FastNet: Fast and Accurate Statistical Inference of Phylogenetic Networks Using Large-Scale Genomic Sequence Data

Hussein A. Hejase[1], Natalie VandePol[2], Gregory M. Bonito[2], and Kevin J. Liu[3(✉)]

[1] Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA
[2] Department of Plant, Soil and Microbial Sciences, Michigan State University, East Lansing, MI 48824, USA
[3] Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824, USA
kjl@msu.edu

Abstract. An emerging discovery in phylogenomics is that interspecific gene flow has played a major role in the evolution of many different organisms. To what extent is the Tree of Life not truly a tree reflecting strict "vertical" divergence, but rather a more general graph structure known as a phylogenetic network which also captures "horizontal" gene flow? The answer to this fundamental question not only depends upon densely sampled and divergent genomic sequence data, but also computational methods which are capable of accurately and efficiently inferring phylogenetic networks from large-scale genomic sequence datasets. Recent methodological advances have attempted to address this gap. However, in the 2016 performance study of Hejase and Liu, state-of-the-art methods fell well short of the scalability requirements of existing phylogenomic studies.

The methodological gap remains: how can phylogenetic networks be accurately and efficiently inferred using genomic sequence data involving many dozens or hundreds of taxa? In this study, we address this gap by proposing a new phylogenetic divide-and-conquer method which we call FastNet. We conduct a performance study involving a range of evolutionary scenarios, and we demonstrate that FastNet outperforms state-of-the-art methods in terms of computational efficiency and topological accuracy.

1 Introduction

Recent advances in biomolecular sequencing [30] and evolutionary modeling and inference [10,34] set the stage for a new era of phylogenomics. One major outcome is the discovery that interspecific gene flow has played a major role in the

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-00834-5_14) contains supplementary material, which is available to authorized users.

M. Blanchette and A. Ouangraoua (Eds.): RECOMB-CG 2018, LNBI 11183, pp. 242–259, 2018.
https://doi.org/10.1007/978-3-030-00834-5_14

evolution of many different organisms across the Tree of Life [1,23,29], including humans and ancient hominins [15,39], butterflies [44], mice [28], and fungi [14]. These findings point to new directions for phylogenetics and phylogenomics: to what extent is the Tree of Life not truly a tree reflecting strict vertical divergence, but rather a more general graph structure known as a phylogenetic network where reticulation edges and nodes capture gene flow? And what is the evolutionary role of gene flow? In addition to densely sampled and divergent genomic sequence data, one additional ingredient is needed to make progress on these questions: computational methods which are capable of accurately and efficiently inferring phylogenetic networks on large-scale genomic sequence datasets.

Recent methodological advances have attempted to address this gap. Solís-Lemus and Ané proposed SNaQ [42], a new statistical method which seeks to address the computational efficiency of species network inference using a pseudo-likelihood approximation. The method of Yu and Nakhleh [45] (referred to here as MPL, which stands for maximum pseudo-likelihood) substitutes pseudo-likelihoods in place of the full model likelihoods used by the methods of Yu et al. [48] (referred to here as MLE, which stands for maximum likelihood estimation, and MLE-length, which differ based upon whether or not gene tree branch lengths contribute to model likelihood). Two of us recently conducted a performance study which demonstrated the scalability limits of SNaQ, MPL, MLE, MLE-length, and other state-of-the-art phylogenetic methods in the context of phylogenetic network inference [17]. The scalability of the state of the art falls well short of that required by current phylogenetic studies, where many dozens or hundreds of divergent genomic sequences are common [34]. The most accurate phylogenetic network inference methods performed statistical inference under phylogenomic models [42,47,48] that extended the multi-species coalescent model [16,24]. MPL and SNaQ were among the fastest of these methods while MLE and MLE-length were the most accurate. None of the statistical phylogenomic inference methods completed analyses of datasets with 30 taxa or more after many weeks of CPU runtime – not even the pseudo-likelihood-based methods which were devised to address the scalability limitations of other statistical approaches. The remaining methods fell into two categories: split-based methods [4,7] and the parsimony-based inference method of Yu et al. [46] (which we refer to as MP in this study). Both categories of methods were faster than the statistical phylogenomic inference methods but less accurate.

The methodological gap remains: how can species networks be accurately and efficiently inferred using large-scale genomic sequence datasets? In this study, we address this question and propose a new method for this problem. We investigate this question in the context of two constraints. We focus on dataset size in terms of the number of taxa and the number of reticulations in the species phylogeny. We note that scalability issues arise due to other dataset features as well, including population-scale allele sampling for each taxon in a study.

2 Methods

One path forward is through the use of divide-and-conquer. The general idea behind divide-and-conquer is to split the full problem into smaller and more closely related subproblems, analyze the subproblems using state-of-the-art phylogenetic network inference methods, and then merge solutions on the subproblems into a solution on the full problem. Viewed this way, divide-and-conquer can be seen as a computational framework that "boosts" the scalability of existing methods (and which is distinct from boosting in the context of machine learning). The advantages of analyzing smaller and more closely related subproblems are two-fold. First, smaller subproblems present more reasonable computational requirements compared to the full problem. Second, the evolutionary divergence of taxa in a subproblem is reduced compared to the full set of taxa, which has been shown to improve accuracy for phylogenetic tree inference [11,19,26]. We and others have successfully applied divide-and-conquer approaches to enable scalable inference in the context of species tree estimation [26,27,33].

Here, we consider the more general problem of inferring species phylogenies that are directed phylogenetic networks. A directed phylogenetic network $N = (V, E)$ consists of a set of nodes V and a set of directed edges E. The set of nodes V consists of a root node $r(N)$ with in-degree 0 and out-degree 2, leaves $\mathcal{L}(N)$ with in-degree 1 and out-degree 0, tree nodes with in-degree 1 and out-degree 2, and reticulation nodes with in-degree 2 and out-degree 1. A directed edge $(u, v) \in E$ is a tree edge if and only if v is a tree node, and is otherwise a reticulation edge. Following the instantaneous admixture model used by Durand et al. [9], each reticulation node contributes a parameter γ, where one incoming edge has admixture frequency γ and the other has admixture frequency $1-\gamma$. The edges in a network N can be labeled by a set of branch lengths ℓ. A directed phylogenetic tree is a special case of a directed phylogenetic network which contains no reticulation nodes (and edges). An unrooted tree can be obtained from a directed tree by ignoring edge directionality.

