Version dated: December 2, 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 called Total Evidence, 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 consensus trees recovers the right topology better than Maximum Likelihood or than a subset of Bayesian posterior tree distributions and this regardless of the amount of missing data (minimal tree similarity of 0.69 in the worst scenario). Regarding the data, good topological recovery is not related to the amount of missing data *per se* but to the amount of data overlap. The two factors that influences the most this data overlap are the parameters 1 (proportion of coded living taxa) and 3 (number of morphological characters).

(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. 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; Wright and Hillis 2014). 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 (e.g. Pyron 2011; Ronquist et al. 2012a; Schrago et al. 2013; Slater 2013; Beck and Lee 2014), 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; Sansom 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 (50) of highly-covered genes (i.e. a number of genes that are available for all taxa ; 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 (Kearney 2002; Wiens 2003; Pattinson et al. 2014).

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; Pattinson et al. 2014), 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 are interested in assessing the effect of missing data on the ability to recover a correct topology. Even though branch length is a second aspect of phylogenetic inferences equally important, 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 a simulation approach to assess the effect of missing data on tree topologies inferred from Total Evidence matrices. Since 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 (e.g. O'Leary et al. 2013).

We find that the ability of recovering the correct topology can mainly be increased when using the Bayesian methods. Also minimizing the proportion of living taxa with no morphological data and the proportion of missing morphological characters improves the ability to recover the correct more than minimizing the proportion of missing data in the fossil record.

