IPS Instituto Politécnico de Setúbal
Escola Superior de Tecnologia do Barreiro

# Assignment 1

# Deep into phylogenetics


Professor:
Francisco Pina Martins

Students:
2675 Daniel Marçal
2691 Marcelo Pereira
2814 Bernardo Augusto

# Contents

# Introduction

In this assignment, we were given multiple tasks of which the first was choosing a scientific article that matched the required conditions. The main one was it had to have a phylogenetic analysis about a certain topic and we also decided to reduce the search to an article that used mainly the same programs and ideas as we learned and used during classes so it was easier for us to fully understand what was done and why they did it. The article that we eventually choose was a research article named *Evolution of the mitochondrial genome in snakes: Gene rearrangements and phylogenetic relationships*.

A team of 3 people, Jie Yan, Hongdan Li and Kaiya Zhou wanted to further investigate the gene organizations, the evolution of the snake mitochondrial genome, and phylogenetic relationships among several major snake families. This was since the phylogeny of snakes is somewhat controversial as mitochondrial and nuclear genes are limited. The team choose four different snake families and used their entire mitochondrial DNA sequences to achieve the main objective. Spreading their data also helped to understand the evolution of the mitochondrial genome as well as to determine the phylogenetic relationships between all species that were taken into consideration.

The first step of the process was gathering the samples of the already mentioned snake families (three alethinophidian families and one scoleophidian family). There was no need to have big quantities as short fragments of mitochondrial DNA were easily amplified via Takara DNA polymerase. Then with the following conditions: 5 min at 95°C trailed by 35 cycles of 95°C for 30 s, 50–55°C for 30 s, and 72°C for 90 s PCRs were done. PCR products were after sequenced with the help of ABI310 or 3700 systems and multiple internal primers. Depending on what was needed to identify different approaches were used for individual genes they looked for correspondence between the samples and homolog ones from other vertebrates. For tRNA genes, their secondary structures were used with the assistance of DNA-SIS 2.5 and finally, boundaries of rRNA genes and control regions were defined by the boundaries of adjacent coding genes.

Following these procedures, 18 groups were assembled being 14 of them ingroups and 4 outgroups. Assisted by the program Gblocks multiple alignments were analyzed which were later used as a backbone to align the corresponding nucleotide sequences. For the phylogenetic analyses, the team used maximum likelihood, Bayesian, maximum parsimony, and neighbour—joining methods and PAUP 4.0, Modeltest 3.7, MrBayes3.1, and PUZZLE 5.2 programs as the corresponding programs for each method previously mentioned.

In conclusion to all the process, a few statements could be done as well as understanding that for the general objective related to the mitochondrial genomes of snakes more samples and analyses need to be gathered and done to refine the phylogenetic relationships between major groups. A quick mention of the fact that the monophyly of Scolecophidia wasn´t fully rejected during the study. However some ideas were also well supported by the analyses such as the placement of Enhydris plumbea outside of the (Colubridae + Elapidae) cluster, the gene arrangement in ramphotyphlops braminus mtDNA was identical to that found in typical vertebrates, meaning an ancestral arrangement and finally, all the data helped to reconstruct the evolution of mitochondrial gene arrangements, in snakes, which was the main objective of the study even though, as mentioned it could use some more refining.

