that enables it to have good accuracy on small to moderate sized datasets. ASTRAL run in default mode sets X to contain at least the bipartitions from the input gene trees. Given enough gene trees, every bipartition of the true species tree appears in at least one gene tree with high probability; hence, ASTRAL run in default mode is statistically consistent under the multi-species coalescent model [80].

4.2.2 Proposed research

The research described above shows recent improvements for coalescent-based species tree estimation, but the theoretical understanding of these methods is limited. Our proposed research focuses on theoretical guarantees, and in the development of methods with improved theoretical properties.

The impact of the set X on ASTRAL: While we know ASTRAL is statistically consistent under the multi-species model whenever the set X contains the bipartitions from the input gene trees [80], we know nothing about the impact of X on the number of genes needed to recover the true supertree tree with high probability. This is a theoretical question that can have substantial impact on empirical performance, and which we will investigate.

Impact of gene tree estimation error on coalescent-based species tree estimation. There are many coalescent-based methods that estimate species trees by combining gene trees; however, all proofs to date of statistical consistency require true gene trees. It is therefore a critically important open problem whether any of these summary methods are guaranteed to converge to the true species tree, given an unbounded number of gene trees with bounded estimation error. We also do not know if we can bound the error in the computed species tree as a function of the error in the estimated gene trees. These are theoretical questions that we will seek to answer.

Using phylogenetic invariants to estimate the rooted species tree from unrooted gene trees. Rooting phylogenetic trees is challenging, and for this reason most estimated gene trees are either unrooted or have uncertain locations for the root. Furthermore, when the input gene trees are unrooted, the current summary methods produce unrooted species trees. However, theory established in [5] shows that the *rooted species tree is identifiable from the distribution on five-leaf gene trees*. The theory suggests that statistically consistent estimation methods can be designed to estimate the rooted species tree. We will explore this possibility and evaluate the accuracy of rooted species trees estimated using these invariant-based methods.

4.3 Species tree estimation in the presence of HGT

4.3.1 Background

When HGT is present, the construction of the underlying species trees from multiple loci (sequence alignments and/or trees from throughout the genome) is difficult but of great interest [67, 9, 10].

Recent theoretical advances in [102, 90] establish that under random models of horizontal gene transfer (HGT), in which the amount of HGT is not too high (almost linear per gene), the species tree contained within the phylogenetic network *is identifiable from the gene trees*. Conversely, a proof of non-recoverability of the underlying species tree when there is too much HGT was also provided (Theorem 2 in [90]).

The proof of identifiability of the underlying species tree when HGT is low enough is based on the following observation: when the HGT level is low enough, then for every set A of four species, the most probable quartet tree on A will be the true species tree on A. Consequently, statistically consistent methods for estimating the species tree can be developed that operate by determining the dominant quartet trees, and then combining them using quartet-based tree estimation methods. In other words, methods such as ASTRAL and the methods we will address for Aim 2 will be *statistically consistent* techniques for estimating the species tree in the presence of bounded rates of random HGT events.

4.3.2 Proposed research

Constructing species trees in the presence of ILS and HGT. The theoretical results in [102, 90] establish that the underlying species tree can be identified when HGT is present, but does not establish identifiability when both ILS and HGT are present. We will seek to extend the theory in [102, 90] to address the case where gene trees differ from the species tree due to multiple causes, including ILS, duplication and loss, and HGT.

Understanding the impact of gene tree estimation error on species tree estimation in the presence of HGT. The theory in [102, 90] only establishes identifiability given true gene trees; hence, extending the theory to estimated gene trees with bounded estimation error is needed. Thus, we would like to know if the true species tree can be recovered from a large enough number of gene trees, when each has bounded error. More generally, we would like to develop species tree estimation methods whose error can be bounded by a function of the error in the input gene trees.

5 Aim 3: Divide-and-conquer strategies to scale statistical methods

5.1 Background

Statistical methods, generally maximum likelihood or Bayesian MCMC methods, are mainstream techniques for phylogeny estimation [60]. These techniques are statistically consistent under standard models of sequence evolution (see [105] for the proof for Bayesian methods), and the best methods of their types have good accuracy on simulated data. The simplest and most well known such methods are maximum likelihood phylogeny estimation heuristics, such as RAxML [101], under sequence evolution models such as the General Time Reversible (GTR) model (see, for example, [50]). RAxML can analyze datasets with several thousand sequences but is computationally intensive (both memory and time) on datasets with large numbers of sequences. MrBayes [61] and other Bayesian MCMC methods are more computationally expensive, because they use MCMC to perform a random walk through an exponentially-sized space, and must converge to the stationary distribution before their analyses can be considered reliable. These Bayesian MCMC methods are not even attempted on large datasets.

Statistical estimation is also used in other ways. For example, methods that co-estimate multiple sequence alignments and trees have been developed (e.g., BAli-Phy [87] and BayesCAT [95]), but these are limited to even smaller datasets than MrBayes (perhaps 50 sequences), and can run for weeks even on small datasets with only 25 sequences. *BEAST and BEST co-estimate gene trees and species trees under the multi-species coalescent model, but do not converge well on datasets with more than about 20 species and perhaps 50 to 100 genes [123]. The method in [23] co-estimates gene trees and species trees under a model that incorporates gene duplication and loss, and is even more intensive than *BEAST. These are just some of the examples of statistical methods that promise great accuracy but are enormously expensive.

