2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3503-8622-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/BIBM62325.2024.10822795

A Statistical Optimization Technique to Inform Statistical Resampling Assessments of Phylogenetic Reconstruction Reliability

1st Byungho Lee
Department of Computer Science and Engineering
Michigan State University
East Lansing, MI, USA
byungho@msu.edu

2nd Kevin J. Liu
Department of Computer Science and Engineering
Michigan State University
East Lansing, MI, USA
kjl@msu.edu

***Abstract*—Phylogenies represent the evolutionary history of a set of taxa and are typically reconstructed using computational analysis of genomic or other genetic sequence data. A common question is to ask whether a phylogenetic estimate is trustworthy. Felsenstein introduced the influential application of non-parametric statistical resampling to place confidence intervals on a reconstructed phylogenetic tree. Since then, this task – referred to as phylogenetic support estimation – has become a de facto requirement in systematics and wherever reconstructed phylogenies are reported. And algorithmic development efforts have continued to build upon Felsenstein's original method.**

In this study, we propose a novel statistical optimization technique to improve statistical resampling assessments of phylogenetic tree support. The technique uses a statistical criterion as a means to assess re-estimation difficulty of different resampled replicate datasets and more intelligently allocate computational effort, given the computationally difficult optimization problems that are addressed during re-estimation in this context. We couple the new statistical optimization technique with a recently introduced sequence-aware resampling method named RAWR, and we evaluate the performance of these methods using simulated and empirical benchmarking datasets. The new method is also used to conduct a case study of Darwin's finches. We find that the resulting phylogenetic support estimates offer comparable or often improved type I and II error compared to the original RAWR method, at the cost of additional computational runtime. Our study outcomes point the way to future algorithmic enhancements and better informed statistical resampling approaches.

***Index Terms*—phylogeny, phylogenetic, tree, multiple sequence alignment, support, confidence interval, resampling, maximum likelihood optimization**

I. Introduction

One of the most widely used resampling methods is bootstrap resampling, which samples an input set of observations uniformly at random with replacement [1]. Bootstrap resampling has found many applications throughout science and engineering. A particularly important application can be found in the field of phylogenetics, where bootstrap resampling is used to assess reliability of phylogenetic reconstruction [2]. This task – sometimes referred to as phylogenetic support estimation – is now routine wherever reconstructed phylogenies are reported, and the original publication describing the phylogenetic bootstrap method is the 41st most cited of all time [3].

As with other non-parametric resampling techniques, a primary advantage of bootstrap resampling is that it does not require a parametric model and attendant assumptions about model appropriateness for a particular dataset. Still, bootstrap resampling requires another important simplifying assumption: that the input observations are independent and identically distributed (i.i.d.). In the context of phylogenetic reconstruction using biomolecular sequence data, a host of evolutionary and other factors can violate this simplifying assumption [2]. These include biomolecular structure and function, genetic recombination, sequence insertion and deletion processes, and many others.

To move beyond the simplifying assumption of i.i.d. data, Wang et al. [4] recently introduced a new sequence-aware resampling technique named RAWR ("RAndom Walk Resampling"). The resampling method takes the form of a random walk conducted directly on an input set of unaligned sequence data. Sequence data are resampled asynchronously between walk reversals, where the latter make use of anchor positions as synchronization points (similar to the concept of barriers in parallel computing). A key property of RAWR is the "neighbor preservation" principle: neighboring site positions (e.g., neighboring bases in the case of DNA sequence data) in a resampled sequence are guaranteed to also appear as neighbors in the corresponding original sequence in the input. The neighbor preservation principle is key to preserving sequential ordering information in the input, which is necessary for critical downstream estimation tasks such as multiple sequence alignment. To date, RAWR has been successfully applied to place confidence intervals on phylogenetic trees [4] as well as other biomolecular sequence analysis tasks [5]. Performance studies have consistently demonstrated that RAWR offers comparable or often better type I and II error compared to bootstrap resampling and other state-of-the-art methods.

But there is a catch. All statistical resampling methods – RAWR included – require additional estimation to be performed on resampled replicates. The added computational

overhead can be onerous for estimation tasks that require computationally difficult optimization.

Part of the reason for this burden is due to the modularized design of statistical resampling applications. In part to mitigate resampling bias, resampling is completely decoupled from re-estimation on resampled replicates. As a consequence, resampling is oblivious to downstream re-estimation. And all replicates are homogeneous from the perspective of re-estimation: all replicates are treated similarly with the same re-estimation method and a priori allocation of computational effort is equal across replicates. We believe that an "uninformed" re-estimation strategy is in fact an opportunity for improved algorithmic design – one which may be exploited to obtain improvements in support estimation accuracy (and potentially other aspects of method performance).

II. Materials and Methods

Our study introduces a novel optimization technique for improving RAWR support estimation. The former method builds on the latter method – which itself requires an initial MSA and tree estimate as input. We now provide a detailed description of these methods.

A. Methods under study

Initial MSA and tree estimates for phylogenetic support annotation. To begin, an initial multiple sequence alignment and phylogenetic tree are estimated on the input set of unaligned sequences, and the latter serves as the "annotation" tree for phylogenetic support annotation purposes. Estimation is performed using a two-phase approach: in the first stage of two-phase analysis, a multiple sequence alignment (MSA) is estimated using the unaligned sequences as input; in the second stage, a phylogenetic tree is estimated using the estimated MSA as input.

Our study utilized MAFFT [6] – a widely used software package for MSA estimation. The software implements several different MSA estimation algorithms, some of which have been shown to offer competitive estimation accuracy compared to the state of the art [7], [8] and others which offer greater computational efficiency at the cost of estimation accuracy. The initial MSA was estimated using MAFFT version 7.490 with default settings, which corresponds to its FFT-NS-2 algorithm – one of the faster algorithms in the software suite which offers less accuracy compared to other implemented algorithms such as MAFFT's L-INS-i algorithm option [6].

Using the MAFFT-estimated MSA as input, a phylogenetic tree was reconstructed using maximum likelihood estimation (MLE) under the GTR+Γ model of nucleotide substitution [9]. We used RAxML version 8.2.12 to conduct phylogenetic MLE analyses [10]. The resulting phylogenetic tree estimate serves as the "annotation" tree for phylogenetic support estimation purposes.

RAWR support estimation. The RAWR method for phylogenetic support estimation provides both a baseline for comparison purposes and an initial resampling analysis that our new optimization technique builds upon. RAWR support estimation consists of multiple steps: (i) RAWR is used to perform sequence-aware resampling and obtain replicate sets of unaligned sequences, (ii) MSA and tree re-estimation is performed on each replicate set of unaligned sequences, and (iii) phylogenetic tree support for the annotation tree is calculated using re-estimated trees. We now recap details for each stage of analysis.

The input to RAWR analysis consists of the initial MSA and tree that were estimated using a two-phase method. The output consists of phylogenetic support values for each internal branch of the input tree – i.e., a real value between 0 and 1 where larger values reflect greater estimation reliability.

