Syst. Biol. 0(0):1–8, 2019

DOI:10.1093/sysbio/syz045

INSTRAL: Discordance-Aware Phylogenetic Placement Using Quartet Scores

Maryam Rabiee[1] and Siavash Mirarab[2,*]

[1]*Department of Computer Science and Engineering, UC San Diego, La Jolla, CA 92093, USA; and* [2]*Department of Electrical and Computer Engineering, UC, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA*
**Correspondence to be sent to: Department of Electrical and Computer Engineering, UC, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA; E-mail: smirarab@ucsd.edu.*

Received 04 November 2018; reviews returned 25 June 2019; accepted 02 July 2019
Associate Editor: David Posada

Abstract.—Phylogenomic analyses have increasingly adopted species tree reconstruction using methods that account for gene tree discordance using pipelines that require both human effort and computational resources. As the number of available genomes continues to increase, a new problem is facing researchers. Once more species become available, they have to repeat the whole process from the beginning because updating species trees is currently not possible. However, the *de novo* inference can be prohibitively costly in human effort or machine time. In this article, we introduce INSTRAL, a method that extends ASTRAL to enable phylogenetic placement. INSTRAL is designed to place a new species on an existing species tree after sequences from the new species have already been added to gene trees; thus, INSTRAL is complementary to existing placement methods that update gene trees. [ASTRAL; ILS; phylogenetic placement; species tree reconstruction.]

Gene trees and species trees can differ (Maddison 1997; Degnan and Rosenberg 2009), and methods for accounting for discordance are now widely available and are adopted by many (Szöllsi et al. 2014; Edwards et al. 2016). Discordance-aware methods come in many forms, such as coestimation of gene trees and species trees (e.g., Liu 2008; Heled and Drummond 2010; Boussau et al. 2013) and site-based methods (e.g., Bryant et al. 2012; De Maio et al. 2013; Chifman and Kubatko 2014; Schrempf et al. 2016). The most scalable approach for species reconstruction has remained what has been called a summary approach: gene trees are inferred independently for all loci and are then combined to build a species tree. Many methods are available for combining gene trees (e.g., Kubatko et al. 2009; Liu et al. 2009, 2010; Chaudhary et al. 2010; Mossel and Roch 2010; Liu and Yu 2011; Wu 2012; Bayzid et al. 2013; Sayyari and Mirarab 2016a), and many of them are statistically consistent under various models of genome evolution. In particular, many methods have been designed to be consistent under the multispecies coalescent model (Pamilo and Nei 1988; Rannala and Yang 2003), which seeks to capture incomplete lineage sorting (ILS). Several statistically consistent summary methods, including ASTRAL (Mirarab et al. 2014a), NJst/ASTRID (Liu and Yu 2011; Vachaspati and Warnow 2015), and MP-EST (Liu et al. 2010) are in wide use.

Despite the progress for the *de novo* inference of species trees, updating trees under the MSC model has received little attention. As new genomes become available, researchers often need to know their position on an existing phylogeny. One solution is to reconstruct the species tree from scratch each time new data becomes available. This process can require excessive computation and may not scale to groups with tens of thousands of genomes (more than a hundred thousand bacterial genomes are currently available).

A more efficient alternative is what has been called phylogenetic placement (Matsen et al. 2010): adding a new *query* species onto an existing phylogeny. For placing a new sequence onto a single tree, we have maximum likelihood (ML) methods such as `pplacer` (Matsen et al. 2010) and `EPA` (Berger et al. 2011; Barbera et al. 2019), distance-based methods such as `APPLES` (Balaban et al. 2018), and divide-and-conquer methods such as `SEPP` (Mirarab et al. 2012). Even earlier, sequential sequence insertion algorithms, which essentially solve the same computational problem, existed (e.g., Felsenstein 1981; Desper and Gascuel 2002).

Existing placement algorithms place a new sequence onto a single tree, which is typically a gene tree. Current methods can be used to place new sequences on an estimate of the species tree using a concatenation of multiple genes, but this approach ignores gene tree discordance. We are not aware of any discordance-aware methods for placement onto species trees. Here, we present INSTRAL (Insertion of New Species using asTRAL) which extends ASTRAL to enable placing a new species onto an existing species tree.

Description

ASTRAL estimates an unrooted species tree given a set of unrooted gene trees and is statistically consistent under the multispecies coalescent model given true gene trees (Mirarab et al. 2014a). ASTRAL seeks to maximize the quartet score: the total number of induced quartet trees in the gene trees that match the species tree. Similar to earlier work (Bryant and Steel 2001), ASTRAL uses dynamic programming to solve this NP-Hard problem (Lafond and Scornavacca 2018). However, to allow scalability, it constrains its search space so that the output draws its clusters from a predefined set X, which consists of clusters from gene trees and others that are heuristically selected (a cluster is one side of a bipartition). The most recent version, ASTRAL-III

(Zhang et al. 2018) guarantees polynomial running time and scales to data sets with many thousands of species.

Problem Statement

Quartet placement problem. Given a set of k unrooted trees labeled with $n+1$ species and a backbone tree on n species, find the tree that includes all $n+1$ species and has the maximum quartet score with respect to the input trees.