5.2 Preliminary Results

PI Warnow has developed several divide-and-conquer methods to improve the accuracy and scalability of statistical estimation methods [62, 115, 15, 81]. The basic idea is to (1) quickly divide the species dataset into small overlapping subsets, (2) estimate subset trees on the small subsets using the statistical method, and then (3) combine the subset trees together into a tree on the full dataset. The first step requires clustering methods, the second step uses the statistical method of choice, and the third step requires supertree methods. Many of these methods are based on chordal graph theory to produce the division of the species into overlapping subsets, combine subset trees, and to prove theoretical properties about the trees that are computed, but the paradigm is quite general. Thus, the first step can be replaced by any clustering method, and the third step can use any supertree method.

We illustrate this idea with DACTAL [81], which was designed to enable “almost alignment-free” phylogeny estimation. The input is a set of unaligned sequences, and the output is a tree on the sequences but no multiple sequence alignment. DACTAL can be initiated in two ways: (1) computing a clustering of the sequences into small overlapping subsets using BLAST [6] or (2) computing a tree on the input sequences (through “two-phase” techniques that first align the sequences and then compute a tree on the alignment), and then using the tree to decompose the dataset into small overlapping subsets using the “pRecDCM3” decomposition (see Figure 1). In each case, the subsets that are produced are small (as small as desired, based on the user-specified threshold) and contain similar sequences. Then, a tree is computed on each of the subsets (again, using a two-phase technique), and the subset trees are combined using a supertree method into a tree on the full set of sequences. This process can iterate several times. Figure 1 presents a cartoon for the DACTAL algorithmic design, and Figure 2 shows results on three biological datasets with about 6,000 to nearly 28,000 sequences, using five iterations of DACTAL, decompositions into overlapping subsets of size 200, and the SuperFine+MRP [107] supertree method. We report results based on the final iteration; note that DACTAL is substantially more accurate than standard two-phase methods that first align and then compute maximum likelihood trees.

The second illustration is the use of DACTAL to improve the scalability of MP-EST, the coalescent-based species tree estimation method discussed earlier. As shown in [15], MP-EST’s running time grows quickly with the number of species. We modified DACTAL for use with MP-EST as follows. The input is a set of gene trees that are estimated on different loci, but where gene tree discord due

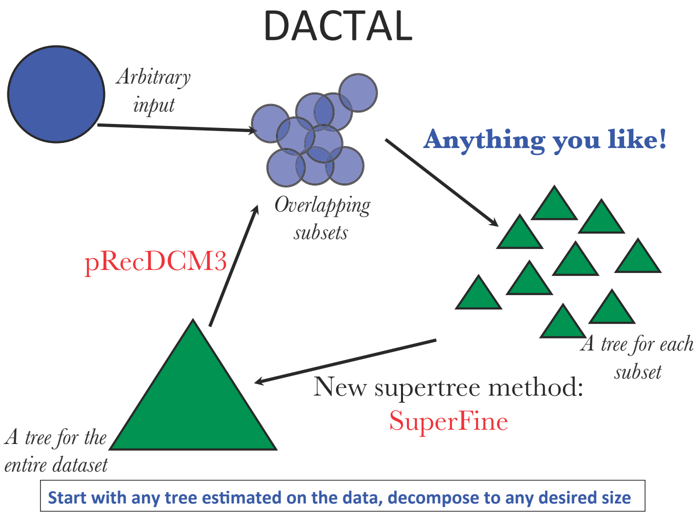

Figure 1: **DACTAL as a generic divide-and-conquer strategy.** The input to DACTAL can be quite general, but in its usual use it is a set of molecular (DNA, RNA, or AA) sequences. In the first step, the sequence dataset is divided into small overlapping subsets, where each subset has relatively similar sequences. Trees are then computed on each subset using a preferred technique (e.g., first compute a multiple sequence alignment, and then compute a maximum likelihood tree on the alignment). Because the subsets are overlapping, the subset trees are also overlapping. A supertree method is then applied to the subset trees, to produce a tree on the entire dataset. This is the end of one iteration. To continue with additional iterations, a padded version of the recursive DCM3 technique [94] (pRecDCM3) decomposition is applied, which decomposes the dataset into small overlapping subsets. In a pRecDCM3 decomposition, a centroid edge e in the tree is obtained, and a subset of the leaves is computed by taking the p closest leaves in each of the four subtrees around edge e; thus, the decomposition depends on the parameter p. This set (A) is added to the four sets of leaves in each of the four subtrees, to produce a decomposition of the dataset into four subsets, each containing the set A and otherwise being disjoint. The decomposition recurses on each subset until all created subsets have size at most M, the user provided maximum subset size. DACTAL iterates until a stopping rule is triggered (typically a fixed number of iterations or amount of time). Note that even if alignments are computed on the subsets, no alignment is returned since the alignments may not be compatible. DACTAL can also be initiated through a user-provided tree, computed using quick and approximate methods, and that the default implementation of DACTAL uses this technique.