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

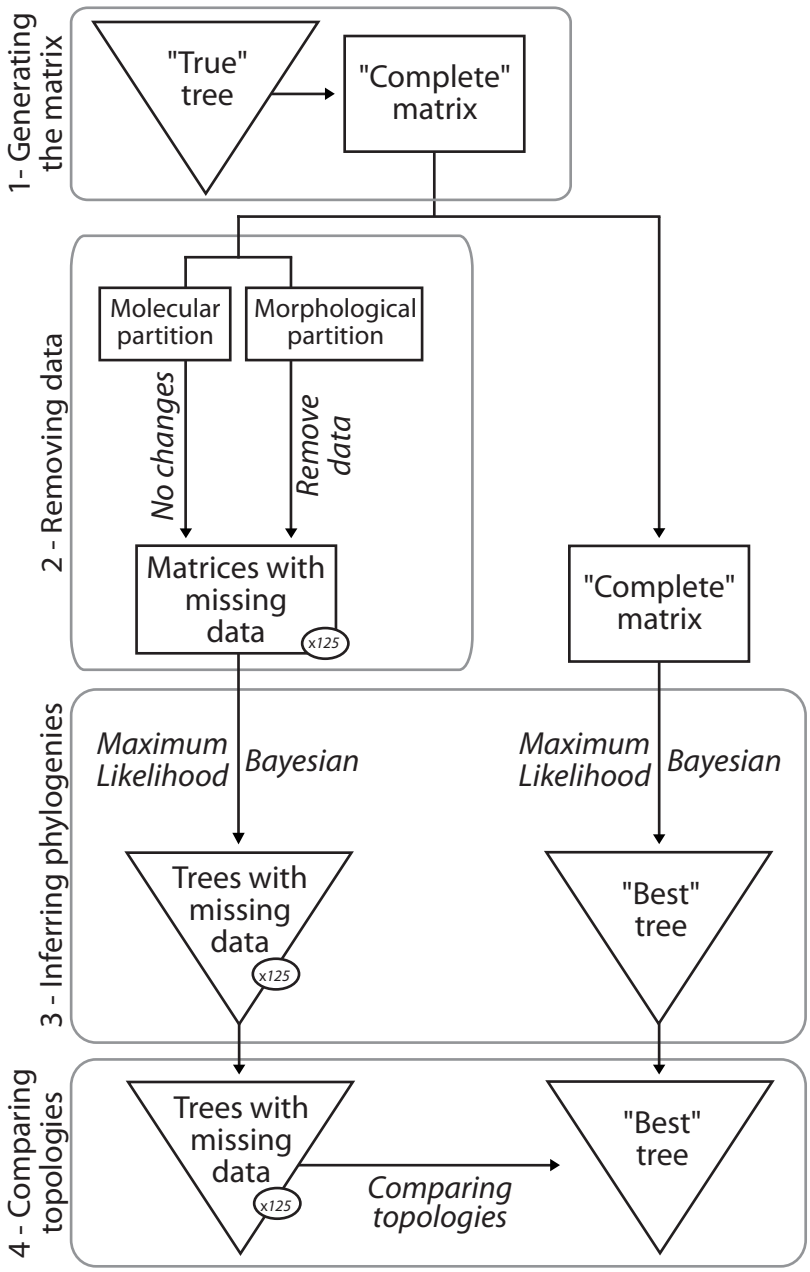

Figure 1: Protocol outline. (1) We randomly generated a birth-death tree (the ″true″ tree) and used it to infer a matrix with no missing data (the ″complete″ matrix). (2) We removed data from the morphological part of the ″complete″ matrix resulting in 125 ″missing-data″ matrices. (3) We built phylogenetic 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 Supplementary Material Section 1). 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 topology: 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. Note that both tree comparisons are representing different aspects of topological similarity and do not perform equaly (see discussion below and Kuhner and Yamato 2014).

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 Section 2). 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).

In order to investigate the effect of the parameters and/or the methods used in

this simulations, we measured the similarity among the different distributions using the Bhattacharyya Coefficient (Bhattacharyya 1943). The Bhattacharyya Coefficient is the probability of overlap between two distributions, ranging from 0 to 1 (Bhattacharyya 1943). For each method (Maximym likelihood trees, Bayesian consensus trees, Bootstraps and Bayesian posterior trees) and metric (Robinson-Foulds and Triplets distance), we used this coefficient in two ways in order to:

1. assess the effect of the method on all parameters:

   We performed a pairwise comparison of each parameter between each pair of methods. This resulted in 125 comparisons per method and per metric (each parameter combination within one method compared to the same parameter combination in the other method). We then compared the distribution of these pairwise comparisons to see the global difference between two methods: if the method have really similar results, we would expect the distribution to be cluster around 1 (high probability of overlap between the distributions) and if the methods are really dissimilar around 0 (low probability of overlap - Fig. 4).

2. assess the effect of the parameters within a method:

   We performed a pairwise comparison per method and per metric for each combination of parameters. This resulted in 7875 pairwise comparisons per method and per metric (triangle of a 125*125 matrix). We represented these results as a triangular matrix with the values of each pairwise comparison coloured according to the value of the Batthacharyya Coefficient (coloured in green when the distributions overlap completely and in red when they don't - Fig. 5).

Results

Effect of missing data on topology

As it would be expected form the literature, the amount of missing data in the morphological matrix does decrease the ability to recover the right topology, regardless of the parameter, the method or the metric used (e.g. Roure and Philippe 2011; Sansom and Wills 2013; Pattinson et al. 2014). However each variable does not affect the topology in the same way (Fig. 2). Regarding the conservation of entire clades (i.e. the Ronbinson-Foulds distance), the Bayesian consensus trees outperform the other methods and the amount of missing data in the living taxa (M_L) decreases the most rapidly the clade conservation when using this method. However, when looking at the position of wild card taxa (i.e. the Triplets distance), the Maximum Likelihood outperforms the other methods but with wider distributions overlap (Fig. 4-B).

When looking at the global trend of the Bayesian consensus trees for all the parameters combinations, this method outperforms Maximum Likelihood for the Ronbinson-Foulds distance and plateaus to a minimal modal tree similarity of 0.69 regardless the amount of missing data (Fig. 3 - see also Supplementary Material Section 3). However, regarding the Triplets distance, Bayesian consensus trees seems to perform poorly but with great confidence intervals overlaps (Fig. 3 and Fig. 5). These greater larger confidence intervals are due to the lower ability of the Triplet method to highlight precise differences in topology (Kuhner and Yamato 2014). Results for the global trend comparison between the uncertainty methods (Bootstrap and Bayesian posterior tree distribution) are available in the supplementary materials (see Supplementary Material Section 3).

