Comparing combinatorial models of interacting evolutionary histories*

Gabriele Di Palma[1,2], Catherine Matias[3,4,5] and Blerina Sinaimeri[2]

[1]Università Campus Bio-Medico di Roma, Rome, Italy
[2]LUISS University, Rome, Italy
[3]Sorbonne Université, Paris, France
[4]Université Paris Cité, Paris, France
[5]CNRS, Laboratoire de Probabilités, Statistique et Modélisation, LPSM, F-75005 Paris, France

Abstract

Cophylogenetic datasets consist of two rooted evolutionary trees together with a bipartite graph describing associations between their leaves. From a combinatorial perspective, generators of such datasets can be viewed as random models for structured objects, in the same spirit as classical random graph models. A natural question is therefore which structural properties are induced by a given generative mechanism, and whether different generators calibrated to comparable evolutionary regimes produce instances with comparable typical structure. This question is particularly relevant because synthetic datasets are widely used as benchmarks for evaluating computational methods, and generator-specific structural properties may therefore affect the instances on which algorithms are tested. We study four representative cophylogenetic generators under cospeciation-dominated, host-switch-dominated, and mixed regimes, comparing their outputs through event frequencies, tree-size and tree-balance measures, and association-network descriptors. Our results show that substantial generator-specific differences remain after calibration. Thus, different generators explore different structural regions of the instance space, making the choice of generator crucial when generating synthetic data for algorithm evaluation.

Keywords

cophylogeny, synthetic benchmarks, host-symbiont systems, coevolution, experimental algorithmics

1. Introduction

Cophylogeny studies the evolutionary histories of two sets of interacting entities, typically hosts and their symbionts. These histories are represented by two rooted trees: a host tree H and a symbiont tree S, whose leaves (denoted by $L(H)$ and $L(S)$, respectively) represent present-day species. A set $A \subseteq L(H) \times L(S)$ records which present-day symbionts are associated with which hosts. A *reconciliation* explains the symbiont tree relative to the host tree through four main events: cospeciation, when host and symbiont speciate together; duplication, when the symbiont speciates independently of the host; host switching, when a symbiont lineage moves to another host lineage; and loss, when a symbiont lineage is absent from a descendant host lineage [2, 3, 4, 5, 6, 7]. Since real host-symbiont datasets are limited in number and vary widely in size, sampling effort and biological context, synthetic data are widely used to evaluate reconciliation algorithms and data-driven pipelines.

This creates a benchmarking problem. A synthetic cophylogenetic instance is not only a pair of trees with associations. It is also the observable output of a particular generative model. Such models differ in whether they simulate the host tree independently or jointly with the symbiont tree, whether they use explicit time, how they implement switches, whether they allow multiple associations, and how extant associations are obtained from latent evolutionary histories. These choices are algorithmically relevant: they determine the size and shape of the generated trees, the degree distribution of the bipartite association graph, and the part of the instance space that an algorithm actually sees during evaluation.

ICTCS 2026: Italian Conference on Theoretical Computer Science, September 7–9, 2026, Udine, Italy
*This paper is a shortened version of work that appeared in RecombCG 2026 [1].
dipalmag@luiss.it (G. D. Palma)
0009-0006-0928-5892 (G. D. Palma); 0000-0001-6665-2421 (C. Matias); 0000-0002-9797-7592 (B. Sinaimeri)

Table 1
Combinatorial and simulation choices that affect the generated benchmark instances. "Cond." means that the symbiont history is generated conditional on a host tree. "MA" denotes multiple associations.

Generator	H/S gen.	Time	Associations / switching
`Coala`	input / cond.	no	one host per symbiont; switch creates a new lineage on another host
`treeducken`	joint / joint	yes	MA; switch/spread changes or expands the host repertoire
`cophylo`	input / cond.	yes	MA; host shift may add a host without symbiont speciation
`AsymmeTree`	birth-death / cond.	yes	no MA; horizontal transfer acts as switching

Table 2
Regime-specific parameter settings. Only parameters varying across R_1-R_3 are shown; parameters fixed across regimes are given in the text.