RAWR begins by conducting sequence-aware resampling to obtain replicate sets of unaligned sequence data. The resampling procedure takes the form of random walk conducted directly on the input MSA. The random walk randomly selects a starting site and then proceeds in a random direction. The direction is reversed with certainty at the first and last position, and with probability γ elsewhere; our study experiments utilized reversal probability $\gamma = 0$. Sites are resampled along the random walk path and the unaligned sequences corresponding to the sampled sites as read in random walk order constitute the replicate set of unaligned sequences; equivalently, resampled aligned sequences are unaligned by omitting indels. The walk concludes once a length criterion is satisfied, where the walk length equals the length of the input MSA.

MSA and tree re-estimation are performed on the replicate set of unaligned sequences. We utilized the same two-phase approach and methods as in the initial MSA and tree estimation procedures.

The resampling and re-estimation procedures are repeated to obtain resampling replication. Our study utilized 100 RAWR replicates for each input dataset analysis. Finally, phylogenetic support for each internal branch of the input annotation tree is calculated as the fraction of re-estimated trees that also display that branch.

Statistical optimization of RAWR support estimation. Our new optimization technique builds upon RAWR support estimation. Figure 1 provides an illustrated example and flowchart for the procedure. (Detailed pseudocode is provided in Algorithm 1 in the Supplementary Appendix.)

The optimization procedure proceeds from where RAWR resampling and re-estimation concludes. (See above for RAWR analysis procedures). An optimization criterion is used to identify RAWR replicates that pose greater challenges to phylogenetic reconstruction and may yield less reliable re-estimates, as these could benefit from more intensive computational analysis to improve MSA and tree reconstruction. We utilize model likelihood under the GTR+Γ substitution model for this purpose, where likelihood score is calculated using the initial MSA estimate as data. The model likelihood calculations were performed using RAxML version 8.2.11.

A data-driven threshold is then calculated based on the best score in the larger of two sets: (a) the low-scoring tail of the optimization score distribution across RAWR replicates, where the tail consists of all optimization scores that are 1

standard deviation below the mean (which is equivalent to a standard score/z-score of 1), and (b) the lowest decile of the optimization score distribution across RAWR replicates. An additional threshold adjustment is performed to ensure that a sufficient number of replicates are selected for re-analysis: if the number of replicates in the lowest decile of scores is less than or equal to the number of replicates that scored more than 1 standard deviation below the mean, the threshold is set to the score with rank that is the average of these two quantities.

The selected replicates are re-analyzed using more computationally intensive and accurate MSA and tree estimation methods (as compared to the original estimation as well as re-estimation on RAWR replicates). MAFFT's L-INS-i algorithm was used for this purpose, which offers improved MSA accuracy compared to the FFT-NS-2 algorithm at the cost of increased computational overhead. RAxML was then used to reconstruct a phylogenetic tree using the resulting MSA estimate as input; RAxML MLE was run using the same model and settings as elsewhere in our study.

The resulting "re-re-estimates" replace the previous re-estimates in corresponding RAWR replicates. The annotation tree support calculation is updated accordingly based upon the set of optimized tree re-estimates/re-re-estimates.

B. Performance study using simulated and empirical benchmarking data

Simulated benchmarking datasets. Our study utilizes simulation conditions and datasets from the previous study of Wang et al. [4]. The model conditions include a wide range of dataset sizes and evolutionary divergence for assessing performance of phylogenetic reconstruction and phylogenetic support estimation. Below we recap the procedures for simulating the synthetic benchmarking datasets.

The 10-taxon and 50-taxon simulation datasets were simulated using INDELible version 1.03 [11]. Non-ultrametric model trees are simulated under a random birth process with branch lengths drawn uniformly at random from the open unit interval. For the 100-taxon simulations, random birth-death model trees were sampled using r8s version 1.7 [12], and model trees were then deviated away from ultrametricity using Nakhleh et al.'s [13] approach with deviation factor $c = 2.0$. All model trees were rescaled based on a height parameter h. Then, nucleotide sequence evolution along each model tree was simulated under a finite-sites models of nucleotide substitutions coupled with a model of sequence insertions/deletions, where root sequence length was set to 1 kb. The nucleotide substitution model consisted of the general time-reversible (GTR) model [9], with base frequency and substitution rate parameters set to empirical NemATol estimates from the previous study of Liu et al. [8]. The 10-taxon and 50-taxon simulations of sequence evolution were performed with INDELible [11] and the indel model of Fletcher and Yang [11]. The 100-taxon simulations of sequence evolution were performed with ROSE [8] and utilized the indel model with medium gap length distribution from the study of Liu et al. [8]. Simulations were repeated for each model condition to obtain 20 replicate datasets. Model conditions and summary statistics for the simulated datasets are listed in Table I.

Empirical benchmarking datasets. Our study utilized empirical benchmarking datasets from the Comparative RNA Website (CRW) database [14]. The CRW rRNA datasets were comprehensively curated using biomolecular sequence data, structural information, and other heterogeneous data. As such, this resource provides a "gold standard" reference for benchmarking multiple sequence alignment [7], [8]. The manually curated MSA provided with each dataset serves as the reference MSA in our experiments. A reference tree was obtained using MLE analysis of the reference MSA; to this end, RAxML was used to perform maximum likelihood phylogenetic inference under the same finite-sites substitution model and the same software settings as in the simulation experiments.

The simulations in our simulation study reflect non-coding nucleotide sequence evolution and dataset sizes up to 100 taxa. We therefore focused on intronic rRNA datasets with at most 250 sequences for the purposes of experimental consistency. Sites with more than 99% missing data were omitted during data preprocessing. Summary statistics for the empirical benchmarking datasets are listed in Table II.

Performance criteria used in evaluations. Phylogenetic support estimation performance was evaluated based on both type I and type II error with respect to the model tree in the simulation experiments and the reference tree in the empirical CRW benchmarking experiments. Accuracy assessments in our study focus on precision-recall curves, receiver operating characteristic (ROC) curves, and area under these curves, as these facilitate simultaneous assessment of type I and type II error and tradeoffs between them.

Confusion matrices for the PR and ROC curves are formed from four classes of error. True positives (TP) are bipartitions of the annotation tree with support values greater than or equal to a given threshold and also appear in the reference tree. False positives (FP) are bipartitions of the annotation tree with support values greater than or equal to a given threshold but do not appear in the reference tree. False negatives (FN) are bipartitions of the annotation tree with support values less than a given threshold but appear in the reference tree. True negatives (TN) are bipartitions of the annotation tree with support values less than a given threshold and do not appear in the reference tree. The PR curve plots precision ($\frac{|\text{TP}|}{|\text{TP}|+|\text{FP}|}$) versus recall or true positive rate ($\frac{|\text{TP}|}{|\text{TP}|+|\text{FN}|}$). The ROC curve plots true positive rate ($\frac{|\text{TP}|}{|\text{TP}|+|\text{FN}|}$) versus false positive rate ($\frac{|\text{FP}|}{|\text{TN}|+|\text{FP}|}$). Custom scripts and the scikit-learn Python library [15] were used to calculate confusion matrices, curves, and AUC values.

Computational runtime and memory usage were also assessed for the phylogenetic support estimation methods under study. The former are reported as serial wall-clock runtimes.

