2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3503-8622-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/BIBM62325.2024.10821822

Estimating confidence intervals on reconstructed cophylogenies using bootstrap resampling

1st Julia Zheng
Computer Science and Engineering
Michigan State University
East Lansing, USA
0009-0003-5177-3345

2nd Yuya Nishida
Integrative Biology
Michigan State University
East Lansing, USA
nishiday@msu.edu

3rd Elizabeth A.C. Heath-Heckman
Integrative Biology
Microbiology, Genetics, & Immunology
Michigan State University
East Lansing, USA
each@msu.edu

4th Kevin J. Liu
Computer Science and Engineering
Michigan State University
East Lansing, USA
kjl@msu.edu

***Abstract*—Cophylogenies represent coevolutionary histories of two or more sets of coevolved taxa, and are used to study coevolution and other fundamental evolutionary processes. As with traditional phylogenies, cophylogenies are primarily reconstructed using computational analysis of DNA and other biomolecular sequence data. An essential question concerns the reliability of reconstructed phylogenies and cophylogenies.**

Statistical resampling offers a principled approach to evaluate statistical confidence for these tasks. We therefore apply bootstrap resampling – one of the most widely used non-parametric resampling techniques – to place confidence intervals on a reconstructed cophylogeny, which is the first such method to our knowledge. We validate the performance of the resulting reliability estimates in a simulation study as well as an empirical case study of bobtail squid and its bioluminescent endosymbionts. The utility of statistical resampling to assess cophylogenetic reconstruction reliability in an automated and data-driven manner points the way forward – both for future methods development and wider adoption in studies of symbioses and other forms of coevolution.

***Index Terms*—cophylogeny, cophylogenetic, bootstrap, resampling, support, confidence intervals, simulation study, squid, symbiosis**

I. Introduction

Bootstrap resampling was first introduced by Efron in 1982 [1] as a non-parametric resampling technique for confidence interval estimation and other statistical tasks. Given an input set of observed data that is assumed to be independent and identically distributed (i.i.d), the basic technique resamples input observations uniformly at random with replacement to obtain a bootstrap replicate. Over the years, bootstrap resampling has become a key fixture in many fields. A particularly important example can be found in phylogenetics: bootstrap resampling is used to calculate confidence intervals on reconstructed phylogenetic trees, as proposed by Felsenstein in 1985 [2]. Given an MSA sequenced from a set of related taxa, we can estimate a phylogeny and then assess bootstrap support for the phylogenetic estimate, based on the following procedure. Bootstrap resampling is applied to a multiple sequence alignment (MSA) A under the i.i.d. assumption to obtain bootstrap replicates $\{A_1^*, A_2^*, \ldots, A_m^*\}$. Phylogenetic estimation is then performed on each bootstrap replicate to obtain an estimate T^* of the phylogeny. For every branch of an annotation phylogeny T estimated on the original MSA A, the bootstrap estimates of the phylogeny $\{T_1^*, T_2^*, \ldots, T_m^*\}$ are used to calculate the fraction of bootstrap estimated phylogenies that also display the same branch.

While bootstrap resampling is a de facto standard for assessing statistical reliability of reconstructed phylogenies, it has yet to be directly applied to the field of cophylogenetics. Cophylogenetics is the study of co-evolutionary histories between coevolved sets of taxa. As in traditional phylogenetics, cophylogenies are typically reconstructed using computational analysis of biomolecular sequence data.

Prior experimental work indicates that different sources of error can strongly impact cophylogenetic reconstruction, including multiple sequence alignment error and tree estimation error [3]. There is therefore a great need to assess cophylogenetic reconstruction reliability in an automated and data-driven fashion. We therefore propose an extension of Felsenstein's phylogenetic bootstrap method to assess cophylogenetic reconstruction reliability. To the best of our knowledge, the resulting method is the first to address this important problem.

II. Methods

We begin with the mathematical background that is necessary to describe the new cophylogenetic support estimation method and experimental procedures. Some definitions and notations were reproduced from [3].

A rooted phylogenetic tree $T_\chi = (V_\chi, E_\chi)$ consists of a set of vertices V_χ and a set of directed edges E_χ that represent the evolutionary history of a set of related taxa χ. Note that many cophylogenetic reconciliation methods require rooted binary phylogenetic trees as input. The rooted binary tree T_χ

979-8-3503-8622-6/24/$31.00 ©2024 IEEE

has a root ρ with in-degree zero and out-degree two, leaves $\mathcal{L}_\chi \subseteq V_\chi$ such that each leaf has out-degree zero and in-degree one and each leaf corresponds to a unique taxon, and internal tree nodes $v \in V_\chi \backslash \mathcal{L}_\chi$ where each inner node has out-degree two and in-degree one. For each directed edge $(u, v) \in E_\chi$, v is a child of u. Each edge $e \in E_\chi$ can also have a branch length $bl(e) \in \mathbb{R}^+$. For vertices $u, v \in V_\chi$, u is an ancestor of v, $u \in anc(v)$, v is a descendant of u, and $v \in desc(u)$ if and only if u lies on the unique path from root ρ to v. The unrooted version of a rooted tree can be obtained by converting each directed edge into an undirected edge (i.e., ignoring directionality) and then collapsing the root node (i.e., omitting the root and connecting its previously outgoing edges into a single undirected edge).

The rooted evolutionary history of a set H of hosts is denoted by T_H and the rooted evolutionary history of a set S of symbionts is denoted by T_S. A mapping function $\pi(s, h) : S \times H \to \{0, 1\}$ denotes known interactions between the extant species of T_H and T_S, where $\pi(s, h) = 1$ means a symbiont is associated with a host, and otherwise $\pi(s, h) = 0$. The tuple (T_H, T_S, ϕ) serves as the input to cophylogenetic methods, and can be visualized using a tanglegram.

The cophylogenetic reconciliation problem is defined as follows. The problem input consists of a host tree T_H, symbiont tree T_S, and extant taxon interaction mapping π. The problem output is the set of coevolutionary event associations $\Phi \subset V_S \times V_H$. Each element $\phi \in \Phi$ associates an internal node $s \in V_S \backslash \mathcal{L}_S$ of the symbiont tree T_S with an internal node $h \in V_H \backslash \mathcal{L}_H$ of the host tree T_H, and represents a coevolutionary event such as cospeciation, duplication, loss, or host switching.

Cophylogenetic reconciliation methods fall into two broad categories: (1) global-fit methods, which evaluate congruence of host and symbiont phylogenies under statistical tests [4]; and (2) event-based methods, which reconcile a symbiont tree and host tree under a variant of the duplication-transfer-loss model (or other cophylogenetic model) to reconstruct a cophylogeny [5]. Event-based cophylogenies account for four broad categories of coevolutionary events [6]: cospeciation, host switch, duplication, and loss. In this study, we apply bootstrap resampling to assess reliability of coevolutionary event estimates in an event-based cophylogeny.

A. *CO^3: a new cophylogenetic support estimation method*

In practice, cophylogenetic reconciliation takes place within a larger computational pipeline. Multiple sequence alignment and phylogenetic reconstruction are typically performed first, where the host tree T_H and symbiont tree T_S are respectively estimated using a host MSA A_H and a symbiont MSA A_S as input. (As an aside, the two phylogenetic analyses are often performed independently – a simplifying assumption that conflicts with a coevolutionary hypothesis). Subsequently, the host tree T_H, symbiont tree T_S, and extant taxon interaction mapping π are reconciled into the cophylogeny Φ. As in traditional phylogenetic tree reconstruction, it is natural to ask: is the reconstructed cophylogeny Φ a reliable estimate?

We propose to answer this question by addressing the following problem, which we refer to as the cophylogenetic support estimation problem. The problem input consists of a cophylogenetic reconciliation Φ for the host tree T_H, symbiont tree T_S, and extant host/symbiont interaction mapping π. The problem output consists of a set of support values $\sigma(\phi) \in [0, 1]$ for each cophylogenetic reconciliation element/event $\phi \in \Phi$.

To address the cophylogenetic support estimation problem, we introduce a new method that performs "COnfidence interval estimation for COphylogenetic reCOnciliation": CO^3 ("CO-cubed"). CO^3 utilizes bootstrap resampling, a statistical resampling technique that has many applications throughout science and engineering [7].