Generator	R_1	R_2	R_3
`Coala`	$\mu_H = .63$; $p_c \sim U(.7, .9)$; $p_s \sim U(0, .05)$	$\mu_H = .24$; $p_c \sim U(.05, .1)$; $p_s \sim U(.5, .7)$	$\mu_H = .45$; $p_c, p_s \sim U(.2, .4)$
`treeducken`	$\lambda_c \sim U(2, 2.5)$; $\chi \sim U(0, .05)$	$\lambda_c \sim U(0, .05)$; $\chi \sim U(2, 2.5)$	$\lambda_c, \chi \sim U(1.5, 1.8)$
`cophylo`	$\mu_H = .63$; $\mu_S = .645$; $c \sim U(.7, .9)$; $s \sim U(0, .05)$	$\mu_H = .24$; $\mu_S = .66$; $c \sim U(.05, .1)$; $s \sim U(.5, .7)$	$\mu_H = .45$; $\mu_S = .60$; $c, s \sim U(.2, .4)$
`AsymmeTree`	$\mu_H = .63$; $\lambda_L = .645$; $\lambda_{HGT} \sim U(0, .05)$	$\mu_H = .24$; $\lambda_L = .66$; $\lambda_{HGT} \sim U(2, 2.5)$	$\mu_H = .45$; $\lambda_L = .60$; $\lambda_{HGT} \sim U(1.5, 1.8)$

From a combinatorial perspective, each generator defines a random model over structured instances (H, S, A), with its outputs restricted to a generator-specific region of the instance space. We ask whether generators calibrated to comparable evolutionary regimes induce comparable typical structures. If not, generator choice becomes very relevant when generating synthetic data for algorithm evaluation. We compare `Coala` [8], `treeducken` [9], the generator associated with Alcala et al. [10], here called `cophylo`, and `AsymmeTree` [11] under three regimes: high cospeciation, high host switching, and mixed coevolution. We also compare the generated instances with 37 real host-symbiont datasets.

2. Models, descriptors and experimental setup

Table 1 summarizes the main modelling differences relevant for benchmark design. In particular, `Coala` and `AsymmeTree` do not produce multiple associations in the configurations considered here, while `treeducken` and `cophylo` can. Similarly, `treeducken` jointly simulates host and symbiont evolution, whereas the other tools condition symbiont evolution on an independently generated host tree.

The comparison used three families of descriptors. *Coevolution-based* descriptors record the relative frequencies of cospeciation and switching events; these were used for calibration. *Tree-based* descriptors record size, height, host-symbiont leaf difference, and normalized versions of the Cherry, Sackin and Colless indices [12, 13]. *Association-based* descriptors are computed on the bipartite graph A and include density, degree assortativity [14], and the fraction of host hotspots, i.e., hosts whose degree is above the host-degree mean plus one standard deviation. These descriptors are simple, but they expose whether generators produce comparable algorithmic inputs after calibration. They are not intended to provide a complete characterization of the cophylogenetic instance generated; rather, they are used to identify properties that remain characteristic of the generator after calibration on evolutionary-event frequencies.

We considered three main coevolutionary scenarios: a cospeciation-dominated regime R_1, a host-switch-dominated regime R_2, and a mixed regime R_3. To make the intended scenarios explicit, we formalize them as regions in the plane of observed event frequencies. For each generated dataset, let $\hat{c}$ and $\hat{s}$ denote the observed proportions of cospeciation and host-switch events, respectively. Since both

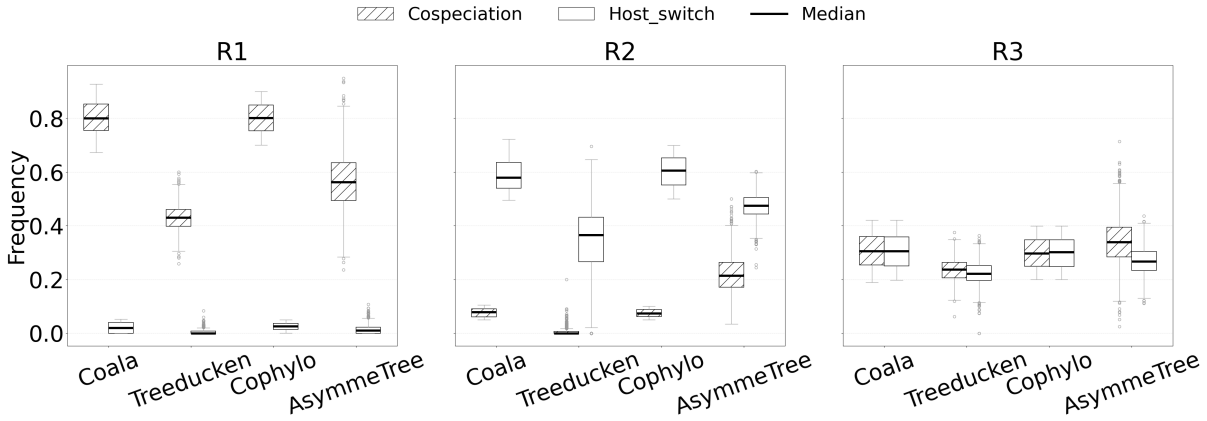

Figure 1: Boxplots of cospeciation and host-switch frequencies across generators in regimes R1–R3, shown after parameter tuning. The figure indicates how closely each generator attains the intended regime.

