

# Assessing traits and phylogenetic signal to unravel the tempo and mode of phenotypic evolution

Diogo da Silva Ribeiro

Mestrado em Bioinformática e Biologia Computacional

Departamento de Biologia

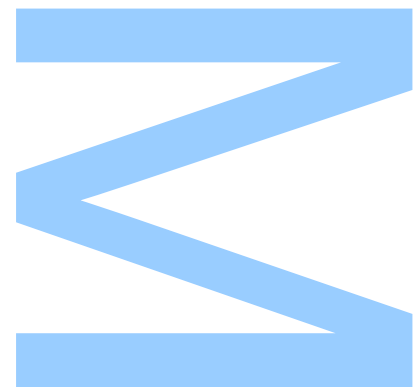
2022

## **Orientador**

Prof. Agostinho Antunes, FCUP and CIIMAR

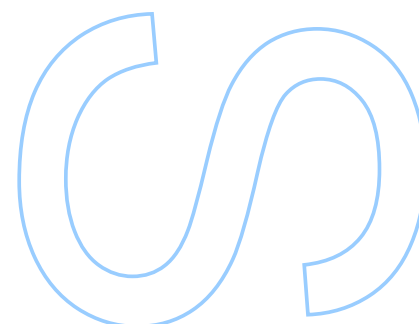
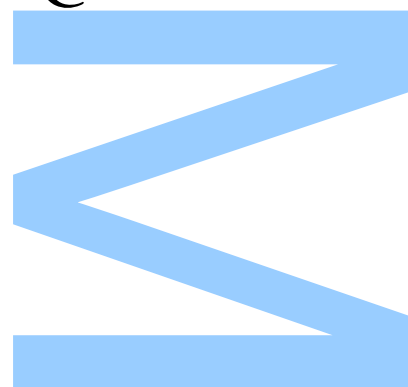
## **Coorientador**

Prof. Ana Paula Rocha, FCUP



**U.** PORTO

**FC** FACULDADE DE CIÊNCIAS  
UNIVERSIDADE DO PORTO



# Sworn Statement

I, Diogo da Silva Ribeiro, enrolled in the Master Degree Bioinformática e Biologia Computacional at the Faculty of Sciences of the University of Porto hereby declare, in accordance with the provisions of paragraph a) of Article 14 of the Code of Ethical Conduct of the University of Porto, that the content of this dissertation reflects perspectives, research work and my own interpretations at the time of its submission.

By submitting this dissertation, I also declare that it contains the results of my own research work and contributions that have not been previously submitted to this or any other institution.

I further declare that all references to other authors fully comply with the rules of attribution and are referenced in the text by citation and identified in the bibliographic references section. This dissertation does not include any content whose reproduction is protected by copyright laws.

I am aware that the practice of plagiarism and self-plagiarism constitute a form of academic offense.

Diogo da Silva Ribeiro

30 of September 2022



# Acknowledgements

The completion of this undertaking could not have been possible without the participation and assistance of so many people whose names may not all be enumerated. Their contributions are sincerely appreciated and gratefully acknowledged. However, I would like to express my deep appreciation and indebtedness particularly to the following:

It is a great pleasure to acknowledge my deepest thanks and gratitude to Prof. Agostinho Antunes, for suggesting the topic of this essay, and his kind supervision. I would also like to express my deepest thanks and sincere appreciation to Prof. Ana Paula Rocha, for her encouragement, comprehensive advice and her kind supervision until this work came to existence. It is a great honour to have worked under their supervision.

I would like to express my deep and sincere gratitude to Prof. Rui Borges, for providing invaluable guidance throughout this research. His dynamism, vision and motivation have deeply inspired me. He has taught me the methodology to carry out the research and to present the research works as clearly as possible. I would also like to thank him for his constant availability and his good mood.

Finally, I would like to thank my family members, my friends, and others for their valuable patience and who in one way or another shared their support.

This work was partially supported by the FCT project PTDC/CTA-AMB/31774/2017 (POCI-01-0145-FEDER/031774/2017).



# Abstract

The phylogenetic signal measures the tendency that species that have recently diverged resemble more than species that are distantly related. Species' observable traits (i.e., phenotypic traits) usually follow the species' evolutionary history and are thus expectedly correlated. However, violations of this expectation occur in nature and may provide clues of how species adaptation proceeds (e.g., convergent evolution). The phylogenetic signal has thus the potential to help us understand why species diverge and become different. As such, several indices to quantify the phylogenetic signal have been proposed over the last 20 years, but while many exist for continuous traits, few were devised for categorical traits. This is because categorical data pose additional challenges as they do not allow calculating variances and covariances.

The recently developed delta-statistic is based on the concept of entropy from information theory. It exploits the uncertainty on the ancestral trait's probability vectors (inferred via maximum likelihood or Bayesian inference) to calculate the degree of phylogenetic signal between a categorical trait and a phylogeny. As several phenotypic traits used in evolutionary research can only be measured in categories (e.g., presence or absence), the delta-statistic allows testing hypotheses that have been intractable to date.

Despite delta-statistic being currently in use, it suffers from both computational and statistical shortcomings that we address in this master project. In particular, we extended the statistic to deal with more evolutionary histories, accounting for sources of error that are currently being ignored and optimizing the algorithm to allow its use in large-scale genomic studies. We increase the accessibility and reproducibility of the delta-statistic, by facilitating its use to the evolutionary community, namely by introducing an easy-to-use web interface.

**Keywords:** Statistics; Evolution; Phylogeny; Categorical traits; Bioinformatics





# Contents

<b>Sworn Statement</b>	<b>iii</b>
<b>Acknowledgements</b>	<b>v</b>
<b>Abstract</b>	<b>vii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Phylogenetic Signal . . . . .	1
1.1.1 Model Approaches . . . . .	2
1.1.2 Brownian motion overview . . . . .	4
1.2 $\delta$ Statistic . . . . .	5
1.2.1 Discrete evolutionary models . . . . .	5
1.2.2 Ancestral character reconstruction . . . . .	6
1.2.3 Shannon's entropy . . . . .	8
1.2.4 Null Hypothesis . . . . .	9
1.3 Computation . . . . .	10
1.3.1 Programming language . . . . .	10
<b>2 Materials and methods</b>	<b>11</b>
2.1 Entropy Calculation . . . . .	11
2.1.1 Sample collection . . . . .	11
2.1.2 Phylogenetic inference . . . . .	12
2.1.3 Entropy Analysis . . . . .	13
2.2 Computation . . . . .	16
2.2.1 Python conversion . . . . .	16
2.2.2 Straightforward interface . . . . .	16
<b>3 Results</b>	<b>17</b>
3.1 Multiple Trees . . . . .	17
3.1.1 Entropy distribution . . . . .	17

3.1.2	Null-Hypothesis distribution . . . . .	19
3.1.3	Probability value . . . . .	21
3.2	Website Interface . . . . .	23
<b>4</b>	<b>Conclusion</b>	<b>25</b>
4.1	Future Remarks . . . . .	26
<b>5</b>	<b>References</b>	<b>27</b>

# List of Tables

1.1	Phylogenetic signal statistics table, customized from [3, 9] . . . . .	2
2.1	Specie's scientific name, family and respective trait information that was analyzed in this project. . . . .	14
3.1	Resulting values of the calculation [Quantile of ST method / Quantile of MT method] .	20
3.2	Quadrants of the 3-Class trait's significant level . . . . .	22



# List of Figures

1.1	In the left column, the plots show 500 replicates of simulated BM with the same starting value (0) and total time frame (10s) but with different evolutionary rates (A = 0.5; B = 1.5). In the right column, it shows histograms with the distributions of ending values from the BM simulations. . . . .	4
1.2	Representation of a 3-state Markov chain with the transition rates between states (A-B-C)	5
1.3	Representation of the ACR calculated for a 2-class trait, where species (A, B = 1) and (C = 2) . . . . .	6
1.4	Representation of the "Linear version of the Shannon's entropy" applied to ancestral nodes of <b>(A)</b> 2 and <b>(B)</b> 3 class discrete traits . . . . .	8
2.1	Histogram of the evolutionary rate values of the 1000 markers obtained from a random uniform sample. . . . .	11
2.2	Visual examination of the CTU2 marker in the Tracer package by analyzing both runs. <b>A.</b> trace plots, <b>B.</b> density plots and <b>C.</b> box-plots . . . . .	13
2.3	Result distribution of the multiple trees method and current method value as a green line (CTU2) in the 2-class trait . . . . .	15
2.4	Current method <b>A.</b> and Multiple trees <b>B.</b> null-hypothesis distributions (CTU2) in the 2-class trait . . . . .	15
2.5	Time comparison of the $\delta$ statistic's code in Python (VS Code) and R (RStudio), measured in a 30 species' random phylogenies with (100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 10000) iterations . . . . .	16
3.1	A comparison of the entropies obtained in the single phylogenetic tree (ST) and multiple trees (MT) methods when analyzing the sample of 1000 markers. . . . .	17
3.2	Phylogenetic tree of the genetic marker MBNL1. Phylogeny visualization made with the FigTree software. . . . .	18
3.3	Distribution of the entropy results in the different methods and k-class traits, when analyzing the sample of 1000 markers. . . . .	19

3.4	Distribution of the null hypothesis results in the different methods and k-class traits, when analyzing the sample of 1000 markers. . . . .	20
3.5	Percentage of trees from the null-hypothesis distributions that have a lower than the one obtained from the methods . . . . .	21
3.6	Image of the home page from the $\delta$ statistic web interface implemented in Django. . .	23
3.7	Image of the result page from the $\delta$ statistic web interface implemented in Django. . .	24
3.8	Image of the data page from the $\delta$ statistic web interface implemented in Django. . . .	24

# Abbreviations

**ACR** Ancestral character reconstruction. xiii, 6, 7, 16

**BM** Brownian motion. xiii, 3–5

**bp** base pair. 11

**CDS** Coding Sequence. 11

**CTMC** Continuous-time Markov chain. 5

**GTR** General time reversible. 12

**JC** Jukes-Cantor. 6

**MAP** Maximum a posteriori. 7, 14, 15, 17, 18, 25

**MCMC** Markov chain Monte Carlo. 7, 9, 12, 13, 17

**ML** Maximum likelihood. 6–8

**MPPA** Marginal Posterior Probabilities Approximation. 7, 8





# Chapter 1

## Introduction

### 1.1 Phylogenetic Signal

Phylogenies (i.e., the ancient relationships among a set of species) are frequently adopted to test past and present relationships between species and their respective adaptive traits. This is a result of the increasing number of studies reporting more accurate and larger phylogenies, as well as ecology studies giving important clues on how adaptation operates [1, 2, 3]. Adaptive traits (i.e., phenotypic traits) may follow the species' evolutionary history and are thus expected to correlate across evolutionary scales. However, violations of this expectation occur in nature and may provide clues to how fast and unpredicted species adaptation proceeds, as well as further enlighten some species' ecological resemblance [4, 5, 6]. This is why it is important to measure the statistical non-independence among species' trait values as a result of their phylogenetic similarity since they cannot be seen as independent data points [7, 3]. An important evolutionary quantity, the phylogenetic signal, measures the tendency that recently diverged species resemble more than species that are distantly related [8], and it is frequently used to find signatures of adaptive evolution or associations between genes and phenotypes that they are likely involved in [9].

The value of the phylogenetic signal in a specie's attribute can vary depending on the chosen measure to calculate it [**Table1.1**]. This has been to be particularly problematic when analyzing inaccurate phylogenies, low sample sizes, and the absence of evolutionary times (i.e., information of the branch lengths) [3]. Nonetheless, a high value for the phylogenetic signal usually indicates that similar trait values are observed in species that are closely related or, equivalently, that share a more-recent common ancestor [2]. On the contrary, a low phylogenetic signal value indicates either a random trait distribution across the phylogeny or distantly-related species that possible due to similar selective pressures, and not by inheritance, developed similar phenotypes (i.e., convergent evolution) [6][10].

### 1.1.1 Model Approaches

Several methods have been developed to measure the phylogenetic signal. These include a multitude of approaches and statistics, which we summarize in **Table 1.1**. They frequently address similar ecological and evolutionary questions (e.g., properties of species, their habitat and phylogenetic relationships) [11, 12, 10], while focusing and quantifying different aspects of the phylogenetic signal [3].

