Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 11, n. 1, 2025.

Trabalho apresentado no XLIII CNMAC, Centro de Convenções do Armação Resort - Porto de Galinhas - PE, 2024

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Diagnosing Markov chain Monte Carlo for phylogenetics: theory and methods

Luiz Max Carvalho[1]
Escola de Matemática Aplicada – Fundação Getulio Vargas, Rio de Janeiro, RJ

Abstract. With the advent of Bayesian phylogenetics, Markov chain Monte Carlo (MCMC) methods became the *de facto* standard for sampling from distributions on the space of phylogenetic trees, or treespace. Treespace is vast and does not admit a canonical representation, posing difficulties to the development of not only efficient MCMC schemes but also sensitive diagnostics. In this talk I will detail recent work on the development of validation and diagnostic tools for assessing the output of phylogenetic MCMC. The talk will cover theoretical/combinatorial results on the lumpability of tree-valued processes as well as empirical/methodological work on simulation-based validation of MCMC samplers.

Keywords. coalescent; diagnostics; lumpability; Markov chain Monte Carlo; phylogenetics.

1 Background

In Bayesian phylogenetics one is usually interested in computing the posterior distribution

$$p(t, \boldsymbol{b}, \boldsymbol{\theta}|D) = \frac{f(D|t, \boldsymbol{b}, \boldsymbol{\theta})\pi(t, \boldsymbol{b}, \boldsymbol{\theta})}{\sum_{t_i \in \mathbb{F}} \int_{\boldsymbol{B}} \int_{\boldsymbol{\Theta}} f(D|t_i, \boldsymbol{b}_i, \boldsymbol{\theta})\pi(t_i, \boldsymbol{b}_i, \boldsymbol{\theta}) d\boldsymbol{\theta} d\boldsymbol{b}_i} \,, \tag{1}$$

where D is observed data and $t \in \mathbb{F}$ is a fully-ranked tree topology associated set of branch lengths $\boldsymbol{b}$. Finally $\boldsymbol{\theta}$ is a set of parameters such as substitution model parameters, migration rates, heritability coefficients, etc. The summation in the denominator of (1) is usually taken over a huge set of trees – for $n = 53$ leaves there are $\approx 10^{80}$ possible trees. In many applications, the aim is to construct time-calibrated phylogenies, i.e. phylogenetic trees whose branch lengths are measured in units of calendar time. In particular, one might have sequences sampled through time (heterochronous/serially-sampled) which enable direct estimation of the rate of evolution and reconstruction of past population dynamics [5, 6]. These types of data sets pose additional challenges to inference because they impose constraints[2] on the space of valid trees [9]. A crucial insight is that for phylogenetics one needs to check for convergence in treespace as well, rather than rely on convergence in the space of continuous parameters ($\boldsymbol{\theta}$ and $\boldsymbol{b}$). See Brusselmans et al [4] for a thorough discussion.

2 Assessing correctness

We now discuss both exact and simulation-based diagnostics to ascertain whether a given MCMC algorithm produces (approximately) correct samples from the target p.

[1]luiz.fagundes@fgv.br
[2]More specifically temporal precedence constraints.

2.1 Analytical results

In this section we detail some exact results about phylogenetic distributions that can be used to check correctness of phylogenetic MCMC algorithms. We start by describing the special properties of the so-called exchangeable distributions on trees.

2.1.1 Exchangeable phylogenetic distributions and their binary projections

One way to cope with the vastness of treespace is to project down to lower dimensions. A natural such projection is the so-called clade. A clade is a partition of the set of leaves and two clades $A = A_1|A_2$ and $B = B_1|B_2$ are said to be compatible if at least one of $A_i \cap B_j$, $i, j = 1, 2$ is empty. Let $\boldsymbol{C}_n$ be the space of all possible clades and let $\mathcal{C} : \boldsymbol{T}_n \to \boldsymbol{C}_n$ be a function that outputs the constituent clades of a given tree. Define the indicator $\mathbb{I}_c(T) : \boldsymbol{T}_n \to \{0, 1\}$ such that

$$\mathbb{I}_c(T) = \begin{cases} 1, c \in \mathcal{C}(T) \\ 0, \text{ otherwise,} \end{cases} \tag{2}$$

and let $X_c = \mathbb{I}_c(T)$ for $T \sim F$ and F a distribution on $\boldsymbol{T}_n$. It turns out that for exchangeable distributions on the space of trees (see [1]) one can compute the marginal probability of each clade exactly ([12, 13]).