Thus, one species, called the *query*, is not present in the backbone tree, and the goal is to insert the query species into the backbone. A typical use of this problem is placing a new species onto an existing species tree (Fig. 1). Imagine a previous analysis has already produced a set of k gene trees on n species and an ASTRAL tree (inferred from those k gene trees). Now, a new species with genome-wide data has become available. To insert the new species onto a given ASTRAL tree, we first add it to each of the k gene trees using tools such as `SEPP`, `pplacer`, or `EPA`. Then, we use the updated gene trees in addition to the existing ASTRAL tree as input to the quartet placement problem; the output will be a species tree with the new species included. Just like ASTRAL, the use of the quartet score ensures that the inferred position of the new species is a statistically consistent estimator of its true position under the MSC model given true gene trees.

INSTRAL (Single Query)

INSTRAL finds the optimal solution to the quartet placement problem. Unlike ASTRAL, the number of possible solutions to the placement problem is small (grows linearly with n), and thus, INSTRAL can solve the problem exactly even for large trees. In principle, it is possible to develop algorithms that compute the quartet score for all possible branches, one at a time, and to select the optimal solution at the end. However, the ASTRAL dynamic programming allows for a more straight-forward algorithm.

The ASTRAL algorithm can solve the placement problem if we define the search space (set X) such that *all* trees that induce the backbone tree and *only those trees* are allowed. To achieve this, X should be the set of all clusters in the backbone tree both with and without the new species added. More precisely, let q be a set with the new species and let $\mathcal{B}(T)$ denote the set of all (including trivial) bipartitions of the backbone tree T on the leaf-set $\mathcal{L}$ with each bipartition represented as a tuple: $(A,\mathcal{L}\setminus A)$ where $A\subset\mathcal{L}$. Then

$$X=\{q,\mathcal{L}\cup q,\mathcal{L}\}\cup\bigcup_{(A,\mathcal{L}\setminus A)\in\mathcal{B}(T)}\{A,\mathcal{L}\setminus A,A\cup q,(\mathcal{L}\setminus A)\cup q\}. \tag{1}$$

With this set X, the search space will include all possible placement of the query on the backbone tree (due to $A\cup q$ and $(\mathcal{L}\setminus A)\cup q$). Moreover, every bipartition built from X is one that existed in the backbone tree once q is removed and thus only trees that induce the backbone are allowed. Since ASTRAL finds the optimal placement restricted to the search space, this algorithm is guaranteed to solve the quartet placement problem exactly. The number of clusters in this search space is $3+4(2n-3)=\Theta(n)$. Thus, its running time increases as $\Theta(nD)=O(n^2k)$ where D is the sum of degrees of all *unique* nodes in the input gene trees (see for details, Zhang et al. 2018).

Adding Multiple New Species

If multiple queries are available, we can still attempt to use the basic INSTRAL algorithm in one of two ways (Fig. 1). (*i*) *Independent placement:* We add all the new queries independently without trying to find the relationship among the queries. This approach is reasonable if the goal is to detect the identity of some unknown species or if the set of new species are expected not to belong to the same branches of the backbone tree. If needed, we can merge separate placements into a single tree, introducing polytomies wherever multiple queries are placed on the same branch. (*ii*) *Ordered placement:* We order the queries (e.g., arbitrarily) and then add them to the backbone one at a time, updating the backbone tree each time to include the latest query. This ordered placement approach gives us the relationships between queries. However, like similar greedy algorithms (Desper and Gascuel 2002), it is not guaranteed to find the optimal tree at the end.

The advantage in using the independent insertion approach is that adding m queries requires time that increases linearly with m, whereas the time needed for the ordered placement increases proportionally to m^3. The *de novo* execution of ASTRAL-III on $n+m$ species requires $O(((m+n)k)^{2.73})$ time in the worst case (Zhang et al. 2018). In contrast, INSTRAL-independent would run in $\Theta(m.D.n)=O(mn^2k)$ and INSTRAL-ordered would require $O(m^3k+(n+m)nmk)$. Thus, the relative running time of ASTRAL-III and INSTRAL-ordered depend on values of n, m, and k, while INSTRAL-independent is always faster than ASTRAL-III.

We can also ask a statistical consistency question: Starting from a correct backbone tree and placing several new species using INSTRAL, is the output tree guaranteed to be correct with high probability as the number of error-free gene trees drawn under MSC goes to infinity? In the independent placement scenario, the output is an unresolved tree and cannot be a statistically consistent estimate of the species tree. However, due to the consistency of each placement, with arbitrarily high probability, *all* placements are on the correct branch of the backbone given enough gene trees, and therefore, the output tree will not have wrong branches with high probability (but it will have missing branches). In the ordered placement scenario, since each placement is