# Materials and methods

This assignment is a replication of the original paper "Evolution of the mitochondrial genome in snakes: Gene rearrangements and phylogenetic relationships Jie Yan, Hongdan Li and Kaiya Zhou". In this assignment, we didn't extract the DNA from the organism. On the other hand, we acquired the mitochondrial genomes from NCBI. NCBI is The National Center for Biotechnology Information (NCBI) is part of the United States National Library of Medicine (NLM), a branch of the National Institutes of Health (NIH). The NCBI houses a series of databases relevant to biotechnology and biomedicine and is an important resource for bioinformatics tools and services. Major databases include GenBank for DNA sequences and PubMed, a bibliographic database for biomedical literature. Other databases include the NCBI Epigenomics database. All these databases are available online through the Entrez search engine (Wikipedia, 2020). From NCBI, we were able to collect the mitochondrial genomes needed for the replication of the paper once referred, visible in figure1 and figure2. For the search on NCBI, we have searched for the GenBank Accession no. of the organisms, available in the original paper and downloaded the mitochondrial genomes to our Virtual Machine (VM) with Ubuntu. The names of the organisms and their accession numbers are in a table (Table 1). From there we started the funny part of the project. After we assembled all mitochondrial genomes needed, we move forward to the alignment of the sequences. For this procedure, we have used the tool, MAFFT. MAFFT (for multiple alignments using fast Fourier transform) is a program used to create multiple sequence alignments of amino acid or nucleotide sequences. First, the typed "maft", in the terminal, to open the alignment program MAFFT. We also needed to insert the input file with the genome sequences of our organisms (not aligned) and the output file to store our sequences after the alignment done by MAFFT (figure 3). After we aligned the sequences, we moved to the next step, visualize the data aligned. For the visualization, we have used Aliview. AliView is yet another alignment viewer and editor, but this one is probably one of the fastest and most intuitive to use, less bloated, and hopefully to your liking. The general idea when designing this program has always been usability and speed, all new functions are optimized so they do not affect the general performance and capability to work swiftly with large alignments (andersla, n.d.). As can be observed in figure4, we can observe the aligned genome sequences. After visualized our aligned sequences, in Aliview, we moved to the construction of phylogenetic trees. For this, we have used 2 different tools, Mega X, and Raxml. Molecular Evolutionary Genetics Analysis (MEGA) is computer software for conducting statistical analysis of molecular evolution and for constructing phylogenetic trees. It includes many sophisticated methods and tools for phylogenomics and phylomedicine. It is licensed as proprietary freeware (Wikipedia, 2019). RAxML-HPC (randomized accelerated maximum likelihood for high-performance computing) is a sequential and parallel program for the inference of large phylogenies with maximum likelihood (ML). Low-level technical optimizations, a modification of the search algorithm, and the use of the GTR+CAT approximation as a replacement for GTR+Γ yield a program that is between 2.7 and 52 times faster than the previous version of RAxML (bwHPC Wiki, 2016). We have used the tool, Mega

X for the construction of Maximum Parsimony (MP) and neighbour-joining (NJ). On the other hand, RAxML was used to build the Maximum Likehood tree. We have defined "N" as missing data, "-" as alignment gap, and "." as identical Symbol, as we can see in figure 5. With sequences, we obtained the trees represented in figure 6 and figure 7. In the RAxML we have used command to obtain the desire tree (Maximum Likehood). The command can be sawed in block 8. We first call the directory (cd) where our aligned sequences were. Then we inserted the command shown in block 8. In order to execute the RAxML we have to use ./raxmlHPC in Linux environment. The "-f a" tell RAxML to conduct a rapid Bootstrap analysis and search for the best-scoring ML tree in one single program run. The GTRCAT approximation is a computational workaround for the widely used General Time Reversible model of nucleotide substitution under the Gamma model of rate heterogeneity. CAT servers the analogous purpose, that is, to accommodate searches that incorporate rate heterogeneity (Stamatakis, 2016). The main idea behind GTRCAT is to allow for the integration of rate heterogeneity into phylogenetic analyses at a significantly lower computational cost (about 4 times faster) and memory consumption (4 times lower). Essentially, GTRCAT represents a rather un-mathematical quick & dirty approach to rapidly navigate into portions of the tree space, where the trees score well under GTRGAMMA (Stamatakis, 2016). The only thing that will be extracted from the string passed via "-m" is the model of rate heterogeneity you want to use. The "-s", specify the name of the alignment data file which can be in a relaxed PHYLIP format. Relaxed means that you don't have to worry if the sequence file is interleaved or sequential and that the taxon names are too long (Stamatakis, 2016). And finally, but not least, the "-n" specifies the name of the output file. This option must be always specified. The arbitrary name passed via n will be appended to all RAxML output files such that you know which files have been generated by which invocation (Stamatakis, 2016).

