Chris Deardeuff
1236847562
BIO-549 Phylogenetic Biology & Analysis
Dr. Nate Upham
2025-10-24

Background

This Project will analyze and build a phylogeny for the family Spheniscidae, commonly referred to as penguins. They are part of the order Sphenisciformes, and are flightless birds adapted to live near, and operate in, aquatic environments. Spheniscidae often have countershading to assist with camouflaging in their aquatic habitats, and their wings resemble paddles [1]. The majority of Spheniscidae live only in the southern hemisphere [2]. Some of the smaller species of penguins occupy temperate or even tropical climates, while the larger penguins are mostly in the cooler subantarctic regions. There are eighteen extant species composing six genera recognized by the International Ornithologists' Union [3].

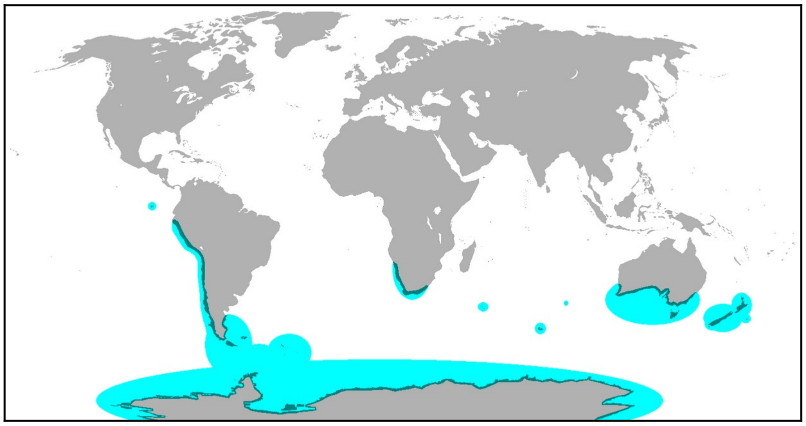

Figure 1*: Breeding range of all penguin species*

Spheniscidae was chosen for this project because they are a widely recognized species, but evolutionarily still have room for understanding of their adaptation to their environments [4] which makes them appealing for further investigation. They also are unique in features as they are the only flightless birds that also dive. Overall, Spheniscidae make an ideal group to analyze as they have unique and specialized adaptations that give them a distinct morphology while also being sensitive to even minor changes in the climate [5].

Materials

For the data used in this project, GenBank returns 381 total entries for Spheniscidae cytochrome b (cytB) gene. From these results one is a parasite that infects them so that has been removed. The results also include multiple samples of the same species (e.g. *Eudyptes* chrysocome has 5 samples). One sample of *Macronectes giganteus* (Southern Giant Petrel) is included as an outgroup as it is the closest living relatives to Spheniscidae. GenBank returns 15 entries for *Macronectes giganteus*, but only one is used. Overall, there are 19 species in the dataset with 18 unique penguin species, and one outgroup species. The downloaded datasets were manually cleaned (refactored names and removed duplicate taxa), and MAFFT software was used to align the sites and fill missing sites with dashes.

Methodology

A partitioned data analysis will be used in RevBayes. The partitioned data analysis allows for different rates to be applied to different codon positions, in this case the first and second codons vs the third codon, which often evolve at different rates. RevBayes uses a Markov chain Monte Carlo (MCMC) algorithm to explore the parameter space and then use Bayes factors to compare and choose which partitioning scheme best fits. The tutorial being referenced for this analysis is the Partitioned Data Analysis [6].

Results

Overall, 50,000 generations of two independent MCMC chains were run with a sample frequency of 10. A burn-in of ten percent was used and thus discarded the first 5,000 trees. This gave the runs an effective sample size (ESS) greater than 3,500 which is vastly over the greater than 500 'great' mark. The total tree length (TL) is at 1.256 which, because no node dating was used, represents the average total number of substitutions per site.

Another key parameter from this analysis is the invariable sites represented by pinvar. The total number of estimated sites that did not change, pinvar[1], was quite high at 81%. This shows that the cytB gene region is likely not the best to solely look at and the dataset might need more samples of the entire gene. Part_rate_multi[1] and part_rate_multi[2] are interesting parameters that highlight the rate of substitution between codon positions 1 and 2 vs the 3rd codon position. In this analysis, the 1st and 2nd codons have a substitution rate of 0.122, while the 3rd codon has a substitution rate of 0.878. This means that the 3rd codon is estimated to evolve at a rate 7.2 times faster than the 1st and 2nd codons. The high rate of change between the codons supports the use of partitioned data analysis with this dataset.

The topology generated by this analysis overall has low posterior probability (PP) for every node, as shown in Figure 2. The highest is 38% for the clade including the outgroup (Giant Southern Petrel) and *Spheniscidae's*, while the lowest values are around 1% for some members of the genus *Eudyptes*. Generally, greater than or equal to 95% is considered good support for a clade.

