Dorian Desblancs
260722712

COMP 561
December, 2019

Tracing the Evolution of Cytochrome c Oxidase-Producing Genes in Cats: A Phylogenetic Study

I. Introduction:

"Genes are rarely about inevitability, especially when it comes to humans, the brain, or behavior. They're about vulnerability, propensities, tendencies."

— Robert M Sapolsky, *Why Zebras Don't Get Ulcers*

Elusive and mysterious, the Felidae family has been roaming this planet for longer than humans. Their lineage is diverse, and sprawls all continents and habitats. This report presents a phylogenetic study of three protein-coding genes present in the family's species today: MT-COX1, MT-COX2, and MT-COX3. These genes are at the heart of the production of cytochrome c oxidase, an enzyme essential for breathing in all mammalian species. In this project, ancestral reconstructions of all three genes were generated. These were then analyzed to uncover the mysteries behind their evolution through time, and more importantly their adaptations to various environments.

In this report, some background information is first presented. This section notably introduces the Felidae family, cytochrome c oxidase and its role in mammals, and the gathering of nucleotide sequences for phylogenetic study. From there, the methodologies behind ancestral reconstruction and sequence comparison are outlined. The results produced by these methods are then displayed, interpreted and discussed.

II. Background Information:

a. The Felidae Family

Felidae is a family of mammals comprised today of 41 species and commonly referred to as cats. The family is split into two subfamilies: Pantherinae, also known as big cats, and Felinae, also known as small cats. The former is comprised of one lineage and seven species, while the latter is comprised of seven lineages and 34 species [1].

Many factors have contributed to the evolution of the Felidae family through time. First, Pantherinae and Felinae are believed to have diverged approximately 11 million years ago [1]. After that, the rise and fall of see levels throughout the past 10 million years allowed the Pantherinae and Felinae ancestors to travel away from their home continent Asia [2]. The species first migrated to Africa, then North America, and finally South America around 9 million years ago in what is known as the first wave of Feline migrations [2]. This migration was due to unusually low sea levels, which gave rise to patches of land connecting Asia, the Americas, and Africa [2]. The second wave of migrations occurred between 1 and 4 million years ago, when sea levels dropped again to expose land bridging continents such as Asia and North America. These migrations resulted in a flourishing of subspecies, spread throughout the world.

In 2007, O'Brien and Johnson [2] published a phylogenetic tree based on the Felidae family's DNA sequences and fossils (Figure 1). At the time, the Felidae family was comprised of 37 species. The phylogeny provided in the paper was used throughout this project.

