

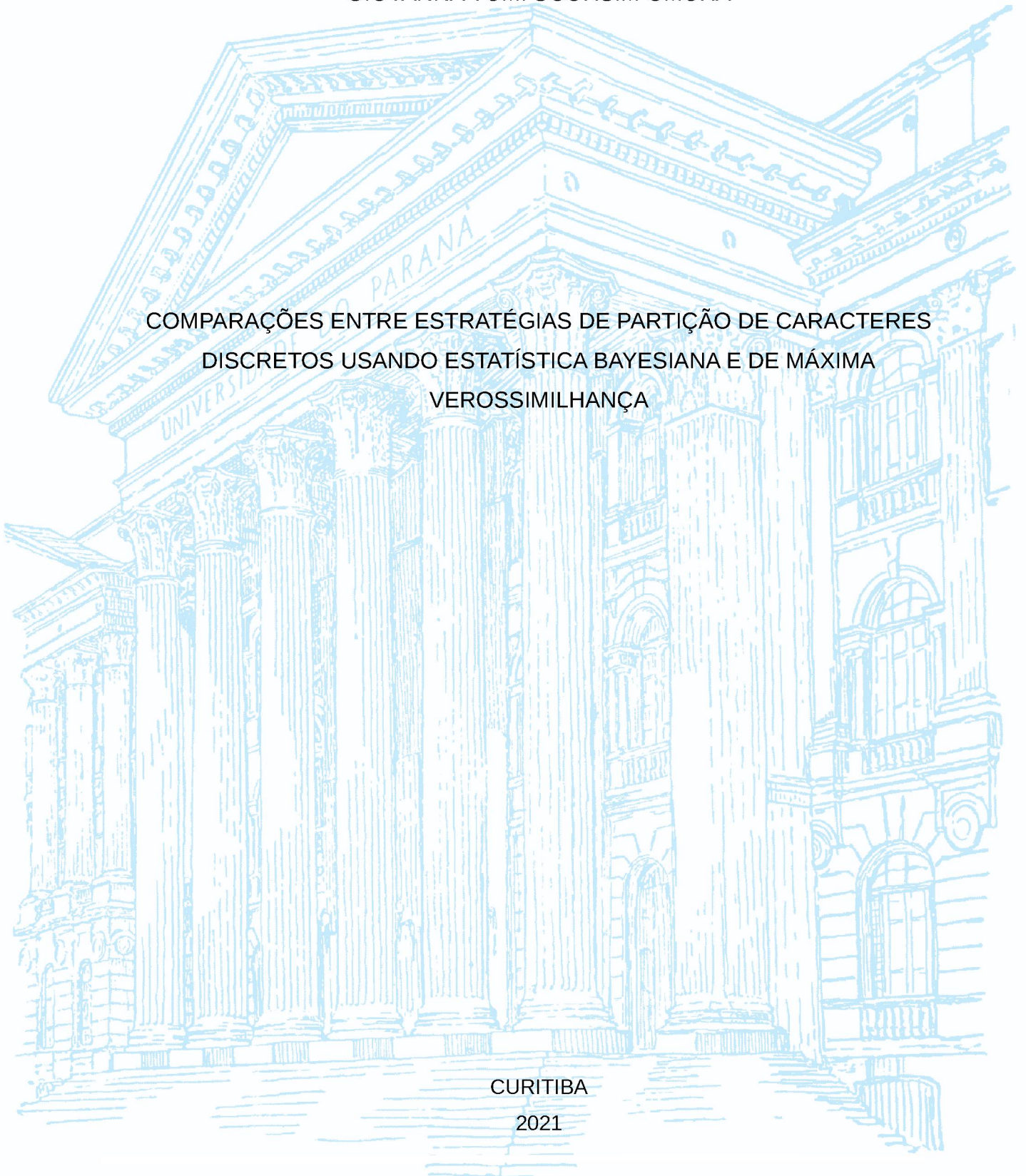
UNIVERSIDADE FEDERAL DO PARANÁ

GIOVANNA YUMI SCORSIM OMURA

COMPARAÇÕES ENTRE ESTRATÉGIAS DE PARTIÇÃO DE CARACTERES
DISCRETOS USANDO ESTATÍSTICA BAYESIANA E DE MÁXIMA
VEROSSIMILHANÇA

CURITIBA

2021



GIOVANNA YUMI SCORSIM OMURA

COMPARAÇÕES ENTRE ESTRATÉGIAS DE PARTIÇÃO DE CARACTERES
DISCRETOS USANDO ESTATÍSTICA BAYESIANA E DE MÁXIMA
VEROSSIMILHANÇA

Dissertação apresentada ao Curso de Pós-Graduação em Zoologia, Setor de Ciências Biológicas, da Universidade Federal do Paraná, como requisito parcial à obtenção do título de Mestre em Zoologia.

Orientador: Prof. Dr. Fabricius Maia Chaves Bicalho Domingos

Coorientador: Prof. Dr. Prof. Dr. Luiz Ricardo Lopes de Simone & Prof. Dr. Marcos Soares Barbeitos

CURITIBA

2021

Universidade Federal do Paraná. Sistema de Bibliotecas.
Biblioteca de Ciências Biológicas.
(Rosilei Vilas Boas – CRB/9-939).

Omura, Giovanna Yumi Scorsim.

Comparações entre estratégias de partição de caracteres discretos usando estatística Bayesiana e de Máxima Verossimilhança. / Giovanna Yumi Scorsim Omura. – Curitiba, 2021.

59 f. : il.

Orientador: Fabricius Maia Chaves Bicalho Domingos.

Coorientador: Luiz Ricardo Lopes de Simone.

Coorientador: Marcos Soares Barbeitos.

Dissertação (Mestrado) – Universidade Federal do Paraná, Setor de Ciências Biológicas. Programa de Pós-Graduação em Zoologia.

1. Bioinformática. 2. Filogenia. 3. Gastropoda. 4. Particionamento. I. Título. II. Domingos, Fabricius Maia Chaves Bicalho. III. Simone, Luiz Ricardo Lopes de. IV. Barbeitos, Marcos Soares. V. Universidade Federal do Paraná. Setor de Ciências Biológicas. Programa de Pós-Graduação em Zoologia.

ATA DE SESSÃO PÚBLICA DE DEFESA DE Mestrado PARA A OBTENÇÃO DO GRAU DE MESTRA EM ZOOLOGIA

No dia vinte e nove de setembro de dois mil e vinte e um às 14:00 horas, na sala Virtual, Plataforma TEAMS, foram instaladas as atividades pertinentes ao rito de defesa de dissertação da mestranda **GIOVANNA YUMI SCORSIM OMURA**, intitulada: **COMPARAÇÕES ENTRE ESTRATÉGIAS DE PARTIÇÃO DE CARACTERES DISCRETOS USANDO ESTATÍSTICA BAYESIANA E DE MÁXIMA VEROSSIMILHANÇA**, sob orientação do Prof. Dr. FABRICIUS MAIA CHAVES BICALHO

DOMINGOS. A Banca Examinadora, designada pelo Colegiado do Programa de Pós-Graduação em ZOOLOGIA da Universidade Federal do Paraná, foi constituída pelos seguintes Membros: FABRICIUS MAIA CHAVES BICALHO DOMINGOS (UNIVERSIDADE FEDERAL DO PARANÁ), TAUANA CUNHA (SMITHSONIAN TROPICAL RESEARCH INSTITUTE), JULIA KLACZKO

(UNIVERSIDADE DE BRASÍLIA). A presidência iniciou os ritos definidos pelo Colegiado do Programa e, após exarados os pareceres dos membros do comitê examinador e da respectiva contra argumentação, ocorreu a leitura do parecer final da banca examinadora, que decidiu pela APROVAÇÃO. Este resultado deverá ser homologado pelo Colegiado do programa, mediante o atendimento de todas as indicações e correções solicitadas pela banca dentro dos prazos regimentais definidos pelo programa. A outorga de título de mestra está condicionada ao atendimento de todos os requisitos e prazos determinados no regimento do Programa de Pós-Graduação. Nada mais havendo a tratar a presidência deu por encerrada a sessão, da qual eu, FABRICIUS MAIA CHAVES BICALHO DOMINGOS, lavrei a presente ata, que vai assinada por mim e pelos demais membros da Comissão Examinadora.

CURITIBA, 29 de Setembro de 2021.

Assinatura Eletrônica

04/10/2021 11:41:38.0

FABRICIUS MAIA CHAVES BICALHO DOMINGOS

Presidente da Banca Examinadora

Assinatura Eletrônica

04/10/2021 20:47:29.0

TAUANA CUNHA

Avaliador Externo (SMITHSONIAN TROPICAL RESEARCH INSTITUTE)

Assinatura Eletrônica

04/10/2021 11:37:29.0

JULIA KLACZKO

Avaliador Externo (UNIVERSIDADE DE BRASÍLIA)

Avenida Cel Francisco H Santos, s/n - CURITIBA - Paraná - Brasil

CEP 81531-980 - Tel: (41) 3361-1641 - E-mail: pgzoo@ufpr.br Documento assinado eletronicamente de acordo com o disposto na legislação federal Decreto 8539 de 08 de outubro de 2015.

Gerado e autenticado pelo SIGA-UFPR, com a seguinte identificação única: 116739

Para autenticar este documento/assinatura, acesse <https://www.prppg.ufpr.br/siga/visitante/autenticacaoassinaturas.jsp> e insira o código 116739

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em ZOOLOGIA da Universidade Federal do Paraná foram convocados para realizar a arguição da dissertação de Mestrado de **GIOVANNA YUMI SCORSIM OMURA** intitulada: **COMPARAÇÕES ENTRE ESTRATÉGIAS DE PARTIÇÃO DE CARACTERES DISCRETOS USANDO ESTATÍSTICA BAYESIANA E DE MÁXIMA VEROSSIMILHANÇA**, sob orientação do Prof. Dr. FABRICIUS MAIA CHAVES BICALHO DOMINGOS, que após terem inquirido a aluna e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de mestra está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

CURITIBA, 29 de Setembro de 2021.

Assinatura Eletrônica

04/10/2021 11:41:38.0

FABRICIUS MAIA CHAVES BICALHO DOMINGOS

Presidente da Banca Examinadora

Assinatura Eletrônica

04/10/2021 20:47:29.0

TAUANA CUNHA

Avaliador Externo (SMITHSONIAN TROPICAL RESEARCH INSTITUTE)

Assinatura Eletrônica

04/10/2021 11:37:29.0

JULIA KLACZKO

Avaliador Externo (UNIVERSIDADE DE BRASÍLIA)

Avenida Cel Francisco H Santos, s/n - CURITIBA - Paraná - Brasil

CEP 81531-980 - Tel: (41) 3361-1641 - E-mail: pgzoo@ufpr.br Documento assinado eletronicamente de acordo com o disposto na legislação federal Decreto 8539 de 08 de outubro de 2015.