Effect of the method on topology

When looking at the comparisons between the methods using the Robinson-Foulds distance, the Bayesian consensus trees have the least overlap with

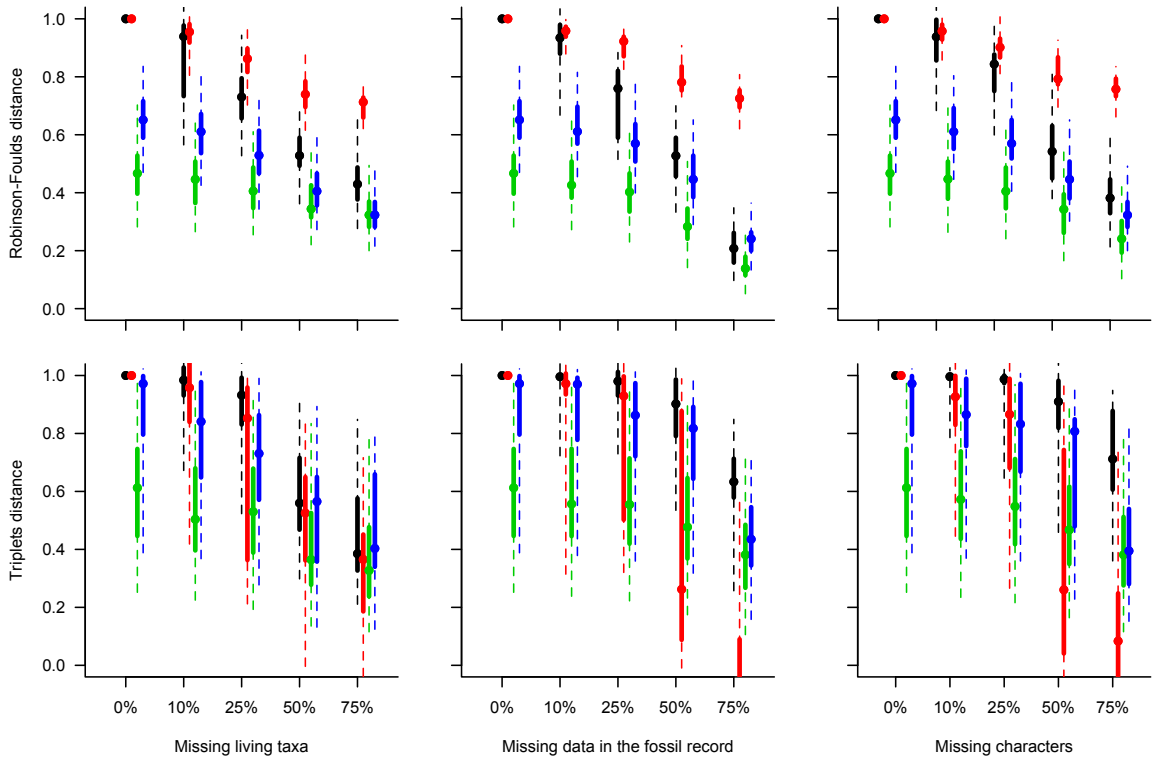

Figure 2: Comparison between the effect of missing data and the tree inference method on topology. The amount of missing data for each parameter is represented on the x axis. The topology is represented on the y axis, both using Robinson-Foulds distance (upper row) and Triplets distance (lower row). 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.