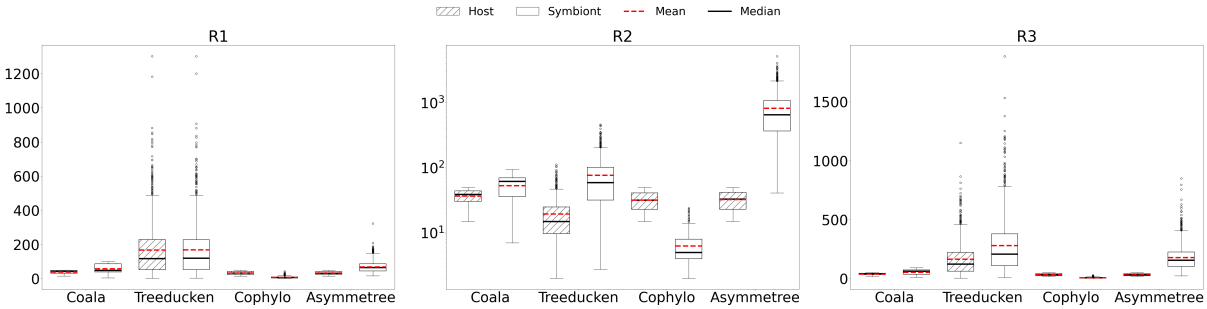

Figure 2: Boxplots of the number of leaves for each regime R1-R3

are computed with respect to the same total number of events, $\hat{c} + \hat{s} \leq 1$. We define R_1 by $\hat{c} \geq 0.50$ and $\hat{s} \leq 0.10$, R_2 by $\hat{c} \leq 0.25$ and $\hat{s} \geq 0.40$, and R_3 by $0.20 \leq \hat{c} \leq 0.40$ and $0.20 \leq \hat{s} \leq 0.40$.

Because the generators use different parameterizations, their input parameters cannot be matched directly. We therefore used generator-specific settings designed to target the same regions of the observed $(\hat{c}, \hat{s})$ plane; the corresponding settings are reported in Table 2. Parameters fixed across regimes are as follows. For `Coala`, $\lambda_H = 0.7$, the number of extant host leaves is sampled in $[15, 50]$, and (p_d, p_l) is drawn from the remaining probability mass according to Dirichlet$(1, 1)$. For `treeducken`, $\lambda_H = \lambda_S = 0.7$, $\mu_H \sim U(0.3, 0.8)(\lambda_H + \lambda_c)$, $\mu_S \sim U(0.3, 0.8)(\lambda_S + \lambda_c + \chi)$, $T = 2$, and `hs-mode="switch"`. For `cophylo`, $\lambda_H = \lambda_S = 0.7$, the number of extant host leaves is sampled in $[15, 50]$, $f_d(k) \propto k^{-1.6}$, and $T = T_{\text{MRCA}} = 2$. For `AsymmeTree`, $\lambda_H = \lambda_D = 0.7$ and $T = 2$.

For each generator and regime, we generated $N = 1000$ independent datasets.

3. Results

The generators do not reproduce the three regimes equally well (Fig. 1). `Coala` and `cophylo` are generally closer to the intended cospeciation and host-switch frequencies, whereas `treeducken` and `AsymmeTree` show larger variability, especially in R_1 and R_2. The differences are smaller in the mixed regime R_3.

Strong generator effects are also visible in tree size (Fig. 2). In particular, host and symbiont size distributions are relatively similar for `treeducken`, whereas larger differences arise for `AsymmeTree`. The tree-balance descriptors (Cherry, Sackin and Colless indices) show the same qualitative pattern: generator-specific differences remain after event-frequency calibration. Thus, comparable evolutionary regimes do not imply comparable tree structures.

Association structure provides an additional source of generator-specific variation. Here we report

Table 3
Association-network density, reported as mean/median (standard deviation), for the three regimes.