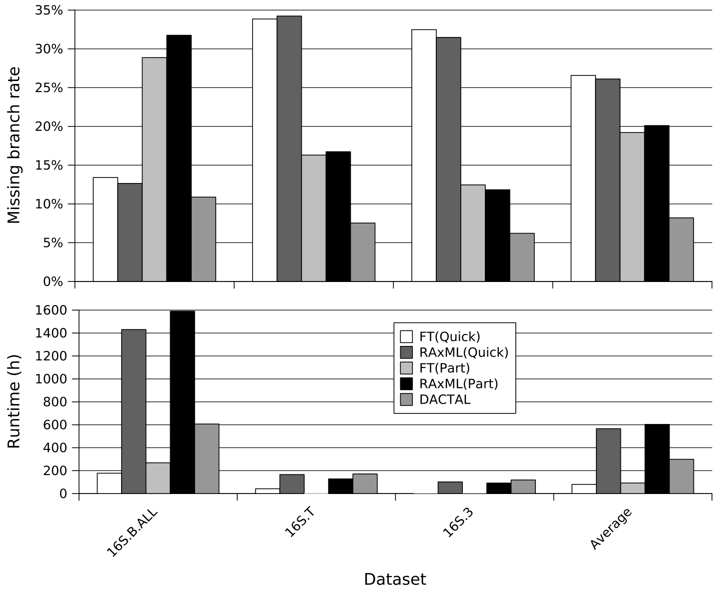

Figure 2: **Tree error rates and running time for DACTAL and two-phase methods on three large RNA datasets.** We evaluate DACTAL as a method for estimating trees from unaligned sequences, in comparison to maximum likelihood (ML) on estimated alignments. We use RNA datasets ranging in size from approximately 6,000 to nearly 28,000 sequences, from the Comparative Ribosomal Website [26], each of which has a structurally-based reference alignment and phylogenetic tree. We use either RAxML or FastTree-2 (FT) [86] for ML estimation, and compute alignments using Clustalw's Quick-Tree (Quick) command or MAFFT's PartTree (Part) command, two methods that are able to analyze datasets of this size. We evaluate accuracy of the estimated trees with respect to reference trees for these datasets, recording the *missing branch rate*. DACTAL is much more accurate than the maximum likelihood trees computed on estimated alignment. DACTAL achieves this accuracy while only computing alignments on subsets of 200 sequences, and so does not need to compute on alignment on the full dataset.

to ILS is expected. We initiated the DACTAL strategy by computing a greedy consensus tree for the input gene trees. In each iteration, we used the pRecDCM3 decomposition (see Figure 1) to divide the species dataset into overlapping subsets of at most 15 to 20 species. Each subset of species defines a new input to MP-EST, since it induces a set of smaller gene trees. We then applied MP-EST to each set of smaller gene trees, to compute an estimated species tree for the small subset of species. These smaller subset trees are then combined into a species tree using the SuperFine supertree method [107]. We ran DACTAL multiple times (later analyses showed three iterations gave very good results) to produce a set of candidate species trees. We scored each tree in the set with respect to the total quartet distance to the input gene trees, and selected the tree that had the lowest quartet distance.

As expected, DACTAL-boosting of MP-EST improved the speed of MP-EST; however, we were surprised to see that it also resulted in improved accuracy on simulated and biological benchmarks [15]. Interestingly, the improvement is not because the method is able to find better solutions to the maximum pseudo-likelihood problem that MP-EST tries to solve, and at this time we do not fully understand why it is more accurate. However, one conjecture for why it is more accurate is that the DACTAL-boosted version of MP-EST relies less on parametric optimization, and this makes it more robust. In other words, MP-EST tries to optimize numeric model parameters to fit the observed data, but its model assumes true gene trees; hence, gene tree estimation error may have a larger impact on it than on non-parametric methods.

5.3 Proposed Research

Fast Clustering Methods. These results show the feasibility of techniques like this, but also reveal limitations of the specific techniques used in DACTAL. From a running time perspective, both ways of initiating DACTAL are expensive – computing all-pairs BLAST scores is fast for small datasets but not for large datasets, and even "quick-and-dirty" techniques for computing trees are not fast enough. Thus, we will need to develop a fast approach for clustering the input species into small overlapping subsets of closely related species. Hence, we will develop new methods for clustering that are fast and produce reasonable initial settings for DACTAL, and which do not rely on existing multiple sequence alignments or calculations of pairwise distances for all pairs of sequences. Many clustering techniques are available, and we will explore these for use in this context (for example, DNACLUST [54]).

Better supertree methods. Improved supertree methods are the main focus of Aim 1, and so success for Aim 1 will directly benefit Aim 3.

Other divide-and-conquer techniques. Alternative divide-and-conquer techniques will also be considered, including the methods based on chordal graph theory from [114]. Some of these have very strong theoretical guarantees; for example, the "DCM1" method [62, 115] transforms exponentially converging methods into *absolute fast converging* methods, which are phylogeny estimation methods that are guaranteed to recover the true tree from polynomial length sequences (see [45, 46, 38] for early literature on this concept). These divide-and-conquer strategies can also be used to improve the scalability of statistical estimation methods to large datasets, and may be able to improve their theoretical properties.