Besides the split between the genus *Aptenodytes* and the rest of the family, most branches have shorter lengths indicating more rapid substitutions in the cytB gene, but there are still a few like *Eudyptula minor* with longer branches indicating less substitutions.

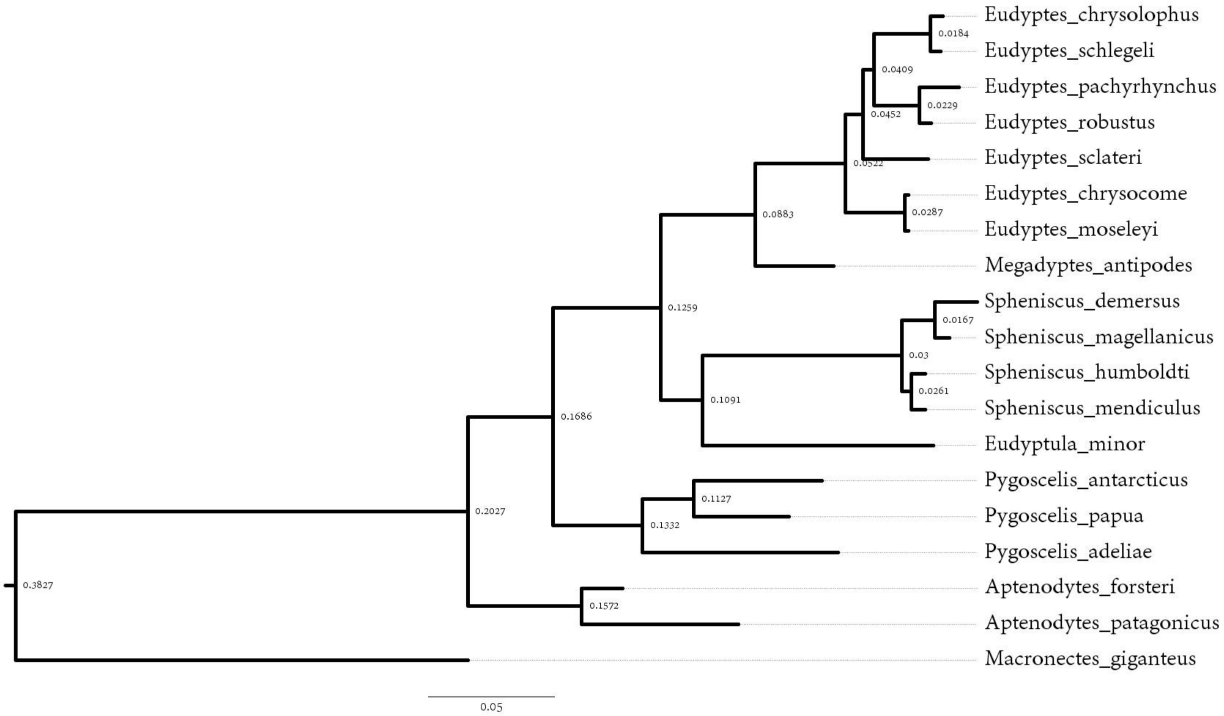

Figure 2*: Phylogenetic tree of penguin species and outgroup*

Discussion

Overall, the posterior probability for the nodes on the tree are very low. This is possible for multiple reasons, the most likely due to not enough coverage in the gene from the dataset. Out of the 19 taxa, only three: *Macronectes giganteus*, *Eudyptes chrysolophus*, and *Eudyptes chrysocome*, had the full 1,162 sites from the cytB gene, while the rest contain between 600 and 800 sites. The analysis was re-run with 500,000 generations with a burn-in of 10% and the posterior probability for the nodes were not more certain with changes only ranging in the plus or minus 0.01%. This further supports the hypothesis that more complete datasets are needed for the taxa to get a more certain result. Despite lower posterior probability, my findings are similar with more recent analyses that use whole genomes.

Similar to previous publications [7-11], this analysis has King (*A. patagonicus*) and Emperor (*A. forsteri*) penguins diverging early and being sister taxa to all extant penguins. It also has similar clades like *Eudyptes* and *Megadyptes* being sister genera. The biggest differences in topology from this analysis compared with other analyses is the placement of *Eudyptes chrysolophus* (Macaroni penguin) and *Eudyptes schlegeli* (Royal penguin) as sister taxa to *Eudyptes pachyrhynchus* (Firodland penguin) and *Eudyptes robustus* (Snares penguin) within the overall clade of *Eudyptes* rather than Macaroni and Royal penguins being sister taxa to all *Eudyptes* [12].