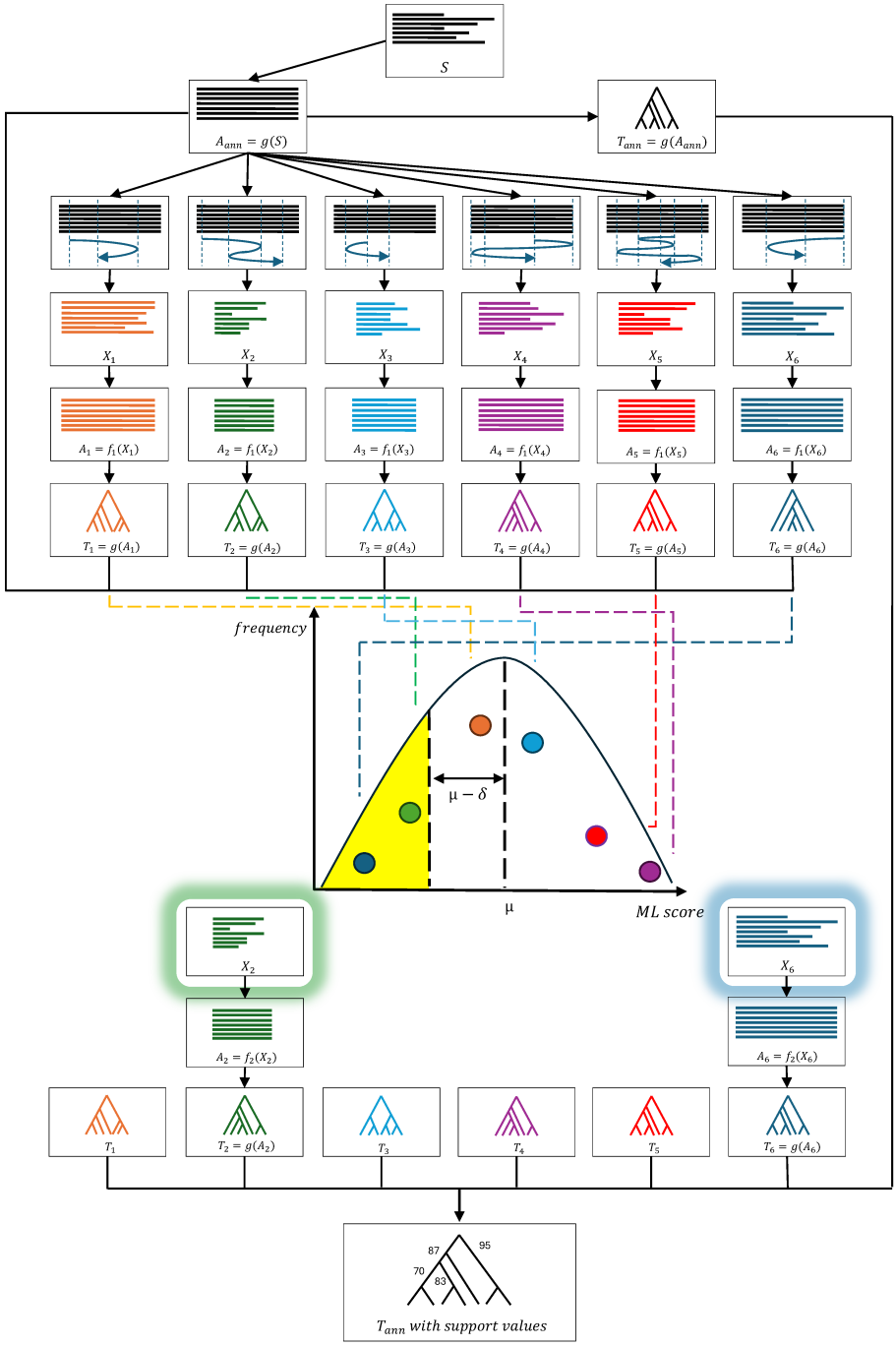


Fig. 1: *Illustrated example with flowchart of Optimized RAWR method for phylogenetic support estimation.* The Optimized RAWR method begins with RAWR resampling and re-estimation. Beginning with an input set of unaligned sequences (row 1), the input is aligned to obtain an initial multiple sequence alignment (MSA) and a phylogenetic tree is reconstructed using the initial MSA estimate as input (row 2). The latter serves as the "annotation" tree for phylogenetic support estimation purposes. Sequence-aware resampling of the initial MSA is then performed using RAWR: the MSA is resampled via random walk resampling and a replicate set of unaligned sequences is obtained (rows 3 and 4, respectively). Then, an MSA and phylogenetic tree are re-estimated on the replicate set of unaligned sequences (rows 5 and 6, respectively); sequence resampling and MSA/tree re-estimation are repeated, resulting in a set of re-estimated trees (row 6). The re-estimated trees are scored and ranked using the model likelihood criterion on the initial MSA estimate (row 7). The likelihood score serves as an indicator of re-estimation difficulty. Low-scoring replicates based on the optimization score distribution's low-scoring tail are re-analyzed using more computationally intensive approaches than in preceding steps, resulting in a set of re-re-estimated MSAs and trees for the low-scoring replicates (rows 8 and 9). The re-re-estimated trees replace previously re-estimated trees for the low-scoring replicates (row 10). Finally, phylogenetic support for each non-leaf branch of the annotation tree is calculated based on the proportion of re-estimated/ re-re-estimated trees that also display that branch (row 11).

TABLE I: *Simulation study: model condition parameters and summary statistics.* The following table is reproduced from [4]. Model condition parameters consists of number of taxa, model tree height, and insertion/deletion probability. The 10-taxon model conditions are named 10.A through 10.E in order of generally increasing evolutionary divergence, and the 50- and 100-taxon model conditions are named similarly. The following summary statistics are reported for each model condition ($n = 20$): "ANHD" is the average normalized Hamming distance of a pair of aligned sequences in an MSA, "Gappiness" is the proportion of an MSA matrix that consists of indels, "length" is the number of MSA columns, and "SP-FN" and "SP-FP" are the proportions of nucleotide-nucleotide homologies that appear in the true alignment but not in the estimated alignment or vice versa, respectively. The average normalized Robinson–Foulds distance ("nRF") between the model tree and the RAxML(MAFFT)-inferred tree is also reported for each model condition ($n = 20$).

