IPS Instituto Politécnico de Setúbal
Escola Superior de Tecnologia do Barreiro

BEATRIZ OLIVEIRA DE LA FUENTE DOS SANTOS

POSITIVE SELECTION IN CICHLIDS

Internship Report
Bachelor's in Bioinformatics

ADVISOR

Professor, Francisco Pina Martins

SUPERVISORS

Jingtao Lilue, Rui Oliveira

June 2022

Acknowledgment

Firstly, I would like to thank my internship advisor, Francisco Pina Martins, from Escola Superior de Tecnologia do Barreiro, for the support and availability during the internship period.

Thank you, Rui Oliveira and Jingtao Lilue, for being my supervisores and welcomed me into Instituto Gulbenkian de Ciência, and also for the trust and support you have placed in me.

A special thanks to my colleague, Pol Sorigué, for including me in your research project, trust in me for it's development and for always showing availability to help.

Last but not least, I would also like to thank all my colleagues from the Bioinformatics Unit, for including me in the team group and helped me during the internship.

Index

Figure Index

Table Index

List of Acronyms and Abbreviations

aBSREL - Adaptive branch-site random effects likelihood

BLAST – Basic Local Alignment Search Tool

DNA – Deoxyribonucleic acid

IGC – Instituto Gulbenkian de Ciência

MCMC – Markov Chain Monte Carlo

ML – Maximum likelihood

MP – Maximum parsimony

MRCA – Most recent common ancestor

mRNA – messenger RNA

MSA – Multiple Sequence Analysis

NCBI – National Center for Biotechnology Information

OT or oxt – Oxytocin

OTU – Operational Taxonomic Unit

RAxML – Randomized Axelerated Maximum Likelihood

RNA – Ribonucleic acid

rRNA – ribossomic RNA

tRNA – transfer RNA

Abstract

In this report, all the activities developed by me, Beatriz Oliveira de La Fuente dos Santos, a student in Bioinformatics from Instituto Politécnico de Setúbal and an intern at Instituto Gulbenian de Ciência are precisely described.

The activities were developed in the Bioinformatics Unit with the main goal of finding the evidence of positive selection of behavior related genes in Lake Tanganyika cichlids. The construction of phylogenetic trees, the analysis of synonymous and non-synonymous substitutions, the analysis of the splicing of messenger RNA sequences and the development of a pipeline were methods used to achieve the objective of this project.

The results obtained demonstrate that most species of the same tribe group together in phylogenetic trees. In the analysis of synonymous and non-synonymous substitutions, it is observed that the synonymous substitutions occur more than non-synonymous substitutions and the ratio between them is smaller than 1. By the splicing analysis, it is verified that the sequences encode the same region, both in cichlids and in the reference specie. The results suggest that there is no evidence of positive selection.

The internship accomplishment was fundamental to my academic formation because it provided an excellent opportunity for the practical application of the theories studied in the Bioinformatics course. Highlighting some curricular units such as Biological Sequence Analysis, Laboratory of Bioinformatics, Structural and Evolutionary Genomics, Programming Languages, Databases and General Biology.

Key words: Cichlids; Phylogenetics; Positive Selection.

Introduction

The species of cichlid fishes from Lake Tanganyika in East Africa comprise the greatest diversity of cichlid fishes in terms of morphology, ecology and breeding styles. The origin of cichlids has been characterized by a process called adaptive radiation [1], which is believed to be induced by extrinsic environmental factors such as geological and climatic events and intrinsic genetic factors. Extrinsic factors interact with the biological characteristics of the species involved, such as specialization, site fidelity, territoriality, mating behavior and social organization ([2], [3], [4], [5]).

It is intended to understand how the brain and behavior can be shaped by extrinsic factors, for that, in this project, two neurotransmitters, oxytocin and vasotocin, and their respective receptors had been studied, which regulate a wide range of bio-functions such as the nesting behavior and social vocalizations. In the present project, responsibility was given for the study of the neurotransmitter oxytocin in cichlid fish. OXT-OXTR signaling regulates the hypothalamic-pituitary-adrenal axis that modulates the behavioral response to stress and social behavior [6]. Oxytocin receptors belong to the G-protein-linked hetero-trimeric receptor family and can be found on various types of cells such as neurons, bone cells, myoblasts, cardiomyocytes and endothelial cells [7].

Taking into account the knowledge in the area of Bioinformatics, the internship project is aiming to study the evolutionary relationships of oxytocin paralogues in different species of cichlids, and looking for evidence of positive selection during the process of adaptive radiation. Phylogenetic trees were constructed to observe evidence of selection, analysis and statistical calculation of the ratio of synonymous and non-synonymous substitutions to observe which type of mutations occur most frequently, and analysis of splicing variation of RNA sequences in order to observe whether the cichlid sequences encode the same region as the reference sequence.

1 Internship Goals

The main objective of the internship was to study whether positive selection occurs in different species of cichlids at the level of the gene that encodes Oxytocin.

To achieve this objective, it was necessary to study:

- DNA sequences for phylogenetic analysis and determination of the ratio of synonymous and non-synonymous substitutions.
- Messenger RNA sequences to analyze splicing variation.
- Development of a pipeline to automate the analysis process.

2 Institute Characterization

Founded in 1961, the Instituto Gulbenkian de Ciência, which is part of the Calouste Gulbenkian Foundation, is dedicated to biological and biomedical research, innovative postgraduate training and the transformation of society through science.

The IGC has around 400 scientists, from 44 nationalities, dedicated to understanding the fundamental principles of Biology to explain how organisms are formed and interact with their environment, leading to new perspectives on how to treat diseases and how to promote a sustainable world.

Its organization is designed to have an open, stimulating and collaborative culture. It is organized into several groups, such as scientific research groups, scientific support units, management and technical support units.

The Bioinformatics Unit is included in the scientific support group that aims to provide technical and human resources to support research. This group has the latest technologies and highly qualified personnel, focused on providing the latest technical advances and developing innovative protocols. The Bioinformatics Unit team is composed of the Coordinator, Jingtao Lilue, three Bioinformaticians, Gonçalo Leiria, João Costa and Hugo Lainé, a Post-Doctorate, Francisco Cerqueira, and a Master's student, Paulo Peres. Their objective is to provide support on data analysis and computational biology issues to research groups inside and outside the IGC.

The main interest of Integrative Biology of Behavior scientific research group is the integrated study of social behavior, which combines the study of proximate causes and final effects. It is composed by the Principal Investigator, Rui Oliveira, and the members of the group that include post-doctoral and master's students, along with technicians. This group aims to understand how the brain and the behavior of the environment can be shaped by the social environment and how the cognitive mechanisms, neural and genetic underlying for plasticity in the expression of social behavior evolved.

The mission of Instituto Gulbenkian de Ciência is to face the global challenges of science, its vision is to be a top scientific research institution where individuals pose original scientific questions in a multidisciplinary, international and collaborative environment. Its values include independence, integrity and ethics, scientific excellence, flexibility and cooperation, generosity and responsibility.

3 Activities developed

In global terms, during the internship period, a collaborative project was developed with the research group of Integrative Biology of Behavior. I worked with PhD student Pol Sorigué to carry out his project where the main objective is to investigate the hypothesis of positive selection in cichlids and the evolutionary mechanisms underlying the appearance of mutations in the oxytocin gene. The main role in this project was to support data analysis using Bioinformatics tools and methods.

3.1 Internship Cronogram

The activities were developed according to the following schedule (Table 1).

	Weeks															
Activities	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
A																
B																
C																
D																
E																
F																
G																

Table 1 - Cronogram of the internship developed at Instituto Gulbenkian de Ciência, from March 1 to June 30, 2022.

Activity A: Obtaining data.

Activity B: Alignment of biological sequences.

Activity C: Construction of phylogenetic trees.

Activity D: Analysis of synonymous and non-synonymous substitutions.

Activity E: Precision Oncology Course.

Activity F: Construction of a pipeline for phylogenetic analysis and ratio of synonymous and non-synonymous substitutions.

Activity G: Analysis of splicing of messenger RNA sequences.

4 Theoretical Foundations

To fully understand this work, it is necessary to address some essential concepts of phylogenetics, general biology and computational biology in order to analyze whether positive selection occurs.

4.1 Phylogenetics

4.1.1 Concept

Phylogenetics is the set of methods for inferring relationships between operational taxonomic units (OTUs). It is used in taxonomy, molecular dating, molecular evolution and gene transfer detection. The evolutionary relationships between a group of organisms which are normally illustrated by phylogenetic trees.