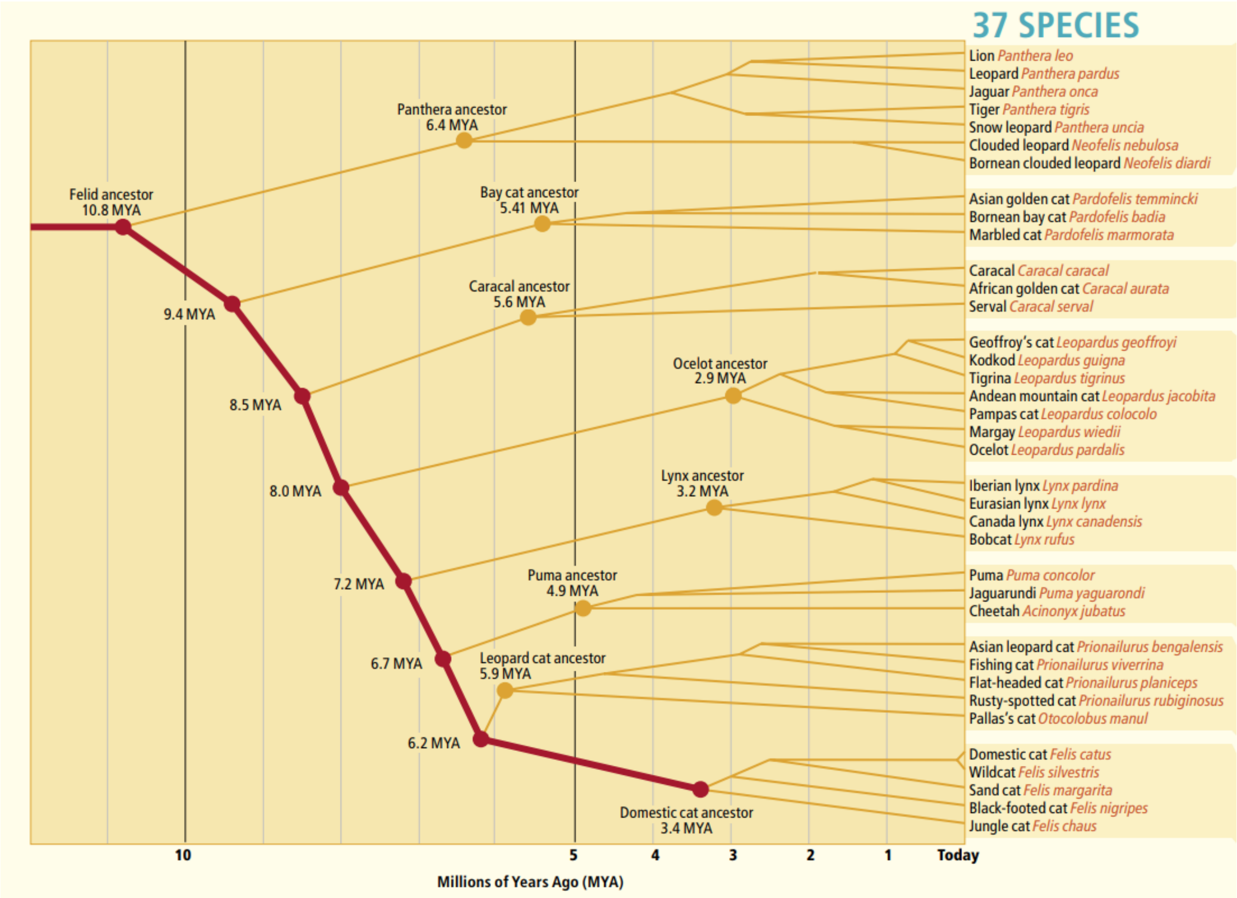
37 SPECIES
Lion Panthera leo
Leopard Panthera pardus
Jaguar Panthera onca
Tiger Panthera tigris
Snow leopard Panthera uncia
Clouded leopard Neofelis nebulosa
Bornean clouded leopard Neofelis diardi
Asian golden cat Pardofelis temmincki
Bornean bay cat Pardofelis badia
Marbled cat Pardofelis marmorata
Caracal Caracal caracal
African golden cat Caracal aurata
Serval Caracal serval
Geoffroy's cat Leopardus geoffroyi
Kodkod Leopardus guigna
Tigrina Leopardus tigrinus
Andean mountain cat Leopardus jacobita
Pampas cat Leopardus colocolo
Margay Leopardus wiedii
Ocelot Leopardus pardalis
Iberian lynx Lynx pardina
Eurasian lynx Lynx lynx
Canada lynx Lynx canadensis
Bobcat Lynx rufus
Puma Puma concolor
Jaguarundi Puma yaguarondi
Cheetah Acinonyx jubatus
Asian leopard cat Prionailurus bengalensis
Fishing cat Prionailurus viverrina
Flat-headed cat Prionailurus planiceps
Rusty-spotted cat Prionailurus rubiginosus
Pallas's cat Otocolobus manul
Domestic cat Felis catus
Wildcat Felis silvestris
Sand cat Felis margarita
Black-footed cat Felis nigripes
Jungle cat Felis chaus
Felid ancestor 10.8 MYA
Panthera ancestor 6.4 MYA
Bay cat ancestor 5.41 MYA
Caracal ancestor 5.6 MYA
Ocelot ancestor 2.9 MYA
Lynx ancestor 3.2 MYA
Puma ancestor 4.9 MYA
Leopard cat ancestor 5.9 MYA
Domestic cat ancestor 3.4 MYA
9.4 MYA
8.5 MYA
8.0 MYA
7.2 MYA
6.7 MYA
6.2 MYA
10
5
4
3
2
1
Today
Millions of Years Ago (MYA)

Figure 1: The Cat Family Tree [2]

b. Cytochrome c oxidase

In mammals, cytochrome c oxidase (COX) is a complex enzyme composed of 13 subunits [3]. Three of these subunits are encoded by mitochondrial DNA [3]. These three subunits are referred to as cytochrome c oxidase Subunit I (COX1), cytochrome c oxidase Subunit II (COX2), and cytochrome c oxidase Subunit III (COX3). The protein-coding genes encoding these three subunits are referred to as MT-COX1, MT-COX2, and MT-COX3 in this paper. The remaining subunits are encoded within the nuclear genome [5].

