Version dated: November 25, 2014

RH: Missing data and topology in total evidence approach

Effect of missing data on topological inference using a total evidence approach

Thomas Guillerme[1,2], and Natalie Cooper[1,2]

[1]*School of Natural Sciences, Trinity College Dublin, Dublin 2, Ireland;*

[2]*Trinity Centre for Biodiversity Research, Trinity College Dublin, Dublin 2, Ireland;*

Corresponding author: Thomas Guillerme, School of Natural Sciences, Trinity College Dublin, Dublin 2, Ireland; E-mail: guillert@tcd.ie; Fax: +353 1 6778094; Tel: +353 1 896 2571.

Abstract

Living species represent a marginal part of all species that have ever lived. Ignoring fossil taxa may lead to misinterpretation of macroevolutionary patterns and processes such as trends in species richness, biogeographical history or paleoecology. This fact has led to an increasing consensus among scientists that both living and fossil taxa must be included in macroevolutionary studies. One approach, the Total Evidence approach, uses molecular data from living taxa and morphological data from both living and fossil taxa to infer phylogenies with both living and fossil taxa at the tips. Although the Total Evidence approach seems very promising, it requires a lot of data and is therefore likely to suffer from missing data issues which may affect its ability to infer correct phylogenies.

In this study we assess the effect of missing data on tree topologies inferred from Total Evidence matrices. Using simulations we investigate three major factors that directly affect the completeness of the morphological part of the matrix: (1) the proportion of living taxa with no morphological data, (2) the amount of missing data in the fossil record, and (3) the overall number of morphological characters in the matrix.

We find that, when using a clade conservative metric such as Robinson-Foulds distance, Bayesian method recovers the right topology better than Maximum Likelihood method. However, when using triplets method, there is less significant difference between both methods.

-Overall missing data is doesn't affect topological recovery in Bayesian (around minimum 70% of topological topology in the worth scenario).

-However, one essential way to improve it would be to code missing living taxa.

(Keywords: missing data, Total Evidence, Bayesian, Maximum Likelihood, topology)

Introduction

Although most species that have ever lived are now extinct (Novacek and Wheeler 1992; Raup 1993), the majority of macroevolutionary studies focus solely on living species (e.g. Meredith et al. 2011; Jetz et al. 2012). Ignoring fossil taxa may lead to misinterpretation of macroevolutionary patterns and processes such as the timing of diversification events (e.g. Pyron 2011), relationships among lineages (e.g. Manos et al. 2007) or niche occupancy (e.g. Pearman et al. 2008). This has led to increasing consensus among scientists that fossil taxa must be included in macroevolutionary studies (Jackson and Erwin 2006; Quental and Marshall 2010; Dietl and Flessa 2011; Slater and Harmon 2013; Fritz et al. 2013). However, to do this we need to be able to place living and fossil taxa into the same phylogenies; a task that remains difficult despite recent methodological developments (e.g. Pyron 2011; Ronquist et al. 2012a; Schrago et al. 2013).

Up to now, three main approaches have been used to place both living and fossil taxa into phylogenies. These approaches differ mainly in whether they treat fossil taxa as tips or as nodes in the phylogeny, and in which part of the available fossil data is used (i.e. the age of the fossil only or both its age and morphology). Classical cladistic methods use matrices containing morphological data from both living and fossil taxa and treat each taxon as a tip in the phylogeny. Relationships among the taxa are then inferred using optimality criteria such as maximum parsimony (Simpson 1945). This approach is commonly used by paleontologists but it ignores the additional molecular data available from living species and does not allow use of probabilistic methods for dealing with phylogenetic uncertainty (but see Spencer and Wilberg 2013). Neontologists, on the other hand, more commonly use probabilistic approaches (e.g. Maximum Likelihood or Bayesian methods) based on matrices containing only

molecular data from living species. Because fossil taxa do not usually have available DNA, fossils are used as nodes rather than tips in these phylogenies and their occurrence date are used to time calibrate phylogenies (Zuckerkandl and Pauling 1965). There have been great improvements in the theory and application of these two approaches (e.g. Bapst 2013; Stadler and Yang 2013; Heath et al. 2013) as well as much debate about the "best" approach to use (e.g. Spencer and Wilberg 2013). However neither approach uses all the available data.

A final approach, known as the Total Evidence method, uses matrices containing molecular data from living taxa and morphological data from both living and fossil taxa (Eernisse and Kluge 1993). This approach treats every taxon as a tip in the phylogeny, uses the occurrence age of the fossils to time calibrate the phylogeny, and allows the use of probabilistic methods for estimating phylogenetic uncertainty (Ronquist et al. 2012a). Total Evidence methods have been successfully applied to empirical data (Pyron 2011; Ronquist et al. 2012a; Schrago et al. 2013), and are becoming an increasingly popular way of adding fossil taxa to phylogenies. However, although the Total Evidence approach seems very promising, there is one big drawback in using this approach: it requires a lot of data that can be difficult (or impossible) to collect. The morphological data for living taxa is rarely collected when molecular data is available (e.g. O'Leary et al. 2013 vs. Meredith et al. 2011), and, for fossil taxa, the scarcity of the fossil record only allow to collect the data available (for example, in vertebrates, the hardest parts of the skeleton; Sansom and Wills 2013). Therefore Total Evidence matrices are likely to contain a lot of missing data that may affect the method's ability to infer correct topologies, branch lengths and support values (Salamin et al. 2003).

The effect of missing data on phylogenetic inferences has been widely studied (Wiens 2003, 2006; Wiens and Moen 2008; Lemmon et al. 2009; Roure and Philippe 2011; Sansom and Wills 2013; Pattinson et al. 2014). Missing molecular data has been seen by

some authors as an issue because it can, in some part of the tree, decrease phylogenetic signal (i.e. the evolutionary information contained within the matrix allowing to infer topology and branch length), especially when using large matrices (Lemmon et al. 2009). However, this may not be a major issue because phylogenetic signal is easily increased by: (i) including a "modest" number of highly-covered genes (i.e. approximatively of the genes that are coded for most of the species; Roure and Philippe 2011) (ii) adding a greater number of taxa (especially slowly-evolving taxa or taxa close to the outgroup; Roure and Philippe 2011); and (iii) choosing more appropriate models of sequence evolution (Wiens 2006; Wiens and Moen 2008; Roure and Philippe 2011). Similarly, missing morphological data might be seen as either a major or minor issue for accurately inferring phylogenies depending on the study in question (Wiens 2003; Sansom and Wills 2013; Pattinson et al. 2014). Because soft-tissue characters are rarely preserved in the fossil record, missing data is mainly found in these characters, and is therefore not randomly distributed which can lead to biased placement of fossil taxa in phylogenies (e.g. Sansom and Wills 2013 but see Pattinson et al. 2014). However, the phylogenetic signal is not related to the amount of missing data *per se* but to the number of informative characters for each taxon, therefore missing data is less of an issue than the number of shared informative characters (Wiens 2003).

Although missing data does not appear be a major problem in molecular and morphological matrices separately (Wiens 2003, 2006; Wiens and Moen 2008; Roure and Philippe 2011), it may become more of an issue in Total Evidence matrices containing both molecular and morphological data for living and fossil species. This may be particularly problematic as fossil taxa (generally) do not have molecular data, resulting in a large section of missing data. Until now, no attempt has been made to study the impact of this issue on phylogenetic inference from Total Evidence methods.

In this study, we focused only on topology as one of the two aspects of the

phylogenetic signal (topology and branch length). Even though both aspects are equally important, branch topology is the first and most straightforward aspect reflecting phylogenetic signal (i.e. topological changes are discrete opposed to branch length changes are continuous). Also, interestingly, the effect of Total Evidence method has not been formally assessed in previous studies using fixed topology (Ronquist et al. 2012a; Schrago et al. 2013).

Here we use simulations to assess the effect of missing data on tree topologies inferred from Total Evidence matrices. The molecular part of a Total Evidence matrix acts like a "classical" molecular matrix containing only the living taxa (Ronquist et al. 2012a). The effect of missing data on such matrices is well known (Wiens 2006; Wiens and Moen 2008; Roure and Philippe 2011), therefore, we focus only on missing data in the morphological part of the matrix. We investigate three major parameters that directly affect the completeness of the morphological part of the matrix:

1. the proportion of living taxa with no morphological data;
2. the proportion of missing data in the fossil taxa; and
3. the proportion of missing morphological characters for both living and fossil taxa in the matrix.

We remove data from a Total Evidence matrix by changing the values of these three parameters and then assess how this affects the topology of trees inferred using Maximum Likelihood and Bayesian methods. We chose these parameters because they reflect empirical biases in data availability. The advent of molecular phylogenetics means that morphological data for living species is rarely collected, and few people have the skills to identify characters needed for detailed phylogenetic analysis. Missing data in fossil taxa is very common due to preservation biases (Sansom and Wills 2013),

and the overall number of characters depends on the effort of the people identifying them.

We find that when using a Maximum Likelihood approach, as missing data increases, the likelihood of recovering the correct tree topology decreases. However, even with no missing data, Total Evidence matrices dramatically reduce the performance of Bayesian methods for inferring tree topology. We propose that this drastic difference between Bayesian and Maximum Likelihood methods is due to a flattening of the likelihood landscape caused by the unavoidable amount of missing molecular data for fossil taxa in a Total Evidence matrix. We make suggestions for how best to deal with this issue when inferring phylogenies from Total Evidence matrices.

Methods

To explore how missing data in the morphological sections of Total Evidence matrices influences tree topology, we used the following protocol (note that we explain each step in detail below this general outline; Fig. 1).

1. Generating the matrix

   We randomly generated a birth-death tree (hereafter called the ″true″ tree) and used it to infer a matrix containing both molecular and morphological data for living and fossil taxa (hereafter called the ″complete″ matrix).

2. Removing data

   We removed data from the morphological part of the ″complete″ matrix to simulate the effects of missing data by modifying three parameters (i) the proportion of living taxa with no morphological data (M_L), (ii) the proportion of missing data in the fossil taxa (M_F) and (iii) the proportion of missing morphological characters (M_C) (the resulting matrices are called hereafter ″missing-data″ matrices).