The goals of phylogenetics include:

- Reconstructing the correct genealogy between biological entities.
- Estimate the divergence time between biological entities.
- Report the sequence of events along evolutionary lineages.

Within these goals, the most important one for this work was the reporting of events along evolutionary lineages in order to understand whether the Oxytocin gene was conserved over time or underwent mutations that alter the protein's function, thus offering information about the possible occurrence of positive selection and, in turn, alteration of social behavior.

Different types of data can be used to study phylogenetics, such as morphological traits, behavioral traits or molecular sequences [8].

4.1.2 Phylogenetic Trees

A phylogenetic tree is a tree-shaped diagram used to visualize the evolutionary relationships between a set of OTUs [8]. The OTU can represent species, individual organisms of a population, a gene or a protein sequence. The tree is composed of nodes and branches (Figure 1).

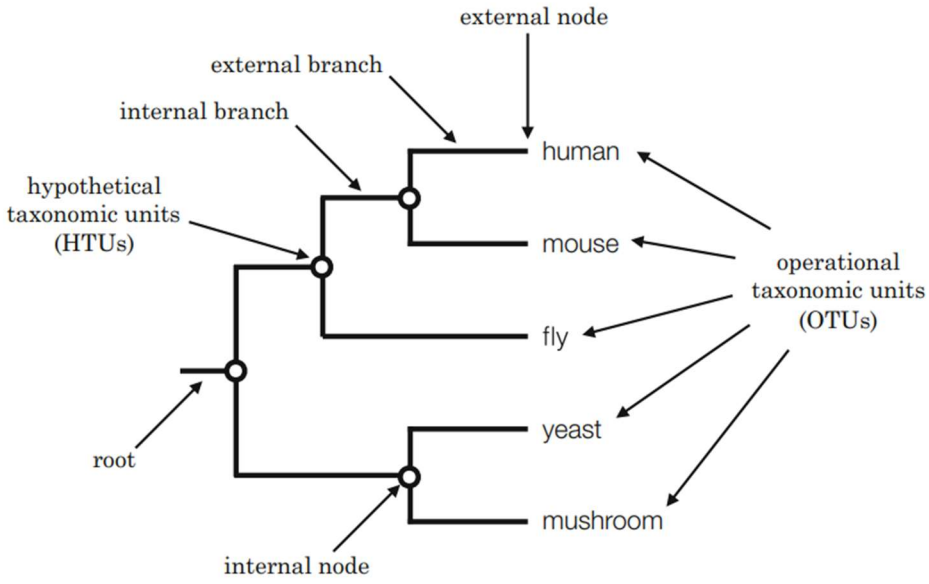


Figure 1 - Composition of a phylogenetic tree.

External nodes represent OTUs and the internal nodes represent a common ancestor. Among these are the branches that are used to connect nodes and show the evolutionary relationships between them.

An inner branch that connects two inner nodes, demonstrates an ancient relationship. On the other hand, an outer branch connecting an inner node and an outer node represents a recent relationship.

The deepest branch of the tree represents the root or most recent common ancestor (MRCA) of all taxonomic units in the tree. There are two types of trees, rooted and unrooted, that is, one has a root and the other has no root. The rooted tree in an evolutionary context has more meaning than a rootless tree, because it identifies the ancestral origin of all taxonomic units. The best way to add a root to the tree is to add an outgroup. The best outgroup is the organism that has recently diverged from the rest of the organisms in the tree.

OTUs that descend from the same ancestor form a “monophyletic group”. However, a group of organisms that share the same common ancestor but does not include all members descended from that ancestor is called a "paraphyletic group". The group of OTUs that are not derived from the same common ancestor is called a “polyphyletic group” [8] (Figure 2).

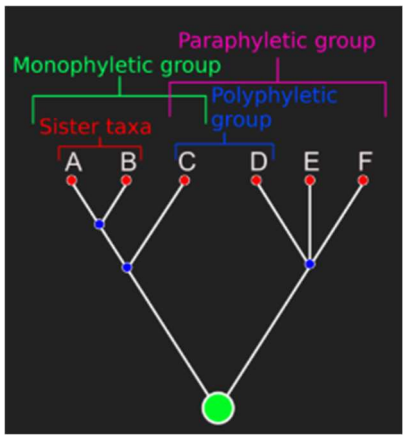


Figure 2 - Examples of monophyletic, paraphyletic and polyphyletic groups.

4.1.3 Phylogenetic Reconstruction Methods

Phylogenetic reconstruction methods can be classified into two approaches: distance-based methods and character-based methods.

Distance-based methods transform the information from all sequences into a distance matrix, which is then analyzed using an algorithm to cluster the OTUs. Building a tree with this method is faster but the sequence information is lost in the process. Character-based methods are more time consuming because all sequences information is used for the evolution of the best phylogenetic tree [8]. The calculation

of phylogenetic trees using this method can be done using several approaches, such as Maximum Parsimony (MP), Maximum Likelihood (ML) or Bayesian Inference (BI).

The maximum parsimony method is a substitution model-free method for phylogenetic tree reconstruction that assumes as few changes as possible. It is used to build trees from morphological data, making it difficult to measure the rate of evolutionary change. This method looks for the “most parsimonious tree”. It is a time-consuming method and is not recommended when sequences of multiple genes are concatenated or have high levels of variation [8].

The maximum likelihood method estimates branch lengths and tree topology based on the substitution model and sequence alignment. The numerical result of the ML analysis is the probability that the topology of a tree and model fit the sequences. The tree with the highest value of likelihood is considered the best tree. ML method claims to be very accurate, because the analysis depends heavily on the evolutionary model. However, it has a long calculation time, due to using all the information from the sequences to calculate the highest value of probability, and it is impractical for large data sets, because the calculation is robust and requires significant computational resources [8].

The Bayesian inference method relies on Bayesian statistics, and uses the Markov Chain Monte Carlo (MCMC) algorithm to find the best tree. This technique is more sophisticated than the one used in the ML method because each new tree explored can produce a lower score than the tree in the previous step. This allows the algorithm to efficiently find the best tree [8].

The bootstrap values correspond to the values at the internal nodes and support the phylogenetic analysis produced. The higher the bootstrap value, the higher the confidence of the represented node.

4.1.4 DNA and RNA

DNA (deoxyribonucleic acid) [9] and RNA (ribonucleic acid) [10] are nucleic acids that have different structures and functions. While DNA is responsible for storing the genetic information of living beings, RNA acts in the production of proteins and in the

regulation of gene expression. There are several types of RNA, the most common are messenger RNA (mRNA), ribosomal RNA (rRNA) and transfer RNA (tRNA).

4.1.5 Splicing

Splicing is the process of maturation of the pre-mRNA, in this process the introns are removed from the pre-mRNA, and the open reading frame (ORF) become continuous [11].

4.1.6 Genome

The genome represents all the hereditary information of an organism that is encoded in the DNA. Including genes, intergenic regions, mitochondrial DNA and plastid DNA [12].

4.1.7 Transcriptomes

Transcriptomes are the set and quantity of transcripts (messenger RNA, ribosomal RNA, transfer RNA and microRNA) of an organism, organ, tissue or cell lineage at a specific developmental stage or physiological condition [13]. Transcriptomes are important because they are a direct reflection of gene expression. Being mRNA sequences the protein coding sequences, they are important for the study of functional genomics, therefore, it is an important point for the analysis of positive selection because it is important to analyze the splicing variation and expression level of genes, in addition to sequence mutations.

4.1.8 Sequence Alignment

Before proceeding with the construction of phylogenetic trees, it is necessary to align the sequences. Phylogenetic analysis is about making comparisons and so it is necessary to make sure that the data are comparable. Therefore, homology is necessary. Homology is similarity due to shared ancestry between a pair of structures or genes in different taxonomic units [14].

Sequence alignment is a way of organizing primary structures of DNA, RNA or proteins to identify similar regions that may be a consequence of functional, structural or evolutionary relationships between them. There are two types of alignments, Pairwise Sequence Alignment and Multiple Sequence Alignment (MSA). Pairwise alignment is used to identify regions of similarity that may indicate functional, structural and/or

evolutionary relationships between two biological sequences. In contrast, MSA corresponds to the alignment of three or more biological sequences of similar length. From the output data of the MSA method, it is possible to infer the homology and evolutionary relationship between the sequences under study [15].

4.2 Adaptive Radiation and Cichlid Fishes

Adaptive radiation corresponds to the evolutionary phenomenon where, in a short period of time, several species are formed from the same ancestral species [16].

Adaptive radiation is probably the source of much of the ecological and morphological diversity. However, how adaptive radiations proceed and what determines their extent remains a mystery in most cases [17].