For clades $A, B \in \boldsymbol{C}_n$, the correlation between the clade indicators X_A and X_B is

$$\rho_n(A, B) = \frac{p_n(A, B) - p_n(A)p_n(B)}{\sqrt{p_n(A)[1 - p_n(A)]p_n(B)[1 - p_n(B)]}}. \tag{3}$$

We can now define the clade correlation matrix, $\boldsymbol{\Xi}_n$, as the matrix whose (i, j) entry is $\rho_n(C_i, C_j)$, $i, j = 1, 2, \ldots, |\boldsymbol{C_n}|$. Code for computing marginal and joint probabilities as well as correlations using Theorem 4.5 in [12] is provided at `https://github.com/maxbiostat/cladeCorrelation`.

We now give a few results concerning the properties of $\boldsymbol{\Xi}_n$. Let $c(n)$ be the proportion of entries in the clade correlation matrix that are **positive**.

Proposition 2.1 (**Minimum and maximum correlation**). *For $n \geq 4$, the minimum and maximum values for $\rho_n(A, B)$ are, respectively*

$$\rho_{\min}(n) = -\frac{2}{3n - 5}, \tag{4}$$

$$\rho_{\max}(n) = \frac{2u(n)k(n) - 4n^2(n-1)}{2n(n-1)\sqrt{\left[\left\lfloor \frac{n}{2} \right\rfloor \left(\left\lfloor \frac{n}{2} \right\rfloor + 1\right) k(n) - 2n\right]\left[\left\lceil \frac{n}{2} \right\rceil \left(\left\lceil \frac{n}{2} \right\rceil + 1\right) k(n) - 2n\right]}}, \tag{5}$$

whence

$$u(n) := \left\lfloor \frac{n}{2} \right\rfloor \left(\left\lfloor \frac{n}{2} \right\rfloor + 1\right) \left\lceil \frac{n}{2} \right\rceil \left(\left\lceil \frac{n}{2} \right\rceil + 1\right) = \begin{cases} \frac{n^2(n+2)^2}{16}, \ n \text{ is even}, \\ \frac{(n-1)(n+1)^2(n+3)}{16}, \ n \text{ is odd}, \end{cases}$$

$$k(n) := \binom{n}{\lfloor \frac{n}{2} \rfloor} = \binom{n}{\lceil \frac{n}{2} \rceil}.$$

Proposition 2.2 (**Sparsity of exchangeable priors**). *The following facts imply that the exchangeable PDA prior induces a "flat" correlation matrix as n grows:*

i) $\lim_{n\to\infty} \rho_{\min}(n) = 0$;

ii) $\lim_{n\to\infty} c(n) = 0$.

Additionally, $\lim_{n\to\infty} \rho_{\max}(n) = 1/4$.

2.1.2 Simulation-based calibration

The results in the last section rely on sampling from a measure that is uniform on $\mathbb{T}_n$, which is unrealistic – even though the results remain valuable as a first correcntess check. To remedy this limitation, one can employ the so-called simulation-based calibration (SBC) [10]. The approach consists of exploiting the fact that if the sampler is correct, simulating data from the hierarchical model implied by the prior and the likelihood will lead to a replicate-averaged posterior (RAP) that should match the target posterior $p(\theta \mid y)$. See Figure 1 and Mendes et al. [8] for more details.

The main idea is as follows

0. Generate a reference tree from the prior $\bar{\tau}_0 \sim \pi(\tau|\boldsymbol{\gamma})$;

   for each iteration in 1:N, **do**:

1. Generate $\bar{\tau} \sim \pi(\tau|\boldsymbol{\gamma})$;
2. Compute the distance $\bar{\delta} = d_\sigma(\bar{\tau}, \bar{\tau}_0)$ according to the metric of choice;
3. Generate some (aligment) data $y \sim p(y|\bar{\tau}, \boldsymbol{\alpha})$;
4. Compute the posterior $\pi(\tau|y)$ and draw $\boldsymbol{\tau}_s = \{\tau_s^{(1)}, \tau_s^{(2)}, \ldots, \tau_s^{(L)}\}$ from it;
5. Compute distances $\boldsymbol{\delta}_s = \{\delta_1, \delta_2, \ldots, \delta_L\}$ with $\delta_i = d_\sigma(\tau_i, \bar{\tau}_0)$;
6. Compute the rank $r(\boldsymbol{\delta}_s, \bar{\delta}) = \sum_{i=1}^{L} \mathbb{I}(\delta_i < \bar{\delta})$.

and then we may check the distribution of the ranks for uniformity [10].