3. Inferring phylogenies

   We inffered phylogenetic trees from the ″complete″ matrix and from the ″missing-data″ matrices resulting in one tree generated from a matrix containing no missing data (hereafter called the ″best″ tree) and multiple trees inferred from matrices with missing morphological data (hereafter called the ″missing-data″ trees). Phylogenies were inferred via both Maximum Likelihood and Bayesian approaches.

4. Comparing topologies

We compared the ”best” tree to the ”missing-data” trees to assess the influence of each parameter (M_L, M_F, M_C) and their interactions on the topologies of our phylogenies

We repeated steps 1 to 4 50 times.

Generating the matrix

First we randomly generated a ”true” tree of 50 taxa in R v3.0.2 (R Core Team 2014) using the package diversitree v0.9-6 (FitzJohn 2012). We generated the tree using a birth death process by sampling speciation (λ) and extinction (μ) rates from a uniform distribution but maintaining $\lambda > \mu$ (Paradis 2011). We implemented a rejection sampling algorithm to select only trees with 25 living and 25 fossil taxa to ensure that we had enough taxa of each type for our missing data simulations to work. We then added an outgroup to the tree, using the mean branch length of the tree to separate the outgroup from the rest of the taxa, and with the branch length leading to the outgroup set as the sum of the mean branch length and the longest root-to-tip length of the tree.

Next, we generated a molecular and a morphological matrix from the ”true” tree. The molecular matrix was inferred from the ”true” tree using the R package phyclust v0.1-14 (Chen 2011). The matrix contained 1000 character sites for 51 taxa and was generated using the seqgen algorithm (Rambaut and Grassly 1997) and using the HKY model (Hasegawa et al. 1985) with random base frequencies and transition/transversion rate of 2 (Douady et al. 2003). The substitution rates were distributed following a gamma distribution with an alpha (α) shape of 0.5 (Yang 1996). We chose a low value of α to reduce the number of sites with high substitution rates, thus avoiding too much homoplasy and a decrease in phylogenetic signal. We selected the parameters above to generate data with no special assumption about how the

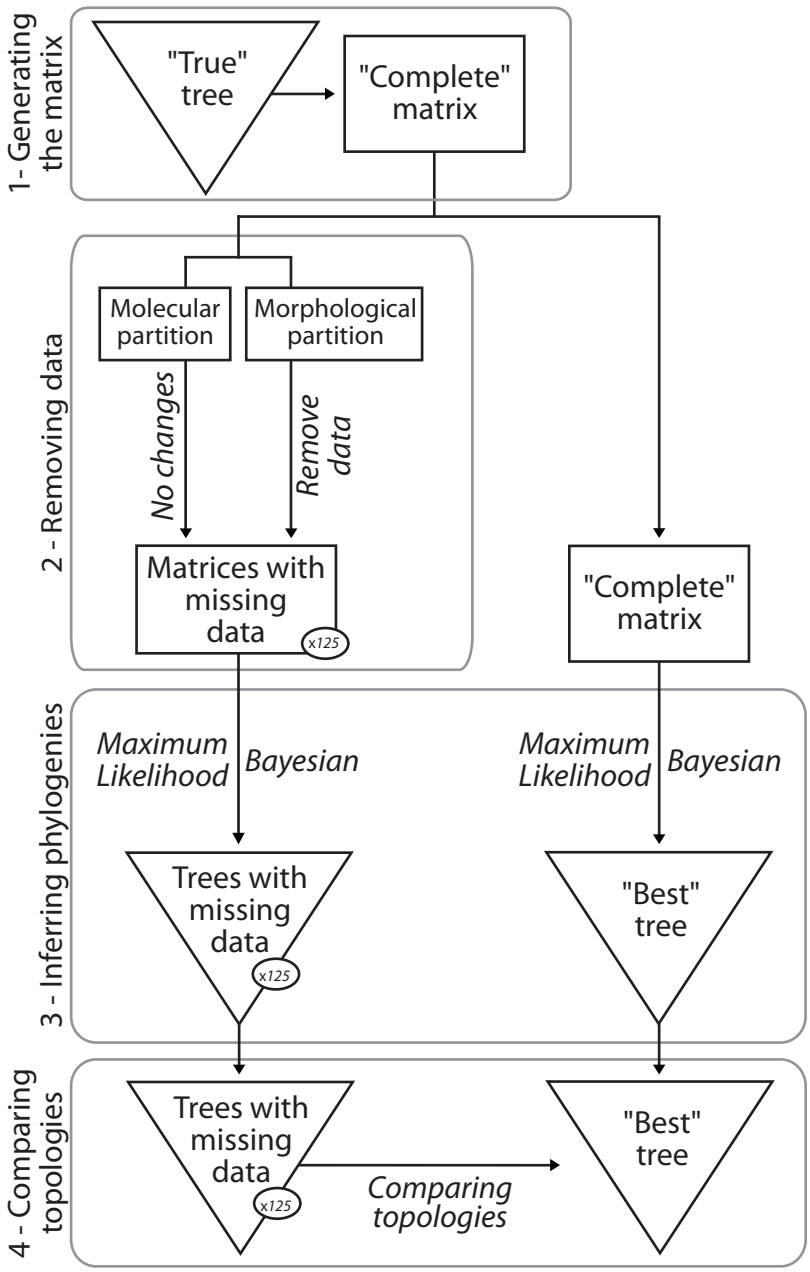

Figure 1: Protocol outline. (1) We randomly generated a birth-death tree (the ″true″ tree) and used it to infer a matrix with no missing data (the ″complete″ matrix). (2) We removed data from the morphological part of the ″complete″ matrix resulting in 125 ″missing-data″ matrices. (3) We built phylogenetic trees from each matrix using both Maximum Likelihood and Bayesian methods. (4) We compared the ″missing-data″ trees to the ″best″ tree. We repeated steps 1-4 50 times.

characters evolved, and to reduce the computational time required if these parameters were estimated rather than defined in the tree building part of the analysis (even with the parameters defined, total computational time for the whole analysis was over 150 CPU years). All the molecular information for fossil taxa was replaced by missing data ("?").

We inferred the morphological matrix using the R package ape v3.0-11 (Paradis et al. 2004) to generate a matrix of 100 character sites for 51 taxa. We assigned the number of character states (either two or three) for each morphological character by sampling with a probability of 0.85 for two states characters and 0.15 for three state characters. These probabilities were selected using the overall distribution of character states extracted from 100 published empirical morphological matrices (See supplementaries). We then ran an independent discrete character simulation for each character using the "true" tree with the character's randomly selected number of states (two or three) and assuming an equal rate of change (i.e. evolutionary rate) from one character state to an other (Pagel 1994). This method allows us to have only two parameters for each character: the number of states and the evolutionary rate. For each character, the evolutionary rate was sampled from a gamma distribution with $\alpha = 0.5$. We used low evolutionary rate parameters (i.e. α) to avoid homoplasy in the morphological part of the matrix and create a clear phylogenetic signal (Wagner 2000; Davalos et al. 2014).

Finally, we combined the morphological and molecular matrices obtained from the "true" tree. Hereafter we call this the "complete" matrix: the matrix with no missing data except for the molecular data of the fossil taxa.

Removing data

We modified the "complete" matrix to get matrices with missing data by randomly

replacing data with "?" in the morphological part of the matrices according to the following parameters:

1. The proportion of living taxa with no morphological data (M_L): 0%, 10%, 25%, 50% or 75%. This parameter illustrates the number of living taxa that are present in the molecular part of the matrix but not in the morphological part. This reflects the fact that because of the increasing availability of DNA sequences for living taxa, detailed morphological data is scarce.

2. The proportion of missing data in the fossil taxa (M_F): 0%, 10%, 25%, 50% or 75%. This parameter illustrates the quality of the fossil record.

3. the proportion of missing morphological characters for both living and fossil taxa (M_C): 0%, 10%, 25%, 50% or 75%. This parameter illustrates the number of available morphological characters for both living and fossil taxa.

In practice, each parameter represents a different way of removing data from the matrix: M_L removes rows from the living taxa; M_F removes cells from the fossil taxa; and M_C removes columns across both living and fossil taxa. Note that M_L is different to M_F not only because of the region of the matrix affected: for M_L, all the morphological data of a percentage of living taxa is removed, but for M_F, a percentage of the data is removed at random from across the whole of the morphological matrix for fossil taxa.

We tested all parameters combinations resulting in 125 (5^3) matrices. Note that one of these combinations has no missing data so is equivalent to the "complete" matrix, thus we have one effectively complete matrix in our 125 "missing-data" matrices. Because some parameter combinations introduce a lot of missing data (e.g. M_L=75%, M_F=75% and M_C=75%), some matrices contained fossil taxa without any data at all. When this occurred we repeated the random deletion of characters until every taxa had at least 5% data across the whole morphological part of the matrix.

Building phylogenies

From the resulting matrices we generated two types of trees, the ″best″ tree inferred from the ″complete″ matrix and the ″missing-data″ trees inferred from the 125 matrices with various amounts of missing data. The ″true″ tree was used to generate the ″complete″ matrix and reflects the ″true″ evolutionary history in our simulations. The ″best″ tree, on the other hand, is the best tree we can build using state-of-the-art phylogenetic methods. In real world situations, the ″true″ tree is never available to us because we cannot know the true evolutionary history of a clade (except in very rare circumstances, e.g. Rozen et al. 2005). Therefore, here we focus on comparing the trees inferred from the matrices with missing data to the ″best″ tree, rather than the ″true″ tree, as the ″best″ tree is generally what biologists have to work with.

Maximum Likelihood.— The ″best″ tree and the ″missing-data″ trees were inferred using RAxML v8.0.20 (Stamatakis 2014). For the molecular data, we used the GTR + Γ_4 model (Tavaré 1986; default GTRGAMMA in RAxML v8.0.20; Stamatakis 2014) as a generalization of the HKY + Γ_4 model (Hasegawa et al. 1985) for the molecular data. For the morphological data, we used the implemented Markov k state model (Lewis 2001) assuming an equal state frequency and a unique overall substitution rate (μ) following a gamma distribution of the rate variation with four distinct categories (Mk + Γ_4; -K MK option in RAxML v8.0.20; Stamatakis 2014).