Evolutionary radiations are referred to in this way if new life forms rapidly evolve through adaptive diversification into a variety of ecological niches, which normally presupposes ecological opportunity. The occurrence or not of an adaptive radiation depends on a variety of extrinsic and intrinsic factors. The cichlid fish species present in Lake Tanganyika in East Africa is considered by some to be the “most outstanding example of adaptive radiation” [17]. This group of cichlids comprises about 240 species, which together show an extraordinary degree of morphological, ecological and behavioral diversity. In the scientific paper “Drivers and dynamics of a massive adaptive radiation in cichlid fishes” a phylogenetic tree of Lake Tanganyika cichlid species had been constructed based on genome-wide data in order to demonstrate the adaptive nature of radiation (Figure 3).

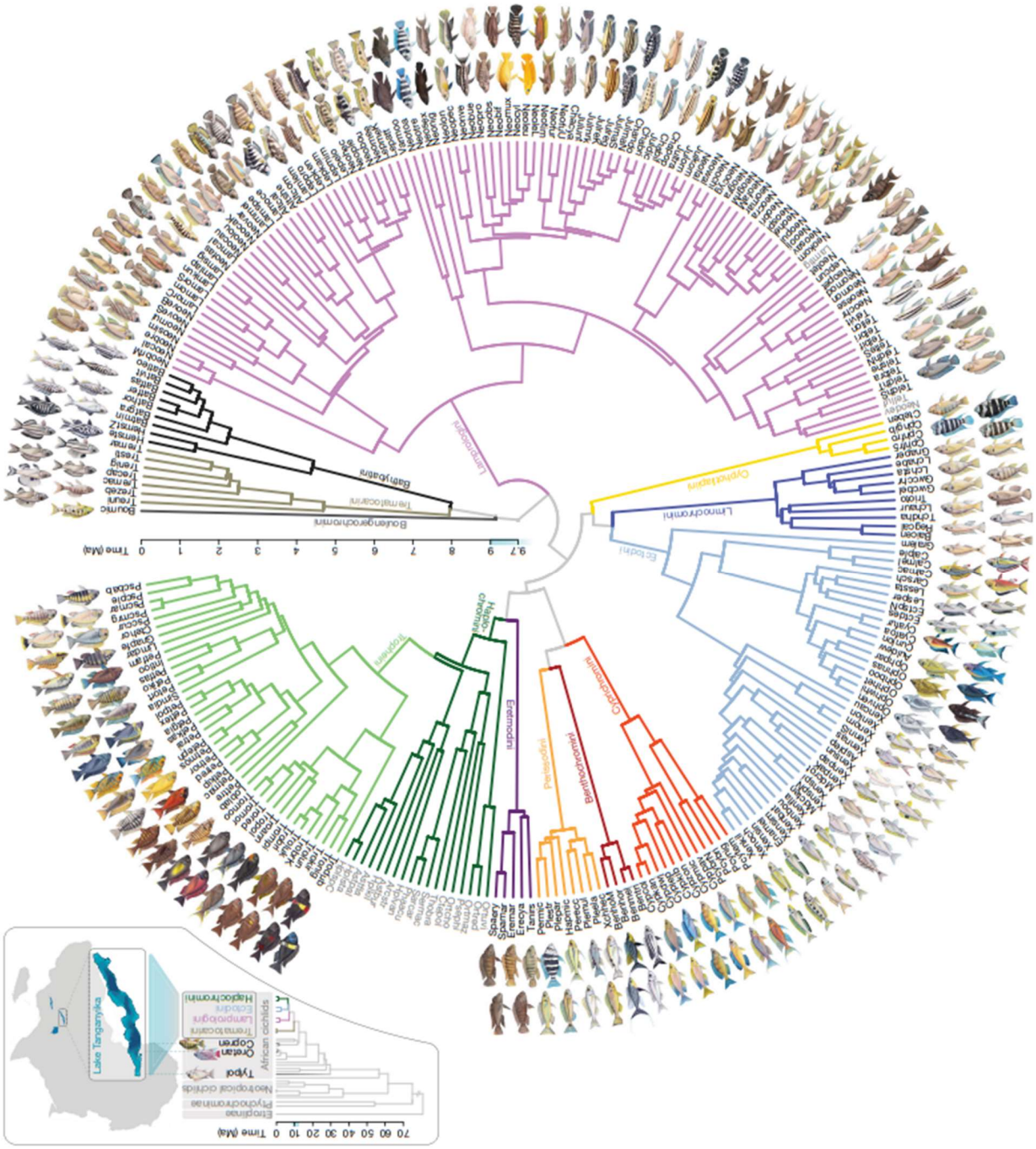


Figure 3 - Phylogenetic tree of Lake Tanganyika cichlid species based on entire genome. Source: "Drivers and dynamics of a massive adaptive radiation in cichlid fishes" [17].

There are several factors that promote radiation, such as:

- The conquest of a new ecological niche in which there are no competitors, allowing the ancestral species to radiate itself in different ways, capable of exploiting existing resources.
- The extinction of competitors, opening space for the emergence of new species.

- Substitution of competing species.
- Adaptive barriers that end up forcing species to differentiate themselves so that they can exploit new resources.

4.3 Positive Selection

Selective pressure is any set of environmental conditions that cause certain alleles or genes to be favored over others in a given population. These genes do not necessarily bring advantages to individuals in other environmental conditions, the continuity of the environment will lead to directional selection increasing the frequency of favored alleles [18]. Therefore, directional or positive selection is the process by which new genetic variants confer an advantage on individuals being expanded in the population.

Positive selection can be analyzed through the ratio of synonymous and non-synonymous substitutions. A non-synonymous substitution is a nucleotide mutation that changes the amino acid sequence of a protein. Synonymous substitutions differ from non-synonymous substitutions, which do not change the amino acid sequences. As non-synonymous substitutions result in a biological change in the organism, they are subject to natural selection [19].

An excess of non-synonymous over synonymous substitutions at certain specific sites on individual amino acids is an important indicator that positive selection has affected the evolution of a protein between the sequences under study and its most recent common ancestor. There are several methods for detecting the presence, and sometimes the location, of positively selected sites in protein-coding sequence alignments [20].

Non-synonymous substitutions at certain specific locations in the sequence, or "loci", can be compared to synonymous substitutions at those same locations to obtain the K ratio. This ratio is used to measure the evolutionary rate of gene sequences. If a gene has lower levels of non-synonymous than synonymous nucleotide substitution, then it can be inferred as functional due to the Ka/Ks ratio.

$$\mathbf{K = Ka/Ks}$$

Ka, corresponds to a nucleotide substitution that changes the corresponding amino acid in the protein, that is, a non-synonymous substitution. Ks, corresponds to a nucleotide substitution that does not change the amino acid in the protein, that is, a synonymous substitution.

If a site has K > 1, it suggests that it is variable and that it evolved under positive selection, or directional selection. If a site has K < 1 it means that it has been conserved and will have been subjected to purifying selection (Massingham & Goldman, 2005). Synonymous mutations do not change the encoded protein and are therefore often considered selectively neutral. If a non-synonymous mutation does not affect the fitness of a protein, it drifts in the population at the same rate as a synonymous mutation, giving a non-synonymous/synonymous substitution ratio (K) of 1.

4.4 Pipelines

4.4.1 Concept

Pipelines are high-level components of continuous integration, delivery and deployment that enable the automation of analysis processes. They comprise jobs, which define what to do, and stages, which define when to run the jobs.

The advantage of using pipelines is that they are executed automatically and do not require intervention after they are created, making the analysis process faster and automated [21].

4.4.2 Git and GitLab

Git [22] is a decentralized version control system and indicates what, who, when and why changed. It does not depend on a central infrastructure, making it ideal for asynchronous collaborations where each programmer has a complete copy of the project. Projects in *Git* are called repositories and contain all files including their respective edit history, where each change made is called a commit. Each commit is linked and organized into branches.

GitLab is an online *Git* platform where you can publish and share code. For creating a public pipeline, this platform is ideal.

4.4.3 Docker

Docker is a container manager that can start or stop containers, run them and create container file systems, i.e., images. Its objective is to create the pipeline image itself and an automatic file that executes several instructions in the command line in succession, thus installing the programs necessary for the analysis [23].

4.4.4 Snakemake

The *Snakemake* workflow management system is a tool for creating reproducible and scalable data analytics. Workflows are described using a human-readable language based on Python. They can scale seamlessly for server, cluster, grid, and cloud environments without having to modify the workflow definition [24].

A *Snakemake* workflow is defined with a *Snakefile*. The *Snakefile* indicates the order of execution of the program, following the set of implemented rules.