Gerado e autenticado pelo SIGA-UFPR, com a seguinte identificação única: 116739

Para autenticar este documento/assinatura, acesse <https://www.prppg.ufpr.br/siga/visitante/autenticacaoassinaturas.jsp> e insira o código 116739

RESUMO

Atualmente, existe um claro renascimento no uso de matrizes morfológicas em análises filogenéticas modernas e complexas, com crescente importância em diferentes abordagens na Biologia Evolutiva. Um dos grandes avanços nas análises de dados morfológicos têm sido o uso de abordagens analíticas particionadas e, em particular, usando pesos implícitos de caracteres homoplásicos para determinar tais partições. Nesta dissertação, avaliamos o desempenho das abordagens de Inferência Bayesiana (BI) e Máxima Verossimilhança (ML) na análise de dados morfológicos sob diferentes estratégias de particionamento e usando diferentes conjuntos de modelos evolutivos. Também testamos a implementação do ModelFinder, um procedimento automatizado de seleção de partições e modelos de evolução implementado no software IQTree. Nossos resultados sugerem que os modelos de evolução inferidos pelo ModelFinder possuem desempenho particularmente bom quando a análise leva em conta diferentes partições que utilizam diferentes modelos evolutivos e onde os comprimentos de ramo podem ser estimados separadamente. O uso do ModelFinder e estimativas filogenéticas usando ML pode ser considerada uma alternativa viável e muito mais eficiente computacionalmente em comparação à implementação de pesos implícitos em reconstruções filogenéticas usando BI, pelo menos considerando análises de caracteres discretos de pequenas matrizes morfológicas.

Palavras-chave: Bioinformática. Filogenia. Gastropoda. Partição Homoplásica. Sistemática.

ABSTRACT

There is, currently, a renaissance in the use of morphological matrices in modern and complex phylogenetic analyses, and a growing importance in many Evolutionary Biology approaches. One type of such advances in the analyses of morphological data has been in the use of partitioned analytical approaches and, in particular, by using implicit weights of homoplastic characters to determine such partitions. In this dissertation, we evaluated the performance of Bayesian Inference (BI) and Maximum Likelihood (ML) approaches on the analyses of morphological data under different partitioning strategies and using different sets of evolutionary models. We also tested the implementation of ModelFinder, an automated partition selection procedure implemented in the software IQTree. Our results suggest that the evolutionary models inferred by ModelFinder are particularly good when different partitions are allowed to have different evolutionary models and branch lengths are separately estimated. The use of ModelFinder and phylogenetic estimation using ML approaches can be considered an important and more computationally efficient alternative to implementing implied weights under BI phylogenetic reconstructions, at least when applied to the analyses of discrete characters of small morphological matrices.

Keywords: Bioinformatics. Gastropoda. Homoplastic Partitioning. Phylogeny. Systematics.

LIST OF FIGURES

FIGURE 1 - Flow chart of study design outlining details of each step regarding the used datasets and phylogenetic analyses approaches.....	25
--	----

LIST OF TABLES

TABLE 1 - Maximum Likelihood values of ML analyses on the empirical datasets that contained invariable sites. The best model (higher ML value) for each model selection strategy is in boldface. The models and partition schemes selected by ModelFinder can be found in Table S3-1.....	26
TABLE 2 - Maximum Likelihood values of ML analyses on the empirical datasets that contained variable sites only. The best model (higher ML value) for each model selection strategy is in boldface. The models and partition schemes selected by ModelFinder can be found in Table S3-2.....	27
TABLE 3 - Maximum Likelihood values of ML analyses on the simulated datasets. The best model (higher ML value) for each model selection strategy is in boldface. The models and partition schemes selected by ModelFinder can be found in Table S3-3.....	27
TABLE 4 - Marginal Likelihood of Bayesian Inference analyses on the empirical datasets that contained invariable sites. The best model (higher MgL value) for each model selection strategy is in boldface. The models and partition schemes selected by ModelFinder can be found at Table S3-1.....	28
TABLE 5 - Marginal Likelihood of Bayesian Inference analyses on the empirical datasets that contained variable sites only. The best model (higher MgL value) for each model selection strategy is in boldface. The models and partition schemes selected by ModelFinder can be found at Table S3-2.....	29
TABLE 6 - Marginal Likelihood of Bayesian Inference analyses on the simulated datasets. The best model (higher MgL value) for each model selection strategy is in boldface. The models and partition schemes selected by ModelFinder can be found at Table S3-3.....	30
TABLE 7 - Normalized (Generalized) Jaccard Robinson-Foulds and Normalized (value=1) Matching Splits distances among the 'true' trees and the inferred tree using the simulated datasets under Maximum Likelihood and Bayesian Inference using the MK Model and ModelFinder. 'True' trees were named according to the respective empirical tree used to determine the number of taxa and characters, and tree format (symmetric or asymmetric). Smaller distances (depicting better analytical performance) of nGRF and higher similarity of nMS for each comparison are highlighted in boldface for each distance metric.....	32
TABLE 8 - Normalized (Generalized) Jaccard Robinson-Foulds and Normalized (value=1) Matching Splits distances of simulated datasets with Maximum Likelihood (ML) and Bayesian Inference (BI) comparing tree strategies using the MK Model and ModelFinder. Smaller	

distances (depicting better analytical performance) of nGRF and higher similarity of nMS for each comparison are highlighted in boldface for each distance metric.....33

TABLE 9 - Normalized (Generalized) Jaccard Robinson-Foulds and Normalized (value=1) Matching Splits distances of empirical datasets (invariable characters) with Maximum Likelihood (ML) and Bayesian Inference (BI) comparing tree strategies using the MK Model and ModelFinder. Smaller distances (depicting better analytical performance) of nGRF and higher similarity of nMS for each comparison are highlighted in boldface for each distance metric..... 34

TABLE 10 - Normalized (Generalized) Jaccard Robinson-Foulds and Normalized (value=1) Matching Splits distances of empirical datasets (variable characters only) with Maximum Likelihood (ML) and Bayesian Inference (BI) comparing tree strategies using the MK Model and ModelFinder. Smaller distances (depicting better analytical performance) of nGRF and higher similarity of nMS for each comparison are highlighted in boldface for each distance metric..... 35

SUMMARY

ATA DE SESSÃO PÚBLICA DE DEFESA DE MESTRADO PARA A OBTENÇÃO DO GRAU DE MESTRA EM ZOOLOGIA.....	4
TERMO DE APROVAÇÃO.....	5
1 INTRODUCTION.....	16
2 MATERIAL AND METHODS.....	21
2.1 DATASETS.....	21
2.1.1 <i>Empirical datasets</i>	21
2.1.2 Simulated datasets.....	21
2.2 PHYLOGENETIC ANALYSES.....	21
2.2.1 Homoplasy Partitioning.....	21
2.2.2 Maximum Likelihood.....	22
2.2.3 Bayesian Inference.....	22
2.3 TREE DISTANCES.....	23
3 RESULTS.....	26
3.1 EMPIRICAL AND SIMULATED DATASETS.....	26
3.2 TREE DISTANCES.....	31
4 DISCUSSION.....	36
5 CONCLUSION.....	40
REFERENCES.....	41
APPENDIX 1 – SUPPLEMENTARY TABLES.....	50
APPENDIX 2 – COMPUTATIONAL SCRIPTS.....	52

1 INTRODUCTION

Knowledge on phylogenetic relationships among organisms is currently one of the fundamental bases of evolutionary biology and life sciences studies (Smith et al, 2020). During the past four decades, the use of DNA sequences to reconstruct phylogenetic relationships became the norm in evolutionary biology (Lartillot, 2012). The progress made in terms of data collection (automated sequencing) was accompanied by the development of robust and accurate methods of phylogenetic reconstructions (Huelsenbeck & Ronquist, 2001; Stamatakis, 2014; Nguyen et al, 2015; Kozlov, 2019). However, when genetic data is missing, or impossible to be collected, different approaches to phylogenetic reconstructions using morphological data are necessary (O'Reilly, 2016), and these implementations depend on a different set of evolutionary models and analytical strategies that are usually quite different from DNA-based approaches (Lewis, 2001). Testing the efficacy of different evolutionary models should allow for stronger and trustworthy analytical approaches using morphological matrices, especially because different morphological characters will have different statistical properties (Puttick et al, 2017;2019). Hence, modern phylogenetic analyses should be capable of accommodating different evolutionary processes underlying the morphological data (Wright & Hillis, 2014).

Using morphological matrices in complex phylogenetic analyses has, thus, shown a renaissance in the past few years (Lee & Palci, 2015). Such implementations are invaluable when fossils or organisms that are only represented in biological collections are needed, such as to include mollusk shells, exoskeleton data, and detailed anatomical descriptions (O'Reilly et al, 2018). The use of this type of data alone, or in combination with genetic data, can provide many benefits to phylogenetic analytical approaches (Wortley & Scotland, 2006). Recent developments on the use of morphological data have facilitated the test of tree weights (Wright & Hills, 2014), improved divergence time estimates (Warnock & Wright, 2020; Álvarez-Carretero et al, 2019), provided better estimates of phylogenetic signal (May & Moore, 2019), helped in building stronger macroevolutionary models (Jia et al, 2014), and especially contributed to solving the relationships among groups that have consistently shown difficult estimations using genetic data alone (Berger & Stamatakis, 2010; Dornellas, 2020).

Thus, this renewed attention and increased importance on the use of morphological datasets led to a concomitant development of analytical methods in recent years (Lee & Palci, 2015). In particular, methodological advancements were made on the use of concatenated matrices of morphological and molecular data (McInerney et al, 2019; Rio et al, 2019; Álvarez-Carretero et al, 2019; Chazot et al, 2019), on the strategies used to infer phylogenies using discrete characters (Wright & Hills, 2014; Wright, 2019) and morphometric data (Parins-Fukuchi, 2018). Not surprisingly, many studies also focused on comparing the performance of the three most used phylogenetic reconstruction methods, i.e., Maximum Parsimony (MP), Maximum Likelihood (ML), and Bayesian Inference (BI), considering different evolutionary scenarios (Puttick et al, 2017; 2019; Vernygora et al, 2020; Spade, 2020).

One specific type of advance made in the analyses of morphological data has been in the use of partitioned analytical approaches (Rosa et al, 2019) and, in particular, by using implicit weights of homoplastic characters to determine these partitions (Goloboff et al, 2018b; Goloboff & Arias, 2019; Rosa et al, 2019). This protocol is performed by pondering the characters and generating different weights for groups of characters (partitions), which are based on their comparative levels of homoplasy. This strategy has been shown to produce trees that are more similar to the 'true' tree (Goloboff, 1993), likely because characters with a higher homoplastic tendency can produce a higher phylogenetic signal or evolve faster than non-homoplastic ones. Because equally weighted parsimony considers that all characters share the same rates of evolution on an MP analysis, it can produce less accurate and under-resolved trees (Smith et al, 2019). Thus, although implicit weights can increase the proportion of incorrect groups in relation to equal weights under MP, additional groups added to the matrix are more likely to be correct if a large number of characters is used (Goloboff, 2018b).