In order to measure the phylogenetic signal of our simulations, we first ran a fast bootstrap analysis with 500 replicates on the ″complete″ matrix. We removed all the simulations that had a median bootstrap support lower than 50 as a proxy for weak phylogenetic signal (Zander 2004). We repeated this selection until we obtained 50 sets of simulations (i.e. 50 ″complete″ and 50*125 ″missing-data″ matrices) with a relative good phylogenetic signal (median bootstrap $>$ 50).

On these selected simulations, we used the fast bootstrap algorithm and performed 1000 bootstraps per tree inference to assess the topological support (Pattengale et al. 2010). When using these parameters, it took 6 CPU years to build 50 sets of 125 bootstrapped Maximum Likelihood trees (8 core nodes 2.30GHz clock speed).

Bayesian.— The "best" tree and the "missing-data" trees were inferred using MrBayes v3.2.1 (Ronquist et al. 2012b). We partitioned the data to treat the molecular part as a non-codon DNA partition and the morphological part as a multi-state morphological partition. The molecular evolutionary history was inferred using the HKY model with a transition/transversion ratio of two (Douady et al. 2003) and a gamma distribution for the rate variation with four distinct categories (HKY + Γ_4). For the morphological data, we used the Markov *k* state model (Lewis 2001), with equal state frequency and a unique overall substitution rate (μ) with four distinct rates categories (M*k* + Γ_4). We chose these models to be consistent with the parameters used to generate the "complete" matrix.

Each Bayesian tree was estimated using two runs of four chains each for a maximum of 50×1^6 generations. We used the average standard deviation of split frequencies (ASDS) as a proxy to estimate the convergence of the chains and used a stop rule when the ASDS went below 0.01 (Ronquist et al. 2012b). The effective sample size (ESS) was also checked on a random sub-sample of runs in each simulation to ensure that ESS $>>$ 200 (Drummond et al. 2006). For each run, we removed 25% of the iterations as burn-in. We used the following priors for each tree (see Supplementary Material S1):

1. the "true" trees topology as a starting tree (with a starting value for each branch length of 1),

2. an exponential prior on the shape of the gamma distribution of $\alpha = 0.5$ for both partitions, and

3. a transition/transversion ratio prior of two sampled from a strong beta distribution ($\beta(80,40)$).

We used these prior to speed up the Bayesian estimation process. These priors biased the way the Bayesian process calculated branch lengths by giving non-random starting points and boundaries for parameter estimation, however, here we are focusing on the effect of missing data on tree topology and not branch lengths. Even using these priors, it took 140 CPU years to build 50 sets of 125 Bayesian trees (8 core nodes 2.30GHz clock speed).

Comparing topologies

We compared the topology of the "missing-data" trees to the "best" tree to measure the effect of the three parameters M_L, M_F and M_C on tree topology. We used the Robinson-Foulds distance (Robinson and Foulds 1981) to identify conserved clade positions and the Triplets distance (Dobson 1975) to assess the number of conserved taxa across trees. We then used Normalized Tree Similarity index (Bogdanowicz et al. 2012) to generalize our results for any n number of taxa. These metrics are described in detail below.

Robinson-Foulds distance.— Robinson-Foulds distance (Robinson and Foulds 1981), or "path difference", measures the number of shared clades across two trees. The metric reflects the distance between the distributions of tips among clades in the two trees (Robinson and Foulds 1981 ; see Supplementary Material S2). This metric is bounded between 1 when the two trees are identical and $n - 2$ (for two trees with n taxa) when

there is not one single shared clade between both trees. This metric is sensitive to the exact clade conservation: if the trees are composed of two clades of three taxa (*((((a,b),c),((d,e),f))*), the swap of two taxa will lead to a maximal score of the Robinson-Foulds distance indicating a bad tree similarity.

Triplets distance.— The Triplets distance (Dobson 1975) measures the number of sub-trees made up of three taxa that differ between two given trees (Critchlow et al. 1996 ; see Supplementary Material S2). This metric measures the position of each taxon and clade towards it's closest neighbours. It is bounded between 0 when the two trees are identical and $\binom{n}{4}$ (for two trees with n taxa) when there is not one single position of taxon/clade identical between both trees. Therefore this metric sensitive to the conservation of individual taxa towards the neighbouring trees.

Normalized Tree Similarity.— We used the Normalized Tree Similarity index, NTS_m (Bogdanowicz et al. 2012) to be able to compare the two metrics for any n taxa. This index allows to scale the value of any metric m (either Robinson-Foulds or Triplets distance in our study) to the expected value of the metric m when comparing two random trees (see Supplementary Material S2). When NTS_m=1, the two trees are strictly identical, when NTS_m=0 the trees are no more different than expected when comparing two random trees and when NTS_m<0, the difference between the two trees is greater than when comparing two random trees. In our study we used the NTS_m index as a proxy for topological recovery: a high score of this index (i.e. towards 1) means that the topology is highly conserved between the two trees; on the opposite, a low score of this index (i.e. towards 0) means that the topological difference between the two trees is as much as expected when comparing two random trees.

Tree comparisons.— For the Maximum Likelihood and Bayesian consensus trees we

performed pairwise comparisons between the ”best” tree and each ”missing-data” tree using both the Robinson-Foulds and Triplets metrics with the TreeCmp java script (Bogdanowicz et al. 2012). For each metric, we then normalized the value using the Normalized Tree Similarity scaled by the mean value of 1000 pairwise random tree comparisons for the metric in question and n = 51 taxa (see Supplementary Material S2). We compared each ”missing-data” tree with the ”best” tree for each of our 50 simulation runs resulting in 50 comparisons for each ”missing-data” tree. We calculated the mode and the 50% and 95% confidence intervals from the resulting distribution using the hdrcde R package v3.1 (with contributions from Jochen Einbeck and Wand 2013).

Also, to take into account the uncertainty of tree inference in both Maximum Likelihood and Bayesian (i.e node support), we ran 1000 random pairwise comparison between respectively the bootstrapped trees from the Maximum Likelihood analysis and the posterior tree distribution of the Bayesian analysis. In the same way that we compared a single ”missing-data” tree to the ”best” tree (whether the trees are Maximum Likelihood or Bayesian consensus): we randomly selected 1000 trees from the ”missing-data” tree sets (either the Bootstrapped trees or the posterior tree distribution) and did a pairwise comparison with 1000 randomly selected trees from the ”best” tree set.

For each of the 125 ”missing-data” tree, we obtained Robinson-Foulds and Triplets distance distributions from either 50 or 50*1000 pairwise comparisons. We then calculated the mode and the 50% and 95% confidence intervals from the resulting distribution using the hdrcde R package v3.1 (with contributions from Jochen Einbeck and Wand 2013). Finally we tested for significant differences among the different distributions of trees using non-parametric group and pairwise comparisons by using respectively Kruskal-Wallis and Nemenyi-Damico-Wolfe-Dunn test (CITE R

PACKAGE)Ruxton and Beauchamp (2008).

Results

Effect of the method on topological recovery

-With RF Bayesian significantly better than ML for clade conservation.

-With RF Bayesian consensus trees plateaus around 0.8 topological recovery at worst whatever combination of missing data.

-With Tr however, ML is more performant but not significantly. When looking at uncertainty, both results tend to just overlap.

Effect of the missing data on topological recovery

-No big difference between the

Effect of M_L

In our Maximum Likelihood analyses, topological recovery (i.e. tree similarity) decreased as the percentage of missing data increased, using both the Robinson-Foulds and Triplets metrics. However, topological recovery was not significantly different for trees with 25%, 50% and 75% missing morphological data for living taxa (Fig 1b???), suggesting that although tree topology is sensitive to this missing data, once it is greater than 25% it makes little difference. Overall, the Robinson-Foulds metric shows lower tree similarities than the Triplets metric. This indicates that the positions of individual taxa are generally more conserved than the membership of entire clades.

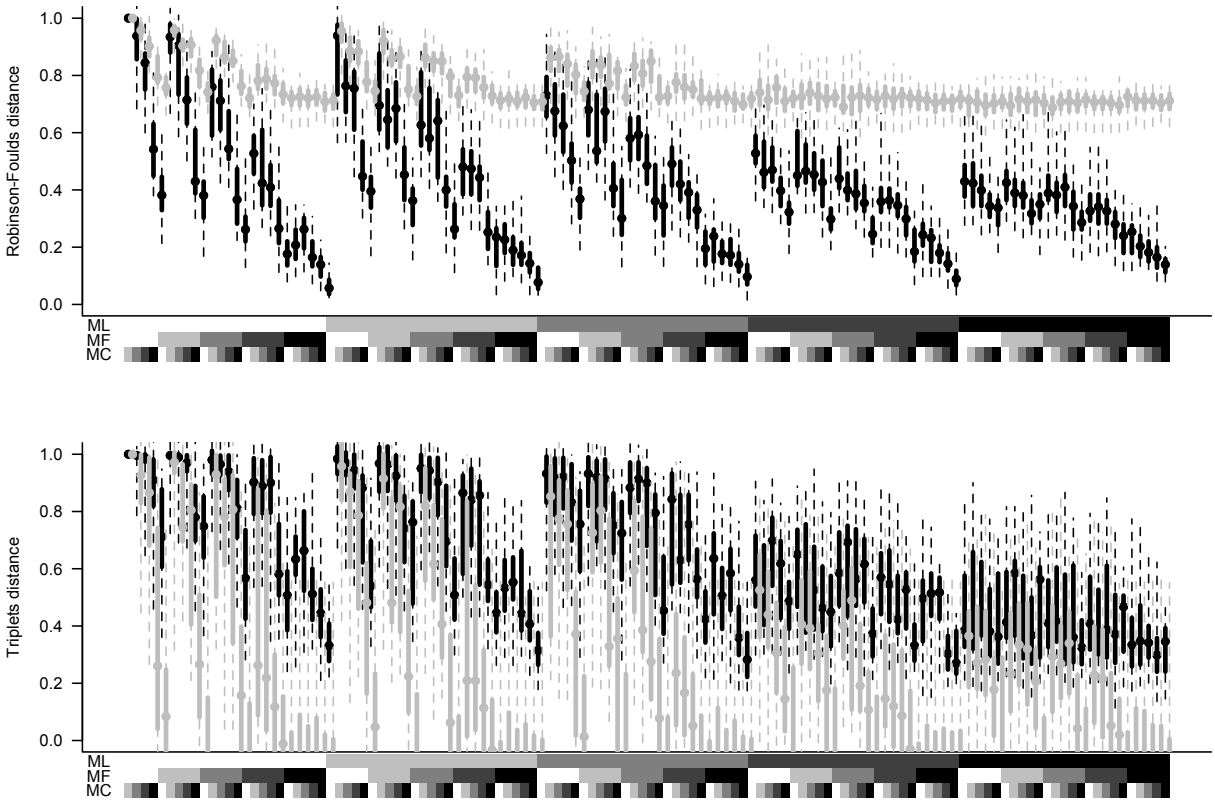

Figure 2: Trend of the effect of missing data on topological recovery on ML and consensus trees. The amount of missing data per parameter (M_L, M_F and M_C) is represented along the x axis. The colour gradient from white to black represents respectively, 0%, 10%, 25%, 50% and 75% of missing data. The topological recovery is represented on the y axis, both using Robinson-Foulds distance (upper row) and Triplets distance (lower row). Points represent the modal value of each distribution ; thick solid and thin dashed lines represents respectively the 50% and 95% confidence intervals or the distributions. The Maximum Likelihood trees are represented in black and the Bayesian consensus trees in grey.