6 Evaluation Plan

Aims 1-3 involve theoretical research, based on optimization problems that are relevant to practice. However, phylogenetic accuracy (i.e., how close the estimated tree is to the true tree) can only be assessed through extensive testing on benchmark datasets. The phylogenetics research community largely relies on simulations to assess accuracy, since the evolutionary history of a group of species is rarely known exactly (and when the true tree is known, this is usually because the dataset is easy to analyze correctly). Hence, careful use of biologically realistic simulations is necessary.

Performance analysis of phylogeny estimation methods is an area of research that PI Warnow has expertise in, and she will provide the benchmark datasets and guide additional simulations. We will use genome evolution simulation tools such as DendroPy [106], which produce gene trees that can differ from the species tree due to ILS. We will use [53] and similar tools to simulate gene trees that can differ from the species tree due to ILS and also horizontal gene transfer. We will include sequence evolution simulation tools such as Indelible [51] to evolve sequences down gene trees under Markov models of evolution, such as GTR. We will compute source trees using maximum likelihood methods (RAxML for small datasets, and FastTree [86] for larger datasets). We will compute supertrees or species trees using methods we develop as well as popular alternative methods (e.g., MP-EST for coalescent-based species tree estimation and SuperFine+MRP or SuperFine+MRL for supertree estimation). We will evaluate computed species trees based on topological distance to the true tree (known, since these are for simulated datasets), measured using the Robinson-Foulds (RF) rate [88].

Finally, we will also collaborate with biological research groups, especially the 1KP [113] and avian phylogenomics project [63], each of which is assembling large biological datasets that will need analyses of the sort we are developing. Our biologist colleagues on these projects will give us feedback about the accuracy of the trees computed using our methods, which will enable us to develop better methods and potentially better optimization criteria.

7 Broader Impacts

Human Resource Development. Two PhD students will be supported by this grant, one at Illinois and one at Berkeley. Warnow (PI, Illinois) and Rao (PI, Berkeley) have a prior research collaboration and were both PI's on the CIPRES (NSF EF-0715370) project, which contributed to the training of 16 postdoctoral fellows and 73 graduate students to work in computational and mathematical phylogenetics. Chekuri and Warnow are in the same department and will co-supervise a PhD student. The joint supervision of students planned in this project will benefit from the prior experience in multi-institutional research graduate training.

Open source software and training. The methods developed by the project will be tested extensively on benchmark datasets, and the most accurate methods will then be made available in open source software through github. We will provide free training in the use of the software at biology conferences (e.g., the annual Evolution meeting, SMBE, etc.) to biology graduate students and postdoctoral fellows.

Online educational materials. Warnow teaches an annual graduate course in the mathematical and computational foundations of phylogenetic estimation, and these course materials are available online. She is also writing a textbook in the area, which is available online through the course webpage.

Impact on both biology and theoretical computer science. Success in the aims of this project will bring new computational methods with improved accuracy to biology, and also new computational and mathematical problems to theoretical computer science and statistical inference. Improved phylogeny estimation methods may also produce more accurate phylogenies for important parts of the Tree of Life through the collaborations we plan with different biological research groups (the Thousand Plant Transcriptome Project and the Avian Phylogenomics Project in particular).

References

[1] R. Agarwala, V. Bafna, M. Farach, B. Narayanan, M. Paterson, and M. Thorup. On the approximability of numerical taxonomy: fitting distances by tree metrics. In *Proc. 7th Ann. ACM/SIAM Symp. on Discr. Algs. SODA96*, pages 365–372. SIAM Press, 1996.

[2] N. Ailon and M. Charikar. Fitting tree metrics: hierarchical clustering and phylogeny. In *Proc. FOCS 2005*, pages 73–82, 2005. doi:10.1109/SFCS.2005.36.

[3] W. A. Akanni, C. J. Creevey, M. Wilkinson, and D. Pisani. L.U.St: a tool for approximated maximum likelihood supertree reconstruction. *BMC Bioinformatics*, 15:183, 2014.

[4] D. J. Aldous, M. A. Krikun, and L. Popovic. Five Statistical Questions about the Tree of Life. *Systematic Biology*, 60(3):318–328, 2011.

[5] E.S. Allman, J.H. Degnan, and J.A. Rhodes. Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. *Journal of Mathematical Biology*, 62:833–862, 2011.

[6] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215(3):403–10, 1990.

[7] E Avni, R Cohen, and S Snir. Weighted quartets phylogenetics. *Systematic Biology*, 2014. syu087.

[8] M. S. Bansal, J. G. Burleigh, O. Eulenstein, and D. Fernández-Baca. Robinson-foulds supertrees. *Algorithms for Molecular Biology*, 5(18), 2010. doi:10.1186/1748-7188-5-18.

[9] MS Bansal, G Banay, JP Gogarten, and R Shamir. Detecting highways of horizontal gene transfer. *Journal of Computational Biology*, 18(9):1087–1114, 2011.

[10] MS Bansal, G Banay, TJ Harlow, JP Gogarten, and R Shamir. Systematic inference of highways of horizontal gene transfer in prokaryotes. *Bioinformatics*, 29(5):571–579, 2013.

[11] B. R. Baum and M. A. Ragan. The MRP method. In Olaf R. P. Bininda-Emonds, editor, *Phylogenetic Supertrees: combining information to reveal The Tree Of Life*, pages 17–34. Kluwer Academic, Dordrecht, the Netherlands, 2004.