Historically, morphological matrices have been traditionally analyzed using MP analytical approaches (Felsenstein, 1978), and the implicit weights strategy was developed under this scenario (Goloboff, 1993). However, this homoplastic partitioning strategy was shown to be highly effective for model inference on empirical morphological datasets with different rates of evolution under a BI analytical approach (Rosa et al, 2019). Bayesian inference is, currently, probably the most used

methodology for the analyses of both molecular and morphological data (Kapli et al, 2020). This approach is based on estimating the posterior probability of tree parameters, in order to infer an optimal result in the stochastic field in the midst of observed conditional trees (Rannala & Yang, 1996). Its process consists of determining an *a priori* probability (prior probability), specified by the user, for a set of tree parameters, and then updating this probability to obtain the posterior probability according to the Bayes Theorem. Finally, a set of final trees that portray the best inferences is statistically obtained from the posterior distribution (Huelsenbeck & Ronquist, 2001). There are several parameters available to calibrate the Bayesian search, and modern implementations offer a wide range of possibilities for the use of morphological and/or molecular matrices to perform robust, complex, and highly customizable analyses (Álvarez-Carretero et al, 2019).

Many studies have compared the performance of BI and MP under different evolutionary scenarios (O'Reilly et al, 2016; 2018), and using both empirical and simulated morphological datasets (Puttick et al, 2017). Recent analytical approaches demonstrated that BI outperforms MP in most cases (Schrägo et al, 2018), such as using implied weighting (Rosa et al., 2019). However, comparisons with Maximum Likelihood (ML) analytical approaches using modern implementations are still scarce. The ML methods are performed in order to maximize the likelihood function and find the best tree in a topological and branch length search process. Evidence suggests that the analyses of discrete characters using ML methods will also perform better than parsimony searches (Lee & Worthy, 2012), but these results have been put into question (Xu & Pol, 2014). Modern ML implementations, such as those implemented in the software IQTree (Nguyen et al, 2015), have the advantage of enabling the incorporation of different evolutionary models, distributing different evolutionary rates among characters, and allowing for the inference of branch lengths at different tree levels (Wright & Hills, 2014).

In order to perform model-based phylogenetic reconstructions (BI and ML) in modern software, users will need to search for and evaluate the best-fitting evolutionary models to be used under partitioned or unpartitioned approaches (Kalyaanamoorthy et al, 2017). Compared to the use of genetic data, the application of ML and BI analyses on morphological characters is particularly limited in terms of the availability of different evolutionary models to estimate tree parameters. The most

commonly used model for morphological analyses is the Mk model, where 'M' stands for "Markov", and 'k' refers to the number of observed states that have the assumption of uniform equilibrium frequencies and equal substitution rates between all pairs of states (Lewis, 2001). Also, it is particularly common for morphological matrices to be composed exclusively of variable characters, which might lead to artificial increases in branch lengths during phylogenetic reconstruction (Vernygora et al, 2020). Hence, combining the Mk model with ascertainment bias correction is a common practice in order to alleviate this overestimation of divergence among taxa in morphological phylogenetic analyses (Harrison & Larsson, 2015). Other parameters that are known to influence the analyses like nucleotide or character frequencies (Abadi et al, 2019), the number of gamma rate distributions (Jia et al, 2014), and whether or not the model considers a proportion of invariable sites (Bromham et al, 2018) were mainly tested using molecular sequences, hampering our understanding of the statistical properties of morphological data under Bayesian and ML estimates.

In light of the above-mentioned advancements on phylogenetic inferences using morphological data, Bayesian Inference usually performs better when partitioning matrices applying implicit weights based on homoplasy, which is an adequate strategy for both large and small matrices. However, Rosa and collaborators (2019) based their performance comparisons on the overall parameter inference under distinct phylogenetic models and did not evaluate possible impacts on topology estimation. Moreover, the performance of homoplasy weighting and other inference-based partitioning strategies have not been evaluated under Maximum Likelihood approaches. Finally, other evolutionary models, apart from the overall used Mk model, might be applied to different partitions, and the performance of such models on matrices of discrete morphological characters has not been consistently evaluated.

Thus, in this dissertation, we evaluated the performance of BI and ML approaches on the analyses of morphological data under different partitioning strategies and using different sets of evolutionary models. We used three empirical Gastropoda matrices and simulated morphological datasets to compare the performance of both methods under different experimental combinations. Specifically, we aimed at 1) comparing the analytical efficacy of Maximum Likelihood (ML) and

Bayesian Inference (BI) approaches; 2) investigating the performance of those phylogenetic methodologies using different partitioning schemes and; 3) analyzing the effect that different evolutionary models might have on the accuracy of phylogenetic estimations.

2 MATERIAL AND METHODS

2.1 DATASETS

2.1.1 *Empirical datasets*

We analyzed three published phenotype multistate matrices of Gastropoda (Mollusca, Caenogastropoda): superfamily Cerithioidea J. Fleming 1822, composed of 23 species and 122 characters (Simone, 2001); order Architaenoglossa Haller 1890, with 20 species and 143 characters (Simone, 2004); and superfamily Stromboidea Rafinesque 1815, with 21 species and 102 characters (Simone, 2005). Those matrices were previously published and generated through detailed morphological studies, with descriptions of internal anatomy and shell morphology (Simone 2001; 2004; 2005). They were analyzed in the original manuscripts using Maximum Parsimony but were never subjected to re-inspections using modern analytical approaches.

2.1.2 *Simulated datasets*

Simulations were conducted using custom R scripts (R Core Team, 2020) following the procedures described in Puttick et al (2017), and specifically modified to allow for comparisons with our empirical datasets. We generated one set of matrices simulated with three different datasets, one simulated the evolution of multistate discrete characters on a fully symmetrical tree, and the other two the evolution of multistate characters on fully asymmetric trees. To ensure that our simulated data were realistic, we generated each set containing the same number of characters and taxa as the empirical datasets, since the different conformation of each dataset can provide a better assessment of the impacts of different taxa and characters sampling on the phylogenetic estimation. The empirical tree of Stromboidea is more similar to an asymmetrical tree, while Cerithioidea and Architaenoglossa better resemble symmetrical trees. The script used for simulation is available in Appendix 2.

2.2 PHYLOGENETIC ANALYSES

2.2.1 *Homoplasy Partitioning*

We analyzed both simulated and empirical matrices following the approach of Rosa et al. (2019) using the software TNT version 1.5 (Goloboff et al, 2008) to infer homoplasy classes based on the character's unbiased measure of homoplasy using

the method of Goloboff, 1997, with the default concavity parameter ($k = 3$) (Goloboff et al, 2008). In this approach, characters are partitioned according to their homoplasy classes, i.e., there are as many partitions as homoplasy classes. The classified values returned by the implied weight analysis were normalized between 0 and 1, using a Perl 5 programming script (Rosa et al, 2019) where the lowest value corresponds to no homoplasy, and the other classes are given increasing homoplasy values (i.e., weights).

2.2.2 Maximum Likelihood

All ML analyses were performed in IQTree 1.6.12 (Nguyen et al, 2015). Initially, we performed thorough ML searches using the Mk model on homoplasy partitioned and non-partitioned schemes. To incorporate different evolutionary models in addition to the Mk model, we used the IQTree tool ModelFinder (Kalyaanamoorthy et al, 2017), which automatically searches for the best-fit evolution models and character partitions. In short, this model-selection method can be used with both molecular and morphological alignments, and searches for models that will improve the accuracy of phylogenetic estimates (Kalyaanamoorthy et al, 2017) by proposing partitions that can share the same set of branch lengths, while allowing for the partitions to have their own evolutionary models. Previous comparisons on the performance of ModelFinder indicated that the Bayesian Information Criterion (BIC) estimates more accurate partition scheme models in relation to the other two available information criteria (Akaike Information Criterion and Corrected Akaike Information Criterion) (Kalyaanamoorthy et al, 2017). Therefore, we used the partition schemes as proposed by the BIC in all ML analyses. Finally, partitioned analyses were performed using both linked and unlinked branch lengths.

2.2.3 Bayesian Inference

We performed BI using MrBayes 3.2.7a (Huelsenbeck & Ronquist, 2001; Ronquist et al, 2012) at the Cipres Science Gateway (Miller et al, 2011). We ran standard analyses using the Mk model with 5 million generations sampled at every 2.000th step using both MCMC (*Markov Chain Monte Carlo*) and SS (*Stepping stone*) sampling algorithms. Different from MCMC, SS directly estimates marginal probabilities, instead of estimating a marginal probability ratio (Xie et al, 2011), and might provide more accurate parameter estimations. After preliminary analyses and

troubleshooting, the temperature parameter for heating the chains was set to *0.150* to improve chain mixing. Four independent runs were used to assess convergence, even considering that simulated datasets tend to reach stationarity and convergence between runs in fewer generations compared to empirical datasets (Vernygora et al, 2020). For both analyses types, MCMC and SS, we used the Marginal Likelihood (MgL) Harmonic Mean to compare models.

To implement the analyses with partitioned and non-partitioned schemes we used the inferred ModelFinder character sets, adapted to only consider evolutionary models that can be implemented in MrBayes (Huelsenbeck & Ronquist, 2001; Ronquist et al, 2012) using the same settings as described above. As in ML, partitioned analyses were also performed using linked and unlinked branch lengths.

In summary, for each of our three empirical matrices and our three simulated matrices we ran ML and BI analyses using the following strategies: a) unpartitioned analyses; b) partitioned according to homoplasy, c) partitioned according to ModelFinder with linked and unlinked branch lengths. Furthermore, because two of our original matrices had invariant character sites (Architaenoglossa = 28 constant sites; Stromboidea = 24 constant sites) we also ran analyses to compare results regarding the presence and absence of these sites. Initially, we excluded those sites from the alignment and ran analyses as described above and, in another step, we kept the invariant sites in all analyses as described above. Analytical steps are summarized in Figure 1. We only used an ascertainment bias correction (+ASC) model (Lewis, 2001) if ModelFinder pointed to correcting the likelihood condition regarding the number of variable sites.

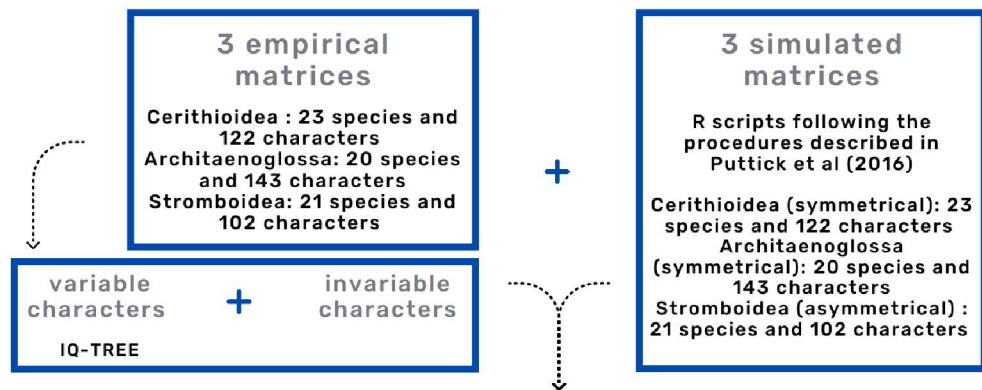
2.3 TREE DISTANCES

Calculating the distance among trees can lend support to those with higher reliability in phylogenetic analyses. However, evaluating tree distances between different methodological approaches is not a trivial task. In order to compare the performance of our analytical strategies and to understand whether two phylogenetic trees inferred under different algorithms would produce similar trees when the same evolutionary models and partitioning schemes were used, we compared: (1) the distances between the known ‘true’ tree used to simulate the datasets and the inferred tree based on these datasets; (2) trees inferred from simulated datasets and

analyzed under Maximum Likelihood and Bayesian Inference; (3) trees inferred from our empirical datasets under Maximum Likelihood and Bayesian Inference.