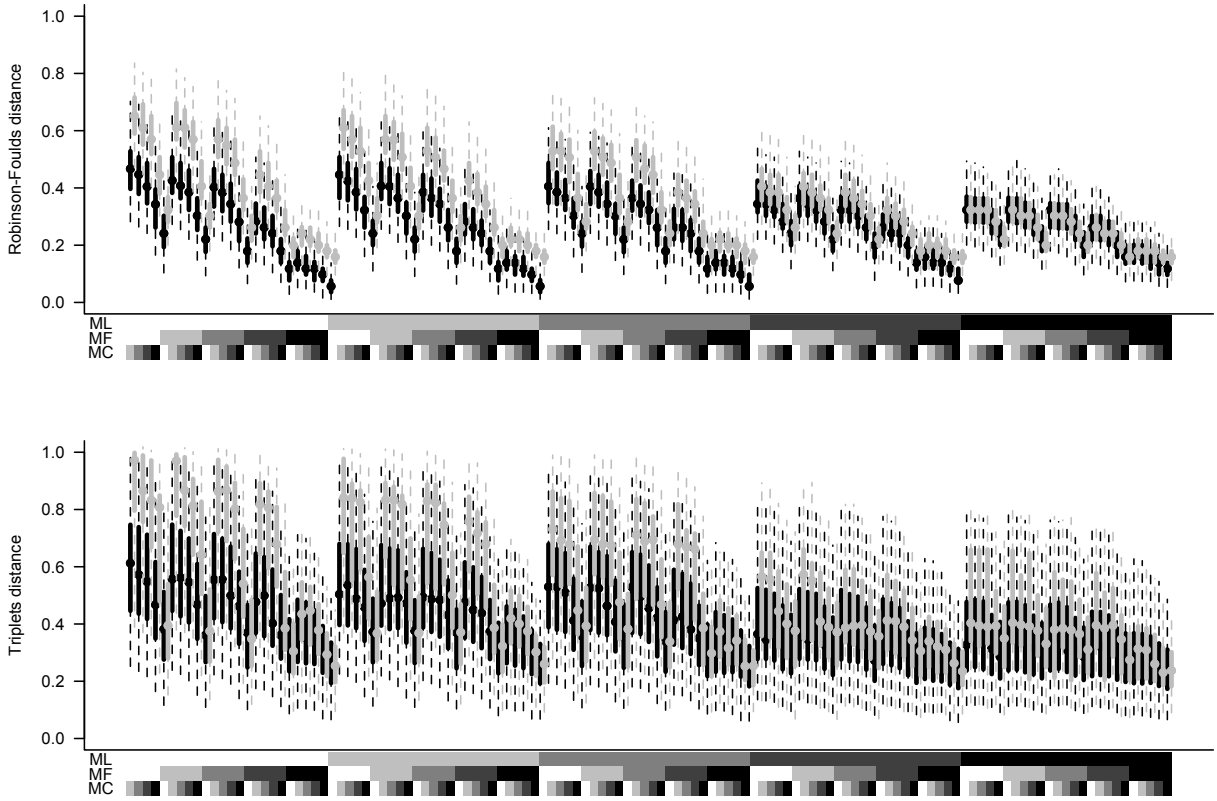

Figure 3: Trend of the effect of missing data on topological recovery on ML and consensus trees. The amount of missing data per parameter (M_L, M_F and M_C) is represented along the x axis. The colour gradient from white to black represents respectively, 0%, 10%, 25%, 50% and 75% of missing data. The topological recovery is represented on the y axis, both using Robinson-Foulds distance (upper row) and Triplets distance (lower row). Points represent the modal value of each distribution ; thick solid and thin dashed lines represents respectively the 50% and 95% confidence intervals or the distributions. The Bootstrap trees pairwise comparisons are represented in black and the Bayesian posterior trees distribution pairwise comparisons in grey.

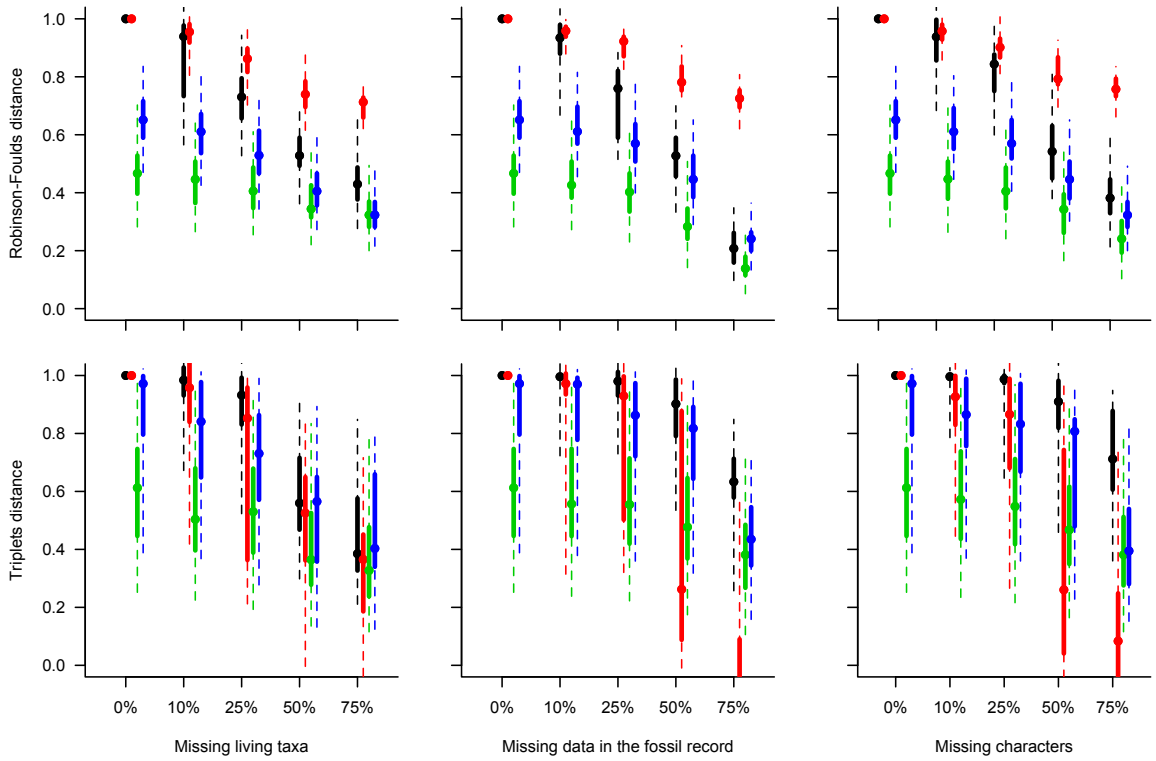

Figure 4: Comparison between the effect of missing data and the tree inference method on topological recovery. The amount of missing data for each parameter is represented on the x axis. The topological recovery is represented on the y axis, both using Robinson-Foulds distance (upper row) and Triplets distance (lower row). Points represent the modal value of each distribution ; thick solid and thin dashed lines represents respectively the 50% and 95% confidence intervals or the distributions. The Maximum Likelihood trees are represented in black, the Bayesian consensus trees in red, the bootstrap trees in green and the posterior tree distribution in blue.

Topological recovery was poor in our Bayesian analyses, regardless of the amount of missing data (Figures???), although topological recovery did decrease slightly as missing data increased. Overall, tree similarity index was only slightly above zero which means that trees are only slightly more similar than we would expect if we compared two completely random trees.

Effect of M_F.— In Maximum Likelihood framework, the effect of missing data in the fossil record appeared to be constant and led to a constant decrease in topological recovery when the missing data increase. The difference between the two metrics used is more important than for the M_L parameter: clades are less conserved than individual taxa placement as the amount of missing data increases. Interestingly, 10% or 25% missing data for the M_F parameter does not seems to affect the apparition of unstable taxa .

Similarly than for the effect of M_L, Bayesian tree recovery was only slightly better than expected by chance. In the same way as described above, there seems to be a small difference depending on the metric used. There is no significant effect of the M_F parameter when using the Triplet metric but the effect is significant when using the Robinson-Foulds metric . However, as mentioned previously the effect of the parameter M_F is still minimal in a Bayesian framework .

Effect of M_C.— The number of missing morphological characters (M_C), however, seems to be more affecting topological recovery than M_L and M_F . The Robinson-Foulds metric shows a rapid drop in tree recovery from 10% of missing data (Robinson-Foulds NTS mode $<$ to 0.5 from 10% missing data . This decrease is however slower when using the Triplets metric with still a good topological recovery at 10% (Triplets NTS modes = 0.95 .

