TEM simulations paper

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

[1]TCD

[2]tcbr

[*]Corresponding author

Abstract

Abstract

Introduction

Living species represent less than 1% of all species that have ever lived **??**. However, the majority of macroevolutionary studies focus solely on living species (e.g. **??**). Ignoring fossil taxa may lead to misinterpretation of macroevolutionary patterns and processes such as species richness gradients, e.g. extant clade diversity may ignore past diversification patterns (e.g. **?**); biogeographical history, e.g. extant biogeographical patterns are improved when integrating fossil data (**?**, e.g.[]); or paleoecology, e.g. extant clades niches might have differed greatly through time (e.g. **?**). These factors have led to increasing consensus among scientists that fossil taxa must be included

in macroevolutionary studies (**?????**). However, to do this we need to be able to place living and fossil taxa into the same phylogenies which still remains complex. Three main approaches have been used for combining fossil and living taxa data in phylogenies. These approaches differ in whether they treat fossil taxa as nodes or tips and how much of the available data is used (i.e. age only or age and morphology). Classical cladistic methods use morphological data from both fossil and living taxa and treat each taxon as a tip **?**. This approach is commonly used by paleontologists but it ignores the majority of the additional molecular data available from living species. Neontologists, on the other hand, more commonly use only molecular data from living species to build phylogenies. Because fossil taxa do not usually have available DNA, fossils are used as nodes rather than tips in these analyses and only their occurrence age is used to time calibrate phylogenies (**?**). There have been great improvements in the theory and application of these two approaches (e.g. **???**) as well as much debate about the best approach to use (e.g. **?**). A final class of methods, known as total evidence methods use molecular data from living taxa and morphological data from both living and fossil taxa, and treats every taxon as a tip on the phylogeny (**?**). Here we focus on these total evidence methods because they have been recently successfully developed and applied to empirical data (**???**). Although the total evidence approach seems very promising, there is one big drawback in using this approach: it requires a lot of data. In particular it requires morphological data form both living and fossil taxa, both of which are scarce. Therefore total evidence approaches are likely to suffer from aving lots of missing data which may affect their ability to infered correct phyogenies. The effect of missing data on phylogenetic

inferences has been widely studied (??????). Missing molecular data has been seen by some authors as an issue because it can decrease the phylogenetic signal in some parts of the tree, especially when using large supermatrices (?). However other authors do not see missing molecular data as a major issue because the phylogenetic signal is more likely to increase by having at least a modest number of highly covered genes (50% - ?), a higher number of taxa (especially slowly evolving taxa or taxa close to the outgroup) and by choosing more adequate models of sequence evolution rather than by reducing the amount of missing data ???. Similarly, missing morphological data might be seen as major or minor issue for accurately infering phylogenies (??). Because soft-tissues characters are rarely preserved in the fossil record, missing data is found in soft tissues, i.e. it is not randomly distributed, which can lead to biased placement of fossil taxa in phylogenies (?). However, the phylogenetic signal is not related to the level of missing data per se but to the number of informative characters per taxa, therefore missing data is less an issue than the number of shared informative characters (?). Although not a major problem separately (????), missing data in molecular and morphological supermatrices may become an issue when combined for example in a total evidence type supermatrices and no attempt has been made to study the impact of this issue until now.

Here we assess the effect of missing data on tree topology infered from total evidence supermatrices. The molecular part of a total evidence supermatrix contains no fossil taxa so will act like a classical molecular supermatrix (??). The effect of missing data on such matrices is well known, therefore, we only focus on the missing data issue in the morphological part of the supermatrix.

Using simulations we investigate three major factors that directly affect the completeness of the morphological part of the supermatrix: (1) the proportion of living taxa with no morphological data, (2) the amount of missing data in the fossil taxa (i.e. the preservation quality of the fossil record) and (3) the overall number of morphological characters for both living and fossil taxa in the supermatrix. We assess how changing the values of these three parameters affects the topology of total evidence method phylogenies.

Materials and Methods

To explore the effect of missing data on total evidence trees topologies we used the following protocol (note that we explain each step in detail below this general outline - Fig. **??**)

1. Generating the complete matrix We built a randomly generated birth death tree to infer a complete matrix containing both molecular and morphological data for living and fossil taxa.

2. Removing data We removed data from the complete matrix to simulate the effects of missing data by modifying three parameters (1) the proportion of missing living taxa (M_L), (2) the proportion of missing data in the fossil taxa (M_F) and (3) the proportion of missing morphological characters (M_C).

3. Building phylogenies We infered Bayesian phylogenetic trees from the complete matrix and from the matrices containing missing data.

