Tractable time tree distributions

Tobia Ochsner[1, 2*], Jonathan Klawitter[1], Alexei J. Drummond[1*],

1 University of Auckland, Auckland, Aotearoa/New Zealand
2 ETH Zurich, Basel, Switzerland

* a.drummond@auckland.ac.nz

Abstract

In Bayesian phylogenetics, Markov Chain Monte Carlo (MCMC) methods generate samples of time trees to capture the posterior distribution of the phylogenetic tree. We explore novel approaches to fit distributions on these time trees samples. In particular, we extend the family of Conditional Clade Distributions (CCD) to model the tree topology and branch lengths simultaneously. Our distributions are designed to capture both central tendencies and dependencies within this high-dimensional and non-Euclidean space. They facilitate a straightforward reconstruction of a summary tree and manage to capture the tails of the time tree distribution. This property extends their utility beyond the robust reconstruction of a single summary tree, enabling more sophisticated downstream analyses like exploring credible sets or evaluating alternative evolutionary hypotheses.

Author summary

Introduction

Markov Chain Monte Carlo (MCMC) methods have become a cornerstone of Bayesian phylogenetics. Given a set of aligned molecular sequences and an evolutionary model (consisting of a tree model, a molecular clock model, a substitution model, and prior distributions on parameters), MCMC algorithms like Beast 2 [1] or MrBayes [2] generate a set of phylogenetic trees sampled from the posterior distribution. This posterior sample usually consists of thousands of trees and contains a wealth of information about the evolutionary relationships encoded in the provided sequences.

For simple continuous or indicator parameters of the phylogenetic model (like substitution or clock rates), analyzing the marginal distributions of the posterior sample is relatively straightforward. Tools like Tracer [3] provide readily available summaries and visualizations of these marginal distributions.

The space of phylogenetic trees presents a significantly greater challenge. The sheer number of possible rooted trees explodes super-exponentially with increasing taxa, making it highly unlikely that the MCMC samples contain the true tree topology [4]. The *curse of dimensionality* [5] diminishes the effectiveness of common sample-based statistical methods, particularly distance-based approaches. We demonstrate in the Supplementary Information that MCMC samples can exhibit strikingly similar pairwise distances, contradicting the expectation that regions of higher probability density should yield samples with closer proximity. Furthermore, the space of possible node times is non-Euclidean, as the times are inherently constrained by the topology [6–9].

These challenges make it difficult to directly analyze the posterior sample of trees and branch lengths. One strategy to extract meaningful information from the sample involves dedicated post-processing techniques. The raw sample is distilled into representations that enable a more meaningful analysis and interpretation.

A common approach is to infer a point estimate, a single "best" tree that summarizes the posterior sample [10]. Several heuristics have been developed for this purpose, including DensiTree visualizations [11], Maximum Clade Credibility (MCC) trees, Hipster [12], and the Maximum A Posteriori tree based on Conditional Clade Distributions (CCD) [4, 13]. These methods often focus primarily on recovering a likely tree topology. The determination of branch lengths is treated as an afterthought, potentially overlooking important information about evolutionary timescales.

Another class of algorithms considers both tree topology and branch lengths simultaneously by embedding trees in a particular metric space. Examples include the Billera-Holmes-Vogtmann (BHV) space [9], the Wald space [6], the t and τ space [7], the palm tree space [8], the RNNI space [14], and the recently developed Wald space [6]. Within these spaces, a natural approach to obtain a point estimate is to calculate the Fréchet mean of the posterior tree sample [15]. However, these spaces often suffer from limitations such as the *sticky mean problem* (where the mean tree is almost independent of the actual tree samples) [16], high computational complexity associated with distance calculations, and the inherent difficulties of performing statistical inference in complex geometric spaces [17].

While obtaining a point estimate from the posterior sample is crucial, it is important to recognize that the MCMC sample embodies information about the entire posterior distribution. Tractable tree distributions [4] aim to capture the posterior distribution using a well-defined approximate distribution. This provides an elegant framework for testing alternative evolutionary hypotheses, determining whether a specific tree falls within a credible set, calculating statistics like information content [18], and analyzing dependencies between different parts of the tree (e.g., relationships between branch lengths and topological features). Existing methods that estimate distributions on the posterior sample for rooted trees focus on modeling tree topologies [4, 13], ignoring time information entirely. The family of Bayesian subsplit networks [19] has been extended to model branch lengths in multiple ways [20, 21]. However, Bayesian subsplit methods focus on direct variational inference, rather than posterior sample summarization.