Model condition	Number of taxa	Model tree height	Insertion/deletion Probability	True alignment			MAFFT alignment			RAxML (MAFFT) nRF
				ANHD	Gappiness	Length	Length	SP-FN	SP-FP	
10.A	10	0.47	0.13	0.380	0.591	2466	1543	0.566	0.629	0.186
10.B	10	0.7	0.1	0.479	0.618	2691	1602	0.687	0.750	0.243
10.C	10	1.2	0.06	0.591	0.645	2832	1588	0.811	0.850	0.443
10.D	10	2	0.031	0.642	0.591	2490	1583	0.815	0.841	0.464
10.E	10	4.4	0.013	0.696	0.578	2390	1623	0.904	0.913	0.664
50.A	50	0.45	0.06	0.415	0.607	3070	2053	0.340	0.336	0.084
50.B	50	0.73	0.03	0.513	0.605	2525	1834	0.451	0.431	0.146
50.C	50	1.2	0.02	0.598	0.620	2646	1950	0.731	0.704	0.322
50.D	50	2	0.012	0.667	0.671	2720	2171	0.902	0.881	0.517
50.E	50	4.3	0.005	0.715	0.574	2474	2385	0.974	0.965	0.755
100.A	100	4	1×10^{-5}	0.454	0.439	1682	1533	0.054	0.046	0.075
100.B	100	7	1×10^{-5}	0.479	0.439	2263	1861	0.209	0.176	0.119
100.C	100	15	5×10^{-5}	0.646	0.571	2317	2418	0.680	0.603	0.470
100.D	100	25	2×10^{-5}	0.683	0.614	2837	2799	0.899	0.853	0.607
100.E	100	20	4×10^{-5}	0.672	0.614	2487	2701	0.848	0.796	0.661

TABLE II: *Empirical study: summary statistics for intronic rRNA datasets.* The following table is reproduced from [4]. The empirical benchmarking datasets used in the study were obtained from the Comparative RNA Website (CRW) database [14]. The CRW database provides manually curated MSAs for each dataset, and these are used as reference alignments. A reference tree was then constructed using MLE analysis of the reference alignment. (See Methods section for details.) Summary statistics for each dataset are provided ($n = 1$), and the description of summary statistics are identical to Table I.

Dataset	Number of taxa	Reference alignment			MAFFT alignment			RAxML (MAFFT) nRF
		ANHD	Gappiness	Length	Length	SP-FN	SP-FP	
IGIA	110	0.606	0.915	10368	6065	0.732	0.780	0.645
IGIB	202	0.579	0.910	16233	7070	0.825	0.863	0.678
IGIC2	32	0.533	0.700	4243	3530	0.691	0.716	0.517
IGID	21	0.719	0.782	5061	3063	0.874	0.905	0.778
IGIE	249	0.451	0.838	2751	2847	0.406	0.389	0.585
IGIIA	174	0.668	0.814	6406	6945	0.817	0.800	0.450

C. Empirical study of Darwin's finches

Genomic sequence dataset. We re-analyzed genomic sequence data from Lamichanney et al.'s [16] study of Darwin's finches, which was also re-analyzed by Wang et al. [4]. The dataset consisted of genomic sequence data for 25 samples from different species in the clade and 34,972 loci in total. The genomic loci included annotated genes, intergenic regions, and scaffolds without gene annotations.

Initial MSA and tree estimation. The unaligned sequences for each locus were aligned using MAFFT with default settings to obtain an initial MSA estimate (as in the rest of the study). MSAs were concatenated across loci for subsequent phylogenetic tree reconstruction.

As in the rest of our study, maximum likelihood estimation (MLE) was used to reconstruct a phylogenetic tree; for the finch dataset analysis, MLE was performed on a concatenated and partitioned MSA. In terms of sequence length, the concatenated and partitioned MSA is multiple orders of magnitude larger than the simulated datasets. We therefore used ExaML version 3.0.22 [17] – a variant of RAxML that supports distributed-memory parallelism – to conduct highly parallelized phylogenetic MLE on a high-performance computing cluster. To facilitate the ExaML analysis, RAxML version 8.2.9 was used to perform maximum parsimony (MP) optimization and obtain an initial starting tree; ExaML version 3.0.22 was also used to specify additional analysis metadata in the form of a partition configuration file. The resulting tree estimate served as the annotation tree for the purposes of phylogenetic support estimation.

Phylogenetic support estimation using RAWR. For each locus, RAWR resampling was conducted on the initial lo-

cus MSA to obtain a replicate set of unaligned sequences. The replicate set of unaligned sequences was aligned using MAFFT with default settings to obtain a re-estimated MSA for each locus. Re-estimated MSAs were then concatenated across loci to obtain a concatenated and partitioned MSA. The latter MSA was used to perform phylogenetic tree re-estimation using the same approach as in the initial tree estimation step. The process was repeated to obtain a set of 100 re-estimated trees, which were then used to calculate phylogenetic support for the annotation tree.

Phylogenetic support estimation using Optimized RAWR. As in the RAWR analysis of Darwin's finches, Optimized RAWR was adapted to perform multi-locus data analysis. The procedure begins with RAWR analysis using the steps described above. All replicates are scored under the optimization criterion where model likelihood under the finite-sites substitution model is calculated using the concatenated and partitioned MSA as data. Low-scoring replicates are selected for re-analysis using more intensive MSA and tree estimation methods: MAFFT L-INS-i was used to re-re-estimate MSAs for each locus, re-re-estimated MSAs are concatenated across loci, and ExaML was used to perform parallelized phylogenetic MLE on the concatenated and partitioned MSA. The resulting tree replaces the previously re-estimated tree for the corresponding replicate. Phylogenetic support for the annotation tree is calculated using the updated set of re-estimated/re-re-estimated trees, following the procedure used throughout our study.

III. Results

A. Performance study using simulated and empirical benchmarking data

Performance benchmarking using simulated datasets. Table III compares type I and type II error of RAWR versus Optimized RAWR. On the 10-taxon model conditions, Optimized RAWR returned PR-AUC improvements over the baseline RAWR method in all cases. The smallest improvement was seen on the least divergent 10.A model condition. As model conditions increased in evolutionary divergence and MSA and tree reconstruction became more challenging, Optimized RAWR's PR-AUC and ROC-AUC advantage grew to as much as 7.9% and 7.2%, respectively, on the most divergent 10-E model condition. Average PR-AUC and ROC-AUC improvement of Optimized RAWR over RAWR on the 10-taxon model conditions amounted to 3.0% and 2.6%, respectively. These differences were not statistically significant based on a one-tailed pairwise t test, with the exception of both AUC comparisons on the most divergent 10.E model condition and the ROC-AUC comparison on the 10.C model condition.

Similar outcomes were observed on the 50- and 100-taxon model conditions, with one difference: AUC improvement of Optimized RAWR over baseline RAWR was statistically significant for all model conditions other than the ROC-AUC comparison on the 50.A model condition and both AUC comparisons on the 100.A model condition – both of which have lowest evolutionary divergence among the 50- and 100-taxon model conditions, respectively. PR-AUC and ROC-AUC improvement of Optimized RAWR over baseline RAWR averaged around 1% and 5%, respectively. The largest AUC improvements were generally seen on the most divergent model conditions.

Our study also compared computational runtime and peak memory usage of the phylogenetic support estimation methods under study. Runtimes are reported in Figure 2. Baseline RAWR and Optimized RAWR runtimes amounted to under 10 minutes across the 10-taxon datasets, on average. As expected due to the computational difficulty of the MSA and tree estimation problems under study, runtimes on 50-taxon datasets grew to a few hours; the largest runtimes in our study were observed on the 100-taxon datasets – amounting to as much as a half a day or so per dataset. Runtimes tended to increase as evolutionary divergence increased from the 50.A to 50.E model conditions, and similarly on the 100-taxon model conditions.

