Materials and methods section of "Baraminology data filtering method based on entropy measurement and its application in dinosaur and cephalopod data sets" (Matthew Cserhati, Journal of Creation 33(3), 2019)

The entropy filter algorithm was applied to 90 cephalopod and 227 dinosaur species from four data sets. The goal was to refine baraminic predictions which have possibly been over-lumped into a smaller number of possible holobaramins. This was done on the Lindgren[1] and Sutton[2] cephalopod data sets, and the Brusatte[3], Lee[4], van der Reest[5], and Lamanna[6] dinosaur data sets. The BDIST results of the Zanno[7] dinosaur data set showed that the discovered putative baramins were well-segregated enough to make further analysis using the entropy filter unnecessary.

For each data set, a small subset of species was identified which made up what was deemed to be an over-lumped cluster. Character entropy filtering was performed on these species within the data set. The BDIST method was re-run on the remaining filtered data set to see if the entropy filtering was able to split up the selected species into a larger number of baramins, each with a smaller species membership.

The results of data filtering and re-clustering are reported here. For the analysis of each data set the original data set is included as well as a list of species for which entropy filtering was done. Furthermore, the filtered character set and the BDIST results are also provided for each analysis in a separate Excel file, which are available on github.com/csmatyi/EntropyFilter2.

Data sets for the cephalopod and dinosaur baraminology studies listed in table 1 were downloaded. The script EntropyFilter.R was written in R studio, version 1.1.442. The script itself, as well as supplementary figures and data files, are available on the github web page.

The script applies several filters to the data. First, it filters out those species which have a percentage of undefined ("?") characters above a certain cutoff. Next, it selects those species which are over-lumped. The names of these species should be listed in a separate txt file. The third filtering step is the most important and is essential to the whole method. This filtering step involved calculating entropy for each character. A column of character values is extracted from the double-filtered data set. Those characters are filtered out, and contain a certain percentage of undefined characters, just as with the row filtering criterion. Shannon entropy is calculated for each of the characters, minus the undefined states of a given character. Mixed characters, such as {0,1} are treated as separate characters (thus, 0, 1, and {0,1} count as three states of a given character). Shannon entropy is calculated in the following manner for a given character j:

$$H_j = \sum_{i=0}^{n} -p_i log_n p_i$$

Where n is equal to the number of states for character j, pi is equal to the probability of observing state i of the given character, and is equal to the number of occurrences of state i/the total number of occurrences for a character j. A minimum undefined character ratio for rows and columns and a minimal entropy value was selected for all data sets.

Figures 3, 4, and 10 were made using Cytoscape version 3.7.1. The bootstrapping values of the BDIST results of the entropy-filtered Brusatte, Lee, Lamanna, and van der Reest analyses were combined. An edge was placed between two species (vertices) if their bootstrapping value was ≥ 95+. Edge thickness was adjusted to reflect the number of BDIST studies which showed continuity between a given pair of species.

The baraminic distance correlation matrix as well as the stress graphs for all studies were generated using the BDIST software at coresci.org/bdist.html. The Venn diagram (figure 9) was created using the software available at bioinformatics.psb.ugent.be/webtools/Venn.

References

[1] Lindgren, A.R., Giribet, G., and Nishiguchi, M.K., A combined approach to the phylogeny of Cephalopoda (Mollusca), Cladistics 20:454–486, 2004.

[2] Sutton, M., Perales-Raya, C., and Gilbert, I., A phylogeny of fossil and living neocoleoid cephalopods, Cladistics (32):297–307, 2016 | doi:10.1111/cla.12131.

[3] Brusatte, S.L., Lloyd, G.T., Wang, S.C., and Norell., M.A., Gradual assembly of avian body plan culminated in rapid rates of evolution across the dinosaur-bird transition, Current Biology 24:2386–2392, 2014.

[4] Lee, M.S.Y., Cau, A., Naish, D., and Dyke, G.J., Sustained miniaturization and anatomical innovation in the dinosaurian ancestors of birds, Science 345(6196):562–566, 2014.

[5] van der Reest, A.J. and Currie, P.J., Troodontids (Theropoda) from the Dinosaur Park Formation, Alberta, with a description of a unique new taxon: implications for deinonychosaur diversity in North America, Canadian J. Earth Sciences 54(9):919–935, 2017.

[6] Lamanna, M.C., Sues, H.-D., Schachner, E.R., and Lyson, T.R., A new large-bodied oviraptorosaurian theropod dinosaur from the latest Cretaceous of western North America PLoS ONE 9(3):e92022, 2014.

[7] Zanno, L.E., A taxonomic and phylogenetic re-evaluation of Therizinosauria (Dinosauria: Maniraptora), J. Systematic Palaeontology 8(4):503–543, 2010.