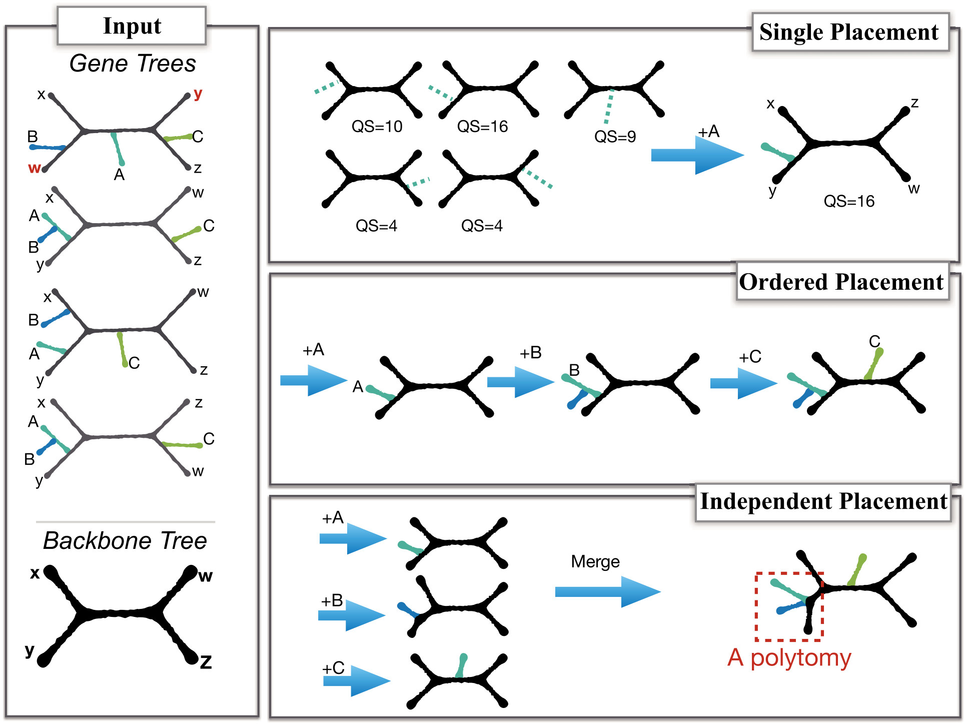

FIGURE 1. Left: The quartet placement problem. A backbone species tree with four leaves ($\{x,y,w,z\}$) and $k=4$ gene trees are given; each gene tree also has new species (here, $\{A,B,C\}$). Note that the first gene tree is discordant with the species tree. Top right: placing a single new species (A) on the backbone tree requires computing the quartet score (QS) for each placement and finding the maximum. Here, the optimal placement is on the terminal branch of y, which matches 16 out of 20 quartets on $\{x,y,w,z,A\}$ in the gene trees. Middle right: placing multiple species can be done by ordering them and placing them one at a time. Bottom right: alternatively, all new species can be placed independently, and the results can be merged at the end (creating polytomies when multiple new species are placed on the same branch).

correct with an arbitrarily high probability given enough genes, we can make *all* placements be correct with an arbitrarily high probability. Thus, the ordered placement result is a statistically consistent estimate of the species tree (see Supplementary material available at Dryad at http://dx.doi.org/10.5061/dryad.cs59t13 for proof). Note that this consistency is despite the fact that the ordered placement is *not* an optimal solution to the problem that ASTRAL seeks to solve.

BENCHMARK

Data Sets

We first benchmark INSTRAL on a simulated data set previously generated by Mirarab and Warnow (2015). This data set has 200 ingroup taxa and an outgroup species and is generated using SimPhy (Mallo et al. 2015). By setting the maximum tree heights to 10^7, 2×10^6, or 5×10^5 generations, this data set has created three model conditions with respectively, moderate, high, or very high levels of ILS; the average normalized (Robinson and Foulds, 1981) distance (RF) between true gene trees and the true species tree are 15%, 34%, and 69%, respectively. In our experiments, we use gene trees inferred using FastTree-II (Price et al. 2010) from sequence data. These inferred trees have relatively high levels of gene tree error (25%, 31%, and 47% for the three model conditions). For each replicate, we also have estimates of the species tree using both ASTRAL-II and concatenation with ML (CA-ML) performed using FastTree-II. We have 100 replicates per condition, and each replicate has 1000 gene trees, from which we have randomly sampled 200 and 50 gene trees to create three different input sets. Thus, in total, we have 9 model conditions (ILS level × # Gene). Following (Mirarab and Warnow, 2015), three replicates are removed because their gene trees are extremely unresolved; this leaves us with $9\times100-3\times3=891$ data sets in total.

Leave-One-Out Experiments

Comparision to ASTRAL. For each data set, an ASTRAL species tree inferred from gene trees is available. For each of the 200 ingroup species in each data set, we prune it from the ASTRAL tree and we use INSTRAL to add it back onto the tree, using FastTree gene trees as input. Thus, overall, we have $891\times200=1.782\times10^5$ independent placements. When there are multiple placements with equal quartet scores (happens in only 63 cases), we break ties identically to the full backbone ASTRAL tree.

TABLE 1. For each condition, we show the number of cases where (left) the INSTRAL tree has a different (i.e., higher) quartet score than the full ASTRAL tree, (middle) the (Robinson and Foulds, 1981) distance (RF) of the INSTRAL tree to the true tree is different than the RF distance of the full ASTRAL tree to the true tree, and (right) the INSTRAL tree has a *reduced* RF distance to the true tree compared with the full ASTRAL tree

	50 genes	200 genes	1000 genes
Moderate ILS	11; 8; 1	5; 3; 0	4; 4; 0
High ILS	41; 31; 13	12; 8; 5	5; 3; 2
Very high ILS	178; 140; 26	41; 33; 7	19; 12; 6

Note: All numbers are out of 20,000 insertions, except for very high ILS, which is out of 19,400.

Among all of these placements, in only 316 cases (<0.2%) the output trees have different quartet scores compared with the original ASTRAL tree. Note that INSTRAL is guaranteed to find the optimal placement, and therefore, its quartet score is always at least as good as the ASTRAL tree. Thus, these 316 cases are those where ASTRAL has failed to find the optimal placement for a species. We note that 178 out of 316 cases correspond to the model condition with very high ILS and only 50 genes. Increasing the number of genes and reducing the amount of ILS both decrease the number of cases where ASTRAL is sub-optimal (Table 1). For example, with moderate ILS/1000 genes, only four out of 20,000 placements using INSTRAL improved quartet scores compared with ASTRAL. Only 176 of 316 cases result in any change in the RF distance of the inferred tree compared with the true tree, and only in 59 out of 176 cases did INSTRAL reduce the RF distance compared with ASTRAL. Thus, removing and reinserting a species using INSTRAL is generally consistent with the ASTRAL tree but in rare cases improves quartet scores.