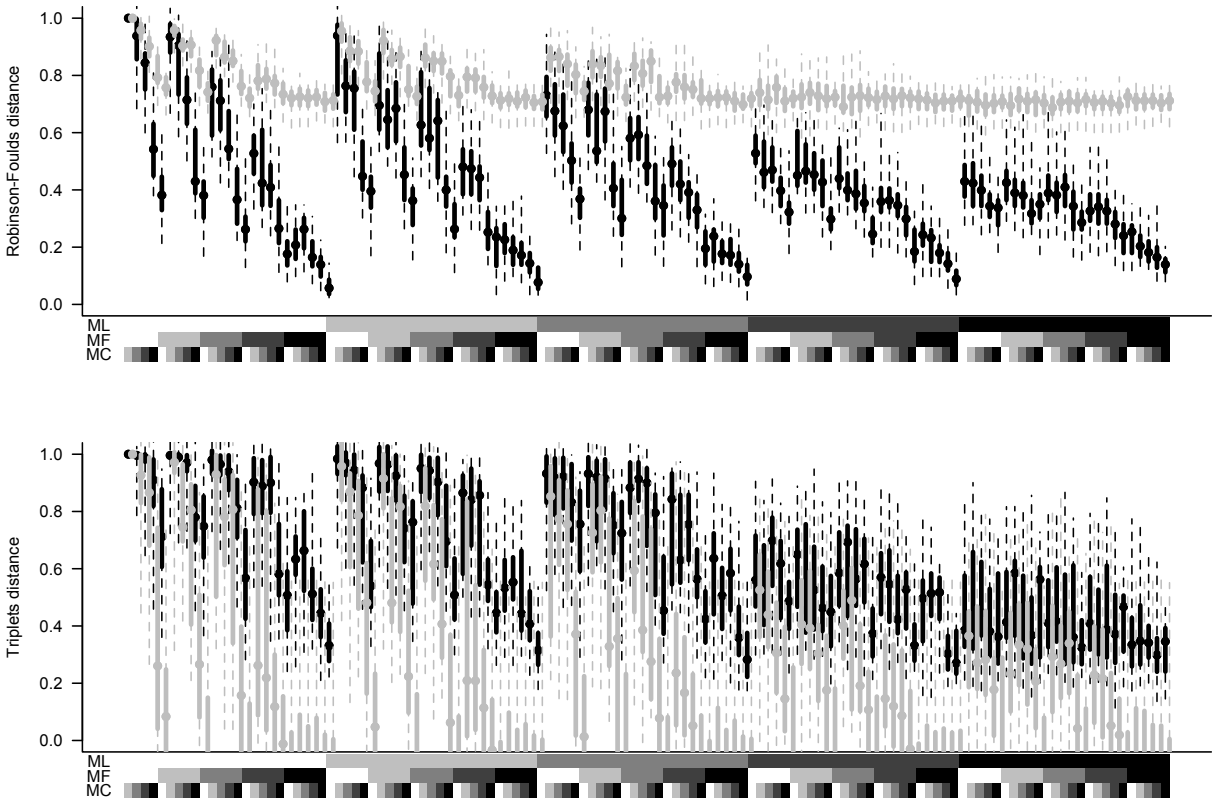

Figure 3: Trend of the effect of missing data on topology on ML and consensus trees. The amount of missing data per parameter (M_L, M_F and M_C) is represented along the x axis. The colour gradient from white to black represents respectively, 0%, 10%, 25%, 50% and 75% of missing data. The topology is 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.