The phylogenetic network inference problem consists of the following. One input is a partitioned multiple sequence alignment $\boldsymbol{A}$ containing data partitions a_i for $1 \leq i \leq k$, where each partition corresponds to the sequence data for one of k genomic loci. Each of the n rows in the alignment $\boldsymbol{A}$ is a sample representing taxon $x \in X$, and each taxon is represented by one or more samples. Similar to other approaches [42,48], we also require an input parameter C_r which specifies a hypothesized number of reticulations. We note that increasing C_r for a given input alignment $\boldsymbol{A}$ results in a solution with either better or equal likelihood under the evolutionary models used in our study and others [42,48]. As is common practice for this and many other statistical inference/learning problems, inference can be coupled with standard model selection techniques (e.g., information criteria [2,3,20,41], cross-validation, etc.) to balance model fit to the observed data against model complexity, thereby determining a suitable choice for parameter C_r in an automated manner. The output consists of a directed phylogenetic network N where each leaf in $\mathcal{L}(N)$ corresponds to a taxon $x \in X$.

2.1 The FastNet Algorithm

We now describe our new divide-and-conquer algorithm, which we refer to as FastNet. A flowchart of the algorithm is shown in Fig. 1. (Detailed pseudocode can be found in the Appendix's Supplementary Methods section.)

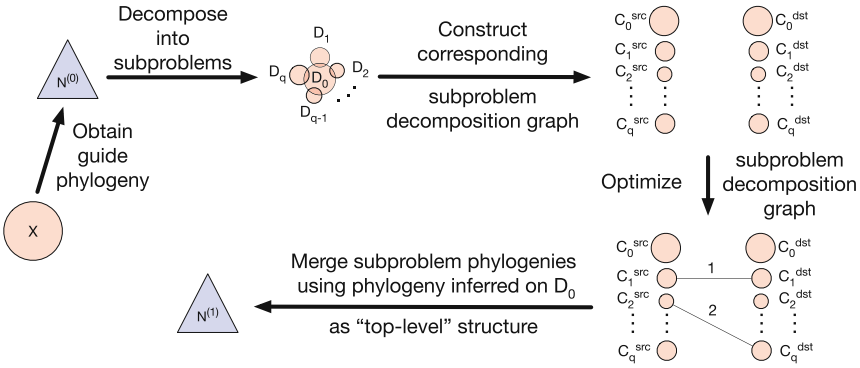

Fig. 1. A high-level illustration of the FastNet algorithm. First, a guide phylogeny $N^{(0)}$ is inferred on the full set of taxa X. Next, the guide phylogeny $N^{(0)}$ is used to decompose X into subproblems $\{D_0, D_1, D_2, \ldots, D_{q-1}, D_q\} = D$. Then, the subproblem decomposition D is used to construct a bipartite graph $G_D = (V_D, E_D)$, which is referred to as the subproblem decomposition graph. The set of vertices V_D consist of two partitions: source vertices $V_D^{\text{src}} = \{C_0^{\text{src}}, C_1^{\text{src}}, \ldots, C_q^{\text{src}}\}$ where each subproblem D_i has a corresponding source vertex C_i^{src}, and destination vertices $V_D^{\text{dst}} = \{C_0^{\text{dst}}, C_1^{\text{dst}}, \ldots, C_q^{\text{dst}}\}$ similarly. The subproblem decomposition graph G_D is optimized to infer subproblem phylogenies and reticulations, where the latter are inferred based on the placement of weighted edges $e \in E_D$. Finally, the subproblem phylogenies are merged using the phylogeny inferred on D_0 as the "top-level" structure.

Step Zero: Obtaining Local Gene Trees. FastNet is a summary-based method for inferring phylogenetic networks. Subsequent steps of the FastNet algorithm (i.e., steps one and three) therefore utilize a list of gene trees G as input, where the ith gene tree g_i in list G represents the evolutionary history of data partition a_i. The experiments in our study utilized either true or inferred gene trees as input to summary-based inference methods, including FastNet (see below for details). We used FastTree [37] to perform maximum likelihood estimation of local gene trees. Our study made use of an outgroup, and the unrooted gene trees inferred by FastTree were rooted on the leaf edge corresponding to the outgroup.

Step One: Obtaining a Guide Phylogeny. The subsequent subproblem decomposition step requires a rooted guide phylogeny $N^{(0)}$. The phylogenetic relationships need not be completely accurate. Rather, the guide phylogeny needs to be sufficiently accurate to inform subsequent divide-and-conquer steps.

Another requirement is that the method used for inferring the guide phylogeny must have reasonable computational requirements.

A range of different methods for obtaining guide phylogenies can satisfy these criteria. One option is the parsimony-based algorithm proposed by Yu et al. [46] to infer a rooted species network. The algorithm is implemented in the PhyloNet software package [43]. We refer to this method as MP. In a previous simulation study [17], we found that MP offers a significant runtime advantage relative to other state-of-the-art species network inference methods, but had relatively lower topological accuracy. Another option is using ASTRAL [31,32], a state-of-the-art phylogenomic inference method that infers species trees, to infer a guide phylogeny that is a tree rather than a network. A primary reason for the use of species tree inference methods is their computational efficiency relative to state-of-the-art phylogenetic network inference methods. ASTRAL effectively infers an unrooted and undirected species tree. We rooted the species tree using outgroup rooting. Another consideration is that, while ASTRAL accurately infers species trees for evolutionary scenarios lacking gene flow, the assumption of tree-like evolution is generally invalid for the computational problem that we consider. As we show in our performance study, our divide-and-conquer approach can still be applied despite this limitation, suggesting that FastNet is robust to guide phylogeny error. For this reason, the FastNet experiments in our study exclusively use ASTRAL to infer guide phylogenies.

