Version dated: February 13, 2016

Assessment of available anatomical characters for phylogenetic analysis among living mammals

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

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

[2] *Department of Life Sciences, Natural History Museum, Cromwell Road, London, SW7 5BD, UK.*

***Corresponding author.** *t.guillerme@imperial.ac.uk*

Abstract

ABSTRACT

Analyses of living and fossil taxa are crucial for understanding biodiversity through time. The Total Evidence method allows living and fossil taxa to be combined in phylogenies, using molecular data for living taxa and morphological data for living and fossil taxa. With this method, substantial overlap of coded anatomical characters among living and fossil taxa is vital for accurately inferring topology. However, although molecular data for living species are widely available, scientists generating morphological data mainly focus on fossils. Therefore, there are few coded anatomical characters in living taxa, even in well-studied groups like mammals.

We investigated the number of coded anatomical characters available in phylogenetic matrices for living mammals and how these were phylogenetically distributed across orders. 22 of 28 mammalian orders have <25% species with available characters; this has implications for the accurate placement of fossils, although the issue is less pronounced at higher taxonomic levels. In most orders, species with available characters are randomly distributed across the phylogeny, which may reduce the impact of the problem. We suggest that increased morphological data collection efforts for living taxa are needed to produce accurate Total Evidence phylogenies.

Key words: Total Evidence method, phylogenetic clustering, cladistic matrix, extinct, topology

Introduction

There is an increasing consensus among biologists that studying both living and fossil taxa is essential for fully understanding macroevolutionary patterns and processes [1, 2]. To perform such analyses it is necessary to combine living and fossil taxa in phylogenetic trees. One increasingly popular method, the Total Evidence method [3], ccombines molecular data from living taxa and morphological data from both living and fossil taxa in a supermatrix that can then be used with the tip-dating method (e.g. [4, 3, 5, 1, 6]), producing a chronogram with living and fossil taxa at the tips. A downside of this method is that it requires molecular data for living taxa and morphological/anatomical data for both living and fossil taxa. Sections of this data can be difficult, or impossible, to collect for every taxon in the analysis. For example, fossils rarely have molecular data and incomplete fossil preservation may reduce the number of anatomical characters available. Additionally, it has become less common to collect anatomical characters for living taxa when molecular data is available (e.g. in [7], only 13% of living taxa have coded anatomical characters). Unfortunately this missing data can lead to errors in phylogenetic inference. Simulations show that the ability of the Total Evidence method to recover the correct topology decreases when there is little overlap between coded anatomical characters in living and fossil taxa, and that the effect of missing data on topology is greatest when living taxa have few anatomical characters available [8]. This is because (1) fossils will not be placed accurately within the correct clade if it contains no coded anatomical characters for living taxa; and (2) fossils have a higher probability of being placed within clades with more coded anatomical characters available for living taxa, regardless of whether this is the correct clade [8].

The issues above highlight that it is crucial to have sufficient coded anatomical characters available for living taxa in a clade before using Total Evidence approaches. However, it is unclear how many coded anatomical characters are actually available for living taxa, i.e. already coded from museum specimens and deposited in phylogenetic matrices accessible online, and how this data is distributed across clades. Intuitively, most people assume this kind of data has already been collected, but empirical data suggest otherwise (e.g. in [3, 7, 6]). To investigate this further, we assess the number of available coded anatomical characters for living mammals to determine whether enough data exists to build reliable Total Evidence phylogenies. We also determine whether the characters are phylogenetically overdispersed or clustered across mammalian orders.

Materials and Methods

Data collection and standardisation

We downloaded all cladistic matrices containing any living and/or fossil mammal taxa from three major public databases: MorphoBank (`morphobank.org` [9]), Graeme Lloyd's website (`graemetlloyd.com/matrmamm.html`) and Ross Mounce's GitHub repository (github.com/rossmounce/cladistic-data). We also performed a systematic Google Scholar search for matrices that were not uploaded to these databases (see Electronic Supplementary Material (ESM) for details). In total, we downloaded 286 matrices containing 5228 unique operational taxonomic units (OTUs). We used OTUs rather than species because entries in the matrices ranged from species to families. We standardised the taxonomy as described in the ESM and excluded OTUs that were not present in the phylogeny of [10] or the taxonomy of [11] to remove fossil species. This resulted in 1601 unique OTUs from 286 matrices.

Data availability and distribution

To assess the availability of coded anatomical characters for each mammalian order and across mammals, we calculated the percentage of OTUs with cladistic data at three different taxonomic levels: family, genus and species. We consider orders with <25% of living taxa with available anatomical characters as having low data coverage, and orders with >75% of living taxa with available anatomical characters as having high data coverage.

For each order and for all mammals, we investigated whether the available coded anatomical characters were (i) randomly distributed, (ii) overdispersed or (iii) clustered, with respect to phylogeny, using two metrics from community phylogenetics: the Nearest Taxon Index (NTI; [12]) and the Net Relatedness Index (NRI; [12]). NTI is most sensitive to clustering or overdispersion near the tips, whereas NRI is more sensitive to it across the whole phylogeny [13]. Both metrics were calculated using the `picante` package in R [14, 15].

