Zoe Aiello

3.16.22

An Analysis of Xenarthra Clade Relationships

# 1. Introduction

## 1.1 Background

Xenarthra is a small but morphologically diverse clade of placental mammals consisting of anteaters, armadillos, and sloths (Delsuc et al., 2001, Gaudin et al, 2015). Between 65.5 and 55.8 million years ago, twelve distinct Xenarthran families emerged in South America (Delsuc et al., 2001). Today, only one surviving group lives outside of South America: the nine-banded armadillo (*Dasypus novemcinctus*) which takes up residence in the Southern United States (Delsuc et al., 2001). The fossil record of Xenarthrans is confined to the Americas except for three fossil discoveries: a pre-edentate (AKA a pre-Xenarthran) from the Upper Paleocene era of China, an anteater from the Middle Eocene era of Europe, and a sloth from the Eocene Era of Antarctica (Gaudin et al, 2015). These geographical exceptions to the fossil record of Xenarthrans along with their diverse morphology make uncovering their evolutionary history especially challenging (Gaudin et al., 2015). Despite these challenges, the Xenarthrans are a widely recognized monophyletic group – meaning that they share a single common ancestor which is not shared by any other group (Gaudin et al., 2015; Möller-Krull et al., 2007). There are many hypotheses which attempt to explain the Xenarthrans relationship with other placental mammals (Möller-Krull et al., 2007). However, one classical morphology-based hypothesis makes Xenarthrans the sister clade to all other placental mammals (Möller-Krull et al., 2007).

## 1.2 Purpose

As discussed in the background, many unique characteristics of the Xenarthra clade make their evolutionary relationships especially difficult to resolve. The purpose of this investigation is to use mitochondrial DNA (mtDNA) sequences of the large ribosomal RNA subunit gene (16S rRNA) to resolve evolutionary relationships between taxa in the Xenarthra clade. The 16S rRNA gene, is highly conserved – meaning it doesn’t mutate often – but 16S rRNA contains hypervariable regions – regions of the gene that mutate frequently (Janda, 2007). Since this gene is mostly the same across the tree of life but has regions which vary widely from species to species, it is ideal for species-specific comparisons (Janda, 2007). Additionally, mitochondrial DNA – the genetic information used by the powerhouse of the cell – is ideal for phylogenetic analysis due to it having a single parent and, therefore, a single evolutionary history (Corneli, 2000). Since mtDNA is only inherited by the mother, it never undergoes genetic recombination where the genes from each parent are ‘scrambled up’ before being inherited by the child (Corneli, 2000). Genetic recombination can make it difficult to trace the evolutionary history of nuclear DNA (nDNA) – the DNA housed in the nucleus which receives genes from both parents (Corneli, 2000). Furthermore, mtDNA has a high mutation rate which helps resolve relationships between closely related organisms (Corneli, 2000).

# 2. Methods

## 2.1 Data Preparation

Fasta files of mitochondrial DNA (mtDNA) sequences of the large ribosomal RNA subunit gene (16S rRNA) for twenty-three Xenarthra were obtained from NCBI’s database website. The twenty-three files contained 12 armadillo, 4 anteater, and 7 sloth sequences. Once all Fasta files were downloaded:

cat \*.fa\* > xenarthra.fasta

was used in the command line to combine all the files into one for use in statistical analysis. A table containing identifying information for the sequences can be found in Table 1.

Table 1. Mitochondrial DNA (mtDNA) sequences of the 16S rRNA gene from 12 armadillo, 4 anteater, and 7 sloth were found on the NCBI database website. The group identity and GenBank accession number is listed for each taxon.