This paper addresses the challenge of inferring tractable distributions on time trees, encompassing both topology and branch lengths. We compare different methods and tree embeddings, and demonstrate that the described methods exhibit significant differences in terms of capturing central tendencies and the tails of the distribution. Finally, we introduce a software package that enables researchers to fit and analyze these distributions, facilitating more nuanced insight into phylogenetic uncertainty. This tool allows for a more complete utilization of the information contained within the posterior tree sample, moving beyond simple point estimates.

Materials and methods

In this section, we introduce the concept of tractable tree distributions on time trees and discuss desirable properties of such distributions. We then introduce a family of distributions satisfying these properties and show how to perform common operations such as parameter estimation and obtaining a point estimate. Finally, we outline the procedure used to assess and compare the distributions on both synthetic and real datasets.

Tractable tree distributions

Berling et al. [4] introduced *tractable tree distributions* as a set of properties of a distribution on tree topologies. A distribution is tractable if a range of operations can be performed efficiently in practice. We apply the same concept to distributions on *time trees*.

Tractable time tree distributions are used as a tool to analyze a posterior tree sample. Therefore, we should be able to efficiently infer its parameters based on that tree sample. In order to compare different tree hypotheses, we need to evaluate the probability of any time tree under the model. Another important operation is sampling, as it enables to test if a tree is in the 95% credible set and to inspect the dependencies of different node heights. Lastly, a tractable time tree distribution should provide us with a *summary tree*, for instance by retrieving the highest-probability or expected tree.

Mathematical framework

A *time tree* $(T, \{H_v\}_{v \in V})$ for a set of taxa X and root R consists of a undirected tree $T = (V, E)$ and a set of random variables $\{H_v\}_{v \in V} \in \mathbb{R}^+$ denoting the vertex heights (the divergence times). The following properties have to be satisfied:

1. (V, E) describe a fully connected tree.
2. The root $R \in V$ has degree 2. The taxa $X \subset V$ have degree 1. All other vertices are *internal* vertices and have degree 3.
3. We only consider contemporary sampling of the taxa: $H_x = 0$ for all taxa x.
4. Each edge e in E consists of a parent p and child c such that the parent is closer to the root. The heights must describe a valid tree and evolution runs backward in time: $H_c < H_p$ for all edges (p, c) in E.

The probability of a time tree can be factorized into the probability of its topology T and the probability of the vertex heights $\{H_v\}$ given the topology:

$$P\left(T, \{H_v\}_{v \in V}\right) = P\left(T\right) P\left(\{H_v\}_{v \in V} | T\right).$$

To model $P\left(T\right)$, we use the conditional clade distributions *CCD0* and *CCD1* [4, 13]. They already exhibit the desired properties and have shown to work well in practice [4, 12, 22]. Thus, we focus on modelling the conditional probability $P\left(\{H_v\}_{v \in V} | T\right)$. It is crucial to note that we do not restrict ourselves to a single topology T—we simply model $P\left(\{H_v\}_{v \in V} | T\right)$ as a function of the topology T.

Clade parameterisation

There are two main challenges when designing a distribution on the heights:

First, $P\left(\{H_v\}_{v \in V} | T\right)$ is a function of the topology T. However, we cannot assume that every possible topology is found in the MCMC sample. Hence, $P\left(\{H_v\}_{v \in V} | T\right)$ needs to be defined even for unknown topologies. Moreover, as the number of possible topologies can be extremly large, we have to share parameters amongst different topologies. In order to overcome this challenge, we employ the same strategy as the CCD0 model [4]: instead of having random variables corresponding to vertices of a topology, they correspond to clades:

$$P\left(\{H_v\}_{v \in V} | T\right) = P\left(\{H_c\}_{c \in C(T)} | T\right),$$

where $C(T)$ is the set of clades in topology T. A clade is a subset of taxa that share a common ancestor. While every vertex corresponds to exactly one clade, modeling random variables on a clade-level allows to share parameters amongst different topologies that have common clades. We call a clade *non-trivial* if it is neither a singleton nor the root clade. In the following, we exclude the singleton clades corresponding to the taxa from $C(T)$, as H_c equals 0 for all singleton clades and does not need to be modelled.