5 Materials and Methods

5.1 Obtaining the Sequences

The sequences of cichlid species are available at NCBI in the *Bioproject* with accession number PRJNA550295. It contains the complete genome sequencing of cichlid fish from Lake Tanganyika for the construction of phylogenetic relationships and analysis of sequence evolution.

These sequences were the same used in the article “Drivers and dynamics of a massive adaptive radiation in cichlid fishes” [17].

In the internship project, the species present in the “List of spp.xlsx” file referred to in the annexes were used. The *NCBI BLAST* tool [25] was used to locate the oxytocin gene in the entire cichlid genome. The sequence containing only the oxytocin gene of the ancestral species *Oreochromis niloticus* (Nile Tilapia) was used as a *query*, and the sequences of the genomes of the cichlid species as *subject*.

Up to 6% of query cover were found to have multiple repeat regions that corresponded to random parts of the genome, probably due to a conserved domain between different

genes. Therefore, all sequences with *query cover* above 6% were considered, as they could represent complementary regions. To obtain the cichlid sequences only with the oxytocin gene, the *range* was observed. These sequences were obtained manually. All sequences were obtained in *fasta* format. For the reverse sequences, the "Reverse Complement Tool" was used.

5.2 Sequence Alignment

The *Aliview* program version 1.28 [26] was used in order to obtain a *pairwise* alignment. This type of alignment was used instead of the *MSA* alignment due to noting the similarities between the cichlid species and the reference, and the mutations that occurred in the oxytocin gene over time. The *default* alignment of *Aliview*, *Muscle* version 3.8.425 [27] was used.

After performing the alignment, the coding sequences were obtained. Through Ensembl it was possible to observe the regions of the *oxt* gene that are expressed, that is, the exons. Using *Aliview*, only the regions of the cichlid sequences corresponding to the exons were selected. All sequences with *query cover* equal to 6% and *e-value* lower than e-50 were eliminated, as they did not represent complementary regions of the sequences, only repeated parts of the genome. This alignment resulted in *gaps* that were later manually eliminated.

5.3 Construction of Phylogenetic Trees

For the construction of phylogenetic trees, two different programs were used in order to build two different types of trees.

The program MrBayes version 3.2.6 [28] was used for Bayesian inference. *MrBayes* needs the files in *nexus* format. In the first approach, the files were converted directly through *Aliview*. The following parameters were used:

- *Nst=6*, defines the number of substitution types.
- *Rates=invgamma*, defines the model for rate variation between sites.
- *Ngen=1800000*, choose the number of cycles for the MCMC algorithm.
- *Samplefreq=100*, specifies how often the Markov chain is sampled.

- *Printfreq=100*, specifies how often information about the string is shown on the screen.
- *Diagnfreq=1000*, indicates the number of generations between the calculation of MCMC diagnoses.

The remaining parameters were left in *default*.

For the inference based on the maximum likelihood of phylogenetic trees, in *Cipres* [29] the *RAxML-HPC* v.8 [30] tool was used in *XSEDE* (8.2.12) for fast *bootstrapping*. The parameters were left in *default*.

To visualize the phylogenetic trees obtained, the program *Figtree* version 1.4.4 was used.

5.4 Synonymous and Non-Synonymous Substitutions

For the analysis of synonymous and non-synonymous substitutions, in a first manual approach, *Aliview* was used, in order to observe the mutations in each cichlid species.

5.5 Pipeline for construction of phylogenetic trees and analysis of synonymous and non-synonymous substitutions

In order to automate the analysis process, a second approach to the project was made. To this end, a pipeline was created in GitLab.

A Git repository was created with the files and folders: *Snakefile*, *Dockerfile*, *README.md*, *data*, *hyphy*, *nexus*, *outputs*, *scripts* and *trees*. A *Dockerfile* (Docker), to install the necessary programs for the analysis: *RAxML*, *MrBayes*, *Conda*, *HyPhy*, among others. And a *Snakefile* (Snakefile), containing rules for building phylogenetic trees and analyzing the ratio of synonymous and non-synonymous substitutions.

To run the *Dockerfile* the commands were used:

- *docker run*, in order to create the container with the image.
- *docker build*, to build the image from the *Dockerfile*.
- *docker images*, to manage all container images.

To run the pipeline, the following command was used:

$ *INPUT="input file name" OUTPUTDIR="output folder name" snakemake --cores "number of cores" "rule name"*

5.5.1 Phylogenetic Trees

For the phylogenetic tree based on maximum likelihood, the program *RAxML* with *PTHREADS* were used with the following parameters:

- *-f a*, for quick bootstrap analysis and a search for the best maximum likelihood score.
- *-m GTRCAT*, defines the model that will be used, in this case it will be the CAT model.
- *-p 112358*, specifies a seed for parsimony inferences to help reproduce the results.
- *-x 112358*, do a quick bootstrap and specify a random seed.
- *-N 100*, sets the number of bootstraps analysis to be performed.

The remaining parameters were left at *default*.

Fasta to Nexus Converter

A python script was created to convert *fasta* files to *nexus*, and a rule in *Snakefile* to add the necessary parameters to build the phylogenetic tree in *MrBayes*.

The phylogenetic tree based on Bayesian inference was made in the same way as in the first approach. *MrBayes* and the same parameters were used.

5.5.2 Ratio of Synonymous and Non-Synonymous Substitutions

In order to infer the selection pressure acting on the oxytocin gene, the ratio of non-synonymous (Ka) and synonymous (Ks) substitutions was estimated using *Hyphy* version 2.5.29 [31]. The *aBSREL* method was chosen for the test of specific lineage evolution through the locations of the branches of the phylogenetic tree of maximum likelihood. The remaining parameters used can be found in the *hyphy_parameters* file available on *GitLab*.

5.6 Transcriptomes analysis

To analyze the splicing variation of messenger RNA sequences, *IGV* was used [32].

The cichlids RNA sequences were made available by Instituto Gulbenkian de Ciência in *BAM* format. To obtain the sequences, the Institute's remote server was accessed and the sftp protocol was used in order to access the remote files on the local computer.

The reference genome corresponding to the specie *Oreochromis niloticus* was downloaded in *IGV*. Then, genome annotations were imported to identify the various genes present in the genome. Finally, an mRNA sequence from a cichlid was downloaded.

Having the location of the oxytocin gene in the genome of *Nile Tilapia*, it was possible to observe the reads of each cichlid sequence. The location of oxytocin gene is *NC_031977.2:24,070,811-24,071,039*. Through *IGV* it is possible to observe the nucleotide mutations and analyze the splicing of the sequences. The splicing analysis is important to confirm that the oxytocin gene from cichlids and the oxytocin gene from the reference specie, *Nile Tilapia*, encode the same region and produce a similar and functional protein.

6 Results

6.1 Phylogenetic Trees

The species present in the phylogenetic trees are represented according to the tribe theybelong to by color, as in the species tree (Figure 3). The values present in the branches of the trees correspond to the bootstrap values and the *posterior probability* values in percentage, in the case of the maximum likelihood tree and Bayesian inference, respectively. The attachments section contains the original files for a better visualization of the phylogenetic trees obtained.

MrBayes

The phylogenetic trees obtained using the Bayesian inference method in MrBayes were exactly the same when using the command line interface or the pipeline. In Figure 4, the phylogenetic tree obtained by the Bayesian inference method is represented.