|  |  |  |  |
| --- | --- | --- | --- |
| **Taxa** | **Group Identity** | **GenBank Accession #** | **Citation** |
| Zaedyus pichiy | Armadillo | NC\_028577.1 | Gibb et. al. (2016) |
| Tolypeutes tricinctus | Armadillo | NC\_028576.1 | Gibb et. al. (2016) |
| Myrmecophaga tridactyla | Anteater | NC\_028572.1 | Gibb et. al. (2016) |
| Megatherium americanum | Sloth | NC\_042737.1 | Desulc et. al. (2019) |
| Tolypeutes matacus | Armadillo | NC\_028575.1 | Gibb et. al. (2015) |
| Tamandua tetradactyla | Anteater | NC\_004032.1 | Arnason et. al. (2002) |
| Tamandua mexicana | Anteater | NC\_028574.1 | Gibb et. al. (2016) |
| Priodontes maximus | Armadillo | NC\_028573.1 | Gibb et. al. (2016) |
| Mylodon darwinii | Sloth | NC\_037941.1 | Delsulc et. al. (2018) |
| Euphractus sexcinctus | Armadillo | NC\_028571.1 | Gibb et. al. (2016) |
| Dasypus yepesi | Armadillo | NC\_028570.1 | Gibb et. al. (2016) |
| Dasypus septemcinctus | Armadillo | NC\_028569.1 | Gibb et. al. (2016) |
| Dasypus sabanicola | Armadillo | NC\_028568.1 | Gibb et. al. (2016) |
| Dasypus pilosus | Armadillo | NC\_028567.1 | Gibb et. al. (2016) |
| Dasypus novemcinctus | Armadillo | KT818542.1 | Gibb et. al. (2016) |
| Dasypus kappleri | Armadillo | NC\_028566.1 | Gibb et. al. (2016) |
| Dasypus hybridus | Armadillo | NC\_028565.1 | Gibb et. al. (2016) |
| Cyclopes didactylus | Anteater | NC\_028564.1 | Gibb et. al. (2016) |
| Choloepus hoffmanni | Sloth | NC\_027964.1 | Song et. al. (2015) |
| Bradypus variegatus | Sloth | NC\_028501.1 | Gibb et. al. (2016) |
| Bradypus torquatus | Sloth | NC\_028555.1 | Gibb et. al. (2015) |
| Bradypus pygmaeus | Sloth | NC\_028554.1 | Gibb et. al. (2016) |
| Acratocnus ye | Sloth | NC\_042752.1 | Desulc et. al. (2019) |

## 2.1 Analysis

R was used to create a distance matrix using kmers of the DNA sequences. Kmers are short subsequences of a specified length, k (Blaisdell, 1989). Using kmers to estimate evolutionary distance entails comparing the subsequences to determine their similarity (Blaisdell, 1989). It is an alternative to performing multiple sequence alignment – where one aligns the positions of the nucleotides in the different sequences (Blaisdell, 1989). Multiple sequence alignment is a computationally expensive task which makes kmer-based distance estimation an attractive substitute (Blaisdell, 1989). K-values that are too small or too large may lead to inaccuracies in estimation of evolutionary distance (Blaisdell, 1989). However, since our sequences are relatively closely related, a smaller k-mer is appropriate for estimating their evolutionary distance, therefore a k-value of 3 was used in the R function kdistance from the kmer package. The resulting distance matrix was used in the hclust function to cluster the sequences using the ‘average’ method which defines the cluster centers (centroids) as the average of all datapoints within the cluster. The hierarchical clustering results were mapped to a dendroid and cut into 3 clusters using the ‘cutree’ function in R (Figure 1). Three clusters were determined to be appropriate from previous literature (Delsulc et. al., 2015). The true clustering results separated the species into three groups: armadillo, sloth, and anteater. The classError() function from the mclust package in R was used to calculate the clustering error of the hierarchical clustering results.

Since the clustering results from the hierarchical clustering method were high, several packages from the Bioconductor module were imported to create a more accurate phylogenetic tree of the sequences. The msa package was used to import the sequences as a DNAbin object which allowed them to be aligned using the msa() function. The msa() function uses ClustalW, an algorithm for multiple sequence alignment in bioinformatics. An identity distance matrix was calculated from the alignment using dist.alignment(). The identity distance matrix counts the number of nucleotides that are identical between the sequences (n.d., 2010). The function nj() from the ape package was used to estimate a phylogenetic tree given the identity distance matrix. The nj() function uses a neighbor-joining algorithm which defines the best tree as the one which minimizes total branch length (Saitou and Nei, 1987). The branch length represents evolutionary distance and is measured in the number of nucleotide substitutions per site (i.e. between the sequences GCA and GTA there is one substitution in one site (C 🡪 T)) (Saitou and Nei, 1987). The aardvark, Orycteropus afer, was used as an outgroup as seen in Gibb et. al.’s construction of a Xenarthran phylogeny (Gibb et. al., 2016). Since we’re interested in comparing Xenarthra species, the importance of a non-xenarthran outgroup species is to serve as a reference point and allow for effective ingroup comparison (Baum, 2008). The groupClade() function from the ggtree package was used to group together the corresponding clades and a phylogenetic tree was plotted (Figure 2). A data frame was created with the information identifying each species and their corresponding group and was used to calculate the classification error of the phylogeny.

