Version dated: July 23, 2015

Assessment of cladistic data availability for 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.** *guillert@tcd.ie*

Abstract

Analyses of living and fossil taxa are crucial for understanding changes in biodiversity through time. The Total Evidence method allows living and fossil taxa to be combined in phylogenies, by using molecular data for living taxa and morphological data for both living and fossil taxa. With this method, substantial overlap of morphological data among living and fossil taxa is crucial for accurately inferring topology. However, although molecular data for living species is widely available, scientists using and generating morphological data mainly focus on fossils. Therefore, there is a gap in our knowledge of neontological morphological data even in well-studied groups such as mammals.

We investigated the amount of morphological (cladistic) data available for living mammals and how this data was phylogenetically distributed across orders. 22 of 28 mammalian orders have <25% species with available morphological data; this has implications for the accurate placement of fossil taxa, although the issue is less pronounced at higher taxonomic levels. In most orders, species with available data 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, morphological 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, 4], combines molecular data from living taxa and morphological data from both living and fossil taxa in a supermatrix (e.g. [5, 4, 6, 1, 7]), producing a phylogeny with living and fossil taxa at the tips. A downside of this method is that it requires molecular data for living taxa and morphological data for both living and fossil taxa. Chunks 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 restrict the amount of morphological data available. Additionally, it has become less common to collect morphological characters for living taxa when molecular data is available (e.g. in [8], only 13% of living taxa have coded morphological data). 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 morphological data in living and fossil taxa, and that the effect of missing data on topology is greatest when living taxa have few morphological data [9]. This is because (1) fossils cannot branch in the correct clade if it contains no morphological data for living taxa; and (2) fossils have a higher probability of branching within clades with more morphological data for living taxa, regardless of whether this is the correct clade [9].

The issues above highlight that it is crucial to have sufficient morphological data for living taxa in a clade before using a Total Evidence approach. However, it is unclear how much morphological data for living taxa is actually available, 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 [4, 8, 7]). To investigate this further, we assess the amount of available morphological data for living mammals to determine whether sufficient data exists to build reliable Total Evidence phylogenies in this group. We also determine whether the available cladistic data is 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 (`http://www.morphobank.org/` [10]), Graeme Lloyd's website (`graemetlloyd.com/matrmamm.html`) and Ross Mounce's GitHub repository (`https://github.com/rossmounce/cladistic-data`). We also performed a systematic Google Scholar search for matrices that were not uploaded to these databases (see Supplementary Materials Section 1 for a detailed description of the search procedure). In total, we downloaded 286 matrices containing 5228 unique operational taxonomic units (OTUs). We used OTUs rather than species since entries in the matrices ranged from species to families, and standardised the taxonomy as described in Supplementary Materials (section 1). We designated as "living" all OTUs that were either present in the phylogeny of [11] or the taxonomy of [12].

Matrices with few characters are problematic when comparing available data among matrices because (1) they have less chance of having characters that overlap with those of other matrices [13] and (2) they are more likely to contain a higher proportion of specific characters that are not-applicable across large clades (e.g. "antler ramifications" is a character that is only applicable to Cervidae not all mammals [14]). Therefore we selected only matrices containing >100 characters for each OTU. This threshold was chosen to correspond with the number of characters used in [9] and [15]. Results of analyses with no threshold are available in Supplementary Material. After removing matrices with <100 characters, we retained 1074 unique living mammal OTUs from 126 matrices.

Data availability and distribution

To assess the availability of cladistic data for each mammalian order, 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 cladistic data as having low data coverage, and orders with >75% of living taxa with cladistic data as having high data coverage.

We investigated whether the available cladistic data for each order was (i) randomly distributed, (ii) overdispersed or (iii) clustered, with respect to phylogeny, using two metrics from community phylogenetics: the Nearest Taxon Index (NTI; [16]) and the Net Relatedness Index (NRI; [16]). NTI is most sensitive to clustering or overdispersion near the tips, whereas NRI is more sensitive to clustering or overdispersion across the whole phylogeny [17]. Both metrics were calculated using the `picante` package in R [18, 19].