3 Measuring performance

One way in which performance can be measured by estimating the effective sample size. To begin, we consider estimators of univariate ESS, for each clade indicator. We assume the process is a Markov chain consider a reparametrisation of the Markov chain $(X_i)_{i\geq 0}$ in terms of the marginal success probability p and a transition probability α which controls the "flipping rate" of the chain. The autocorrelation in a two-state Markov chain can also be written in closed-form, which enables closed-form computation of the effective sample size (ESS). First, recall that under the reparametrisation considered here, $\mathrm{Cor}(X_i^{(t)}, X_i^{(t+k)}) = (1 - \alpha/p)^k$. Next, consider the expression for the effective sample size for a sample of size M:

$$\begin{aligned} \mathrm{ESS} &= \frac{M}{1 + 2\sum_{t=1}^{\infty} \rho_t}, \\ &= \frac{M}{1 + 2\frac{p-\alpha}{\alpha}}, \\ &= \frac{\alpha}{2p - \alpha} M. \end{aligned} \tag{6}$$

Under independent sampling, we have $\alpha = p$ and thus $\mathrm{ESS} = M$, as expected. See Proposition 4.1 and Figure 2 for a justification of this parametric choice.

Magee et al. [7] point out that trees are fundamentally multivariate objects. Thus, a more appropriate summary might be the multivariate effective sample size (mESS):

$$\mathrm{mESS} = M \left(\frac{\det(\boldsymbol{\Lambda})}{\det(\boldsymbol{\Sigma})} \right)^{1/p}, \tag{7}$$