The effect of M_C in Bayesian framework is similar as the one described for M_L

and M_F: tree recovery is only slightly better than expected by chance and only the Robinson-Foulds metric shows a significant effect of the M_C parameter on tree recovery.

Discussion

Building the trees

Simulating evolutionary history matrices still remains a big drawback in theoretical phylogenetics. The size of our simulated matrices was at least two orders orders of magnitude lower than usual matrices, both for the molecular part (e.g. Springer et al. 2012) and the morphological part (e.g. Ni et al. 2013). This configuration probably lead to globally low phylogenetic signal as well as the intrinsic difficulties to simulated characters with phylogenetic signal. Even though molecular characters evolution (and therefore simulation) only depends of a small number of parameters (i.e. the base frequencies and the substitution matrix), simulating molecular matrix with a strong evolutionary signal is still complicated when generating unrealistically small matrices. For morphological characters, the underlying pattern of their evolution are often more complex and ruled by more parameters than molecular characters (i.e the number of character states, the states frequencies, the substitution matrix and the statistical model used) (Pagel 1994; Wagner 2000; Lewis 2001).

Also morphological characters studies involve many potential statistical pitfalls (e.g. independent characters violation, rate variation - Davalos et al. (2014)) and especially (i) incongruence with molecular signal and (ii) homoplasy.

1. First, morphological data can display a different signal than molecular data, especially in small matrices. This might lead to a controversial phylogenetic signal in the overall matrix and lower down the support values. However, regarding empirical data studies, most of the groups shows fairly congruent morphological and molecular phylogenetic signal (e.g. Lee et al. 2013).

2. Secondly, in this study, we made the assumption that theoretically, morphological characters are randomly distributed on an organism however it seems clear that empirical morphological data does not act randomly (Sansom and Wills 2013). However, following our simulation assumption of random character distributions, if they accumulate through time in the same way as the majority of the molecular characters then, homoplasic characters are expected to appear randomly through time (Davalos et al. 2014). Therefore, homoplasy is expected to be more important (by chance) in bigger morphological matrices (Davalos et al. 2014). After a reaching a critical amount of morphological characters, adding new ones increases homoplasy (Wagner 2000).

Our simulation parameters both decrease the phylogenetic signal (difficulty to simulate morphological characters and size of the matrix) as well as it increases it (reduction of homoplasy in the morphological part of the matrix). Therefore, these drawbacks seems to have only a minor impact on the main results of this study: the incapacity of recovering any topology in Bayesian inference.

Comparing topologies

Comparing topologies is a crucial question in phylogenetics but has always been hard to normalise because of the vast amount of different metrics used as proxies for different aspects of tree similarity/dissimilarity (Agapow and Purvis 2002). Because

our global framework is to study how to include efficiently fossils into phylogenies, we chose metrics reflecting the most interesting aspect of this global question: where do individual taxa (i.e. the fossils) branch in the tree. The Robinson-Foulds (Robinson and Foulds 1981) and the Triplets (Critchlow et al. 1996) metrics are more sensible to taxa and clade placement than other tree comparison metrics (e.g. Kirkpatrick and Slatkin 1993, Imbalance metric) and where therefore favored. Also by using the Normalized Tree Similarity index (Bogdanowicz et al. 2012) we emphasize the aspect of "good" phylogenetic signal *versus* random phylogenetic signal because it normalized the values of the metric by correcting for the expected value when comparing random trees (NTS = 0).

The Robinson-Foulds metric is a conservative tree topology metric, it is more sensible to single taxon displacement because it will count clades as similar only if they are composed of the same number of taxa with the same topologies (Robinson and Foulds 1981). When getting closer to the root of the tree, displacement of single taxon makes the clades not being exactly identical any more even if the clade still contains all the other taxa in both trees. On the other hand, the Triplets method is measuring the position of each taxon towards to other reference taxa (Critchlow et al. 1996). It will penalise only trees where taxa get removed furthest from their original clade. Regarding our problematic (how does missing data influence topological recovery in trees containing both living and fossil taxa) we are more interested the placement of taxa (i.e. where does the fossil branch) than the exact conservation of clades.

Although the idea of comparing two single trees is straightforward, it becomes more complex when comparing trees distributions (e.g. Bayesian posterior distributions). The introduction of our Random Pairwise Bayesian Tree Comparison methods allows to summarize the comparison of two trees distributions by picking up the most frequent signal in the distribution of the pairwise comparisons (i.e. the mode).

Even though this method is subject to randomness and might artificially increase or decrease the score of the studied metric, simulations showed that when a sufficient amount of random comparisons is performed, this method doesn't seem to be subject to randomness (e.g. 1000 random pairwise comparisons - see **??**).

Maximum Likelihood versus Bayesian

The main results of our analysis shows that Bayesian inference fails to recover any Topology in a Total Evidence method framework (Fig. **??**). This results is surprising regarding the behaviour of Maximum Likelihood inference (Fig. **??**) as well as the Bayesian inferences performed on empirical data (Ronquist et al. 2012a; Schrago et al. 2013). However, it is important to note that this effect was not mentioned in the aforementioned empirical tests because both studies used a Bayesian approach with fixed topology (Ronquist et al. 2012a; Schrago et al. 2013).

In our case, we suspect that this inability of Bayesian methods to recover topology is due to the intrinsic structure of a Total Evidence matrix. In fact, as previously studies have shown, missing data doesn't seem to be a major drawback as long as the missing data is randomly distributed (e.g. Wiens 2003; Roure and Philippe 2011; Sansom and Wills 2013). However, in Total Evidence matrices, the majority of the missing data is not randomly distributed but concentrated in the molecular part of the matrix for the fossil taxa (i.e. the missing-data sub-matrix). This leads to high decrease of topological recovery when using non optimal criterion approach (i.e. Bayesian inference) because of the high variance in the near-likeliest solutions sampled in the Bayesian posterior distribution. In opposition, when applying optimal criterion approach (i.e. Maximum Likelihood), it is still possible to sample the likeliest tree.

Effect of missing data

The three parameters we selected in this study account for three potential pitfalls in collecting the data for Total Evidence analysis: M_L represents the living taxa for which there is no available morphological data, M_F represents the quality of the fossil record, and M_C represents the general coding effort and the overall of morphological data available. Ideally, the lowest possible amount of missing data is wanted in any phylogenetic analysis leading to a data collection part prior to the phylogenetic analysis. Each of the three parameters can be improved in a different way prior to the analysis: for M_L, one should put more effort in using natural museums history collections for coding the missing morphological data for living taxa if possible for the M_F parameter, the amount of missing data unfortunately depends on the quality of the fossil record and can not be actively improved and depends on exceptional discoveries (e.g. Ni et al. 2013); finally for the M_C parameter, improvement can be done by vast collaborative projects in order to gather as much characters as possible (e.g. O'Leary et al. 2013).

Our results shows that the M_F parameter have less influence on recovering the good tree topology in Maximum Likelihood framework which is fortunate since it is the parameter that is the more difficult to fix for practical reasons. Regarding the fact that we are more interested in taxa placement than in clade conservation (Triplets *versus* Robinson-Fould), the parameter that affects the more the decrease in topological recovery is the number of missing living taxa in the morphological part of the matrix (M_L) and the overall number of morphological characters M_C. This makes sense since the living taxa are bearing both the information to build the tree backbone (the molecular data) and the information used for branching the fossil on this backbone (the morphological information). Therefore, we advocate the importance of coding morphological characters for the most living taxa possible and with the most characters possible, potentially by using collaborative projects portals such as morphobank

(OLeary and Kaufman 2011).

Conclusion

A Bayesian approach fails to recover accurate topology in a Total Evidence approach framework, whatever the amount of missing data. We think this failure is due to the intrinsic structure of the Total Evidence matrices that includes a vast amount of non randomly distributed missing data (i.e. the molecular part of the matrix for the fossil taxa). This missing data is equal to the number of fossil taxa $\times$ the number of molecular characters leading to a vast amount of trees to be sampled in Bayesian posterior distribution. However, one can use a Maximum Likelihood approach to fix the likeliest topology. If so, an effort should be made prior to the phylogenetic analysis on collecting as much morphological data as their is available from living taxa in order to efficiently improve the quality of the trees.

Acknowledgements

Thanks to Frédéric Delsuc, Emmanuel Douzery, Trevor Hodkinson and Andrew Jackson, Gavin Thomas, April Wright and the members of the Macro Journal Club for useful comments on our simulation protocol. Thanks to Paddy Doyle, Graziano D'Innocenzo and Sean McGrath for assistance with the computer cluster. Simulations used the Lonsdale cluster maintained by the Trinity Centre for High Performance Computing and funded through grants from Science Foundation Ireland. This work was funded by a European Commission CORDIS Seventh Framework Programme (FP7) Marie Curie CIG grant (proposal number: 321696).

*

References

Agapow, P. and A. Purvis. 2002. Power of eight tree shape statistics to detect nonrandom diversification: a comparison by simulation of two models of cladogenesis. Systematic Biology 51:866–872.

Bapst, D. W. 2013. A stochastic rate-calibrated method for time-scaling phylogenies of fossil taxa. Methods in Ecology and Evolution 4.

Bogdanowicz, D., K. Giaro, and B. Wrbel. 2012. Treecmp: Comparison of trees in polynomial time. Evolutionary Bioinformatics 8:475–487 10.4137/EBO.S9657.

Chen, W.-C. 2011. Overlapping codon model, phylogenetic clustering, and alternative partial expectation conditional maximization algorithm. Ph.D. thesis.

Critchlow, D. E., D. K. Pearl, and C. Qian. 1996. The triples distance for rooted bifurcating phylogenetic trees. Systematic Biology 45:323–334.

Davalos, L. M., P. M. Velazco, O. M. Warsi, P. D. Smits, and N. B. Simmons. 2014. Integrating incomplete fossils by isolating conflicting signal in saturated and non-independent morphological characters. Systematic Biology .

Dietl, G. and K. Flessa. 2011. Conservation paleobiology: putting the dead to work. Trends in Ecology & Evolution 26:30–37.

Dobson, A. J. 1975. Comparing the shapes of trees Pages 95–100. Springer Berlin Heidelberg.