We used the Normalized Generalized Jaccard Robinson-Foulds distance (GdRF) (Smith, 2020) and Normalized Matching Splits (MS) (Bogdanowicz & Giaro, 2011) distance metrics. GdRF distance calculates pairing splits in one tree with similar splits in the other where each pair is assigned a similarity score, the sum of which summarizes the similarity between two trees and, hence, smaller values depict more similar trees. On the other hand, the MS metric returns distances between each pair of trees based on each bipartition split and, thus, higher values depict more similar trees. Accordingly, these distances should be interpreted as different measures of similarity, and can not be directly compared between them. The MS metric was normalized using a constant normalization parameter of value 1. Both metrics were calculated using the R package TreeDist (Smith, 2020).

Morphological datasets (multistate)



Methods

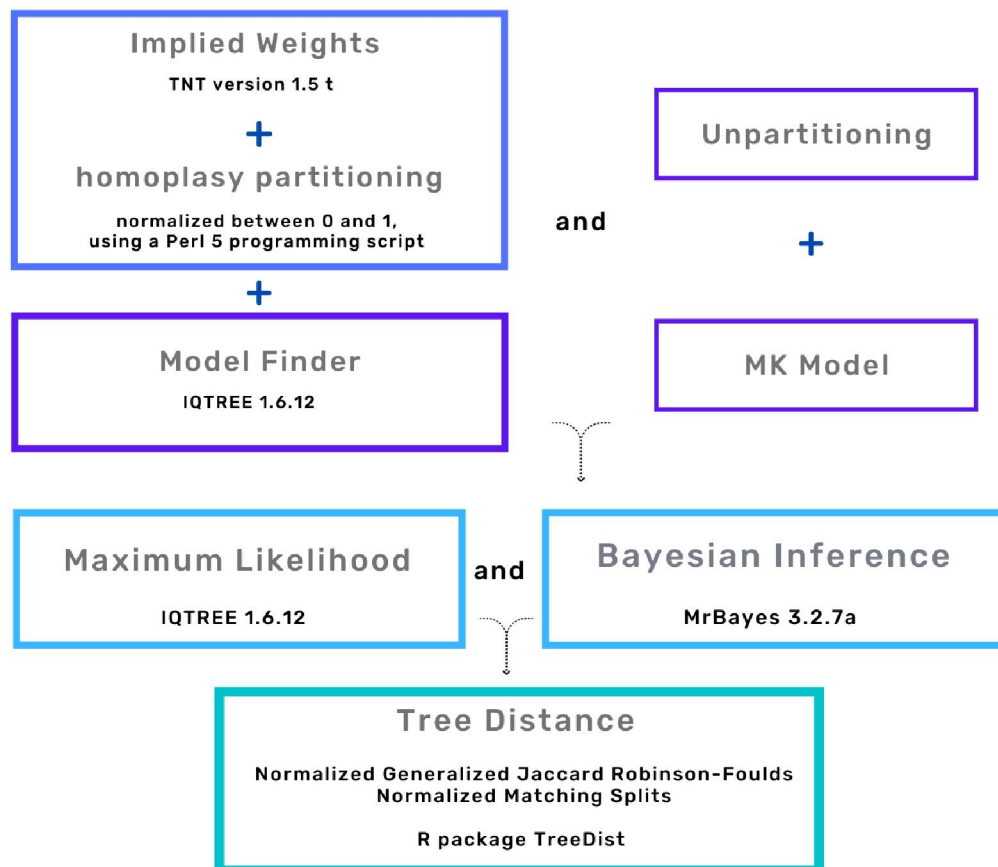


FIGURE 1 - Flow chart of study design outlining details of each step regarding the used datasets and phylogenetic analyses approaches.

3 RESULTS

3.1 EMPIRICAL AND SIMULATED DATASETS

We systematically found a higher relative maximum likelihood value in Maximum Likelihood analyses that used partitioned schemes estimated by ModelFinder with unlinked branch lengths (Tables 1, 2, and 3). The partitioned data with unlinked branch lengths also produced values significantly higher using an Mk model, and the accuracy of the model under unpartitioned schemes was lower (Tables 1, 2, and 3). There were no major differences comparing analyses that used invariant and only variant sites using the ascertainment bias correction (+ASC) model estimated by ML analyses (Tables 1 and 2).

TABLE 1 - Maximum Likelihood values of **ML** analyses on the **empirical** datasets that contained **invariable sites**. The best model (higher ML value) for each model selection strategy is in boldface. The models and partition schemes selected by ModelFinder can be found in Table S3-1.

	Criteria	Branch Lengths	Partitions	Mk Model	ModelFinder
Stromboidea	Partitioned	linked	6	-1080.341	-1061.868
	Partitioned	unlinked	6	-1006.754	-987.525
	Unpartitioned		1	-1136.911	-1131.285
Architaenoglossa	Partitioned	linked	5	-1016.568	-990.175
	Partitioned	unlinked	5	-890.402	-873.127
	Unpartitioned		1	-1068650	-1068.65

TABLE 2 - Maximum Likelihood values of **ML** analyses on the **empirical** datasets that contained **variable sites only**. The best model (higher ML value) for each model selection strategy is in boldface. The models and partition schemes selected by ModelFinder can be found in Table S3-2.

	Criteria	Branch Lengths	Partitions	Mk Model	ModelFinder
Stromboidea	Partitioned	linked	6	-1018.134	-1001.636
	Partitioned	unlinked	6	-943.584	-928.063
	Unpartitioned		1	-1031.54	-1020.317
Cerithioidea	Partitioned	linked	9	-1732.96	-1709.61
	Partitioned	unlinked	9	-1538.302	-1518.06
	Unpartitioned		1	-1793.04	-1777.255
Architaenoglossa	Partitioned	linked	5	-954.668	-930.105
	Partitioned	unlinked	5	-829.218	-812.522
	Unpartitioned		1	-973.524	-959.066

TABLE 3 - Maximum Likelihood values of **ML** analyses on the **simulated** datasets. The best model (higher ML value) for each model selection strategy is in boldface. The models and partition schemes selected by ModelFinder can be found in Table S3-3.

	Criteria	Branch Lengths	Partitions	Mk Model	ModelFinder
<i>"Stromboidea"</i> (asymmetric)	Partitioned	linked	8	-616.986	-615.301
	Partitioned	unlinked	8	-503.663	-499.732
	Unpartitioned		1	-731.656	-679.575
<i>"Cerithioidea"</i> (symmetric)	Partitioned	linked	12	-2835.976	-2831.928
	Partitioned	unlinked	12	-2607.859	-2598.679
	Unpartitioned		1	-2972.279	-2940.685
<i>"Architaenoglossa"</i> (symmetric)	Partitioned	linked	6	-1669.999	-1658.617
	Partitioned	unlinked	6	-1554.17	-1544.365
	Unpartitioned		1	-1741.45	-1738.202

The performance of Bayesian Inference analyses, as inferred by marginal likelihood (MgL) values, showed a few differences according to the used dataset. The empirical datasets with invariable sites only and “Architaenoglossa” simulated dataset had a higher-relative value using partitioned schemes with the MK Model and unlinked branch lengths (Tables 4 and 5). Simulated “Stromboidea” had higher values with the Mk Model using partitioned schemes and linked branch lengths (Table 6). Datasets that contained variable sites resulted in higher values using partitioned schemes with ModelFinder and using unlinked branch lengths, except for Cerithioidea that had a better performance using a partitioned scheme and unlinked branch lengths with the Mk Model (Table 5).

Analyses using BI usually scored higher under partitioned approaches, in combination with the Mk Model, and using unlinked branch lengths compared to the partition strategies inferred by ModelFinder that had only a higher MgL with variable sites in the “Cerithioidea” simulated dataset (Tables 4, 5 and 6).

TABLE 4 - Marginal Likelihood of **Bayesian Inference** analyses on the **empirical** datasets that contained **invariable sites**. The best model (higher MgL value) for each model selection strategy is in boldface. The models and partition schemes selected by ModelFinder can be found at Table S3-1.

	Criteria	Branch Lengths	Partitions	Mk Model		ModelFinder	
				mcmc	ss	mcmc	ss
Stromboidea	Partitioned	linked	6	-677	-1989.7	-1110.91	-2013.65
	Partitioned	unlinked	6	-682.45	-1861.68	-1857.83	-1896.45
	Unpartitioned		1	-672.34	-2048.55	-670.52	-1952.45
Architaenoglossa	Partitioned	linked	5	-775.03	-2898.05	-833.78	-2856.95
	Partitioned	unlinked	5	-719.57	-2644.63	-787.66	-2700.97
	Unpartitioned		1	-771.18	-2836.55	-769.2	-2787.61

TABLE 5 - Marginal Likelihood of **Bayesian Inference** analyses on the **empirical** datasets that contained **variable sites only**. The best model (higher MgL value) for each model selection strategy is in boldface. The models and partition schemes selected by ModelFinder can be found at Table S3-2.

	Criteria	Branch Lengths	Partitions	Mk Model		ModelFinder	
	Models			mcmc	ss	mcmc	ss
Stromboidea	Partitioned	linked	6	-673.13	-2092.12	-680.28	-2025.7
	Partitioned	unlinked	6	-683.34	-1850.34	-692.87	-1799.44
	Unpartitioned		1	-672.53	-1990.12	-685.22	-1947.36
Cerithioidea	Partitioned	linked	9	-1322.05	-2892.54	-1326.01	-2833.88
	Partitioned	unlinked	9	-1213.02	-2541.34	-1236.9	-2547.85
	Unpartitioned		1	-1319.62	-2743.04	-1323.61	-2698.95
Architaenoglossa	Partitioned	linked	5	-769.56	-2854.93	-783.93	-2914.67
	Partitioned	unlinked	5	-769.37	-2670.51	-740.61	-2628.63
	Unpartitioned		1	-774.98	-2881.83	-786.3	-2591.12

TABLE 6 - Marginal Likelihood of **Bayesian Inference** analyses on the **simulated** datasets. The best model (higher MgL value) for each model selection strategy is in boldface. The models and partition schemes selected by ModelFinder can be found at Table S3-3.