Another difficulty is the fact that every topology enforces a different set of constraints on the heights. We expect $P\left(\{H_c\}_{c\in C(T)}|T\right)$ to vanish for heights invalid under the topology. We solve this by introducing the concept of a *clade embedding*. Given a topology T, a clade embedding f_T induces a bijection from the heights into another set of random variables in the *embedding space*:

$$f_T : \{H_c\}_{c\in C(T)} \mapsto \{Z_c\}_{c\in C(T)}.$$

The embedding can be chosen such that the Z_c have nice constraints independent of the topology (for instance, $Z_c \in \mathbb{R}$ or $Z_c \in [0,1]$). Equipped with an embedding, we can define a probability distribution on $\{Z_c\}_{c\in C(T)}$ as well as the corresponding pulled-back probability distribution on $\{H_c\}_{c\in C(T)}$:

$$P\left(\{H_c\}_{c\in C(T)}|T\right) = P\left(\{Z_c\}_{c\in C(T)}|T\right)\left|\det\left(J_{f_t^{-1}}(\{Z_c\}_{c\in C(T)})\right)\right|,$$

where $J_{f_t^{-1}}(\{Z_c\}_{c\in C(T)})$ is the Jacobian of the inverse of the embedding.

In order to get the probability of a set of heights under a topology, we first transform the set of heights into the embedding space. Then, we can use the pulled-back probability distribution to calculate the probability. If we want to sample from the distribution, we can sample from the embedding space and transform the samples back into the original height space.

In the following sections, we introduce three different clade embeddings equipped with different strategies for assigning probabilities in the embedding space. Our general approach is to find the best approximation of the marginal time tree posterior distribution of the MCMC model. Thus, biological interpretation of the designed distributions is not a primary goal. Instead, we focus on a wide range of embeddings and distributions that fulfill the properties of tractable time tree distributions and then compare their fit with the MCMC posterior distribution on a variety of real and simulated datasets.

Last divergence embedding

Under the birth death model, internal branch lengths are assumed to be independent realizations of a stochastic evolutionary process. In contrast, pendant branch lengths are primarily influenced by the time of sampling. We propose to model internal branch lengths using independent distributions, while separatly accounting for the time between sampling and the last divergence event. This approach aligns with theoretical findings suggesting that internal and pendant branch lengths follow notably different distributions [23, 24].

Specifically, we consider the branch length leading to a vertex v corresponding to a non-trivial clade c:

$$Z_c = H_{p_v} - H_v.$$

Furthermore, we consider the time between the present and the last divergence event and assign it to the random variable corresponding to the root clade X:

$$Z_X = \min_{x \in X} H_x.$$

This embedding allows to choose arbitrary positive Z_c while always producing a valid tree. Thus, we propose to use either independent lognormal, independent gamma, or independent Weibull distributions:

$$P(\{Z_c\}_{c \in C(T)}) = \begin{cases} \prod_{c \in C(T)} LogNormal(Z_c; \mu_c, \sigma_c), \\ \prod_{c \in C(T)} Gamma(Z_c; \alpha_c, \beta_c), \\ \prod_{c \in C(T)} Weibull(Z_c; k_c, \lambda_c). \end{cases}$$

Different models for divergence suggest different distributions on the branch lengths [25]. For instance, a constant rate of divergence results in an exponential distribution, which is a special case of the gamma distribution. If the rate of divergence changes with the amount of time since the last divergence, the resulting distribution might be closer to a Weibull distribution.