Douady, C., F. Delsuc, Y. Boucher, W. Doolittle, and E. Douzery. 2003. Comparison of bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. Molecular Biology and Evolution 20:248–254.

Drummond, A. J., S. Y. Ho, M. J. Phillips, and A. Rambaut. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biology 4:e88.

Eernisse, D. and A. Kluge. 1993. Taxonomic congruence versus total evidence, and amniote phylogeny inferred from fossils, molecules, and morphology. Molecular Biology and Evolution 10:1170–1195.

FitzJohn, R. G. 2012. Diversitree : comparative phylogenetic analyses of diversification in r. Methods in Ecology and Evolution 3.

Fritz, S., J. Schnitzler, J. Eronen, C. Hof, B. Katrin, and C. Graham. 2013. Diversity in time and space: wanted dead and alive. Trends in Ecology & Evolution .

Hasegawa, M., H. Kishino, and T. A. Yano. 1985. Dating of the human ape splitting by a molecular clock of mitochondrial-dna. Journal of Molecular Evolution 22:160–174.

Heath, T., J. Huelsenbeck, and T. Stadler. 2013. The fossilized Birth-Death process: a coherent model of fossil calibration for divergence time estimation .

Jackson, J. and D. Erwin. 2006. What can we learn about ecology and evolution from the fossil record? Trends in Ecology & Evolution 21:322–328.

Jetz, W., G. Thomas, J. Joy, K. Hartmann, and A. Mooers. 2012. The global diversity of birds in space and time. Nature 491:444–448.

Johnson, L. A. and D. E. Soltis. 1998. Assessing congruence: empirical examples from molecular data chap. 11, Pages 297–348. Springer US.

Kirkpatrick, M. and M. Slatkin. 1993. Searching for evolutionary patterns in the shape of a phylogenetic tree. Evolution Pages 1171–1181.

Lee, M., J. Soubrier, and G. Edgecombe. 2013. Rates of phenotypic and genomic evolution during the cambrian explosion. Current Biology 23:1889 – 1895.

Lemmon, A., J. Brown, S. Kathrin, and E. Lemmon. 2009. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and bayesian inference. Systematic Biology 58:130–145.

Lewis, P. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. Systematic Biology 50:913–925.

Manos, P., P. Soltis, D. Soltis, S. Manchester, S. Oh, C. Bell, D. Dilcher, and D. Stone. 2007. Phylogeny of extant and fossil juglandaceae inferred from the integration of molecular and morphological data sets. Systematic Biology 56:412–430.

Meredith, R., J. Janečka, J. Gatesy, O. Ryder, C. Fisher, E. Teeling, A. Goodbla, E. Eizirik, T. L. Simão, T. Stadler, D. Rabosky, R. Honeycutt, J. Flynn, C. Ingram, C. Steiner, T. Williams, T. Robinson, B. Angela, M. Westerman, N. Ayoub, M. Springer, and W. Murphy. 2011. Impacts of the cretaceous terrestrial revolution and KPg extinction on mammal diversification. Science 334:521–524.

Ni, X., D. Gebo, M. Dagosto, J. Meng, P. Tafforeau, J. Flynn, and K. Beard. 2013. The oldest known primate skeleton and early haplorhine evolution. Nature 498:60–64.

Novacek, M. J. and Q. Wheeler. 1992. Extinction and phylogeny. Columbia University Press.

O'Leary, M. A., J. I. Bloch, J. J. Flynn, T. J. Gaudin, A. Giallombardo, N. P. Giannini, S. L. Goldberg, B. P. Kraatz, Z.-X. Luo, J. Meng, X. Ni, M. J. Novacek, F. A. Perini, Z. S. Randall, G. W. Rougier, E. J. Sargis, M. T. Silcox, N. B. Simmons, M. Spaulding, P. M.

Velazco, M. Weksler, J. R. Wible, and A. L. Cirranello. 2013. The placental mammal ancestor and the postk-pg radiation of placentals. Science 339:662–667.

OLeary, M. A. and S. Kaufman. 2011. Morphobank: phylophenomics in the "cloud". Cladistics 27:529–537.

Pagel, M. 1994. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. Proceedings of the Royal Society of London. Series B: Biological Sciences 255:37–45.

Paradis, E. 2011. Time-dependent speciation and extinction from phylogenies: a least squares approach. Evolution 65:661–672.

Paradis, E., J. Claude, and K. Strimmer. 2004. APE: analyses of phylogenetics and evolution in r language. Bioinformatics 20:289–290.

Pattengale, N. D., M. Alipour, O. R. Bininda-Emonds, B. M. Moret, and A. Stamatakis. 2010. How many bootstrap replicates are necessary? Journal of Computational Biology 17:337–354.

Pattinson, D. J., R. S. Thompson, A. K. Piotrowski, and R. J. Asher. 2014. Phylogeny, paleontology, and primates: do incomplete fossils bias the tree of life? Systematic biology .

Pearman, P., A. Guisan, O. Broennimann, and C. Randin. 2008. Niche dynamics in space and time. Trends in Ecology & Evolution 23:149–158.

Pyron, R. 2011. Divergence time estimation using fossils as terminal taxa and the origins of lissamphibia. Systematic Biology 60:466–481.

Quental, T. and C. Marshall. 2010. Diversity dynamics: molecular phylogenies need the fossil record. Trends in Ecology & Evolution 25:434–441.

R Core Team. 2014. R: a language and environment for statistical computing. R Foundation for Statistical Computing Vienna, Austria.

Rambaut, A. and N. C. Grassly. 1997. Seq-gen: an application for the monte carlo simulation of dna sequence evolution along phylogenetic trees. Computer Application in the Biosciences 13:235–8.

Raup, D. 1993. Extinction: bad genes or bad luck? Oxford University Press.

Robinson, D. F. and L. R. Foulds. 1981. Comparison of phylogenetic trees. Mathematical Biosciences 53:131–147.

Ronquist, F., S. Klopfstein, L. Vilhelmsen, S. Schulmeister, D. Murray, and A. Rasnitsyn. 2012a. A total-evidence approach to dating with fossils, applied to the early radiation of the hymenoptera. Systematic Biology 61:973–999.

Ronquist, F., M. Teslenko, P. van der Mark, D. L. Ayres, A. Darling, S. Hohna, B. Larget, L. Liu, M. A. Suchard, and J. P. Huelsenbeck. 2012b. Mrbayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. Systematic Biology 61:539–42.

Roure, B. and H. Philippe. 2011. Site-specific time heterogeneity of the substitution process and its impact on phylogenetic inference. BMC Evolutionary Biology 11.

Rozen, D. E., D. Schneider, and R. E. Lenski. 2005. Long-term experimental evolution in escherichia coli. xiii. phylogenetic history of a balanced polymorphism. Journal of Molecular Evolution 61:171–80.

Ruxton, G. D. and G. Beauchamp. 2008. Time for some a priori thinking about post hoc testing. Behavioral Ecology 19.

Salamin, N., M. W. Chase, T. R. Hodkinson, and V. Savolainen. 2003. Assessing internal support with large phylogenetic dna matrices. Molecular Phylogenetics and Evolution 27:528–539.

Sansom, R. and M. Wills. 2013. Fossilization causes organisms to appear erroneously primitive by distorting evolutionary trees. Scientific Reports 3.

Schrago, C., B. Mello, and A. Soares. 2013. Combining fossil and molecular data to date the diversification of new world primates. Journal of Evolutionary Biology 26:2438–2446.

Simpson, G. G. 1945. Tempo and mode in evolution. Transactions of the New York Academy of Sciences 8:45–60.

Slater, G. J. and L. J. Harmon. 2013. Unifying fossils and phylogenies for comparative analyses of diversification and trait evolution. Methods in Ecology and Evolution 4.

Spencer, M. R. and E. W. Wilberg. 2013. Efficacy or convenience? model-based approaches to phylogeny estimation using morphological data. Cladistics 29.

Springer, M., R. Meredith, J. Gatesy, C. Emerling, J. Park, D. Rabosky, T. Stadler, C. Steiner, O. Ryder, J. Janeka, C. Fisher, and W. Murphy. 2012. Macroevolutionary dynamics and historical biogeography of primate diversification inferred from a species supermatrix. PLOS ONE 7.

Stadler, T. and Z. Yang. 2013. Dating phylogenies with sequentially sampled tips. Systematic Biology .

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics .

Tavaré, S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences vol. 17 of *Some Mathematical Questions in Biology*. American Mathematical Society.

Wagner, P. J. 2000. Exhaustion of morphologic character states among fossil taxa. Evolution 54:365–386.

Wiens, J. 2006. Missing data and the design of phylogenetic analyses. Journal of Biomedical Informatics 39:34–42.

Wiens, J. J. 2003. Missing data, incomplete taxa, and phylogenetic accuracy. Systematic Biology 52.

Wiens, J. J. and D. S. Moen. 2008. Missing data and the accuracy of bayesian phylogenetics. Journal of Systematic Evolution 46:307–314.

with contributions from Jochen Einbeck, R. J. H. and M. Wand. 2013. hdrcde: Highest density regions and conditional density estimation. R package version 3.1.

Yang, Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends in Ecology & Evolution 11:367–372.

Zander, R. 2004. Minimal values for reliability of bootstrap and jackknife proportions, decay index, and bayesian posterior probability. Phyloinformatics 2:1–13.

Zuckerkandl, E. and L. Pauling. 1965. Molecules as documents of evolutionary history. Journal of Theoretical Biology 8:357–366.

Supplementary Material

Supplementary material Section 1

Morphological characters states

In order to obtain a realistic probabilistic value for of k characters states for each simulated morphological character, we downloaded 100 random morphological characters (with more than 100 characters each) from TreeBASE database (http://treebase.org/) published between 1985 and 2013 and covering 19 taxomomic classess (Chordata, Arthropoda, Annelida, Angiosperm, Gymnosperm and Pteridophyta). We selected a total of 22563 characters ranging from 2 to 10 states. We calculated the proportion of characters with 2, 3, 4, 5, 6, 7, 8, 9 or 10 states. We then sampeled 22563 k values between 2 and 10 with the same proportion of characters from the empirical data. We then used a simple t-test to check if our simulation was equal to the empirical data. In this study, we only simulated characters with 2 or 3 states because of the high proportion of ordered characters ecountered on characters with more than 3 states and the difficulties of simulate biologicaly sensible ordered characters.