4. Comparing topologies We then compared the trees obtained from the

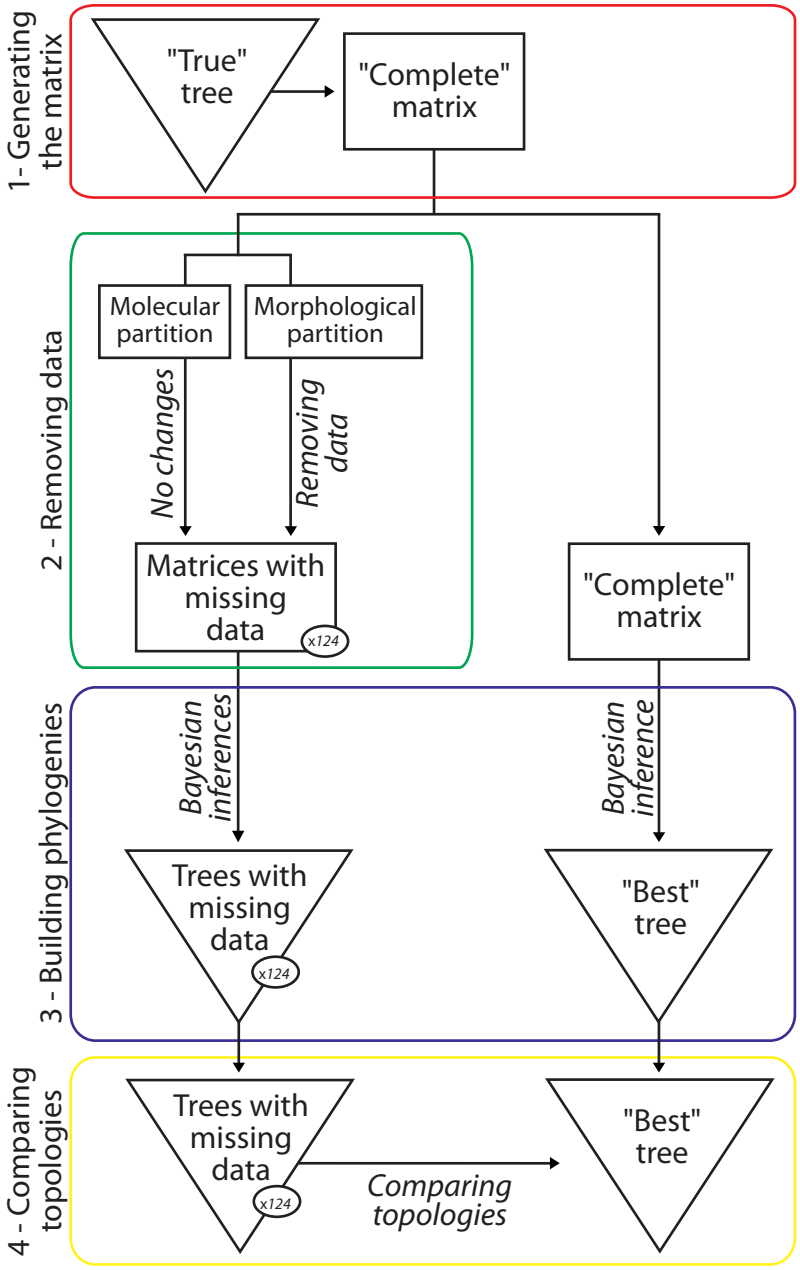

Figure 1: Protocol outline. (1) We generated a random tree (hereafter called the true tree) to infere a matrix with no missing data (hereafter, the complete matrix). (2) We removed data fron the morphological partition of the Complete matrix resulting in 124 matrices with missing data. (3) We build a bayesian phylogenetic tree from each matrix. (4) We then compared the trees with missing data (from the matrices with missing data) to the tree with no missing data (hereafter, the best tree infered from the complete matrix). We repeated step (1) to (4) 100 times.

matrices containing missing data to the ones obtained from the complete matrix to assess the influence of each parameter (M_L, M_F, M_C and their interactions) on the topologies of the phylogenies we estimated.

To measure the effect of missing datas distribution, we repeated steps (1) to (4) with the exact same fixed parameters 100 times.

We then repeated steps (1) to (4) with different numbers of taxa and characters to determine how this affected our results. Finaly we compared our results from simulated data with those from two empirical matrices (treated as in step (2) to (4)) and to fully random trees.