To begin, the standard bootstrap resampling method is used to resample sites uniformly at random with replacement from the host MSA A_H and thereby obtain a host replicate MSA $A_H^{(1)}$. Standard bootstrap resampling is similarly performed on the symbiont MSA A_S to obtain a symbiont replicate MSA $A_S^{(1)}$. The process is repeated to obtain a total of m replicates of host and symbiont MSA pairs $(A_H^{(1)}, A_S^{(1)}), (A_H^{(2)}, A_S^{(2)}), \ldots, (A_H^{(m)}, A_S^{(m)})$.

On each replicate pair of MSAs $A_H^{(i)}$ and $A_S^{(i)}$ for $1 \leq i \leq m$, re-estimation is performed using a cophylogenetic reconciliation pipeline. The first stage takes as input the host MSA $A_H^{(i)}$ and reconstructs a host tree $T_H^{(i)}$; the symbiont MSA $A_S^{(i)}$ is similarly used to construct a symbiont tree $T_S^{(i)}$. The subsequent stage reconciles the host and symbiont trees $T_H^{(i)}$ and $T_S^{(i)}$ (along with an additional input consisting of the extant taxon interaction mapping π), resulting in a reconstructed cophylogeny $\Phi^{(i)}$.

The final step annotates the originally estimated cophylogeny Φ with cophylogenetic support values. Specifically, a support value $\sigma(\phi)$ for each cophylogenetic element/event $\phi \in \Phi$ is calculated as the proportion of re-estimated cophylogenies in the set $\{\Phi^{(i)} : 1 \leq i \leq m\}$ that also display ϕ (i.e. $\phi \in \Phi^{(i)}$).

B. *Simulation experiments*

Following the approach of Zheng et al. [3], the simulation experiments utilized one of two different simulation procedures. Each "forward" simulation was performed purely in silico under Treeducken's cophylogenetic birth-death model [8], as implemented in the R package Treeducken v1.0.0. Each "mixed" simulation utilized a reference cophylogeny that was based on an empirical estimate from a previously published empirical study. Each simulation condition included a pair of model species trees, a set of associations for host and symbiont taxa, and a reference/model cophylogeny. The model conditions and simulation procedures were reproduced from the study of [3]. Here we recap these procedures.

a) Forward simulations: Forward simulations were performed using a custom-modified version of Treeducken v1.0.0 [8] under its forward-time cophylogenetic birth-death model. Treeducken was modified to output historical association matrices, cophylogenetic events, and host/symbiont lineages to

TABLE I

SUMMARY STATISTICS FOR FORWARD SIMULATION CONDITIONS. *For each model condition ("Model conditions"), modified Treeducken was used to simulate under its cophylogenetic birth-death model with dataset characteristics based on a previously published cophylogenetic study ("Source"). Every model condition comprises of a model cophylogeny, model species trees, host-symbiont associations, and MSAs. For host and symbiont taxa, the number of taxa ("# taxa"), mean model tree height ("tree height"), true MSA length ("aln lengths"), mean and standard error of normalized Hamming distance of true MSAs ("ANHD Avg" and "ANHD SE", respectively) are reported. The reference cophylogenies were simulated with Treeducken. The number of coevolutionary events in the reference cophylogenies are listed by event type: cospeciations ("# CSP"), symbiont speciations ("# SSP"), symbiont host expansion/ symbiont host switch ("# SHE/SHS"), symbiont extinction/ missing the boat ("# SX/MTB"), and host speciation ("# HSP").*

Model condition	Source	Taxa	# taxa	tree height	aln lengths	ANHD Avg	ANHD SE	# CSP	# SSP	# SHE/SHS	# SX/MTB	# HSP
forward_gopher	[9]	hosts	11	1.4006	400	0.4683	0.1994	10	3	0	0	0
		symbionts	14	1.4035	400	0.4524	0.1521					
forward_stinkbug	[10]	hosts	4	0.3886	1,000	0.2610	0.1129	3	5	0	0	0
		symbionts	9	0.7478	1,000	0.4354	0.1901					
forward_primate	[11]	hosts	40	1.3804	400	0.4652	0.0796	39	2	15	0	0
		symbionts	57	1.2220	400	0.5579	0.0923					
forward_damselfly	[12]	hosts	18	1.7693	1,000	0.4992	0.1228	17	9	3	12	0
		symbionts	30	1.5837	1,000	0.5170	0.1124					
forward_bird	[13]	hosts	27	1.6187	5,000	0.6019	0.1084	23	20	1	4	3
	[14]	symbionts	45	2.1712	5,000	0.6595	0.1103					
forward_moth	[15]	hosts	65	2.4827	3,000	0.5724	0.0941	60	16	24	28	4
		symbionts	101	2.4773	3,000	0.6239	0.0948					

TABLE II

TREEDUCKEN PARAMETERS USED IN FORWARD SIMULATIONS. *We used a modified version of Treeducken [8] (supplementary Section S6) to simulate cophylogenies, their constituent species phylogenies, and the host-symbiont associations. Treeducken's cophylogenetic birth-death model specifies the following parameters: the symbiont speciation rate λ_S, the symbiont extinction rate μ_S, the cospeciation rate λ_C, the host speciation rate λ_H, the host extinction rate μ_H, the expected number of host taxa H_{tips}, the expected number of symbiont taxa S_{tips}, the host switch or host expansion rate hs_rate, and length of coevolutionary time.*

Model condition	H_{tips}	S_{tips}	λ_H	λ_C	λ_S	μ_H	μ_S	hs_rate	time
forward_gopher	5	1	0.0010	1.2000	0.5776	0.0000	0.4010	0.0099	1.0000
forward_stinkbug	12	12	0.0031	1.0000	0.8996	0.0000	0.2031	0.0099	1.0000
forward_primate	155	155	0.0800	2.6660	0.0920	0.0003	0.0280	1.0873	0.7500
forward_damselfly	155	155	0.0040	1.1000	0.9000	0.0000	1.0100	0.4990	1.3000
forward_bird	35	55	0.3974	1.0000	0.9000	0.0000	0.0100	0.4990	1.3000
forward_moth	100	100	0.1239	2.0043	0.5200	0.0005	0.7200	0.9730	1.2000

obtain model species trees, host-symbiont associations, and a ground truth cophylogeny (see Supplementary Section S6). The simulation outputs were further processed to remove extinct lineages as well as map Treeducken's event types to the four DTL event categories: cospeciation, duplication, host switch, and loss. Model condition parameter settings were based on empirical estimates from previously published studies, including the dataset size (number of taxa and sequence length), substitution rates, base frequency rates, and coevolutionary event frequencies. Model species trees were deviated away from ultrametricity using the approach in Moret et al. [16] with deviation factor $c = 2.0$.

An additional set of experiments varied evolutionary divergence for the forward-gopher model condition. In these experiments, branch lengths in the model species trees were scaled by factor $h \in \{0.5, 1.0, 2.0\}$. (In the rest of the forward simulation experiments, the scaling factor was effectively $h = 1$.)

Seq-Gen v1.3.4 [17] was then used to simulate DNA sequence evolution along model host and symbiont species trees to produce true host and symbiont MSAs. The simulations were performed under the General Time Reversible model of nucleotide substitution with Γ-distributed rate heterogeneity (GTR+Γ). The substitution model parameter settings were based on MLE analyses of the empirical dataset on which each model condition was based (Table I column "Source").

b) Mixed simulations: Each mixed simulation used a set of empirical estimates (i.e., a host tree, symbiont tree, and cophylogeny) as a model instance to perform parametric simulations. To begin, six empirical datasets from past studies were selected to span a range of coevolutionary scenarios (Table III column "Source"). The empirical datasets were preprocessed so that extant host-symbiont associations were subsampled to be one-to-one, as required by the cophylogenetic reconciliation method under study (i.e., eMPRess [18]). Each set of unaligned sequence data was then aligned with MAFFT v7.221 using default settings [19]. Using the host and symbiont MSAs as input, we estimated MLE trees under the GTR+Γ model with RAxML v8.1.12 [20] for the hosts and symbionts, respectively; the trees were then midpoint rooted. The rooted trees served as the model trees for sequence evolution simulations. As in the forward simulation experiments, branch lengths in the model trees were scaled by a parameter h; the default setting was effectively $h = 1$ throughout the mixed simulation experiments, with the exception of an additional set of experiments for the mixed-gopher model condition where $h \in \{0.5, 1.0, 2.0\}$. Next, the host and symbiont trees (and host-symbiont associations) were reconciled to obtain a cophylogeny using eMPRess command line interface (CLI) v1.2.1 [18] under default cost settings; the latter served as the reference cophylogeny for subsequent performance assessments. DNA sequence evolution along model host and symbiont trees was simulated using the same procedure as in the forward simulation experiments, yielding true host and symbiont MSAs.