Cytochrome c oxidase is a key element of the aerobic metabolism [4]. Aerobic organisms generate energy through a respiratory chain that uses oxygen as the terminal acceptor of electrons [4]. Cytochrome c oxidase is the enzyme that catalyzes the reduction of oxygen to its molecular form in breathing systems, allowing the production of energy [4]. It is therefore an essential component of the respiratory system in mammals.

COX1 is the largest and most important subunit. The reduction of oxygen to its molecular form takes place in this subunit. COX2 also plays a major role in the respiratory chain. It provides the substrate-binding site for cytochrome c subunit I, allowing the latter to catalyze its reaction with oxygen [5]. Finally, the third subunit of cytochrome c oxidase remains somewhat of a mystery. García-Horsman et al. [5] outline that the subunit likely has some features that are different from the other two subunits encoded by DNA as it is not believe to play a primary role in the reaction of oxygen to its molecular form [6]. The third subunit is also absent from a number of species, though present in the Felidae family.

One notable condition associated with cytochrome c oxidase in mammals is a metabolic disorder associated with COX deficiency [7]. Deficiency of COX often leads to an inability to produce skeletal muscle tissues, and sometimes even cognitive tissue. Most notably, Leigh's disease is due to a COX deficiency in the brain causing developmental delays in infants. Cytochrome c oxidase deficiencies are most often inherited. In rare cases, these are the results of mutations in a mitochondrial gene.

This project provides a phylogenetic analysis of the three genes encoding subunits I, II, and III in the Felidae family of mammals.

c. Gene Collection

All the gene sequences used for this project were collected from the NCBI database [8]. These were found by cycling through all cat species and extracting the nucleotide sequences associated with cytochrome c oxidase subunits I, II, and III production. These sequences were then exported to FASTA and NEXUS files for later analysis[1].

Note that the following four species were not present in the database[2]:

- Caracal Aurata (African Golden Cat).
- Caracal Serval (Serval).
- Neofelis Diardi (Bornean Clouded Leopard).
- Lynx Pardina (Iberian Lynx).

To account for the missing sequences, a revised phylogenetic tree was generated (Figure 2). Note that all branch lengths are equal, and node numbers denote ancestral species' names in this project.

III. Methodology:

a. Ancestral Sequence Generation

All ancestral sequences in the tree nodes of Figure 2 were generated using Mesquite [9]. After opening a NEXUS file containing nucleotide sequences of each gene for each species and their phylogeny in NEWICK form, Mesquite's 'Trace All Characters' feature was used to output the most parsimonious ancestral states at each node[3].

In this project, the ancestral state reconstruction method used was maximum parsimony. Each character in each node is associated with the nucleotide(s) that outputs the maximum parsimony score for the whole input tree. In the future, using Maximum Likelihood ancestral state reconstruction could also be used for more precise ancestral reconstruction (see Discussion).

a. Sequence Comparison

Since the ancestral sequences used in this project were generated using maximum parsimony, multiple nucleotides can sometimes suit a particular location of an ancestral sequence (for example, in the ancestral sequence 10, character 1495 can either be an A, a G, or a T without affecting the maximum parsimony score of the tree overall).

[1] These files can be found in the 'Nexus Files' and the 'Species Fasta Files' directories.

[2] A complete list of the species and the presence of MT-COX1, MT-COX2, and MT-COX3 nucleotide sequences in the NCBI Database can be found in the 'Species List and Gene Presence Summary' directory.

[3] The ancestral sequences generated and used for this project can be found in Excel form in the 'Ancestral Sequences' directory.