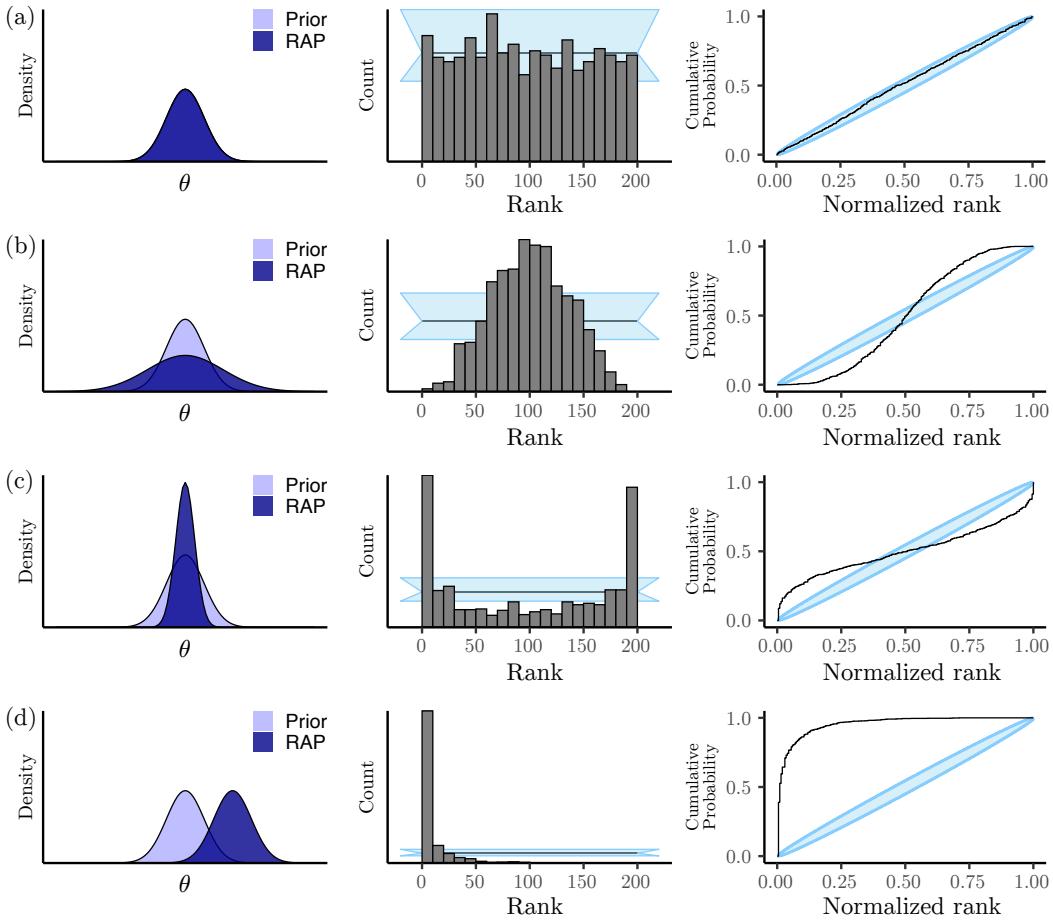

Figura 1: Patterns observable after inference in rank-uniformity validation (RUV). We explain how to interpret the histogram of ranks (middle column) and ECDF plots (right-hand side column) in the main text. (a) Model implementation is correct. (b) Parameter estimates are overdispersed relative to their true values. (c) Parameter estimates are underdispersed relative to their true values. (d) Parameter estimates are consistently overestimated relative to their true values. In the left-hand side column, the prior and replicate-averaged posterior (also known as the data-averaged posterior) distributions over some parameter θ are shown in light blue and dark blue, respectively. In the middle graphs, light-blue bands represent the 95%-confidence interval about the expected rank count, and horizontal black lines mark the rank count mean. Light-blue ellipses in the rightmost graphs represent confidence intervals about the empirical cumulative distribution function (ECDF).

where $\mathbf{\Lambda}$ is the covariance matrix of the binary indicators, which depends on the (pushforward) target distribution and $\mathbf{\Sigma}$ is the long-term covariance matrix, which depends on the sampling efficiency of the chosen algorithm. The advantage of employing exchangeable phylogenetic distributions in this context is that $\mathbf{\Lambda}$ does not need to be estimated from the data but can instead be exactly computed with a straightforward modification of (3). The long-run covariance matrix $\mathbf{\Sigma}$ must be estimated, which is usually done using batch means or the lugsail estimator – see [11].

DOI: 10.5540/03.2025.011.01.0380

4 Tree-valued Markov chains: lazy Metropolis-Hastings

In the talk I will show results from a few simulated and real-world examples, all pertaining to the coalescent family of prior measures (which are a subset of PDA models). Here I will describe only a simple toy model that permits greater control over sampling efficiency on treespace and thus allows one to study the effectiveness of the proposed projection techniques.

For $t \in \mathbb{T}_n$ let $N(t)$ be the set of all trees $u \in \mathbb{T}_n$ which are on subtree prune-and-regraft operation away from t. Define $a(x) := 1 - \sum_{z \in N(x)} \frac{1}{|N(x)|} \min\left\{1, \frac{|N(x)|}{|N(z)|}\right\}$.

$$p_{\mathrm{MH}}(x,y) = \begin{cases} \frac{1}{|N(x)|} \min\left\{1, \frac{|N(x)|}{|N(y)|}\right\}, y \in N(x), \\ a(x), y = x \\ 0, y \notin N(x). \end{cases}$$

The invariant distribution is

$$\pi(t) = 1/|\mathbb{T}_n|, \quad \forall t \in \mathbb{T}_n.$$

We can (artificially) change the mixing of the original MH by adding a probability $\rho \in (0,1)$ of staying in the same place. The new transition matrix is

$$\boldsymbol{P}_{\mathrm{Lazy}}(\rho) = (1-\rho)\boldsymbol{P}_{\mathrm{MH}} + \rho \boldsymbol{I}_{|\mathbb{T}_n|}.$$

This process targets the same invariant distribution as the original MH. Data from a comprehensive simulation study can be found in `https://doi.org/10.5281/zenodo.8168349`

Definition 4.1 (**Lumpability**). *Let $(X_k)_{k\geq 0}$ be Markov chain on a finite state-space $S = \{e_1, e_2, \ldots, e_r\}$ with initial distribution μ_0 and matrix of transition probabilities $\boldsymbol{P}$. We say $(X_k)_{k\geq 0}$ is **lumpable** with respect to a partition of $\bar{S} = \{E_1, E_2, \ldots, E_v\}$ of the state space if a new chain on $\bar{S}$ induced by $(X_k)_{k\geq 0}$, $(Y_k)_{k\geq 0}$, is also a Markov chain for any μ_0.*

Definition 4.2 (**Lumping error and ϵ-lumpability**). *Consider again a partition $\bar{S} = \{E_1, \ldots, E_K\}$ of $\mathcal{S}$. For $x, y \in E_i$, define the **lumping error** as*

$$R(x,y) = \sum_{z \in E_j} p(x,z) - \sum_{z \in E_j} p(y,z). \tag{8}$$

When $|R(x,y)| \leq \epsilon$ for every pair x, y and every E_j, $j \neq i$, we say the Markov chain is ϵ-almost lumpable with respect to $\bar{S}$ [3].