NTI [12] is based on mean nearest neighbour distance ($MNND$) and is calculated as follows:

$$\text{NTI} = -\left(\frac{\overline{MNND}_{obs} - \overline{MNND}_{n}}{\sigma(MNND_{n})}\right) \tag{1}$$

where $\overline{MNND}_{obs}$ is the observed mean sum of the branch lengths between each of n taxa with available coded anatomical characters and its nearest neighbour with available coded anatomical characters in the phylogeny, $\overline{MNND}_{n}$ is the mean of 1000 $MNND$ between n randomly drawn taxa, and $\sigma(MNND_{n})$ is the standard deviation of these 1000 random $MNND$ values. NRI is calculated in the same way, but using the mean phylogenetic distance (MPD):

$$\text{NRI} = -\left(\frac{\overline{MPD}_{obs} - \overline{MPD}_{n}}{\sigma(MPD_{n})}\right) \tag{2}$$

where $\overline{MPD}_{obs}$ is the observed mean phylogenetic branch length of the tree containing only the n taxa with available coded anatomical characters. Negative NTI and NRI values show that the focal taxa are more overdispersed across the phylogeny than expected by chance, and positive values reflect clustering.

We calculated NTI and NRI values for all mammals or each mammalian order separately, at each different taxonomic level. For each analysis our focal taxa were those with available coded anatomical characters at that taxonomic level and the phylogeny was the order pruned from [10].

Results

Across mammals, species coverage was low (<25% species with available coded anatomical characters) but family coverage was high (>75% families with available coded anatomical characters). For each order, 11 out of 28 had low coverage and seven had high coverage at the species-level. At the genus-level, one order had low coverage and 15 had high coverage, and at the family-level, no orders had low coverage and 25 had high coverage (Table**??**).

Across mammals, taxa with available coded anatomical characters were significantly clustered using NTI at the species- and genus-level. For each order, only seven showed significant clustering (Cetartiodactyla, Cingulata, Pilosa and Rodentia at the species-level and Carnivora, Chiroptera and Soricomorpha at both species- and genus-level) and none showed significant overdispersion (Table **??**).

Figure **??** shows randomly distributed OTUs with available coded anatomical characters in Primates (Figure **??**A) and phylogenetically clustered OTUs with available coded anatomical characters in Carnivora (mainly Canidae and Urisdae but no Herpestidae; Figure **??**B).

Discussion

Our results show that although phylogenetic relationships among living mammals are well-resolved (e.g. [10, 16]), most of the data used to build these phylogenies is molecular, and few coded anatomical characters are available for living mammals compared to fossils (e.g. [17, 18]). This has implications for building Total Evidence phylogenies, as without sufficient overlapping anatomical characters for living and fossil species, fossil placements in these trees may be unreliable [8].

The number of living mammalian OTUs with available coded anatomical characters was surprisingly low at the species-level: only 17%. Only seven out of 28 orders have a high coverage of taxa with available coded anatomical characters. This high coverage threshold of 75% of taxa with available characters represents the minimum amount of data required before missing data has a significant effect on the topology of Total Evidence trees [8]. Beyond this threshold, there is considerable displacement of wildcard taxa and decreased clade conservation [8]. Therefore we expect difficulties in placing fossils at the species-level in most mammalian orders, but fewer issues at higher taxonomic levels. This is important in practice because of the slight discrepancy between neontological and palaeontological species concepts. While neontological species are described using morphology, genes, distribution etc.; palaeontological species can be based only on morphological/anatomical, spatial and temporal data (e.g. [18]). Therefore, many palaeontological studies use the genus (or monospecific genera) as their smallest OTU (e.g. [18, 17]), so data availability at the genus-level in living mammals should be our primary concern when building phylogenies of living and fossil taxa.

When few species have available coded anatomical characters, the ideal scenario is for them to be evenly distributed (as measured by phylogenetic overdispersion) to maximize the possibilities of a fossil being placed in the correct clade. The second best scenario is that species with available characters are randomly distributed across the phylogeny. Here we expect no bias in the placement of fossils [8], it is therefore encouraging that for most orders, species with available coded anatomical characters were randomly distributed across the phylogeny. The worst case scenario for fossil placement is that species with available characters are phylogenetically clustered. Then we expect two major biases: first, fossils will not be placed within a clade containing no data, and second, fossils will have higher probability of being placed within the most sampled clade by chance. Our results suggest that this may be problematic at the genus-level in Carnivora, Chiroptera and Soricomorpha. For example, a Carnivora fossil will be unable to be placed in the Herpestidae clade because they have no coded anatomical characters available. Instead the fossil will have a high probability of being placed on a branch that contains many anatomical characters such as within the Canidae or Ursidae (Figure **??**B). This is analogous to the problem of long-branch attraction/short branch repulsion, as one can think of herpestids as having zero-length branches for anatomical characters, and canids and ursids having long branches and thus ”attracting” fossil placements.