NTI [16] is based on mean nearest neighbour distance (MNND) and is calculated as follows:

$$NTI = -\left(\frac{\overline{MNND}_{obs} - \overline{MNND}_n}{\sigma(MNND_n)}\right) \quad (1)$$

where $\overline{MNND}_{obs}$ is the observed mean distance between each of n taxa with cladistic data and its nearest neighbour with cladistic data in the phylogeny, $\overline{MNND}_n$ is the mean of 1000 mean 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 MNND is replaced by mean phylogenetic distance (MPD) as follows:

$$NRI = -\left(\frac{\overline{MPD}_{obs} - \overline{MPD}_n}{\sigma(MPD_n)}\right) \quad (2)$$

where $\overline{MPD}_{obs}$ is the observed mean phylogenetic distance of the tree containing only the n taxa with cladistic data. 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 each mammalian order separately, at each different taxonomic level. For each analysis our focal taxa were those with available cladistic data at that taxonomic level and the phylogeny was that of the order pruned from [11].

Results

22 of 28 orders have low coverage (<25% species with cladistic data) and six have high coverage (>75% species with cladistic data) at the species-level. At the genus-level, three orders have low coverage and 12 have high coverage, and at the family-level, no orders have low coverage and 23 have high coverage (Table1).

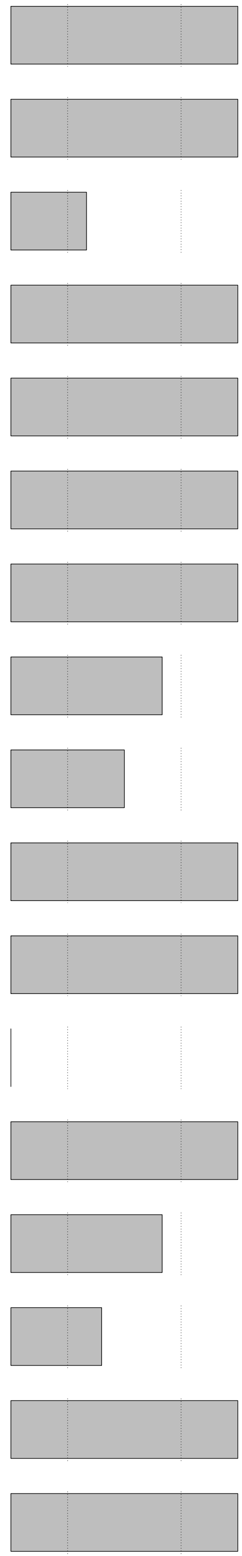
Table 1: Number of taxa with available cladistic data for mammalian orders at three taxonomic levels. The left vertical bar represents low coverage (<25%); the right vertical bar represents high coverage (>75%). Negative Net Relatedness Index (NRI) and Nearest Taxon Index (NTI) values indicate phylogenetic overdispersion; positive values indicate phylogenetic clustering. Significant NRI or NTI values are in bold. *p <0.05; **p <0.01; ***p <0.001.

Order	Taxonomic level	Proportion of taxa	Coverage	NRI	NTI
Afrosoricida	family	2/2			
Afrosoricida	genus	17/17			

Afrosoricida	**species**	**23/42**	**1.89***	1.19
Carnivora	family	11/15	0.43	1.68
Carnivora	**genus**	**30/125**	**4.14****	**1.81***
Carnivora	**species**	**42/283**	**18.64****	**3.02****
Cetartiodactyla	family	21/21		
Cetartiodactyla	**genus**	**77/128**	0.87	**1.77***
Cetartiodactyla	**species**	**129/310**	**2.72***	0.04
Chiroptera	family	13/18	0.55	0.63
Chiroptera	**genus**	**85/202**	**16.91****	**2.85****
Chiroptera	**species**	**165/1053**	**14.55****	**3.44****
Cingulata	family	1/1		
Cingulata	genus	8/9	1.49	-1.63
Cingulata	species	6/29	1.43	0.36
Dasyuromorphia	family	2/2		
Dasyuromorphia	genus	7/22	-1	-1.45
Dasyuromorphia	species	8/64	-1.15	-0.62
Dermoptera	family	1/1		

Dermoptera	genus	1/2		
Dermoptera	species	1/2		
Didelphimorphia	family	1/1		
Didelphimorphia	genus	16/16		
Didelphimorphia	species	40/84	-0.94	0.36
Diprotodontia	family	9/11	-0.8	0.56
Diprotodontia	genus	20/38	-1.36	-0.73
Diprotodontia	species	16/126	-2.29	-1.55
Erinaceomorpha	family	1/1		
Erinaceomorpha	genus	10/10		
Erinaceomorpha	species	21/22	-1.1	-0.3
Hyracoidea	family	1/1		
Hyracoidea	genus	1/3		
Hyracoidea	species	1/4		
Lagomorpha	family	1/2		
Lagomorpha	genus	1/12		
Lagomorpha	species	1/86		