Step Two: Subproblem Decomposition. The rooted and directed guide phylogeny $N^{(0)}$ is then used to produce a subproblem decomposition D. The decomposition D consists of a "bottom-level" component and a "top-level" component, which refers to the subproblem decomposition technique. The bottom-level component is comprised of disjoint subsets D_i for $1 \leq i \leq q$ which partition the set of taxa X such that $\bigcup_{1 \leq i \leq q} D_i = X$. We refer to each subset D_i as a bottom-level subproblem. The top-level component consists of a top-level subproblem D_0 which overlaps each bottom-level subproblem D_i where $1 \leq i \leq q$.

The bottom-level component of the subproblem decomposition is obtained using the following steps. First, for each reticulation node in $N^{(0)}$, we delete the incoming edge with lower admixture frequency. Let $T^{(0)}$ be the resulting phylogeny, which contains no reticulation edges and is therefore a tree. Removal of any single edge in $T^{(0)}$ disconnects the tree into two subtrees; the leaves of the two subtrees will form two subproblems. We extend this observation to obtain decompositions with two or more subproblems. The decomposition is defined by S, a set of nodes in $T^{(0)}$. Each node $s \in S$ induces a corresponding subproblem D_i for $1 \leq i \leq q$ which consists of the taxa corresponding to the leaves that are reachable from s in $T^{(0)}$. Of course, not all decompositions are created equal. In this study, decompositions are constrained by the maximum subproblem size c_m; we also required a minimum of two subproblems in a decomposition. We obtained a decomposition using a greedy algorithm which is similar to the Center-Tree-i decomposition used by Liu et al. [26] in the context of species tree inference. The two methods differ primarily due to their decomposition criteria. Initially

the set S consists of the root node $r(T^{(0)})$. The set S is iteratively updated as follows: each iteration greedily selects a node $s \in S$ with maximal corresponding subproblem size, the node s is removed from the set S and replaced by its children. Iteration terminates when both decomposition criteria (the maximum subproblem size criterion and the minimum number of subproblems) are satisfied. If no decomposition satisfies the criteria, then the search is restarted using a maximum subproblem size of $c_m - 1$. In practice, the parameter c_m is set to an empirically determined value which is based upon the largest datasets that state-of-the-art methods can analyze accurately within a reasonable timeframe [17]. The output of the search algorithm is effectively a search tree $T^{(0)}_{\text{top}}$ with a root corresponding to $r(T^{(0)})$, leaves corresponding to $s \in S$, and the subset of edges in $T^{(0)}$ which connect the root $r(T^{(0)})$ to the nodes $s \in S$ in $T^{(0)}$. The decomposition is obtained by deleting $T^{(0)}_{\text{top}}$'s corresponding structure in $T^{(0)}$, resulting in q sub-trees which induce bottom-level subproblems as before.

The top-level component augments the subproblem decomposition with a single top-level subproblem D_0 which overlaps each bottom-level subproblem. Phylogenetic structure inferred on D_0 represents ancestral evolutionary relationships among bottom-level subproblems. Furthermore, overlap between the top-level subproblem D_0 and bottom-level subproblems is necessary for the subsequent merge procedure (see "Step four" below). The top-level subproblem D_0 contains representative taxa taken from each bottom-level subproblem D_i for $1 \leq i \leq q$: for each bottom-level subproblem D_i, we choose the leaf in $T^{(0)}$ that is closest to the corresponding node $s \in S$ to represent D_i, and the corresponding taxon is included in the top-level subproblem D_0.

Step Three: Subproblem Decomposition Graph Optimization. Tree-based divide-and-conquer approaches reduce evolutionary divergence within subproblems by effectively partitioning the inference problem based on phylogenetic relationships. Within each part of the true phylogeny corresponding to a subproblem, the space of possible unrooted sub-tree topologies contributes a smaller set of distinct bipartitions (each corresponding to a possible tree edge) that need to be evaluated during search as compared to the full inference problem. The same insight can be applied to reticulation edges as well, except that a given reticulation is not necessarily restricted to a single subproblem.

We address the issue of "inter-subproblem" reticulations through the use of an abstraction which we refer to as a subproblem decomposition graph. A subproblem decomposition graph $G_D = (V_D, E_D)$ is a bipartite graph where the vertices V_D can be partitioned into two sets: a set of source vertices V_D^{src} and a set of destination vertices V_D^{dst}. There is a source vertex $C_i^{\text{src}} \in V_D^{\text{src}}$ for each distinct subproblem $D_i \in D$ where $0 \leq i \leq q$, and similarly for destination vertices $C_i^{\text{dst}} \in V_D^{\text{dst}}$. An undirected edge $e_{ij} \in E_D$ connects a source vertex C_i^{src} to a destination vertex C_j^{dst} where $i \leq j$ and has a weight $w(e_{ij}) \in \mathcal{N}^+$. If an edge e_{ii} connects nodes C_i^{src} and C_i^{dst} that correspond to the same subproblem $D_i \in D$, then the edge weight $w(e_{ii}) > 0$ specifies the number of reticulations in the phylogenetic network to be inferred on subproblem D_i; otherwise, a phylogenetic tree is to be inferred on subproblem D_i. If an edge e_{ij} connects nodes C_i^{src} and

C_j^{dst} where $i < j$, then the edge weight $w(e_{ij}) > 0$ specifies the number of "inter-subproblem" reticulations between the subproblems D_i and D_j (where an inter-subproblem reticulation is a reticulation with one incoming edge which is incident from the phylogeny to be inferred on subproblem D_i and the other incoming edge which is incident from the phylogeny to be inferred on D_j); otherwise, no reticulations are to be inferred between the two subproblems. A subproblem decomposition graph is constrained to have a total number of reticulations such that $\sum_{e \in E_D} w(e) = C_r$.

Given a subproblem decomposition D, FastNet's search routines make use of the correspondence between a subproblem decomposition graph G_D and a multiset with cardinality C_r that is chosen from $\binom{q+1}{2} + (q+1)$ elements, where q is the number of bottom-level subproblems and there are $(q+1)$ subproblems in D. Enumeration over corresponding multisets is feasible when the number of subproblems and C_r are sufficiently small; otherwise, perturbations of a corresponding multiset can be used as part of a local search heuristic. See Algorithm 1 in the Appendix's Supplementary Methods section for detailed pseudocode.