Comparison to concatenation using ML (CA-ML). An alternative to INSTRAL is to simply concatenate all the genes and use ML to place the query on an existing tree. We compare INSTRAL to this CA-ML approach using EPA-ng (Barbera et al. 2019) v0.3.5 for ML placement. To avoid biasing results towards one method, we use the true species tree as the backbone, both for INSTRAL and CA-ML. For CA-ML, we use RAxML (Stamatakis 2014) to compute branch lengths of backbone and GTR+Γ model parameters based on true alignment. We test INSTRAL with two types of input. In one case, gene trees are computed *de novo* using FastTree-II. In leave-one-out experiments, we approximate this scenario by simply removing each species from the species tree but keeping it in all our estimated gene trees. In the second case, gene trees are updated using EPA-ng; thus, we first remove the query species from all gene trees and then place it on each gene tree using EPA-ng. Due to memory requirements of EPA-ng (up to 35GB), we could only run it for up to 200 genes and we restrict leave-one-out tests to only 50 randomly selected leaves.

In terms of accuracy, INSTRAL outperforms the CA-ML using EPA-ng regardless of the amount of ILS or the number of genes (Supplementary Table S1 available on Dryad and Fig. 2a). For example, with high ILS and 200 genes, CA-ML fails to find the correct placement in 17% of cases, while INSTRAL is incorrect in 5% and 8%, respectively, with *de-novo* and EPA-ng gene trees (Supplementary Table S1 available on Dryad). When methods are wrong, there are typically off by one edge and only rarely by two or more edges (Supplementary Fig. S1 available on Dryad). As the level of discordance goes up, the error increases for all methods, and contrary to our expectations, the relative performance of methods does not change. However, as the number of genes increases from 50 to 200, INSTRAL enjoys a substantial reduction in error but CA-ML benefits less from increased gene sampling for moderate to high levels of ILS (e.g., for moderate ILS, mean error drops from 0.08 edges to 0.04 for INSTAL+*de novo* but only from 0.18 to 0.17 for CA-ML). In all conditions, using *de novo* gene trees resulted in improved accuracy compared with using EPA-ng for updating gene trees; however, INSTRAL+EPA-ng is still substantially more accurate than CA-ML.

Comparing the total running time, INSTRAL+EPA-ng takes twice as much time as CA-ML (Fig. 2b). INSTRAL+EPA-ng took on average about 200 s, of which, on average only 3 s were spend by INSTRAL and the rest was used up by EPA-ng on gene trees. However, note that gene tree updating using EPA-ng enjoys trivial parallelism (each gene tree can be assigned to a different CPU), whereas CA-ML does not enjoy trivial parallelism. Finally, using INSTRAL+EPA-ng requires a lot less memory than CA-ML using EPA-ng. CA-ML needed up to 35GB of memory (mean 19GB), while INSTRAL+EPA-ng runs with less than 0.5GB of memory in every case (Supplementary Fig. S2 available on Dryad).

Ordered Placement

To see if the agreement with ASTRAL remains if more species are placed using INSTRAL, we perform a second experiment. Here, we prune a portion ($\frac{1}{4}$, $\frac{1}{2}$, or $\frac{3}{4}$) of species from the ASTRAL species tree, order removed species randomly, and then place them one after another on the backbone tree, updating the backbone tree each time (Ordered Placement in Fig. 1). In the end, we have a tree on the full leaf-set; this tree, which we call the INSTRAL tree, can be thought of as a greedy solution to the same problem ASTRAL seeks to solve.

ASTRAL and INSTRAL trees have similar RF distances to the true tree, but ASTRAL is somewhat more accurate in the hardest conditions (Fig. 3a). Overall, the normalized RF error of ASTRAL is on average 0.3% lower than INSTRAL (corresponding to roughly half an edge), and these improvements are statistically significant ($P \ll 10^{-6}$ according to a paired t-test). Among all $891 \times 3 = 2673$ INSTRAL trees that we have computed, 1470 have RF distances to the true tree that are identical to the ASTRAL tree. Differences in the RF distance are seen