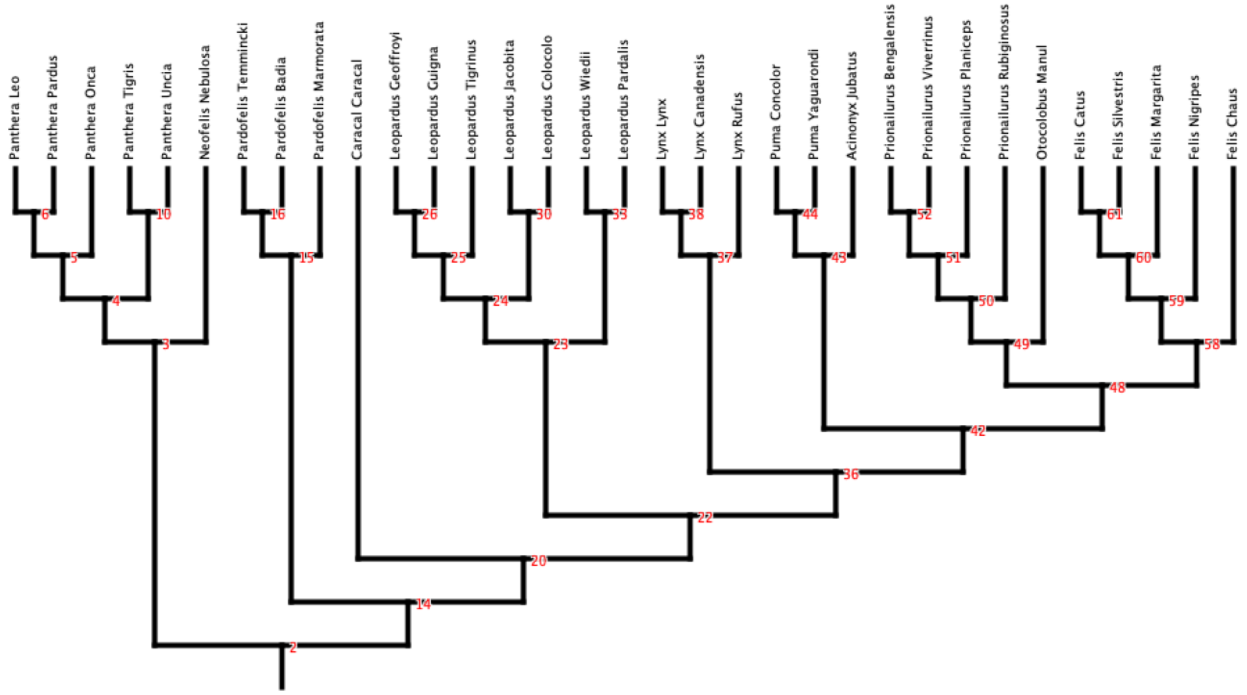

Figure 2: An Adapted Phylogeny of the Felidae Family

In order to compare any ancestral sequence to another sequence, regular or ancestral, on could generate all possible sequences of the latter and compare them to the former. This would however be extremely time-consuming and the results would be hard to interpret.

Instead, a more efficient method based on the Needleman-Wunsch dynamic program [10] was used. The algorithm is altered such that, for sequence characters that have more than one nucleotide option, one iterates through the set of possible states to generate the best score in the dynamic programming table. Hence, for the sequence 10 example outlined previously, one would compare the other sequence's nucleotide(s) with A, T, and G. All the possible match and mismatch, deletion, and insertion scores are hence computed. The maximal score is then used in the dynamic programming table (see Appendix for pseudocode).

Note that the Needleman-Wunsch trace back algorithm was not used or implemented. All one needs to get an idea of the similarity between two sequences is their optimal alignment score. The alignments themselves, however, were unused. The scoring parameters used for genetic analysis were: +5 for a match, -2 for a mismatch, and -30 for a gap (creation and extension). This scoring scheme was established in order to avoid the creation and extension of gaps as much as possible, as the genes studied in this project almost always have the same lengths throughout the 33 species.

It is important to note that the methodology used for sequence comparisons is far from ideal. Indeed, this is somewhat of a hacky way to get around the problem posed by multiple nucleotide possibilities in ancestral sequences. However, the Needleman-Wunsch alteration used provides a way of analyzing ancestral sequences in an unbiased fashion, without favoring one ancestral reconstruction over another. The scoring discrepancies observed throughout gene comparisons are hence the result of actual differences in the sequences analyzed and not the product of guesswork.