A subproblem decomposition graph G_D facilitates phylogenetic inference given a subproblem decomposition D. The resulting inference is evaluated with respect to a pseudo-likelihood-based criterion. Pseudocode for the pseudo-likelihood calculation is shown in Algorithm 2 in the Appendix's Supplementary Methods section.

The first step is to analyze each individual subproblem $D_i \in D$ where $0 \leq i \leq q$. If an edge e_{ii} exists, then a phylogenetic network with $w(e_{ii})$ reticulations is inferred on the corresponding subproblem D_i; otherwise, a phylogenetic tree is inferred. We used one of three different summary-based methods to perform phylogenetic inference on subproblems, which we refer to as a base method: two likelihood-based methods – MLE and MLE-length – as well as MPL, a pseudo-likelihood-based method. Due to the modular design of FastNet's divide-and-conquer algorithm, topological constraints on a base method's inference will also apply to FastNet. To simplify discussion, the remainder of the algorithm description will assume the use of MLE as a base method.

Next, reticulations are inferred "between" pairs of subproblems as follows. Let N_i and N_j where $i \neq j$ be the networks inferred on subproblems D_i and D_j, respectively, using the above procedure. Construct the cherry given by the Newick-formatted [12] string "$(N_i : b_i, N_j : b_j)$ANC;", which consists of a new root node ANC with children $r(N_i)$ and $r(N_j)$ where N_i and N_j are respectively retained as sub-phylogenies. Then, infer branch lengths b_i and b_j and add $w(e_{ij})$ reticulations under the maximum likelihood criterion used by the base method. For pairs of subproblems not involving the top-level subproblem D_0, we used the base method to perform constrained optimization. For pairs of subproblems involving the top-level subproblem D_0, we used a greedy heuristic: initial placements were chosen arbitrarily for each reticulation, the source node for each reticulation edge was exhaustively optimized, and then the destination node for each reticulation edge was exhaustively optimized.

Inferred phylogenies and likelihoods were cached to ensure consistency among individual and pairwise subproblem analyses, which is necessary for the subsequent merge procedure. Caching also aids computational efficiency.

Finally, the subproblem decomposition graph and associated phylogenetic inferences are evaluated using a pseudo-likelihood criterion:

$$\prod_{0 \leq i \leq q} \delta[i, w(G_D, i, i)] \prod_{\substack{0 \leq i \leq q \\ i < j \leq q}} \psi[i, j, w(G_D, i, j), w(G_D, i, i), w(G_D, j, j)] \quad (1)$$

where $w(G_D, i, j)$ is the weight of edge e_{ij} if it exists in $E(G_D)$ or 0 otherwise, $\delta[i, w(G_D, i, i)]$ is the cached likelihood for an individual subproblem D_i, and $\psi[i, j, w(G_D, i, j), w(G_D, i, i), w(G_D, j, j)]$ is the cached likelihood for a pair of subproblems D_i and D_j where $i < j$. The pseudo-likelihood calculation effectively assumes that subproblems are independent, although they are correlated through connecting edges in the model phylogeny. The choice of optimization criterion in this context represents a tradeoff between efficiency and accuracy, and several other state-of-the-art phylogenetic inference methods also use pseudo-likelihoods to analyze subsets of taxa (e.g., MPL and SNaQ). Other choices are possible. For example, an alternative would be to merge subproblem inferences into a single network hypothesis and calculate its likelihood under the multi-species network coalescent (MSNC) model.

We optimize subproblem decomposition graphs under the pseudo-likelihood criterion. Exhaustive enumeration of subproblem decomposition graphs is possible for the datasets in our study. Pseudocode to obtain a global optimum is shown in Algorithm 3 in the Appendix's Supplementary Methods section. For larger datasets with more reticulations, heuristic search techniques can be used to obtain local optima as a more efficient alternative.

Step Four: Merge Subproblem Phylogenies into a Phylogeny on the Full Set of Taxa. Given an optimal subproblem decomposition graph G'_D returned by the previous step, the final step of the FastNet algorithm merges the "top-level" phylogenetic structure inferred on D_0 and "bottom-level" subproblem phylogenies D_i for $1 \leq i \leq q$ (Algorithm 4 in the Appendix's Supplementary Methods section). First, the phylogeny inferred on the top-level subproblem D_0 serves as the top-level of the output phylogeny N'. Next, the ith taxon in N' is replaced with the phylogeny inferred on bottom-level subproblem D_i, which was cached during the evaluation of G'_D. Finally, each "inter-subproblem" reticulation that was inferred for a pair of subproblems D_i and D_j where $i < j$ is added to the output phylogeny N', which is compatible by construction of the decomposition D and the optimal subproblem decomposition graph G'_D. The result of the merge procedure is an output phylogeny N' on the full set of taxa X.

2.2 Performance Study

We conducted a simulation study to evaluate the performance of FastNet and existing state-of-the-art methods for phylogenetic network inference. The perfor-

mance study utilized the following procedures. Detailed commands and software options are given in the Supplementary Material.

We also conducted an empirical study to evaluate FastNet's performance. Details about the empirical study are provided in the Appendix.

Simulation of Model Networks. For each model condition, random model networks were generated using the following procedure. First, r8s version 1.7 [40] was used to simulate random birth-death trees with n taxa where $n \in \{15, 20, 25, 30\}$, which served as in-group taxa during subsequent analysis. The height of each tree was scaled to 5.0 coalescent units. Next, a time-consistent level-r rooted tree-based network [13,21,49] was obtained by adding r reticulations to each tree, where $r \in [1, 4]$. The procedure for adding a reticulation consists of the following steps: based on a consistent timing of events in the tree, (1) choose a time t_M uniformly at random between 0 and the tree height, (2) randomly select two tree edges for which corresponding ancestral populations existed during time interval $[t_A, t_B]$ such that $t_M \in [t_A, t_B]$, and (3) add a reticulation to connect the pair of tree edges. Finally, an outgroup was added to the resulting network at time 15.0.