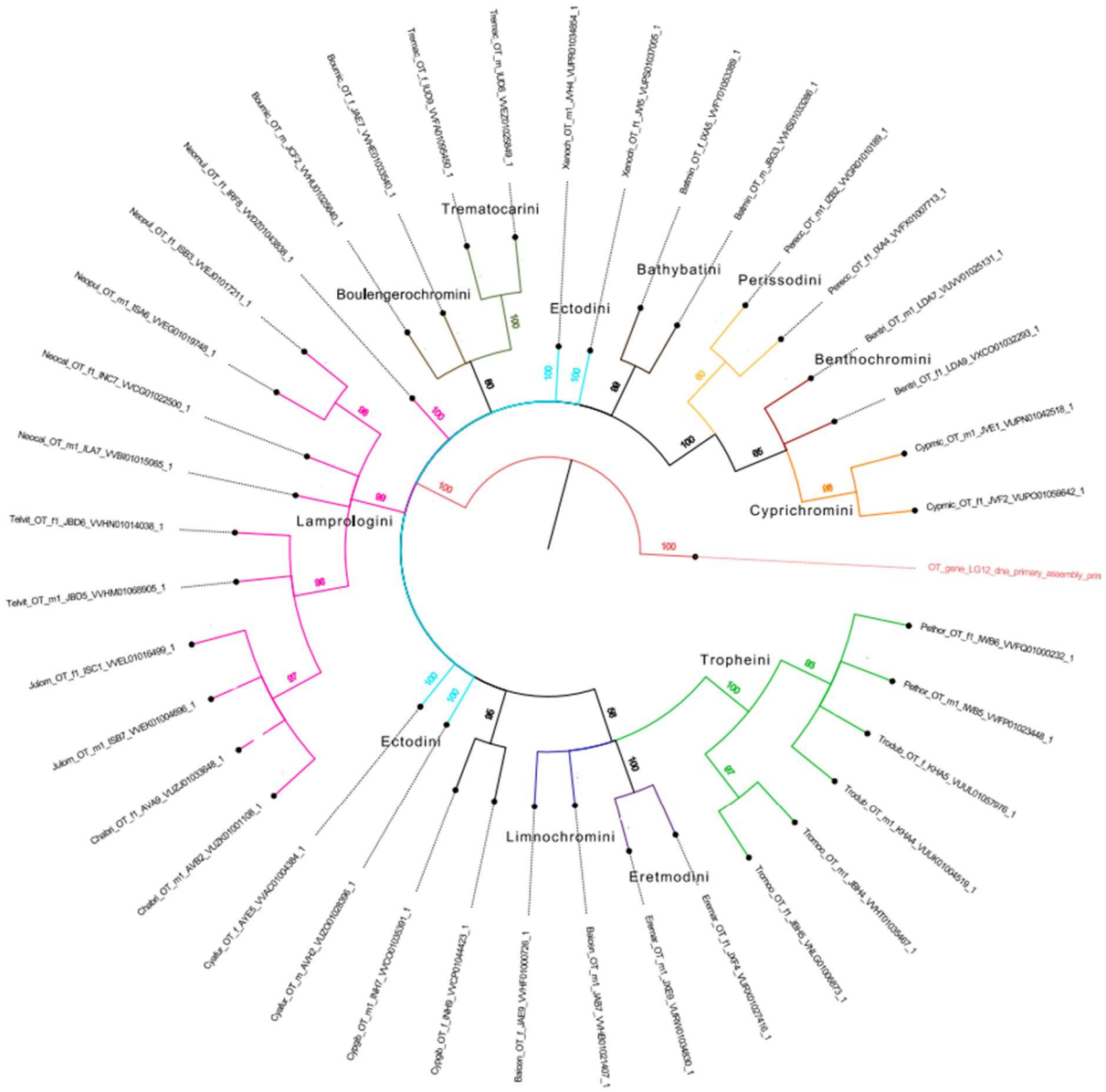


Figure 4 - Phylogenetic tree obtained by the Bayesian inference method of cichlid species from Lake Tanganyika based on the oxt gene. The values in the branches correspond to the "posterior probability" values in percentage. The colors represent the tribe to which they belong.

It is possible to observe that in the tree obtained by Bayesian inference method, the species of *Ectodini* tribe, *Xenoch) e Cyafur*, do not group together while the species of

the other tribe's group together. In the species tree (Figure 3), the species *Xenoch* and *Cyafur* group together. *Cypgib*, *Xenoch*, *Cyafur* and *Batmin* species form polytomies.

RAxML

The RAxML-HPC tool, executed in XSEDE of the *Cipres* portal, was used to obtain a maximum likelihood tree (Figure 5). The RAxML and PTHREADS method was also used and resulted in a similar tree (Supplementary Figure 1), available in the attachments section.

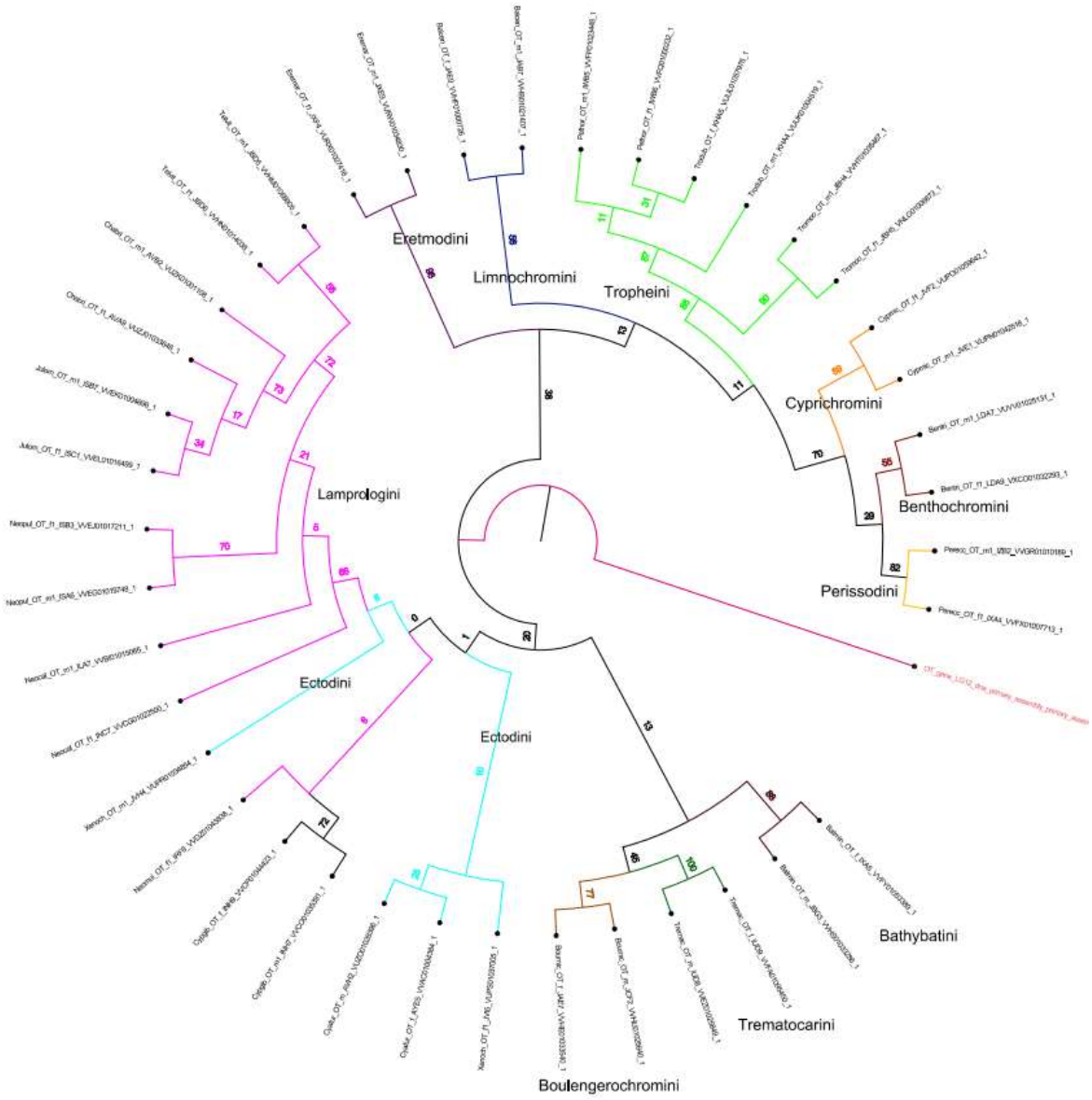


Figure 5 - Phylogenetic tree obtained by ML method of cichlids species from Lake Tanganyika based on the oxt gene using RAxML-HPC. The values in the branches correspond to the support values (bootstrap) in percentage. The colors represent the tribe to which they belong.

In the trees obtained using maximum likelihood method, it is possible to observe that the species of the tribe *Ectodini*, *Xenoch* and *Cyafur*, group together with the species of the tribe *Lamprologini*. However, in the species tree (Figure 3), *Xenoch* and *Cyafur* group together.

6.2 Synonymous and Non-Synonymous Substitutions

Non-synonymous substitutions observed manually using Aliview are shown in the following table (Table 2).

Table 2 - Table of non-synonymous substitutions observed directly from Aliview.