To further improve upon this analysis two major factors could be improved. Firstly, this tree is using relative dating and node time constraints. By adding strict fossil and geological calibrations it can better estimate substitution rates which can improve posterior probabilities, which this model would greatly benefit from. The next major improvement would be a more complete cytB sample for all the taxa which are lacking the complete gene. From a more in-depth search of GenBank, some of the taxa included in this analysis do have more complete Coding DNA Sequence (CDS), but most still have only partial CDS which will still limit the analysis. Another possibility is to run this analysis as both a partitioned data analysis between codons of cytB, but also with other gene regions substituting at different rates. Other mitochondrial gene regions could offer more insight into the more recent splits which have the lowest posteriors in this analysis. But, like the cytB options, either are only partial or don't exist at all within GenBank for *Spheniscidae*.

FIGURES

Figure 1: File: Penguin range.png - Wikimedia Commons. (2022). Wikimedia.org [Wikimedia]
Figure 2: Phylogenetic tree of penguin species including outgroup

REFERENCES

1. Prevost, J. (2020). Penguin - Natural history | Britannica. In Encyclopedia Britannica. [Britannica]
2. Prevost, J. (2020). Penguin - Natural history | Britannica. In Encyclopedia Britannica. [Britannica]
3. Kagu, Sunbittern, tropicbirds, loons, penguins – IOC World Bird List. (n.d.). [World Bird Names]
4. Vianna, J. A., Fernandes, F. A. N., Frugone, M. J., Figueiró, H. V., Pertierra, L. R., Noll, D., Bi, K., Wang-Claypool, C. Y., Lowther, A., Parker, P., Le Bohec, C., Bonadonna, F., Wienecke, B., Pistorius, P., Steinfurth, A., Burridge, C. P., Dantas, G. P. M., Poulin, E., Simison, W. B., & Henderson, J. (2020). Genome-wide analyses reveal drivers of penguin diversification. Proceedings of the National Academy of Sciences, 117(36), 22303–22310. [PNAS]
5. Pan, H., Cole, T. L., Bi, X., Fang, M., Zhou, C., Yang, Z., Ksepka, D. T., Hart, T., Bouzat, J. L., Argilla, L. S., Bertelsen, M. F., Boersma, P. D., Bost, C.-A., Cherel, Y., Dann, P., Fiddaman, S. R., Howard, P., Labuschagne, K., Mattern, T., & Miller, G. (2019). High-coverage genomes to elucidate the evolution of penguins. GigaScience, 8(9). [Gigascience]
6. Partitioned data analysis. Github.io. [Github]
7. Vianna, J. A., Fernandes, F. A. N., Frugone, M. J., Figueiró, H. V., Pertierra, L. R., Noll, D., Bi, K., Wang-Claypool, C. Y., Lowther, A., Parker, P., Le Bohec, C., Bonadonna, F., Wienecke, B., Pistorius, P., Steinfurth, A., Burridge, C. P., Dantas, G. P. M., Poulin, E., Simison, W. B., & Henderson, J. (2020). Genome-wide analyses reveal drivers of penguin diversification. Proceedings of the National Academy of Sciences, 117(36), 22303–22310. [PNAS]
8. A. J. Baker, S. L. Pereira, O. P. Haddrath, K. A. Edge, Multiple gene evidence for expansion of extant penguins out of Antarctica due to global cooling. *Proc. Biol. Sci.* 273, 11–17 (2006). [PUBMED]

9. J. A. Clarke et al., Paleogene equatorial penguins challenge the proposed relationship between biogeography, diversity, and Cenozoic climate change. *Proc. Natl. Acad. Sci. U.S.A.* 104, 11545–11550 (2007). [PUBMED]

10. D. T. B. S. Ksepka, N. Giannini, The phylogeny of the living and fossil Sphenisciformes (penguins). *Cladistics* 22, 412–441 (2006). [Cladistics]

11. S. Bertelli, N. P. Giannini, A phylogeny of extant penguins (Aves: Sphenisciformes) combining morphology and mitochondrial sequences. *Cladistics* 21, 209–239 (2005). [Cladistics]

12. Vianna, J. A., Fernandes, F. A. N., Frugone, M. J., Figueiró, H. V., Pertierra, L. R., Noll, D., Bi, K., Wang-Claypool, C. Y., Lowther, A., Parker, P., Le Bohec, C., Bonadonna, F., Wienecke, B., Pistorius, P., Steinfurth, A., Burridge, C. P., Dantas, G. P. M., Poulin, E., Simison, W. B., & Henderson, J. (2020). Genome-wide analyses reveal drivers of penguin diversification. Proceedings of the National Academy of Sciences, 117(36), 22303–22310. [PNAS]

DATA AND MATERIALS

all scripts, datasets, outputs, and deliverables can be found at
https://github.com/ChrisDeardeuff/BIO-549-Phylogeny-Project-1