Reticulations in our study have the same interpretation as in the study of Leaché et al. [25]. Gene flow is modeled using an isolation-with-migration model, where each reticulation is modeled as a unidirectional migration event with rate 5.0 during the time interval $[t_A, t_B]$. We focus on paraphyletic gene flow as described by Leaché et al.; their study also investigated two other classes of gene flow – both of which involve gene flow between two sister species after divergence. Our simulation study omits these two classes since several existing methods (i.e., MLE and MPL) have issues with identifiability in this context; thus, the model networks in our study are a proper subset of the class of level-r rooted tree-based networks. We note that FastNet makes no assumptions about the type of gene flow to be inferred, and identifiability depends on the model used for inference by FastNet's base method.

As in the study of Leaché et al., we further classify simulation conditions based on whether gene flow is "non-deep" or "deep" based on topological constraints. Non-deep reticulations involve leaf edges only, and all other reticulations are considered to be deep. Similarly, model conditions with non-deep gene flow have model networks with non-deep reticulations only; all other model conditions include deep reticulations and are referred to as deep.

Simulation of Local Genealogies and DNA Sequences. We used ms [18] to simulate local gene trees for independent and identically distributed (i.i.d.) loci under an extended multi-species coalescent model, where reticulations correspond to migration events as described above. Each coalescent simulation sampled one allele per taxon. The primary experiments in our study simulated 1000 gene trees for each random model network. Our study also investigated data requirements of different methods by including additional datasets where either 200 or 100 gene trees were simulated for each random model network.

Sequence evolution was simulated using seq-gen [38], which takes the local genealogies generated by ms as input and simulates sequence evolution along

each genealogy under a finite-sites substitution model. Our simulations utilized the Jukes-Cantor substitution model [22]. We simulated 1000 bp per locus, and the resulting multi-locus sequence alignment had a total length of 1000 kb.

Replicate Datasets. A model condition in our study consisted of fixed values for each of the above model parameters. For each model condition, the simulation procedure was repeated twenty times to generate twenty replicate datasets.

Species Network Inference Methods. Our simulation study compared the performance of FastNet against existing methods which were among the fastest and most accurate in our previous performance study of state-of-the-art species network inference methods [17]. Like FastNet, these methods perform summary-based inference – i.e., the input consists of gene trees inferred from sequence alignments for multiple loci, rather than the sequence alignments themselves. The methods are broadly characterized by their statistical optimization criteria: either maximum likelihood or maximum pseudo-likelihood under the multi-species network coalescent (MSNC) model [47]. The maximum likelihood estimation methods consisted of two methods proposed by Yu et al. [48] which are implemented in PhyloNet [43]. One method utilizes gene trees with branch lengths as input observations, whereas the other method considers gene tree topologies only; we refer to the methods as MLE-length and MLE, respectively. Our study also included the pseudo-likelihood-based method of [46], which we refer to as MPL. For each analysis in our study, all species network inference methods – MLE, MLE-length, MPL, and FastNet – were provided with identical inputs.

Our study included two categories of experiments. The "boosting" experiments in our simulation study compared the performance of FastNet against its base method; we refer to all other experiments in our study as "non-boosting". To make boosting comparisons explicit, each boosting experiment will refer to "FastNet(BaseMethod)" which is FastNet run with a specific base method "BaseMethod" – either MLE-length, MLE, or MPL. The input for each boosting experiment consisted of either true or inferred gene trees for all loci. The inferred gene trees were obtained using FastTree [37] with default settings to perform maximum likelihood estimation under the Jukes-Cantor substitution model [22]. The inferred gene trees were rooted using the outgroup. The non-boosting experiments focused on the performance of FastNet using MLE as a base method and inferred gene trees as input, where gene trees were inferred using the same procedure as in the boosting experiments.

Performance Measures. The species network inference methods in our study were evaluated using two different criteria.

The first criterion was topological accuracy. For each method, we compared the inferred species phylogeny to the model phylogeny using the tripartition fraction [35], which counts the proportion of tripartitions that are not shared between the inferred and model network. It has been shown that the tripartition fraction is not a metric on rooted phylogenetic networks in general [8]. However, the model networks in our study satisfy the tree-child condition (i.e., every

internal node has at least one child that is a tree node) since the simulation procedure stipulates that reticulation placements can only connect tree edges; the reticulation placement procedure also naturally gives a temporal representation [5] and ensures that the parents of a reticulation node cannot be connected by a path. Cardona et al. [8] showed that the tripartition fraction is a metric for the subset of rooted phylogenetic networks that satisfy these constraints.

The second criterion was computational runtime. All computational analyses were run on computing facilities in Michigan State University's High Performance Computing Center. We used compute nodes in the intel16 cluster, each of which had a 2.5 GHz Intel Xeon E5-2670v2 processor with 64 GiB of main memory. All replicates completed with memory usage less than 32 GiB.

3 Results

FastNet's use of phylogenetic divide-and-conquer is compatible with a range of different methods for inferring rooted species networks on subproblems, which we refer to as "base" methods. From a computational perspective, FastNet can be seen as a general-purpose framework for boosting the performance of base methods. We began by assessing the relative performance boost provided by FastNet when used with two different state-of-the-art network inference methods. We evaluated two different aspects of performance: topological error as measured by the tripartition fraction [35] between an inferred species network and the model network, and computational runtime. The initial set of boosting experiments focused on species network inference in isolation of upstream inference accuracy by providing true gene trees as input to all of the summary-based inference methods.

In the performance study of Hejase and Liu [17], the probabilistic network inference methods were found to be the most accurate among state-of-the-art methods, and MPL was among the fastest methods in this class. MPL utilized a pseudo-likelihood-based approximation for increased computational efficiency compared with full likelihood methods [45]. However, the tradeoff netted efficiency that was well short of current phylogenomic dataset sizes [17].

Table 1 shows the performance of FastNet(MPL) relative to MPL on model conditions with increasing numbers of taxa and non-deep reticulations. On model conditions with dataset sizes ranging from 15 to 30 taxa and from 1 to 4 reticulations, FastNet(MPL)'s improvement in topological error relative to its base method was statistically significant (one-sided pairwise t-test with Benjamini-Hochberg correction for multiple tests [6]; $\alpha = 0.05$ and $n = 20$) and substantial in magnitude – an absolute improvement that amounted to as much as 41%. Furthermore, the improvement in topological error grew as datasets became larger and involved more reticulations: the largest improvements were seen on the 30-taxon 4-reticulation model condition. Runtime improvements were also statistically significant and represented speedups which amounted to as much as a day and a half of runtime.