IV. Results:

Ancestral reconstruction of all three genes were generated for all nodes. Overall, as one moves up throughout the tree, the scores generated by the sequence comparison algorithm get more and more similar for all three genes. This is due to the fact that the maximum parsimony algorithm reconstructs sequences at each node that suit all sub-trees, and by extension all cats present at the leaves of the tree. Also, genes within a family of species are usually fairly similar, even after millions of years of evolution.

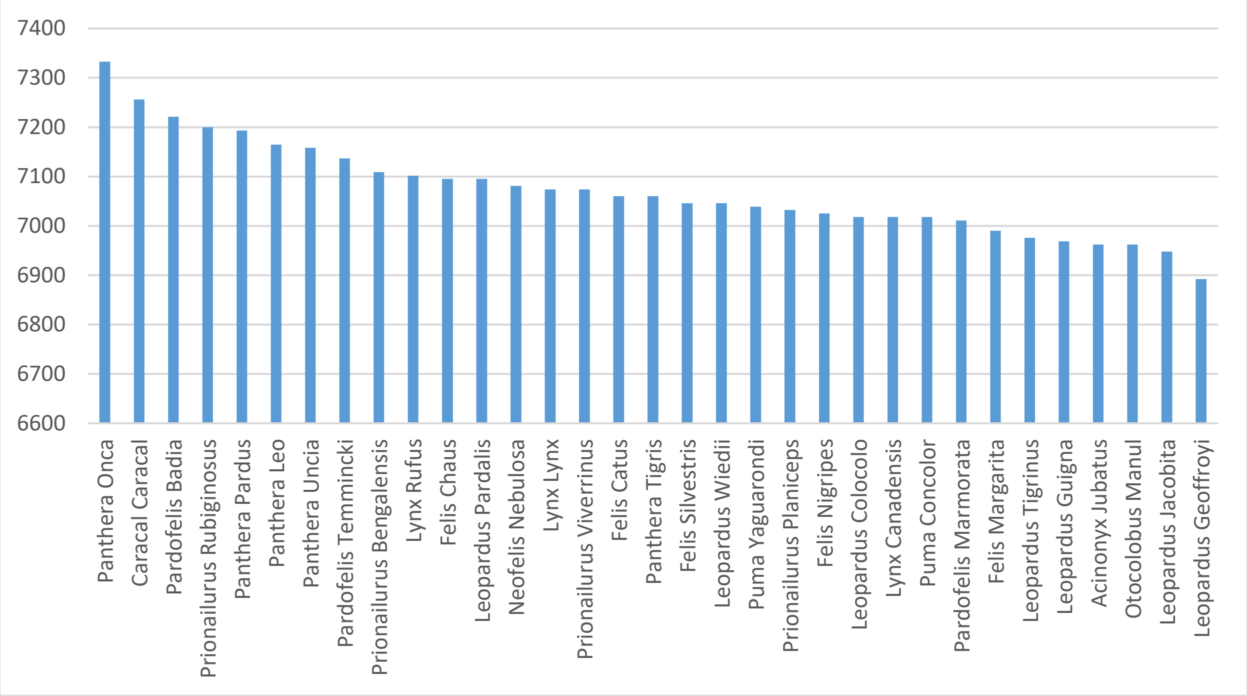

Figure 3: Global Alignment Scores for each Present-Day Species with Ancestor 2 (MT-COX1)