Tree Building Software settings

Maximum Likelihood - RAxML v8.0.20 (Stamatakis 2014).—

Model:

Molecular data:

GTR + Γ_4 (-m GTRGAMMA)

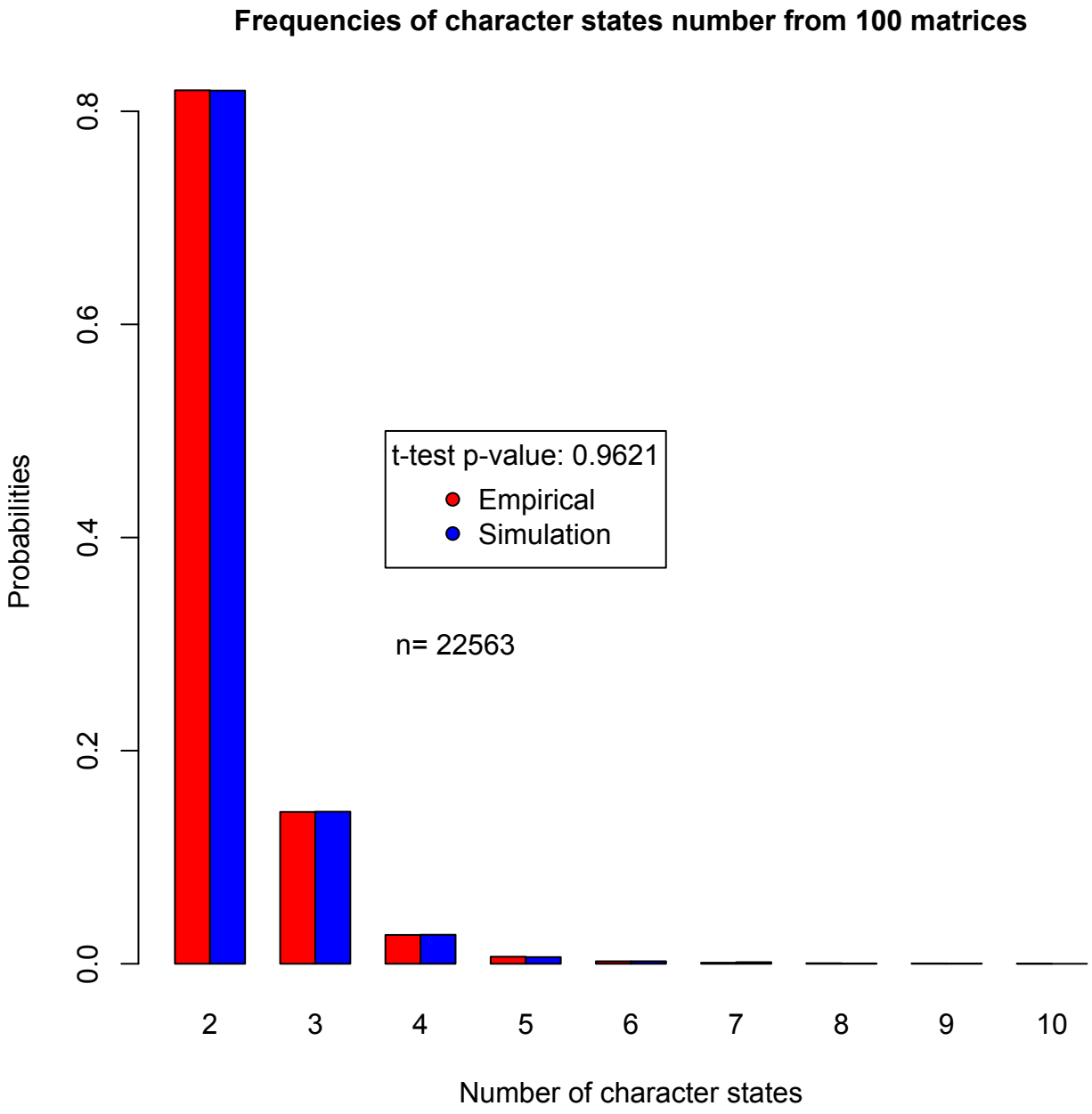

Figure 5: Character states distribution in empirical matrices. Characters states number distribution extracted from 100 random morphological matrices downloaded from RreeBase.

Morphological data:

Mk + Γ_4 (-K MK)

Support:

Rapid Boostrap algorithm (LSR), 1000 replicates

Bayesian - MrBayes v3.0.2 (Ronquist et al. 2012b).—

Priors:

Molecular data:

rates distribution shape (α) = 0.5

Transition/Transversion ratio = 2 (β(80,40))

Starting tree: "True" tree topology with each branch length = 1

Morphological data:

rates distribution shape (α) = 0.5

Models:

Molecular data: HKY + Γ_4

Morphological data: Mk + Γ_4

MCMC:

2 runs

4 chains per run

generations ¡ 50×1^6

sample frequency = 1050×1^3

ASDS diagnosis frequency = 50×1^3

ASDS $<$ 0.01

ESS $>>$ 200

Burnin = 25%

Tree Comparisons

Supplementary material Section 2

Triplets distance details ($T_{x,y}$)

Triples distance ($T_{x,y}$; Dobson 1975) measures the number of sub-trees made up of three taxa (triplets) that differ between two given trees. Each triplet can be written as I_{ijk}=(ijk). Where I_{ijk} is equal to zero if the the two triplets (ijk) are the same in the two trees otherwise I_{ijk} is equal to one. For any rooted binary tree there are only three possible combinations for each triplet: ((j,k),i);, ((i,k),j); and ((i,j),k); (Johnson and Soltis 1998). If the trees used are not fully binary, a fourth triplet combination is possible: (i,j,k). We can calculate the triplet distance between two trees, S_n, as:

$$S_n = \sum_{ijk} I_{ijk} \tag{1}$$

Where:

$$\sum_{ijk} = \binom{n}{4} = \frac{n!}{4!(n-4)!} \tag{2}$$

And where n is the total number of taxa in both trees (modified from Critchlow et al. (1996)). If S_n = 0, the trees are identical; when $S_n = \binom{n}{4}$, the trees are as different as possible (i.e. every taxon has a different placement in the two trees). Because the possible number of triplets per clade is a finite number, the probability of two random trees with the same n taxa to have the same triplet is:

$$P(I_{ijk} = 0) = \frac{1}{4} \tag{3}$$

Therefore one can calculate the probability of two random trees having the same triplets:

$$P(S_n = 0) = \sum_{ijk} P_{I_{ijk}=0} \tag{4}$$

$$P(S_n = 0) = \frac{n!}{4(3!(n-3)!} \tag{5}$$

And in the same way:

$$P(S_n = 1) = \frac{3n!}{4(3!(n-3)!} \tag{6}$$

Robinson-Foulds distance details

Robinson-Foulds distance (*RF*; Robinson and Foulds 1981), or "path difference" , measures the number of shared clades across two trees. The metric reflects the distance between the distributions of tips among clades in the two trees (Robinson and Foulds 1981) and can be expressed as following:

$$RF_{x,y} = N_x + N_y - 2C_{x,y} \tag{7}$$

Where $C_{x,y}$ is the number of clades in common in the two trees. C is one if the two trees have the same n taxa; the maximal value is $C = n - 2$. This metric is more sensitive to taxon displacement than Triples distance (i.e. if one taxon moves out of a clade, then the clades are no longer considered similar; Critchlow et al. (1996); Johnson and Soltis (1998); Wiens (2003)). The minimal value of C is equal to 1 if the two trees have the same n taxa; the maximal value in $C = n - 2$. For a fully unresolved tree (star tree) N=1 and for a fully resolved tree (binary tree) $N = n - 2$. The minimal and maximal topological distance for taxa is:

$$RF_{min} = 1 + 1 - 2C_{x,y} \tag{8}$$

And:

$$RF_{max} = 2(n-2) - 2 \tag{9}$$

One can then rescale *RF.scaled* by using the maximal and minimal value for any n taxa:

$$RF.scaled_{x,y} = \frac{RF_{x,y} - RF_{max}}{RF_{max}} \tag{10}$$

This metric is more sensitive to taxa displacement than the Triplet distance (Critchlow et al. 1996; Johnson and Soltis 1998; Wiens 2003) and therefore a low value will show a good clade conservation between two trees and a high value will show a bad recovery of common clades.

Normalised Tree Similarity

For any tree with n taxa compared using a tree distance metric m, Normalized Tree Similarity, NTS_m (Bogdanowicz et al. 2012), represents the similarity score for the two trees given the expected distance between two random Yule trees with n taxa. If $\bar{d}_{m,n}(rand)$ is the average distance between two random Yule trees with n taxa and $d_{m,n}(x,y)$ the distance between the two trees x and y containing each n taxa, then:

$$NTS_{m,n}(x,y) = \frac{\bar{d}_{m,n}(rand) - d_{m,n}(x,y)}{\bar{d}_{m,n}(rand)} \tag{11}$$

NTS ranges from one to -∞. For any m, n, when $NTS = 1$, the trees are identical, when $NTS = 0$ the trees are no more different than expected by chance, and when $NTS < 0$, the trees are more different than expected when comparing two random trees.

Tree comparisons

Random tree comparison scaling.— We used the comparison of 1000 random trees to obtain the mean comparison value $\bar{d}_{m,n}(rand)$ for the NTS metric. We randomly generated two sets of 1000 trees of n taxa using the rmtree function of ape package (v3.0-11 Paradis et al. (2004)) that generates a given number of random Yule trees. We

calculated the $\bar{d}_{m,n}$*(rand)* value using an approach similar to the RPCBTC (described below) by performing 1000 random pairwise comparisons using the TreeCmp java script (Bogdanowicz et al. 2012).