[12] M. S. Bayzid, S. Mirarab, and T. Warnow. Inferring optimal species trees under gene duplication and loss. In *Proc. of Pacific Symposium on Biocomputing (PSB)*, volume 18, pages 250–261, 2013.

[13] M. S. Bayzid and T. Warnow. Estimating optimal species trees from incomplete gene trees under deep coalescence. *Journal of Computational Biology*, 19(6):591–605, 2012.

[14] Md. S. Bayzid, S. Mirarab, and T. Warnow. Weighted statistical binning: enabling statistically consistent genome-scale phylogenetic analyses. arXiv:1412.5454v2 [q-bio.QM], 2014.

[15] M.S. Bayzid, T. Hunt, and T. Warnow. Disk covering methods improve phylogenomic analyses. *BMC Genomics*, 15(Suppl 6):S7, 2014. A preliminary version appeared in RECOMB-Comparative Genomics.

[16] M.S. Bayzid and T. Warnow. Naive binning improves phylogenomic analyses. *Bioinformatics*, 29(18):2277–84, 2013.

[17] U. Bergthorsson, K. Adams, B. Thomason, and J. Palmer. Widespread horizontal transfer of mitochondrial genes in flowering plants. *Nature*, 424:197–201, 2003.

[18] U. Bergthorsson, A. Richardson, G. Young, L. Goertzen, and J. Palmer. Massive horizontal transfer of mitochondrial genes from diverse land plant donors to basal angiosperm Amborella. *Proc. Natl Acad. Sci., USA*, 101:17,747–17,752, 2004.

[19] O. R. P. Bininda-Emonds, J. L. Gittleman, and M. A. Steel. The (super)tree of life: procedures, problems, and prospects. *Annu. Rev. Ecol. Syst.*, 33:265–289, 2002.

[20] O.R.P. Bininda-Emonds, editor. *Phylogenetic Supertrees: Combining information to reveal the Tree of Life*. Kluwer Academic Publishers, 2004.

[21] A. Bouchard-Côté and M. I. Jordan. Evolutionary inference via the Poisson indel process. *Proceedings of the National Academy of Sciences*, 110(4):1160–1166, 2013.

[22] B. Boussau and M. Guoy. Efficient likelihood computations with non-reversible models of evolution. *Syst Biol*, 55(5):756–68, 2006.

[23] Bastien Boussau, GJ Szöllsi, and Laurent Duret. Genome-scale coestimation of species and gene trees. *Genome Research*, 23(2):323–330, December 2013.

[24] M. Brinkmeyer, T. Griebel, and S. Böcker. Polynomial supertree methods revisited. *Advances in Bioinformatics*, 2011. Article ID 524182, doi=10.1155/2011/524182.

[25] Gruia Călinescu, Chandra Chekuri, Martin Pál, and Jan Vondrák. Maximizing a monotone submodular function subject to a matroid constraint. *SIAM J. Comput.*, 40(6):1740–1766, 2011.

[26] J.J. Cannone, S. Subramanian, M.N. Schnare, J.R. Collett, L.M. D'Souza, Y. Du, B. Feng, N. Lin, L.V. Madabusi, K.M. Muller, N. Pande, Z. Shang, N. Yu, and R.R. Gutell. The Comparative RNA Web (CRW) Site: An Online Database of Comparative Sequence and Structure Information for Ribosomal, Intron and Other RNAs. *BioMed Central Bioinformatics*, 3(15), 2002. `http://www.rna.ccbb.utexas.edu`.

[27] Deeparnab Chakrabarty, Chandra Chekuri, Sanjeev Khanna, and Nitish Korula. Approximability of capacitated network design. In *Integer Programming and Combinatoral Optimization - 15th International Conference, IPCO 2011, New York, NY, USA, June 15-17, 2011. Proceedings*, pages 78–91, 2011.

[28] Chandra Chekuri and Alina Ene. Approximation algorithms for submodular multiway partition. In *Proceedings of the IEEE Symposium on Foundations of Computer Science (FOCS)*, 2011.

[29] Chandra Chekuri and Alina Ene. Submodular cost allocation problem and applications. In *International Colloquium on Automata, Languages and Programming (ICALP (1))*, pages 354–366, 2011. A longer version is available on the arXiv, abs/1105.2040.

[30] Chandra Chekuri, Alina Ene, and Ali Vakilian. Node-weighted network design in planar and minor-closed families of graphs. In *Automata, Languages, and Programming - 39th International Colloquium, ICALP 2012, Warwick, UK, July 9-13, 2012, Proceedings, Part I*, pages 206–217, 2012.

[31] Chandra Chekuri, Alina Ene, and Ali Vakilian. Prize-collecting survivable network design in node-weighted graphs. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques - 15th International Workshop, APPROX 2012, and 16th International*

Workshop, RANDOM 2012, Cambridge, MA, USA, August 15-17, 2012. Proceedings, pages 98–109, 2012.

[32] Chandra Chekuri, Sreeram Kannan, Adnan Raja, and Pramod Viswanath. Multicommodity flows and cuts in polymatroidal networks. In Shafi Goldwasser, editor, *ITCS*, pages 399–408. ACM, 2012.