Oldest child embedding

We introduce an alternative embedding that directly utilizes absolute branch lengths. For a clade c corresponding to vertex v, we define Z_c as the branch length leading to the older child of v.

$$Z_c = H_v - \max\left(H_{a_v}, H_{b_v}\right),$$

where a_v and b_v represent the two child vertices of v. This parameterisation ensures that any set of positive Z_c, given a topology, results in a valid contemporary sampled time tree. To assign probabilities to these branch lengths, we consider same three probabilistic models:

$$P(\{Z_c\}_{c \in C(T)}) = \begin{cases} \prod_{c \in C(T)} LogNormal(Z_c; \mu_c, \sigma_c), \\ \prod_{c \in C(T)} Gamma(Z_c; \alpha_c, \beta_c), \\ \prod_{c \in C(T)} Weibull(Z_c; k_c, \lambda_c). \end{cases}$$

Height ratio embedding

Inspired by TreeFlow [26], we introduce the height ratio embedding to capture the relative temporal structure of the vertices in time trees. For any non-trivial clade c correspoding to vertex v, we define Z_c as the ratio of the height of v to the height of its parent vertex p_v.

$$Z_c = \frac{H_v}{H_{p_v}}.$$

In addition to the relative positions of the vertices, we model the overall height of the time tree using the random variable corresponding to the root clade X:

$$Z_X = H_R.$$

Consequently, Z_X takes values in $\mathbb{R}^+$. For any non-trivial clade c, Z_c is constrained to the interval $[0, 1]$. The primary advantage of the height ratio embedding lies in its ability to mitigate the effects of dependencies introduced by contemporary sampling.

We model the ratios and the tree height independently. For the height ratios, we consider the following probability distributions:

$$P(\{Z_c\}_{c\in C(T)\setminus X}) = \begin{cases} \prod_{c\in C(T)\setminus X} Beta(Z_c;\alpha_c,\beta_c), \\ \prod_{c\in C(T)\setminus X} LogitNormal(Z_c;\mu_c,\sigma_c), \\ TreeDirichlet(Z_c;\{\alpha_c\}_{c\in C(T)\setminus X}). \end{cases}$$

The *TreeDirichlet* distribution is a generalized Dirichlet distribution, parameterized by a single parameter for each non-trivial clade. It is a special case of the independent beta distributions, defined such that the marginal distribution of any path from the root vertex to a leaf follows a Dirichlet distribution. A formal definition is provided in the supporting information.

We model the overall tree height with a lognormal distribution:

$$P(Z_X) = \text{LogNormal}(Z_X;\mu_X,\sigma_X).$$

Operations on tractable time tree distributions

We use the maximum likelihood estimator to estimate the parameters for each embedding and the corresponding distributions. The stated independence assumptions significantly simplify this process, reducing it to the maximum likelihood estimation of basic univariate continuous distributions. For the lognormal and logitnormal distribution, the maximum likelihood estimators exhibit well-known closed form solutions. For the gamma, beta and weibull distributions, we apply Newton's method as implemented in the julia package *Distributions.jl* [27]. The maximum likelihood estimator for the TreeDirichlet distribution is more involved and is detailed in the supporting information.

In order to calculate the probability for a given time tree, we first transform the set of heights into the embedding space. Then, we can use the pulled-back probability distribution to calculate the probability.

Sampling from the distribution is achieved by drawing samples from the embedding space and mapping them back to the original height space using the inverse transformation. To assess if a given time tree is within the 95 % credible set, we can generate a large number of samples from the fitted distribution. If the model probability of the given tree ranks within the top 95 % of the sampled tree probabilities, it is in the credible set.

TODO: point estimator

Validation

To evaluate the quality of our proposed tractable time tree distributions, we use two complementary methods: model comparison and goodness-of-fit (GoF) testing.

Given a posterior set of time trees, we calculate the Akaike Information Criterion (AIC) to compare the relative performance of our distributions on a given set of time trees. AIC balances model likelihood with parameter count, providing a ranking of models. However, it does not determine if any model accurately reflects the posterior distribution.

We perform a goodness-of-fit test to assess how well our tractable tree distributions capture the true distribution. We frame this as a two-sample hypothesis test, where we compare the MCMC time trees with a set of time trees sampled from our approximate distribution. Standard two sample GoF tests do not work in the space of time trees due to its high-dimensional, continuous nature and the lack of a natural order. To overcome this, we transform the trees into a univariate continuous space. Specifically, we calculate the probability of each tree under the fitted model, transforming it into the space $[0, 1]$. We then employ a two-sample Anderson-Darling test on the set of probabilities of the

two samples to determine if they are drawn from the same distribution. It is important to acknowledge that this does not prove the equality of the time tree distributions: Distinct tree distributions could, in theory, produce similar distributions of probabilities. However, it does provide a necessary condition for the distributions to be identical.