Year	Statistic	Approach	Null Hypothesis	Data Type	Ref.
1950	Moran's $I$	Autocorrelation	Permutation	Continuous	[13]
1999	Abouheif's $C_{\text{mean}}$	Autocorrelation	Permutation	Continuous	[14]
1999	Pagel's $\lambda$	Evolutionary	Maximum Likelihood	Continuous	[1]
2003	Blomberg's $K$	Evolutionary	Permutation	Continuous	[15]
2010	$D$ statistic	Evolutionary	Permutation	Categorical	[16]
2018	$\delta$ statistic	Evolutionary	Bayesian	Categorical	[9]

**Table 1.1:** Phylogenetic signal statistics table, customized from [3, 9]

The first indices were originally developed with the intention of detecting the presence of systematic spatial variation in a mapped variable (spatial autocorrelation), having the purpose of finding signatures in the distribution [17]. Later on, these were adopted to be used in phylogenetic analyses [18]. The autocorrelation indices, Moran's  $I$  [13] and Abouheif's  $C_{\text{mean}}$  [14], are based on summary statistics of correlation. When compared to evolutionary approaches, they show better robustness to inaccurate phylogenetic information and impose less restrictive assumptions [19, 20]. However, since the statistic under the given model is unknown beforehand, comparing the values obtained between multiple phylogenetic trees cannot provide quantitative interpretations. On the other hand, all recent methods use evolutionary approaches that rely on an evolutionary model. This is extremely beneficial since it allows an easier evolutionary interpretation of character evolution [21, 22].

Another instance where these methods diverge is in how they generate random trait distributions with the aim of testing the null hypothesis (no phylogenetic signal). While most analyses numerically simulate random traits by relying on random permutations of the trait values among the tips of the phylogenetic tree, the Pagel's  $\lambda$  [1] and  $\delta$  statistic [9] can attain it analytically by leveraging previously made assumptions as a means to create a distribution where the alternative hypothesis is tested [3]. A comparison between the two methods still has not reached a clear conclusion [23, 24].

Another crucial aspect is the fact that, although several indices to quantify the phylogenetic signal have been proposed, most were only devised for continuous traits. This is because, following Pagel's  $\lambda$  [1], several statistics have their indices directly related to a Brownian motion (BM) model [25]. Hence, the traits evolve following a random walk along the branches of the phylogenetic tree, and the variance in the distribution of trait values is directly proportional to the branch length. On the other hand, since variance and covariance cannot be directly calculated in categorical data, the number of available alternatives to compute the phylogenetic signal are significantly lower. Two methods have nevertheless been proposed to overcome this issue:

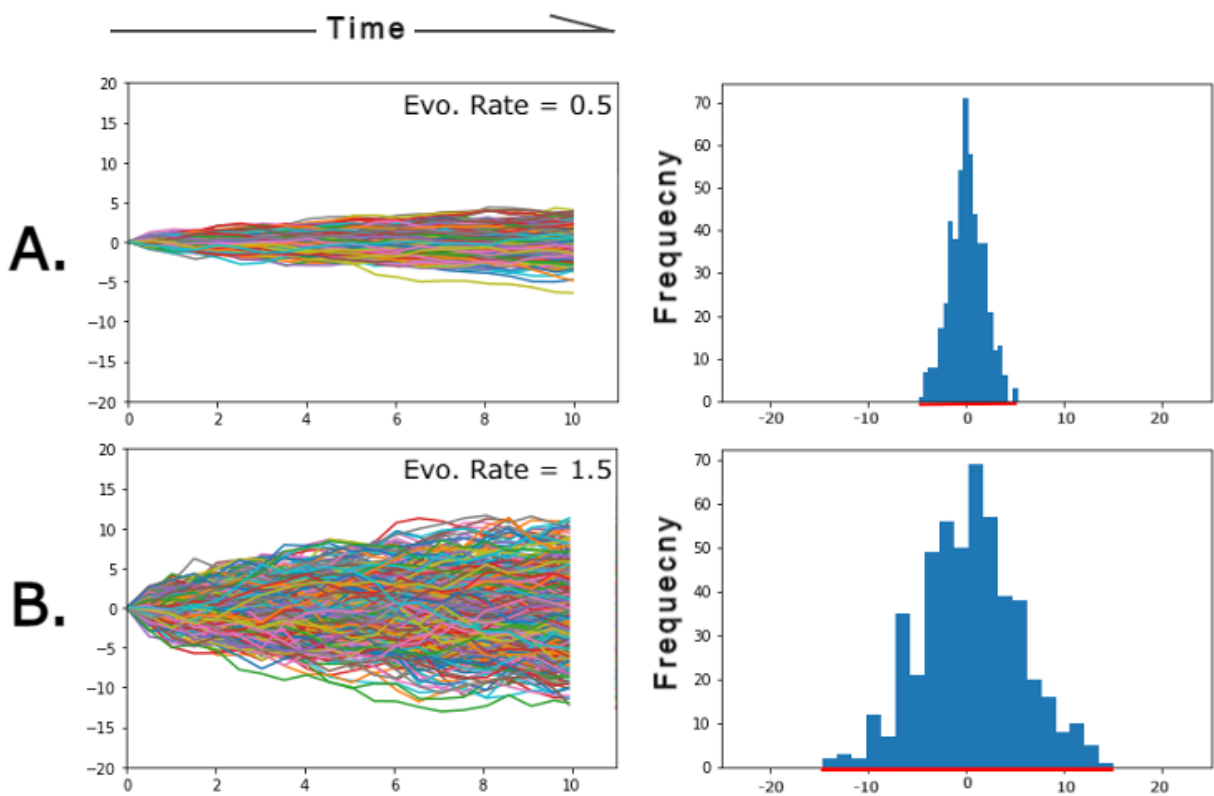
- **D statistic** [16] circumvents this problem by defining a trait discretization based on a continuous trait that evolves under the BM.
- **$\delta$  statistic** [9] uses Shannon entropy for measuring the degree of phylogenetic signal between a categorical trait and a phylogeny.

Multiple aspects of species' characteristics can only be explored with categorical traits, but the phylogenetic signal in these characters is still relatively new and understudied due to the statistical difficulties in treating this type of data [26, 9].

Nonetheless, it is known that discrete traits are affected heavily by the evolutionary rate, becoming more pronounced when the rate increases or the number of potential states decreases [27, 28, 29]. The validity of this in continuous traits is something that has recently been heavily questioned after it was noticed that the evolutionary rate does not impact the phylogenetic signal when the evolutionary process approximates BM. Furthermore, the analyses of different evolutionary processes show that the resulting phylogenetic signal value does not seem to be affected by the similarity of these methods when testing the results in Blomberg's  $K$  [21, 30].

### 1.1.2 Brownian motion overview

The BM model, originally developed to describe the motion of particles suspended in a fluid, is a popular model in comparative biology since it captures the way traits might evolve in basic evolutionary circumstances. It has very practical statistical properties that make tree analyses and computations pretty straightforward. Overall, in this model, traits are affected by a lot of extremely small "forces" that when combined result in a normal distribution of the trait values regardless of how they are distributed or what forces caused them [25, 7, 21].



**Figure 1.1:** In the left column, the plots show 500 replicates of simulated BM with the same starting value (0) and total time frame (10s) but with different evolutionary rates (A = 0.5; B = 1.5). In the right column, it shows histograms with the distributions of ending values from the BM simulations.

BM is a simplistic model that can be completely described by only the initial mean trait value that is seen in the ancestral population and the evolutionary rate parameter, that determines how fast traits will randomly walk through time and how varied the species could become. Furthermore, it can be represented simply by a variance-covariance matrix, where the values are proportional to the branches of species in the phylogenetic tree [Figure 1.1] [31, 10].

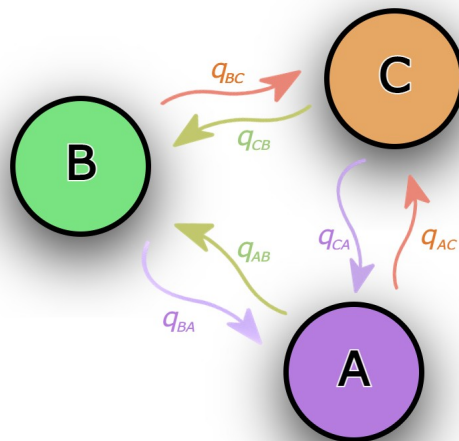
However, a BM model corresponds to a few simple scenarios of trait evolution, not fitting all biological traits. The phylogenetic signal may also greatly deviate from what the BM predicts, changing drastically as a result of some evolutionary processes (e.g., genetic drift with different rates across the tree) [32, 21, 30].

## 1.2 $\delta$ Statistic

As aforementioned, the  $D$  statistic is based on the BM to model the evolution of categorical traits, and thus might be unfitting in some evolutionary scenarios [16, 9]. Therefore, the  $\delta$  statistic manages to fill an important void in the literature since it is currently the only method that can accurately calculate the phylogenetic signal in categorical traits without assuming the BM model.

### 1.2.1 Discrete evolutionary models

The continuous-time Markov chain (CTMC) model is the usual approach for analyzing the evolution of discrete traits in trees. These characters can be modeled similarly to how the models of sequence evolution work (e.g., JC [33], HKY [34], and GTR [35]). The models thus provide a rate matrix that, instead of defining the nucleotide or amino acid changes, represent the state changes in the character space [Figure 1.2].



**Figure 1.2:** Representation of a 3-state Markov chain with the transition rates between states (A-B-C)

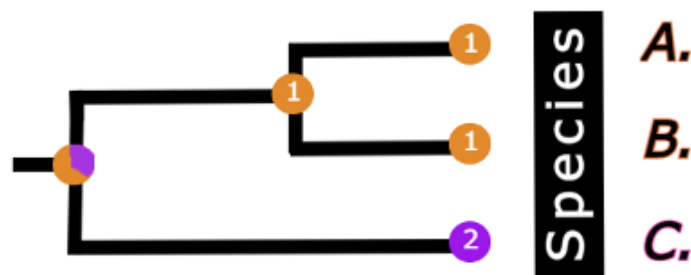
Furthermore, through an exponentiation of the instantaneous rate matrix, the probability distribution of the  $k$ -states of a trait can be calculated for any time interval [36].

The Markov chain model is the most basic model for discrete character evolution, being an analog of the Jukes-Cantor (JC) [33] model. All of its transitions have the same rate, with the probability of changing from one state to another depending only on the current state, regardless of their start or end states. As such, it makes no difference if a lineage has had the feature for a long time or has only recently evolved it [37, 38]. This approach could be extended by incorporating more complex models that account for rate variations between characters:

- **(a)** On one hand, this can be incorporated as a symmetric model, where the rate of change in 2 character states is the same in both forward and reverse.
- **(b)** On the other hand, a more complex model where every possible type of transition can have a different rate, (all-rates-different model) can also be incorporated.

### 1.2.2 Ancestral character reconstruction

For the purpose of calculating the phylogenetic signal, an ancestral character reconstruction (ACR) of the discrete traits is first necessary. This inference of historical data from measured characteristics in species [Figure 1.3] is an important application of phylogenetics and is still widely used [39, 40]. The ACR is done probabilistically, still having some uncertainty regarding the possible state at a certain ancestral node. Progress is constantly being made by the recent development of efficient computer techniques and the exponential expansion of computing power. However, numerous challenges still persist [38].



**Figure 1.3:** Representation of the ACR calculated for a 2-class trait, where species (A, B = 1) and (C = 2)

The ACR relies on a multitude of factors that need to be carefully considered when analyzing the data. For instance, the incorporation of known-age historical samples for calibration can help resolve some problems of reconstructions. Generally speaking, the more evolutionary time passes between the ancestor and its descendants, the less accurate these reconstructions become. Using this calibration is shown to improve the ACR compared to only using contemporaneous data [41, 42, 43].

Throughout the years, several methods (Maximum parsimony [44], maximum likelihood (ML) [1, 45, 39] and Bayesian methods [46, 47]) with multiple approaches for ACR have been proposed,

consequently having different complexity and performance [38, 37]. It is important to note that the parsimony-based ACR cannot be used to calculate the  $\delta$  statistic, as it is necessary to produce probability vectors for the ancestral nodes [9]. Moreover, although this approach is fast, it is not very accurate, having various problems when compared to the other two methods and lacking versatility [48, 49, 50, 38].

- **ML methods** treat character states at the internal nodes as parameters and attempt to provide point estimates (instead of distributions, like in Bayesian) by using the observed leaf states, model of evolution (an inaccurate model might drastically compromise the accuracy of ACR) and phylogeny. Their objective is to try to maximize the probability of the data and, at each node, all potential ancestral character states are used to determine the likelihood of its descendants [1].

The most well-known practices to analyze the data are: (a.) calculating marginal ML at each ancestral node and then choosing the combination of states that gives the highest ML value. This selection could induce globally inconsistent scenarios because of the independence between nodes [45, 51, 37]; (b.) or using a more computationally complex method to find the joint combination of ancestral character states from across the tree that together maximizes the likelihood of the data. This joint method selects a unique state for every tree node, not taking into account that multiple scenarios may have similar uncertainties [52, 38].