c) Phylogenetic and cophylogenetic reconstruction: Using RAxML v8.1.12 [20], the simulated host and symbiont MSAs were used to reconstruct MLE trees under the GTR+Γ model, and the estimated trees were then midpoint rooted. The rooted host tree, rooted symbiont tree, and the extant host-symbiont associations were then reconciled to obtain a reconstructed cophylogeny using eMPRess v1.2.1 [18] via

TABLE III

SUMMARY STATISTICS FOR MIXED SIMULATION MODEL CONDITIONS. *Mixed simulation conditions utilized a reference cophylogeny, host tree, and species tree that were obtained via empirical estimates from previously published cophylogenetic studies ("Source"); the latter two trees then served as model trees to perform sequence evolution simulations and obtain true MSAs. (See Methods section for details.) The number of cospeciation, duplication, host switch, and loss events in each reference cophylogeny are reported as "# cosp", "# dup", "# hs", and "# loss", respectively. Table description is otherwise identical to Table I.*

Model conditions	Source	Taxa	# taxa	aln length	ANHD Avg	ANHD SE	tree height	# cosp	# dup	# hs	# loss
mixed-gopher	[9]	Host	15	379	0.2239	0.0722	0.3575	10	0	6	5
		Symbiont	17	379	0.5245	0.0802	2.4764				
mixed-stinkbug	[10]	Host	7	1,745	0.2694	0.0904	0.2751	6	5	0	0
		Symbiont	12	1,583	0.0689	0.0470	0.0860				
mixed-primate	[11]	Host	63	696	0.2752	0.0980	1.5920	19	3	20	14
		Symbiont	43	425	0.3877	0.1266	1.3659				
mixed-damselfly	[12]	Host	24	1,051	0.1726	0.0680	0.4897	5	2	15	3
		Symbiont	23	3,297	0.1320	0.0622	0.1777				
mixed-bird	[13]	Host	200	5,000	0.1757	0.0587	0.5007	6	15	35	7
	[14]	Symbiont	57	5,000	0.3519	0.0532	0.5017				
mixed-moth	[15]	Host	82	1,404	0.1024	0.0631	0.1574	14	0	26	15
		Symbiont	41	4,326	0.0250	0.0143	0.0393				

its command line interface (CLI). The latter serves as the "annotation" estimate for CO^3 support estimation.

d) Phylogenetic and cophylogenetic support estimation: CO^3 was used to estimate cophylogenetic support as follows. Standard bootstrap resampling was performed on the host MSA to obtain 100 bootstrap replicate MSAs, and resampling was performed similarly on the symbiont MSA to obtain 100 bootstrap replicates for symbionts. RAxML v8.1.12 [20] was used to perform all bootstrap resampling.

Re-estimation was then performed on each bootstrap replicate. Using a bootstrap replicate MSA as input, an MLE tree was re-estimated under the GTR+Γ model and midpoint rooted using RAxML [20]. The rooted host and symbiont trees and the extant host-symbiont associations were then used to re-estimate a cophylogenetic reconciliation using eMPRess CLI [18] with default cost settings.

A custom script was used to annotate the original "annotation" cophylogeny with cophylogenetic support values. The latter utilized the cophylogenetic support calculated against the re-estimated cophylogenies across all bootstrap replicates, as described above. Phylogenetic support values for "annotation" trees were similarly calculated using re-estimated trees on bootstrap replicates.

e) Performance assessments: On each simulated dataset, CO^3 support estimates on the annotation cophylogeny were compared against the reference cophylogeny to assess type I and type II error. The comparison results were categorized into four classes, each of which is represented by a cell in a confusion matrix. True positives (TP) consist of elements/events that appear in the annotation cophylogeny with support value greater than or equal to a given threshold and also appear in the reference cophylogeny. False positives (FP) consist of elements/events in the annotation cophylogeny that appear with support greater than or equal to a given threshold, but do not exist in the reference cophylogeny. False negatives (FN) consist of elements/events that appear in the annotation cophylogeny with support value less than a given threshold but appear in the reference cophylogeny. Finally, true negatives (TN) consist of elements/events that appear in the annotation cophylogeny with support value less than a given threshold and do not appear in the reference cophylogeny.

Receiver operator characteristic (ROC) curves and precision-recall (PR) curves and their respective area under curves (AUCs) were then calculated. The ROC curve plots true positive rate $\frac{|\mathrm{TP}|}{|\mathrm{TP}|+|\mathrm{FN}|}$ versus false positive rate $\frac{|\mathrm{FP}|}{|\mathrm{FP}|+|\mathrm{TN}|}$ as a support threshold is varied to calculate data points along the curve. The PR curve plots precision $\frac{|\mathrm{TP}|}{|\mathrm{TP}|+|\mathrm{FP}|}$ against recall (also known as true positive rate) $\frac{|\mathrm{TP}|}{|\mathrm{TP}|+|\mathrm{FN}|}$ as a support threshold is varied. We used the scikit-learn Python library [21] to calculate curves and AUCs.

Type I and II error of phylogenetic tree support was assessed using a similar approach. A confusion matrix was constructed by comparing bootstrap support estimates on the annotation tree versus the reference tree. True positives consist of bipartitions that appear in the annotation tree with support greater than or equal to a given threshold and that also appear in the reference tree. False positives consist of bipartitions in the annotation tree that appear with support greater than or equal to a given threshold, but do not exist in the reference tree. False negatives consist of bipartitions that appear in the annotation tree with support less than a given threshold but actually appear in the reference tree. True negatives consist of bipartitions in the annotation tree that appear with support less than a given threshold and do not exist in the reference tree. ROC and PR curves and their respective AUCs were calculated in the same manner as for cophylogenetic support assessments.

Topological accuracy of an estimated unrooted tree versus a reference unrooted tree is reported based on pairwise Robinson-Foulds (RF) distance between the two trees [22]. The Robinson-Foulds (RF) distance is defined as the symmetric difference between the two trees' bipartition sets. The normalized RF distance divides the RF distance by its maximum (i.e., $2(n-3)$). Topological accuracy of a reconstructed cophylogeny versus a reference cophylogeny is assessed using cophylogenetic precision [23]. Given cophylogenies Φ_A and

Φ_B, cophylogenetic precision is defined as $\frac{|\Phi_B \cap \Phi_A|}{|\Phi_A|}$, which is the proportion of reconciled coevolutionary events in Φ_A that were also found in Φ_B.

f) Experimental replication: The simulation procedures were repeated for each model condition to obtain 100 experimental replicates. Simulation study results are reported across all replicates in a model condition. In particular, ROC and PR curves and AUC values are reported in aggregate across experimental replicates for each model condition.

C. Empirical study of bobtail squids and their bioluminescent bacteria

a) Data acquisition: The genomic data for 22 bobtail squids were sourced from Sanchez et al. [24], and the genomic sequences of 37 bioluminescent *Vibrio* samples were sourced from Bongrand et al. [25]. The authors of [24] used genome skimming to identify ultraconserved loci and concatenated the MSAs to obtain total host MSA length of 37,512 bp. We post-processed and concatenated the multilocus gene alignments for *Vibrio* in an earlier paper [3] to produce a symbiont MSA length of 2,722,691 bp. The known host-symbiont associations data were obtained from Bongrand et al. [25].

b) Phylogenetic and cophylogenetic reconstruction: Species trees for bobtail squids and bioluminescent bacteria were reconstructed using maximum likelihood estimation (MLE) under the GTR+Γ model and then midpoint rooted using RAxML v8.1.12 [20]. Cophylogenetic reconciliation of the estimated species trees and the known squid-*Vibrio* associations was performed using eMPRess CLI v1.2.1 [18].

c) Phylogenetic and cophylogenetic support estimation: The bobtail squid and bioluminescent *Vibrio* MSAs were resampled to obtain 100 bootstrap replicates. For each bootstrap replicate, a maximum likelihood tree was estimated under the GTR+Γ model and then midpoint rooted using RAxML analysis of the replicate's MSA. [20]. The rooted host and symbiont trees were then reconciled into a species cophylogeny using eMPRess CLI [18] with default event cost settings. Confidence intervals were calculated for the empirically-estimated cophylogeny using the bootstrap re-estimated cophylogenies, following the same approach as in the simulation experiments. The reconstructed cophylogeny with support values was visualized using eMPRess v1.2.1 [18] and its graphical user interface (GUI). The estimation procedure was repeated to obtain a total of 10 independent analyses.