[33] Chandra Chekuri, Guyslain Naves, and F. Bruce Shepherd. Maximum edge-disjoint paths in k-sums of graphs. In *Automata, Languages, and Programming - 40th International Colloquium, ICALP 2013, Riga, Latvia, July 8-12, 2013, Proceedings, Part I*, pages 328–339, 2013.

[34] Chandra Chekuri, Jan Vondrák, and Rico Zenklusen. Multi-budgeted matchings and matroid intersection via dependent rounding. In *Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2011, San Francisco, California, USA, January 23-25, 2011*, pages 1080–1097, 2011.

[35] Chandra Chekuri, Jan Vondrák, and Rico Zenklusen. Submodular function maximization via the multilinear extension and contention resolution schemes. In *Proceedings of the ACM Symposium on Theory of Computing (STOC)*, pages 783–792, 2011.

[36] J. Cracraft and M.J. Donoghue. *Assembling the Tree of Life*. Oxford University Press, 2004.

[37] A. Criscuolo, V. Berry, E. Douzery, and O. Gascuel. SDM: A fast distance-based approach for (super) tree building in phylogenomics. *Syst. Biol.*, 55:740–755, 2006.

[38] M. Csürős and M. Y. Kao. Recovering evolutionary trees through harmonic greedy triplets. *Proc. 10th Ann. ACM/SIAM Symp. Discr. Algs. SODA99*, pages 261–270, 1999.

[39] M. P. Cummings. Transmission patterns of eukaryotic transposable elements: arguments for and against horizontal transfer. *Trends Ecol. Evol.*, 9:141–145, 1994.

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

[41] James H Degnan. Anomalous unrooted gene trees. *Syst Biol*, 62(4):574–590, 2013.

[42] T Dobzhansky. Nothing in biology makes sense except in the light of evolution. *American Biology Teacher*, 35:125–129, 1973.

[43] J. J. Doyle, J. L. Doyle, and J. D. Palmer. Multiple independent losses of two genes and one intron from the legume chloroplast genome. *Syst. Bot.*, 20:272–294, 1995.

[44] J.A. Eisen and C.M. Fraser. Phylogenomics: intersection of evolution and genomics. *Science*, 300(5626):1706–1707, 2003.

[45] P.L. Erdos, M.A. Steel, L Szekely, and T Warnow. A few logs suffice to build (almost) all trees (i). *Random Structures and Algorithms*, 14:153–184, 1999.

[46] P.L. Erdos, M.A. Steel, L Szekely, and T Warnow. A few logs suffice to build (almost) all trees (ii). *Theoretical Computer Science*, 221:77–118, 1999.

[47] J. Fakcharoenphol, S. Rao, and K. Talwar. A tight bound on approximating arbitrary metrics by tree metrics. In *Proceedings 35th Annual STOC (ACM Symp. on Theory of Computing)*, pages 448–455. ACM Press, 2003.

[48] M. Fellows, M. Hallet, and U. Stege. On the multiple gene duplication problem. pages 347–356. Springer-Verlag, 1998. in *LNCS* **1533**.

[49] J. Felsenstein. Evolutionary trees from dna sequences: a maximum likelihood approach. *J Mol Evol*, 17(6):368–376, 1981.

[50] J. Felsenstein. *Inferring Phylogenies*. Sinauer Associates, Sunderland, Massachusetts, 2004.

[51] W. Fletcher and Z. Yang. INDELible: a flexible simulator of biological sequence evolution. *Mol. Biol. and Evol.*, 26(8):1879–1888, 2009.

[52] L. R. Foulds and R. L. Graham. The Steiner problem in phylogeny is NP-complete. *Advances in Applied Mathematics*, 3:43–49, 1982.

[53] N. Galtier. A model of horizontal gene transfer and the bacterial phylogeny problem. *Systematic Biology*, 56:633642, 2007.

[54] M. Ghodsi, B. Liu, and M. Pop. DNACLUST: accurate and efficient clustering of phylogenetic marker genes. *BMC Bioinformatics*, 12:271, 2011. doi:10.1186/1471-2105-12-271.

[55] JP Gogarten, WF Doolittle, and JG Lawrence. Prokaryotic evolution in light of gene transfer. *Molecular Biology and Evolution*, 19(12):2226–2238, 2002.

[56] JP Gogarten and JP Townsend. Horizontal gene transfer, genome innovation and evolution. *Nature Reviews Microbiology*, 3(9):679–687, 2005.

[57] D. Grauer and W.-H. Li. *Fundamentals of Molecular Evolution*. Sinauer Publishers, 2000.

[58] M. T. Hallett and J. Lagergren. New algorithms for the duplication-loss model. In *Proc. ACM Symp. Comput. Biol. RECOMB2000*, pages 138–146, New York, 2000. ACM Press.

[59] J. Heled and A. J. Drummond. Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution*, 27(3):570–580, 2010.

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

[61] J.P. Huelsenbeck and R. Ronquist. MrBayes: Bayesian inference of phylogeny. *Bioinformatics*, 17:754–755, 2001.

[62] D. Huson, S. Nettles, and T. Warnow. Disk-covering, a fast converging method for phylogenetic tree reconstruction. *Journal of Computational Biology*, 6(3):369–386, 1999.