# 3. Results

## 3.1 Hierarchical Clustering

The hierarchical clustering resulted in the dendrogram shown in Figure 1. The first cluster (black) contained the following taxa: Bradypus torquatus (sloth), Priodontes maximus (armadillo), Choloepus hoffmanni (sloth), Bradypus pygmaeus (sloth), Megatherium americanum (sloth), Bradypus variegatus (sloth), Acratocnus ye (sloth), and Mylodon darwinii (sloth). The second cluster (red) included Cyclopes didactylus (anteater). The third cluster (green) included Dasypus kappleri (armadillo), Tolypeutes tricinctus (armadillo), Tolypeutes matacus (armadillo), Myrmecophaga tridactyla (anteater), Tamandua Mexicana (anteater), Tamandua tetradactyla (anteater), Dasypus hybridus (armadillo), Dasypus septemcinctus (armadillo), Dasypus pilosus (armadillo), Dasypus novemcinctus (armadillo), Dasypus sabanicola (armadillo), Dasypus yepesi (armadillo), Euphractus sexcinctus (armadillo), and Zaedyus pichiy (armadillo). The by-clade breakdown is as follows:

Table 2. A breakdown of how many taxa of each clade were grouped into each cluster.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Clade | Armadillo | Anteater | Sloth | Total |
| Cluster 1 | 1 | 0 | 7 | 8 |
| Cluster 2 | 0 | 1 | 0 | 1 |
| Cluster 3 | 11 | 3 | 0 | 14 |
| Total | 12 | 4 | 7 | 23 |

Cluster 1 appears to correspond to the sloth clade since most cases within the cluster are sloths while cluster 2 corresponds to the anteater clade and cluster 3 corresponds to the armadillo clade (Table 2). Cluster 1 includes one misclassification: the armadillo Priodontes maximus. Cluster 2 includes 0 misclassifications, but only features one taxon: the anteater: Cyclopes didactylus (Table 2). Cluster 3 includes 3 misclassifications: the anteaters: Myrmecophaga tridactyla, Tamandua Mexicana, and Tamandua tetradactyla (Table 2). The anteaters seemed to be the most difficult taxa to cluster using this method. Interestingly, all the anteaters were closest to each other in the dendrogram tree (Figure 1). The anteaters in the tamandua genus were grouped together and sister to Myrmecophaga tridactyla (Figure 1). Additionally, all taxa within the Dasypus lineage were closest to each other in the tree aside from Dasypus kappleri which was positioned sister to the Tolypeutes lineage (Figure 1). However, the genera in cluster 1 were not well grouped together as the sloths in the Bradypus lineage were separated from each other (Figure 1). Since 4 out of the 23 taxa were misclassified, the classError function resulted in a classification error of 17.39% for the hierarchical clustering method.

Chart

Description automatically generated with low confidence

Figure . Dendrogram of hierarchical clustering from 23 16S rRNA mtDNA Xenarthra sequences. The coloring corresponds to the resulting clusters when the number of clusters is 3.

## 3.2 Phylogenetic Clustering

The phylogeny method resulted in the tree shown in Figure 2. The first cluster (green) included the following taxa: Dasypus yepesi (armadillo), Dasypus sabanicola (armadillo), Dasypus pilosus (armadillo), Dasypus novemcinctus (armadillo), Dasypus septemcinctus (armadillo), Dasypus hybridus (armadillo), Dasypus kappleri (armadillo), Tolypeutes tricinctus (armadillo), Tolypeutes matacus (armadillo), Priodontes maximus (armadillo), Zaedyus pichiy (armadillo), and Euphractus sexcinctus (armadillo). The second cluster (blue) included the following taxa: Mylodon darwinii (sloth), Choloepus hoffmanni (sloth), Acratocnus ye (sloth), Megatherium americanum (sloth), Bradypus variegatus (sloth), Bradypus pygmaeus (sloth), and Bradypus torquatus (sloth). The third cluster (purple) included the following taxa: Tamandua tetradactyla (anteater), Tamandua Mexicana (anteater), Myrmecophaga tridactyla (anteater), and Cyclopes didactylus (anteater). The by-clade breakdown is as follows:

Table 3. A breakdown of how many taxa of each clade were grouped into each cluster.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Clade | Armadillo | Anteater | Sloth | Total |
| Cluster 1 | 12 | 0 | 0 | 12 |
| Cluster 2 | 0 | 0 | 7 | 7 |
| Cluster 3 | 0 | 4 | 0 | 4 |
| Total | 12 | 4 | 7 | 23 |

A picture containing chart

Description automatically generated

Figure . A phylogeny of 23 16S rRNA mtDNA Xenarthran sequences produced using the neighbor joining algorithm and rooted using Orycteropus afer.