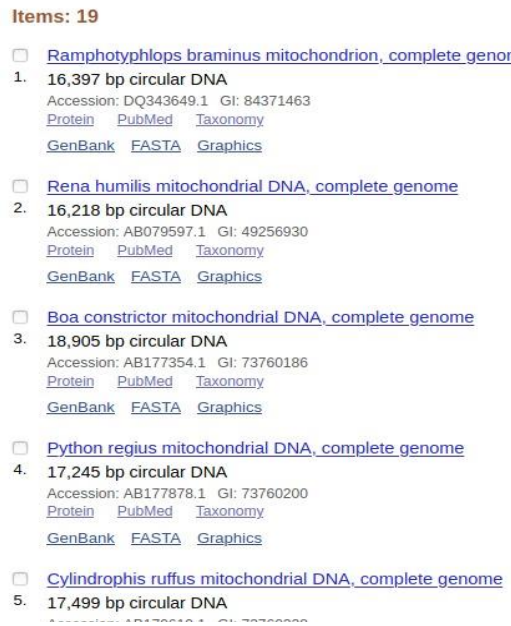


*Figure 1*

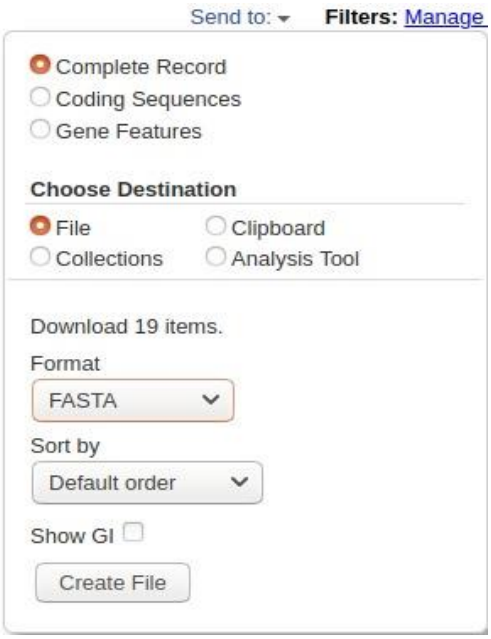


*Figure 2*

| Family | Species | GenBank Accession no. | Reference |
|---|---|---|---|
| **Ingroup** | | | |
| Scolecophidia | | | |
| Typhlopidae | Ramphotyphlops braminus | DQ343649 | Jie Yan, Hongdan Li and Kaiya Zhou, 2008 |
| Leptotyphlopidae | Leptotyphlops dulcis | AB079597 | Jie Yan, Hongdan Li and Kaiya Zhou, 2008 |
| Alethinophidia | | | |
| Henophidia | | | |
| Boidae | Boa constrictor | AB177354 | Jie Yan, Hongdan Li and Kaiya Zhou, 2008 |
| Pythonidae | Python regius | AB177878 | Jie Yan, Hongdan Li and Kaiya Zhou, 2008 |
| Cylindrophiidae | Cylindrophis ruffus | AB179619 | Jie Yan, Hongdan Li and Kaiya Zhou, 2008 |
| Xenopeltidae | Xenopeltis unicolor | AB179620 | Jie Yan, Hongdan Li and Kaiya Zhou, 2008 |
| Caenophidia | | | |
| Colubridae | Dinodon semicarinatus | AB008539 | Jie Yan, Hongdan Li and Kaiya Zhou, 2008 |
| | Pantherophis slowinskii | DQ523162 | |
| Elapidae | Naja naja | DQ343648 | Jie Yan, Hongdan Li and Kaiya Zhou, 2008 |
| Homalopsidae | Enhydris plumbea | DQ343650 | Jie Yan, Hongdan Li and Kaiya Zhou, 2008 |
| Viperidae | Deinagkistrodon acutus | DQ343647 | Jie Yan, Hongdan Li and Kaiya Zhou, 2008 |
| | Ovophis okinavensis | AB175670 | |
| | Agkistrodon piscivorus | DQ523161 | |
| Acrochordidae | Acrochordus granulatus | AB177879 | Jie Yan, Hongdan Li and Kaiya Zhou, 2008 |
| **Outgroup** | | | |
| Amphisbaenidae | Amphisbaena schmidti | AY605475 | Jie Yan, Hongdan Li and Kaiya Zhou, 2008 |
| Scincidae | Eumeces egregius | NC_000888 | Jie Yan, Hongdan Li and Kaiya Zhou, 2008 |
| Iguanidae | Iguana iguana | AJ278511 | Jie Yan, Hongdan Li and Kaiya Zhou, 2008 |
| Varanidae | Varanus komodoensis | AB080275/AB080276 | Jie Yan, Hongdan Li and Kaiya Zhou, 2008 |