Moreover, we use probability–probability plots for a visual inspection of the goodness of fit. They work by plotting the empirical cumulative distribution function (ECDF) of the two probability-transformed samples against each other. If the approximate distribution closely resembles the posterior distribution, the resulting plot should be close to a straight line.

Datasets

We conduct a comprehensive comparison of different distributions using both simulated and real datasets. For the simulated datasets, we analyze posterior samples from the CCD study [28], focusing on the Yule tree simulations with 10, 20, 50, 100, 200, and 400 taxa. These samples were generated using LPhyStudio, LPhyBEAST [29], and BEAST2 [1]. We refer to the CCD paper [4] for more details on the simulations.

To assess the performance on biological datasets, we obtained 99 posterior time tree distributions from Dryad [30]. Specifically, we filtered the database for files with a *.trees* extension and manually selected those containing posterior sample trees obtained by BEAST2 or a similar software. An exhaustive list of the studies considered and potential reasons for exclusion are provided in the supporting information.

Results

Discussion

Conclusion

Supporting information

The curse of dimensionality

This section examines the susceptibility of time trees to the *curse of dimensionality* [5], a phenomenon where the complexity of a problem increases exponentially with the number of dimensions. Specifically, we look at the pairwise Robinson–Foulds (RF) distances between the MCMC trees of the synthetic Yule datasets. We find that the distances increase with the number of taxa, while the relative variation of the distances decreases (see figures 1 and 2). This defies the intuition that regions of higher probability density should yield samples with closer proximity and might complicate the application of conventional distance-based methods to analyse the posterior sample. Previous works suggests that the curse of dimensionality might not be limited to simulated trees—the datasets analysed by [31] show a median intrinsic dimension of around 7 when using the RF distance.

Fig 1. The average pairwise Robinson–Foulds distance between the MCMC trees of the synthetic Yule datasets. The distances increase with the number of taxa.

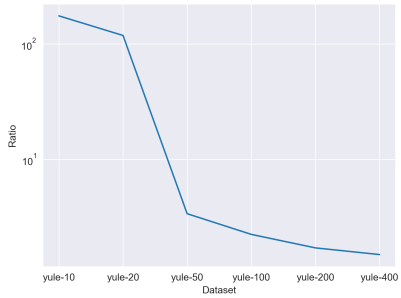

Fig 2. The average ratio of the minimum to maximum pairwise Robinson–Foulds distance between the MCMC trees of the synthetic Yule datasets. The ratios decrease with the number of taxa.

Tree dirichlet distribution

Datasets

References

1. Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH, Xie D, et al. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS computational biology. 2014;10(4):e1003537.
2. Ronquist F, Teslenko M, Van Der Mark P, Ayres DL, Darling A, Höhna S, et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Systematic biology. 2012;61(3):539–542.
3. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. Systematic biology. 2018;67(5):901–904.
4. Berling L, Klawitter J, Bouckaert R, Xie D, Gavryushkin A, Drummond AJ. Accurate Bayesian phylogenetic point estimation using a tree distribution parameterized by clade probabilities. PLOS Computational Biology. 2025;21(2):e1012789.
5. Altman N, Krzywinski M. The curse (s) of dimensionality. Nat Methods. 2018;15(6):399–400.
6. Lueg J, Garba MK, Nye TM, Huckemann SF. Wald space for phylogenetic trees. In: Geometric Science of Information: 5th International Conference, GSI 2021, Paris, France, July 21–23, 2021, Proceedings 5. Springer; 2021. p. 710–717.
7. Gavryushkin A, Drummond AJ. The space of ultrametric phylogenetic trees. Journal of theoretical biology. 2016;403:197–208.
8. Monod A, Lin B, Yoshida R, Kang Q. Tropical geometry of phylogenetic tree space: a statistical perspective. arXiv preprint arXiv:180512400. 2018;.
9. Billera LJ, Holmes SP, Vogtmann K. Geometry of the space of phylogenetic trees. Advances in Applied Mathematics. 2001;27(4):733–767.

10. Heled J, Bouckaert RR. Looking for trees in the forest: summary tree from posterior samples. BMC evolutionary biology. 2013;13:1–11.