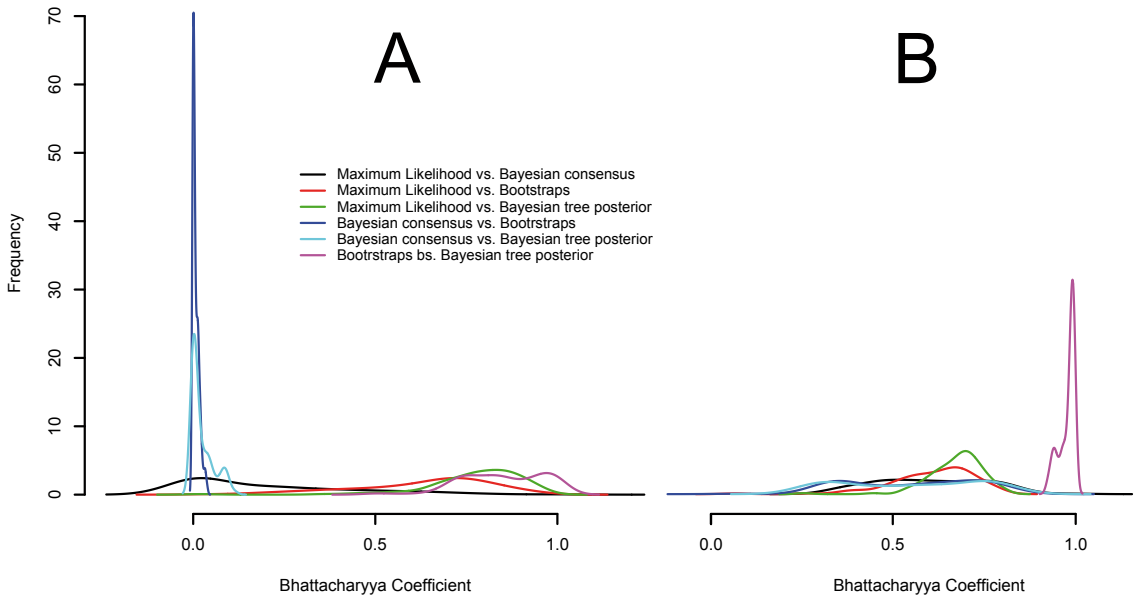

Figure 4: Distribution of the Bhattacharyya Coefficients between methods. Curves represents the kernel density estimations of the BC between the four different methods. A. Results for the Normalised Robinson-Foulds distance. B. Results with the for Normalised Triplets distance. Pikes around the two different extremes values of the BC show the two pairs of most dissimilar methods (Bayesian consensus vs. Bootstraps and Bayesian consensus vs. Bayesian posterior trees) for the Normalised Robinson-Foulds distance ; and the most similar methods (Bootstraps and Bayesian posterior trees) for the Normalised Triples methods.

respectively the Bayesian posterior trees and the Bootstrap (Fig. 4-A). When regarding at the Triplets distance, the two method that have the most distribution are the Bootstraps and the Bayesian posterior trees (Fig. 4-B). These results are due to the slight difference in comparing topologies between Maximum Likelihood and Bayesian consensus trees and Bootstraps and Bayesian posterior trees: the two last ones are based on a random pairwise comparison of a sub-sample of 1000 trees. This is likely to add noise to the data.

Effect of the missing data parameters on topology

When looking at the pairwise Bhattacharyya Coefficients between all the parameters combinations for Bayesian consensus trees, the number of missing living taxa (M_L) results the lowest probabilities of overlap (right lower corner in Fig. 5-A). However, when looking at the Triplets distance, because of the high overlap in confidence intervals, there is no major region with low probability of overlap. One should note however that the bottom line of the matrix where the Bhattacharyya Coefficients are very low is due to comparing the distribution of the pairwise comparison of the best tree vs. the best tree, leading to a Normalised triplet score of 1 every time (with no variance, see Fig. 2 and Fig. 3). All the other pairwise comparisons are available in the supplementary materials (see Supplementary Material Section 3).

Discussion

The results of our simulations are in adequacy with recent analysis on the effect of missing data in morphological matrices (Pattinson et al. 2014; Wright and Hillis 2014; Sansom 2014). Our results are showing that ability to recover the correct topology decreases with the amount of missing data, regardless the metric, the method or the