Table of non-synonymous substitutions				
Species	Reference codon	Reference aminoacid	Mutation	Aminoacid formed
Pethor	CGC	Arg	CAC	His
Trodub				
Tromoo				
Eremar	AGT	Ser	AAT	Asn
Pethor	GTG	Val	ATG	Met
Trodub				
All species	GCC	Ala	CCC	Pro
All species	GCC	Ala	ACC	Thr
Cypmic	TCC	Ser	TTC	Phe
Tremac				

Tremac	ACC	Thr	ATC	Ile
Bentri	GGC	Gly	GTC	Val
Perecc				
Cypmic				
Batmin	GGC		AGC	Ser
Boumic	GGT	Gly	AGT	Ser
Tremac				
Xenoch male	CAT	His	CGT	Arg
Chabri	GCT	Ala	GTT	Val
Julorn				
Batmin male	CCT	Pro	TCT	Ser
Eremar	CAA	Gln	TAA (stop codon)	X

It was observed that the number of synonymous substitutions is greater than the number of non-synonymous substitutions.

6.3 Transcriptomes

The splicing of messenger RNA sequences of the oxytocin gene from cichlids species and the reference specie, *Oreochromis niloticus*, was observed. The following figure shows the splicing of the specie *Boulengerochromis microlepis* (Figure 6).

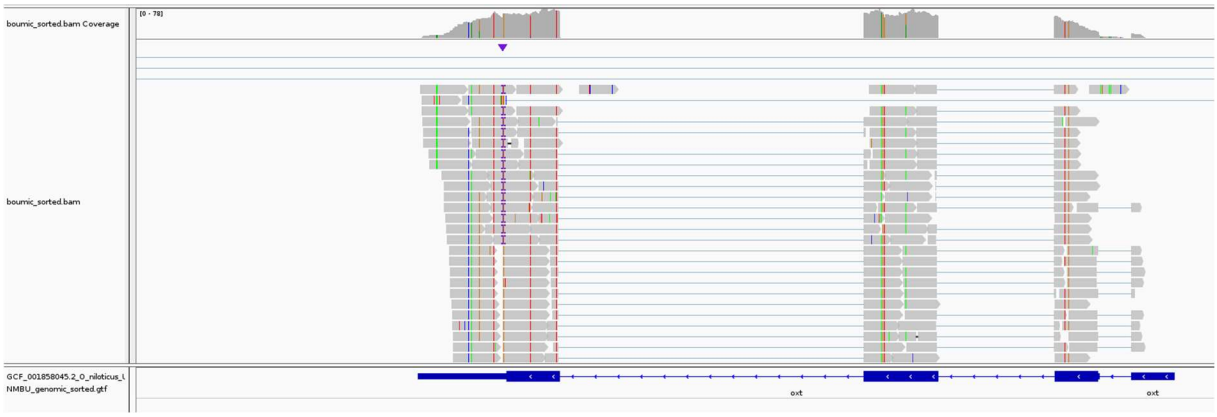


Figure 6 - Splicing of cichlid species sequences.

It was observed that the cichlid sequences and the reference encode the same region for the oxytocin gene.

Figures 7, 8 and 9 show some examples of mutations that occur in the *mRNA* sequence of *Boulengerochromis microlepis* specie observed in *IGV*.

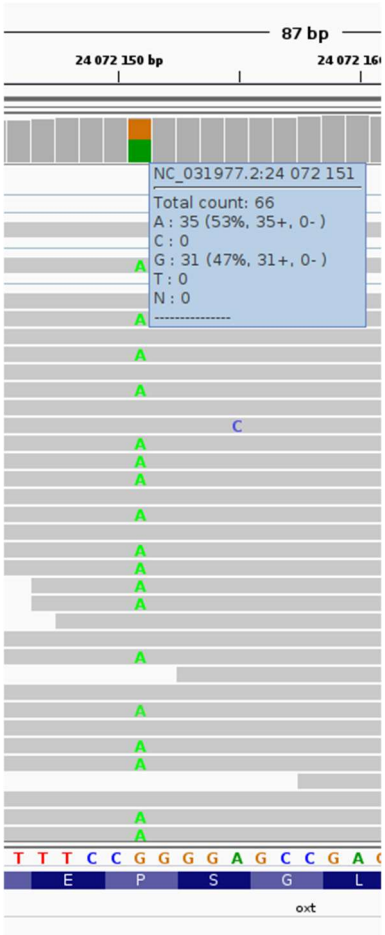


Figure 7 - Heterogozity in Boulengerochromis microlepis.

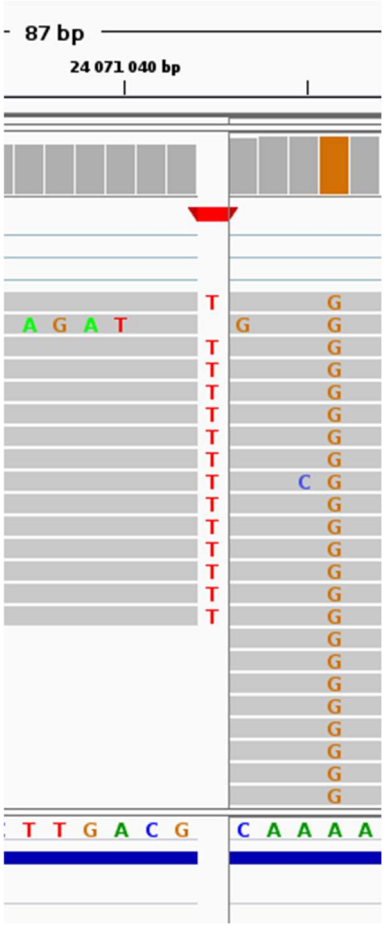


Figure 8 - Insertion in Boulengerochromis microlepis.

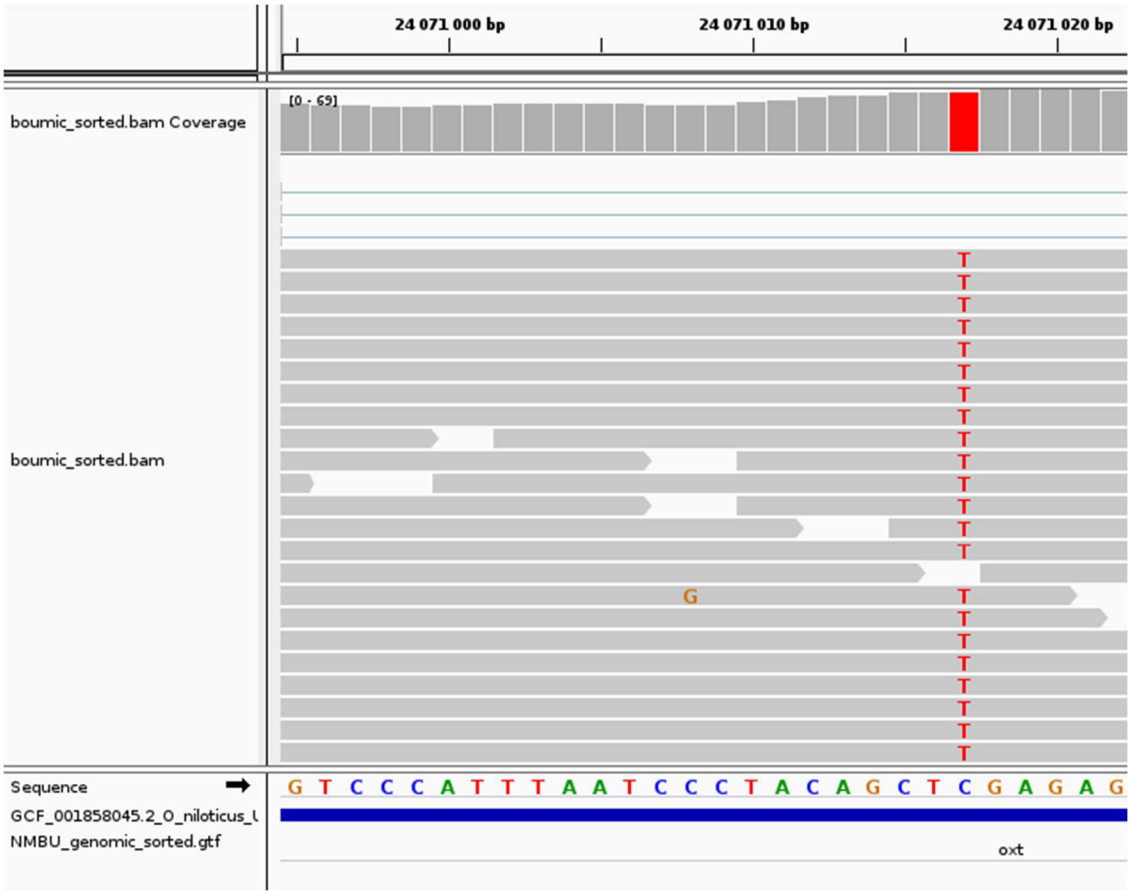


Figure 9 - Non-synonymous nucleotide substitution (mutation) in the species Boulengerochromis microlepis.