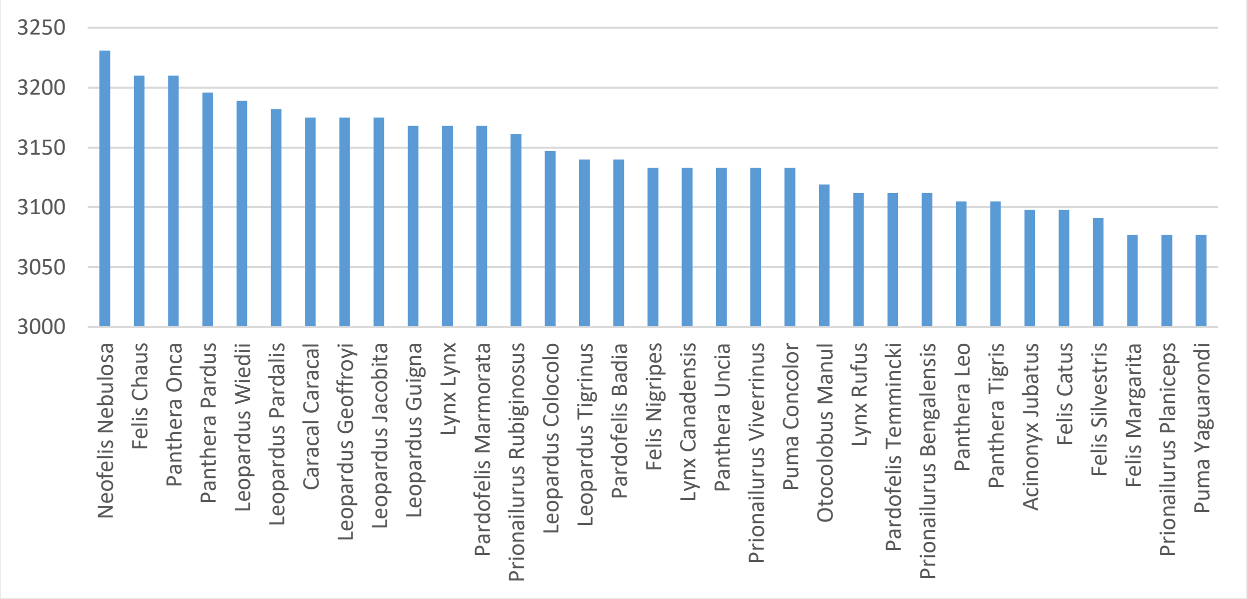

Figure 4: Global Alignment Scores for each Present-Day Species with Ancestor 2 (MT-COX2)

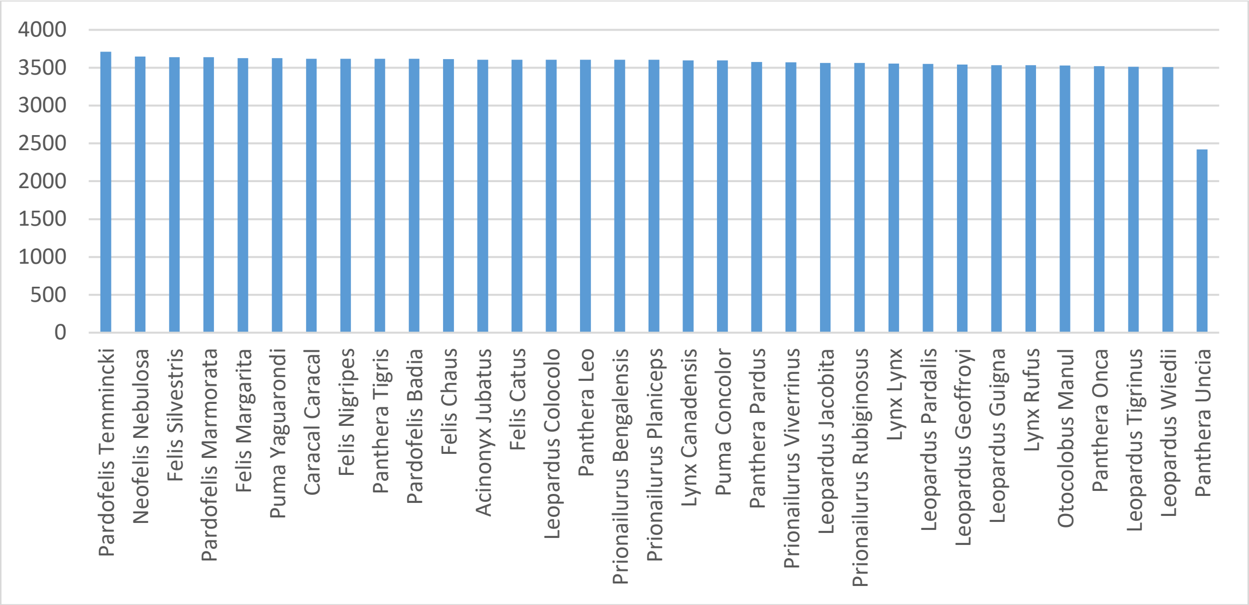

Figure 5: Global Alignment Scores for each Present-Day Species with Ancestor 2 (MT-COX3)

Figures 4, 5, and 6 outline the alignment scores of all present-day genes with those of the inferred ancestor of all cats (node 2 in Figure 2). The graphs are ordered from largest to smallest score for clarity. Results for all ancestral alignment scores with today's cats can be found in the 'Results' directory.