*Table 1: List of taxa used in this assignment*

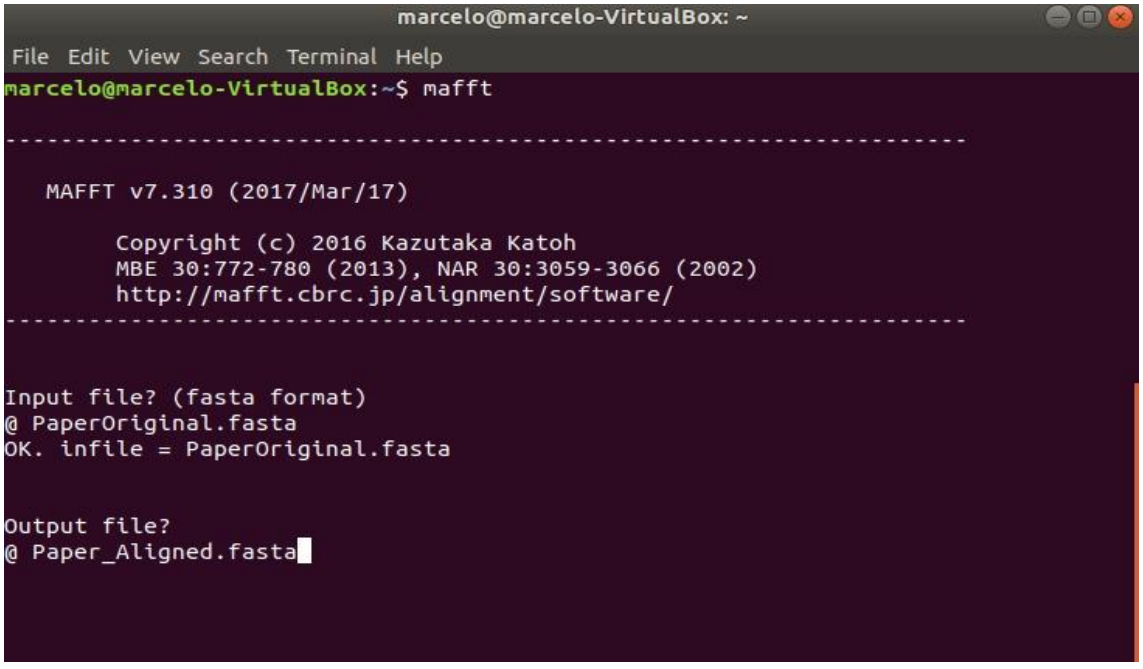


*Figure 3*

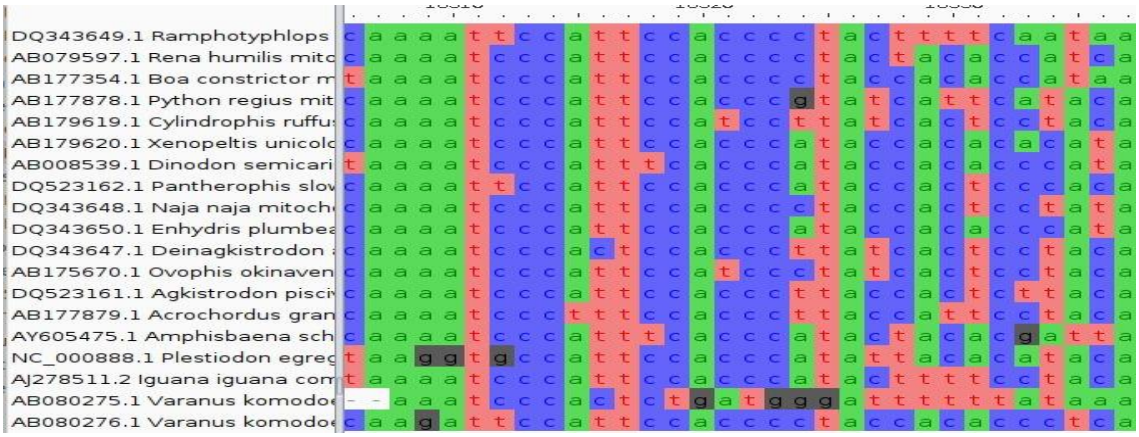


*Figure 4*

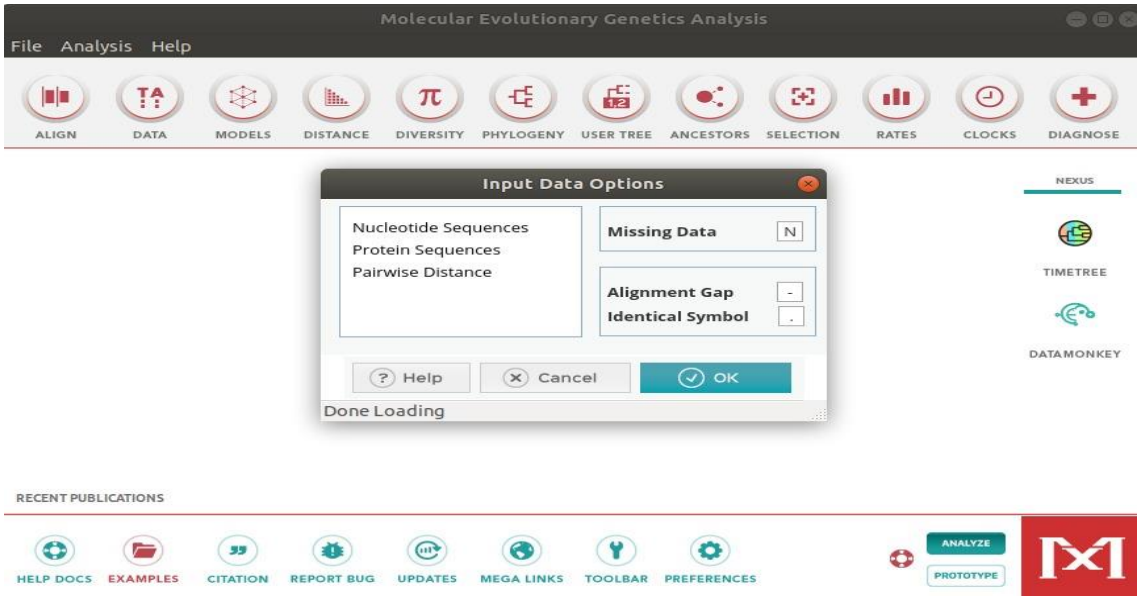


*Figure 5*

# Results

The obtained trees we obtained are strongly supporting Ramphotyphlops being the sister's group to the remaining ingroup. The scolecophidans do not form a monophyletic group. Alethinnophidia formed a big monophyly group. Henophidia and Caenophidia are sisters group as well.

The Viperidae group may have some divergence to the Colubridea group, but it forms a sister clade with Elapide and Homalopsidae.

Another aspect that our tree supports is the position of the Enhydris, because it forms a new group (Homalopsidae), but traditionally it places in the Colubridae group.

Our BI tree is very well supported with the boost strap values obtained.

All threes are based just in the mitochondrial DNA of the snakes, and none was based on the morphology.