	Criteria	Branch Lengths	Partitions	Mk Model		ModelFinder	
				mcmc	ss	mcmc	ss
<i>"Stromboidea"</i> (asymmetric)	Partitioned	linked	8	-512.83	-765.93	-690.81	-1064.74
	Partitioned	unlinked	8	-514.58	-786.21	-693.72	-1090.45
	Unpartitioned		1	-513.92	-795.78	-516.82	-769.84
<i>"Cerithioidea"</i> (symmetric)	Partitioned	linked	12	-2658.19	-4414.75	-2647.39	-4566.54
	Partitioned	unlinked	12	-2656.99	-4336.51	-2545.04	-4032.4
	Unpartitioned		1	-2656.29	-4537.03	-2640	-4127.59
<i>"Architaenoglossa"</i> (symmetric)	Partitioned	linked	6	-1522.33	-2955	-1522.33	-2955
	Partitioned	unlinked	6	-1452.71	-2642.4	-1497.64	-2691.46
	Unpartitioned		1	-1523.31	-2874.33	-1519.68	-3092.74

Comparing empirical and simulated datasets analyses, partition schemes generated by ModelFinder demonstrated expected but only slight differences on the inferred partition character sets (charsets). In the empirical Cerithioidea we had 9 partitioned charsets, whereas 12 were estimated for the simulated matrix; empirical Architaenoglossa had 5 charsets with 6 for the simulated dataset and; lastly, empirical Stromboidea had 6 charsets and the simulated dataset had 8, even though they had the same number of taxa and characters (Tables S3 1, 2 and 3). Even when datasets had differences between numbers of charsets, the ML or MgL (in BI) values demonstrated higher probabilities at the same analytical strategy, indicating that the best analytical strategy might reside in the nature of the data. The most present model in all partitioned datasets was MK+FQ+ASC (MK Model, Frequency State Characters and Ascertainment Bias Correction) and MK+FQ+I+G4 (MK Model, Frequency State Characters, Allowing for a proportion of invariable sites and Gamma Rate Model with 4 rate categories). It is important to emphasize that all partitioned and non-partitioned schemes inferred by ModelFinder had the Mk model, thus demonstrating its importance for morphological datasets (Tables S3 1, 2 and 3).

3.2 TREE DISTANCES

The analytical performance on simulated datasets in terms of the distances among the inferred trees and the 'true' trees with ML and BI as estimated by GdRF and nMS varied among the three datasets. For BI analyses, most results showed higher distances compared to ML, which showed more accurate smaller distances among trees. "Stromboidea" (asymmetric) was more biased under ML using the MK Model (nGRF= 0.411) but varied in relation to partitioned approaches. However, partitioned datasets with unlinked branch lengths showed an increase in accuracy under ML as showed by nMS (nMS=59) (Table 7). In "Cerithioidea" (symmetric) the distance for ML using unpartitioned scheme with the MK Model was lower (nGRF=0.085), whereas partitioned unlinked branch length showed a better performance according to MS (nMS=28). "Architaenoglossa" (symmetric) trees resulted in equal distances among most analyses under ML (nGRF=0.046; nMS=6), except at partitioned unlinked branch lengths (nGRF=1; nMS=0) (Table 7).

TABLE 7 - Normalized (Generalized) **Jaccard Robinson-Foulds** and Normalized (value=1) **Matching Splits** distances among the 'true' trees and the inferred tree using the **simulated datasets** under Maximum Likelihood and Bayesian Inference using the MK Model and ModelFinder. 'True' trees were named according to the respective empirical tree used to determine the number of taxa and characters, and tree format (symmetric or asymmetric). Smaller distances (depicting better analytical performance) of nGRF and higher similarity of nMS for each comparison are highlighted in boldface for each distance metric.

Distance Metrics	Jaccard Robinson Foulds				Matching Splits			
Inference Method	Maximum Likelihood		Bayesian Inference		Maximum Likelihood		Bayesian Inference	
	MK	Model Finder	MK	Model Finder	MK	Model Finder	MK	Model Finder
<i>"Stromboidea"</i> (asymmetric)								
Partitioned linked bl	0.412	0.411	0.758	0.665	58	53	12	20
Partitioned unlinked bl	0.427	0.437	0.758	0.614	59	49	12	21
Unpartitioned	0.468	0.58	0.758	0.682	52	79	12	12
<i>"Cerithioidea"</i> (symmetric)								
Partitioned linked bl	0.113	0.113	0.142	0.144	20	20	0	5
Partitioned unlinked bl	0.123	0.178	0.177	1	17	28	5	0
Unpartitioned	0.085	0.113	0.177	0.144	11	20	5	5
<i>"Architaenoglossa"</i> (symmetric)								
Partitioned linked bl	0.046	0.046	0.046	0.046	6	6	6	6
Partitioned unlinked bl	0.088	0.046	1	0.941	8	6	0	6
Unpartitioned	0.046	0.046	0.046	0.046	6	6	6	6

The distance of the inferred trees using the simulated datasets comparing ML and BI, was lower using the inferred ModelFinder schemes. 'Stromboidea' had the lower (nGRF=0.307) distance value in partitioned unlinked branch lengths, and Cerithioidea in partitioned linked branch lengths (nGRF=0.111) and unpartitioned (nGRF=0.111). Most nMS distances did not depict higher values, which indicates that

the resulting trees were somewhat not similar between both BI and ML inferences, except for *Architaenoglossa* (nMS=6).

TABLE 8 - Normalized (Generalized) **Jaccard Robinson-Foulds** and Normalized (value=1) **Matching Splits** distances of **simulated datasets** with Maximum Likelihood (ML) and Bayesian Inference (BI) comparing tree strategies using the MK Model and ModelFinder. Smaller distances (depicting better analytical performance) of nGRF and higher similarity of nMS for each comparison are highlighted in boldface for each distance metric.

Distance Metrics	Jaccard Robinson Foulds		Matching Splits	
Comparison	ML - BI		ML - BI	
Model	MK	Model Finder	MK	Model Finder
<i>"Stromboidea"</i> (asymmetric)				
Partitioned linked bl	0.556	0.396	1	2
Partitioned unlinked bl	0.556	0.307	1	0
Unpartitioned	0.545	0.525	0	3
<i>"Cerithioidea"</i> (symmetric)				
Partitioned linked bl	0.142	0.111	0	0
Partitioned unlinked bl	0.142	1	0	0
Unpartitioned	0.142	0.111	0	0
<i>"Architaenoglossa"</i> (symmetric)				
Partitioned linked bl	0	0	0	0
Partitioned unlinked bl	1	0.941	0	6
Unpartitioned	0	0	0	0

Considering the empirical datasets with invariable characters, distances varied according to the inference method and employed distance metric. *Stromboidea* had lower distance values for the MK Model with unpartitioned data (nGRF=0.152) and partitioned linked branch lengths (nMS=6) (Table 9). *Architaenoglossa* also showed the smallest distance values under partitioned schemes with linked branch lengths (nGRF=0.111). For nMS values, the highest

result was shared among models of unpartitioned MK and ModelFinder (nMS=12) (Table 9).

TABLE 9 - Normalized (Generalized) **Jaccard Robinson-Foulds** and Normalized (value=1) **Matching Splits** distances of **empirical datasets (invariable characters)** with Maximum Likelihood (ML) and Bayesian Inference (BI) comparing tree strategies using the MK Model and ModelFinder. Smaller distances (depicting better analytical performance) of nGRF and higher similarity of nMS for each comparison are highlighted in boldface for each distance metric.

Distance Metrics	Jaccard Robinson Foulds		Matching Splits	
Comparison	ML - BI		ML - BI	
Model	MK	Model Finder	MK	Model Finder
Stromboidea				
Partitioned linked bl	0.163	0.273	6	4
Partitioned unlinked bl	0.214	0.214	0	0
Unpartitioned	0.152	0.172	4	0
Architaenoglossa				
Partitioned linked bl	0.166	0.111	12	4
Partitioned unlinked bl	0.351	0.351	8	8
Unpartitioned	0.171	0.171	12	12

In empirical datasets with variable characters only, Stromboidea had the lowest distance in ModelFinder with partitioned linked branch lengths (nGRF=0.146) and MK Model with partitioned linked branch lengths (nMS=6). Cerithioidea showed lower distances under the MK Model with unpartitioned data (nGRF=0.052), however, it shared the same highest values for MS distance in partitioned linked branch lengths (nMS=11). For Architaenoglossa, the ModelFinder scheme with partitioned linked branch lengths had the lowest distance (nGRF=0.111) and the same higher values for MS in unpartitioned schemes (nMS=12) (Table 10).

TABLE 10 - Normalized (Generalized) **Jaccard Robinson-Foulds** and Normalized (value=1) **Matching Splits** distances of **empirical datasets (variable characters only)** with Maximum Likelihood (ML) and Bayesian Inference (BI) comparing tree strategies using the MK Model and ModelFinder. Smaller distances (depicting better analytical performance) of nGRF and higher similarity of nMS for each comparison are highlighted in boldface for each distance metric.

Distance Metrics	Jaccard Robinson Foulds		Matching Splits	
Comparisons	ML - BI		ML - BI	
Model	MK	Model Finder	MK	Model Finder
Stromboidea				
Partitioned linked bl	0.163	0.146	6	2
Partitioned unlinked bl	0.214	0.214	0	0
Unpartitioned	0.152	0.152	4	4
Cerithioidea				
Partitioned linked bl	0.123	0.123	11	11
Partitioned unlinked bl	0.108	0.108	2	2
Unpartitioned	0.052	0.121	0	3
Architaenoglossa				
Partitioned linked bl	0.066	0.111	0	4
Partitioned unlinked bl	0.365	0.351	11	8
Unpartitioned	0.171	0.171	12	12

Overall, some of the compared trees indicated that both analyses of ML and IB shared similar tree topologies. Nonetheless, the performance of each analytical approach varied between models and partitioned schemes. In particular, greater resolution of trees may be found by using ModelFinder in both analysis of ML and BI, since the distance metrics indicated that a more similar topology between them.

4 DISCUSSION

Evaluating the performance of morphological datasets under different probabilistic phylogenetic analyses provides a reasonable strategy to improve and select the best analytical framework. In a previous study using discrete morphological characters, Rosa et al. (2019) showed that phylogenetic Bayesian model estimation, as based on SS, performed better when using homoplasy partitioning. Here, we built on this framework and took the opportunity to both compare the performance of BI and Maximum Likelihood, and test for different analytical strategies considering character evolution models and branch lengths linkage among partitions. In short, we demonstrate that homoplasy partitioning and the inclusion of more parameter-rich models with complex evolution models and unlinked branch lengths provides more accurate phylogenetic inferences, while also being particularly robust when applied to small-sized matrices of discrete morphological characters. We tested the implementation of ModelFinder (Kalyaanamoorthy et al, 2017) on empirical and simulated matrices, providing evidence of its performance in partitions that apply different evolutionary models and while estimating parameters, particularly branch lengths, separately for each partition (Tables 1 – 3). Our results indicate that this approach is a stronger alternative when implementing implied weights on discrete morphological characters. The ModelFinder tool, although originally aimed at molecular analyses, showed to be a good tool for morphological matrices, adding power to homoplasy-partitioning strategies. When applied to more complex evolutionary models and different sets of branch lengths, it informed parameter-rich analyse that provided better performance both with empirical and simulated datasets (Tables 1–4).

Besides the fact that ascertainment bias can possibly cause an inflation of estimated evolutionary rates in small datasets (Wright & Hillis, 2014), the incorporation of a few invariant sites in the analyses did not improve the results. Phylogenetic analyses of both matrices that had invariable sites showed a weaker performance compared to the matrices that only had variable sites (Tables 4 - 5). The IQTree tool pointed to such matrix enhancement with ascertainment bias, which possibly improved the morphological matrices and also the partition models that resulted with the +ASC model for all datasets (Table S3-1-3) (Nguyen et al, 2014). Invariable characters can be automatically excluded from morphological matrices on

IQTree and MrBayes (MKv or Mk-pars). However, we also obtained the model that allows for invariable characters in one of the partitioned simulated datasets (Table S3-3). This may demonstrate that even with constructions carried out in simulations, some characters may be subject to non-informative variations (Abadi et al, 2019). This might be a factor that highly influences analyses of molecular datasets under BI (Bouckaert & Drummond, 2017), but our results depicted no important differences regarding the best results using morphological datasets.