III. Results

A. Simulation experiments

Type I and type II error of CO^3 support estimation was assessed based on area under receiver operating characteristic curve and precision-recall curve (ROC AUC and PR AUC, respectively). CO^3 AUC values are reported for the forward and mixed simulation conditions in Table IV. Across the mixed simulation conditions, CO^3 returned average PR-AUC of 93.9% (with a maximum of 99.8% and a minimum of 84.4%); average ROC-AUC was 87.5% (with a maximum of 96.5% and a minimum of 78.7%). In the mixed simulation experiments, the highest AUC values were observed on the mixed-bird model condition; the mixed-primate, mixed-stinkbug, mixed-damselfly, and mixed-gopher appeared in the next highest category of AUC values, in descending order, and the lowest AUC values were observed on the mixed-moth model condition. Across the forward simulation conditions, CO^3 returned average PR-AUC of 83.7% (with range between 98.1% and 70.3%) and average ROC-AUC of 79.7% (range between 94.8% and 63.1%). In the forward simulation experiments, the highest AUC values were observed on the forward-bird and forward-gopher model conditions; the other four model conditions – forward-stinkbug, forward-primate, forward-damselfly, and forward-moth – returned lower CO^3 AUC values.

As phylogenetic tree re-estimation is an essential step of CO^3 analysis, we also assessed the performance of bootstrap support for host and symbiont trees. Table V shows PR-AUC and ROC-AUC as an average for host and symbiont trees for each mixed and forward simulation condition. The mixed simulation experiments resulted in mean PR-AUC of 99.5% (with range between 100.0% and 97.1%) and mean ROC-AUC of 95.0% (with range between 100.0% and 68.9%). The forward simulation experiments returned mean PR-AUC of 99.6% (with range between 100.0% to 98.3%) and mean ROC-AUC of 96.9% (with range between 99.8% and 93.6%).

Finally, we examined the impact of the annotation cophylogeny's accuracy on type I/II error of CO^3 support estimation. Our experiments included a wide range of annotation cophylogeny accuracy (Table VI), where cophylogenetic precision ranged between 40.4% and 94.1%. The annotation cophylogeny in the mixed and forward simulation experiments had average precision of 75.5% and 62.6%, respectively – a difference of 13.0%. Scatterplots and linear regression analyses were used to quantify the relationship between annotation cophylogeny accuracy and type I/II error returned by CO^3 (Figures 1 and 2). The scatterplot of CO^3 ROC-AUC versus annotation cophylogeny precision is shown in Figure 1. Accompanying linear regression analyses yielded a linear model with slope near 0 for both the mixed and forward simulation experiments; statistical testing indicated that the fitted model's slope coefficient was not significantly different from a null hypothesis of 0. The relationship between CO^3 PR-AUC and annotation cophylogenetic precision was also weak, with linear model slope of 0.211 and 0.434 for the mixed and forward simulation experiments, respectively (Figure 2).

An additional set of experiments examined the performance of CO^3 as evolutionary divergence varied (Table VII). Varying evolutionary divergence as the reference tree height was scaled by a factor of 0.5 or 2 resulted in a small decrease in ROC-AUC by 4.46% and 4.50%, respectively, and a similarly small decrease in PR-AUC of 3.32% and 3.22%, respectively.

CO^3 runtime and main memory usage is reported in Supplementary Figure S1. For one experimental replicate consisting of 100 bootstrap replicates for the host and symbiont taxa respectively, average runtime on each simulated dataset from

TABLE IV

TYPE I AND TYPE II ERROR OF CO^3 COPHYLOGENETIC SUPPORT ESTIMATES IN THE MIXED AND FORWARD SIMULATION EXPERIMENTS. *Type I and type II error was assessed using area under curve (AUC) for receiver operating characteristic (ROC) curves and precision-recall (PR) curves ("ROC AUC" and "PR AUC", respectively). AUC values are reported as an aggregate across experimental replicates in each model condition ($n = 100$).*

CO^3 performance		
Model condition	ROC AUC	PR AUC
mixed-gopher	0.8441	0.9152
mixed-stinkbug	0.9029	0.9709
mixed-primate	0.9049	0.9774
mixed-damselfly	0.8451	0.9294
mixed-bird	0.9654	0.9979
mixed-moth	0.7872	0.8435
forward-gopher	0.9482	0.9740
forward-stinkbug	0.6310	0.8276
forward-primate	0.7948	0.7031
forward-damselfly	0.7097	0.7084
forward-bird	0.9428	0.9806
forward-moth	0.7544	0.8309

TABLE V

TYPE I AND TYPE II ERROR OF PHYLOGENETIC BOOTSTRAP SUPPORT ESTIMATES IN THE MIXED AND FORWARD SIMULATION EXPERIMENTS. *As in the CO^3 performance assessments, type I and type II error was assessed using area under curve (AUC) for receiver operating characteristic (ROC) curves and precision-recall (PR) curves ("ROC AUC" and "PR AUC", respectively). Reported AUC values are aggregated across experimental replicates in each model condition ($n = 100$).*

	Host tree		Symbiont tree	
Model condition	ROC AUC	PR AUC	ROC AUC	PR AUC
mixed-gopher	0.9759	0.9984	0.9689	0.9977
mixed-stinkbug	1.0000	1.0000	0.9569	0.9929
mixed-primate	0.9440	0.9964	0.9606	0.9988
mixed-damselfly	0.9696	0.9991	0.6885	0.9707
mixed-bird	0.9910	0.9999	1.0000	1.0000
mixed-moth	0.9627	0.9903	0.9598	0.9945
forward-gopher	0.9470	0.9920	0.9586	0.9947
forward-stinkbug	0.9980	0.9999	0.9969	1.0000
forward-primate	0.9448	0.9920	0.9560	0.9934
forward-damselfly	0.9358	0.9832	0.9868	0.9995
forward-bird	0.9772	0.9985	0.9704	0.9978
forward-moth	0.9683	0.9983	0.9848	0.9993

the mixed-bird condition was under 2.5 hours and had peak memory usage of less than 200 MB. All other datasets required comparatively less runtime and peak memory usage.

B. Empirical study

The histograms of event frequencies for bobtail squids and *Vibrio* are reported in Figure 3. The average cumulative frequency for all four event types was higher than 90. On average, each bootstrapped replicate's cophylogeny contained 2 cospeciations, 2 host switches, 3 losses, and 30 duplications.

The visualized tanglegram with phylogenetic bootstrap values for bobtail squid and *Vibrio* is presented in Supplementary Figure S2. The phylogenetic bootstrap values for bioluminescent bacteria was above 96% except for two clades of *Vibrio* with 70% and 77%. Similarly for the bobtail squid phylogeny,

TABLE VI

COPHYLOGENETIC PRECISION OF ANNOTATION COPHYLOGENIES. *Average ("Mean") and standard error ("SE") of cophylogenetic precision are reported across experimental replicates for each model condition ($n = 100$).*

Cophylogenetic precision					
Model condition	Mean	SE	Model conditions	Mean	SE
mixed-gopher	0.6567	0.0205	forward-gopher	0.7415	0.0132
mixed-stinkbug	0.7818	0.0120	forward-stinkbug	0.8525	0.0051
mixed-primate	0.8136	0.0079	forward-primate	0.5120	0.0037
mixed-damselfly	0.7660	0.0183	forward-damselfly	0.4042	0.0063
mixed-bird	0.9410	0.0045	forward-bird	0.6938	0.0033
mixed-moth	0.5718	0.0165	forward-moth	0.5496	0.0029

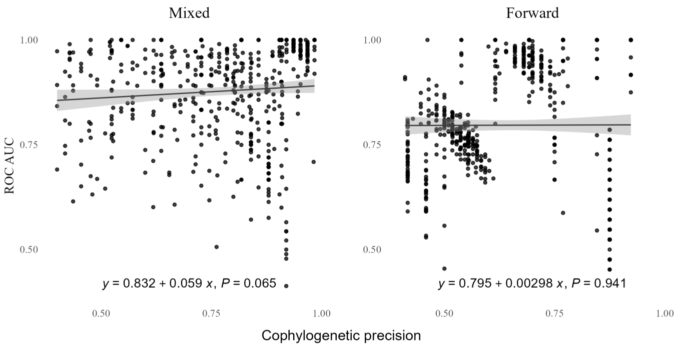


Fig. 1. *The relationship between CO^3 support estimation error (as assessed using area under receiver operating characteristic curve or ROC AUC) versus annotation cophylogeny precision.* A scatterplot is shown, where each data point shows CO^3 ROC AUC and precision of the annotation cophylogeny for an experimental replicate from a model condition ($n = 1200$). A linear regression analysis was also performed. The fitted model is shown as a trendline and equation; a p-value from a statistical test is also reported.