Next, we evaluated FastNet's performance when boosting MLE-length, the most accurate state-of-the-art method from the performance study of

Table 1. FastNet(MPL) "boosts" MPL's runtime and topological accuracy, where a greater performance boost occurs as dataset sizes increase. The relative performance of FastNet(MPL) and MPL is compared on model conditions with 15–30 taxa and 1–4 non-deep reticulations. The performance measures consisted of topological error as measured by the tripartition fraction between an inferred species network and the model network and computational runtime in hours. Average ("Avg") and standard error ("SE") of FastNet(MPL)'s performance improvement over MPL is reported ($n = 20$). All methods were provided with true gene trees as input. The statistical significance of FastNet(MPL)'s performance improvement over MPL was assessed using a one-sided t-test. Corrected q-values are reported where multiple test correction was performed using the Benjamini-Hochberg method [6].

Number of taxa	Number of reticulations	Improvement in topological error			Improvement in runtime (h)		
		Avg	SE	Corrected q-value	Avg	SE	Corrected q-value
15	1	0.087	0.036	3.3×10^{-2}	2.8	0.3	7.2×10^{-5}
20	2	0.346	0.036	1.1×10^{-5}	9.6	0.1	1.1×10^{-2}
25	3	0.281	0.024	7.9×10^{-5}	35.6	5.6	8.5×10^{-4}
30	4	0.413	0.001	8.8×10^{-12}	30.3	6.5	2.8×10^{-2}

Hejase and Liu [17]. On model conditions with non-deep reticulations, FastNet (MLE-length) had a similar boosting effect as compared to FastNet (MPL) (Table 2). On the 15-taxon single-reticulation model condition, FastNet's average improvement in topological error was greater when MLE-length was used as a base method rather than MPL. An even greater improvement in computational runtime was seen: FastNet(MLE-length)'s runtime improvement over MLE-length was over an order of magnitude greater than FastNet(MPL)'s improvement over MPL. As the number of taxa increased from 15 to 20 (but the number of reticulations was fixed to one), FastNet(MLE-length)'s advantage in topological error and runtime relative to its base method nearly doubled. In all cases, FastNet(MLE-length)'s performance improvements were statistically significant (Benjamini-Hochberg-corrected one-sided pairwise t-test; $\alpha = 0.05$ and $n = 20$). Although FastNet(MLE-length) successfully completed analysis of larger datasets (i.e., model conditions with more than 20 taxa and/or more than one reticulation), we were unable to quantify FastNet(MLE-length)'s performance relative to its base method due to MLE-length's scalability limitations.

We further evaluated FastNet's performance in the context of additional experimental and methodological considerations. On model conditions with deep gene flow (Table 3), FastNet returned significant improvements in topological accuracy and runtime relative to its base method – either MPL or MLE-length – with one exception: on the 15-taxon single-reticulation model condition, FastNet(MPL) returned a small and statistically insignificant improvement in topological error over MPL. Otherwise, FastNet's performance boost was robust to the choice of base method. As dataset sizes increased, the average performance boost increased when MPL was the base method; a similar finding applied to runtime improvements when MLE-length was the base method, whereas

Table 2. FastNet(MLE-length) "boosts" MLE-length's runtime and topological accuracy, where a greater performance boost occurs as dataset sizes increase. The relative performance of FastNet(MLE-length) and MLE-length is compared on model conditions with 15–20 taxa and 1–2 non-deep reticulations. Note that, for the model condition with 20 taxa and 2 reticulations, MLE-length did not finish analysis of any replicates after a week of runtime. Otherwise, table layout and description are identical to Table 1.

Number of taxa	Number of reticulations	Improvement in topological error			Improvement in runtime (h)		
		Avg	SE	Corrected q-value	Avg	SE	Corrected q-value
15	1	0.103	0.021	8.8×10^{-4}	49.4	6.9	9.1×10^{-7}
20	1	0.195	0.024	6.1×10^{-5}	114.3	14.7	3.3×10^{-7}
20	2	Base method DNF					

topological error improvements were largely unchanged. We note that FastNet's performance boost was somewhat smaller on model conditions involving deep gene flow as opposed to non-deep gene flow. When maximum-likelihood-estimated gene trees were used as input to summary-based inference in lieu of true gene trees (Table 4), FastNet boosted the topological accuracy and runtime of its base method in all cases and the improvements were statistically significant. As dataset sizes increased, FastNet's improvement in topological accuracy and runtime grew when MPL was its base method; runtime improvements grew and topological error improvements were largely unchanged when MLE-length was the base method. Finally, we conducted an additional experiment to evaluate FastNet's statistical efficiency when given a finite number of observations in terms of the number of loci (Table 5). As the number of loci ranged from genome-scale (i.e., on the order of 1000 loci) to sizes that were smaller by up to an order of magnitude, FastNet's average topological error increased by less than 0.02.

Table 3. Boosting experiments on model conditions with deep gene flow. The performance improvement of FastNet over its base method (either MPL or MLE-length) is reported for two different performance measures: topological error as measured by tripartition fraction and computational runtime in hours. The simulation conditions involved either 15 or 20 taxa and a single deep reticulation. Otherwise, table layout and description are identical to Table 1.

Number of taxa	Boosted method	Improvement in topological error			Improvement in runtime (h)		
		Avg	SE	q-value	Avg	SE	q-value
15	MPL	0.015	0.017	3.8×10^{-1}	2.3	0.2	5.1×10^{-4}
20	MPL	0.166	0.035	3.2×10^{-3}	8.0	1.5	3.2×10^{-3}
15	MLE-length	0.066	0.001	1.5×10^{-2}	35.0	4.1	1.3×10^{-7}
20	MLE-length	0.070	0.014	1.1×10^{-2}	71.1	7.7	8.7×10^{-8}

Table 4. Boosting experiments using inferred gene trees. The performance improvement of FastNet over its base method (either MPL or MLE-length) is reported for two different performance measures: topological error as measured by tripartition fraction and computational runtime in hours. For each replicate dataset, all summary-based methods were provided with the same input: a set of rooted gene trees that was inferred using FastTree and outgroup rooting (see Methods section for more details). The simulation conditions involved either 15 or 20 taxa and 1–2 non-deep reticulations. Otherwise, table layout and description are identical to Table 1.