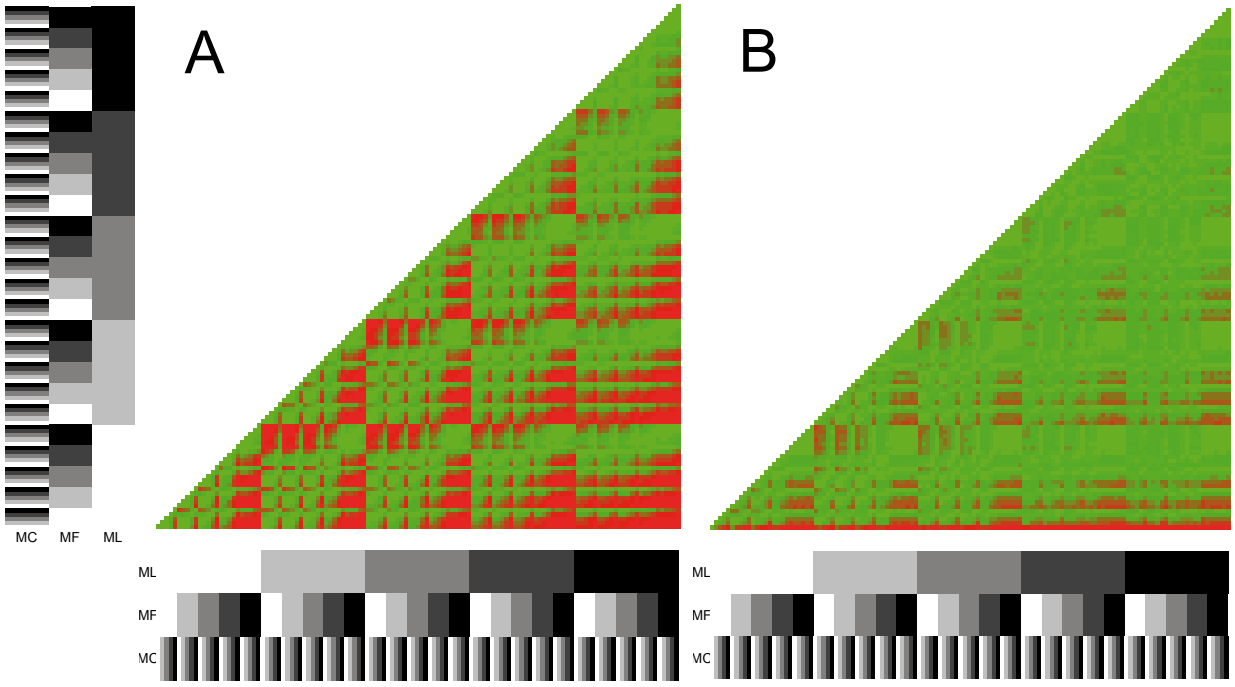

Figure 5: Pairwise Bhattacharyya Coefficients within the Bayesian consensus trees. The pairwise trees comparisons are represent on both axis. The colour gradient from white to black represents respectively, 0%, 10%, 25%, 50% and 75% of missing data. The matrix represents the values of pairwise Bhattacharyya Coefficients going from green (BC=1) to red (BC=0). A. Results for the Normalised Robinson-Foulds distance. B. Results with the for Normalised Triplets distance.

parameters considered. However, the effect of missing data on topological recovery varies with different methods to infer topology and different metrics to compare topologies. Additionally, within a method and a metric, the parameters used to remove data have different effect on the ability to recover the correct topology. Thus, we found that using the topologies of Bayesian consensus trees performed consistently better than other methods for recovering the correct topology (Fig. 3 and Fig. 5) and that both the amount of available characters (M_C) and the amount of data for the living taxa (M_L) are crucial missing data parameters.

As proposed in numerous previous studies, when recovering topology, the amount of missing data *per se* is not a problem as long as enough data overlaps in the matrix (e.g. Roure and Philippe 2011; Pattinson et al. 2014). To improve topological recovery in a Total Evidence framework, one can efficiently increase the amount of overlapping data by sampling a maximum of characters and by increasing the number of living data coded in the morphological part of the matrix.

For each missing data parameter used in this study, there is a different way to reduce the amount of missing data:

1. M_L: One should put more effort in using natural museums history collections for coding the missing morphological data for living taxa if possible

2. M_F: 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)

3. M_C: improvement can be done by vast collaborative projects in order to gather as much characters as possible (e.g. O'Leary et al. 2013). 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).

Our results are consistent with previous works and are showing that the amount of missing data *per se* does not reduce the ability to recover the correct topology (e.g. Kearney 2002; Wiens 2003; Pattinson et al. 2014). In fact, as shown previously, the amount of overlap in the matrix is more crucial (e.g. Roure and Philippe 2011). In our case, the overlap in data is greatly reduced when different parameters are combined with a high amount of missing data (e.g. M_L=75%, M_F=75% and M_C=75% - Fig. 2 and Fig. 3). In his scenario the overall number of missing characters M_C affects the proportion of overlapping data by reducing the overall size of the matrix (Fig. 3). However it is important to note that in the pairwise Bhattacharyya Coefficient analysis, the area with the lowest probability of distribution overlap is led by comparing the trees with no missing data for the living taxa to the trees with 75% missing data for the living data (Right lower corner in Fig. 5).

Following Kuhner and Yamato (2014)'s conclusions on the superiority of the Robinson-Foulds distance, the best method to use for recovering the correct topology in a Total Evidence framework is using the Bayesian consensus tree (Fig. 2 and Fig. 3). In fact, regardless of the amount of missing data, this method always displays the highest normalised Robinson-Foulds score (see Supplementary Material Section 3). When comparing this the probability of overlap in this method between the other onces, it is counter intuitive to note that the Bayesian consensus tree shows the highest dissimilarity with the Bayesian posterior trees from which the consensus tree is build (Fig. 5-A). Note that it is also highly dissimilar to the Bootstraps. However that can be linked to the fact that both Bootstraps andBayesian posterior trees are measured as 125*1000 random pairwise comparisons instead of just 125 tree comparisons for the Bayesian consensus trees. The fact that the 1000 trees to be compared for each of the 125 pairwise comparisons can add noise to the results for these two methods due to the

random factor in the pairwise comparisons (see Methods - Tree comparisons).