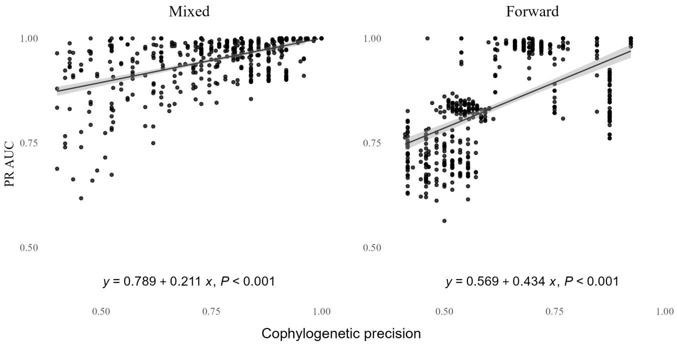


Fig. 2. *The relationship between CO^3 support estimation error (as assessed using area under precision-recall curve or PR AUC) versus annotation cophylogeny precision.* Figure layout and description are otherwise identical to Figure 1.

TABLE VII

TYPE I AND TYPE II ERROR OF CO^3 COPHYLOGENETIC SUPPORT ESTIMATES IN THE MIXED-GOPHER AND FORWARD-GOPHER SIMULATION EXPERIMENTS WITH VARYING EVOLUTIONARY DIVERGENCE. *In these experiments, branch lengths in the model trees were scaled by a factor $h \in \{0.5, 1, 2.0\}$. Type I and type II error was assessed using area under receiver operating characteristic curve and precision-recall curve ("ROC AUC" and "PR AUC", respectively). For each model condition, reported AUC values are aggregated across experimental replicates ($n = 100$).*

Evolutionary divergence on CO^3 performance						
	height x 0.5		original (height x 1)		height x 2	
Model condition	ROC AUC	PR AUC	ROC AUC	PR AUC	ROC AUC	PR AUC
mixed-gopher	0.7711	0.8543	0.8441	0.9152	0.7620	0.8623
forward-gopher	0.9319	0.9686	0.9482	0.9740	0.9402	0.9625

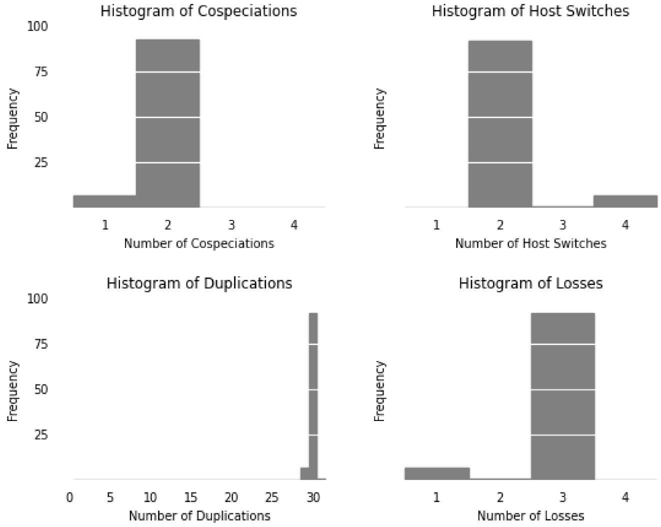


Fig. 3. *Histogram of event frequencies from CO^3 re-estimated cophylogenies for bobtail squids and bioluminescent bacteria dataset.* Y-axis labels are shared across all subplots. Frequency was normalized to one estimation replicate after aggregating across the 10 estimation replicates. Each independent estimation replicate included analyses of $m = 100$ bootstrap replicates.

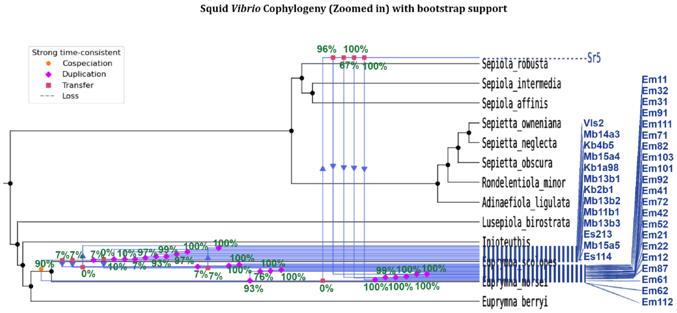


Fig. 4. *Estimated cophylogeny with CO^3 support values for the bobtail squid and bioluminescent bacteria dataset.* The estimated cophylogeny was visualized using a customized version of the eMPRess GUI software [18]. (See Supplementary Online Materials Section S5 for details). CO^3 support values for coevolutionary events are shown in green. CO^3 support estimation was run using 100 resampled replicates.

the bootstrap values were above 96% within the *Euprymna* clade and at the lowest 64% near the root of the tree.

Figure 4 shows the reconstructed bobtail squid and *Vibrio* cophylogeny (visualized using eMPRess GUI [18]). The bootstrap values showed clusters of high support (93% to 100%) in different regions of the cophylogeny.

IV. Discussion

To our knowledge, CO^3 is the first method for cophylogenetic support estimation. CO^3 provides a practical, data-driven assessment of cophylogenetic reconstruction reliability.

Across the model conditions in our simulation study, CO^3 returned average ROC-AUC of 83.6% and average PR-AUC of 88.8%. The comparison of mixed simulation experiments versus forward simulation experiments showed an average increase of 7.8% ROC-AUC and 10.2% PR-AUC of the former over the latter. Some of the highest CO^3 AUC values in simulation study were observed on forward-bird and mixed-bird model conditions. This model condition also had longest sequence length by several kb, and the increased amount of character data benefits statistical MLE and subsequent cophylogenetic reconciliation – during the initial stage to obtain an annotation MSA and tree, as well as subsequent MSA and tree re-estimation on bootstrap replicates.

We emphasize that the novelty of CO^3 implies a lack of existing methods to compare against. A tempting and imperfect point of comparison is to look to other resampling applications such as phylogenetic tree support estimation, but we caution that such cross-task comparisons bring challenges. As an example, we note that performance assessments of phylogenetic bootstrap support estimation for mixed and forward simulation experiments in our study had average ROC-AUC of 95.8% and mean PR-AUC of 99.5% – AUC values that would suggest lower type I and type II error than those observed for cophylogenetic support estimation. However, we note that the latter problem is more complex than – and in fact contains – the former problem, for the problem formulations under study. As a result, methods for addressing the latter problem tend to be more complicated (e.g., pipeline-based methods tend to have more estimation stages, each of which introduces additional opportunities for estimation error to arise and propagate to downstream stages) as compared to methods for addressing the former problem. Comparisons of methods for solving the two different problems tend to be fraught for these reasons.

We also examined the impact of several experimental factors on CO^3 performance. One of the additional experimental factors was annotation cophylogeny accuracy. The scatterplots in Figure 1 showed CO^3's performance as measured by ROC-AUC was not correlated with cophylogenetic precision of the annotation cophylogenies. This indicated that the performance of CO^3 was largely robust to the annotation cophylogeny quality. The finding indicates wide applicability of CO^3 across a range of annotation cophylogenies – from inaccurate to accurate. Another experimental factor concerned evolutionary divergence, which is known to impact phylogenetic reconstruction [26], [27] as well as cophylogenetic reconciliation [3]. We examined the effect of evolutionary divergence on CO^3 by scaling model tree heights in our simulation experiments, and we found that on average CO^3 performance was mostly unaffected (within 0.05 AUC value).

The reconstructed cophylogeny for bobtail squid and *Vibrio* and its CO^3 support value annotation (Figure 4) showed that there were parts of the cophylogeny that were highly supported and parts that were not well supported. Earlier coevolutionary events that occurred closer to the root of the squid phylogeny were less well supported (0% to 7%). In contrast, cophylogenetic events estimated near the leaves of the squid tree were highly supported (99% to 100%) duplications. The observation of lower cophylogeny support near the root of the squid phylogeny is not mirrored in the squid phylogeny or *Vibrio* phylogeny (see tanglegram in Figure S2), which indicates that the uncertainty may have arisen due to the combination of phylogenetic rooting and cophylogenetic reconciliation. Low cophylogenetic bootstrap confidence limits near the root of the squid phylogeny indicates that the reconciled squid-*Vibrio* cophylogeny is less reliable the further back in time.