6.4 Ratio of Synonymous and Non-Synonymous Substitutions

It was observed that the ratio of synonymous and non-synonymous substitutions is equal to 0.05, that is, K < 1, through the probability test of adaptive branching random effects (Figure 10).

```
|          Node38          |     1     |  1.00 (100.00%)   |        0.00         |       1.00000       |
|          Node51          |     1     |  1.00 (100.00%)   |        0.00         |       1.00000       |
|          Node59          |     1     |  1.00 (100.00%)   |        0.00         |       1.00000       |
|          Node63          |     1     |  1.00 (100.00%)   |        0.00         |       1.00000       |
|          Node57          |     1     |  1.00 (100.00%)   |        0.00         |       1.00000       |
|          Node52          |     1     |  1.00 (100.00%)   |        0.00         |       1.00000       |
|          Node54          |     1     |  1.00 (100.00%)   |        0.00         |       1.00000       |
----
### Adaptive branch site random effects likelihood test
Likelihood ratio test for episodic diversifying positive selection at Holm-Bonferroni corrected _p =   0.0500_ found **0** branches under selection among **71** tested.
```

Figure 10 - ML ratio test

7 Discussion of Results

Phylogenetic Trees

Comparing the phylogenetic trees obtained in the results section (Figures 4 and 5) referring to oxytocin gene with the species tree referring to cichlid genomes (Figure 3), it was observed that the trees are very similar due to all individuals of the same species group together. This indicates that the oxytocin gene is well conserved, and as such, the occurrence of positive selection signatures is not observable.

The tree obtained through the Bayesian inference method does not define all speciation events, for this reason it does not represent a very descriptive tree. This is possibly due to the fact that the oxytocin gene is very conserved and also because it is a gene of relatively small size. If the analysis had taken into account a less conserved gene, the evidence of positive selection would possibly be remarkable. However, the trees obtained by the Bayesian inference method do not correspond to a true tree, but to a consensus among the best maximum likelihood trees. This method calculates a likelihood value that indicates how well the model matches the data that is not necessarily true.

The trees obtained through the maximum likelihood method are better defined and correspond to a true tree, thus representing an advantage to use this method instead of the Bayesian inference method.

However, despite the trees corresponding to oxytocin gene being very similar to the species tree, some differences were observed. These differences are due to the fact that the trees obtained in the results represent only one gene, some individuals that were previously grouped together in the entire genome tree, in the trees corresponding to the Oxytocin gene are separated, as is the case of species from the tribe *Ectodini* and *Lamprologini*, as they share the most similar sequences of oxytocin gene. It is believed that Oxytocin affects behavior at the level of stress and social behavior, so if the species share similar characteristics at the level of the sequences of the respective gene, it may be possible to say that these species share similar behaviors at the level of the oxytocin gene.

In the trees obtained through the method of maximum likelihood, it is possible to observe that the species of the tribe *Ectodini*, *Xenoch* and *Cyafur*, group with the species of the tribe *Lamprologini*. This event could represent a possible positive selection event, and as such, represents something to be analyzed in more detail. In order to investigate further, the similarities between the respective branches should have been observed and the analysis of how these differences translate into a different protein should have been carried out.

Synonymous and Non-Synonymous Substitutions

When looking at the sequences manually in Aliview, more synonymous than non-synonymous substitutions were found, this means that the percentage of synonymous substitutions is higher, so it is more likely not to change the final protein. Again, this is possibly due to the fact that oxytocin is a highly conserved gene, as previously seen in phylogenetic trees.

Transcriptomes

Through the results of the IGV it is possible to confirm that the oxytocin gene is highly conserved. By analyzing the splicing of the mRNA transcriptomes, it is verified that the oxt gene of cichlids and the oxt gene of the ancestral reference species encode the same region, thus producing a similar protein. By observing the mutations in the RNA sequences of some species, it was found that some non-synonymous substitutions occur. However, it is observed that the number of this type of mutation is relatively low to affirm that the protein is altered and that, in turn, there is evidence of positive selection.

Ratio of Synonymous and Non-Synonymous Substitutions

According to the maximum likelihood model, it can be assumed that there are no signs of ORF-wide positive selection, since the value of K is less than 1. This means that the locations of the branches of the maximum likelihood tree were conserved and will have been subjected to a purifying selection. As oxytocin is a gene subject to purifying selection, it has a reduced amino acid substitution rate, so it is more conserved among different species, thus assuming that positive selection does not occur. However, the *aBSREL* selection pressure inference method does not test selection at specific

locations, it only tests selection for each branch of interest of the maximum likelihood tree.

The analyzes performed and the results obtained suggest that there is no positive selection at the level of the oxytocin gene in the cichlid fish species studied. This is because oxytocin is a very well conserved gene. To be sure that positive selection occurs, should be include more cichlid species in the analysis, use longer or less conserved genes. Furthermore, include the use of alternative approaches, such as *sweeps* localization tools and F_{ST} tests to indicate regions of interest of the gene in order to provide information about specific regions of the gene that are under selection pressure.

On the other hand, if more data were to be included in the analysis, the maximum likelihood method would likely be impractical because it uses all the information from the sequences to calculate the highest probability value, thus being impractical for large data sets, because the calculation is robust and require significant computational resources that will hardly be solved by mere hardware improvement. In this case, the alternative would be to use the Bayesian inference method, as it is faster. However, it is not as precise a method as the maximum likelihood method and does not represent a true tree, but a consensus among the best maximum likelihood trees. If the maximum likelihood method is slow but feasible, it remains the best choice to look for evidence of positive selection as it represents a more accurate and a true tree.

If we consider that the oxytocin gene is involved in the behavior of cichlid fish, perhaps the OXT-OXTR signaling receptors are changing and undergoing selective pressure instead of the neurotransmitters. Therefore, the next step of the analysis would be to study the oxytocin receptors.

8 Conclusions

In conclusion, all analyzes and results obtained indicate that there is no evidence of signatures of positive selection at the level of the oxytocin gene in cichlid species, in order to fulfill the objective of the project. This is due to the fact that the oxytocin gene is very conserved, possibly because it plays a very important role in the regulation of the hypothalamic-pituitary-adrenal axis that modulates the behavioral response to stress and social behavior and gives rise to proteins that are essential for the survival of the species. However, these results are not sufficient to confirm the occurrence of positive selection. To improve the analysis, more biological data should be included, more tests should be done including specific sequence sites and the study of oxytocin receptors.

The internship project carried out for Instituto Gulbenkian de Ciência was important for it mainly due to the automation of the process of analyzing phylogenetic trees and calculating the ratio of synonymous and non-synonymous substitutions through the creation of a pipeline. The code developed will be used for the actual project of colleague Post-Doctor, Pol Sorigué, and will be available on the GitLab platform for consultation with the name *phylo_analysis* referring in the attachments. The project developed was of great importance for the technical development and programming skills and use of Bioinformatics tools.

Bibliographic References

[7] Ajawatanawong, P. (2017). Molecular Phylogenetics: Concepts for a Newcomer. *Advances in Biochemical Engineering/Biotechnology*, *160*, 185–196. https://doi.org/10.1007/10_2016_49

[24] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

[21] Chacon, S., & Straub, B. (2014). *Pro git*. Apress.

[26] Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

[25] Larsson, A. (2014). AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*, *30*(22), 3276–3278. https://doi.org/10.1093/bioinformatics/btu531

[19] Massingham, T., & Goldman, N. (2005). Detecting amino acid sites under positive selection and purifying selection. *Genetics*, *169*(3), 1753–1762. https://doi.org/10.1534/genetics.104.032144

[28] Miller, M. A., Pfeiffer, W., & Schwartz, T. (2010). Creating the CIPRES Science Gateway for inference of large phylogenetic trees. *2010 Gateway Computing Environments Workshop, GCE 2010*. https://doi.org/10.1109/GCE.2010.5676129

[6] Neumann, I. D. (2002). Involvement of the brain oxytocin system in stress coping: interactions with the hypothalamo-pituitary-adrenal axis. *Progress in Brain Research*, *139*, 147–162. https://doi.org/10.1016/s0079-6123(02)39014-9

[30] Pond, S. L. K., Frost, S. D. W., & Muse, S. V. (2005). HyPhy: hypothesis testing using phylogenies. *Bioinformatics*, *21*(5), 676–679. https://doi.org/10.1093/bioinformatics/bti079

[31] Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. In *Nature biotechnology* (Vol. 29, Issue 1, pp. 24–26). https://doi.org/10.1038/nbt.1754

[16] Ronco, F., Matschiner, M., Böhne, A., Boila, A., Büscher, H. H., El Taher, A., Indermaur, A., Malinsky, M., Ricci, V., Kahmen, A., Jentoft, S., & Salzburger, W. (2021). Drivers and dynamics of a massive adaptive radiation in cichlid fishes. *Nature*, *589*(7840), 76–81. https://doi.org/10.1038/s41586-020-2930-4

[27] Ronquist, F., Huelsenbeck, J., & Teslenko, M. (2018). *Draft MrBayes version 3.2 Manual: Tutorials and Model Summaries*. *July*, 180.