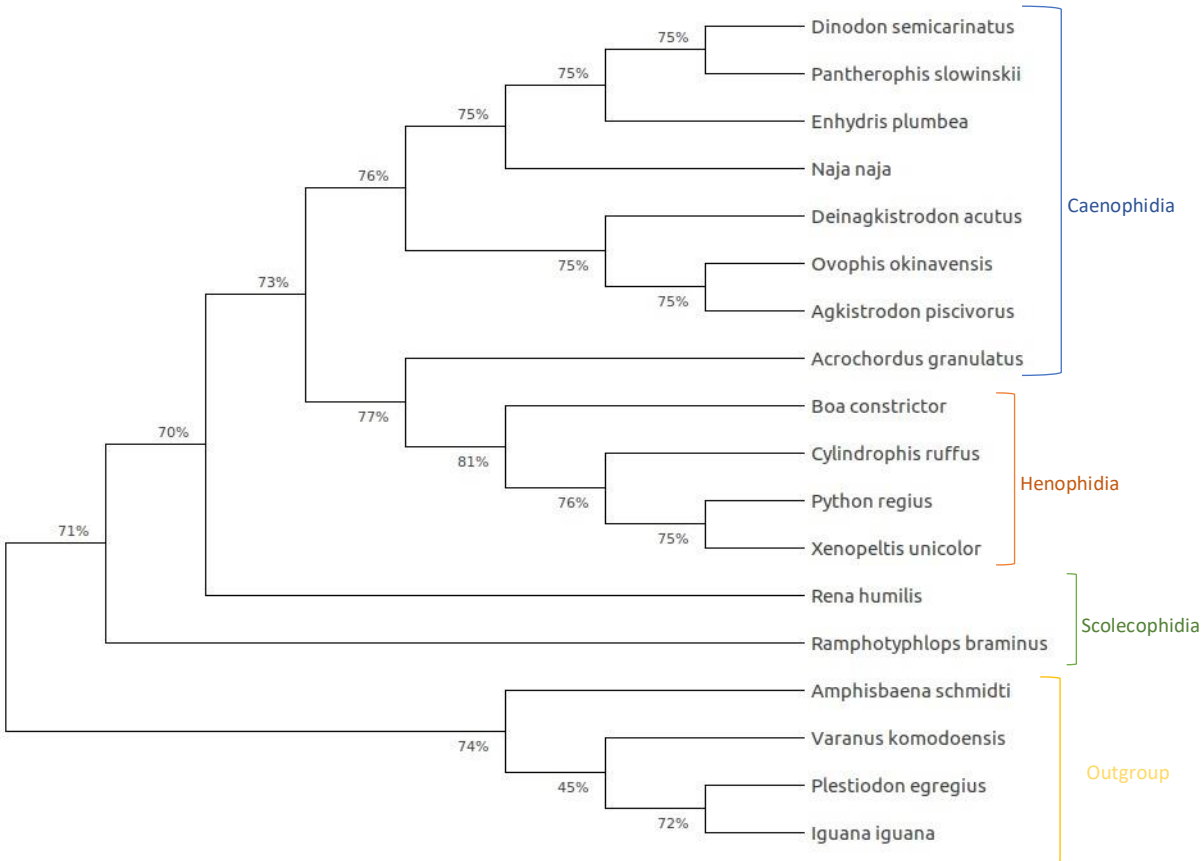


*Figure 6 Neighbor joining (MegaX)*

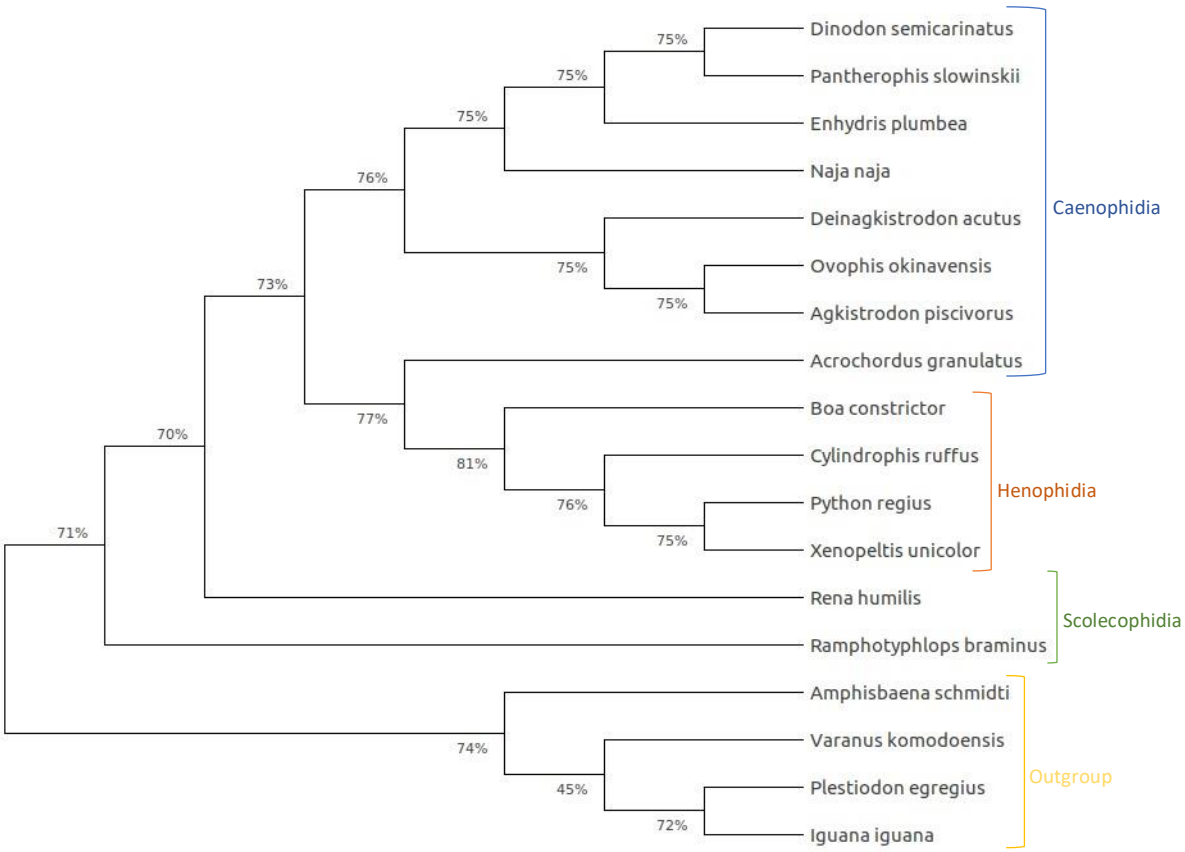


*Figure 7 Maximum Parimony Tree (MegaX)*

```
./raxmlHPC -SSE3 -f a -m GTRCAT -p 112358 -N 100 -s /home/marcelo/PaperAligned.fasta -n raxmlPaper.fasta
```