V. Conclusion

In this study, we introduce CO^3 – the first method for cophylogenetic support estimation, to our knowledge. CO^3 applies bootstrap resampling of MSAs to place confidence intervals on a reconstructed cophylogeny. We assessed the performance of CO^3 support estimation using simulation experiments as well as an empirical study of bobtail squids and their bioluminescent endosymbionts. Overall, CO^3 support estimation provided insight into which regions of a cophylogeny are more trustworthy and provided more granular insights into different event types.

We conclude with thoughts on some future research directions. First, a particular need in today's post-genomic era is to expand this work to next generation sequencing and large-scale genomic sequence datasets. Second, coalescent-based approaches for statistical cophylogenetic reconstruction such as TALE [28] have been recently introduced. The performance of these new statistical methods is not yet fully understood, and statistical resampling may provide important early indicators about their strengths and weaknesses.

Acknowledgments

This research has been supported in part by the National Science Foundation (2144121, 2214038, 1740874, 1737898 to KJL), MSU and the National Science Foundation Research Traineeship Program (DGE-1828149) to JZ, and MSU (EEB and CSE summer fellowship to JZ). All computational experiments and analyses were performed on the MSU High Performance Computing Center, which is part of the MSU Institute for Cyber-Enabled Research.

References

[1] B. Efron, *The jackknife, the bootstrap and other resampling plans*. SIAM, 1982.

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

[3] J. Zheng, Y. Nishida, A. Okrasinska, G. M. Bonito, E. A. Heath-Heckman, and K. J. Liu, "The impact of species tree estimation error on cophylogenetic reconstruction," in *Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2023, pp. 1–10.

[4] I. Blasco-Costa, A. Hayward, R. Poulin, and J. A. Balbuena, "Next-generation cophylogeny: unravelling eco-evolutionary processes," *Trends in Ecology & Evolution*, vol. 36, no. 10, pp. 907–918, 2021.

[5] R. Libeskind-Hadas, "Tree reconciliation methods for host-symbiont cophylogenetic analyses," *Life*, vol. 12, no. 3, p. 443, 2022.

[6] M. Charleston and R. Libeskind-Hadas, "Event-based cophylogenetic comparative analysis," in *Modern phylogenetic comparative methods and their application in evolutionary biology*. Springer, 2014, pp. 465–480.

[7] B. Efron and R. J. Tibshirani, *An introduction to the bootstrap*. CRC press, 1994.

[8] W. Dismukes and T. A. Heath, "treeducken: An R package for simulating cophylogenetic systems," *Methods Ecol. Evol.*, vol. 12, no. 8, pp. 1358–1364, 2021.

[9] M. S. Hafner, P. D. Sudman, F. X. Villablanca, T. A. Spradling, J. W. Demastes, and S. A. Nadler, "Disparate rates of molecular evolution in cospeciating hosts and parasites," *Science*, vol. 265, no. 5175, pp. 1087–1090, 1994.

[10] T. Hosokawa, Y. Kikuchi, N. Nikoh, M. Shimada, and T. Fukatsu, "Strict host-symbiont cospeciation and reductive genome evolution in insect gut bacteria," *PLoS Biology*, vol. 4, no. 10, p. e337, 2006.

[11] W. M. Switzer, M. Salemi, V. Shanmugam, F. Gao, M.-e. Cong, C. Kuiken, V. Bhullar, B. E. Beer, D. Vallet, A. Gautier-Hion *et al.*, "Ancient co-speciation of simian foamy viruses and primates," *Nature*, vol. 434, no. 7031, pp. 376–380, 2005.

[12] M. Lorenzo-Carballa, Y. Torres-Cambas, K. Heaton, G. Hurst, S. Charlat, T. Sherratt, H. Van Gossum, A. Cordero-Rivera, and C. Beatty, "Widespread Wolbachia infection in an insular radiation of damselflies (Odonata, Coenagrionidae)," *Scientific Reports*, vol. 9, no. 1, pp. 1–13, 2019.

[13] R. O. Prum, J. S. Berv, A. Dornburg, D. J. Field, J. P. Townsend, E. M. Lemmon, and A. R. Lemmon, "A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing," *Nature*, vol. 526, no. 7574, pp. 569–573, 2015.

[14] R. S. de Moya, J. M. Allen, A. D. Sweet, K. K. Walden, R. L. Palma, V. S. Smith, S. L. Cameron, M. P. Valim, T. D. Galloway, J. D. Weckstein *et al.*, "Extensive host-switching of avian feather lice following the Cretaceous-Paleogene mass extinction event," *Communications Biology*, vol. 2, no. 1, p. 445, 2019.

[15] Y. Zhang, S. Zhang, Y. Li, S. Ma, C. Wang, M. Xiang, X. Liu, Z. An, J. Xu, and X. Liu, "Phylogeography and evolution of a fungal–insect association on the Tibetan Plateau," *Molecular Ecology*, vol. 23, no. 21, pp. 5337–5355, 2014.

[16] B. M. Moret, U. Roshan, and T. Warnow, "Sequence-length requirements for phylogenetic methods," vol. 2452. Springer, 2002, pp. 343–356.

[17] A. Rambaut and N. C. Grass, "Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees," *Bioinformatics*, vol. 13, no. 3, pp. 235–238, 1997.

[18] S. Santichaivekin, Q. Yang, J. Liu, R. Mawhorter, J. Jiang, T. Wesley, Y.-C. Wu, and R. Libeskind-Hadas, "eMPRess: a systematic cophylogeny reconciliation tool," *Bioinformatics*, vol. 37, no. 16, pp. 2481–2482, 2021.

[19] 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.

[20] 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, publisher: Oxford University Press.

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[22] D. F. Robinson and L. R. Foulds, "Comparison of phylogenetic trees," *Mathematical Biosciences*, vol. 53, no. 1-2, pp. 131–147, 1981.

[23] N. Wieseke, T. Hartmann, M. Bernt, and M. Middendorf, "Cophylogenetic reconciliation with ILP," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 12, no. 6, pp. 1227–1235, 2015.

[24] G. Sanchez, F. Á. Fernández-Álvarez, M. Taite, C. Sugimoto, J. Jolly, O. Simakov, F. Marlétaz, L. Allcock, and D. S. Rokhsar, "Phylogenomics illuminates the evolution of bobtail and bottletail squid (order Sepiolida)," *Communications Biology*, vol. 4, no. 1, p. 819, 2021.

[25] C. Bongrand, S. Moriano-Gutierrez, P. Arevalo, M. McFall-Ngai, K. L. Visick, M. Polz, and E. G. Ruby, "Using colonization assays and comparative genomics to discover symbiosis behaviors and factors in *Vibrio fischeri*," *mBio*, vol. 11, no. 2, pp. e03 407–19, 2020.

[26] 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, p. 90, 2012.

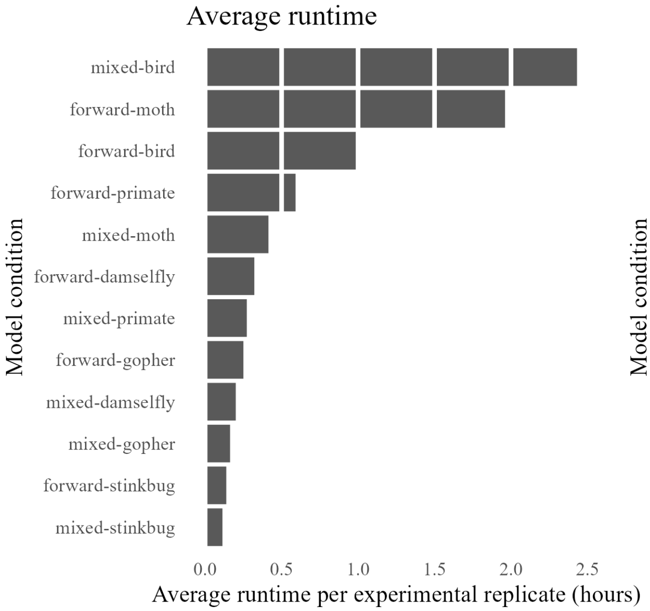
[27] S. Mirarab, N. Nguyen, S. Guo, L.-S. Wang, J. Kim, and T. Warnow, "PASTA: ultra-large multiple sequence alignment for nucleotide and amino-acid sequences," *Journal of Computational Biology*, vol. 22, no. 5, pp. 377–386, 2015.