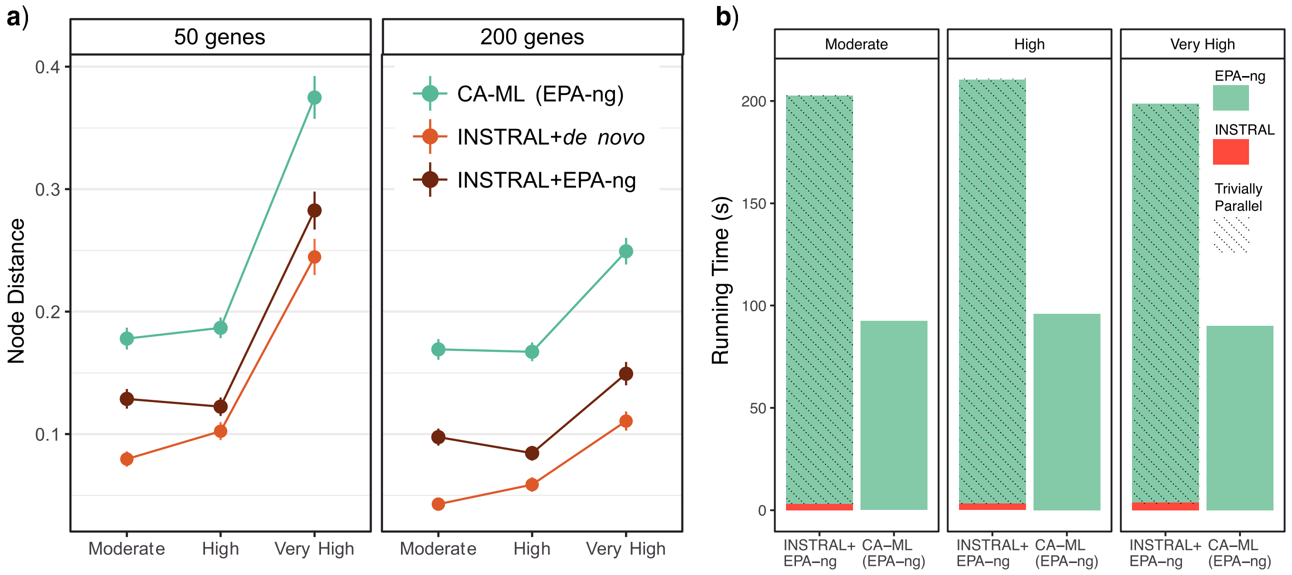

FIGURE 2. Comparison of concatenation using ML (CA-ML) and INSTRAL run on *de novo* gene trees or on gene trees updated using EPA-ng. Both method place on the true species tree in a leave-one-out experiment (50 species per replicate) with 200 or 50 genes. a) Mean and standard error of placement error, measured as the number of nodes between the correct placement and placed edge. Results are over 2500 placements for moderate and high and 2350 placements for very high ILS. b) Total running time in seconds, measured on the same machine, and all methods run with a single core both for EPA-ng and INSTRAL.

more often among replicates with very high ILS (mean RF difference: 0.7%), 50 genes (mean RF difference: 0.7%), or starting trees with 1/4 of the species (mean RF difference: 0.6%). Increasing the number of genes, increasing the size of the starting tree, and reducing the ILS reduce the number of mismatches between ASTRAL and INSTRAL (Fig. 3a).

Unlike leave-one-out tests, for multiple insertions, the quartet score of INSTRAL can be higher or lower than ASTRAL. Overall, when the two trees do not agree, ASTRAL tends to have higher quartet scores (Fig. 3b and Supplementary Fig. S3 available on Dryad). Out of 2673 cases, ASTRAL has higher quartet scores in 1210 cases, while INSTRAL is better in 231 (they tie in the remaining 1232). Reducing the number of genes and increasing the level of ILS both magnify the improvements of ASTRAL compared with INSTRAL.

Scalability

To test the scalability of INSTRAL, we started with a backbone tree of 10,000 species from a previous publication (Zhang et al. 2018), and down-sampled it to smaller trees (down to 250). Each time, we placed 400 to 800 genomes on the backbone and computed the time INSTRAL took for the insertion (Fig. 4). On the backbone of 10,000 species, each placement took close to 16 minutes on average. As the backbone size decreased, the running time rapidly decreased and was close to 8 s on a backbone tree of 250 species. As expected, the running time grows faster than linearly with the size of the backbone (proportional to $n^{1.3}$ in this case).

BIOLOGICAL EXAMPLES

We have tested INSTRAL on three biological data sets: two transcriptomic data sets on insects by (Misof et al., 2014) and plants by (Wickett et al., 2014), and an avian data set by (Jarvis et al., 2014). The insect data set includes 1478 protein-coding genes from 144 species spanning all of the insect diversity and has been recently re-analyzed using ASTRAL by Sayyari et al. 2017. The plant data set includes 103 species and 424 genes, and the original study reported an ASTRAL tree. The avian data set consists of 48 genomes representing all the orders of birds. For this data set, statistical binning was used to build 2022 supergene trees (Mirarab et al. 2014b) and Sayyari and Mirarab 2016b have published an ASTRAL tree on these supergene trees. Among these data sets, the avian data set has extremely high levels of gene tree discordance.

For each of these data sets, we removed species one by one and placed them back onto the species tree using INSTRAL. In every case, INSTRAL found the same position for the new species as the backbone ASTRAL tree. In contrast, EPA-ng on concatenated data of 1KP (the only data set where we were able to test CA-ML) failed to find the same placement as the backbone for 35 out of 103 species and was on average away from the backbone position by 0.53 edges (Supplementary Fig. S4 available on Dryad).

We also tested the ordered placement, where we randomly selected half of the species (20 replicates), removed them, ordered them, and inserted them back on the remaining part of the tree using INSTRAL. The resulting INSTRAL-ordered trees were similar to the full

