Instituto Superior de Engenharia de Lisboa

ML4Phylo: A Machine Learning Approach to Phylogenetic Analysis

Project and Seminary

BSc in Computer Science and Computer Engineering

Miguel Raposo, A49456@alunos.isel.pt, Nº 49456
Gonçalo Silva, A49451@alunos.isel.pt, Nº 49451

Supervisor: Cátia Vaz, cvaz@cc.isel.ipl.pt

March, 2024

1 Introduction

Phylogenetics is the study of evolutionary history and relationships among individuals, groups of organisms or other biological entities with evolutionary histories. [1]

Through phylogenetic analysis, a field of biomedical research that focuses on the study of these relationships, we are able to get a better understanding of the evolution of bacterial and viral epidemics. [2]

The task of inferring this history falls into what we call phylogenetic inference, which can be done by a series of different methods, resulting in a estimation referred to as a phylogenetic tree. Of these methods, the most computationally efficient ones are based on distances. However, to infer the tree it is necessary to traverse, at most, the upper triangle matrix, which implies that these algorithms are at least quadratic. Neighbor Joining (NJ), UPGMA and goeBURST are just a few of some traditional methods used. [1] NJ has a time complexity of $O(n^3)$, UPGMA, depending on the way implemented, has time complexity of $O(n^3)$, $O(n^2 logn)$ or $O(n^2)$ and goeBURST has time complexity of $O(n^2)$.

There are also alternative approaches based on machine learning such as Phyloformer [3], which generates a distance matrix, and Fusang [4], which only generates

quartets, that is, 4-leaf trees. Both receive a multiple sequence alignment (MSA) as input.

Through our project we aim to introduce a range of machine learning techniques with the objective of enhancing performance and efficiency of the inference process compared to other existing methods, like those given as an example.

In order to achieve our objectives we will be using the software Phyloformer, which is designed for fast and accurate phylogeny estimation using self-attention networks, as our basis.

Phyloformer itself is a transformer-based network architecture able to predict all the pairwise evolutionary distances between sequences given a multiple sequence alignment (MSA), allowing the reconstruction of the tree through them.

2 Requirements

In the context of our project, we propose to build a solution that includes the functionality of Phyloformer but accepts both MSAs and Typing Data as input data.

The software is available for use as a library for Python or through command-line tools that are installed with the package.

The following obligatory requirements were defined for this project:

1. Replicate the results obtained by the Phyloformer team.
2. Test Phyloformer, evaluating experimentally its scalability.
3. Implement the ML4Phylo solution, adapting from Phyloformer's.
    - Input can be Typing Data instead of MSA like used in Phyloformer.
    - It must be implemented or adapted an encoder for Typing Data.
    - It must be implemented or adapted a decoder for Typing Data.
4. Evaluate the results obtained by the ML4Phylo solution.
5. Release ML4Phylo as a Conda module or a Docker image.

As an optional requisite it is presented the following one:

1. Compare the generated tree against other phylogenetic inference algorithms not available in Phyloformer, like goeBURST.

In the evaluation part of the results produced by our solution, the goal is, according to the inferred phylogenetic tree, to compare it with others inferred by known algorithms, such as Maximum Likelihood. This comparison can be measured using the Robinson-Foulds metric, which is an integer value that quantifies the topological differences between pairs of trees.