Number of taxa	Number of reticulations	Boosted method	Improvement in topological error			Improvement in runtime (h)		
			Avg	SE	q-value	Avg	SE	q-value
15	1	MPL	0.071	0.021	1.2×10^{-2}	3.8	0.5	7.7×10^{-5}
20	2	MPL	0.134	0.017	1.4×10^{-2}	15.1	1.7	6.9×10^{-6}
15	1	MLE-length	0.231	0.002	1.3×10^{-4}	15.4	2.0	6.7×10^{-7}
20	1	MLE-length	0.195	0.005	5.8×10^{-5}	43.2	7.3	1.7×10^{-5}

Table 5. The impact of the number of observed loci on FastNet(MLE)'s topological error. The inputs to FastNet(MLE) consisted of gene trees that were inferred using FastTree and outgroup rooting (see Methods section for more details). The simulations sampled between 100 and 1000 loci for a single 20-taxon 1-reticulation model condition involving non-deep gene flow. Topological error was evaluated based upon the tripartition fraction between the model phylogeny and the species phylogeny inferred by FastNet(MLE); average ("Avg") and standard error ("SE") are shown ($n = 20$).

Number of loci	Topological error	
	Avg	SE
100	0.094	0.028
200	0.078	0.024
1000	0.075	0.027

4 Discussion

Relative to the state-of-the-art methods that served as base methods, FastNet consistently returned sizeable and statistically significant improvements in topological error and computational runtime across a range of dataset scales and gene flow scenarios. There was only a single experimental condition where comparable error without statistically significant improvements was seen. This exception occurred when FastNet was used to boost a relatively inaccurate base method (MPL) on the smallest dataset sizes in our study and with deep gene flow; even still, large and statistically significant runtime improvements were seen in this case. In contrast, with a more accurate base method (i.e., MLE-length), large and statistically significant performance improvements were seen throughout our simulation study.

FastNet's boosting effect on topological error and runtime were robust to several different experimental and design factors. The boosting performance obtained using different base methods – one with lower computational requirements but higher topological error relative to a more computationally intensive alternative – suggests that, while accuracy improvements can be obtained even using less accurate subproblem inference, even greater accuracy improvements can be obtained when reasonably accurate subproblem phylogenies can be inferred. We note that the base methods were run in default mode. More intensive search settings for each base method's optimization procedures may allow a tradeoff between topological accuracy and computational runtime. We stress that our goal was *not* to make specific recommendations about the nuances of running the base methods. Rather, FastNet's divide-and-conquer framework can be viewed as orthogonal to the specific algorithmic approaches utilized by a base method. In this sense, improvements to the latter accrue to the former in a straightforward and modular manner. Furthermore, FastNet's performance effect was robust to gene tree error and varying numbers of observed loci.

The biggest performance gains were observed on the largest, most challenging datasets. The findings in our earlier performance study [17] suggest that, given weeks of computational runtime, even the fastest statistical methods (including MPL) would not complete analysis of datasets with more than 50 taxa or so and several reticulations. In comparison to MPL, FastNet(MPL) was faster by more than an order of magnitude on the largest datasets in our study, and we predict that FastNet(MPL) would readily scale to datasets with many dozens of taxa and multiple reticulations.

5 Conclusions

In this study, we introduced FastNet, a new computational method for inferring phylogenetic networks from large-scale genomic sequence datasets. FastNet utilizes a divide-and-conquer algorithm to constrain two different aspects of scale: the number of taxa and evolutionary divergence. We evaluated the performance of FastNet in comparison to state-of-the-art phylogenetic network inference methods. We found that FastNet improves upon existing methods in terms of computational efficiency and topological accuracy. On the largest datasets explored in our study, the use of the FastNet algorithm as a boosting framework enabled runtime speedups that were over an order of magnitude faster than standalone analysis using a state-of-the-art method. Furthermore, FastNet returned comparable or typically improved topological accuracy compared to the state-of-the-art-methods that were used as its base method.

Acknowledgments. We gratefully acknowledge the following support: NSF grants no. CCF-1565719 (to KJL), CCF-1714417 (to KJL), and DEB-1737898 (to GMB and KJL), BEACON grants (NSF STC Cooperative Agreement DBI-093954) to GMB and KJL, and computing resources provided by MSU HPCC. We would also like to acknowledge Daniel Neafsey for kindly sending us a processed version of the genomic sequence dataset from [36].