As Optimized RAWR includes baseline RAWR analysis as an initial step, the former naturally requires more runtime than the latter. By comparing average runtime of the former versus the latter, we found that runtime overhead of our new optimization technique was a relatively small fraction of the overall runtime required for baseline RAWR analysis; the outcome was consistent across all model conditions in our simulation study. We note that this occurred despite Optimized RAWR's use of a slower and more accurate two-phase method for re-re-estimation, as compared to RAWR re-estimation.

Peak memory usage was modest and amounted to at most a few hundred MiB in all cases. We note that this amount of main memory is well within the scope of commonly available personal computers and high-performance computing facilities.

Performance benchmarking using empirical CRW datasets. Table IV reports PR-AUC and ROC-AUC results for the phylogenetic estimation methods under study. AUC improvement of Optimized RAWR over baseline RAWR was largest on the IGIB and IGID datasets, with respective PR-AUC improvements of 1.4% and 1.4% and respective ROC-AUC improvements of 1.6% and 1.8%. We note that IGIB is largest dataset by sequence length and second largest dataset by number of taxa, and IGID has highest sequence divergence based on reference MSA ANHD. Smaller AUC differences were observed on the other CRW datasets.

B. Empirical study of Darwin's finches

Figure 3 visualizes the annotation tree that was estimated using the initial two-phase analysis, along with phylogenetic support values that were estimated using either the baseline RAWR method or Optimized RAWR method. The phylogenetic support values returned by the two methods were quite similar, with a maximum difference of at most 10%. There was a directionality trend: Optimized RAWR returned comparable (within 1-2%) or higher phylogenetic support on all branches, as compared to baseline RAWR. Furthermore, the higher support values returned by Optimized RAWR (as

TABLE III: *Simulation study: PR-AUC and ROC-AUC comparisons of baseline RAWR versus Optimized RAWR methods for phylogenetic support estimation.* Average PR-AUC and ROC-AUC values are reported for each method on each model condition ($n = 20$). For each model condition and comparison of either PR-AUC or ROC-AUC values, a one-tailed pairwise t test was used to evaluate statistical significance of AUC value differences between the two methods and a p-values is reported ($n = 20$).

Model condition	PR-AUC		Pairwise t-test	ROC-AUC		Pairwise t-test
	RAWR	Optimized RAWR		RAWR	Optimized RAWR	
10.A	0.9949	0.9988	0.1649	0.9833	0.9833	0.5000
10.B	0.9830	0.9917	0.2109	0.9483	0.9548	0.1678
10.C	0.9197	0.9523	0.0594	0.9020	0.9434	0.0157
10.D	0.8999	0.9270	0.1219	0.9683	0.9808	0.0857
10.E	0.8919	0.9709	0.0429	0.8767	0.9483	0.0038
50.A	0.9981	0.9985	0.0079	0.9654	0.9756	0.0727
50.B	0.9955	0.9969	0.0013	0.9482	0.9620	0.0010
50.C	0.9797	0.9854	3.4e-07	0.8712	0.9762	7.0e-07
50.D	0.9710	0.9854	2.3e-05	0.9810	0.9914	2.4e-07
50.E	0.9429	0.9653	5.8e-07	0.9138	0.9834	6.8e-08
100.A	0.9930	0.9926	0.3698	0.8994	0.8994	0.4894
100.B	0.9899	0.9912	6.5e-06	0.9100	0.9215	2.5e-05
100.C	0.9621	0.9734	1.2e-08	0.7896	0.8560	2.0e-08
100.D	0.9409	0.9664	2.7e-08	0.7787	0.8855	2.8e-11
100.E	0.9445	0.9655	3.2e-08	0.7509	0.8579	2.5e-10

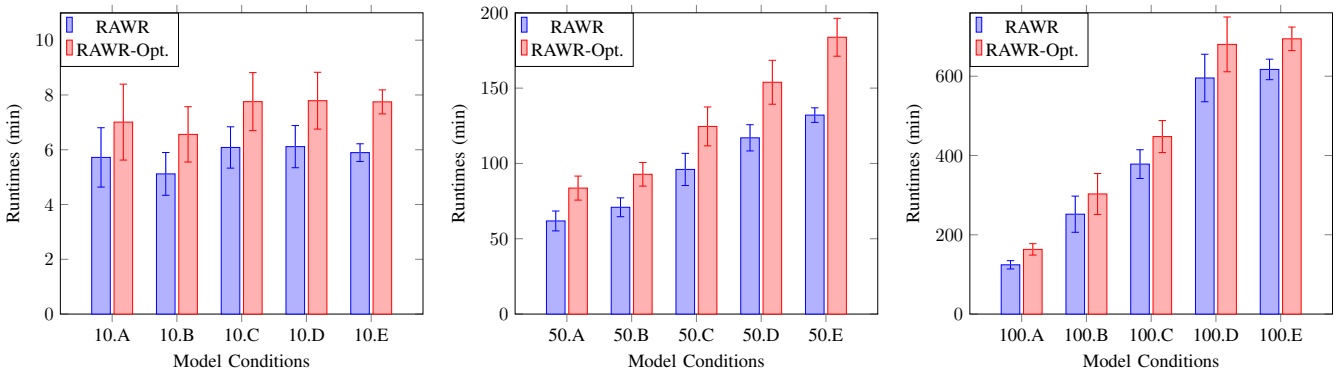


Fig. 2: *Simulation study: runtime comparison of phylogenetic support estimation methods.* Average runtime of baseline RAWR ("RAWR") and Optimized RAWR ("RAWR-Opt.") are reported for each model condition ($n = 20$). Standard error bars are shown.

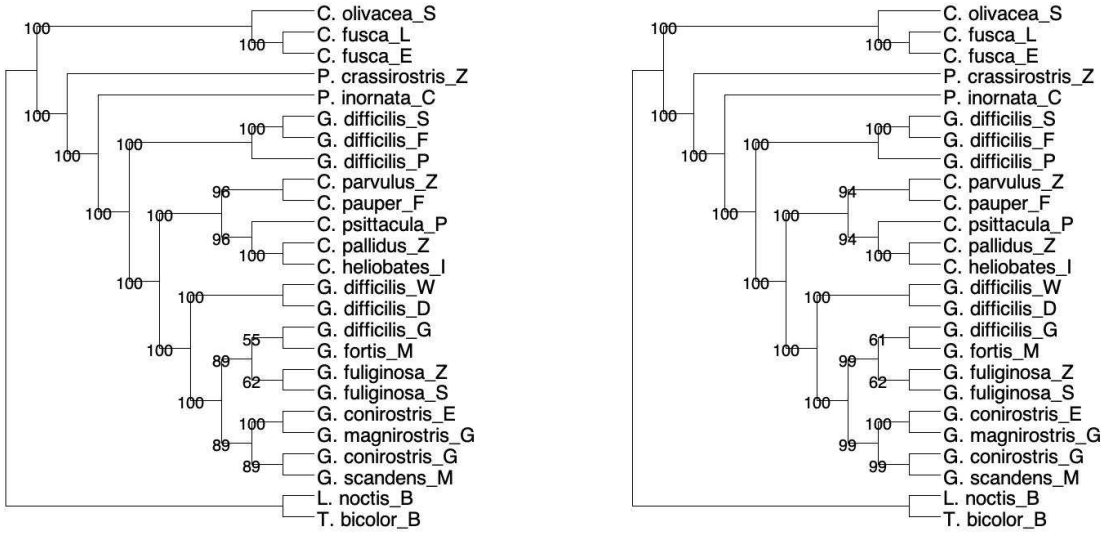
TABLE IV: *PR-AUC and ROC-AUC returned by phylogenetic support estimation methods on empirical CRW datasets.* AUC values are reported for baseline RAWR ("RAWR") and Optimized RAWR ("Optimized") on each CRW dataset ($n = 1$).