Additionally, a novel method between the MAP and joint approaches was recently proposed. This method uses Marginal Posterior Probabilities Approximation (MPPA) [37] and has high accuracy and robustness.

- **Bayesian methods** use the likelihood of observed data and a prior distribution (evolutionary model and the phylogenetic tree) to infer the posterior probabilities of ancestral character states at each internal node of a given tree. According to some researchers, using a Bayesian approach and evaluating ancestral reconstructions over many trees to account for uncertainty in the tree reconstruction should be done [47]. However, this has a high computational cost and, because of it, cannot be achieved for large data sets [38].

Currently, there are two main approaches: (a.) The Empirical Bayes approach requires both the evolutionary model parameters and the tree to be known without error, using them to calculate the probabilities of ancestral states [53]. (b.) The Hierarchical Bayes approach averages all potential evolutionary trees and models in proportion to how plausible they are by using MCMC, leading to the need for a much greater computer cost [46].

There are several software packages available dedicated to ACR that are mainly maintained by the

scientific community, implementing these methods in various programming languages. For instance, the `PastML` software can use the above-stated MPPA method in the Python programming language [37] and the original  $\delta$  statistic code was applied in the package `ape` for the programming language R. This R package uses scaled conditional likelihoods, which, unlike other ML-based methods, might negatively affect the accuracy [54, 38].

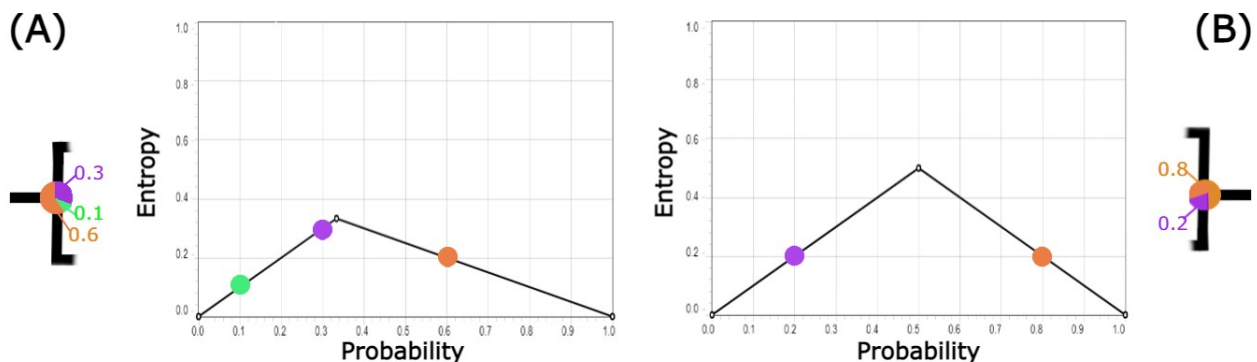
### 1.2.3 Shannon's entropy

We now introduce in deeper detail the delta statistic, as it represents the core of this thesis.

After the ancestral discrete characters are estimated for all nodes in the phylogeny, a linear version of Shannon's entropy (1.1) is applied to calculate the expected information that is obtained from the node probabilities in the ancestral traits [9].

$$e_i^j = \begin{cases} p_i^j & , \text{ if } p_i^j \leq 1/k \\ \frac{1}{1-k} p_i^j - \frac{1}{1-k} & , \text{ if } p_i^j > 1/k \end{cases} \quad (1.1)$$

For each of a trait's states, the linear version of Shannon's entropy is calculated. This equation (1.1) is changed to account for the maximum amount of characters that may be contained in a single trait [Figure 1.4], and it is then implemented so that the maximum value of a state entropy can be simply calculated as  $1/k$  [9].



**Figure 1.4:** Representation of the "Linear version of the Shannon's entropy" applied to ancestral nodes of (A) 2 and (B) 3 class discrete traits



The entropy is maximized when all states have identical probability and, on the other hand, when a state occurrence in a node is certain (i.e., it has probability 1 and all the others 0), the entropy becomes null. Finally, combining all state entropies in a node, will create a node entropy defined in the  $(0, 1)$  interval [9].

$$e^j = \sum_{i=1}^k e_i^j \quad (1.2)$$

#### 1.2.4 Null Hypothesis

In the original approach of the statistic, the node entropies' beta distribution was used to calculate the shape parameters  $\alpha$  and  $\beta$  and then, they were implemented in an MCMC scheme to obtain random samples from the posterior distribution [9]. The Beta distribution is often used in Bayesian inference, describing the distribution of a probability in the  $(0, 1)$  interval. As our entropies also varied in that range, the beta distribution is a suitable option. The delta statistic can then be calculated as the expected ratio between the obtained posterior distributions of the previously calculated  $\alpha$  and  $\beta$ :

$$\delta = E \left[ \frac{p(\beta|\alpha, e)}{p(\alpha|\beta, e)} \right] \quad (1.3)$$

In this work, the null hypothesis was obtained by randomly assigning trait values at the tree tips multiple times and then calculating their entropy distribution. In other words, the whole null distribution and not just a simple value represented by it, was simulated. This approach was preferred because it provides a better assessment of the features that are responsible for false positives and negatives.

## 1.3 Computation

Currently, computation has become an essential tool to quickly deal with previous time-consuming procedures and be able to manage the ever-growing volume of available data (i.e. easily obtained in existing databases [55, 56]). These changes provide an immense opportunity for scientific discovery, but, especially when there is not an established guide or supportive software to handle data more easily, computational resources can be difficult to use and feel overloaded [57].

Having access to an easy-to-use interface is an advantage for the currently numerous researchers without informatics or programming expertise in data analysis. Although this is usually less flexible than working directly in a programming language, restricting these analyses to only researchers that can program ultimately leads to a missed opportunity for breakthrough, limiting ideas and/or science development or, in the worst case, leading to confusion and difficulty in reproducibility of results when the available data is improperly analyzed [58].

### 1.3.1 Programming language

There are currently multiple programming languages used in bioinformatics. Each of them has varying degrees of computing speed, coding flexibility, and memory consumption [59]. Among them, the programming languages R and Python have been widely used in genomics.

The current code of the  $\delta$  statistic is implemented in R, an open-source programming language that has data analysis and visualization as its primary purposes, but, for this project, this code was converted into Python. Python is the current most popular coding language as measured by both the TIOBE [60] and PYPL [61] indexes, having also grown the most in the last five years. In addition to reaching a larger audience, it is actively getting better with each new release, having often changed the language to improve its syntax and performance.

The use of frameworks in web development is becoming crucial, being used by countless developers worldwide to build rich and dynamic web applications. Of those, Django is a framework that uses Python for web development. This framework was chosen because it offers a ton of high-quality documentation and is highly secure, giving little room for security vulnerabilities by being constantly updated and kept up to an elevated standard. Additionally, although it mainly assists in the back-end (server-side, for the website components and server creation), it also helps the user in the front-end (client-side, improving some visual aspects in the website view) [62].

# Chapter 2

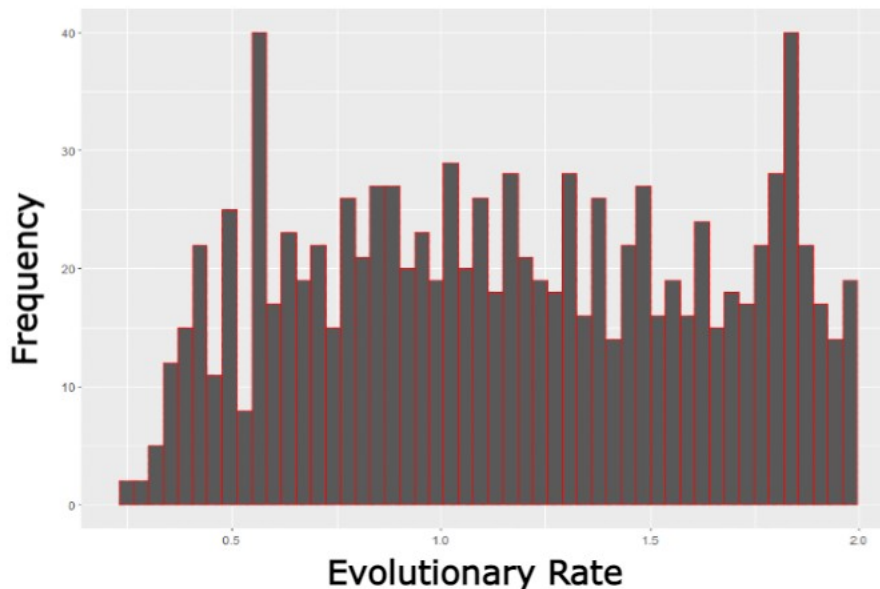
## Materials and methods

### 2.1 Entropy Calculation

#### 2.1.1 Sample collection

The OrthoMAM [63], a database of orthologous mammalian markers, was employed to obtain a group of protein-coding sequence (CDS) alignments. With the aim of obtaining alignments representative of multiple mammal groups, 30 species whose genomes are particularly well known were selected based on their varied taxonomy [Table 2.1]. As such, only the genetic markers containing all of these selected species were considered for further analyses.

Furthermore, to avoid possible biases due to gene length, the alignments were trimmed to 1000 base pair (bp) and, to keep their original data as much as possible, this adjustment was not made in alignments with more than 2500 bp. After this, markers with resulting sequences consisting mostly of gaps were also discarded.



**Figure 2.1:** Histogram of the evolutionary rate values of the 1000 markers obtained from a random uniform sample.

Finally, since this project also has the intention of studying if the rate at which new substitutions arise within a species or lineage influences the estimated  $\delta$ , the relative evolutionary rate values were obtained and a uniform random sample of 1000 was collected, represented in **Figure 2.1**.

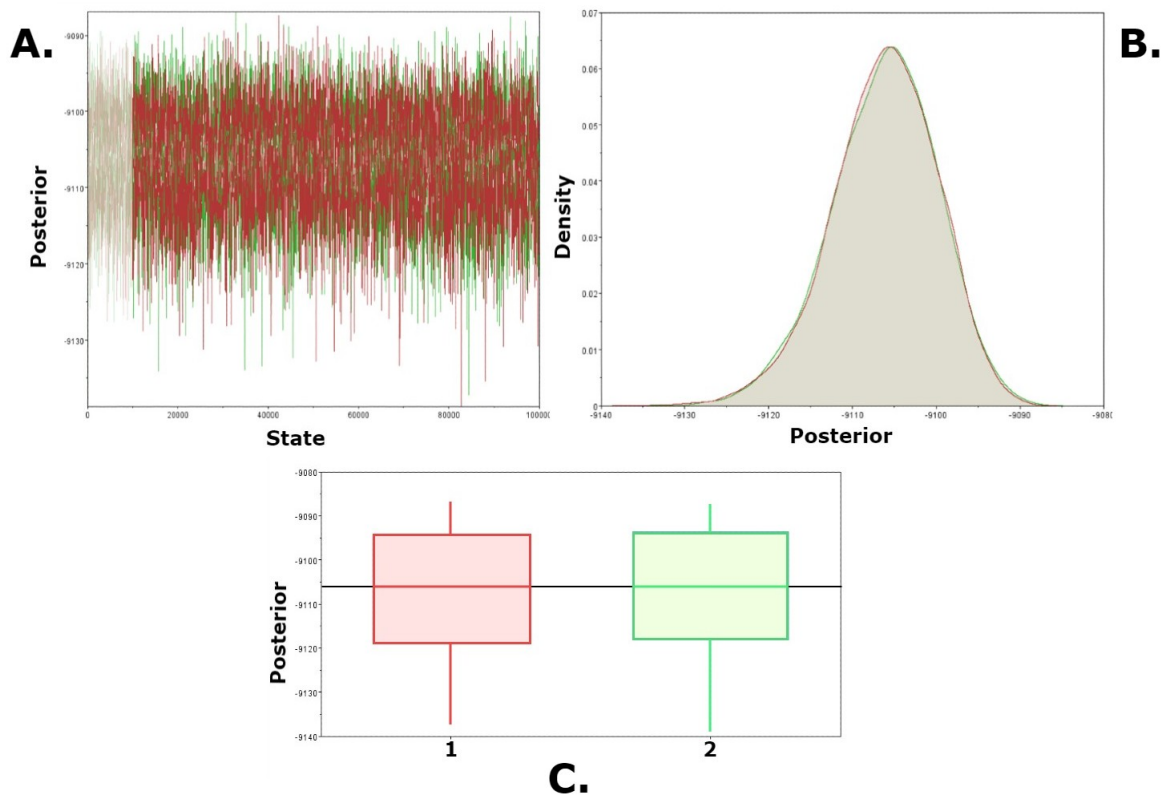
### 2.1.2 Phylogenetic inference

Phylogenetic inference was performed in a Bayesian MCMC framework implemented in `RevBayes` [64], a software that allows for the implementation of more complex phylogenetic models, permitting adapting substitution models to particular biological scenarios.