Generator	R_1	R_2	R_3
`Coala`	0.032/0.028 (0.014)	0.041/0.032 (0.018)	0.041/0.033 (0.019)
`treeducken`	0.039/0.016 (0.067)	0.030/0.025 (0.019)	0.033/0.018 (0.054)
`AsymmeTree`	0.031/0.027 (0.011)	0.022/0.019 (0.008)	0.026/0.023 (0.010)

network density as a representative association-based descriptor (Table 3), while the complete results for density, degree assortativity, and host hotspots are reported in the full study [1]. `cophylo` is excluded from the association-based analysis because its output may contain hosts or symbionts that are absent from the corresponding phylogenetic trees, which can bias network descriptors. `cophylo` is therefore retained only for the tree-based analyses. All three generators produce sparse association networks, but their distributions differ. `Coala` shows relatively stable densities across regimes, whereas `AsymmeTree` produces lower densities in R_2 and R_3. `treeducken` shows the largest variability, particularly in R_1 and R_3, where the mean is substantially larger than the median. This is consistent with the different association mechanisms of the generators: `Coala` assigns exactly one host to each symbiont, `treeducken` allows multiple associations and spread events, whereas `AsymmeTree` does not allow multiple associations in the setting considered here. Thus, even similar average densities can arise from substantially different association structures.

Finally, we compared the synthetic datasets with 37 real host-symbiont datasets in a feature space of tree-size and tree-shape summaries. Features were standardized using the real datasets as reference, and PCA was fitted on the 37 real datasets only; synthetic datasets were then projected using the same transformation (see Figure 3). No generator covers the empirical region completely: `cophylo` overlaps it most strongly but also extends outside it; `treeducken` is the most dispersed; `Coala` covers only part of the region; and `AsymmeTree` is shifted relative to most real datasets. Thus, generator identity remains associated with substantial structural variation even within comparable evolutionary regimes.

4. Discussion and future work

Our results show that cophylogenetic generators are not interchangeable implementations of a single synthetic-data model: they define different random models over structured instances (H, S, A) and retain different structural signatures after event-level calibration. From the viewpoint of experimental algorithmics, the generator and its parameters should therefore be treated as part of the benchmark specification, together with basic summaries of instance size, tree shape, and association structure.

A natural next step is to move from this empirical comparison toward a more analytical study of cophylogenetic generators. For example, how does joint rather than conditional generation affect the relative sizes or shapes of the host and symbiont trees? How do switching and multiple-association mechanisms affect the density or degree distribution of the association graph? Another question is whether two generators can be distinguished only from the datasets they produce. Finally, the comparison with real datasets raises a coverage problem: how should generators and their parameters be combined to cover the empirical instance space as broadly as possible? Answering these questions would help explain why the generators produce different types of instances, rather than only observing that they do.

The comparison with real datasets raises a related question. Different generators cover different parts of the empirical structural region, suggesting that no single model necessarily provides a representative benchmark on its own. From the viewpoint of experimental algorithmics, this is important: algorithms evaluated under the same nominal evolutionary regime may still be tested on structurally different instances depending on the generator used. The generator and its parameters should therefore be considered part of the benchmark specification, together with structural summaries of the generated instances.