[28] H. Menet, A. N. Trung, V. Daubin, and E. Tannier, "Host-symbiont-gene phylogenetic reconciliation," *Peer Community Journal*, vol. 3, 2023.

Supplementary Online Materials

Contents

S1. Runtime and peak CPU usage



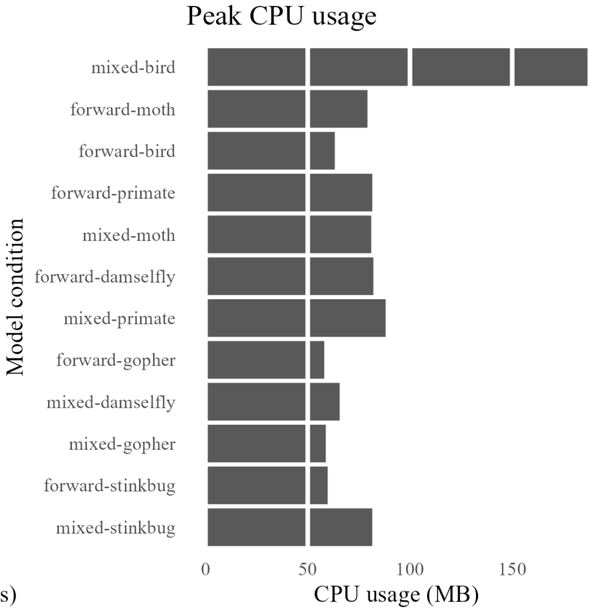


Fig. S1: Barplots of simulation model conditions with respect to average runtime and peak CPU usage for one experimental replicate. Each experimental replicate consists of the CO^3 method run in a sequential pipeline.

S2. Empirical dataset tanglegram

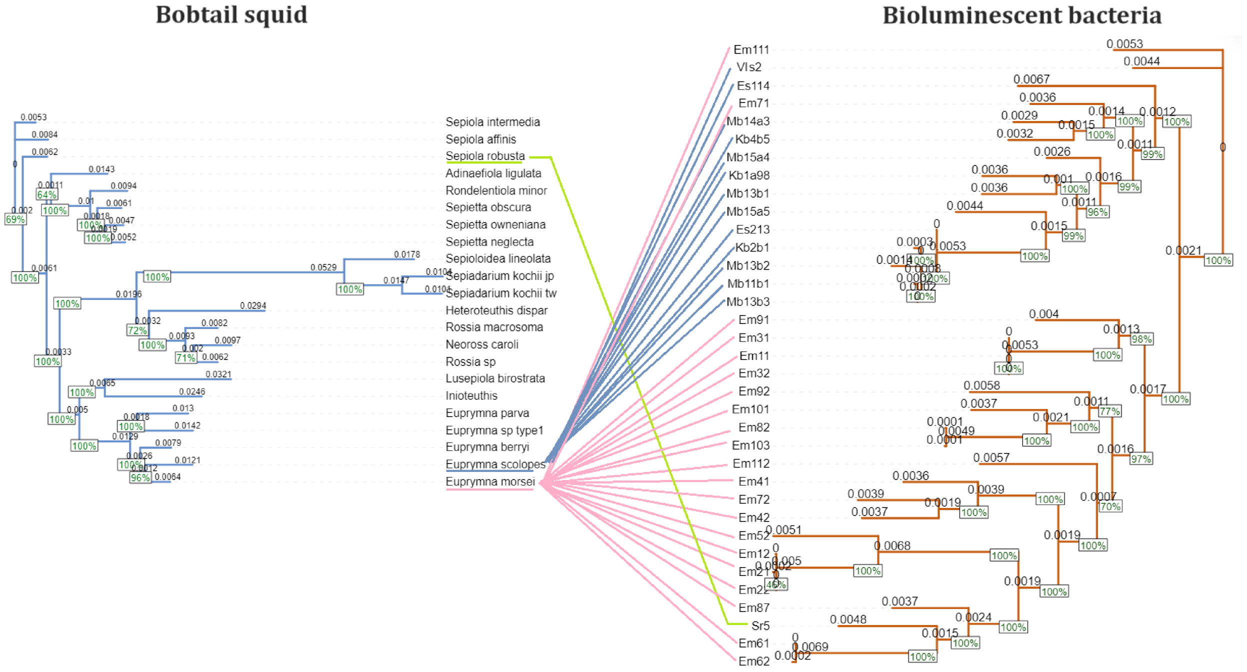


Fig. S2: Tanglegram with phylogenetic bootstrap values for the full bobtail squids (left) and bioluminescent bacteria (right) dataset. The lines between the phylogenies represent host-symbiont associations from Bongrand et al. [1]. Phylogenies were visualized using [2]. Host-symbiont associations were hand-annotated, such that blue lines match to *Euprymna scolopes*, pink lines match to *Euprymna morsei*, and the green line matches to *Sepiola robusta*.

S3. ADDITIONAL SUMMARY STATISTICS

TABLE S1: Summary statistics for mixed and forward simulations on the annotation alignments, trees, and cophylogenies. (n=100 experimental replicates, m=100 bootstrap resamples in each replicate).

			annotation aln		aln	annotation vs reference tree rSPR		annotation vs reference tree nRF		Annotation vs reference cophylogenetic precision	
Model conditions	Taxa	reps	ANHD avg	ANHD SE	length	avg	SE	nRF avg	nRF SE	prec avg	prec SE
mixed-gopher	host	100	0.2239	0.0722	379	1.3600	0.1170	0.0658	0.0060	0.6567	0.0205
	symb		0.5245	0.0802	379	1.5800	0.1343	0.0707	0.0059		
mixed-stinkbug	host	100	0.2694	0.0904	1745	0.5000	0.0714	0.0000	0.0000	0.7818	0.0120
	symb		0.0689	0.0470	1583	1.7200	0.0859	0.1644	0.0095		
mixed-primate	host	100	0.2752	0.0980	696	3.7600	0.1728	0.0568	0.0026	0.8136	0.0079
	symb		0.3877	0.1266	425	2.0400	0.1427	0.0270	0.0025		
mixed-damselfly	host	100	0.1726	0.0680	1051	0.6600	0.1054	0.0195	0.0028	0.7660	0.0183
	symb		0.1320	0.0622	2029	0.5600	0.0765	0.0865	0.0023		
mixed-bird	host	100	0.1749	0.0584	5000	2.3000	0.2199	0.0063	0.0006	0.9410	0.0045
	symb		0.3520	0.0533	5000	0.0400	0.0280	0.0002	0.0002		
mixed-moth	host	100	0.1024	0.0631	1404	14.4600	0.3159	0.2058	0.0037	0.5718	0.0165
	symb		0.0250	0.0143	4326	4.8800	0.1993	0.1224	0.0039		
forward-gopher	host	100	0.4683	0.1994	300	1.9800	0.1443	0.1413	0.0108	0.7415	0.0132
	symb		0.4524	0.1521	300	1.8000	0.1370	0.1209	0.0082		
forward-stinkbug	host	100	0.2610	0.1129	1000	0.1600	0.0524	0.0800	0.0273	0.8525	0.0051
	symb		0.4354	0.1901	1000	0.2000	0.0700	0.0150	0.0048		
forward-primate	host	100	0.4652	0.0796	400	5.1000	0.2025	0.1284	0.0046	0.5120	0.0037
	symb		0.5579	0.0923	400	8.3200	0.2625	0.1307	0.0034		
forward-damselfly	host	100	0.4992	0.1228	1000	1.4600	0.1463	0.0760	0.0058	0.4042	0.0063
	symb		0.5170	0.1124	1000	1.0600	0.1157	0.0380	0.0040		
forward-bird	host	100	0.6019	0.1084	5000	1.9200	0.1240	0.0633	0.0033	0.6938	0.0033
	symb		0.6595	0.1103	5000	3.8200	0.1800	0.0698	0.0030		
forward-moth	host	100	0.5724	0.0941	3000	4.2800	0.1960	0.0519	0.0023	0.5496	0.0029
	symb		0.6239	0.0948	3000	5.0000	0.2020	0.0474	0.0021		

TABLE S2: Summary statistics for mixed and forward simulations' bootstrap resampled results. (n=100 experimental replicates, m=100 bootstrap resamples in each replicate).