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

[64] T Jiang, P Kearney, and M Li. A polynomial-time approximation scheme for inferring evolutionary trees from quartet topologies and its applications. *SIAM J. Comput.*, 30(6):1924–1961, 2001.

[65] J. F. C. Kingman. The coalescent. *Stochastic Processes and Their Applications*, 13(3):235–248, 1982.

[66] J.F.C. Kingman. On the genealogy of large populations. *Journal of Applied Probability*, pages 27–43, 1982.

[67] P Lapierre, E Lasek-Nesselquist, and JP Gogarten. The impact of HGT on phylogenomic reconstruction methods. *Briefings in Bioinformatics*, 15(1):79–90, 2014.

[68] F.-J. Lapointe and G. Cucumel. The average consensus procedure: combination of weighted trees containing identical or overlapping sets of taxa. *Systematic Biology*, 46(2):306–312, 1997.

[69] F.-J. Lapointe, M. Wilkinson, and D. Bryant. Matrix representations with parsimony or with distances: two sides of the same coin? *Systematic Biology*, 52:865–868, 2003.

[70] Bret R Larget, Satish K Kotha, Colin N Dewey, and Cécile Ané. BUCKy: gene tree/species tree reconciliation with Bayesian concordance analysis. *Bioinformatics*, 26(22):2910–2911, 2010.

[71] L. Liu. BEST: Bayesian estimation of species trees under the coalescent model. *Bioinformatics*, 24(21):2542–2543, 2008.

[72] L. Liu, L. Yu, and S. V. Edwards. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolutionary Biology*, 10(1):302, 2010.

[73] G. Lunter, A.J. Drummond, I. Miklós, and J. Hein. Statistical alignment: Recent progress, new applications, and challenges. In Rasmus Nielsen, editor, *Statistical Methods in Molecular Evolution (Statistics for Biology and Health)*, pages 375–406, Berlin, 2005. Springer.

[74] B. Ma, M. Li, and L. Zhang. On reconstructing species trees from gene trees in terms of duplications and losses. In *Proc. 2nd Ann. Int'l Conf. Comput. Mol. Biol. (RECOMB98)*, 1998.

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

[76] L.D.O. Martins, D. Mallo, and D. Posada. A Bayesian supertree model for genome-wide species tree reconstruction. *Systematic Biology*, 2014.

[77] F.A. Matsen. Phylogenetics and the human microbiome. *Systematic Biology*, 64(1):e26–e41, 2015.

[78] S. Mirarab, M. S. Bayzid, and T. Warnow. Evaluating summary methods for multi-locus species tree estimation in the presence of incomplete lineage sorting. *Systematic Biology*, 2014. doi:10.1093/sysbio/syu063.

[79] S. Mirarab, Md. S. Bayzid, B. Boussau, and T. Warnow. Statistical binning improves phylogenomic analysis. *Science*, 346(6215):1250463, 2014.

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

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

[82] D.T. Neves, T. Warnow, J. Sobral, and K. Pingali. Parallelizing SuperFine. In *27th Symposium on Applied Computing (ACM-SAC)*, 2012.

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

[84] N. Nguyen, S. Mirarab, and T. Warnow. MRL and SuperFine+MRL: new supertree methods. *Algorithms for Molecular Biology*, 7(3), 2012.

[85] Swati Patel, Rebecca T Kimball, and Edward L Braun. Error in phylogenetic estimation for bushes in the tree of life. *Journal of Phylogenetics and Evolutionary Biology*, 1(2):110, 2013.

[86] M N Price, P S Dehal, and A P Arkin. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS ONE*, 5(3):e9490. doi:10.1371/journal.pone.0009490, 2010.

[87] B. Redelings and M. Suchard. Joint Bayesian estimation of alignment and phylogeny. *Syst. Biol.*, 54(3):401–418, 2005.

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

[89] S. Roch. A short proof that phylogenetic tree reconstruction by maximum likelihood is hard. *IEEE Trans. Comput. Biol. and Bioinformatics*, 3(1):92–94, 2006.

[90] S. Roch and S. Snir. Recovering the tree-like trend of evolution despite extensive lateral genetic transfer: a probabilistic analysis. *Journal of Computational Biology*, 20(2):93–112, 2012.

[91] S. Roch and M. Steel. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. *Theoretical Population Biology*, 100:56–62, December 2014.

[92] S. Roch and M.A. Steel. Likelihood-based tree reconstruction on a concatenation of alignments can be positively misleading, 2014. Arxiv publication arXiv:1409.2051.

[93] U. Roshan, B.M.E. Moret, T.L. Williams, and T. Warnow. Performance of supertree methods on various dataset decompositions. In O.R.P. Bininda-Emonds, editor, *Phylogenetic Supertrees: combining information to reveal The Tree of Life*, pages 301–328, 2004. Volume 3 of Computational Biology, Kluwer Academics, (Andreas Dress, series editor).

[94] U. Roshan, B.M.E. Moret, T.L. Williams, and T. Warnow. Rec-I-DCM3: A fast algorithmic technique for reconstructing large phylogenetic trees. In *Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference*, pages 98–109, 2004.

[95] H Shim and B Larget. BayesCAT: Bayesian Co-estimation of Alignment and Tree. arXiv preprint arXiv:1411.6150.