Random Pairwise Bayesian Tree Comparison (RPBTC).— We assessed the power of the Random Pairwise Bayesian Tree Comparison (RPBTC) method by comparing 1000 random trees from a posterior distribution trees set to another 1000 random trees from the same posterior distribution trees set. We repeated this 100 times independently using the same posterior distribution trees set each time resulting in 100 replicates of the same posterior distribution trees set compared 1000 times. We used an ANOVA to test if there was no significant difference between the replicates so that the RBTC can be replicated. We applied this protocol on a poorly resolved tree (Low Score), a resolved tree with low support value (Medium Score) and a resolved tree with high support values (High Score). Results are available in table 1.

Table 1: Group comparison results: difference between 100 replicates using the RPBTC method

Tree.Type	Used.metric	Replicates	Df	F.value	p.value
Low Score	RF	100.00	99.00	0.74	0.98
Low Score	Tr	100.00	99.00	0.97	0.58
Medium Score	RF	100.00	99.00	0.64	1.00
Medium Score	Tr	100.00	99.00	0.45	1.00
High Score	RF	100.00	99.00	0.20	1.00
High Score	Tr	100.00	99.00	0.37	1.00

Codes

All codes are available at: *https : //github.com/TGuillerme/Total_Evidence_Method − Missing_data/tree/master/Functions*

The tree comparison results analysis can be repeated for more details at: *https : //github.com/TGuillerme/Total_Evidence_Method − Missing_data/tree/master/Analysis*

Full results

Bootstraps distribution

Table 2: Tree similarity values per parameter in ML framework

Parameter	amount of missing data	metric	mode	50%CI lower	50%CI upper	95%CI lower	95%CI upper
M_L	0%	Robinson-Fould	1	NA	NA	NA	NA
		Triples	1	NA	NA	NA	NA
	10%	Robinson-Fould	0.6714878	0.5692308	0.7894649	0.4689069	1.03615
		Triples	0.9424133	0.8683742	0.9998559	0.6121399	1.06352
	25%	Robinson-Fould	0.5167595	0.4403147	0.6307692	0.2664429	0.89291
		Triples	0.5599494	0.7084874	0.8105781	0.2912950	0.99907
	50%	Robinson-Fould	0.4321852	0.3641026	0.5215627	0.2313246	0.64238
		Triples	0.5143009	0.4197127	0.5991498	0.2991432	0.91045
	75%	Robinson-Fould	0.3727090	0.3230769	0.4342249	0.2035474	0.57762
		Triples	0.4449563	0.3385128	0.5329149	0.1929395	0.85515
M_L	0%	Robinson-Fould	1	NA	NA	NA	NA
		Triples	1	NA	NA	NA	NA
	10%	Robinson-Fould	0.9072419	0.7668415	1.0000000	0.4070844	1.04780
		Triples	0.9763617	0.9448100	1.0039460	0.7158118	1.10433
	25%	Robinson-Fould	0.5722382	0.4883560	0.6923077	0.3253239	0.93674
		Triples	0.9370152	0.8207407	0.9938846	0.5394443	1.08680
	50%	Robinson-Fould	0.4338159	0.3025641	0.4760777	0.1704016	0.66832
		Triples	0.7327464	0.5947211	0.8739630	0.3418679	1.02515
	75%	Robinson-Fould	0.1534487	0.1155128	0.2000000	0.0683485	0.32307
		Triples	0.4588292	0.3905631	0.6111484	0.2903676	0.85682
M_L	0%	Robinson-Fould	1	NA	NA	NA	NA
		Triples	1	NA	NA	NA	NA
	10%	Robinson-Fould	0.473285	0.3842965	0.8358974	0.3564753	0.95897
		Triples 44	0.9592557	0.8691580	1.026095	0.5720847	1.06752
	25%	Robinson-Fould	0.4548038	0.3572132	0.6307692	0.2425339	0.91612
		Triples	0.9071925	0.7516934	0.9976224	0.4462168	1.07052
	50%	Robinson-Fould	0.4344262	0.2755493	0.4666667	0.2404937	0.83873

Table 3: Tree similarity values per parameter in Bayesian framework

Parameter	amount of missing data	mode	50%CI lower	50%CI upper	95%CI lower	95%CI upper
M_L	0%	Robinson-Fould	0.05641026	0.03589744	0.06337695	0.0313920
		Triples	0.04104231	0.00830962	0.05441173	-0.039333
	10%	Robinson-Fould	0.03584309	03584196	03589744	0282564
		Triples	0.03399849	0.003196469	0.04981018	-0.0499133
	25%	Robinson-Fould	0.03589744	0.03589744	0.03589744	0.0153673
		Triples	0.03510953	0.002113358	0.04939228	-0.0412022
	50%	Robinson-Fould	0.03589744	0.03589744	0.03589744	0.0153846
		Triples	0.02849433	0.003263098	0.05298036	-0.0383447
	75%	Robinson-Fould	0.03589744	0.03589744	0.03589744	0.0106999
		Triples	-0.0001940566	-0.01801301	0.04058783	-0.060309
M_L	0%	Robinson-Fould	0.05641026	0.03589744	0.06337695	0.0313920
		Triples	0.04104231	0.008309622	0.05441173	-0.0393337
	10%	Robinson-Fould	0.03593712	0.03198256	0.05641026	0.0279886
		Triples	0.02639309	0.002185408	0.05061314	-0.0465630
	25%	Robinson-Fould	0.03588134	0.03587629	0.03589744	0.0153846
		Triples	0.03856296	0.002185408	0.05061314	-0.0465630
	50%	Robinson-Fould	0.03589744	0.03589744	0.03589744	0.0153628
		Triples	0.01403716	-0.01100608	0.03972324	-0.038003
	75%	Robinson-Fould	0.03589206	0.03588825	0.03589744	0.0153846
		Triples	0.02939363	0.01902351	0.04238907	-0.030879
M_L	0%	Robinson-Fould	0.05641026	0.03589744	0.06337695	0.0313920
		Triples	0.04104231	0.008309622	0.05441173	-0.0393337
	10%	Robinson-Fould	0.05637057	0.03589744	0.05927956	0.0281329
		Triples 45	0.03008132	0.01176800	0.04847400	-0.0245140
	25%	Robinson-Fould	0.03593743	0.03589744	0.03596018	0.0233079
		Triples	0.02010221	-0.001145118	0.0377779	-0.0601833
	50%	Robinson-Fould	0.03589744	0.03589744	0.03589744	0.0305038

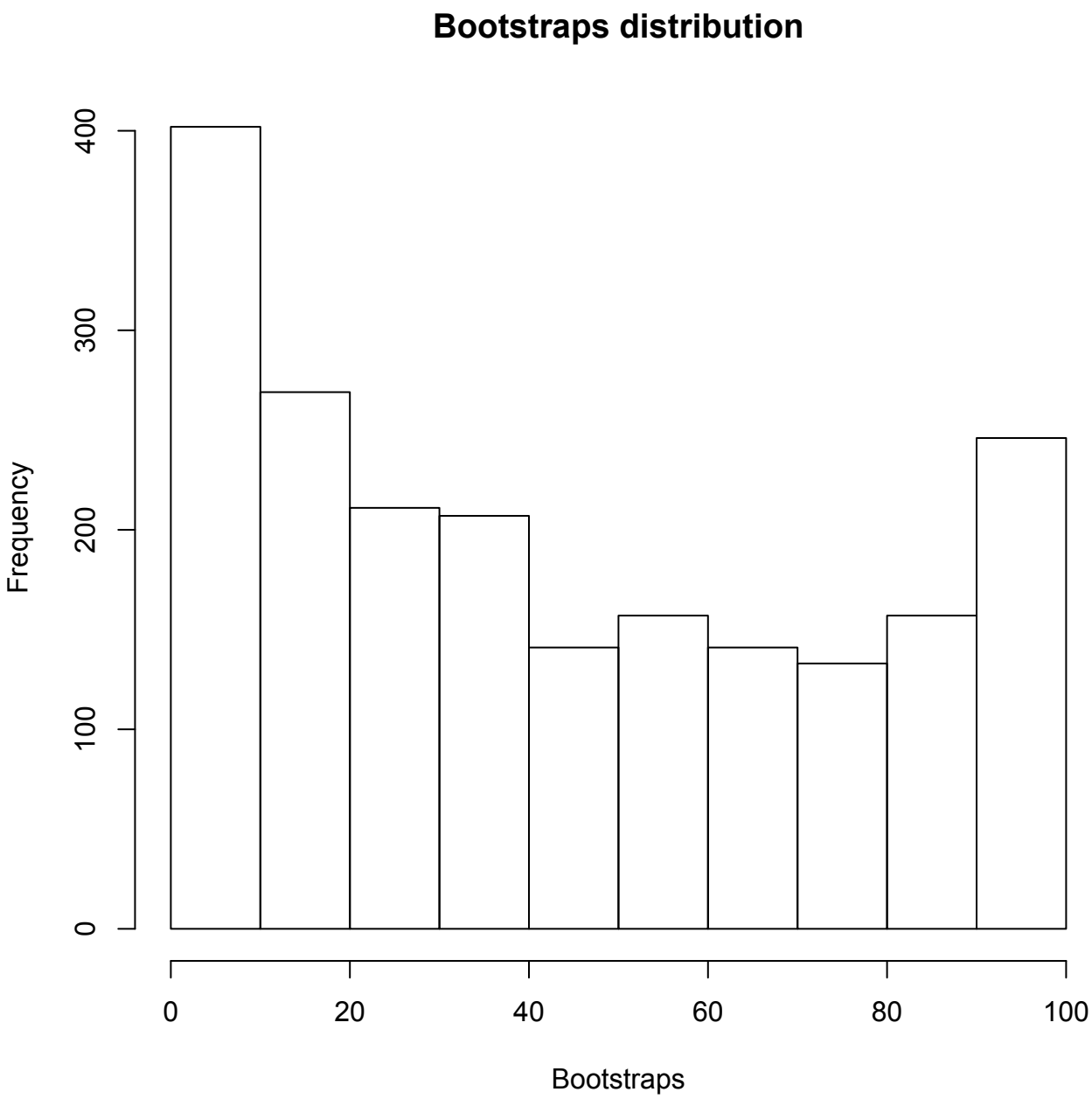

Figure 6: Bootstraps distribution across the ″best″ trees.