The samples were analyzed under the general time reversible (GTR) model [35], which assumes different rates of substitution for each pair of nucleotides and different frequencies of occurrence of nucleotides. Also, both the discrete gamma model [65] and the possibility of invariant sites [34] were included with the GTR model. According to the discrete gamma model, the substitution rate at each site is a random variable, described by a discretized gamma distribution. Additionally, the possibility of invariant sites enables the substitution rate of a site to be zero. Afterwards, since there was no strong prior knowledge about the pattern of the relative rates, the GTR model was defined with a flat Dirichlet distribution (1,1,1,1,1,1), describing equal exchangeabilities between nucleotides. The MCMC ran with two independent chains for a total of 100000 generations and, to control for autocorrelation between the sampled parameters, only every 10th iteration was considered for the estimation.

Although there is not a universally established approach for convergence assessment, the currently most used methods were used in this project: visual examination of the chain trajectory [**Figure 2.2**] and numerical quantification of convergence. Additionally, even though the quantity of discarded samples as a burn-in has divergent opinions, 15% of samples were discarded [66].

- For the visual examination, the `Tracer` package [67] was used to assess the reliability of the MCMC convergence by superimposing the two chains in a trace plot and by checking if they meander fairly smoothly and overlap each other. The resulting density plots were also checked to verify if they overlapped well after the burn-in period, and the distribution of values was quickly inspected [**Figure 2.2**].
- To check numerically for convergence, the python package `ChainConsumer` [68] was deployed to quickly calculate both the Gelman-Rubin statistic [69] (the most popular method for assessing samples, obtained from running MCMC algorithms, which compares the estimated between-chains and within-chain variances) and the Gweke statistic [70] (which is based on a test for equality of the means of the first and last part of a Markov chain), two convergence diagnostics for Markov chains.



**Figure 2.2:** Visual examination of the CTU2 marker in the Tracer package by analyzing both runs. **A.** trace plots, **B.** density plots and **C.** box-plots

The resulting 8500 sampled trees represent topologies that were visited during the MCMC; however, they may have been visited several times. We created an exhaustive code that uses the `ace` package in R, which compares each topology and determines their frequency. A sample of 1000 trees and their sample frequency was created for each gene.

### 2.1.3 Entropy Analysis

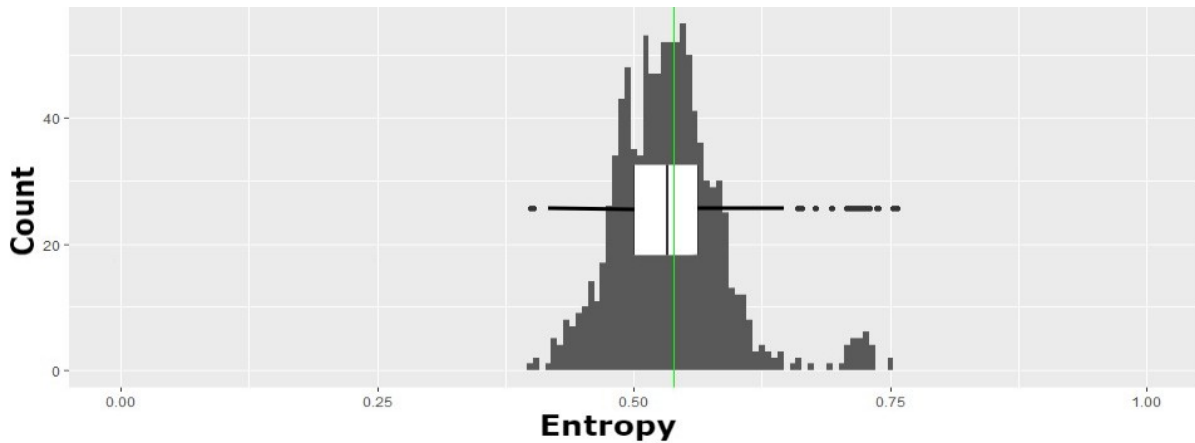
To test the newly devised  $\delta$ -statistic, the characteristic primary diet of the species was arbitrarily chosen. For the 30 mammalian species in our alignments, we created trait vectors that consisted of 2-class [**Traits.A**] (presence/absence of meat in primary diet) and 3-class [**Traits.B**] (carnivorous, omnivorous, and herbivorous) traits that were defined based on existing scientific literature collected by the Animal Diversity Web database [71] [**Table 2.1**].

Specie	Family	Traits.A	Traits.B
<i>Ailuropoda melanoleuca</i>	Ursidae	No	Herbivorous
<i>Bos mutus</i>	Bovidae	No	Herbivorous
<i>Camelus bactrianus</i>	Camelidae	Yes	Omnivorous
<i>Canis familiaris</i>	Canidae	Yes	Carnivorous
<i>Capra hircus</i>	Bovidae	No	Herbivorous
<i>Castor canadensis</i>	Castoridae	No	Herbivorous
<i>Cebus capucinus</i>	Cebidae	Yes	Omnivorous
<i>Ceratotherium simum</i>	Rhinocerotidae	No	Herbivorous
<i>Chlorocebus sabaeus</i>	Cercopithecidae	No	Herbivorous
<i>Colobus angolensis</i>	Cercopithecidae	No	Herbivorous
<i>Cricetulus griseus</i>	Cricetidae	Yes	Omnivorous
<i>Enhydra lutris</i>	Mustelidae	Yes	Carnivorous
<i>Eptesicus fuscus</i>	Vespertilionidae	Yes	Carnivorous
<i>Equus asinus</i>	Equidae	No	Herbivorous
<i>Homo sapiens</i>	Hominidae	Yes	Omnivorous
<i>Macaca fascicularis</i>	Cercopithecidae	Yes	Omnivorous
<i>Mus musculus</i>	Muridae	Yes	Omnivorous
<i>Nomascus leucogenys</i>	Hylobatidae	No	Herbivorous
<i>Odobenus rosmarus</i>	Odobenidae	Yes	Carnivorous
<i>Odocoileus virginianus</i>	Cervidae	No	Herbivorous
<i>Orcinus orca</i>	Delphinidae	Yes	Carnivorous
<i>Orycteropus afer</i>	Orycteropodidae	Yes	Carnivorous
<i>Otolemur garnettii</i>	Galagidae	Yes	Omnivorous
<i>Panthera pardus</i>	Felidae	Yes	Carnivorous
<i>Papio anubis</i>	Cercopithecidae	Yes	Omnivorous
<i>Physeter catodon</i>	Physeteridae	Yes	Carnivorous
<i>Rhinolophus sinicus</i>	Rhinolophidae	Yes	Carnivorous
<i>Sus scrofa</i>	Suidae	Yes	Omnivorous
<i>Trichechus manatus</i>	Trichechidae	No	Herbivorous
<i>Tupaia chinensis</i>	Tupaiaidae	Yes	Omnivorous

**Table 2.1:** Specie's scientific name, family and respective trait information that was analyzed in this project.

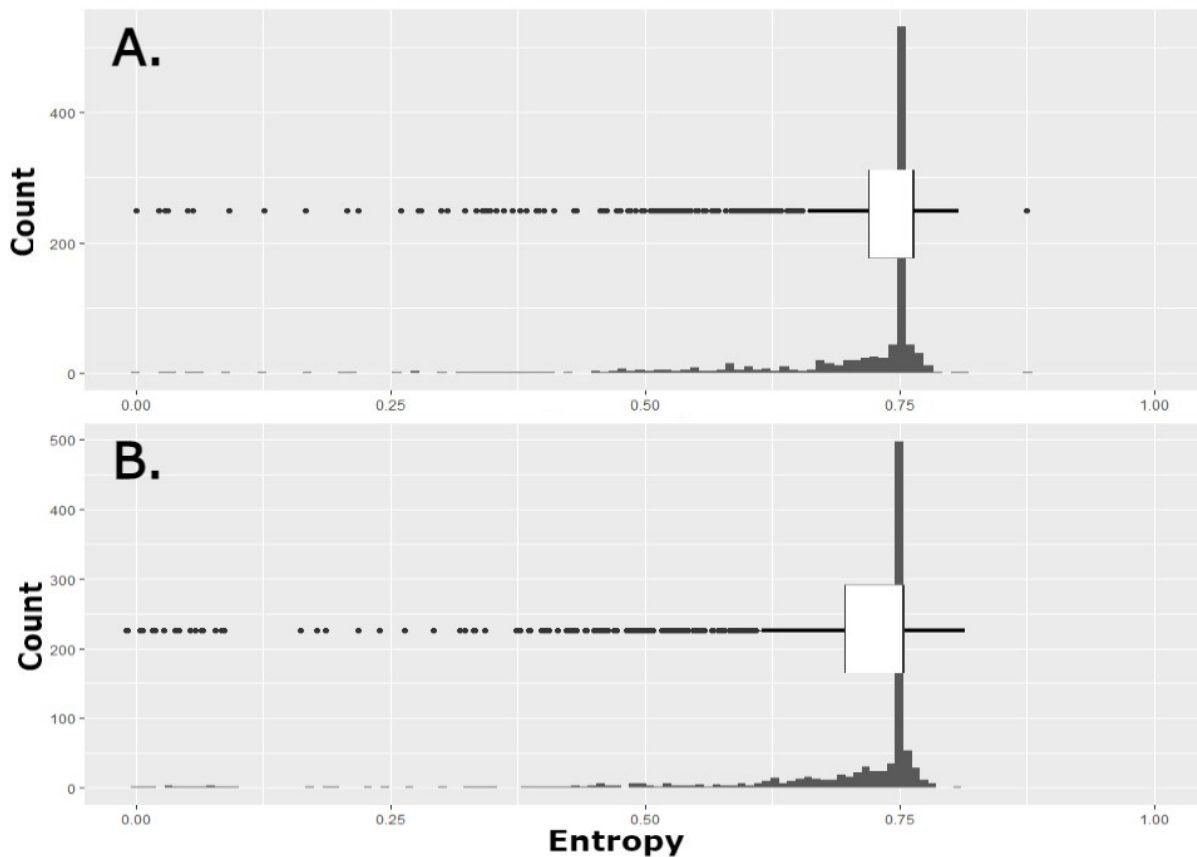
Finally, for the purpose of comparing if the  $\delta$  statistic varies significantly when multiple phylogenetic trees are accounted for, we compared the estimated entropies under the current method and the extention proposed in this thesis [Figure 2.3], plus their null distributions [Figure 3.4]:

- **Current method:** The data uses the maximum a posteriori (MAP) tree and the trait vector.
- **Multiple trees method:** This method calculates the entropies based on the trait vector and multiples trees, thus accounting for variations in the species tree. These trees were sampled from the posterior distribution and their mean entropy value was obtained.



**Figure 2.3:** Result distribution of the multiple trees method and current method value as a green line (CTU2) in the 2-class trait

- **Null-Hypothesis with the current method:** The data uses the MAP tree, but the trait values among the tips of the phylogenetic tree are randomly assigned multiple times. It is represented as a distribution of 1000 data points.
- **Null-Hypothesis with the multiple trees method:** This distribution is obtained by sampled the tree from the posterior distribution, but the trait values among the tips of the phylogenetic tree are randomly assigned in each instance. It is represented as a distribution of 1000 data points.