Model condition	PR-AUC		ROC-AUC	
	RAWR	Optimized	RAWR	Optimized
IGIA	0.8800	0.8825	0.9356	0.9401
IGIB	0.7999	0.8135	0.8939	0.9099
IGIC2	0.8773	0.8715	0.8691	0.8548
IGID	0.8524	0.8661	0.9107	0.9286
IGIE	0.8056	0.8089	0.8741	0.8766
IGIIA	0.8996	0.9043	0.9000	0.9077

compared to baseline RAWR) largely appeared within the sharp-beaked ground finch clade, which also included some other ground finch species.

IV. Discussion

Type I and type II error improvement of optimized RAWR support estimates versus baseline RAWR support tended to grow as two key experimental factors were varied: (1) dataset sizes increased from 10 to 100 taxa, and (2) evolutionary divergence increased. Both support estimation methods perform re-estimation on RAWR replicates by addressing computationally difficult optimization problems [19], [20], and the first experimental factor directly determines combinatorial growth of the optimization space. Theoretical and experimental work has also confirmed the second factor's primary contribution to MSA and phylogenetic tree estimation [7], [8], [21]. The optimized RAWR method's PR-AUC and ROC-AUC improvements over non-optimized RAWR amounted to nearly 10% on the largest and most divergent (and therefore most challenging) model conditions. This magnitude of improvement approaches that observed for RAWR in comparison to state-of-the-art phylogenetic support estimation methods [4].



(a) RAWR

(b) Optimized RAWR

Fig. 3: *Empirical study of Darwin's finches: comparison of phylogenetic support estimates.* The annotation tree that was initially estimated by the two-phase method is visualized along with either (a) baseline RAWR support values or (b) Optimized RAWR support values. We note that the annotation tree topology reported in our study differs from the estimate reported by Wang et al. [4], but is identical to the estimate reported by Lamichhaney et al. [16]. Dendroscope [18] was used to visualize phylogenetic trees and phylogenetic support values.

A similar outcome was observed in the empirical study experiments. Type I and II error improvements returned by RAWR optimization versus baseline were greater on larger and more divergence intronic CRW datasets. Phylogenetic support returned by the former method was comparable or slightly higher than the latter on the Darwin's finch genomic sequence dataset. We note that traditional phylogenetic analysis of genomic sequence data can aggregate over disparate phylogenetic signal across different loci [22], and the sheer amount of sequence data in genomes can benefit statistical reproducibility.

As is typical for any statistical resampling analysis, RAWR optimization's type I and II error improvements come at a cost: namely, added computational runtime. Despite Optimized RAWR's use of a more intensive MSA method compared to baseline non-optimized RAWR analysis, the runtime overhead contributed by the former was relatively small compared to the runtime requirements of the latter. Peak memory usage of optimized and baseline RAWR methods were similar, and main memory requirements did not prove to be a scalability bottleneck for the model conditions under study.

We attribute type I and II error improvements to the model likelihood optimization criterion and its ability to identify resampling replicates that may benefit from more intensive re-estimation. We note that the use of a model likelihood criterion under a finite-sites substitution-only model for *unaligned* sequence data is unconventional. Prior theoretical and experimental work has provided some insights into its application in the study context of MSA and tree reconstruction using unaligned sequence data [8]. While atypical, the use of the optimization approach in this study has paid similar dividends for the problem of simultaneous MSA and tree estimation using unaligned sequence data [7], [8]. As in the earlier studies, we conjecture that the unconventional but effective use of phylogenetic MLE under finite-sites substitution-only models relies on the constrained set of sequence position homologies that are explored when re-estimating MSAs using accurate MSA estimation methods. On the other hand, this study utilizes sequence-aware statistical resampling to obtain resampled sequences that are then re-aligned, unlike the earlier studies that performed re-alignment on subsets of the original input sequences using a phylogenetic divide-and-conquer approach.

Our proposed approach improved phylogenetic support estimation accuracy despite the use of a simple threshold-based rule for selecting RAWR-resampled replicates to perform more intensive additional optimization. This aspect of our method presents a promising algorithmic design opportunity. We conjecture that more sophisticated techniques may return further improvements (see future research directions below).

V. Conclusions

In this study, we introduce a new optimization approach for improving sequence-aware resampling analysis of phylogenetic tree support. The optimization technique is used to identify resampled replicates that may benefit from more intensive optimization during re-estimation. Using simulated and empirical benchmarking datasets, we validate the performance

of the resulting phylogenetic tree support estimates and we find that the estimates show favorable type I and type II error improvements compared to RAWR, a state-of-the-art method. We also demonstrate the utility of the new method in a case study of Darwin's finches.

We conclude with thoughts on future research directions. As mentioned above, both resampling and re-estimation may benefit from informed techniques that draw on interdependence of these two tasks. Also, the statistical optimization approach in our study may benefit from the use of more complex models of sequence evolution, such as the combination of finite-sites substitution models [9] with models of sequence insertion and deletion events [23]–[25]. However, joint modeling will further exacerbate computational scalability challenges. The latter is a fundamental topic in computational biology and bioinformatics, and a host of scalability-enhancing techniques have been explored. These include distributed and coordinated optimization across parallelized re-estimations, divide-and-conquer algorithms [7], [8], [26], and statistical approximation techniques.

Acknowledgment

This work has been supported by the NSF (CCF-1565719, CCF-1714417, DEB-1737898, and IOS-1740874 to KJL). and MSU faculty startup funds (to KJL). Computational experiments and analyses were performed on the MSU High Performance Computing Center, which is part of the MSU Institute for Cyber-Enabled Research.

Conflict of Interest: none declared.

References

[1] B. Efron, "Bootstrap methods: Another look at the jackknife," *Ann. Statist.*, vol. 7, no. 1, pp. 1–26, 01 1979.

[2] J. Felsenstein, "Confidence limits on phylogenies: an approach using the bootstrap," *Evolution*, vol. 39, no. 4, pp. 783–791, 1985.

[3] R. Van Noorden, B. Maher, and R. Nuzzo, "The top 100 papers," *Nature News*, vol. 514, no. 7524, p. 550, 2014.

[4] W. Wang, A. Hejasebazzi, J. Zheng, and K. J. Liu, "Build a better bootstrap and the RAWR shall beat a random path to your door: Phylogenetic support estimation revisited," *Bioinformatics*, vol. 37, no. Supplement_1, pp. i111–i119, 2021.