3 Methodology

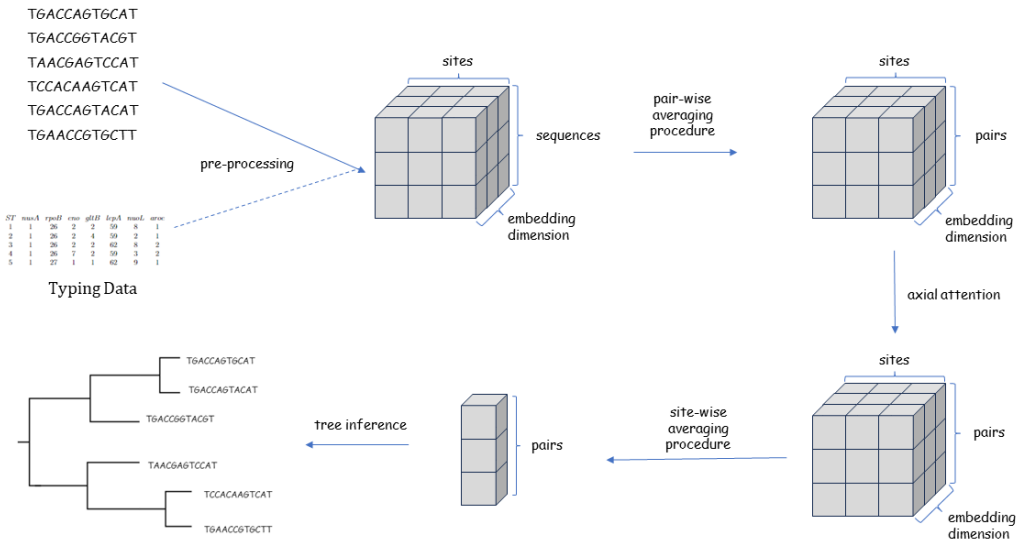

Figure 1: ML4Phylo/Phyloformer Pipeline - In the context of our project, the sites symbolize the identification of a position. If the input is sequences, it represents a letter (ACTG), but if it is Typing Data, it represents the identifier of a gene.

The process that leads to the inference of the phylogenetic tree is represented in the pipeline present in figure 1. This is divided into 4 main steps: the Input, the Pre-processing, the Model, and the Tree Inference.

The Input represents the initial data that is provided, which are sequences of the nucleic alphabet. Before reaching the model, these are subjected to a pre-processing phase where they are aligned, in MSAs, and transformed.

The model, upon receiving the MSAs, will apply an embedding of the data so that they can be understood by the model. These are then passed to the Axial Attention, which has a transformer-based architecture. A transformer consists of two major parts: the encoder and the decoder. Both work with a mechanism called self-attention, which is essential for the model's learning as it is through it that the model has the ability to focus on certain areas of the input in order to accelerate and facilitate the encoding or decoding process. In our case, the axial attention is composed of two layers of linear self-attention, one applied to the various positions of the pair representations and another applied to all pairs in a specific position, and a feed-forward network layer that represents a type of neural network that will propagate forward all the information learned in the previous layers. Linear self-attention is used because of its scalability, compared with common self-attention.

Finally, the data from the output of the transformer are grouped in order to condense the information contained in them, obtaining the evolutionary distances. Through these, in the Tree Inference step, the phylogenetic tree is inferred through a distance-based algorithm, such as, the Neighbor Joining (NJ).

4 Schedule

Date	Duration (weeks)	Assignments
21/02/2024	3	- Study and analyze Phyloformer and phylogenetic inference
06/03/2024	1	- Write project proposal
13/03/2023	2	- Finish and deliver the project proposal - Prepare project proposals presentation - Exploring Python Language with PyTorch package
20/03/2024	4	- Look into and get familiarized with Phyloformers code - Replicate Phyloformers team results and evaluate experimentally its scalability
17/04/2024	1	- Begin writing final report
24/04/2024	1	- Progress presentation
01/05/2024	3	- Implement or adapt an encoder and decoder for Typing Data
22/05/2024	2	- Evaluate the results obtained by the ML4Phylo solution - Prepare Beta version for delivery
05/06/2024	1	- Deliver Beta version
12/06/2024	2	- Release ML4Phylo as a Conda module or a Docker image - Finalize report
26/06/2024	1	- Final delivery

References

[1] Cátia Vaz, Marta Nascimento, João A Carriço, Tatiana Rocher, and Alexandre P Francisco. Distance-based phylogenetic inference from typing data: a unifying view. *Briefings in Bioinformatics*, 22(3):bbaa147, 07 2020.

[2] André Jesus, Nyckollas Brandão, and André Páscoa. Phyloviz web platform. Available: `https://github.com/phyloviz/phyloviz-web-platform/blob/master/docs/final-presentation.pdf`, 2023.

[3] Luca Nesterenko, Bastien Boussau, and Laurent Jacob. Phyloformer: towards fast and accurate phylogeny estimation with self-attention networks. *bioRxiv*, pages 2022–06, 2022.

[4] Zhicheng Wang, Jinnan Sun, Yuan Gao, Yongwei Xue, Yubo Zhang, Kuan Li, Wei Zhang, Chi Zhang, Jian Zu, and Li Zhang. Fusang: a framework for phylogenetic tree inference via deep learning. *Nucleic Acids Research*, 51(20):10909–10923, 2023.