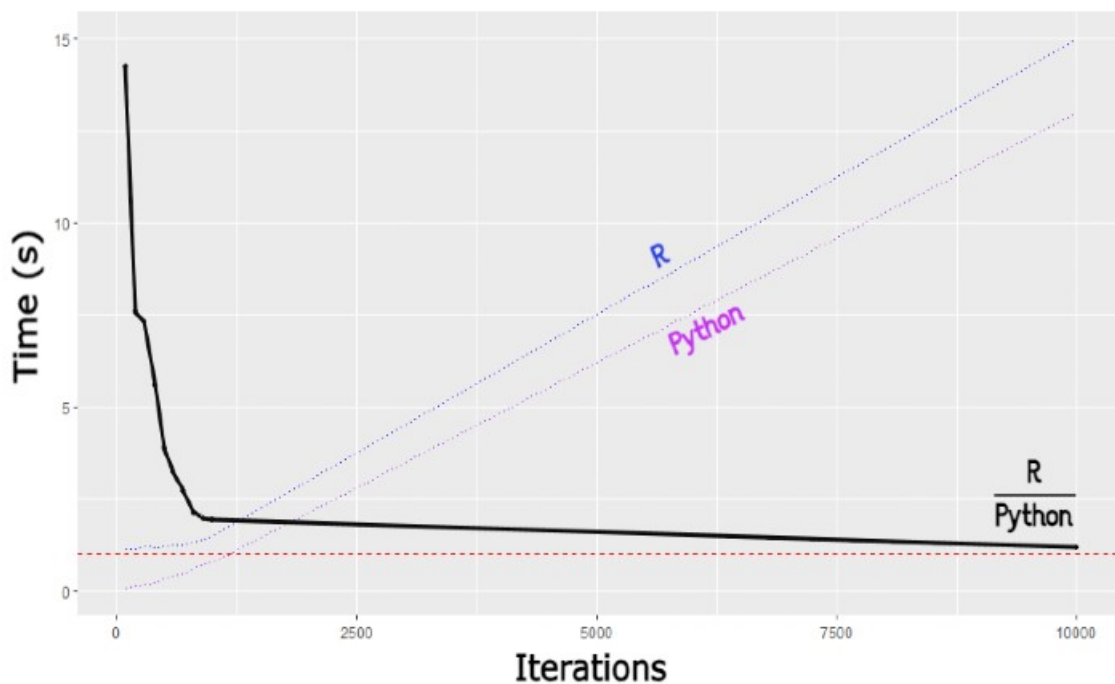


**Figure 2.4:** Current method **A.** and Multiple trees **B.** null-hypothesis distributions (CTU2) in the 2-class trait

## 2.2 Computation

### 2.2.1 Python conversion

As previously stated, the  $\delta$  statistic code was converted to the Python programming language to take advantage of its straightforward code and current fast growth. Array operations were also implemented using the NumPy [72] library, resulting in more compact and faster reading and writing of data items. Then, the final code's velocity was measured to compare it to the existing delta statistic in R [Figure 2.5].



**Figure 2.5:** Time comparison of the  $\delta$  statistic's code in Python (VS Code) and R (RStudio), measured in a 30 species' random phylogenies with (100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 10000) iterations

The PastML package is currently available in Python for the ACR. It was chosen to produce probability vectors for the ancestral nodes and its code was slightly modified and then applied to be used with the delta statistic.

### 2.2.2 Straightforward interface

As previously mentioned, we used Django because of its assistance with both the front-end and back-end which allowed for the swift creation of a website. Additionally, one of the most widely used CSS frameworks, Bootstrap, was used for the front-end development, contributing to a unified appearance of the website by providing basic style definitions for all HTML components and ready-to-use CSS style-sheets. Overall, the website was created using Python, CSS, JavaScript, and HTML.



# Chapter 3

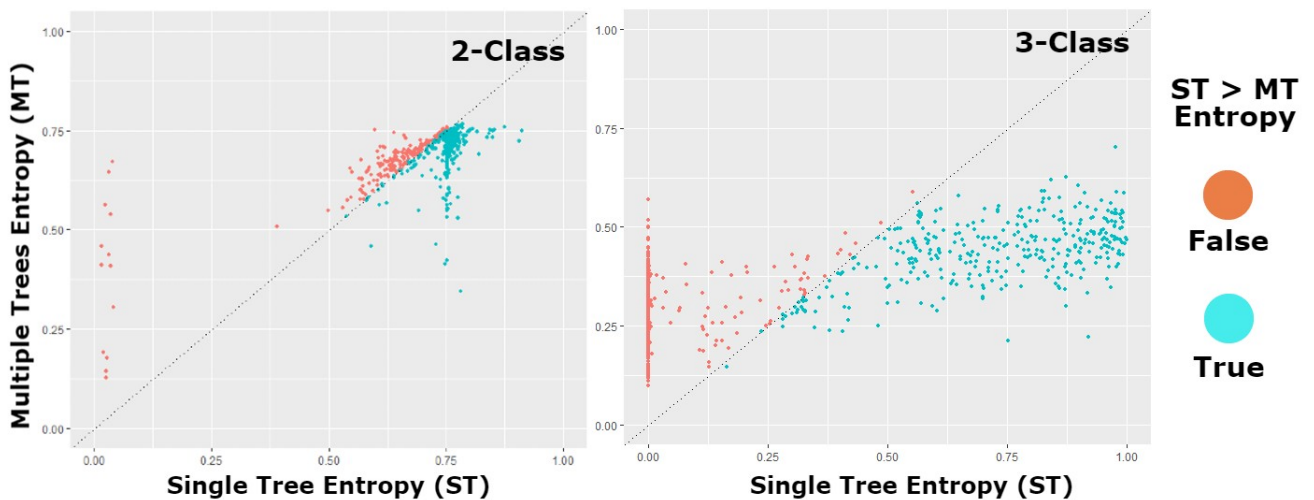
## Results

### 3.1 Multiple Trees

The  $\delta$  statistic still has some statistical flaws, one of which is the assumption that the phylogenetic tree's topology and branch lengths are known. To solve this unaccounted-for source of error, when calculating the statistic, we also examined the entropy values taking into account the overall uncertainty of a phylogeny. For that, we calculated  $\delta$  with several trees (i.e., those that were visited during the MCMC analyses) and compared its value, as well as its statistical significance, to the current method, which uses a single tree (i.e., the MAP tree).

#### 3.1.1 Entropy distribution

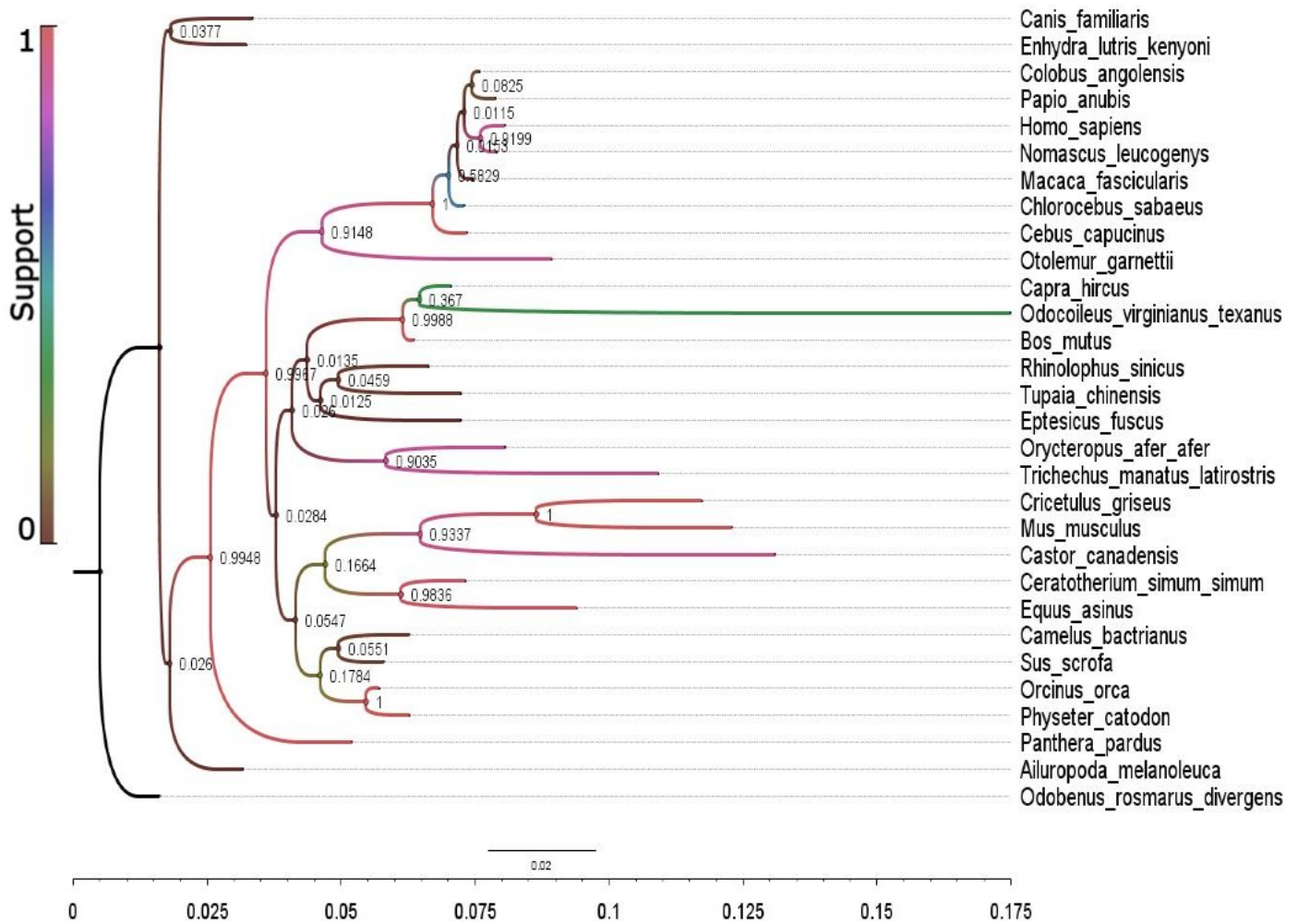
We calculated the entropy in a 2-class and 3-class trait [Table 2.1] for the single and multiple trees methods. When comparing both entropies together, we expect that if the entropies under both methods behave similarly, they fall within the identity line, which is clearly not the case.



**Figure 3.1:** A comparison of the entropies obtained in the single phylogenetic tree (ST) and multiple trees (MT) methods when analyzing the sample of 1000 markers.

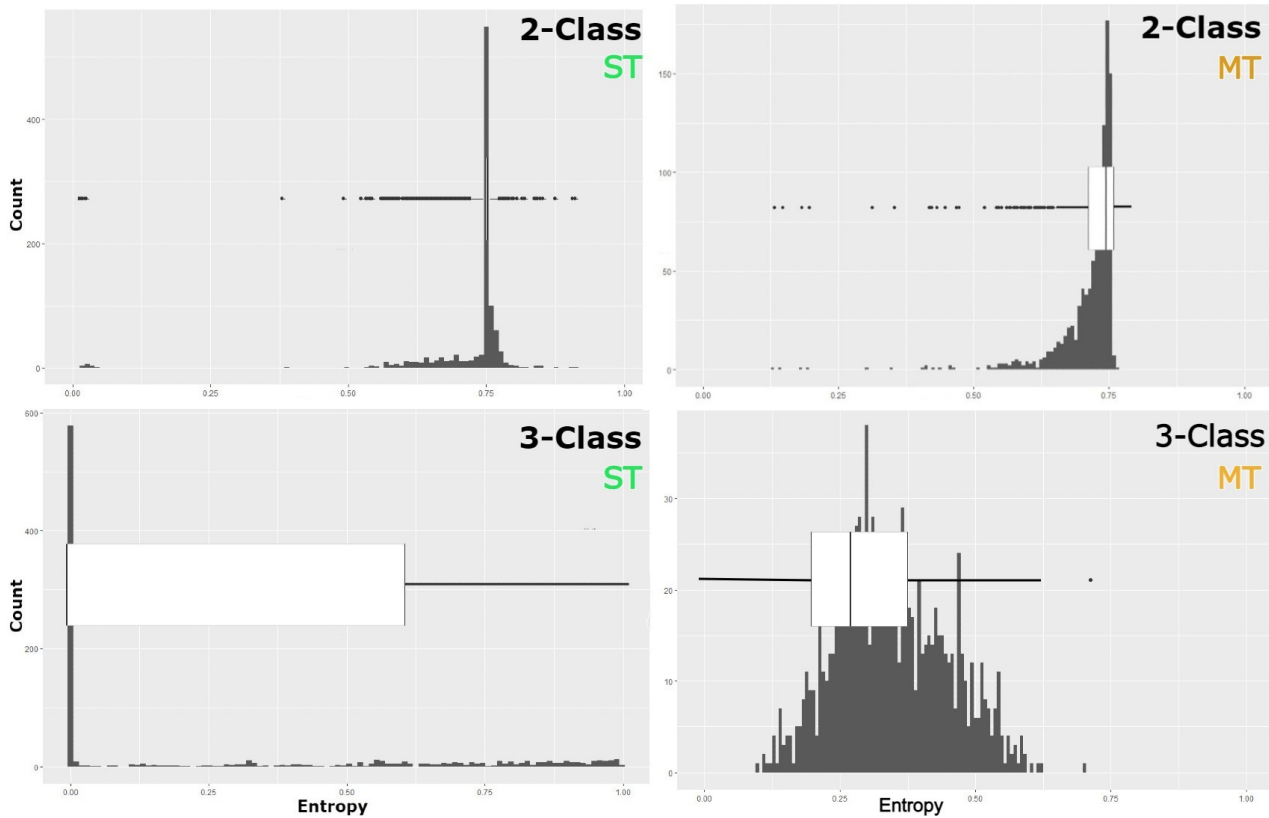
When analyzing **Figure 3.1**, there exists a set of markers that, with the old method had an entropy close to zero but, this was changed when they were analyzed with multiple trees, increasing the entropy. A subset of these outliers' phylogeny was then analyzed and, as expected, these phylogenies show low posterior clade probabilities for the majority of clades. As such, the MAP does not represent a likely species past history and the resulting statistics for the entropy are, in a lot of cases incorrect [**Figure 3.2**]. As such, when the MAP phylogenetic tree has a high uncertainty value, the single tree method needs to make more assumptions and selects a possibly wrong phylogeny but, when the statistic is analyzed through multiple trees, even if low support is verified, some of the alternative trees might explain the trait resulting in a higher entropy value.