[5] R. Doko and K. Liu, "Reconstructing phylogenies using branch-variable substitution models and unaligned biomolecular sequences: A performance study and new resampling method," in *Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2023, pp. 1–10.

[6] K. Katoh and D. M. Standley, "MAFFT multiple sequence alignment software version 7: improvements in performance and usability," *Molecular Biology and Evolution*, vol. 30, no. 4, pp. 772–780, 2013.

[7] K. Liu, S. Raghavan, S. Nelesen, C. R. Linder, and T. Warnow, "Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees," *Science*, vol. 324, no. 5934, pp. 1561–1564, 2009.

[8] K. Liu, T. J. Warnow, M. T. Holder, S. M. Nelesen, J. Yu, A. P. Stamatakis, and C. R. Linder, "SATé-II: Very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees," *Systematic Biology*, vol. 61, no. 1, pp. 90–106, 2012.

[9] F. Rodriguez, J. Oliver, A. Marin, and J. Medina, "The general stochastic model of nucleotide substitution," *Journal of Theoretical Biology*, vol. 142, pp. 485–501, 1990.

[10] A. Stamatakis, "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies," *Bioinformatics*, vol. 30, no. 9, pp. 1312–1313, 2014.

[11] W. Fletcher and Z. Yang, "INDELible: a flexible simulator of biological sequence evolution," *Molecular Biology and Evolution*, vol. 26, no. 8, pp. 1879–1888, 2009.

[12] M. J. Sanderson, "r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock," *Bioinformatics*, vol. 19, no. 2, pp. 301–302, 2003.

[13] L. Nakhleh, B. M. E. Moret, U. Roshan, K. S. John, J. Sun, and T. Warnow, "The accuracy of fast phylogenetic methods for large datasets," *Proceedings of the 7th Pacific Symposium on BioComputing (PSB02)*, pp. 211–222, 2002.

[14] J. Cannone, S. Subramanian, M. Schnare, J. Collett, L. D'Souza, Y. Du, B. Feng, N. Lin, L. Madabusi, K. Muller, N. Pande, Z. Shang, N. Yu, and R. Gutell, "The Comparative RNA Web (CRW) Site: An Online Database of Comparative Sequence and Structure Information for Ribosomal, Intron and Other RNAs." *BMC Bioinformatics*, vol. 3, no. 15, 2002, http://www.rna.ccbb.utexas.edu.

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[16] S. Lamichhaney, J. Berglund, M. S. Almén, K. Maqbool, M. Grabherr, A. Martinez-Barrio, M. Promerová, C.-J. Rubin, C. Wang, N. Zamani, B. R. Grant, P. R. Grant, M. T. Webster, and L. Andersson, "Evolution of Darwin's finches and their beaks revealed by genome sequencing," *Nature*, vol. 518, no. 7539, pp. 371–375, Feb 2015.

[17] A. M. Kozlov, A. J. Aberer, and A. Stamatakis, "ExaML version 3: a tool for phylogenomic analyses on supercomputers," *Bioinformatics*, vol. 31, no. 15, pp. 2577–2579, 2015.

[18] D. H. Huson and C. Scornavacca, "Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks," *Systematic Biology*, vol. 61, no. 6, p. 1061, 2012.

[19] L. Wang and T. Jiang, "On the complexity of multiple sequence alignment," *Journal of Computational Biology*, vol. 1, no. 4, pp. 337–348, 1994, pMID: 8790475.

[20] S. Roch, "A short proof that phylogenetic tree reconstruction by maximum likelihood is hard," *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, vol. 3, no. 1, pp. 92–, Jan. 2006.

[21] P. Lockhart and M. Steel, "A tale of two processes," *Systematic Biology*, vol. 54, no. 6, pp. 948–951, 2005.

[22] S. V. Edwards, "Is a new and general theory of molecular systematics emerging?" *Evolution*, vol. 63, no. 1, pp. 1–19, 2009.

[23] J. L. Thorne, H. Kishino, and J. Felsenstein, "An evolutionary model for maximum likelihood alignment of DNA sequences," *Journal of Molecular Evolution*, vol. 33, no. 2, pp. 114–124, 1991.

[24] ——, "Inching toward reality: an improved likelihood model of sequence evolution," *Journal of Molecular Evolution*, vol. 34, no. 1, pp. 3–16, 1992.

[25] ——, "Erratum, an evolutionary model for maximum likeliihood alignment of DNA sequences," *Journal of Molecular Evolution*, vol. 34, no. 2, pp. 91–92, 1992.

[26] S. Mirarab, N. Nguyen, S. Guo, L.-S. Wang, J. Kim, and T. Warnow, "PASTA: ultra-large multiple sequence alignment for nucleotide and amino-acid sequences," *Journal of Computational Biology*, vol. 22, no. 5, pp. 377–386, 2015.

Additional Files

Additional file 1 — Appendix with Supplementary Material

Supplementary Material - A Statistical Optimization Technique to Inform Statistical Resampling Assessments of Phylogenetic Reconstruction Reliability

1st Byungho Lee
Department of Computer Science and Engineering
Michigan State University
East Lansing, MI, USA
byungho@msu.edu

2nd Kevin J. Liu
Department of Computer Science and Engineering
Michigan State University
East Lansing, MI, USA
kjl@msu.edu

I. Supplementary Methods

Darwin's finches dataset.

The study re-analyzed genomic sequence data originally studied by Lamichhaney et al. (2015) using datasets used by Wang et al. (2021). Original Illumina HiSeq2000 paired-end read data was downloaded from the NCBI SRA database (accession number PRJNA263122 at http://www.ncbi.nlm.nih.gov/sra). For each of 25 different species, one sample was randomly selected (accession numbers SRR1607296, SRR1607504, SRR1607439, SRR1607359, SRR1607385, SRR1607440, SRR1607547, SRR1607403, SRR1607458, SRR1607472, SRR1607551, SRR1607494, SRR1607399, SRR1607462, SRR1607343, SRR1607534, SRR1607406, SRR1607485, SRR1607508, SRR1607543, SRR1607365, SRR1607420, SRR1607466, SRR1607529, and SRR1607480).

Software commands used in performance study involving simulated and empirical benchmarking data.

The following commands was used to perform MSA estimation/re-estimation using MAFFT version 7.490:

```
mafft <unaligned sequence file>
> <estimated aligned sequence file>

mafft --maxiterate 4
--localpair <unaligned sequence file>
> <estimated aligned sequence file>
```

RAxML analyses were performed with RAxML version 8.2.12 using the following command:

```
raxmlHPC -f a -s <estimated alignment
file> -n <output file suffix/name> -m
GTRGAMMA -p <random number> -#2

raxmlHPC -f a -s <estimated alignment
file> -n <output file suffix/name> -m
GTRCAT -V -p <random number> -#2
```

The 100-taxon model trees in the study were sampled using r8s version 1.7 and the following script:

```
begin r8s;
simulate diversemodel=bdback seed=<random
see>
nreps=20 ntaxa=<10 or 50> T=0;
describe tree=0 plot=chrono_description;
end;
```

Software commands used in experiments on Darwin's finches dataset.