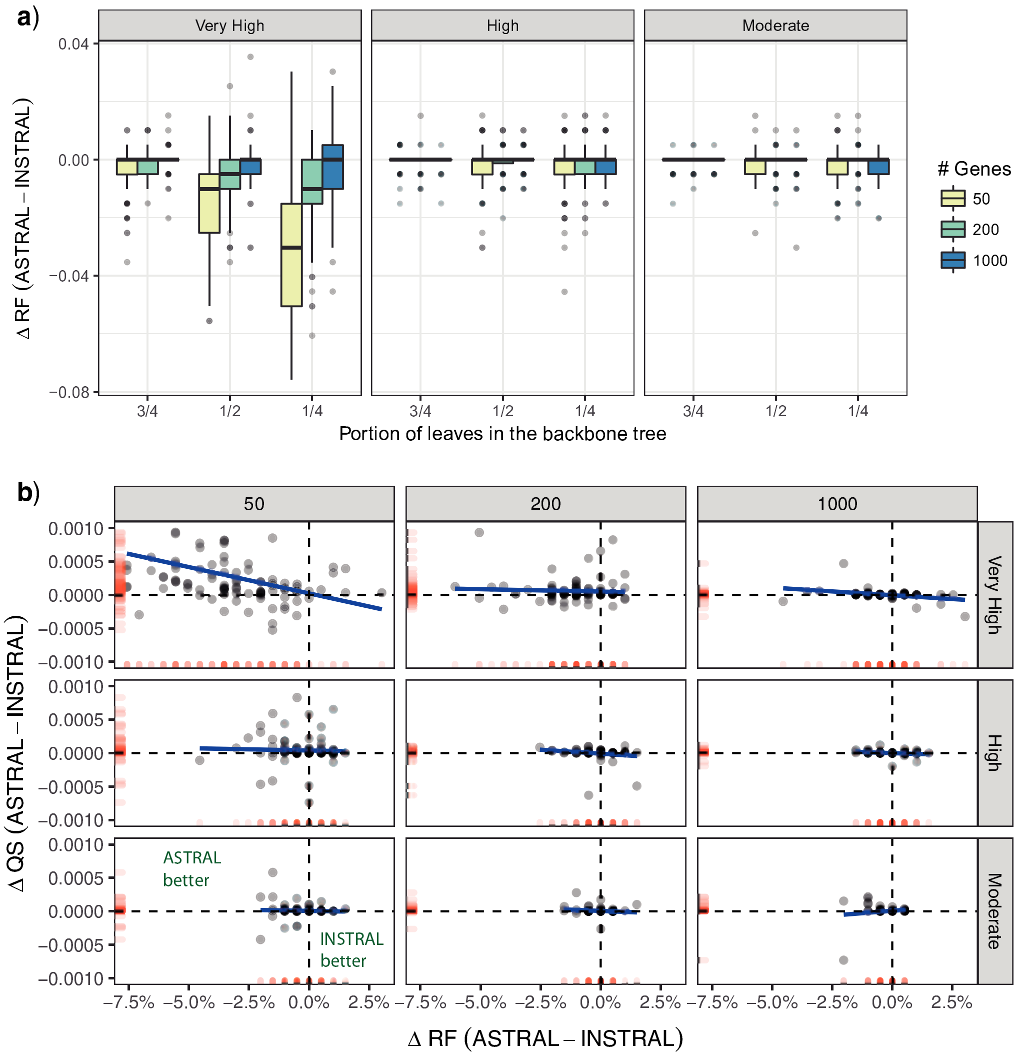

FIGURE 3. Comparison of ASTRAL and INSTRAL. a) Δ RF: The Robinson Foulds (RF) distance of the ASTRAL tree to the true tree minus the RF distance of INSTRAL-ordered tree to the true tree (negative: INSTRAL is better). The size of the starting tree is set to $\frac{1}{4}$, $\frac{1}{2}$, or $\frac{3}{4}$ of species (51, 101, or 151). For three levels of ILS (boxes), each with three numbers of genes, boxplots show distributions of ΔRF (100 points everywhere, except for very high ILS, where it is 97 points.) b) Change in the quartet score (QS) versus the ΔRF for the starting tree with $\frac{1}{4}$ of species (see Supplementary Fig. S3 available on Dryad for others). The marginal bars show the projection of data on each axis.

TABLE 2. The average and standard deviation of RF distance between ASTRAL and INSTRAL trees as well as the change in the quartet score (INSTRAL-ASTRAL) on 20 random sets for each biological data set

	Avian		Insects		1KP	
	Mean	Stdev	Mean	Stdev	Mean	Stdev
RF	0.0311	0.0592	0.0195	0.0126	0.0115	0.0131
QS	−0.0001098	0.0002973	−0.0000026	0.0000361	−0.0000039	0.0000057

Note: For each random set of leaves as a backbone tree, ordered placement has been done.

ASTRAL tree (Table 2), recovering the same tree in one-third of cases and changing by one or two branches in a majority of the remaining cases (Supplementary Fig. S5a available on Dryad). In several replicates, trees changed for five or more branches, including two replicates of the avian data set, where the INSTRAL differed from ASTRAL in nine branches. In both cases, two or three unstable taxa had moved by several branches, causing the high incongruence (Supplementary Fig. S5b available on Dryad). More broadly, changes are mostly

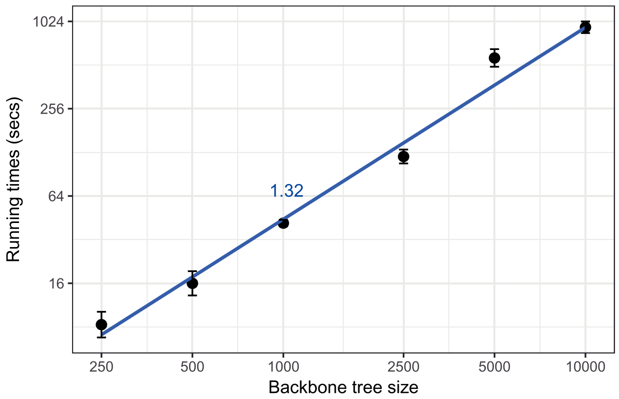

FIGURE 4. The running time scaling of INSTRAL versus backbone size n. Starting with a simulated data set with 10,000 leaves, we prune random sets of leaves to create smaller trees. Dots and bars show the average and standard error of the running times of inserting a new genome to the backbone (800 insertions for $n < 5000$ and 400 insertions for $n \geq 5000$). The slope (1.32) of the line fitted to this log–log plot gives an empirical estimate of the running time increasing proportionally to $n^{1.3}$, which is consistent with the theoretical running time complexity of $\Theta(n.D)$.

among unstable taxa. For example, in the avian tree, Hoatzin, the most challenging taxon, moves by one branch in several replicates. The resulting INSTRAL trees have reduced quartet scores compared to ASTRAL trees (Table 2). Overall, these results indicate that for data sets with very high ILS, using INSTRAL instead of ASTRAL runs the risk of producing sub-optimal trees.