Subsequently, the remaining data indicates that the new technique is more conservative, resulting in a reduction of the overall gene entropies.



**Figure 3.2:** Phylogenetic tree of the genetic marker MBNL1. Phylogeny visualization made with the FigTree software.

The resulting distribution of entropies changes considerably between the two methods. While inspecting their resulting distributions, we can see in both cases that the single tree method, when analyzing a trait vector, gives a fixed point to most of phylogenies, independently of their uncertainty [Figure 3.3].



**Figure 3.3:** Distribution of the entropy results in the different methods and k-class traits, when analyzing the sample of 1000 markers.

In contrast, with the new multiple trees method, its entropy is dependent on multiple possibilities, shifting the result to a more pragmatic interpretation that is dependent on the phylogeny’s uncertainty and the similarities between them (close phylogenies might give similar values). Additionally, extreme values of absolute certainty (Entropy close to 0 or 1), are harder to obtain in this method, needing the possible trees to match entropy values when analyzing the marker.

### 3.1.2 Null-Hypothesis distribution

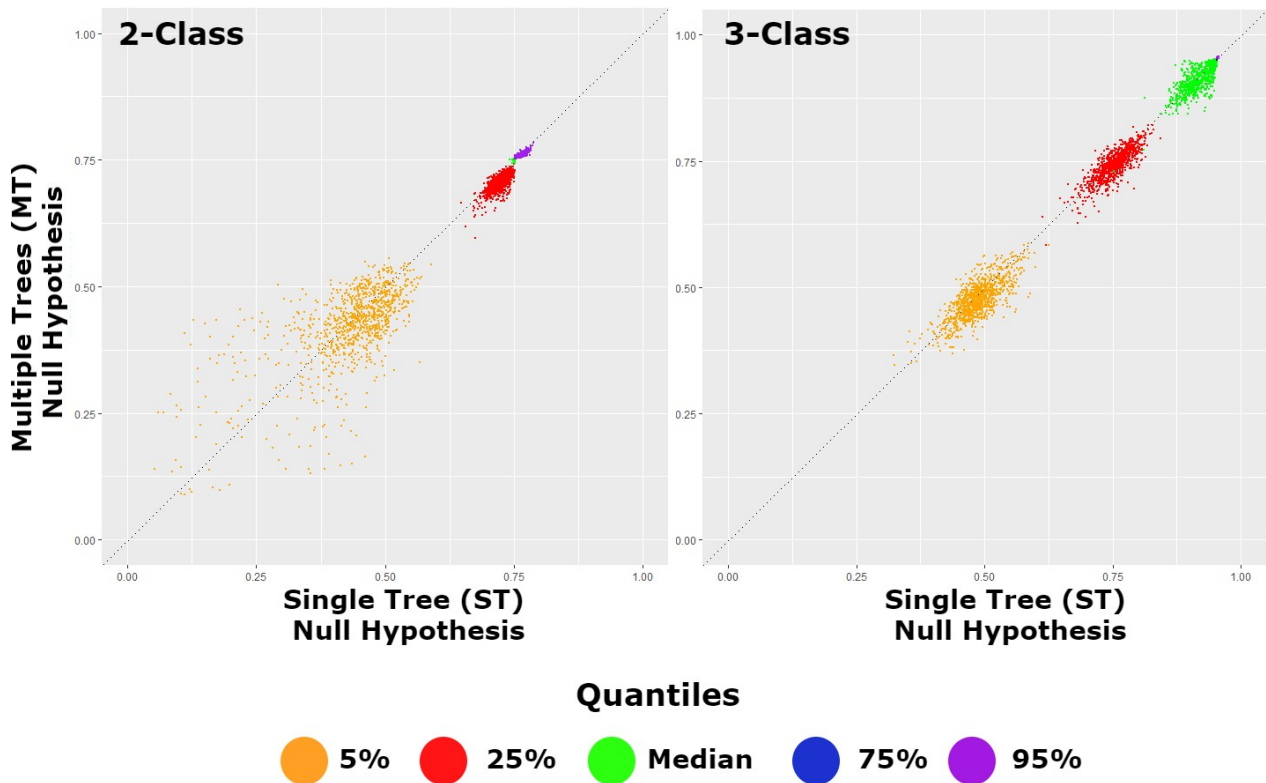
The null distribution is the base for hypothesis testing. To obtain it we randomly assigned character traits to the tree tips and then calculated the resulting entropy. We did this for the standard method and for the the multiple trees. We compare the resulting null-distributions in terms of their 5%, 25%, 75% and 95% and median. The expectations is that if these statistics are similar, then the null distributions generated by these two methods are similar.

Since the quantiles show very similar values across methods [Table 3.1], we conclude that the null distribution might not be responsible for differences in the detected target between the methods and, instead, this variation might be because of the way we calculate the  $\delta$  statistic. As such, both methods can be used to calculate the null-hypothesis, and this would result in similar values.

Quantile divided	2-class mean	2-class median	3-class mean	3-class median
5%	1.0170	0.9688	0.9778	0.9761
25%	0.9750	0.9764	0.9861	0.9874
50%	0.9997	0.9999	0.9944	0.9959
75%	1.0000	1.0000	0.9999	1.0000
95%	0.9976	0.9985	1.0002	1.0003

**Table 3.1:** Resulting values of the calculation [Quantile of ST method / Quantile of MT method]

Additionally, when analyzing their distribution, their results also seem to be relatively congruent, not changing much except the quantile 5% that, in the 2-class trait, seems to vary considerably among methods but a trend was not observed [Figure 3.4].



**Figure 3.4:** Distribution of the null hypothesis results in the different methods and k-class traits, when analyzing the sample of 1000 markers.

The left quantiles determine the significance of the obtained entropy values. The quantile 5% is especially important since it is generally the limit to reject the null hypothesis. To better understand this distribution and its general tendencies, we analyzed the overall probability that the obtained entropy, calculated for the real data, was not due to chance.

### 3.1.3 Probability value

It is important to ascertain if the results obtained under the null hypothesis are not the product of chance. To determine that, we calculate the probability  $p$  of observing the empirical entropy value in the null distribution. If it falls in the left region of this distribution, then the entropies are smaller than expected by chance, and the character has phylogenetic signal. Thus smaller the values of  $p$ , the more associated the character is with the phylogeny.

As such, the number of trees from each genetic marker's null-hypothesis distribution that had a value under the approaches' (mean for the multiple trees method and, in the MAP method, only a value existed) was noted.



**Figure 3.5:** Percentage of trees from the null-hypothesis distributions that have a lower than the one obtained from the methods

Most of the results, when analyzed by the new method, tend to approximate the significance level, by having a lower value. Although we can clearly see this in the 3-class trait distribution, since the lower values are mostly agglomerated, an additional table was created with a 5% significance level to compare them more thoroughly [Table 3.2].

<b>Number of points 3-Class trait</b>	<b>MT &lt;5% [ 893 ]</b>	<b>MT &gt;5% [ 107 ]</b>
<b>ST &lt;5% [ 695 ]</b>	688	7
<b>ST &gt;5% [ 305 ]</b>	205	100

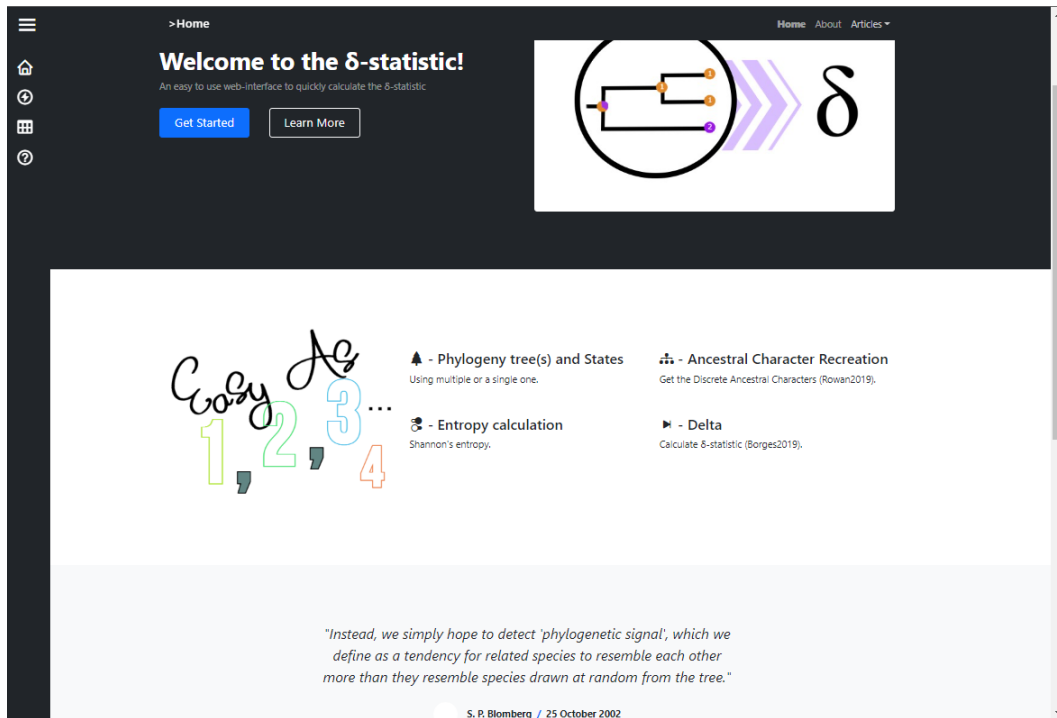
**Table 3.2:** Quadrants of the 3-Class trait's significant level

The results indicate that, by considering the statistical uncertainty present in the phylogenetic trees, we are gaining more biological information and can better explain the data. Consequently, the new method is stronger at assessing phylogenetic signal.

A gene-specific feature that we have tested was the evolutionary rate. Genes evolve differently by incorporating substitutions at different rates; this is expected to have an impact on the estimated entropies. However, no correlation was found between the evolutionary rate and the two methods we employed to measure the entropies. This might indicate that, independently from its values, the new method does not distort the results.

### 3.2 Website Interface

A web interface was created to increase the accessibility and reproducibility of the  $\delta$  statistic, facilitating its use to the evolutionary community. The code of this project was made available at the Github repository: <https://github.com/diogo-s-ribeiro/delta-statistic>



**Figure 3.6:** Image of the home page from the  $\delta$  statistic web interface implemented in Django.

The home page [Figure 3.6] was created with the objective of being a core page that has direct access to multiple relevant links. Additionally, an interactive sidebar was implemented with JavaScript for the user's easier navigation across all the web pages.

Furthermore, a dedicated page with the sole aim to calculate the  $\delta$  statistic was implemented. The user can set the parameters, tuning the overall calculation of the statistic to his preferences and, after this adjustment is made, the website is automatically re-directed to the result page, where the final results are presented [Figure 3.8].

> Calculate > Results


Phylogeny	Lambda0	Se	Sim	Thin	Burn	Date
A	1.0	2.0	100	2	2	Sept. 29, 2022

Delta:  
0.20394790835347462

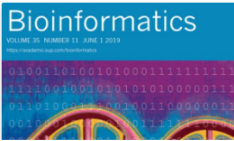
Add Instance

### Important Links


There are multiple important links but these are my recommendation!



**Molecular Biology and Evolution**  
A Fast Likelihood Method to Reconstruct and Visualize Ancestral Scenarios.  
Kevin Rowan



**Bioinformatics**  
Measuring phylogenetic signal between categorical traits and phylogenies.  
Rui Boroos



**Welcome to the  $\delta$ -statistic!**  
Assessing traits and phylogenetic signal to unravel the tempo and mode of phenotypic evolution.  
Dioao Ribeiro

**Figure 3.7:** Image of the result page from the  $\delta$  statistic web interface implemented in Django.

Finally, in the web interface, these results can be added to a data table that stores the calculated  $\delta$  statistics and their respective parameters. These instances can be easily inserted, deleted and updated. The totality of the result table' metadata, can also simply be download in this page.