[96] K Sjolander. Phylogenomic inference of protein molecular function: advances and challenges. *Bioinformatics*, 20(2):170–179, 2004.

[97] Kimmen Sjolander. Getting started in structural phylogenomics. *PLoS Comput Biol*, 6(1):e1000621, 01 2010.

[98] S. Snir and S. Rao. Quartets MaxCut: A divide and conquer quartets algorithm. *IEEE/ACM Transaction of Computational Biology and Bioinformatics*, 7(4):704–718, 2010.

[99] S. Snir and S. Rao. Quartet maxcut: A fast algorithm for amalgamating quartet trees. *Journal of Molecular Phylogenetics and Evolution*, 62:1–8, 2012.

[100] S. Snir and R. Yuster. Reconstructing approximate phylogenetic trees from quartet samples. *SIAM J. Computing*, 41:1466–1480, 2012.

[101] Alexandros Stamatakis. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690, 2006.

[102] M. Steel, S. Linz, D.H. Huson, and M.J. Sanderson. Identifying a species tree subject to random lateral gene transfer. *Journal of Theoretical Biology*, 322:81–93, 2013.

[103] M.A. Steel. The complexity of reconstructing trees from qualitative characters and subtrees. *Journal of Classification*, 9:91–116, 1992.

[104] M.A. Steel and A. Rodrigo. Maximum likelihood supertrees. *Syst. Biol.*, 57(2):243–250, 2008.

[105] Michael A. Steel. Consistency of Bayesian inference of resolved phylogenetic trees. *ArXiv:1001.2684 [q-bio.PE]*, 2010.

[106] Jeet Sukumaran and Mark T. Holder. Dendropy: a Python library for phylogenetic computing. *Bioinformatics*, 26(12):1569–1571, 2010.

[107] M. S. Swenson, R. Suri, C. R. Linder, and T. Warnow. SuperFine: fast and accurate supertree estimation. *Systematic Biology*, 61(2):214–227, 2012.

[108] M Shel Swenson, Rahul Suri, C Randal Linder, and Tandy Warnow. An experimental study of quartets maxcut and other supertree methods. *Algorithms for Molecular Biology*, 6(1):7, 2011.

[109] M.S. Swenson, F. Barbançon, T. Warnow, and C.R. Linder. A simulation study comparing supertree and combined analysis methods using SMIDGen. *Algorithms for Molecular Biology*, 5:8, 2010.

[110] M.S. Swenson, R. Suri, C.R. Linder, and T. Warnow. An experimental study of Quartets MaxCut and other supertree methods. In *Proceedings of the 2010 Workshop on Algorithms in Bioinformatics (WABI)*, 2010.

[111] J.L. Thorne, H. Kishino, and J. Felsenstein. An evolutionary model for maximum likelihood alignment of DNA sequences. *Journal of Molecular Evolution*, 33:114–124, 1991.

[112] J.L. Thorne, H. Kishino, and J. Felsenstein. Inching towards reality: An improved likelihood model of sequence evolution. *Journal of Molecular Evolution*, 34:3–16, 1992.

[113] G K.-S. Wang. The 1KP Project Website. http://onekp.com/project.html.

[114] T. Warnow. Large-scale phylogenetic reconstruction. In S. Aluru, editor, *Handbook of Computational Biology*. Chapman & Hall, 2005. CRC Computer and Information Science Series, 2005.

[115] T. Warnow, B. M. E. Moret, and K. St. John. Absolute phylogeny: true trees from short sequences. In *Proc. 12th Ann. ACM/SIAM Symp. on Discr. Algs. SODA01*, pages 186–195. SIAM Press, 2001.

[116] N. J. Wickett, S. Mirarab, N. Nguyen, T. Warnow, E. Carpenter, N. Matasci, S. Ayyampalayam, M. S. Barker, J. G. Burleigh, M. A. Gitzendanner, et al. A phylotranscriptomics analysis of the origin and early diversification of land plants. *Proc. National Academy of Sciences*, 111(45):E4859–E4868, 2014.

[117] S. J. Willson. Constructing rooted supertrees using distances. *Bulletin of Mathematical Biology*, 66(6):1755–1783, 2004.

[118] Yufeng Wu. Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. *Evolution*, 66:763–775, 2012.

[119] Y. Yu, C. Than, J.H. Degnan, and L. Nakhleh. Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. *Systematic Biology*, 60(2):138–149, 2011.

[120] Y Yu, T Warnow, and L Nakhleh. Algorithms for MDC-based multi-locus phylogeny inference. In *Proc RECOMB 2011*, 2011.

[121] Y Yu, T Warnow, and L Nakhleh. Algorithms for MDC-based multi-locus phylogeny inference: Beyond rooted binary gene trees on single alleles. *J Computational Biology*, 18(11):1543–1559, 2011.

[122] J. Zhang. Evolution by gene duplication: an update. *Trends in Ecology & Evolution*, 18(6):292–8, 2003.

[123] T. Zimmerman, S. Mirarab, and T. Warnow. BBCA: improving the scalability of *BEAST using random binning. *BMC Genomics*, 15:S7, 2014. A preliminary version appeared in the Proc. RECOMB Comparative Genomics.