Cluster 1 corresponds to the armadillo clade, cluster 2 corresponds to the sloth clade, and cluster 3 corresponds to the anteater clade (Table 3). There were 0 misclassifications in each clade which led to a classification error of 0%. Additionally, all genera were properly grouped together. The armadillos in the Dasypus lineage were all deemed to be closely related to each other, in addition to the Tolypeutes lineage, the Bradypus lineage, and the Tamandua lineage (Figure 2).

# 4. Discussion and Conclusion

## 4.1 Discussion

Hierarchical clustering resulted in a very high error of 17.39% while the phylogenetic method resulted in an error of 0%. For the hierarchical clustering method, the sloth cluster (cluster 1), included only one misclassification: the armadillo Priodontes maximus (Table 2). Priodontes maximus is unique as it is the largest species of armadillo by a substantial degree (Carter et. al., 2015). The distinctive nature of Priodontes maximus may have caused the misclassification by the hierarchical clustering method since the clustering was based solely off a K-mer-based distance matrix – a very simplistic algorithm. Additionally, the choice of K is an important one, and is highly dependent on the specific situation. The algorithms used by the phylogenetic clustering method are much more complex. The multiple sequence alignment function in R (msa) uses ClustalW: an algorithm which determines the best alignment of the sequences by taking many parameters into account, including substitution matrices (Pais et al., 2014). Substitution matrices describe the probability of substitution events for every amino acid (amino acids are the building blocks of proteins and are made up by 3 nucleotides) (Altschul et al., 1991). These substitution matrices are based off decades of experimental research and dramatically improve the accuracy of an alignment, and by proxy, a distance matrix (Altschul et al., 1991). Furthermore, the nearest neighbor algorithm used in the phylogenetic clustering analysis takes branch length into account when determining which topology is most appropriate (Saitou and Nei, 1987). As discussed in the methods section, branch length measures evolutionary distance (Saitou and Nei, 1987). The hierarchical clustering method did fairly well considering that it’s algorithms are not tailored to resolve phylogenetic relationships. The phylogenetic clustering method relies on evolutionary models and accounts for biological processes through every step. All things considered; the resulting error rates are reasonable given that the phylogenetic clustering functions were tailored for this purpose while the hierarchical clustering functions were not.

## 4.2 Conclusion

The purpose of this investigation was to resolve evolutionary relationships within the Xenarthra clade. The phylogenetic clustering method proved to be a more accurate way to resolve evolutionary relationships. Through phylogenetic analysis, it was discovered that the sloth and anteater clades are more closely related to each other than either are to armadillos which is consistent with the literature (Delsulc et. al., 2001). Additionally, the finding that the Tolypeutes genus is most closely related to Priodontes maximus (Figure 2) is also congruent with previous findings (Carter et. al., 2015). Although the phylogenetic analysis was very accurate in some respects, there are limitations to consider. ClustalW is a very fast algorithm because it combines local alignment (checking accuracy of alignments subsequences) and global alignment (checking accuracy of alignments through the whole sequence) (Pais et al., 2014). However, while the ClustalW algorithm is very fast, it makes some sacrifices in accuracy (Pais et al., 2014). Inaccurate alignments would lead to inaccurate relationships which is a limitation to consider. A future study might consider using a more computationally expensive alignment algorithm like MAFFT, which may result in more accurate and reliable alignments (Pais et al., 2014). Overall, the Xenarthrans are a unique and morphologically diverse group of mammals that share common ancestry.

References

**[1]** Delsuc F, Catzeflis F, Stanhope M, Douzery E 2001. The evolution of armadillos, anteaters and sloths depicted by nuclear and mitochondrial phylogenies: implications for the status of the enigmatic fossil Eurotamandua. Proceedings of the Royal Society B: Biological Sciences, Royal Society 268: 1605-15.