			bootstrapped aln		aln	Clades nRF bootstrap trees		Discordance bootstrap trees		bootstrap vs annotation trees	
Model conditions	Taxa	reps	ANHD avg	ANHD SE	length	nRF avg	nRF SE	nRF avg	nRF SE	nRF avg	nRF SE
mixed-gopher	host	100	0.2164	0.0003	379	0.1315	0.0074	0.1125	0.0007	0.1275	0.0084
	symb		0.5202	0.0006	379	0.0733	0.0068	0.1407	0.0008	0.1393	0.0076
mixed-stinkbug	host	100	0.2671	0.0002	1745	0.0960	0.0100	0.0000	0.0000	0.0000	0.0000
	symb		0.0720	0.0002	1583	0.1620	0.0081	0.2133	0.0008	0.1967	0.0088
mixed-primate	host	100	0.2787	0.0004	696	0.0808	0.0027	0.1107	0.0004	0.0968	0.0040
	symb		0.3877	0.0001	425	0.0746	0.0035	0.0772	0.0003	0.0675	0.0042
mixed-damselfly	host	100	0.1721	0.0004	1051	0.0977	0.0053	0.0809	0.0005	0.0500	0.0045
	symb		0.1313	0.0004	2029	0.1071	0.0035	0.0374	0.0002	0.0860	0.0027
mixed-bird	host	100	0.1754	0.0002	5000	0.0209	0.0011	0.0329	0.0001	0.0216	0.0012
	symb		0.3517	0.0007	5000	0.0033	0.0007	0.0033	0.0001	0.0007	0.0004
mixed-moth	host	100	0.1042	0.0004	1404	0.2524	0.0034	0.2882	0.0004	0.2639	0.0039
	symb		0.0244	0.0013	4326	0.1356	0.0040	0.1522	0.0005	0.1568	0.0046
forward-gopher	host	100	0.4678	0.0002	300	0.2656	0.0148	0.3221	0.0015	0.2238	0.0157
	symb		0.4507	0.0005	300	0.2350	0.0112	0.2382	0.0012	0.1964	0.0135
forward-stinkbug	host	100	0.2587	0.0009	1000	0.1800	0.0241	0.2533	0.0036	0.1300	0.0338
	symb		0.4387	0.0002	1000	0.1671	0.0068	0.0600	0.0008	0.0517	0.0094
forward-primate	host	100	0.4641	0.0007	400	0.2092	0.0049	0.2072	0.0006	0.2068	0.0061
	symb		0.5562	0.0006	400	0.1567	0.0035	0.1951	0.0005	0.1900	0.0041
forward-damselfly	host	100	0.5013	0.0007	1000	0.1375	0.0063	0.1214	0.0006	0.1300	0.0082
	symb		0.5153	0.0005	1000	0.0638	0.0043	0.0762	0.0004	0.0647	0.0054
forward-bird	host	100	0.6008	0.0003	5000	0.1496	0.0023	0.0517	0.0003	0.0750	0.0037
	symb		0.6590	0.0007	5000	0.0895	0.0022	0.0750	0.0003	0.0834	0.0030
forward-moth	host	100	0.5712	0.0007	3000	0.0711	0.0018	0.0618	0.0002	0.0784	0.0028
	symb		0.6250	0.0005	3000	0.0888	0.0020	0.0772	0.0002	0.0682	0.0023

TABLE S3: Summary statistics for mixed and forward simulations for evolutionary divergence experiment when h = 0.5. (n=100 experimental replicates, m=100 bootstrap resamples in each replicate).

h = 0.5

			annotation aln		aln	annotation vs reference tree nRF		Cophylogenetic precisionn	
Model conditions	Taxa	reps	ANHD avg	ANHD SE	length	nRF avg	nRF SE	prec avg	prec SE
mixed-gopher	host	100	0.1309	0.0477	379	0.0958	0.0080	0.6915	0.0169
	symb		0.4259	0.0844	379	0.0671	0.0052		
forward-gopher	host	100	0.3674	0.1810	300	0.1338	0.0117	0.7692	0.0140
	symb		0.3548	0.1432	300	0.1209	0.0082		

TABLE S4: Summary statistics for mixed and forward simulations for evolutionary divergence experiment when h = 2. (n=100 experimental replicates, m=100 bootstrap resamples in each replicate).

h = 2

			annotation aln		aln	annotation vs reference tree nRF		Cophylogenetic precision	
Model conditions	Taxa	reps	ANHD avg	ANHD SE	length	nRF avg	nRF SE	prec avg	prec SE
mixed-gopher	host	100	0.3479	0.0918	379	0.0633	0.0069	0.6348	0.0215
	symb		0.6057	0.0741	379	0.1007	0.0075		
forward-gopher	host	100	0.5532	0.1965	300	0.2400	0.0181	0.6271	0.0188
	symb		0.5397	0.1458	300	0.1227	0.0083		

S4. Commands to run external software in experiments

Note that texts inside curly brackets {} indicate files and inputs the user passes into the software, thus they are not part of the command.

Seq-Gen v1.3.4 [3] was used to simulate gap-less alignments under model species trees from parameters obtained from running RAxML v8.2.12 [4] on the original empirical alignments. This step was exclusive to simulation experiments.

```
seq-gen -mGTR -r{GTR rate parameters} -z {random number} -or
    -l{simulated alignment length} -f{nucleotide frequencies}
    < {model species tree file} > {simulated alignment file}
```

RAxML version 8.2.12 [4] was used to bootstrap alignments.

```
raxmlHPC -f j -b {random number} -# {number of samples} -m GTRGAMMA
-s {alignment} -n {out file suffix}
```

eMPRess v1.2.1 [5] with default event costs was used to estimate cophylogenies from bootstrapped trees.

```
python empress_cli.py reconcile {host tree file} {symbiont tree file}
{extant species associations} --csv {out file name}.csv
```

RAxML version 8.2.12 [4] was used to midpoint root the phylogenies.

```
raxmlHPC -f I -m GTRCAT -t {unrooted tree} -n {rooted tree file suffix}
-p {random number}
```

RAxML version 8.2.12 [4] was used to reconstruct phylogenies under the GTR model with Γ rates distribution.

```
raxmlHPC -m GTRGAMMA -s {alignment file} -p {random number} -n {tree file suffix}
```

S5. Modified eMPRess code

We modified eMPRess graphical user interface (GUI) to output the corresponding cophylogeny in text format. This modification was made because we needed a cophylogeny visualize and a text format cophylogeny from the same execution. The text format cophylogeny would be used for placing bootstrap support on its coevolutionary events. The modified file is available at https://gitlab.msu.edu/liulab/cophylogenetic-bootstrap-support-data-and-scripts/-/tree/main/Experimental_data_scripts/modified_empress_file.

S6. Modified Treeducken Code

We made modifications to Treeducken v1.0.0 [6] source code to produce a comprehensive coevolutionary history (from time point 0 to the end of simulation time) and made adjustments to what cophylogenetic events are stored on a case-by-case basis. The modified files are available at https://gitlab.msu.edu/liulab/cophylogenetic-bootstrap-support-data-and-scripts/-/tree/main/Experimental_data_scripts/modified_treeducken_files.

S7. Postprocessing scripts

Postprocessing scripts are available at https://gitlab.msu.edu/liulab/cophylogenetic-bootstrap-support-data-and-scripts/-/tree/main/Experimental_data_scripts/scripts.

References

[1] C. Bongrand, S. Moriano-Gutierrez, P. Arevalo, M. McFall-Ngai, K. L. Visick, M. Polz, and E. G. Ruby, "Using colonization assays and comparative genomics to discover symbiosis behaviors and factors in *Vibrio fischeri*," *mBio*, vol. 11, no. 2, pp. e03 407–19, 2020.

[2] I. Letunic and P. Bork, "Interactive tree of life (itol) v6: recent updates to the phylogenetic tree display and annotation tool," *Nucleic Acids Research*, p. gkae268, 2024.

[3] A. Rambaut and N. C. Grass, "Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees," *Bioinformatics*, vol. 13, no. 3, pp. 235–238, 1997.

[4] 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, publisher: Oxford University Press.

[5] S. Santichaivekin, Q. Yang, J. Liu, R. Mawhorter, J. Jiang, T. Wesley, Y.-C. Wu, and R. Libeskind-Hadas, "eMPRess: a systematic cophylogeny reconciliation tool," *Bioinformatics*, vol. 37, no. 16, pp. 2481–2482, 2021.

[6] W. Dismukes and T. A. Heath, "treeducken: An R package for simulating cophylogenetic systems," *Methods Ecol. Evol.*, vol. 12, no. 8, pp. 1358–1364, 2021.