It is worth noting, however, that our analysis did not include all the matrices containing anatomical characters ever published. In fact, our data collection procedure focused on including studies that provided matrices easily accessible, i.e. we did specifically not include any matrices that were only available in paper format (e.g.

printed in books), non-reusable format (e.g. an image of the matrix) or/and matrices available only upon request (e.g. by emailing the authors). The difficulty of obtaining such matrices makes them less likely to be included in Total Evidence analyses, especially where the number of taxa to include in the phylogeny is high.

Despite the absence of good morphological/anatomical data coverage for living mammals, the Total Evidence method still seems to be the most promising way of combining living and fossil species in macroevolutionary analyses. Following the recommendations in [8], we must code anatomical characters for as many living species as possible. Fortunately, mammal specimens are usually readily available in natural history collections, therefore, we propose increased effort into coding anatomical characters from living species, possibly by engaging in collaborative data collection projects. Such efforts would be valuable not only to phylogeneticists, but also to any researcher focusing understanding macroevolutionary patterns and processes.

Ethics statement

N/A

Data accessibility statement

All data and code are available on GitHub (`https://github.com/TGuillerme/Missing_living_mammals`).

Authors' Contributions

TG and NC designed the study. TG analysed the data. TG and NC wrote the the manuscript.

Competing Interests

We have no competing interests.

Acknowledgments

We thank David Bapst, Graeme Lloyd, Nick Matzke, April Wright, Graham Slater, Peter Wagner and one anonymous reviewer.

Funding statement

This was funded by a European Commission CORDIS Seventh Framework Programme (FP7) Marie Curie CIG grant (proposal: 321696).

References

*

References

[1] Slater GJ, Harmon LJ. Unifying fossils and phylogenies for comparative analyses of diversification and trait evolution. Methods Ecol Evol. 2013;4(8):699–702.

[2] Fritz SA, Schnitzler J, Eronen JT, Hof C, Böhning-Gaese K, Graham CH. Diversity in time and space: wanted dead and alive. Trends Ecol Evol. 2013;28(9):509 – 516.

[3] Ronquist F, Klopfstein S, Vilhelmsen L, Schulmeister S, Murray D, Rasnitsyn A. A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera. Syst Biol. 2012;61(6):973–999.

[4] Pyron R. Divergence time estimation using fossils as terminal taxa and the origins of Lissamphibia. Syst Biol. 2011;60(4):466–481.

[5] Schrago C, Mello B, Soares A. Combining fossil and molecular data to date the diversification of New World Primates. J Evolution Biol. 2013;26(11):2438–2446.

[6] Beck RM, Lee MS. Ancient dates or accelerated rates? Morphological clocks and the antiquity of placental mammals. P Roy Soc B-Biol Sci. 2014;281(20141278):1–10.

[7] Slater GJ. Phylogenetic evidence for a shift in the mode of mammalian body size evolution at the Cretaceous-Palaeogene boundary. Methods Ecol Evol. 2013;4(8):734–744.

[8] Guillerme T, Cooper N. Effects of missing data on topological inference using a Total Evidence approach. Mol Phylogenet Evol. 2016;94, Part A:146 – 158. Available from: http://www.sciencedirect.com/science/article/pii/S1055790315002547.

[9] O'Leary MA, Kaufman S. MorphoBank: phylophenomics in the cloud. Cladistics. 2011;27(5):529–537.

[10] Bininda-Emonds ORP, Cardillo M, Jones KE, MacPhee RDE, Beck RMD, Grenyer R, et al. The delayed rise of present-day mammals. Nature. 2007 03;446(7135):507–512.

[11] Wilson DE, Reeder DM. Mammal species of the world: a taxonomic and geographic reference. vol. 1. JHU Press; 2005.

[12] Webb CO, Ackerly DD, McPeek MA, Donoghue MJ. Phylogenies and community ecology. Ann Rev Ecol Syst. 2002;p. 475–505.

[13] Cooper N, Rodríguez J, Purvis A. A common tendency for phylogenetic overdispersion in mammalian assemblages. P Roy Soc B-Biol Sci. 2008;275(1646):2031–2037.

[14] Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, et al. Picante: R tools for integrating phylogenies and ecology. Bioinformatics. 2010;26:1463–1464.

[15] R Core Team. R: a language and environment for statistical computing. Vienna, Austria; 2015. Available from: `http://www.R-project.org`.

[16] Meredith R, Janečka J, Gatesy J, Ryder O, Fisher C, Teeling E, et al. Impacts of the Cretaceous terrestrial revolution and KPg extinction on mammal diversification. Science. 2011;334(6055):521–524.

[17] O'Leary MA, Bloch JI, Flynn JJ, Gaudin TJ, Giallombardo A, Giannini NP, et al. The placental mammal ancestor and the postK-Pg radiation of placentals. Science. 2013;339(6120):662–667.

[18] Ni X, Gebo DL, Dagosto M, Meng J, Tafforeau P, Flynn JJ, et al. The oldest known primate skeleton and early haplorhine evolution. Nature. 2013;498(7452):60–64.