**[2]** Gaudin T, Croft D 2015. Paleogene Xenarthra and the evolution of South American mammals. Journal of Mammalogy 96: 622–634,

**[3]** Möller-Krull M, Delsuc F, Churakov G, Marker C, Superina M, Brosius J, Douzery E, Schmitz J 2007. Retroposed Elements and Their Flanking Regions Resolve the Evolutionary History of Xenarthran Mammals (Armadillos, Anteaters, and Sloths). Molecular Biology and Evolution 24: 2573-2582

**[4]** Davy, S. K., & Wilkinson , S. P. (2018, June 25). *phylogram: an R package for phylogenetic analysis with dendrograms*. Phylogram: An R package for phylogenetic analysis with dendrograms. Retrieved February 18, 2022, from <https://cran.r-project.org/web/packages/phylogram/vignettes/phylogram-vignette.html>

**[5]** Corneli, P.S., & Ward, R.H. (2000, February). Mitochondrial Genes and Mammalian Phylogenies: Increasing the Reliability of Branch Length Estimation, Molecular Biology and Evolution, Volume 17, Issue 2, Pages 224 –234, <https://doi.org/10.1093/oxfordjournals.molbev.a026302>

**[6]** Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution, 4, 406--425.

**[7]** Altschul SF. Amino acid substitution matrices from an information theoretic perspective. J Mol Biol. 1991 Jun 5;219(3):555-65. doi: 10.1016/0022-2836(91)90193-a. PMID: 2051488; PMCID: PMC7130686.

**[8]** Pais, F. S., Ruy, P. C., Oliveira, G., & Coimbra, R. S. (2014). Assessing the efficiency of multiple sequence alignment programs. *Algorithms for molecular biology : AMB*, *9*(1), 4. <https://doi.org/10.1186/1748-7188-9-4>

**[9]** Janda, J. M., & Abbott, S. L. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *Journal of clinical microbiology*, *45*(9), 2761–2764. <https://doi.org/10.1128/JCM.01228-07>

**[10]** Blaisdell, B. E. (1989). Average values of a dissimilarity measure not requiring sequence alignment are twice the averages of conventional mismatch counts requiring sequence alignment for a computer-generated model system.   
*J Mol Evol*, 29(6):538-47.

**[11]** *Sequence alignment*. Sequence alignment - Bioinformatics.Org Wiki. (2010, November 24). Retrieved March 16, 2022, from http://www.bioinformatics.org/wiki/Sequence\_alignment

**[12]** David Baum, "Reading a Phylogenetic Tree: The Meaning of Monophyletic Groups," *Nature Education* 1, no. 1 (2008): 190,

**[13]** Carter, T. S., Superina, M., & Leslie, D. M. (2016). Priodontes Maximus(cingulata: Chlamyphoridae). *Mammalian Species*, *48*(932), 21–34. https://doi.org/10.1093/mspecies/sew002

**Samples:**

**[1]** Gibb,G.C., Condamine,F.L., Kuch,M., Enk,J., Moraes-Barros,N., Superina,M., Poinar,H.N. and Delsuc,F. Shotgun Mitogenomics Provides a Reference Phylogenetic Framework and Timescale for Living Xenarthrans. Mol. Biol. Evol. 33 (3), 621-642 (2016)

**[2]** Delsuc,F., Kuch,M., Gibb,G.C., Karpinski,E., Hackenberger,D., Szpak,P., Martinez,J.G., Mead,J.I., McDonald,H.G., MacPhee,R.D.E., Billet,G., Hautier,L. and Poinar,H.N. Sample Submitted (08-MAY-2019) at Institut des Sciences de l'Evolution, CNRS Universite de Montpellier, Place Eugene Bataillon, Montpellier 34095, France

**[3]** Gibb,G.C., Condamine,F.L., Melanie,K., Enk,J., Moraes-Barros,N., Superina,M., Poinar,H.N. and Delsuc,F. Sample Submitted (23-SEP-2015) Institut des Sciences de l'Evolution, CNRS Universite de Montpellier, Place Eugene Bataillon, Montpellier 34095, France

**[4]** Arnason,U., Adegoke,J.A., Bodin,K., Born,E.W., Esa,Y.B., Gullberg,A., Nilsson,M., Short,R.V., Xu,X. and Janke,A. Mammalian mitogenomic relationships and the root of the eutherian tree. Proc. Natl. Acad. Sci. U.S.A. 99 (12), 8151-8156 (2002)

**[5]** Delsuc,F., Kuch,M., Gibb,G.C., Hughes,J., Szpak,P., Southon,J., Enk,J., Duggan,A.T. and Poinar,H.N. Resolving the phylogenetic position of Darwin's extinct ground sloth (/Mylodon darwinii/) using mitogenomic and nuclear exon data. Proc. Biol. Sci. 285, 20180214 (2018)

**[6]** Song,X., Chen,L. and Chen,X. Direct Submission Submitted (25-JUN-2015) BGI Education Center, University of Chinese Academy of Sciences, No. 11 Building, Beishan Industrial Park, Yantian District, Shenzhen, Guangdong 518083, China