AVAILABILITY

INSTRAL is available on GitHub (https://github.com/maryamrabiee/INSTRAL) in open-source. It is implemented in Java with straight-forward installation (the only dependency is Java 6+). A template tutorial and instructions to run INSTRAL is given there. The generated data, scripts to generate those data and results given in this article are also available on GitHub (https://github.com/maryamrabiee/INSTRAL-results).

SUPPLEMENTARY MATERIAL

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.cs59t13.

FUNDING

This work was supported by the NSF [IIS-1845967 to S.M. and M.R.]. Computations were performed on the San Diego Supercomputer Center (SDSC) through XSEDE allocations, which is supported by the NSF [ACI-1053575].

REFERENCES

Balaban M., Sarmashghi S., Mirarab S. 2018. APPLES: distance-based phylogenetic placement for scalable and assembly-free sample identification. *bioRxiv* 475566.

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

Bayzid M.S.M., Mirarab S., Warnow T. 2013. Inferring optimal species trees under gene duplication and loss. *Pac. Symp. Biocomput.* 18:250–261.

Berger S.A., Krompass D., Stamatakis A. 2011. Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood. *Syst. Biol.* 60:291–302.

Boussau B., Szöllősi G.J., Duret L. 2013. Genome-scale coestimation of species and gene trees. *Genome Res.* 23:323–330.

Bryant D., Steel M. 2001. Constructing optimal trees from quartets. *J. Algorithms* 38:237–259.

Bryant D., Bouckaert R., Felsenstein J., Rosenberg N.A., Roychoudhury A. 2012. Inferring species trees directly from biallelic genetic markers: Bypassing gene trees in a full coalescent analysis. *Mol. Biol. Evol.* 29:1917–1932.

Chaudhary R., Bansal M.S., Wehe A., Fernández-Baca D., Eulenstein O. 2010. iGTP: a software package for large-scale gene tree parsimony analysis. *BMC Bioinformatics* 11:574.

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

De Maio N., Schlötterer C., Kosiol C. 2013. Linking great apes genome evolution across time scales using polymorphism-aware phylogenetic models. *Mol. Biol. Evol.* 30:2249–2262.

Degnan J.H., Rosenberg N.A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.* 24:332–340.

Desper R., Gascuel O. 2002. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *J. Comput. Biol.* 9:687–705.

Edwards S.V., Xi Z., Janke A., Faircloth B.C., McCormack J.E., Glenn T.C., Zhong B., Wu S., Lemmon E.M., Lemmon A.R., Leaché A.D., Liu L., Davis C.C. 2016. Implementing and testing the multispecies coalescent model: a valuable paradigm for phylogenomics. *Mol. Phylogenet. Evol.* 94:447–462.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17:368–376.

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

Jarvis E.D., Mirarab S., Aberer A.J., Li B., Houde P., Li C., Ho S.Y.W., Faircloth B.C., Nabholz B., Howard J.T., Suh A., Weber C.C., da Fonseca R.R., Li J., Zhang F., Li H., Zhou L., Narula N., Liu L., Ganapathy G., Boussau B., Bayzid M.S., Zavidovych V., Subramanian S., Gabaldon T., Capella-Gutierrez S., Huerta-Cepas J., Rekepalli B., Munch K., Schierup M., Lindow B., Warren W.C., Ray D., Green R.E., Bruford M.W., Zhan X., Dixon A., Li S., Li N., Huang Y., Derryberry E.P., Bertelsen M.F., Sheldon F.H., Brumfield R.T., Mello C.V., Lovell P.V., Wirthlin M., Schneider M.P.C., Prosdocimi F., Samaniego J.A., Velazquez A.M.V., Alfaro-Nunez A., Campos P.F., Petersen B., Sicheritz-Ponten T., Pas A., Bailey T., Scofield P., Bunce M., Lambert D.M., Zhou Q., Perelman P., Driskell A.C., Shapiro B., Xiong Z., Zeng Y., Liu S., Li Z., Liu B., Wu K., Xiao J., Yinqi X., Zheng Q., Zhang Y., Yang H., Wang J., Smeds L., Rheindt F.E., Braun M., Fjeldsa J., Orlando L., Barker F.K., Jonsson K.A., Johnson W., Koepfli K.-P., O'Brien S., Haussler D., Ryder O.A., Rahbek C., Willerslev E., Graves G.R., Glenn T.C., McCormack J., Burt D., Ellegren H., Alstrom P., Edwards S.V., Stamatakis A., Mindell D.P., Cracraft J., Braun E.L., Warnow T., Jun W., Gilbert M.T.P., Zhang G. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* 346:1320–1331.

Kubatko L.S., Carstens B.C., Knowles L.L. 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics* 25:971–973.

Lafond M., Scornavacca C. 2018. On the weighted quartet consensus problem. *Theor. Comput. Sci.* 769:1–17.

Liu L. 2008. BEST: Bayesian estimation of species trees under the coalescent model. *Bioinformatics* 24:2542–2543.