Unlinking branch lengths in phylogenetic analyses substantially increases the number of estimated parameters, which might also influence model performance (Wright, 2019). Nonetheless the Marginal Likelihood calculations in BI and the Maximum Likelihood estimations in ML weight down on parameter-rich models in order to provide comparable results among runs with different parameters using the same dataset (Rosa et al, 2019), similarly to what is done when calculating information criteria statistics (e.g., AIC or BIC). Thus, our results indicate a higher model performance using unlinked branch lengths in BI and ML with homoplasy-partitioned schemes, shedding important light on model choice for morphological matrices, especially considering that this method was not shown to provide the best resolution in other model selection studies (Rosa et al, 2019; Goloboff & Arias, 2019). The choice between linking or unlinking branch lengths among partitions when using morphological datasets might have small consequences in topology estimations, and unlinking can also lead to a possible over-fitting of the phylogenetic terraces (Straub et al, 2014; Sanderson et al, 2011). However, our results suggest that allowing for unlinked partitions to have different evolutionary models might possibly better inform on the processes influencing character changes at each partition, and provide a parameter-rich solution for the analysis, not only estimating independent sets of branch lengths for each partition, but also influencing on topology estimations (Tables 1-7) (Goloboff et al, 2018b). These results reinforce the fact that such partitions have different overall rates of evolution, possibly influenced by different relative rates of evolution among lineages (Straub et al, 2014; Lanfear et al, 2017).

Not surprisingly, the relatively small total number of characters had a clear effect on BI results when analyzing the empirical Stromboidea matrix which, using both invariant or variable only characters, had a higher MCMC Harmonic Mean value under non-partitioned schemes (Table 4-5). The dataset size can affect the

performance of the analyses under BI and directly influence the accuracy and precision of tree searches (Vernygora et al, 2020), even considering that chances of homoplasy are lower in datasets with low rates of variation (Goloboff & Arias, 2019).

Some recent advances in the analyses of morphological data showed that BI usually outperforms other methods under different analytical scenarios (O'Reilly et al, 2016; 2018). Bayesian methods have different customisable tools that can be applied to complex evolution models and distinct variation rates (Bromham et al, 2018). Nevertheless, new phylogenetic softwares implementing ML are including important developments to provide more flexible analytical approaches, particularly with the inclusion of new evolution models and approaches for morphological datasets (Kalyaanamoorthy et al, 2017). Importantly, our results showed that ML inference was more robust than BI in terms of retrieving better resolved topologies (Table 7). Both inferences had some similarities in terms of tree distance when using a non-partitioned MK approach. The nGRF values coincide on empirical matrices of Stromboidea with invariable characters and Cerithioidea with variable characters. *Architaenoglossa* resulted in matching values in both datasets for partitioned linked branch lengths for the two inference methods (Table 7-8).

Recovered trees proposed by Simone, 2001; 2004; 2005 have shown slight differences against the position of species that ML and BI resulted in this work. The structure of a good morphological matrix can be an indicator that different analyses can share the same position of groups in results, even when we have a range of trees in the statistical set (Puttick et al, 2017).

In order to test the performance of BI and ML we simulated datasets, because the results from such simulations could provide a baseline for comparison among inference strategies of empirical datasets. Therefore, by testing the statistical power of those datasets with a range of evolutionary models (Rosa et al, 2019) and phylogenetic methodologies (Puttick et al, 2017), we could explore those differences by comparing distances through the pool of derived trees and provide direction to which analytical approach would better perform on empirical datasets (Keating, 2020). Thus, based on our results we recommend not only using partitioning strategies with morphological datasets, but also implement different evolution models at each partition, as selected by information criteria metrics, i.e. using ModelFinder. By directly using homoplasy-based partitioning and BI, researchers may be

neglecting the performance that ML analyses can have, particularly with IQTree. The new tools that can be tested on the IQTree program, such as adding models in each partition, and considering unlinked branch length among partitions, provide a strong new framework for the analyses of morphological characters under Maximum Likelihood. For the past few years, many programs that were previously only used for molecular analysis are being adapted for morphological analysis, which allows for testing and further developing these available tools.

5 CONCLUSION

This dissertation tested evolutionary models in both phylogenetic programs of Maximum Likelihood and Bayesian Inference. Although BI shows to be a solid method under different studies (O'Reilly et al, 2016; Puttick et al, 2017; Rosa et al, 2019) Maximum Likelihood appears to provide better tree resolution compared to BI (Table 7). Overall, our results suggest that tree topology and resolution improves using the implicit weighting homoplasy partitioning strategy as obtained using TNT (Goloboff, 2018b), complemented with the use of ModelFinder on IQTree (Kalyaanamoorthy, 2017). Using small morphological matrices of different sizes and with the implementation of simulated datasets we provide evidence that ML approaches might result in more accurate tree topologies compared to BI approaches. We recommend that the ModelFinder and homoplasy partitioning should not be ignored in phylogenetic analyses of morphological datasets. Another implementation that can improved the results was unlinking branch lengths among partitions, resulting in higher levels of resolution for small datasets. Both ML and BI had their strengths and weaknesses, since the different phylogenetic programs used had their own probabilistic approaches. However, for the past few years we had a higher development on ML programs like IQTree that adapted different methodologies for all types of datasets. Determining the analytical reliability of different partition strategies is essential for studies with discrete characters, and ease of processing data in modern programs suggests that these strategies can be empirically adopted.

REFERENCES

- ABADI, Shiran; AZOURI, Dana; PUPKO, Tal; MAYROSE, Itay. Model selection may not be a mandatory step for phylogeny reconstruction. **Nature communications**, v. 10, n. 1, p. 1-11, 2019.
- ÁLVAREZ-CARRETERO, Sandra; GOSWAMI, Anjali; YANG, Ziheng; *et al.* Bayesian Estimation of Species Divergence Times Using Correlated Quantitative Characters. **Systematic Biology**, v. 68, n. 6, p. 967–986, 2019.
- BERGER, Simon A.; STAMATAKIS, Alexandros. Accuracy of morphology-based phylogenetic fossil placement under Maximum Likelihood. *In: ACS/IEEE International Conference on Computer Systems and Applications - AICCSA 2010*. Hammamet, Tunisia: IEEE, 2010, p. 1–9.
- BOGDANOWICZ, Damian; GIARO, Krzysztof. Matching split distance for unrooted binary phylogenetic trees. **IEEE/ACM Transactions on Computational Biology and Bioinformatics**, v. 9, n. 1, p. 150-160, 2011.
- BOUCKAERT, Remco R.; DRUMMOND, Alexei J. bModelTest: Bayesian phylogenetic site model averaging and model comparison. **BMC Evolutionary Biology**, v. 17, n. 1, p. 42, 2017.
- BROMHAM, L., DUCHÊNE, S., HUA, X; RICHIE, A. M; DUCHÊNE, D. A., & HO, S. Y. Bayesian molecular dating: opening up the black box. **Biological Reviews**, v. 93, n. 2, p. 1165-1191, 2018.
- CHAZOT, Nicolas; WAHLBERG, Niklas; FREITAS, André Victor Lucci; *et al.* Priors and Posteriors in Bayesian Timing of Divergence Analyses: The Age of Butterflies Revisited. **Systematic Biology**, v. 68, n. 5, p. 797–813, 2019.

- DORNELLAS, Ana Paula; COUTO, Diogo Ribeiro; SIMONE, Luiz Ricardo L. Morphological phylogeny of the Tegulinae (Mollusca: Vetigastropoda) reinforces a Turbinidae position. **Cladistics**, v. 36, n. 2, p. 129–163, 2020.
- JIA, Fangzhi; LO, Nathan; HO, Simon YW. The impact of modelling rate heterogeneity among sites on phylogenetic estimates of intraspecific evolutionary rates and timescales. **PLoS One**, v. 9, n. 5, p. e95722, 2014.
- FELSENSTEIN, Joseph. Cases in which parsimony or compatibility methods will be positively misleading. **Systematic zoology**, v. 27, n. 4, p. 401-410, 1978.
- FELSENSTEIN, Joseph. Evolutionary trees from DNA sequences: A maximum likelihood approach. **Journal of Molecular Evolution**, v. 17, n. 6, p. 368–376, 1981.
- GOLOBOFF, Pablo A. Self-weighted optimization: tree searches and character state reconstructions under implied transformation costs. **Cladistics**, v. 13, n. 3, p. 225-245, 1997.
- GOLOBOFF, Pablo A. Calculating SPR distances between trees. **Cladistics**, v. 24, n. 4, p. 591–597, 2008.
- GOLOBOFF, Pablo A. Estimating character weights during tree search. **Cladistics**, v. 9, n. 1, p. 83-91, 1993.
- GOLOBOFF, Pablo A.; ARIAS, J. Salvador. Likelihood approximations of implied weights parsimony can be selected over the Mk model by the Akaike information criterion. **Cladistics**, v. 35, n. 6, p. 695–716, 2019.
- GOLOBOFF, Pablo A.; FARRIS, James S.; NIXON, Kevin C. TNT, a free program for phylogenetic analysis. **Cladistics**, v. 24, n. 5, p. 774–786, 2008.

- GOLOBOFF, Pablo A; PITTMAN, Michael; POL, Diego; *et al.* Morphological Data Sets Fit a Common Mechanism Much More Poorly than DNA Sequences and Call Into Question the MkV Model. **Systematic Biology**, 2018a.
- GOLOBOFF, Pablo A.; TORRES, Ambrosio; ARIAS, J. Salvador. Weighted parsimony outperforms other methods of phylogenetic inference under models appropriate for morphology. **Cladistics**, v. 34, n. 4, p. 407–437, 2018b.
- HARRISON, Luke B.; LARSSON, Hans C. E. Among-Character Rate Variation Distributions in Phylogenetic Analysis of Discrete Morphological Characters. **Systematic Biology**, v. 64, n. 2, p. 307–324, 2015.
- HUELSENBECK, J. P.; RONQUIST, F. MRBAYES: Bayesian inference of phylogenetic trees. **Bioinformatics**, v. 17, n. 8, p. 754–755, 2001.
- HUELSENBECK, John P.; CRANDALL, Keith A. PHYLOGENY ESTIMATION AND HYPOTHESIS TESTING USING MAXIMUM LIKELIHOOD. **Annual Review of Ecology and Systematics**, v. 28, n. 1, p. 437–466, 1997.
- HUELSENBECK, John P.; RANNALA, Bruce. Frequentist Properties of Bayesian Posterior Probabilities of Phylogenetic Trees Under Simple and Complex Substitution Models. **Systematic Biology**, v. 53, n. 6, p. 904–913, 2004.
- KALYAANAMOORTHY, Subha; MINH, Bui Quang; WONG, Thomas K F; *et al.* ModelFinder: fast model selection for accurate phylogenetic estimates. **Nature Methods**, v. 14, n. 6, p. 587–589, 2017.
- KAPLI, P., YANG, Z., TELFORD, M.J.. Phylogenetic tree building in the genomic age. **Nature Reviews Genetics**, v. 21, n. 7, p. 428–444, 2020.