References

1. Abbott, R.J., Rieseberg, L.H.: Hybrid speciation. In: Seligman, E.R.A., Johnson, A. (eds.) Encyclopaedia of Life Sciences. Wiley, Hoboken (2012)
2. Akaike, H.: Information theory and an extension of the maximum likelihood principle. In: Parzen, E., Tanabe, K., Kitagawa, G. (eds.) Selected Papers of Hirotugu Akaike. Springer Series in Statistics (Perspectives in Statistics). Springer, New York (1998). https://doi.org/10.1007/978-1-4612-1694-0_15
3. Akaike, H.: A new look at the statistical model identification. IEEE Trans. Autom. Control **19**(6), 716–723 (1974)
4. Bandelt, H.-J., Dress, A.W.M.: A canonical decomposition theory for metrics on a finite set. Adv. Math. **92**(1), 47–105 (1992)
5. Baroni, M., Semple, C., Steel, M.: Hybrids in real time. Syst. Biol. **55**(1), 46–56 (2006)
6. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B (Methodological) **57**(1), 289–300 (1995)
7. Bryant, D., Moulton, V.: Neighbor-Net: an agglomerative method for the construction of phylogenetic networks. Mol. Biol. Evol. **21**(2), 255–265 (2004)
8. Cardona, G., Rosselló, F., Valiente, G.: Tripartitions do not always discriminate phylogenetic networks. Math. Biosci. **211**(2), 356–370 (2008)
9. Durand, E.Y., Patterson, N., Reich, D., Slatkin, M.: Testing for ancient admixture between closely related populations. Mol. Biol. Evol. **28**(8), 2239–2252 (2011)
10. Edwards, S.V.: Is a new and general theory of molecular systematics emerging? Evolution **63**(1), 1–19 (2009)
11. Felsenstein, J.: Cases in which parsimony or compatibility methods will be positively misleading. Syst. Biol. **27**(4), 401–410 (1978)
12. Felsenstein, J.: Inferring Phylogenies. Sinauer Associates, Sunderland, Massachusetts (2004)
13. Francis, A.R., Steel, M.: Which phylogenetic networks are merely trees with additional arcs? Syst. Biol. **64**(5), 768–777 (2015)
14. Gluck-Thaler, E., Slot, J.C.: Dimensions of horizontal gene transfer in eukaryotic microbial pathogens. PLoS Pathog. **11**(10), e1005156 (2015)
15. Green, R.E., et al.: A draft sequence of the Neandertal genome. Science **328**(5979), 710–722 (2010)
16. Hein, J., Schierup, M., Wiuf, C.: Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory. Oxford University Press, Oxford (2004)
17. Hejase, H.A., Liu, K.J.: A scalability study of phylogenetic network inference methods using empirical datasets and simulations involving a single reticulation. BMC Bioinform. **17**(1), 422 (2016)
18. Hudson, R.R.: Generating samples under a wright-fisher neutral model of genetic variation. Bioinformatics **18**(2), 337–338 (2002)
19. Huelsenbeck, J.P., Hillis, D.M.: Success of phylogenetic methods in the four-taxon case. Syst. Biol. **42**(3), 247–264 (1993)
20. Hurvich, C.M., Tsai, C.-L.: Regression and time series model selection in small samples. Biometrika **76**(2), 297–307 (1989)
21. Huson, D.H., Rupp, R., Scornavacca, C.: Phylogenetic Networks: Concepts Algorithms and Applications. Cambridge University Press, Cambridge, United Kingdom (2010)

22. Jukes, T.H., Cantor, C.R.: Evolution of Protein Molecules, p. 132. Academic Press, New York (1969)
23. Keeling, P.J., Palmer, J.D.: Horizontal gene transfer in eukaryotic evolution. Nat. Rev. Genet. **9**(8), 605–618 (2008)
24. Kingman, J.F.C.: The coalescent. Stoch. Process. Appl. **13**(3), 235–248 (1982)
25. Leaché, A.D., Harris, R.B., Rannala, B., Yang, Z.: The influence of gene flow on species tree estimation: a simulation study. Syst. Biol. **63**, 17–30 (2013)
26. Liu, K., Raghavan, S., Nelesen, S., Linder, C.R., Warnow, T.: Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. Science **324**(5934), 1561–1564 (2009)
27. Liu, K., et al.: SATé-II: Very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. Syst. Biol. **61**(1), 90–106 (2012)
28. Liu, K.J., Steinberg, E., Yozzo, A., Song, Y., Kohn, M.H., Nakhleh, L.: Interspecific introgressive origin of genomic diversity in the house mouse. Proc. Nat. Acad. Sci. **112**(1), 196–201 (2015)
29. McInerney, J.O., Cotton, J.A., Pisani, D.: The prokaryotic tree of life: past, present... and future? Trends Ecol. Evol. **23**(5), 276–281 (2008)
30. Metzker, M.L.: Sequencing technologies - the next generation. Nat. Rev. Genet. **11**(1), 31–46 (2010)
31. Mirarab, S., Warnow, T.: ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics **31**(12), i44–i52 (2015)
32. Mirarab, S., Reaz, R., Bayzid, M.S., Zimmermann, T., Swenson, M.S., Warnow, T.: ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics **30**(17), i541–i548 (2014)
33. Mirarab, S., Nguyen, N., Guo, S., Wang, L.-S., Kim, J., Warnow, T.: PASTA: ultra-large multiple sequence alignment for nucleotide and amino-acid sequences. J. Comput. Biol. **22**(5), 377–386 (2015)
34. Nakhleh, L.: Computational approaches to species phylogeny inference and gene tree reconciliation. Trends Ecol. Evol. **28**(12), 719–728 (2013)
35. Nakhleh, L., Sun, J., Warnow, T., Linder, C.R., Moret, B.M., Tholse, A.: Towards the development of computational tools for evaluating phylogenetic network reconstruction methods. In: Pacific Symposium on Biocomputing, vol. 8, pp. 315–326. World Scientific (2003)
36. Neafsey, D.E.: Highly evolvable malaria vectors: the genomes of 16 Anopheles mosquitoes. Science **347**(6217), 1258522 (2015)
37. Price, M., Dehal, P., Arkin, A.: FastTree 2 - approximately maximum-likelihood trees for large alignments. PLoS ONE **5**(3), e9490 (2010)
38. Rambaut, A., Grassly, N.C.: Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. **13**, 235–238 (1997)
39. Reich, D., et al.: Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature **468**(7327), 1053–1060 (2010)
40. Sanderson, M.J.: r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics **19**(2), 301–302 (2003)
41. Schwarz, G.: Estimating the dimension of a model. Annal. Stat. **6**(2), 461–464 (1978)
42. Solís-Lemus, C., Ané, C.: Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. PLoS Genet. **12**(3), 1–21 (2016)

43. Than, C., Ruths, D., Nakhleh, L.: PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. BMC Bioinform. **9**(1), 322 (2008)
44. The Heliconious Genome Consortium: Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature **487**(7405), 94–98 (2012)
45. Yun, Y., Nakhleh, L.: A maximum pseudo-likelihood approach for phylogenetic networks. BMC Genomics **16**(Suppl 10), S10 (2015)
46. Yu, Y., Cuong, T., Degnan, J.H., Nakhleh, L.: Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. Syst. Biol. **60**(2), 138–149 (2011)
47. Yu, Y., Degnan, J.H., Nakhleh, L.: The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. PLoS Genet. **8**(4), pp. e1002660 (2012)
48. Yu, Y., Dong, J., Liu, K.J., Nakhleh, L.: Maximum likelihood inference of reticulate evolutionary histories. Proc. Nat. Acad. Sci. **111**(46), 16448–16453 (2014)
49. Zhang, L.: On tree-based phylogenetic networks. J. Comput. Biol. **23**(7), 553–565 (2016)