Liu L., Yu L. 2011. Estimating species trees from unrooted gene trees. *Syst. Biol.* 60:661–667.

Liu L., Yu L., Pearl D.K., Edwards S.V. 2009. Estimating species phylogenies using coalescence times among sequences. *Syst. Biol.* 58:468–477.

Liu L., Yu L., Edwards S.V. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evol. Biol.* 10:302.

Maddison W.P. 1997. Gene trees in species trees. *Syst. Biol.* 46:523–536.

Mallo D., de Oliveira Martins L., Posada D. 2015. Simphy: phylogenomic simulation of gene, locus, and species trees. *Syst. Biol.* 65:334–344.

Matsen F.A., Kodner R.B., Armbrust E.V. 2010. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* 11:538.

Mirarab S., Warnow T. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* 31:i44–i52.

Mirarab S., Nguyen N., Warnow T. 2012. SEPP: SATé-enabled phylogenetic placement. *Pac. Symp. Biocomput.* 247–58.

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

Mirarab S., Bayzid M.S., Boussau B., Warnow T. 2014b. Statistical binning enables an accurate coalescent-based estimation of the avian tree. *Science* 346:1250463–1250463.

Misof B., Liu S., Meusemann K., Peters R.S., Donath A., Mayer C., Frandsen P.B., Ware J., Flouri T., Beutel R.G., Niehuis O., Petersen M., Izquierdo-Carrasco F., Wappler T., Rust J., Aberer A.J., Aspock U., Aspock H., Bartel D., Blanke A., Berger S., Bohm A., Buckley T.R., Calcott B., Chen J., Friedrich F., Fukui M., Fujita M., Greve C., Grobe P., Gu S., Huang Y., Jermiin L.S., Kawahara A.Y., Krogmann L., Kubiak M., Lanfear R., Letsch H., Li Y., Li Z., Li J., Lu H., Machida R., Mashimo Y., Kapli P., McKenna D.D., Meng G., Nakagaki Y., Navarrete-Heredia J.L., Ott M., Ou Y., Pass G., Podsiadlowski L., Pohl H., von Reumont B.M., Schutte K., Sekiya K., Shimizu S., Slipinski A., Stamatakis A., Song W., Su X., Szucsich N.U., Tan M., Tan X., Tang M., Tang J., Timelthaler G., Tomizuka S., Trautwein M., Tong X., Uchifune T., Walzl M.G., Wiegmann B.M., Wilbrandt J., Wipfler B., Wong T.K.F., Wu Q., Wu G., Xie Y., Yang S., Yang Q., Yeates D.K., Yoshizawa K., Zhang Q., Zhang R., Zhang W., Zhang Y., Zhao J., Zhou C., Zhou L., Ziesmann T., Zou S., Li Y., Xu X., Zhang Y., Yang H., Wang J., Wang J., Kjer K.M., Zhou X. 2014. Phylogenomics resolves the timing and pattern of insect evolution. *Science* 346:763–767.

Mossel E., Roch S. 2010. Incomplete lineage sorting: consistent phylogeny estimation from multiple loci. *IEEE/ACM Trans. Comput. Biol. Bioinformatics (TCBB)* 7:166–171.

Pamilo P., Nei M. 1988. Relationships between gene trees and species trees. *Mol. Biol. Evol.* 5:568–583.

Price M.N., Dehal P.S., Arkin A.P. 2010. FastTree-2 approximately maximum-likelihood trees for large alignments. *PLoS One* 5:e9490.

Rannala B., Yang Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* 164:1645–1656.

Robinson D., Foulds L. 1981. Comparison of phylogenetic trees. *Math. Biosci.* 53:131–147.

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

Sayyari E., Mirarab S. 2016b. Fast coalescent-based computation of local branch support from quartet frequencies. *Mol. Biol. Evol.* 33:1654–1668.

Sayyari E., Whitfield J.B., Mirarab S. 2017. Fragmentary gene sequences negatively impact gene tree and species tree reconstruction. *Mol. Biol. Evol.* 34:3279–3291.

Schrempf D., Minh B.Q., De Maio N., von Haeseler A., Kosiol C. 2016. Reversible polymorphism-aware phylogenetic models and their application to tree inference. *J. Theor. Biol.* 407:362–370.

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

SzöllÃµsi G.J., Tannier E., Daubin V., Boussau B. 2014. The inference of gene trees with species trees. *Syst. Biol.* 64:e42–e62.

Vachaspati P., Warnow T. 2015. ASTRID: accurate species TRees from internode distances. *BMC Genomics*,16(Suppl 10):S3.

Wickett N.J., Mirarab S., Nguyen N., Warnow T., Carpenter E.J., Matasci N., Ayyampalayam S., Barker M.S., Burleigh J.G., Gitzendanner M.A., Ruhfel B.R., Wafula E., Der J.P., Graham S.W., Mathews S., Melkonian M., Soltis D.E., Soltis P.S., Miles N.W., Rothfels C.J., Pokorny L., Shaw A.J., DeGironimo L., Stevenson D.W., Surek B., Villarreal J.C., Roure B., Philippe H., DePamphilis C.W., Chen T., Deyholos M.K., Baucom R.S., Kutchan T.M., Augustin M.M., Wang J.J., Zhang Y., Tian Z., Yan Z., Wu X., Sun X., Wong G.K.-S., Leebens-Mack J.J. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proc. Natl. Acad. Sci. USA* 111:4859–4868.

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

Zhang C., Rabiee M., Sayyari E., Mirarab S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* 19:153.