Macroscelidea	family	1/1		
Macroscelidea	genus	4/4		
Macroscelidea	species	5/15	-0.98	-1.38
Microbiotheria	family	1/1		
Microbiotheria	genus	1/1		
Microbiotheria	species	1/1		
Monotremata	family	2/2		
Monotremata	genus	2/3	-0.71	-0.71
Monotremata	species	2/4	-1.01	-1.03
Notoryctemorphia	family	1/1		
Notoryctemorphia	genus	1/1		
Notoryctemorphia	species	0/2		
Paucituberculata	family	1/1		
Paucituberculata	genus	2/3	0	0
Paucituberculata	species	2/5	-0.64	-0.65
Peramelemorphia	family	2/2		
Peramelemorphia	genus	7/7		



Peramelemorphia	species	16/18	-0.09	1
Perissodactyla	family	3/3		
Perissodactyla	genus	6/6		
Perissodactyla	species	7/16	0.62	-2.5
Pholidota	family	1/1		
Pholidota	genus	1/1		
Pholidota	**species**	**3/8**	**2.64***	**2.23***
Pilosa	family	3/5	0.94	0.93
Pilosa	genus	3/5	-0.36	-0.31
Pilosa	species	3/29	0.33	0.79
Primates	family	15/15		
Primates	genus	48/68	-0.41	-1.4
Primates	species	56/351	-1.6	-2.04
Proboscidea	family	1/1		
Proboscidea	genus	1/2		
Proboscidea	species	1/3		
Rodentia	family	11/32	-0.46	-1.91

Rodentia	genus	21/450	-2.11	0.3
Rodentia	species	15/2094	-1.65	-2.55
Scandentia	family	2/2		
Scandentia	genus	2/5	-0.77	-0.76
Scandentia	species	2/20	-1.79	-1.99
Sirenia	family	2/2		
Sirenia	genus	2/2		
Sirenia	species	4/4		
Soricomorpha	family	3/4	-0.93	-0.92
Soricomorpha	**genus**	**19/43**	**6.98****	**2.49***
Soricomorpha	**species**	**19/392**	**13.19****	**3.89****
Tubulidentata	family	1/1		
Tubulidentata	genus	1/1		
Tubulidentata	species	1/1		

Only six orders had significantly clustered data (Afrosoricida and Pholidota at the species-level, and Carnivora, Cetartiodactyla, Chiroptera and Soricomorpha at both species- and genus-level) and none had significantly overdispersed data (Table 1).

Figure 1 shows randomly distributed OTUs with cladistic data in Primates (Figure 1A) and phylogenetically clustered OTUs with cladistic data in Carnivora (mainly Canidae; Figure 1B).

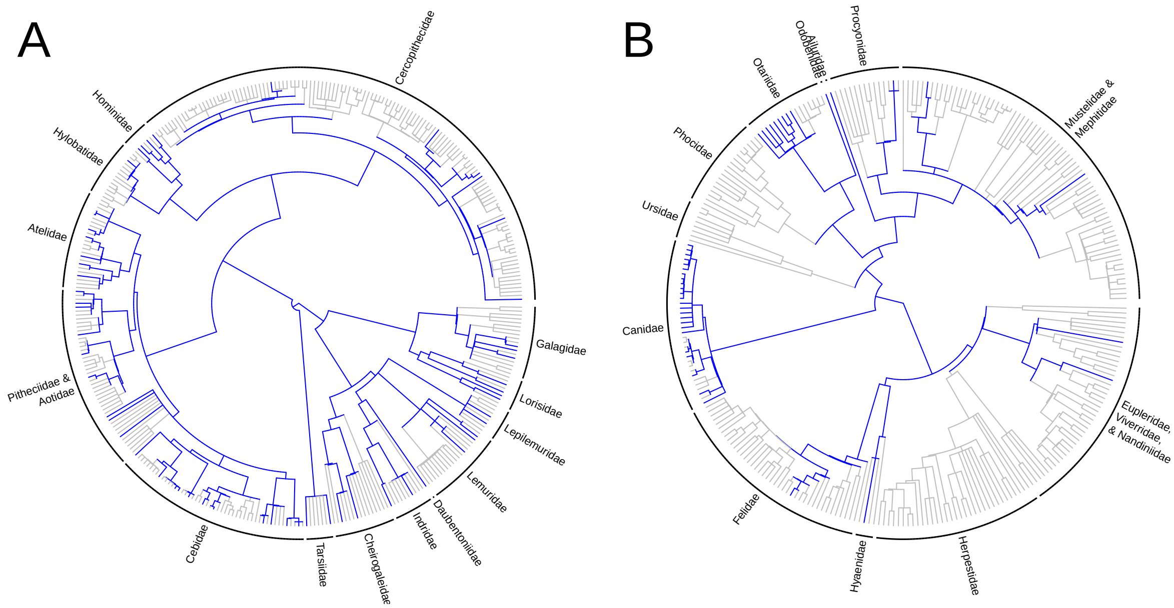


Figure 1: Phylogenetic distribution of species with available cladistic data across two orders (A: Primates; B: Carnivora). Blue branches indicate available cladistic data for the species.