11. Bouckaert RR, Heled J. DensiTree 2: Seeing trees through the forest. BioRxiv. 2014; p. 012401.

12. Baele G, Carvalho LM, Brusselmans M, Dudas G, Ji X, McCrone JT, et al. HIPSTR: highest independent posterior subtree reconstruction in TreeAnnotator X. bioRxiv. 2024; p. 2024–12.

13. Larget B. The estimation of tree posterior probabilities using conditional clade probability distributions. Systematic biology. 2013;62(4):501–511.

14. Gavryushkin A, Whidden C, Matsen IV FA. The combinatorics of discrete time-trees: theory and open problems. Journal of mathematical biology. 2018;76(5):1101–1121.

15. Brown DG, Owen M. Mean and variance of phylogenetic trees. Systematic biology. 2020;69(1):139–154.

16. Lammers L, Van DT, Nye TM, Huckemann SF. Types of Stickiness in BHV Phylogenetic Tree Spaces and Their Degree. In: International Conference on Geometric Science of Information. Springer; 2023. p. 357–365.

17. Said S, Mostajeran C, Heuveline S. Gaussian distributions on Riemannian symmetric spaces of nonpositive curvature. In: Handbook of Statistics. vol. 46. Elsevier; 2022. p. 357–400.

18. Lewis PO, Chen MH, Kuo L, Lewis LA, Fučíková K, Neupane S, et al. Estimating Bayesian phylogenetic information content. Systematic biology. 2016;65(6):1009–1023.

19. Zhang C, Matsen IV FA. Generalizing Tree Probability Estimation via Bayesian Networks. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, editors. Advances in Neural Information Processing Systems. vol. 31. Curran Associates, Inc.; 2018.Available from: `https://proceedings.neurips.cc/paper_files/paper/2018/file/b137fdd1f79d56c7edf3365fea7520f2-Paper.pdf`.

20. Zhang C. Improved variational Bayesian phylogenetic inference with normalizing flows. Advances in neural information processing systems. 2020;33:18760–18771.

21. Zhang C, IV FAM. A Variational Approach to Bayesian Phylogenetic Inference. Journal of Machine Learning Research. 2024;25(145):1–56.

22. Berling L, Ochsner T, Bouckaert R, Drummond AJ. CCD0 revisited; 2025. Available from: `https://www.beast2.org/2025/02/25/CCD0-revisited.html`.

23. Paradis E. The distribution of branch lengths in phylogenetic trees. Molecular phylogenetics and evolution. 2016;94:136–145.

24. Mooers A, Gascuel O, Stadler T, Li H, Steel M. Branch lengths on birth–death trees and the expected loss of phylogenetic diversity. Systematic biology. 2012;61(2):195–203.

25. Venditti C, Meade A, Pagel M. Phylogenies reveal new interpretation of speciation and the Red Queen. Nature. 2010;463(7279):349–352.

26. Swanepoel C, Fourment M, Ji X, Nasif H, Suchard MA, Matsen IV FA, et al. TreeFlow: probabilistic programming and automatic differentiation for phylogenetics. arXiv preprint arXiv:221105220. 2022;.

27. Lin D, White JM, Byrne S, Bates D, Noack A, Pearson J, et al.. JuliaStats/Distributions.jl: a Julia package for probability distributions and associated functions; 2019. Available from: https://doi.org/10.5281/zenodo.2647458.

28. Berling L, Klawitter J, Bouckaert R, Xie W, Gavryushkin A, Drummond A. Posterior tree sets for "Accurate Bayesian phylogenetic point estimation using a tree distribution parameterised by clade probabilities"; 2024. Available from: https://auckland.figshare.com/collections/BEAST2_trees_for_Accurate_Bayesian_phylogenetic_point_estimation_using_a_tree_distribution_parameterized_by_clade_probabilities_/7102354/3.

29. Drummond AJ, Chen K, Mendes FK, Xie D. LinguaPhylo: a probabilistic model specification language for reproducible phylogenetic analyses. PLOS Computational Biology. 2023;19(7):e1011226.

30. Dryad;. http://datadryad.org/.

31. Smith MR. Robust analysis of phylogenetic tree space. Systematic Biology. 2022;71(5):1255–1270.