Our results can be contrasted by looking at the normalised tree similarity when using the Robinson-Foulds or the Triplets distance. When using the second one, the confidence interval of our different parameters combinations per methods seems higher (Fig. 2 and Fig. 3) and the probability of overlap between the distribution of each parameter is high (Bhattacharyya Coefficient close to 1 - Fig. 5-B). The contrast between the results of the two metrics can be explained by the actual nature of both metrics:

1. 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. This metric illustrates therefore the conservation of clades among the compared trees (i.e. when the normalised Robinson-Foulds distance is close to 1, most of the clades in the two trees are identical).

2. On the other hand, the Triplets method is measuring the position of each taxon towards two other reference taxa (Critchlow et al. 1996). It will penalise only trees where taxa get removed furthest from their original clade. This metric illustrates therefore the amount of wildcard taxa (Kearney 2002) between the to compared trees (i.e. when the score is closed to 1, view wildcard taxa are present in the tree).

Therefore, in our study, both metrics are showing two different aspects of tree topology. Also, disregarding their fundamental differences in accounting for topological distance, Kuhner and Yamato (2014) have recently proposed that the Ronbinson-Foulds distance greatly outperforms the Triples distance in signalling differences in topology.

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). 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 (Wright and Hillis 2014).

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).

One way to address this issue would be to run a similar analysis using real data. The amount of studies using Total Evidence data type matrices is steadily increasing (e.g. Ronquist et al. 2012a; Slater 2013; Beck and Lee 2014) and one could analysis these total evidence matrices in the same framework as this study. However, the true topology (obtained from a matrix with no missing data) is unlikely to be known. A second way to address this issue, and especially to incorporate the variation of signal in morphological character would be to simulate morphological at different rates (Wright and Hillis 2014) or to apply patterns of missing data directly obtained from the fossil record (Pattinson et al. 2014). Nevertheless, our simulation, by simplifying an ideal world, already show clear results of the effect of missing data on topology in a Total Evidence framework. Also, our results show how recovering the correct topology can be improved by using Bayesian consensus trees in combination with a maximum amount of characters and a of living taxa with morphological data.

Conclusion

Using Total Evidence matrices with missing data is not a problem for recovering the correct topology as long as enough morphological data is overlapping between the fossil and living taxa. In order to optimize good topological recovery, with propose to use the Bayesian consensus tree over the Maximum Likelihood tree or a subset of the Bayesian posterior trees distribution. Also, in order to improve the overlap in data in the morphological part of the matrix, we advice to use all the morphological characters available and to code them for a maximum of living taxa present in the matrix.

Supplementary material

Supplementary material (code, analysis and full results) can be found in the Dryad data repository at http://dx.doi.org/10.5061/dryad.XXXX.

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

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

Beck, R. M. and M. S. Lee. 2014. Ancient dates or accelerated rates? morphological clocks and the antiquity of placental mammals. Proceedings. Biological sciences / The Royal Society 281.

Bhattacharyya, A. 1943. On a measure of divergence between two statistical populations defined by their probability distributions. Bull. Calcutta Math. Soc. 35:99–109.

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.

Kearney, M. 2002. Fragmentary taxa, missing data, and ambiguity: mistaken assumptions and conclusions. Systematic biology 51:369–381.

Kuhner, M. and J. Yamato. 2014. Practical performance of tree comparison metrics .

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.

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.

Sansom, R. S. 2014. Bias and sensitivity in the placement of fossil taxa resulting from interpretations of missing data. Systematic Biology .

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. 2013. Phylogenetic evidence for a shift in the mode of mammalian body size

evolution at the Cretaceous-Palaeogene boundary. Methods in Ecology and Evolution 4.

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.

Wright, A. M. and D. M. Hillis. 2014. Bayesian analysis using a simple likelihood model outperforms parsimony for estimation of phylogeny from discrete morphological data. PLoS ONE 9.

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.