Proposition 4.1 (**Bounds for the lumping error in the Metropolis-Hastings SPR random walk**). *Consider a random walk as in (4). Then, for all $x, y \in S_1(c)$ and for all $n \geq 4$, we have for $|c| \geq 3$.*

$$\varepsilon(S_1(c), S_0(c)) = \frac{h_{n,|c|}}{3n^2 - 2|c|^2 + 2|c|n - 15n + 16} - \frac{5g_{n,|c|}}{6(4(n-2)^2 - 2\sum_{j=1}^{n-2} \lfloor \log_2(j+1) \rfloor)},$$

where

$$\begin{aligned} h_{n,|c|} &= -8|c|^2 + 8|c|n + 6|c| - 8n - 2 \\ g_{n,|c|} &= -8|c|^2 + 8|c|n - 2(n-3)(|c|-1) + 6|c| - 8n - 2\,, \end{aligned}$$

and for $|c| = 2$

$$\varepsilon(S_1(c), S_0(c)) = \frac{8n - 22}{3n^2 - 11n + 8} - \frac{5(6n-16)}{6(4(n-2)^2 - 2\sum_{j=1}^{n-2} \lfloor \log_2(j+1) \rfloor)}.$$

DOI: 10.5540/03.2025.011.01.0380

See Alves, Saporito and Carvalho [2] for a detailed analysis of lumpability in tree-valued Markov processes.

Proposition 4.1 explains why the projected chain strongly resembles a two-state Markov chain for many chains, as shown in Figure 2.