> Data

## Results

Show 10 entries

Name	Delta	Date
<a href="#">File_13</a>	1.35	June 28, 2022, 10:20 a.m.
<a href="#">File_22</a>	1.34	June 28, 2022, 10:20 a.m.
<a href="#">Mammals in here</a>	1.34	June 28, 2022, 10:20 a.m.
<a href="#">File_12</a>	1.34	June 28, 2022, 10:19 a.m.
<a href="#">221</a>	111	Aug. 29, 2022, 8:58 p.m.
<a href="#">A</a>	1	Aug. 29, 2022, 8:57 p.m.
<a href="#">asdsadsadsad</a>	123	Aug. 29, 2022, 8:57 p.m.
<a href="#">Test</a>	1.3577144564543966	April 26, 2022, 2:10 p.m.
<a href="#">Mammal</a>	1.35	April 21, 2022, 2:47 p.m.
<a href="#">File_3</a>	1.34523	April 21, 2022, 2:20 p.m.

Showing 1 to 10 of 11 entries

Previous 1 2 Next

Add Instance

Download

Download all data:

TXT File

CSV File

**Figure 3.8:** Image of the data page from the  $\delta$  statistic web interface implemented in Django.



# Chapter 4

## Conclusion

In this thesis, we improved computational and statistical shortcomings of the widely used statistic of phylogenetic signal, the  $\delta$  statistic.

We expanded the statistic to allow it to deal with stringent assumptions regarding the phylogenetic history of phenotypic traits. The  $\delta$  statistic was extended to consider the uncertainty present in the MAP phylogenetic tree by instead analyzing multiple possible evolutionary histories. As such, the new entropies reflect several hypothetical phylogenies in a more practical manner, and thus providing a statistical sounder approach for measuring entropy and assess phylogenetic signal with increased accuracy.

Computational changes were also made, improving on one hand the accessibility and reproducibility with the implementation of a easy-to-use interface and, on the other hand, optimizing the algorithm's code to allow its use in current day genomics. Both of these aspects are very important nowadays.

By implementing the web interface, we allow numerous researchers without informatics or programming expertise in data analysis to keep up to date with current methods, which are only available through computational means. This consequently expands the possibility of a science breakthrough and reduces the risk of incorrectly analyzing data.

Furthermore, science nowadays generate an ever-increasing amount of genomic data in both quantity and quality. As such, fast and accurate methods are a must when dealing with large-scale genomic datasets. Thus, more efficient computational implementation of these statistics will considerably expedite the interpretation of genomic data in evolutionary biology.

## 4.1 Future Remarks

Despite strengthening computation and statistics limitations of the  $\delta$  statistic, several improvements can still be considered and tested in the future.

For instance, testing different types of entropy might be beneficial to see if the resulting values in the  $\delta$  statistic change significantly when compared with each other.

Additionally, in the future, if the null hypothesis is obtained through a mathematical way, a lot of time and computational power that is currently dedicated to obtaining these distributions through permutations, can be saved. Currently, this step is the most time- and resource-consuming.

Finally, an implementation of the developed web interface on a stable and fast web server is desired, as this would speed up the calculation of this new statistic and will allow its use even more efficiently than the current implementation.

# Chapter 5

## References

- [1] M Pagel. “The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies”. In: *JSTOR* 48 (3 1999), pp. 612–622. URL: <https://www.jstor.org/stable/2585328>.
- [2] Jonathan B. Losos. “Phylogenetic niche conservatism, phylogenetic signal and the relationship between phylogenetic relatedness and ecological similarity among species”. In: *Ecology Letters* 11 (10 Oct. 2008), pp. 995–1003. ISSN: 1461-0248. DOI: 10.1111/J.1461-0248.2008.01229.X. URL: <https://onlinelibrary.wiley.com/doi/full/10.1111/j.1461-0248.2008.01229.x>  
<https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1461-0248.2008.01229.x>  
<https://onlinelibrary.wiley.com/doi/10.1111/j.1461-0248.2008.01229.x>.
- [3] “How to measure and test phylogenetic signal”. In: *Methods in Ecology and Evolution* 3 (4 Aug. 2012), pp. 743–756. ISSN: 2041-210X. URL: <https://besjournals.onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2012.00196.x>.
- [4] Robin L. Chazdon et al. “Community and phylogenetic structure of reproductive traits of woody species in wet tropical forests”. In: *Ecological Monographs* 73 (3 Aug. 2003), pp. 331–348. ISSN: 00129615. DOI: 10.1890/02-4037.
- [5] Jeannine Cavender-Bares, Adrienne Keen, and Brianna Miles. “PHYLOGENETIC STRUCTURE OF FLORIDIAN PLANT COMMUNITIES DEPENDS ON TAXONOMIC AND SPATIAL SCALE”. In: *Ecology* 87 (7 2006), pp. 109–122. DOI: 10.1890/0012-9658.
- [6] Jason M. Kamilar and Kathleen M. Muldoon. “The Climatic Niche Diversity of Malagasy Primates: A Phylogenetic Perspective”. In: *PLOS ONE* 5 (6 2010), e11073. ISSN: 1932-6203. DOI: 10.1371/JOURNAL.PONE.0011073. URL: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0011073>.
- [7] Felsenstein J. “Phylogenies and quantitative characters”. In: *JSTOR* 19 (1988), pp. 445–471. URL: <https://www.jstor.org/stable/2097162>.
- [8] S. P. Blomberg and T. Garland. “Tempo and mode in evolution: Phylogenetic inertia, adaptation and comparative methods”. In: *Journal of Evolutionary Biology* 15 (6 Nov. 2002), pp. 899–910. ISSN: 1010061X. DOI: 10.1046/J.1420-9101.2002.00472.X.
- [9] Rui Borges et al. “Measuring phylogenetic signal between categorical traits and phylogenies”. In: *Bioinformatics* 35 (11 June 2019), pp. 1862–1869. ISSN: 1367-4803. DOI: 10.1093/BIOINFORMATICS/BTY800. URL: <https://academic.oup.com/bioinformatics/article/35/11/1862/5144670>.

- [10] Jason M. Kamilar and Natalie Cooper. "Phylogenetic signal in primate behaviour, ecology and life history". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 368 (1618 May 2013). ISSN: 14712970. DOI: 10 . 1098 / RSTB . 2012 . 0341. URL: <https://royalsocietypublishing.org/doi/10.1098/rstb.2012.0341>.
- [11] MP Simmons and H Ochoterena. "Gaps as characters in sequence-based phylogenetic analyses". In: *JSTOR* (2000). URL: <https://www.jstor.org/stable/2585224>.
- [12] Peter B. Pearman et al. "Niche dynamics in space and time". In: *Trends in Ecology & Evolution* 23 (3 Mar. 2008), pp. 149–158. ISSN: 0169-5347. DOI: 10.1016/J.TREE.2007.11.005.
- [13] P. A. P. Moran. "Notes on continuous stochastic phenomena". In: *JSTOR* 37 (1 1950), pp. 17–23. URL: <https://www.jstor.org/stable/2332142>.
- [14] Abouheif E. "A method for testing the assumption of phylogenetic independence in comparative data". In: *evolutionary-ecology.com* (1999). URL: <http://www.evolutionary-ecology.com/abstracts/v01/1152.html>.
- [15] Simon P. Blomberg, Theodore Garland, and Anthony R. Ives. "Testing for phylogenetic signal in comparative data: Behavioral traits are more labile". In: *Evolution* 57 (4 Apr. 2003), pp. 717–745. ISSN: 00143820. DOI: 10 . 1111 / J . 0014 - 3820 . 2003 . TB00285 . X. URL: <http://www.biology.ucr.edu/faculty/Garland/>.
- [16] SA Fritz, A Purvis - Conservation Biology, and undefined 2010. "Selectivity in mammalian extinction risk and threat types: a new measure of phylogenetic signal strength in binary traits". In: *Wiley Online Library* 24 (4 Aug. 2010), pp. 1042–1051. DOI: 10 . 1111 / j . 1523 - 1739 . 2010 . 01455 . x. URL: <https://conbio.onlinelibrary.wiley.com/doi/abs/10.1111/j.1523-1739.2010.01455.x>.
- [17] R.P. Haining. "Spatial Autocorrelation". In: *International Encyclopedia of the Social & Behavioral Sciences* (2001), pp. 14763–14768. DOI: 10.1016/B0-08-043076-7/02511-0.
- [18] Daniel A. Griffith. "Spatial Autocorrelation". In: *Encyclopedia of Social Measurement* (Jan. 2004), pp. 581–590. DOI: 10.1016/B0-12-369398-5/00334-0.
- [19] JL Gittleman and M Kot. "Adaptation: statistics and a null model for estimating phylogenetic effects". In: *academic.oup.com* (1990). URL: <https://academic.oup.com/sysbio/article-abstract/39/3/227/1727758>.
- [20] Emília P. Martins. "Phylogenies, spatial autoregression, and the comparative method: A computer simulation test". In: *Evolution* 50 (5 1996), pp. 1750–1765. ISSN: 00143820. DOI: 10.1111/J.1558-5646.1996.TB03562.X.
- [21] LJ Revell, LJ Harmon, and DC Colla. "Phylogenetic signal, evolutionary process, and rate". In: *academic.oup.com* (2008). URL: <https://academic.oup.com/sysbio/article-abstract/57/4/591/1631730>.
- [22] N Cooper et al. "Phylogenetic comparative approaches for studying niche conservatism". In: *Wiley Online Library* 23 (12 Dec. 2010), pp. 2529–2539. DOI: 10.1111/j.1420-9101.2010.02144.x. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1420-9101.2010.02144.x>.

- [23] Marti J. Anderson and Pierre Legendre. “An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model”. In: <http://dx.doi.org/10.1080/00949659908811936> 62 (3 2007), pp. 271–303. ISSN: 00949655. DOI: 10.1080/00949659908811936. URL: <https://www.tandfonline.com/doi/abs/10.1080/00949659908811936>.
- [24] Kellie Ottoboni et al. “An Empirical Comparison of Parametric and Permutation Tests for Regression Analysis of Randomized Experiments”. In: (2017).
- [25] J. Felsenstein. “Phylogenies and the comparative method.” In: *American Naturalist* 125 (1 1985), pp. 1–15. ISSN: 00030147. DOI: 10.1086/284325.
- [26] Bo Xu, Xuyan Feng, and Rebecca D. Burdine. “Categorical Data Analysis in Experimental Biology”. In: *Developmental biology* 348 (1 Dec. 2010), p. 3. ISSN: 1095564X. DOI: 10.1016/j.ydbio.2010.08.018. URL: [/pmc/articles/PMC3021327/%20/pmc/articles/PMC3021327/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3021327/](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3021327/).
- [27] DM Hillis and JP Huelsenbeck. “Signal, noise, and reliability in molecular phylogenetic analyses”. In: *academic.oup.com* (1992). URL: <https://academic.oup.com/jhered/article-abstract/83/3/189/856542>.
- [28] MJ Donoghue and RH Ree. “Homoplasy and developmental constraint: a model and an example from plants”. In: *academic.oup.com* (2000). URL: <https://academic.oup.com/icb/article-abstract/40/5/759/157156>.
- [29] DD Ackerly and R Nyffeler. “Evolutionary diversification of continuous traits: phylogenetic tests and application to seed size in the California flora”. In: *Springer* 18 (3 May 2004), pp. 249–272. DOI: 10.1023/B:EVEC.0000035031.50566.60. URL: <https://link.springer.com/article/10.1023/B:EVEC.0000035031.50566.60>.
- [30] TF Hansen et al. “A comparative method for studying adaptation to a randomly evolving environment”. In: *Wiley Online Library* 62 (8 Aug. 2008), pp. 1965–1977. DOI: 10.1111/j.1558-5646.2008.00412.x. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1558-5646.2008.00412.x>.
- [31] Thomas F. Hansen and Emília P. Martins. “TRANSLATING BETWEEN MICROEVOLUTIONARY PROCESS AND MACROEVOLUTIONARY PATTERNS: THE CORRELATION STRUCTURE OF INTERSPECIFIC DATA”. In: *Evolution* 50 (4 Aug. 1996), pp. 1404–1417. ISSN: 1558-5646. DOI: 10.1111/J.1558-5646.1996.TB03914.X. URL: <https://onlinelibrary.wiley.com/doi/full/10.1111/j.1558-5646.1996.tb03914.x> <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1558-5646.1996.tb03914.x> <https://onlinelibrary.wiley.com/doi/10.1111/j.1558-5646.1996.tb03914.x>.
- [32] Russell Lande. “Natural Selection and Random Genetic Drift in Phenotypic Evolution”. In: *Evolution* 30 (2 June 1976), p. 314. ISSN: 00143820. DOI: 10.2307/2407703.