Assuming the ancestral reconstruction of the ancestor of all cats is somewhat accurate, one can clearly identify that cats today have gone through genetic mutations in their MT-COX1, MT-COX2, and MT-COX3 genes. These are more or less pronounced depending on the species. Most notably, when looking at the tables of scores, one cannot identify a clear separation between the Pantherinae and Felinae lineages. It seems that mutations involving cytochrome-c-oxidase-producing genes have more to do with environmental adaptations and less to do with ancestry. A more in-depth analysis of the environments of each cat is however needed to identify the surrounding factors that lead to genetic mutations in MT-COX1, MT-COX2, and MT-COX3.

The results also help identify one major finding. When comparing the scores of today's cats with ancestral reconstructions, one species particularly stands out for its MT-COX3 gene. This species is Panthera Uncia, or the Snow Leopard. Its score is vastly inferior to all other species, most likely due to its shortened MT-COX3 sequence. In 2015, Janecka et al. [11] demonstrated that Hemoglobin levels in Panthera Uncia were in fact almost identical to its Panthera Leo (lion) counterpart. They argued that the Snow Leopard's adaptation to high altitudes (up to 3500 meters) are due to other, unknown factors in its respiratory system. Though our knowledge on COX3 is somewhat limited, perhaps its mutation through time allowed the Snow Leopard to live in higher altitudes. A more in-depth analysis of the cytochrome c oxidase subunit III is needed to answer this question. It is however very clear that Panthera Uncia's MT-COX3 differs greatly from other species in the Felidae family, and this may not be a coincidence since it is the only cat to live in altitudes where oxygen levels are much lower.

V. Discussion and Future Work:

This project gave me the opportunity to generate ancestral reconstructions of the MT-COX1, MT-COX2, and MT-COX3 genes. The sequences generated are of interest since they hypothesize how cytochrome c oxidase producing genes evolved through time, in a large family of species. The techniques used in this paper can however be generalized to all kinds of starting sequences, from full genomes to individual genes affecting other parts of a system's being.

A few of the techniques used should however be refined. Most notably, inferring ancestral sequences by maximum likelihood would be a huge step forward [12]. This method of ancestral reconstruction is often more accurate, and leads to a much clearer idea of how sequences evolved through time. This would be beneficial for sequence comparison too, as maximum likelihood reconstructions are usually narrower (not as many states with multiple nucleotide options) [12]. Hence, the Needleman-Wunsch algorithm implemented would be closer to the regular one, and the scores generated would be more precise with regards to the actual similarity of modern species with their ancestors.

Also, translating the original nucleotide sequences to amino acid sequences, and then inferring ancestral amino acid sequences would most probably give a clearer idea of why mutations in MT-COX1, MT-COX2, and MT-COX3 happened through time. It is much easier to identify a particular amino acid's role in respiration since its bio-chemical structure can be analyzed to show its affinity with molecules such as oxygen. Hence, amino acid reconstruction could help narrow down the roles of all three COX subunits, especially when comparing their evolutionary adaptations to environments.

Finally, statistical methods such as correlation could be used to identify whether mutations in COX-encoding genes are the product of environmental parameters such as oxygen levels. Correlation global alignment scores with, say, oxygen levels of a species' environment could help identify the reasons behind COX mutations. The results could then be used to gain information about the environments of each continent over the past 11 million years, and further prove or disprove geological hypotheses about our planet's history.

VI. Conclusion:

This project presents a methodology for ancestral sequence generation and comparison in the Felidae family. Its results outline mutations through time that are mostly due to environment. These have very little to do with direct ancestry, as demonstrated with global alignment scores. Its results also show a lot more mutations in the MT-COX3 gene of Panthera Uncia. These could help answer the questions behind its adaptation to high altitudes. Overall, more work needs to be done to analyze the evolutionary reconstructions of MT-COX1, MT-COX2, and MT-COX3. However, one can hope that this project provides a framework for a more in-depth analysis of the genomes of cats and the world that surrounded them in the past 11 million years.