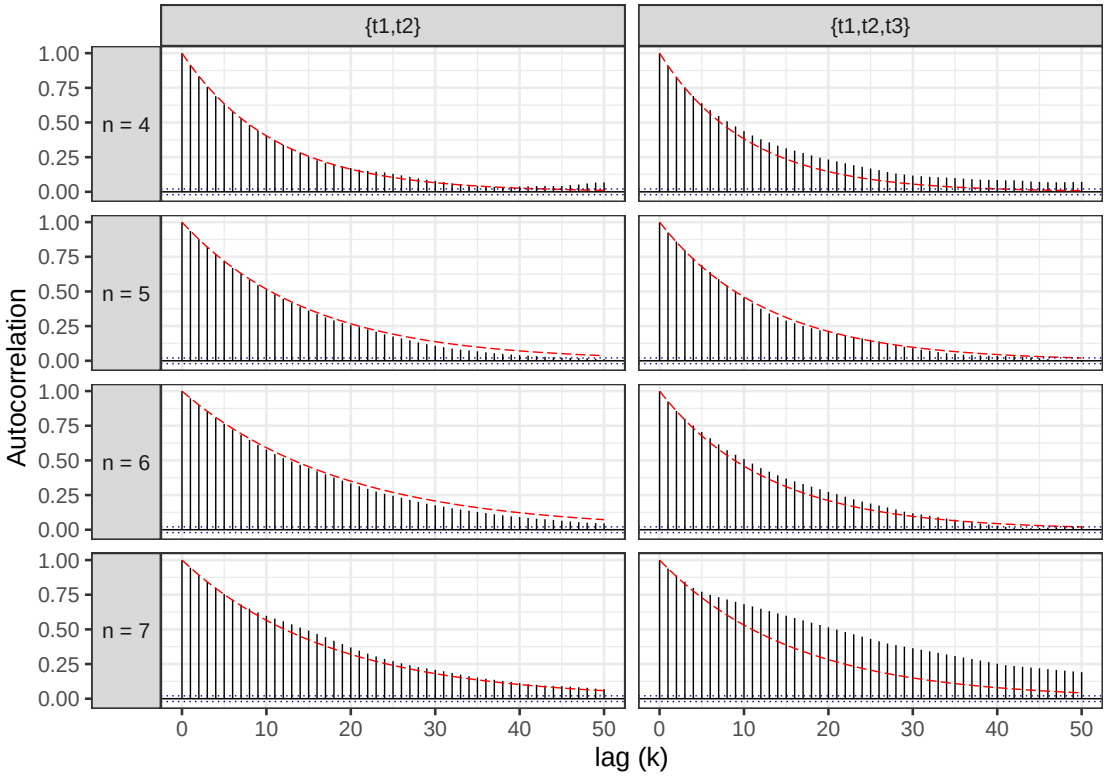

Figura 2: **Autocorrelation spectra of clade indicators for the lazy Metropolis-Hastings**. We show the empirical autocorrelation spectra up to lag $k = 50$ (black bars) for indicators of clades {t1, t2} and {t1, t2, t3} when sampling from a lazy Metropolis-Hastings with $\rho = 0.9$ on a single realisation. The autocorrelation function of the best-fitting two-state Markov chain is also shown (red line).

5 Final Remarks

In this note we have given an overview of modern developments in diagnostics for MCMC in phylogenetic applications. We have discussed how to use analytical and simulation-based techniques to ascertain correctness of a sampler. We have also given theoretical results that bound the lumping error when projecting a tree-valued Markov chain onto the space of clade indicators, providing justification for using a two-state Markov chain estimator for the effective sample size.

DOI: 10.5540/03.2025.011.01.0380

Acknowledgements

The work presented here is joint with many collaborators: Andrew Rambaut, Gytis Dudas, Guy Baele, Remco Bouckaert, Fábio Mendes, Alexei Drummond, Jiansi Gao, Erick Matsen, Marius Brusselmans, Andy Magee and Rodrigo Alves.

Referências

[1] D. Aldous. "Probability distributions on cladograms". Em: **Random discrete structures**. Springer, 1996, pp. 1–18.

[2] R. B. Alves, Y.F. Saporito e L. M. Carvalho. "On lumpability of tree-valued Markov chains". Em: **In preparation** (2024).

[3] A. Bittracher e C. Schütte. "A probabilistic algorithm for aggregating vastly undersampled large Markov chains". Em: **Physica D: Nonlinear Phenomena** 416 (2021), p. 132799.

[4] M. Brusselmans, L. M. Carvalho, S. L. Hong, J. Gao, F.A. Matsen IV, A. Rambaut, P. Lemey, M. A. Suchard, G. Dudas e G. Baele. "On the importance of assessing topological convergence in Bayesian phylogenetic inference". Em: **arXiv preprint arXiv:2402.11657** (2024).

[5] A. J. Drummond, G. K. Nicholls, A. G. Rodrigo e W. Solomon. "Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data". Em: **Genetics** 161.3 (jul. de 2002), pp. 1307–1320.

[6] A. J. Drummond, A. Rambaut, B. Shapiro e O.G. Pybus. "Bayesian coalescent inference of past population dynamics from molecular sequences". Em: **Molecular biology and evolution** 22.5 (2005), pp. 1185–1192.

[7] A. Magee, M. Karcher, F. A. Matsen IV e V. M. Minin. "How trustworthy is your tree? Bayesian phylogenetic effective sample size through the lens of Monte Carlo error". Em: **Bayesian Analysis** 1.1 (2023), pp. 1–29.

[8] F. K. Mendes, R. Bouckaert, L.M. Carvalho e A. J. Drummond. "How to validate a Bayesian evolutionary model". Em: **bioRxiv** (2024). DOI: 10.1101/2024.02.11.579856. eprint: https://www.biorxiv.org/content/early/2024/02/12/2024.02.11.579856.full.pdf. URL: https://www.biorxiv.org/content/early/2024/02/12/2024.02.11.579856.

[9] T. Stadler e Z. Yang. "Dating phylogenies with sequentially sampled tips". Em: **Systematic Biology** 62.5 (2013), pp. 674–688.

[10] S. Talts, M. Betancourt, D. Simpson, A. Vehtari e A. Gelman. "Validating Bayesian Inference Algorithms with Simulation-Based Calibration". Em: **arXiv preprint arXiv:1804.06788** (2018).

[11] D. Vats e J. M. Flegal. "Lugsail lag windows for estimating time-average covariance matrices". Em: **Biometrika** 109.3 (2022), pp. 735–750.

[12] S. Zhu, J. H. Degnan e M. Steel. "Clades, clans, and reciprocal monophyly under neutral evolutionary models". Em: **Theoretical Population Biology** 79.4 (2011), pp. 220–227.

[13] S. Zhu, C. Than e T. Wu. "Clades and clans: a comparison study of two evolutionary models". Em: **Journal of Mathematical Biology** 71.1 (2015), pp. 99–124.