*Figure 8*

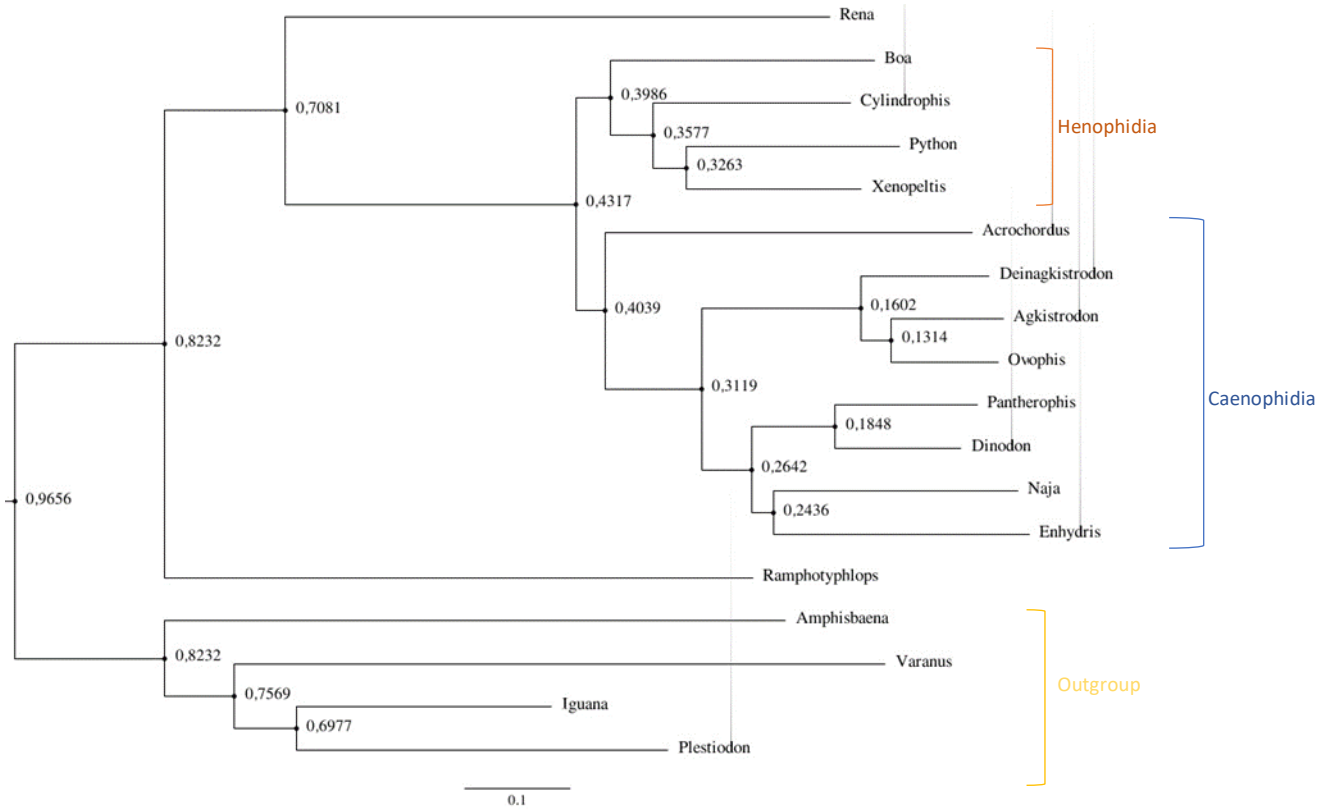


*Figure 9 Maximum likelyhood tree (MegaX)*

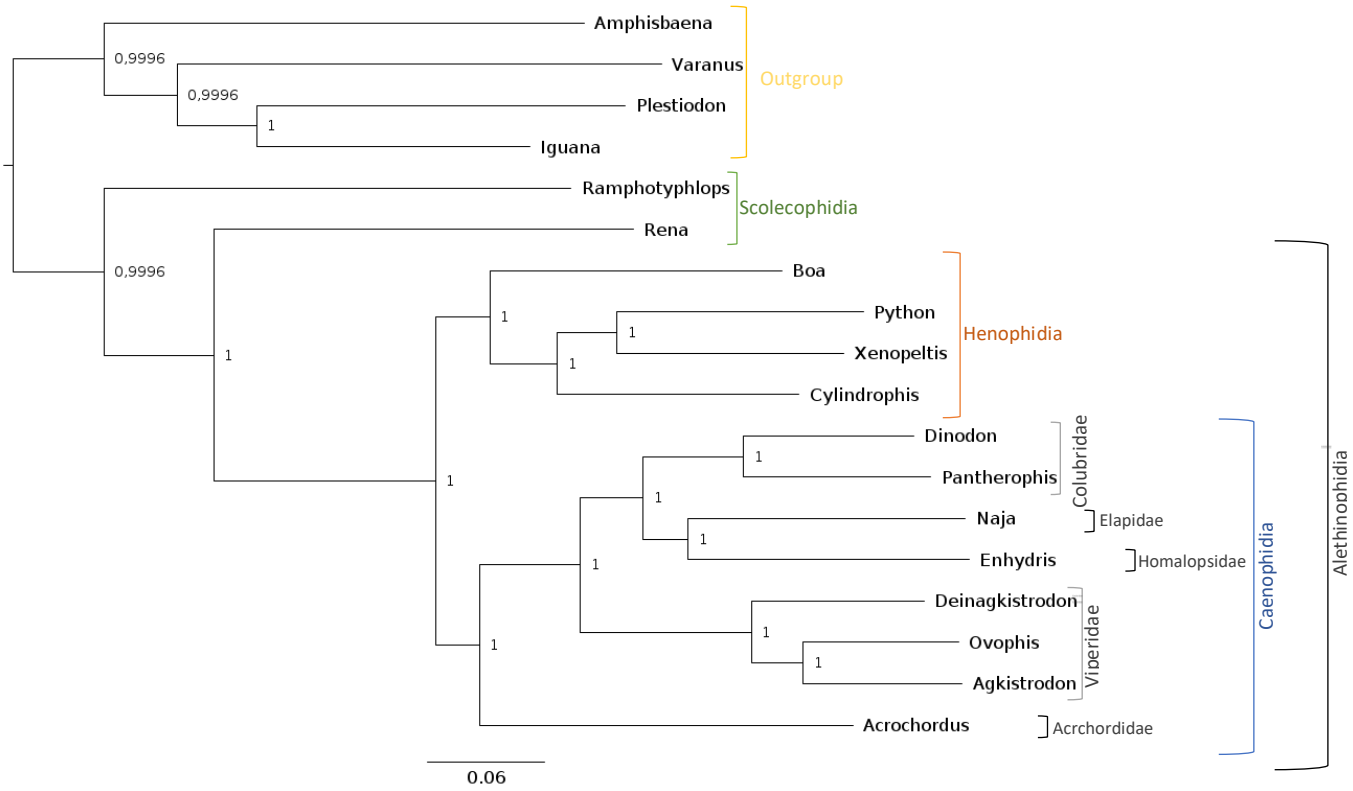


*Figure 10 BI tree (Figtree)*

# Discussion

The main issue in this paper was to understand better the evolution of the snake's mitochondrial genome, studying the gene organizations and the phylogenetic relationships.

We gave our focus on the phylogenetic trees and their relationships.

The tree we obtained is like the one in the paper, with just one difference in the Elapidae group and the Acochordidae group, wherein the paper's tree they are not coming from the same node as in the tree we obtained. Other than that, everything is equal, we got slightly better bootstrap values. But both trees help to understand the evolution of the snakes. If we analyze, we can see and make a conclusion where each group comes from, dividing in the main node of the ingroup, the Scolecophidia, and the Alethinophidia.

In the Alethinophidia group, we can see that Henophidia and Caenophidia were sister clades, and Caenophidia itself evolved and diverged into other groups such as Viperidae, Homalopsidae, Colubridae,Elapidae, and Acrochordidae.

# Bibliography

andersla. (n.d.). Retrieved from https://github.com/AliView/AliView

*bwHPC Wiki*. (2016, March 22 ). Retrieved from RAxML: https://wiki.bwhpc.de/e/RAxML

Jie Yan, H. L. (2008, November 28). Evolution of the mitochondrial genome in snakes: Gene. *Evolution of the mitochondrial genome in snakes: Gene*, p. 7.

Stamatakis, A. (2016, July 20). *RAxML*. Retrieved from The RAxML v8.2.X Manual : https://cme.h-its.org/exelixis/resource/download/NewManual.pdf

*Wikipedia*. (2019, December 4 ). Retrieved from Molecular Evolutionary Genetics Analysis: https://en.wikipedia.org/wiki/Molecular_Evolutionary_Genetics_Analysis

*Wikipedia*. (2020, April 26 ). Retrieved from National Center for Biotechnology Information: https://en.wikipedia.org/wiki/National_Center_for_Biotechnology_Information