- [33] Thomas H. Jukes and Charles R. Cantor. “Evolution of protein molecules”. In: *Mammalian Protein Metabolism*. Ed. by H.N. Munro. New York: Elsevier, 1969, pp. 21–132. DOI: 10.1016/B978-1-4832-3211-9.50009-7. URL: <https://linkinghub.elsevier.com/retrieve/pii/B9781483232119500097>.
- [34] Masami Hasegawa, Hirohisa Kishino, and Taka aki Yano. “Dating of the human-ape splitting by a molecular clock of mitochondrial DNA”. In: *Journal of Molecular Evolution* 1985 22:2 22 (2 Oct. 1985), pp. 160–174. ISSN: 1432-1432. DOI: 10.1007/BF02101694. URL: <https://link.springer.com/article/10.1007/BF02101694>.
- [35] S Tavaré. “Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences”. In: *Lectures on Mathematics in the Life Sciences* 17 (1986), pp. 57–86. ISSN: 00219150. DOI: citeulike-article-id:4801403. URL: <http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20%7B%5C%7Dpath=ASIN/0821811673>.
- [36] Paul O. Lewis. “A Likelihood Approach to Estimating Phylogeny from Discrete Morphological Character Data”. In: *Systematic Biology* 50 (6 Nov. 2001), pp. 913–925. ISSN: 1063-5157. DOI: 10.1080/106351501753462876. URL: <https://academic.oup.com/sysbio/article/50/6/913/1628902>.
- [37] Sohta A. Ishikawa et al. “A Fast Likelihood Method to Reconstruct and Visualize Ancestral Scenarios”. In: *Molecular Biology and Evolution* 36 (9 Sept. 2019), pp. 2069–2085. ISSN: 0737-4038. DOI: 10.1093/MOLBEV/MSZ131. URL: <https://academic.oup.com/mbe/article/36/9/2069/5498561>.
- [38] Jeffrey B. Joy et al. “Ancestral Reconstruction”. In: *PLOS Computational Biology* 12 (7 July 2016), e1004763. ISSN: 1553-7358. DOI: 10.1371/JOURNAL.PCBI.1004763. URL: <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004763>.
- [39] Richard H. Ree and Stephen A. Smith. “Maximum Likelihood Inference of Geographic Range Evolution by Dispersal, Local Extinction, and Cladogenesis”. In: *Systematic Biology* 57 (1 Feb. 2008), pp. 4–14. ISSN: 1063-5157. DOI: 10.1080/10635150701883881. URL: <https://academic.oup.com/sysbio/article/57/1/4/1703014>.
- [40] Michael J. Harms and Joseph W. Thornton. “Analyzing protein structure and function using ancestral gene reconstruction”. In: *Current Opinion in Structural Biology* 20 (3 June 2010), pp. 360–366. ISSN: 0959-440X. DOI: 10.1016/J.SBI.2010.03.005.
- [41] JA Finarelli and JJ Flynn. “Ancestral state reconstruction of body size in the Caniformia (Carnivora, Mammalia): the effects of incorporating data from the fossil record”. In: *academic.oup.com* (2006). URL: <https://academic.oup.com/sysbio/article-abstract/55/2/301/1623326>.
- [42] James S. Albert, Derek M. Johnson, and Jason H. Knouft. “Fossils provide better estimates of ancestral body size than do extant taxa in fishes”. In: *Acta Zoologica* 90 (SUPPL. 1 May 2009), pp. 357–384. ISSN: 00017272. DOI: 10.1111/J.1463-6395.2008.00364.X.
- [43] Graham J. Slater, Luke J. Harmon, and Michael E. Alfaro. “Integrating fossils with molecular phylogenies improves inference of trait evolution”. In: *Evolution* 66 (12 Dec. 2012), pp. 3931–3944. ISSN: 00143820. DOI: 10.1111/J.1558-5646.2012.01723.X.

- [44] DL Swofford and WP Maddison. "Reconstructing ancestral character states under Wagner parsimony". In: *Elsevier* (1987). URL: <https://www.sciencedirect.com/science/article/abs/pii/0025556487900745>.
- [45] J Felsenstein and J Felenstein. "Inferring phylogenies". In: (2004). URL: <https://www.sinauer.com/media/wysiwyg/tocs/InferringPhylogenies.pdf>.
- [46] JP Huelsenbeck and JP Bollback. "Empirical and hierarchical Bayesian estimation of ancestral states". In: *academic.oup.com* 50 (3 2001), pp. 351–366. URL: <https://academic.oup.com/sysbio/article-abstract/50/3/351/1661227>.
- [47] M Pagel, A Meade, and D Barker. "Bayesian estimation of ancestral character states on phylogenies". In: *academic.oup.com* (2004). URL: <https://academic.oup.com/sysbio/article-abstract/53/5/673/2842847>.
- [48] David Sankoff. "MINIMAL MUTATION TREES OF SEQUENCES." In: *SIAM Journal on Applied Mathematics* 28 (1 1975), pp. 35–42. ISSN: 00361399. DOI: 10.1137/0128004.
- [49] Dolph Schluter et al. "Likelihood of ancestor states in adaptive radiation". In: *Evolution* 51 (6 1997), pp. 1699–1711. ISSN: 00143820. DOI: 10.1111/J.1558-5646.1997.TB05095.X.
- [50] CW Cunningham, KE Omland, and TH Oakley. "Reconstructing ancestral character states: a critical reappraisal". In: *Elsevier* (1998). URL: <https://www.sciencedirect.com/science/article/pii/S0169534798013822>.
- [51] Z Yang. "PAML 4: phylogenetic analysis by maximum likelihood". In: *academic.oup.com* (2007). URL: <https://academic.oup.com/mbe/article-abstract/24/8/1586/1103731>.
- [52] T Pupko et al. "A fast algorithm for joint reconstruction of ancestral amino acid sequences". In: *academic.oup.com* (2000). URL: <https://academic.oup.com/mbe/article-abstract/17/6/890/1037793>.
- [53] Z Yang and S Kumar. "A new method of inference of ancestral nucleotide and amino acid sequences." In: *academic.oup.com* (1995). URL: <https://academic.oup.com/genetics/article-abstract/141/4/1641/6061971>.
- [54] Emmanuel Paradis, Julien Claude, and Korbinian Strimmer. "APE: Analyses of Phylogenetics and Evolution in R language". In: *Bioinformatics* 20 (2 Jan. 2004), pp. 289–290. ISSN: 1367-4803. DOI: 10.1093/BIOINFORMATICS/BTG412. URL: <https://academic.oup.com/bioinformatics/article/20/2/289/204981>.
- [55] D. Karolchik et al. "The UCSC Genome Browser Database". In: *Nucleic acids research* 31 (1 Jan. 2003), pp. 51–54. ISSN: 1362-4962. DOI: 10.1093/NAR/GKG129. URL: <https://pubmed.ncbi.nlm.nih.gov/12519945/>.
- [56] Kim D. Pruitt, Tatiana Tatusova, and Donna R. Maglott. "NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins". In: *Nucleic acids research* 33 (Database issue Jan. 2005). ISSN: 1362-4962. DOI: 10.1093/NAR/GKI025. URL: <https://pubmed.ncbi.nlm.nih.gov/15608248/>.
- [57] James Taylor et al. "Using Galaxy to Perform Large-Scale Interactive Data Analyses". In: *Current Protocols in Bioinformatics* 19 (1 Sept. 2007). ISSN: 1934-3396. DOI: 10.1002/0471250953.BI1005S19.

- [58] Jeremy Goecks et al. “Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences”. In: *Genome Biology* 11 (8 Aug. 2010), pp. 1–13. ISSN: 1474760X. DOI: 10.1186/GB-2010-11-8-R86/TABLES/1. URL: <https://link.springer.com/articles/10.1186/gb-2010-11-8-r86> %20https://link.springer.com/article/10.1186/gb-2010-11-8-r86.
- [59] Mathieu Fourment and Michael R. Gillings. “A comparison of common programming languages used in bioinformatics”. In: *BMC Bioinformatics* 9 (1 Feb. 2008), pp. 1–9. ISSN: 14712105. DOI: 10.1186/1471-2105-9-82/TABLES/1. URL: <https://link.springer.com/articles/10.1186/1471-2105-9-82> %20https://link.springer.com/article/10.1186/1471-2105-9-82.
- [60] *TIOBE Index - TIOBE*. Accessed: 2022-09-30. URL: <https://www.tiobe.com/tiobe-index/>.
- [61] *PYPL Popularity of Programming Language index*. Accessed: 2022-09-30. URL: <https://pypl.github.io/PYPL.html>.
- [62] Django Software Foundation. *Django*. Version 2.2. May 5, 2019. URL: <https://djangoproject.com>.
- [63] Celine Scornavacca et al. “OrthoMaM v10: Scaling-Up Orthologous Coding Sequence and Exon Alignments with More than One Hundred Mammalian Genomes”. In: *Molecular Biology and Evolution* 36 (4 Apr. 2019), pp. 861–862. ISSN: 0737-4038. DOI: 10.1093/MOLBEV/MSZ015. URL: <https://academic.oup.com/mbe/article/36/4/861/5303840>.
- [64] Sebastian Höhna et al. “RevBayes: Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification Language”. In: *Systematic Biology* 65 (4 July 2016), pp. 726–736. ISSN: 1063-5157. DOI: 10.1093/SYSBIO/SYW021. URL: <https://academic.oup.com/sysbio/article/65/4/726/1753608>.
- [65] Ziheng Yang. “Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods”. In: *Journal of Molecular Evolution* 1994 39:3 39 (3 1994), pp. 306–314. ISSN: 1432-1432. DOI: 10.1007/BF00160154. URL: <https://link.springer.com/article/10.1007/BF00160154>.
- [66] John K. Kruschke. “Markov Chain Monte Carlo”. In: *Doing Bayesian Data Analysis* (2015), pp. 143–191. DOI: 10.1016/B978-0-12-405888-0.00007-6.
- [67] Andrew Rambaut et al. “Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7”. In: *Systematic Biology* 67 (5 Sept. 2018), pp. 901–904. ISSN: 1063-5157. DOI: 10.1093/SYSBIO/SYY032. URL: <https://academic.oup.com/sysbio/article/67/5/901/4989127>.
- [68] S. R. Hinton. “ChainConsumer”. In: *The Journal of Open Source Software* 1, 00045 (Aug. 2016), p. 00045. DOI: 10.21105/joss.00045.
- [69] Andrew Gelman and Donald B. Rubin. “Inference from Iterative Simulation Using Multiple Sequences”. In: <https://doi.org/10.1214/ss/1177011136> 7 (4 Nov. 1992), pp. 457–472. ISSN: 0883-4237. DOI: 10.1214/SS/1177011136. URL: [https://projecteuclid.org/journals/statistical-science/volume-7/issue-4/Inference-from-Iterative-Simulation-Using-Multiple-Sequences/10.1214/ss/1177011136](https://projecteuclid.org/journals/statistical-science/volume-7/issue-4/Inference-from-Iterative-Simulation-Using-Multiple-Sequences/10.1214/ss/1177011136.full) . full %20https :



//projecteuclid.org/journals/statistical-science/volume-7/issue-4/Inference-from-Iterative-Simulation-Using-Multiple-Sequences/10.1214/ss/1177011136.short.

- [70] John F. Geweke. “Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments”. In: *Staff Report* (1991). URL: <https://ideas.repec.org/p/fip/fedmsr/148.html>.
- [71] Christopher J. Yahnke, Tanya Dewey, and Phil Myers. “Animal Diversity Web as a Teaching & Learning Tool to Improve Research & Writing Skills in College Biology Courses”. In: *The American Biology Teacher* 75 (7 Sept. 2013), pp. 494–498. ISSN: 0002-7685. DOI: 10.1525/ABT.2013.75.7.9. URL: [/abt/article/75/7/494/18647/Animal-Diversity-Web-as-a-Teaching-amp-Learning%20https://online.ucpress.edu/abt/article/75/7/494/18647/Animal-Diversity-Web-as-a-Teaching-amp-Learning](https://online.ucpress.edu/abt/article/75/7/494/18647/Animal-Diversity-Web-as-a-Teaching-amp-Learning%20https://online.ucpress.edu/abt/article/75/7/494/18647/Animal-Diversity-Web-as-a-Teaching-amp-Learning).
- [72] Charles R. Harris et al. “Array programming with NumPy”. In: *Nature* 585.7825 (Sept. 2020), pp. 357–362. DOI: 10.1038/s41586-020-2649-2. URL: <https://doi.org/10.1038/s41586-020-2649-2>.