Discussion

Our results show that although phylogenetic relationships among living mammals are well-resolved (e.g. [11, 20]) , most of the data used to build these phylogenies is molecular, and very little cladistic data is available for living mammals compared to fossil mammals (e.g. [21, 22]). This has implications for building Total Evidence phylogenies containing both living and fossil mammals, as without sufficient cladistic data for living species, fossil placements in these trees are very uncertain [9].

The number of living mammalian taxa with no available cladistic data was surprisingly high at the species-level: only six out of 28 orders have a high coverage of taxa with available cladistic data. This high coverage threshold of 75% of taxa with available cladistic data represents the minimum amount of data required before missing data has a significant effect on the topology of Total Evidence trees [9]. Beyond this threshold, there is considerable displacement of wildcard taxa (*sensu* [23]) and decreased clade conservation [9]. Therefore we expect difficulties in placement of fossil taxa at the species-level in most mammalian orders, but fewer issues at higher taxonomic levels. This point is important from a practical point of view 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, spatial and temporal data (e.g. [22]). Therefore, most palaeontological studies use genus as their smallest OTU (e.g. [22, 21]), 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 cladistic data, the ideal scenario is for them to be phylogenetically overdispersed to maximize the possibilities of a fossil branching from the right clade. The second best scenario is that species with cladistic data are randomly distributed across the phylogeny. Here we expect no special bias in the placement of fossils [9], it is therefore encouraging that for most orders, species with cladistic data were randomly distributed across the phylogeny. The worst case scenario for fossil placement is that species with cladistic data are phylogenetically clustered. Then we expect two major biases to occur: first, fossils will not be able to branch within a clade containing no data, and second, fossils will have higher probability of branching within the most sampled clade by chance. Our results suggest that this may be problematic at the genus-level in Carnivora, Cetartiodactyla, Chiroptera and Soricomorpha. For example, a Carnivora fossil will be unable to branch in the Herpestidae, and will have more chance to randomly branch within Canidae (Figure 1B).

Despite the absence of good cladistic data coverage for living mammals, the Total Evidence method still seems to be the most promising way of combining living and fossil data for macroevolutionary analyses. Following the recommendations in [9], we need to code cladistic characters for as many living species possible. Fortunately, data for living mammals is usually readily available in natural history collections, therefore, we propose that an increased effort be put into coding morphological characters from living species, possibly by engaging in collaborative data collection projects. Such an effort 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 analysis code is available on GitHub (`https://github.com/TGuillerme/Missing_living_mammals`).

Authors' Contributions

T.G. and N.C conceived and designed the experiments. T.G. performed the experiments and analysed the data. T.G. and N.C. contributed to the writing of the manuscript. All authors approved the final version of the manuscript.

Competing Interests

We have no competing interests.

Acknowledgments

We thank David Bapst, Graeme Lloyd, Nick Matzke and April Wright.

Funding statement

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

*

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] Eernisse D, Kluge A. Taxonomic congruence versus total evidence, and amniote phylogeny inferred from fossils, molecules, and morphology. Mol Biol Evol. 1993;10(6):1170–1195.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

[9] Guillerme T, Cooper N. Effects of missing data on topological inference using a Total Evidence approach. Mol Phylogenet Evol. In review;X(X):X.

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

[11] 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.

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

[13] Wagner PJ. Exhaustion of morphologic character states among fossil taxa. Evolution. 2000;54(2):365–386.

[14] Brazeau MD. Problematic character coding methods in morphology and their effects. Biol J Linn Soc. 2011;104(3):489–498.

[15] Harrison LB, Larsson HCE. Among-Character Rate Variation Distributions in Phylogenetic Analysis of Discrete Morphological Characters. Syst Biol. 2015;64(2):307–324.

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

[17] 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.

[18] 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.

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

[20] 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.

[21] 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.

[22] 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.

[23] Kearney M. Fragmentary taxa, missing data, and ambiguity: mistaken assumptions and conclusions. Syst Biol. 2002;51(2):369–381.