- KEATING, Joseph N; SANSOM, Robert S; SUTTON, Mark D; *et al.* Morphological Phylogenetics Evaluated Using Novel Evolutionary Simulations. **Systematic Biology**, v. 69, n. 5, p. 897–912, 2020.
- KOZLOV, Alexey M. *et al.* RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. **Bioinformatics**, v. 35, n. 21, p. 4453-4455, 2019.
- LANFEAR, Robert *et al.* PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. **Molecular biology and evolution**, v. 34, n. 3, p. 772-773, 2017.
- LARTILLOT, Nicolas. The Bayesian Approach to Molecular Phylogeny. p. 18, . LEE, Michael S. Y.; WORTHY, Trevor H. Likelihood reinstates *Archaeopteryx* as a primitive bird. **Biology Letters**, v. 8, n. 2, p. 299–303, 2012.
- LEE, Michael SY; WORTHY, Trevor H. Likelihood reinstates *Archaeopteryx* as a primitive bird. **Biology letters**, v. 8, n. 2, p. 299-303, 2012.
- LEE, Michael SY; PALCI, Alessandro. Morphological phylogenetics in the genomic age. **Current Biology**, v. 25, n. 19, p. R922-R929, 2015.
- LEWIS, Paul O. A Likelihood Approach to Estimating Phylogeny from Discrete Morphological Character Data. **Systematic Biology**, v. 50, n. 6, p. 913–925, 2001.
- MAY, Michael R; MOORE, Brian R. A Bayesian Approach for Inferring the Impact of a Discrete Character on Rates of Continuous-Character Evolution in the Presence of Background-Rate Variation. **Systematic Biology**, p. 15, 2019.
- MCINERNEY, Phoebe L.; LEE, Michael S. Y.; CLEMENT, Alice M.; *et al.* The phylogenetic significance of the morphology of the syrinx, hyoid and larynx,

of the southern cassowary, *Casuarius casuarius* (Aves, Palaeognathae). **BMC Evolutionary Biology**, v. 19, n. 1, p. 233, 2019.

NGUYEN, Lam-Tung; SCHMIDT, Heiko A.; VON HAESELER, Arndt; *et al.* IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. **Molecular Biology and Evolution**, v. 32, n. 1, p. 268–274, 2015.

O'REILLY, Joseph E.; PUTTICK, Mark N.; PARRY, Luke; *et al.* Bayesian methods outperform parsimony but at the expense of precision in the estimation of phylogeny from discrete morphological data. **Biology Letters**, v. 12, n. 4, p. 20160081, 2016.

O'REILLY, Joseph E.; PUTTICK, Mark N.; PISANI, Davide; *et al.* Empirical realism of simulated data is more important than the model used to generate it: a reply to Goloboff *et al.* **Palaeontology**, v. 61, n. 4, p. 631–635, 2018.

O'REILLY, Joseph E.; PUTTICK, Mark N.; PISANI, Davide; *et al.* Probabilistic methods surpass parsimony when assessing clade support in phylogenetic analyses of discrete morphological data. **Palaeontology**, v. 61, n. 1, p. 105–118, 2018.

PARINS-FUKUCHI, Caroline. Use of continuous traits can improve morphological phylogenetics. **Systematic Biology**, v. 67, n. 2, p. 328–339, 2018.

PFEIFFER, Wayne; SCHWARTZ, Terri. The CIPRES science gateway: a community resource for phylogenetic analyses. In: Proceedings of the 2011 TeraGrid Conference: extreme digital discovery. 2011. p. 1–8

PUTTICK, Mark N.; O'REILLY, Joseph E.; PISANI, Davide; *et al.* Probabilistic methods outperform parsimony in the phylogenetic analysis of data simulated without a probabilistic model. **Palaeontology**, v. 62, n. 1, p. 1–17, 2019.

- PUTTICK, Mark N.; O'REILLY, Joseph E.; TANNER, Alastair R.; *et al.* Uncertain-tree: discriminating among competing approaches to the phylogenetic analysis of phenotype data. **Proceedings of the Royal Society B: Biological Sciences**, v. 284, n. 1846, p. 20162290, 2017.
- RANNALA, Bruce; YANG, Ziheng. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. **Journal of molecular evolution**, v. 43, n. 3, p. 304-311, 1996.
- RIO, Jonathan P; MANNION, Philip D; TSCHOPP, Emanuel; *et al.* Reappraisal of the morphology and phylogenetic relationships of the alligatoroid crocodylian *Diplocynodon hantoniensis* from the late Eocene of the United Kingdom. **Zoological Journal of the Linnean Society**, p. zlz034, 2019.
- ROBINSON, D.F.; FOULDS, L.R. Comparison of phylogenetic trees. **Mathematical Biosciences**, v. 53, n. 1–2, p. 131–147, 1981.
- RONQUIST, Fredrik; TESLENKO, Maxim; VAN DER MARK, Paul; *et al.* MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. **Systematic Biology**, v. 61, n. 3, p. 539–542, 2012.
- ROSA, Brunno B; MELO, Gabriel A R; BARBEITOS, Marcos S. Homoplasy-Based Partitioning Outperforms Alternatives in Bayesian Analysis of Discrete Morphological Data. **Systematic Biology**, v. 68, n. 4, p. 657–671, 2019.
- SANDERSON, Michael J.; MCMAHON, Michelle M.; STEEL, Mike. Terraces in phylogenetic tree space. **Science**, v. 333, n. 6041, p. 448-450, 2011.
- SCHRAGO, Carlos G.; AGUIAR, Barbara O.; MELLO, Beatriz. Comparative evaluation of maximum parsimony and Bayesian phylogenetic reconstruction using empirical morphological data. **Journal of Evolutionary Biology**, v. 31, n. 10, p. 1477–1484, 2018.

- SIMONE, Luiz Ricardo L. Phylogenetic analyses of Cerithioidea (Mollusca, Caenogastropoda) based on comparative morphology. **Arquivos de Zoologia**, v. 36, n. 2, p. 147-263, 2001.
- SIMONE, Luiz Ricardo L. Comparative morphological study of representatives of the three families of Stromboidea and the Xenophoroidea (Mollusca, Caenogastropoda), with an assessment of their phylogeny. **Arquivos de Zoologia**, v. 37, n. 2, p. 141, 2005.
- SIMONE, Luiz Ricardo Lopes. Comparative morphology and phylogeny of representatives of the superfamilies of architaenioglossans and the annulariidae (Mollusca, Caenogastropoda). **Arquivos do Museu Nacional**, v. 62, n. 4, p. 387-504, 2004.
- SMITH, Martin R. Bayesian and parsimony approaches reconstruct informative trees from simulated morphological datasets. **Biology Letters**, v. 15, n. 2, p. 20180632, 2019.
- SMITH, Martin R. Information theoretic generalized Robinson–Foulds metrics for comparing phylogenetic trees. **Bioinformatics**, v. 36, n. 20, p. 5007–5013, 2020.
- SMITH, Stacey D.; PENNELL, Matthew W.; DUNN, Casey W.; et al. Phylogenetics is the New Genetics (for Most of Biodiversity). **Trends in Ecology & Evolution**, v. 35, n. 5, p. 415–425, 2020.
- SPADE, David A. An extended model for phylogenetic maximum likelihood based on discrete morphological characters. **Statistical applications in genetics and molecular biology**, v. 19, n. 1, 2020.
- STAMATAKIS, Alexandros. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. **Bioinformatics**, v. 30, n. 9, p. 1312–1313, 2014.

- STRAUB, Shannon C.K; MOORE, Michael J; SOLTIS, Pamela S; SOLTIS, Douglas E; LISTON, Aaron; LIVSHULTZ, Tatyana. Phylogenetic signal detection from an ancient rapid radiation: Effects of noise reduction, long-branch attraction, and model selection in crown clade Apocynaceae. **Molecular Phylogenetics and Evolution**, v. 80, p. 169-185, 2014.
- VERNYGORA, Oksana V; SIMÕES, Tiago R; CAMPBELL, Erin O. Evaluating the Performance of Probabilistic Algorithms for Phylogenetic Analysis of Big Morphological Datasets: A Simulation Study. **Systematic Biology**, v. 69, n. 6, p. 1088–1105, 2020.
- WARNOCK, Rachel CM; WRIGHT, April M. Understanding the tripartite approach to Bayesian divergence time estimation. **Cambridge University Press**, 2020.
- WORTLEY, Alexandra H.; SCOTLAND, Robert W. The effect of combining molecular and morphological data in published phylogenetic analyses. **Systematic Biology**, v. 55, n. 4, p. 677-685, 2006.
- WRIGHT, April M. A Systematist's Guide to Estimating Bayesian Phylogenies From Morphological Data. **Insect Systematics and Diversity**, v. 3, n. 3, p. 2, 2019.
- WRIGHT, April M.; HILLIS, David M. Bayesian Analysis Using a Simple Likelihood Model Outperforms Parsimony for Estimation of Phylogeny from Discrete Morphological Data. **PLoS ONE**, v. 9, n. 10, p. e109210, 2014.
- XIE, Wangang; LEWIS, Paul O.; FAN, Yu; et al. Improving Marginal Likelihood Estimation for Bayesian Phylogenetic Model Selection. **Systematic Biology**, v. 60, n. 2, p. 150–160, 2011.
- XU, Xing; POL, Diego. Archaeopteryx , paravian phylogenetic analyses, and the use of probability-based methods for palaeontological datasets. **Journal of Systematic Palaeontology**, v. 12, n. 3, p. 323–334, 2014.

APPENDIX 1 – SUPPLEMENTARY TABLES

TABLE S1. – EVOLUTIONARY MODELS AND RATE HETEROGENEITY - IQTREE

Model	Explanation
MK	Jukes-Cantor type model for morphological data.
ORDERED ASC	Allowing exchange of neighboring states only. Ascertainment bias
Rate Heterogeneity	Explanation
+I	Allowing for a proportion of invariable sites.
+G +I+G +R +I+R +FQ	Discrete Gamma model Invariable site plus discrete Gamma model FreeRate model Invariable site plus FreeRate model Equal State Frequency

IQTREE Manual (2020).

TABLE S2 – EVOLUTIONARY MODELS AND RATE HETEROGENEITY - MRBAYES

Model	Explanation
MK Iset(coding=all)	Jukes-Cantor type model for morphological data Ascertainment bias
Rate Heterogeneity	Explanation
rates=propinv	Allowing for a proportion of invariable sites.
rate=gamma rate=equal	Discrete Gamma model Equal State Frequency

MrBayes Manual (2017).