Multiple sequence alignments of multi-locus were converted from FASTA format into a binary format, required by ExaML to perform MLE analyses. ExaML version 3.0.22 was used:

```
parse-examl -s <estimated alignment file>
-m DNA -q <partition file> -n <output file
name/sample number>
```

ExaML MLE analyses requires a starting tree and our study used the following commands and RAxML version 8.2.12:

```
raxmlHPC-AVX -y -s <estimated alignment
file> -m GTRGAMMA -n <output file
suffix/replicate ID> -p <random seed> -q
<partition file>
```

ExaML MLE analyses were performed using the binary file and the starting tree using the following command:

```
examl-AVX -t <start tree file> -m GAMMA
```

```
-s <binary file> -n <output file suffix>
```

Software commands used in both simulated and empirical study results analysis.

Maxmium-likelihood analyses requires initially estimated multiple sequence alignments and re-estimated trees and our study used the following commands and RAxML version 8.2.12:

```
raxmlHPC -f N -z <re-estimated trees file>
-s <estimated alignment file> -m GTRGAMMA
-n <output file name>
```

```
raxmlHPC -f N -z <re-estimated trees file>
-s <estimated alignment file> -m GTRCAT -V
-n <output file name>
```

Support value annotated trees require an originally inferred tree and re-estimated trees and our study used the following commands and RAxML version 8.2.12:

```
raxmlHPC -f b -m GTRGAMMA -t <originally
inferred tree> -z <re-estimated trees
file> -n <outputfile suffix/name ID>
```

```
raxmlHPC -f b -m GTRCAT -V -t <originally
inferred tree> -z <re-estimated trees
file> -n <outputfile suffix/name ID>
```

Algorithm 1
Optimized RAWR phylogenetic support estimation

procedure optimizedRAWRsupport(A, T, $f_1()$, $f_2()$ $g()$, γ, k_r)
▷ Input: MSA A,
phylogenetic tree T,
MSA method $f()$
▷ $f_1()$: progressive method
▷ $f_2()$: iterative refinement method,
phylogenetic tree estimation method $g()$,
reversal probability γ,
number of replicates k_r,
▷ Output: phylogenetic support estimates ϵ
re-estimates = <>
maximum-likelihood-scores = $< (score, index) >$
for $i = 1$ **to** k_r **do**
X_i = resampleRAWRreplicate(A, γ)
AddItem(re-estimates <>, $g(f_1(X_i))$)
Y_i = maximum-likelihood-score($g(f_1(X_i))$, A)
AddItem(maximum-likelihood-scores <>, (Y_i, i))
end for
count-threshold = |maximum-likelihood-scores| $*$ 0.1
score-threshold = $mean$(maximum-likelihood-scores)
$-$ $standarddeviation$(maximum-likelihood-scores)
low-ml-scores-indices = <>
sort(maximum-likelihood-scores <>)
for ($score$, $index$) **in** maximum-likelihood-scores **do**
if ($score \leq$ score-threshold) ||
(|low-ml-scores-indices| $\leq$ count-threshold) **then**
AddItem(low-ml-scores-indices, $index$)
end if
end for
if count-threshold $\leq$ |low-ml-scores-indices| **then**
count-threshold = $mean$(count-threshold,
|low-ml-scores-indices|)
end if
for i **in** low-ml-scores-indices<> **do**
if Index(i) $\leq$ count-threshold **then**
re-estimates[i] = $g(f_2(X_i))$
end if
end for
for all non-leaf edge e in T **do do**
$\epsilon(e)$ = proportion of T_i in re-estimates that display
bipartition corresponding to e
end for
return(ϵ)
end procedure

procedure resampleRAWRreplicate(A, γ)
    $Z = <>$
    select $i \in [1, |A|]$ and walkDirection uniformly at random
    while !converged(Z, A) **do**
        $Z = Z \cup A[i]$
        ▷ add ith column of A to Z
        if reversal(γ) || ($i ==$ 1 && walkDirection is left) || ($i == |A|$ && walkDirection is right) **then**
            reverse(walkDirection)
        end if
        i = next column index after i in walkDirection order
    end while
    return unalign(Z)
    ▷ unalign(Z) drops indels from Z
    ▷ resampleRAWRreplicate() is excerpted from Wang et al. (2021).
end procedure

procedure converged(Z, A)
    return (length(Z) $\geq$ length(A))
    ▷ Sequence-length-based convergence criterion requires
        ▷ number of resampled sites $\geq$ input MSA length
    ▷ converged() is excerpted from Wang et al. (2021).
end procedure

II. Supplementary Results

Supplementary Table I: **Simulation study: regression analysis (I) between maximum-likelihood scores and topological errors (i.e., normalized Robinson-Foulds distance).** The following summary statistics are reported on each model condition (n=20): "r" is the correlation coefficient and "R^2" is the coefficient of determination.

Model condition	r		R^2	
	RAWR	**Optimized**	**RAWR**	**Optimized**
10.A	-0.7814	-0.7589	0.6406	0.6049
10.B	-0.7620	-0.7232	0.6074	0.5716
10.C	-0.7756	-0.7810	0.6170	0.6373
10.D	-0.6472	-0.6709	0.4377	0.4792
10.E	-0.2406	-0.2646	0.0960	0.1124
50.A	-0.6951	-0.6945	0.4951	0.4359
50.B	-0.6391	-0.6970	0.4894	0.4992
50.C	-0.6878	-0.8214	0.4858	0.6813
50.D	-0.6269	-0.8668	0.4005	0.7540
50.E	-0.2885	-0.6802	0.1167	0.4757
100.A	-0.7608	-0.7285	0.5996	0.5622
100.B	-0.7476	-0.6985	0.5744	0.5064
100.C	-0.7221	-0.9290	0.5392	0.8636
100.D	-0.6704	-0.9557	0.4683	0.9147
100.E	-0.6540	-0.9720	0.4480	0.9450

Supplementary Table II: **Simulation study: regression analysis (II) between maximum-likelihood scores and topological errors (i.e., normalized Robinson-Foulds distance).**

Model condition	$slope$		$intercept$	
	RAWR	**Optimized**	**RAWR**	**Optimized**
10.A	-2.2278	-2.0685	1.062	1.0786
10.B	-2.6521	-2.604	1.215	1.1865
10.C	-2.0992	-2.1367	1.5314	1.401
10.D	-1.6629	-1.6668	1.5098	1.4505
10.E	-0.6472	-0.6472	1.2404	1.227
50.A	-4.5549	-4.8964	1.1903	1.1416
50.B	-3.0609	-3.7642	1.2288	1.2146
50.C	-2.9225	-2.8127	1.5192	1.3367
50.D	-3.1882	-2.3608	2.3581	1.6787
50.E	-2.1848	-2.4713	2.3299	2.4324
100.A	-5.4115	-1.6959	4.1572	1.3191
100.B	-5.0376	-1.9041	3.7413	1.393
100.C	-3.4486	-1.9763	2.3263	1.2855
100.D	-3.6992	-4.6291	1.2512	1.2111
100.E	-8.7102	-8.5497	1.1585	1.1321