Bibliography:

[1] Kitchener A. C., Breitenmoser-Würsten Ch., Eizirik E., Gentry A., Werdelin L., Wilting A., Yamaguchi N., Abramov A. V., Christiansen P., Driscoll C., Duckworth J. W., Johnson W., Luo S.-J., Meijaard E., O'Donoghue P., Sanderson J., Seymour K., Bruford M., Groves C., Hoffmann M., Nowell K., Timmons Z. & Tobe S. 2017. A revised taxonomy of the Felidae. The final report of the Cat Classification Task Force of the IUCN/ SSC Cat Specialist Group. Cat News Special Issue 11, 80 pp.

[2] O'Brien, S., & Johnson, W. (2007). The Evolution of CATS. Scientific American, 297(1), 68-75.

[3] Dimauro, Salvatore, et al. "Cytochrome c Oxidase Deficiency." *Pediatric Research*, vol. 28, no. 5, 1990, pp. 536–541.

[4] Castresana, J et al. "Evolution of cytochrome oxidase, an enzyme older than atmospheric oxygen." *The EMBO journal* vol. 13,11 (1994): 2516-25.

[5] García-Horsman, J A et al. "The superfamily of heme-copper respiratory oxidases." *Journal of bacteriology* vol. 176,18 (1994): 5587-600. doi:10.1128/jb.176.18.5587-5600.1994

[6] Haltia, T., M. Saraste, and M. Wikstrom. 1991. Subunit III of cytochrome c oxidase is not involved in proton translocation: a site-directed mutagenesis study. EMBO J. 10:2015-2021.

[7] NORD (National Organization for Rare Disorders). (2019). *Cytochrome C Oxidase Deficiency - NORD (National Organization for Rare Disorders)*. [online] Available at: https://rarediseases.org/rare-diseases/cytochrome-c-oxidase-deficiency/

[8] National Center for Biotechnology Information (NCBI)[Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – [cited 2017 Apr 06]. Available from: https://www.ncbi.nlm.nih.gov/

[9] Maddison, W. P. and D.R. Maddison. 2018. Mesquite: a modular system for evolutionary analysis. Version 3.51 http://www.mesquiteproject.org

[10] Needleman, Saul B., and Christian D. Wunsch. "A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins." *Journal of Molecular Biology*, vol. 48, no. 3, 1970, pp. 443–453., doi:10.1016/0022-2836(70)90057-4.

[11] Janecka JE, Nielsen SS, Andersen SD, Hoffmann FG, Weber RE, Anderson T, Storz JF, Fago A. Genetically based low oxygen affinities of felid hemoglobins: lack of biochemical adaptation to high-altitude hypoxia in the snow leopard. J Exp Biol. 21815:2402–2409. 2015.

[12] Elliott Sober, The Contest Between Parsimony and Likelihood, *Systematic Biology*, Volume 53, Issue 4, August 2004, Pages 644–653, https://doi.org/10.1080/10635150490468657

Appendix:

```
$d = Gap\ penalty$
$for\ i = 0\ to\ length(A)$
    $F(i, 0) = d * i$
$for\ j = 0\ to\ length\ (B)$
    $F(i, 0) = d * i$
$for\ i = 1\ to\ length(A)$
    $for\ j = 1\ to\ length(B)$
        $if\ (length(A_i) == 1\ and\ length(B_i) == 1)$
            $Match = F(i-1, j-1) + S(A_i, B_j)$
            $Delete = F(i-1, j) + d$
            $Insert = F(i, j-1) + d$
            $F(i, j) = \max\ (Match, Insert, Delete)$
        $else$:
            $score = -\infty$
            $for\ l1 = 1\ to\ length(A_i)$
                $for\ l2 = 1\ to\ length\left(B_j\right)$
                    $Match = F(i-1, j-1) + S(A_i[l1], B_j[l2])$
                    $Delete = F(i-1, j) + d$
                    $Insert = F(i, j-1) + d$
                    $scoretemp = \max\ (Match, Insert, Delete)$
                    $if\ (scoretemp > score)$
                        $score = scoretemp$
            $F(i, j) = score$
```

[1] A Python implementation can be found in the 'sequence_comparison.py' file.