Generating the complete matrix

First we randomly generated a true tree of 50 taxa in R v3.0.2 (**?**) using the package diversitree v0.9-6 (**?**). We generated the tree using a Birth Death process by sampling the values of the speciation events () and extinction events () from a uniform distribution but maintaining ¿ (**?**). We implemented a rejection sampling algorithm to select only random trees with 25 living and 25 fossil taxa. We then added a species to the resulting Birth-Death tree as the outgroup of the tree. The mean branch length of the tree was used to separate the outgroup from the rest of the taxa and the branch length leading to the outgroup was set as the sum of the mean branch length and the longest root-to-tip length of the tree.

Next, we created a molecular and a morphological matrix from the true tree. The molecular matrix was infered from the true tree using the package phyclust v0.1-14 (**?**). The matrix was made of 1000 characters sites for 51 taxa and generated using the seqgen algorithm (**?**). We used the HKY model **?** with a random base frequencies and with the transition/transversion rate

of 2 (**?**) as parameters for generating the matrix. The substitution rates were distributed following a gamma distribution with an alpha () shape of 0.5 (**?**). We chose a low value of to lower the number of sites with high substitution rates, thus avoiding too much homoplasy and a decrease in phylogenetic signal. These parameters were selected to generate data with no special assumption about how the characters evolved as well as to reduce the computational time required if these parameters were estimated rather than defined (CPU time).

We infered the morphological matrix using the ape package v3.0-8 (**?**) to generate a matrix of 100 character sites for 51 taxa. We assigned the number of character states (either 2 or 3) 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 characters states extracted from 100 published empirical morphological matrices (Supplementary materials). We then ran an independent discrete character simulation for each character with the randomly selected number of states (2 or 3) and assuming an equal rate of change (i.e. evolutionary rate) from one character state to an other. This method allows us to have only two parameters per character i.e. the number of states and the evolutionary rate. For each character, the evolutionary rate was sampled from a gamma distribution with = 0.5. All the molecular information for fossil taxa was replaced by missing data (?). Finally, we combined the morphological and molecular matrices obtained from the true tree. Hereafter we call this the complete matrix, i.e. the matrix with no missing data except for the molecular data of the fossil taxa.

Removing data

Once we obtained the complete matrix we modified it to get a set of matrices with missing data. We randomly replaced data with ? in the morphological part of the matrices according to the following parameters:

1. (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 one.

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

3. (3) The proportion of missing morphological characters across all taxa (living and fossil) (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 represent a way of removing data in the morphological part of the matrix: ML removes a proportion of rows from the living taxa; M_F removes a proportion of cells from the fossil taxa; and M_L removes a proportion of columns across both living and fossil taxa. Note that M_F is different to M_L: for M_F, a proportion of data is removed across the whole of the morphological matrix for fossil taxa (i.e. removing cells) as for M_L, all the morphological data of a proportion of the living taxa is removed (i.e. removing rows). We tested all parameters combinations resulting in 125 (5^3) matrices. Because some parameter combinations introduce a lot of missing

(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 species had at least 5% data.

uilding phylogenies

From the resulting matrices we generated two types of trees, the best tree that is infered from the complete matrix and the trees with missing data infered 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 the 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. (**?**)). Therefore, here we focus on comparing the trees infered 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.

The best tree and the missing-data trees were infered using MrBayes v3.2.2 (**?**). We partitioned the data treating the molecular part as a non-codon DNA partition and the morphological part as a multi-state morphological partition. The molecular evolutionary history was infered using the HKY model (**?**) with a transition/transversion ratio (**?**) of two and a gamma distribution for the rate variation with four distinct categories (HKY + $_4$). For the morphological data, we used the Markov k state model (**?**) wich is a generalisation of the JC69 model (**?**) with k 2, assuming an equal state frequency and a unique overall substitution rate () following a gamma distribution of the rate variation with four distinct categories (Mk + $_4$). 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 5010^6 generations. We used the average standard deviation of split frequencies (ASDS) as a proxy to estimate the convergence of the chains and implemented a stop rule when the ASDS went below 0.01 (**?**). The effective sample size (ESS) was also checked on a random sub-sample of runs in each simulation to check that ESS ¿¿ 200 (**?**). For each run, we removed 25% of the iterations as burnin. We used the following priors for each tree (see Supplementary materials):

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

ii an exponential prior on the shape of the gamma distribution of) = 0.5 for both partitions.

iii a transition/transversion ratio prior of 2 sampled from a strong beta distribution ((80,40)).

We used these priors to speed up the Bayesian process. These priors biased the way the Bayesian process calculated the branch length by giving non-random starting points and boundaries for the parameters estimation process, however, we are focusing on the effect of missing data on the topology and not on the branch length. Even using these priors, it took CPU hours to build sets of 125 Bayesian trees.

Comparing topologies

blablabla

Results

blablabla

Discussion

blablabla

Conclusion

blablabla

Acknowledgements

blablabla