[29] Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

[1] Turner, G. F. (2007). Adaptive radiation of cichlid fish. *Current Biology*, *17*(19), 827–831. https://doi.org/10.1016/j.cub.2007.07.026

[2] Fryer, G. & T. D. Iles, 1972. The Cichlid Fishes of the Great Lakes of Africa. T.H.F., Neptune, NJ.

[3] McKaye, K. R. & W. N. Gray, 1984. Extrinsic barriers to gene flow in rock-dwelling cichlids of Lake Malawi: macrohabitat heterogeneity and reef colonization. In Echelle, A. A. & I. Kornfield (eds), Evolution of Fish Species Flocks. University of Maine at Orono Press, Orono: 169– 183.

[4] Rossiter, A., 1995. The cichlid fish assemblages of Lake Tanganyika: ecology, behavior and evolution of its species flock. Advances in Ecological Research 26: 187–252.

[5] Sturmbauer, C., 1998. Explosive speciation in cichlid fishes of the African Great Lakes: a dynamic model of adaptive radiation. Journal of Fish Biology 53(Supplement A): 18–36.

[7] Gimpl G, Fahrenholz F (2001). The oxytocin receptor system: structure, function, and regulation. Physiol Rev, 81(2):629-83

[9] Bruce Alberts,Alexander Johnson, Julian Lewis, Kazuo, Martin Raff, Keith Roberts, and Peter Walters (2002). *Molecular Biology of the Cell; Fourth Edition*.

[10] Santos, F. P. & Castro, C. S. “RNA”. 2000.

[11] Clancy, S. (2008) RNA splicing: introns, exons and spliceosome. Nature Education 1(1):31

[12] Kevin Davies, Decifrando o Genoma: a corrida para desvendar o DNA Humano ; Companhia das Letras, 2001

[13] REDVET. Revista electrónica de Veterinaria 1695-7504 2007 Volumen VIII Número 10

[14] Panchen, A. L. (1999). "Homology—history of a concept". Novartis Found Symp. Novartis Foundation Symposia.

[15] Markel, Scott; León, Darryl (2003). Sequence Analysis. Beijing: O'Reilly.

[16] A RADIAÇÃO ADAPTATIVAB. Gallavotti, Cientic.

[18] NESCent: Education & Outreach: Examples of Evolution: Selective Pressures and Adaptation

[19] Ting Hu and Wolfgang Banzhaf. “Nonsynonymous to Synonymous Substitution Ratio ka/ks: Measurement for Rate of Evolution in Evolutionary Computation”

[21] https://docs.gitlab.com/ee/ci/pipelines/

[23] https://www.treinaweb.com.br/blog/no-final-das-contas-o-que-e-o-docker-e-como-ele-funciona

[24] https://snakemake.readthedocs.io/en/stable/

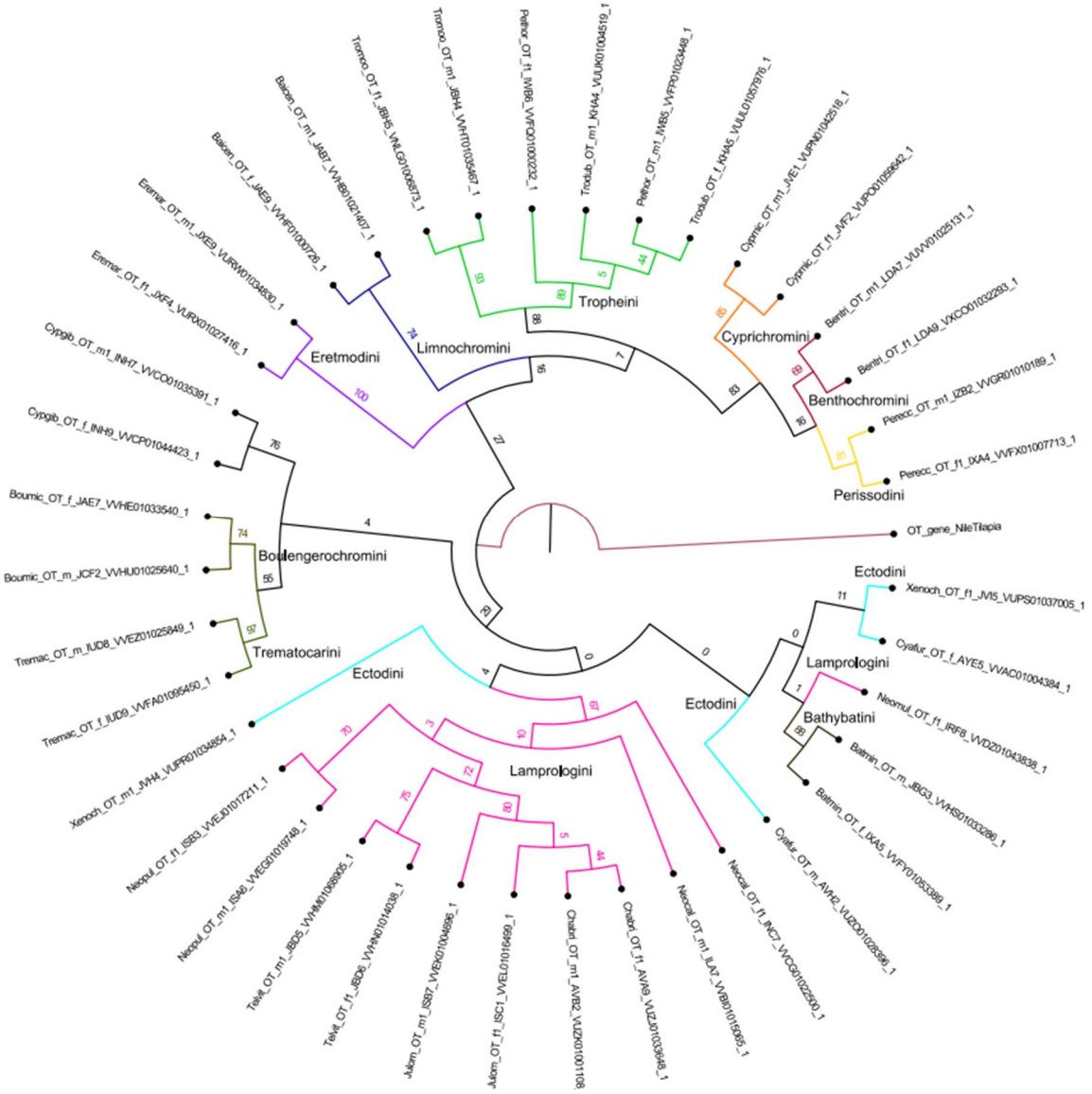
Attachments

The species list (Table 3) used to obtain the cichlid sequences is available in the Git repository: List_of_spp.xlsx

Table 3 - List of species (Excel file)

Species	*tribe (genus)*	*isolate*	*gender*	*Query cover*	*%ID gene*	*e-value*
Baileychromis	Limnochromin	JAB7	male	88%	97,06	0
				6%	87,75	2,00E-60
Baileychromis	Limnochromin	JAE9	female	88%	97,06	0
Bathybates mi	Bathybatini	JBG3	male	97%	95,82	0
Bathybates mi	Bathybatini	IXA5	female	97%	95,74	0
Benthochromi	Benthochromi	LDA7	male	91%	96,65	0
				6%	88,24	4,00E-62
Benthochromi	Benthochromi	LDA9	female	90%	96,65	0
				6%	87,75	2,00E-60

For better observation of the results obtained, the phylogenetic trees represented in Figures 4 and 5 are available in the same Git repository, along with the other phylogenetic trees obtained: https://gitlab.com/beatriz.fuente.santos/phylo_analysis/-/tree/main/trees



Supplementary Figure 1 - Phylogenetic tree obtained by the ML method of cichlid species from Lake Tanganyika based on the oxt gene using RAxML and PTHREADS (pipeline). The values in the branches correspond to the support values (bootstrap) in percentage. The colors represent the tribe to which they belong.

Link to Git *phylo_analysis* repository: https://gitlab.com/beatriz.fuente.santos/phylo_analysis