TABLE S3 – MODEL FINDER CHARSETS

Table S3-1 - Empirical matrices with Invariable Characters

	Architaenoglossa	Stromboidea
Unpartitioned	MK+FQ	MK+FQ+G4
Partitioned charset		
<i>part1</i>	MK+FQ+I+G4	MK+FQ+I+G4
<i>part2</i>	MK+FQ+ASCgm	MK+FQ+ASC+G4
<i>part3</i>	MK+FQ+ASC	MK+FQ+ASC
<i>part4</i>	MK+FQ+ASC	MK+FQ+ASC
<i>part5</i>	MK+FQ+ASC	MK+FQ+ASC
<i>part6</i>		MK+FQ+ASC

Table S3-2 - Empirical matrices with Variable Characters

	Architaenoglossa	Stromboidea	Cerithioidea
Unpartitioned	MK+FQ+ASC	MK+FQ+ASC+G4	MK+FQ+ASC+G4
Partitioned charset			
<i>part1</i>	MK+FQ+ASC	MK+FQ+ASC	MK+FQ+ASC
<i>part2</i>	MK+FQ+ASC	MK+FQ+ASC+G4	MK+FQ+ASC
<i>part3</i>	MK+FQ+ASC	MK+FQ+ASC	MK+FQ+ASC
<i>part4</i>	MK+FQ+ASC	MK+FQ+ASC	MK+FQ+ASC
<i>part5</i>	MK+FQ+ASC	MK+FQ+ASC	MK+FQ+ASC
<i>part6</i>		MK+FQ+ASC	MK+FQ+ASC
<i>part7</i>			MK+FQ+ASC
<i>part8</i>			MK+FQ+ASC
<i>part9</i>			MK+FQ+ASC

Table S3-3 - Simulated matrices

	"Architaenoglossa"	"Stromboidea"	"Cerithioidea"
Unpartitioned	MK+FQ+G4	MK+FQ+ASC+G4	MK+FQ+G4
Partitioned charset			
<i>part1</i>	MK+FQ	MK+FQ+ASC	MK+FQ+I+G4
<i>part2</i>	MK+FQ+ASC	MK+FQ+ASC	MK+FQ+ASC
<i>part3</i>	MK+FQ+ASC	MK+FQ+ASC	MK+FQ+ASC
<i>part4</i>	MK+FQ+ASC	MK+FQ+ASC	MK+FQ+ASC
<i>part5</i>	MK+FQ+ASC	MK+FQ+ASC	MK+FQ+ASC
<i>part6</i>	MK+FQ+ASC+G4	MK+FQ+ASC	MK+FQ+ASC
<i>part7</i>		MK+FQ+ASC	MK+FQ+ASC
<i>part8</i>		MK+FQ+ASC	MK+FQ+ASC
<i>part9</i>		MK+FQ+ASC	
<i>part10</i>		MK+FQ+ASC	
<i>part11</i>		MK+FQ+ASC+G4	
<i>part12</i>		MK+FQ+ASC	

APPENDIX 2 – COMPUTATIONAL SCRIPTS

#1 Code for Homoplasy Partitioned used in Rosa et al, 2019 (code written by Marcos S. Barbeitos), with small modifications by Giovanna Y. S. Omura.

```

package Homoplasy;

use warnings;
use strict;

use Data::Dumper;

use Check;

BEGIN
{
    require Exporter;

    our @ISA = qw( Exporter );

    our @EXPORT = qw( get_hashed_scores get_mb_parts );
}

sub get_hashed_scores
{
    my $file = shift;

    assert_scalar( $file );

    my $fh = _get_file_handle( $file );

    my %hashed_scores;

    my @f_scores;

    my $header_flag = 0;

    while( <$fh> )
    {
        chomp;

```

```

        my $line = $_;

my @tokens = split /\s+|\t/, $_;
shift @tokens;

$header_flag = _update_header_flag( $header_flag, @tokens );
next unless ( $header_flag == 3 );

push @f_scores, @tokens[1 .. @tokens - 1];
    }

    for my $idx ( 0 .. @f_scores - 1 )
    {
        push @{ $hashed_scores{ $f_scores[ $idx ] } }, $idx + 1;
    }

    return \%hashed_scores;
}

sub get_mb_parts
{
    my $hashed_scores = shift;

    assert_ref( $hashed_scores, 'HASH' );

    my $set_count = 1;

    for my $f_score ( sort { $a cmp $b } keys %{ $hashed_scores } )
    {
my @f_scores = @{ $hashed_scores->{ $f_score } };

        print "[f=" . $f_score . "]\ncharset set_" . $set_count . "=";

        my $previous_score = shift @f_scores;

my $score_string = $previous_score;
my $run_flag = 0;

```

```

my $focal_score;

while ( @f_scores )
{
    $focal_score = shift @f_scores;

    if ( $focal_score > ( $previous_score + 1 ) )
    {
        if ( $run_flag )
        {
            $score_string .= '-' . $previous_score . ' ' . $focal_score;
        }
        else
        {
            $score_string .= ' ' . $focal_score;
        }

        $run_flag = 0;
    }
    else
    {
        $run_flag = 1;
    }

    $previous_score = $focal_score;
}

$score_string .= '-' . $focal_score if ( $run_flag );

print $score_string, "\n", "\n";
$set_count++;
}
}

```

```

sub _get_file_handle
{
    my $file = shift;

    assert_scalar( $file );

    open( my $fh, '<:encoding(UTF-8)', $file )

        || die "Cannot open $file";

    return $fh;
}

sub _update_header_flag
{
    my $header_flag = shift;

    assert_scalar( $header_flag );

    my @tokens = @_;

    $header_flag = 1 if ( ( join "", @tokens ) eq '0123456789' && $header_flag == 0 );

    $header_flag = 2 if ( scalar( @tokens ) < 2 && $header_flag == 1 );

    $header_flag = 3 if ( scalar( @tokens ) > 0 && $header_flag == 2 );

    return $header_flag;
}

1;

#!/bin/perl

use warnings;
use strict;

use Data::Dumper;

```

```

use Check;
use Homoplasy;

( $ARGV[0] && -e $ARGV[0] )

    or die "Script requires a valid path to TNT's output file as single argument";

my $hashed_scores = get_hashed_scores( $ARGV[0] );
get_mb_parts( $hashed_scores );

#print scalar( keys %{ $hashed_scores } ), "\n";

```

#2 Code for simulation of datasets proposed in Puttick et al. 2016 (code originally written by Joe O'Reilly and Mark Puttick). Code was heavily modified by Giovanna Y. S. Omura and Fabricius M. C. B. Domingos, and only includes the parts actually used in this study.

```

rm(list=ls(all=T))

#install.packages("phylosim") * phylosim has been archived - install directly from source archive with dependencies
#url <- "https://cran.r-project.org/src/contrib/Archive/phylosim/phylosim_2.1.1.tar.gz"
#pkgFile <- "phylosim_2.1.1.tar.gz"
#download.file(url = url, destfile = pkgFile)
#install.packages(c("R.oo", "ape", "compoisson", "ggplot2"))
#install.packages(pkgs=pkgFile, type="source", repos=NULL)
#unlink(pkgFile)
#install.packages("phangorn")
#install.packages("parallel")
require(phylosim)
require(phangorn)
require(parallel)

#####

# symmetric tree of "Cerithioidea"
# 23 species - 122 characters
library(ape)
library(phytools)
### Template to make a tree
parent <- c(29,35,35,36,36,37,37,38,38,39,39,40,40,41,41,42,42,43,43,44,44,45,45,
            24,24,25,25,26,26,27,27,28,28,29,30,30,31,31,32,32,33,33,34,34)

```

```

child <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10,11,12,13,14,15,16,17,18,19,20,21,22,23,
          25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45)

ematrix <- cbind(parent, child) # This is the edge matrix
ematrix

elength <- c(0.4, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2,
            0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2,
            0.2, 0.2, 0.2, 0.2, 0.4, 0.2, 0.2, 0.2, 0.2, 0.2,
            0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2,
            0.2, 0.2, 0.2, 0.2) # These are the edge.lengths

elength

tlabels <- c("Taxon_1", "Taxon_2", "Taxon_3", "Taxon_4", "Taxon_5",
            "Taxon_6", "Taxon_7", "Taxon_8", "Taxon_9", "Taxon_10",
            "Taxon_11", "Taxon_12", "Taxon_13", "Taxon_14", "Taxon_15",
            "Taxon_16", "Taxon_17", "Taxon_18", "Taxon_19", "Taxon_20", "Taxon_21",
            "Taxon_22", "Taxon_23")

# These are the tip labels

# Create tree

sym <- list (edge = ematrix, edge.lengths = elength, order = "cladewise", tip.label = tlabels, Nnode = 49) # Be very
careful with the spelling of each element to be store in the list

class (sym) <- 'phylo' # Transform in phylogenetic tree
class (sym) <= 'phylo' # check the class

sym

plot.phylo(sym, show.node.label = T)

nodelabels()

#####

# symmetric tree of "Stromboidea"

# 20 species - 102 characters

### Template to make a tree

parent <- c(21, 21, 22, 22, 23, 23, 24, 24, 25, 25,
          26, 26, 27, 27, 28, 28, 29, 29, 30, 30,
          31, 31, 32, 32, 33, 33, 34, 34, 35, 35,
          36, 36, 37, 37, 38, 38, 39, 39)

child <- c(1, 22, 2, 23, 3, 24, 4, 25, 5, 26,

```

```

6, 27, 7, 28, 8, 29, 9, 30, 10, 31,

11, 32, 12, 33, 13, 34, 14, 35, 15, 36,

16, 37, 17, 38, 18, 39, 19, 20)

ematrix <- cbind(parent, child) # This is the edge matrix
ematrix
elength <- c(1, 0.032, 0.968, 0.032, 0.936, 0.032, 0.904, 0.032,
             0.872, 0.032, 0.84, 0.032, 0.808, 0.032, 0.776, 0.032,
             0.744, 0.032, 0.712, 0.032, 0.68, 0.032, 0.648, 0.032,
             0.616, 0.032, 0.584, 0.032, 0.552, 0.032, 0.52, 0.032,
             0.488, 0.032, 0.456, 0.032, 0.444, 0.444)

# 0.392, 0.032)
# These are the edge.lengths
elength
tlabels <- c("Taxon_1", "Taxon_2", "Taxon_3", "Taxon_4", "Taxon_5",
             "Taxon_6", "Taxon_7", "Taxon_8", "Taxon_9", "Taxon_10",
             "Taxon_11", "Taxon_12", "Taxon_13", "Taxon_14", "Taxon_15",
             "Taxon_16", "Taxon_17", "Taxon_18", "Taxon_19", "Taxon_20")

# These are the tip labels

# Create tree
asym <- list (edge = ematrix, edge.lengths = elength, order = "cladewise", tip.label = tlabels, Nnode = 19) # Be very
careful with the spelling of each element to be store in the list
class (asym) <- 'phylo' # Transform in phylogenetic tree
class (asym) <= 'phylo' # check the class
asym

plot.phylo(asym, show.node.label = T)
nodelabels()

##### Set the number of parallel cores on which the code will run (set to nCORES = 1 on windows)

nCORES <- 2

##### Asymmetrical simulation of datasets containing 100, 350, 1000 characters

```

```

mclapply(1:5, mc.cores=nCORES, function(x) sim.and.write(tree=sym, tree_name="sym", length=122, replicates=1,
proportion=0.56, starting_value = x, k = 2, breakPoints=brks,
binTarget=targetDist))

```

```

mclapply(1:5, mc.cores=nCORES, function(x) sim.and.write(tree=sym, tree_name="sym", length=143, replicates=1,
proportion=0.56, starting_value = x, k = 2, breakPoints=brks,
binTarget=targetDist))

```

```

mclapply(1:5, mc.cores=nCORES, function(x) sim.and.write(tree=asym, tree_name="asym", length=102,
replicates=1,
proportion=0.56, starting_value = x, k = 2, breakPoints=brks,
binTarget=targetDist))

```