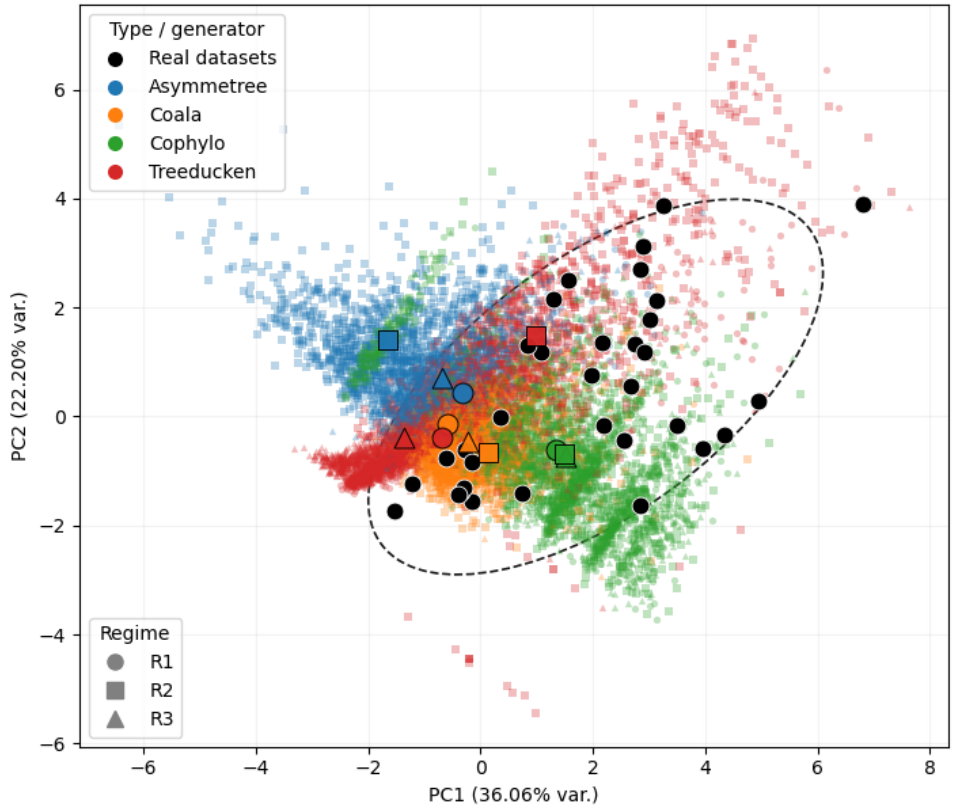

Figure 3: Projection of real (black) and synthetic datasets (colored by generator, shaped by regime) onto PC1–PC2 from standardized feature vectors. Large markers show generator–regime centroids; the dashed ellipse summarizes real-data dispersion.

Code and datasets for the underlying experiments are available at https://github.com/Gabbo240900/synthetic_cophylo.

Declaration on Generative AI

During the preparation of this shortened version, generative AI tools were used for language editing and LaTeX restructuring. The scientific content, experiments, interpretations and responsibility for the final text was totally done by the authors.

Acknowledgments

The authors declare no competing interests. We thank the anonymous reviewers for their constructive comments, which helped improve the manuscript.

References

[1] G. Di Palma, C. Matias, B. Sinaimeri, Similarities, Differences and Biases in Cophylogenetic Models for Host-Symbiont Coevolution, Springer Nature Switzerland, 2026, p. 213–233. URL: http://dx.doi.org/10.1007/978-3-032-26891-4_11. doi:10.1007/978-3-032-26891-4_11.

[2] R. D. Page, Parallel phylogenies: Reconstructing the history of host-parasite assemblages, Cladistics 10 (1994) 155–173. doi:10.1111/j.1096-0031.1994.tb00170.x.

[3] M. Charleston, Recent Results in Cophylogeny Mapping, Elsevier, 2003, pp. 303–330. doi:10.1016/s0065-308x(03)54007-6.

[4] C. Conow, D. Fielder, Y. Ovadia, R. Libeskind-Hadas, Jane: a new tool for the cophylogeny reconstruction problem, Algorithms for Molecular Biology 5 (2010). doi:10.1186/1748-7188-5-16.

[5] B. Donati, C. Baudet, B. Sinaimeri, P. Crescenzi, M. Sagot, Eucalypt: efficient tree reconciliation enumerator, Algorithms for Molecular Biology 10 (2015) 3. doi:https://doi.org/10.1186/s13015-014-0031-3.

[6] Y. Wang, A. Mary, M.-F. Sagot, B. Sinaimeri, Capybara: equivalence class enumeration of cophylogeny event-based reconciliations, Bioinformatics 36 (2020) 4197–4199. doi:10.1093/bioinformatics/btaa498.

[7] T. Calamoneri, B. Sinaimeri, Some Problems Related to the Space of Optimal Tree Reconciliations: (Invited Talk), Springer International Publishing, 2022, p. 3–14. doi:10.1007/978-3-030-96731-4_1.

[8] C. Baudet, B. Donati, B. Sinaimeri, P. Crescenzi, C. Gautier, C. Matias, M.-F. Sagot, Cophylogeny reconstruction via an approximate bayesian computation, Systematic Biology 64 (2014) 416–431. doi:10.1093/sysbio/syu129.

[9] W. Dismukes, T. A. Heath, treeducken: An r package for simulating cophylogenetic systems, Methods in Ecology and Evolution 12 (2021) 1358–1364. doi:10.1111/2041-210x.13625.

[10] N. Alcala, T. Jenkins, P. Christe, S. Vuilleumier, Host shift and cospeciation rate estimation from co-phylogenies, Ecology Letters 20 (2017) 1014–1024. doi:10.1111/ele.12799.

[11] D. Schaller, M. Hellmuth, P. F. Stadler, Asymmetree: A flexible python package for the simulation of complex gene family histories, Software 1 (2022) 276–298. doi:10.3390/software1030013.

[12] M. J. Sackin, “good” and “bad” phenograms, Systematic Biology 21 (1972) 225–226. doi:10.1093/sysbio/21.2.225.

[13] D. H. Colless, E. O. Wiley, Phylogenetics: The theory and practice of phylogenetic systematics., Systematic Zoology 31 (1982) 100. URL: http://dx.doi.org/10.2307/2413420. doi:10.2307/2413420.

[14] M. E. J. Newman, Mixing patterns in networks, Physical Review E 67 (2003). URL: http://dx.doi.org/10.1103/PhysRevE.67.026126. doi:10.1103/physreve.67.026126.