Instituto Superior de Engenharia de Lisboa

BSc in Computer Science and Computer Engineering

PHYLOViZ Web Platform:
A Modular and Web-Based Tool
for Phylogenetic Analysis

Project and Seminary

André Jesus, A48280@alunos.isel.pt, 930 533 735

André Páscoa, A48089@alunos.isel.pt, 962 249 051

Nyckollas Brandão, A48287@alunos.isel.pt, 963 209 503

Supervisors:

Cátia Vaz, ISEL, cvaz@cc.isel.ipl.pt

Alexandre P. Francisco, IST, aplf@ist.utl.pt

March 20, 2023

1   Introduction

Modern biomedical research has seen a remarkable increase in the production and computational analysis of large datasets, leading to an urgent need of tools for data integration, data processing and visualization, as well as the need of sharing standardized analytical techniques. One important field of biomedical research is **phylogenetic analysis**, which allow to understand the evolution of bacterial and viral epidemics, such as SARS-CoV-2 or some E. coli strains, in order to determine their origin, evolution and resistance to the various treatments under study.

Phylogenetic analyses comprise in general the following steps: 1) the **alignment of genetic sequences**, 2) the application of a **typing methodology**, 3) the application of **phylogenetic inference methods**, and finally 4) the **visualization** and exploration of inference results, integrating epidemiological and ancillary data for isolates under study.

Several approaches to support phylogenetic analysis have been proposed, varying from standalone tools to integrative web applications that include tools and/or algorithms for executing the common analysis tasks for these kind of data. **PHYLOViZ** [1, 2] and **PHYLOViZ Online** [3], provide phylogenetic analysis processing, namely inference methods, visualization and integration of epidemiological and ancillary data. Nevertheless, features in both PHYLOViZ versions (desktop and web) are not the same, and the web version is not modular, which makes it difficult to extend and maintain. Also, inference and/or visualization optimization tasks are run on the client side, which becomes often unfeasible for large amounts of data as well as data must also be transferred from existing databases in order to be analysed. A prototype, **phyloDB** [4], was developed for efficiently managing phylogenetic data, supporting, for instance the deployment of algorithms for inferring/detecting patterns and for pre-computing and optimizing phylogenetic visualizations. Other prototype, **FLOWViZ** [5] was developed to support the usage of a workflow system for processing data.

This project aims to develop a new version of PHYLOViZ Online - **PHYLOViZ Web Platform** - with a **modular architecture**, that unifies both versions of PHYLOViZ and allows for the use of tools such as phyloDB and FLOWViZ, not being compromised by these. With its modular architecture, other modules can be used and easily integrated in the future, so that the platform can be extended to support new analysis methods and tools.

The PHYLOViZ Web Platform will allow users to access and perform phylogenetic analyses and visualizations from anywhere with an internet connection, without requiring installation of software or access to high-performance computing resources. This is achievable by performing most computing operations and data management server-side.

The modular architecture will allow for easy integration of new analysis methods and tools, making it a **flexible** and **extensible** platform for biomedical researchers. For instance, FLOWViZ could be used to enable the creation and execution of customized analysis workflows, while the phyloDB prototype for efficient storage and retrieval of large phylogenetic datasets.

Overall, the development of the PHYLOViZ Web Platform with a modular architecture and advanced data management and optimization capabilities represents a significant step forward in the field of phylogenetic analysis. By providing a user-friendly, web-based platform that can handle large and complex datasets with ease, this new tool will help researchers to better understand the evolution of bacterial and viral epidemics and ultimately contribute to the development of new treatments and preventative measures.

2 Requirements

To develop the PHYLOViZ Web Platform, the challenge of designing a modular architecture that allows for the integration of pre-existing prototypes like phyloDB and FLOWViZ needs to be overcome.

The following list of requirements is a summary of the main tasks that need to be accomplished:

1. **Design a modular architecture** - The architecture should allow for the integration of new modules and reuse of existing ones. The backend should contain modules, referred to as **microservices**, for the tasks of **data processing**, **data visualization**, and **data integration**. Each of these microservices will be responsible for a designated function:

   - Creating workflows with tools that execute the phylogenetic algorithms and store the data appropriately.
   - Allowing the user to upload and download data.
   - Providing the user with the data for visualization within the application, supporting multiple output formats and allowing users to customize their views - apply **filters**.

   The implementation of these modules should prioritize the use of existing prototypes, in order to reuse existing tools and to avoid re-implementing the same functionalities. Other microservices may be implemented that encompass features that don't fall under the category of the aforementioned microservices.

2. **Effective data management** - The storage solution should ensure the efficient handling of large datasets:

   - Implement data quality control and validation mechanisms to ensure data integrity and consistency. This step should include error-checking of inputs, validation of outputs, and data normalization. This solution should consider both the storage and retrieval of data, including integration with existing databases.
   - Creation of a personal project by the user and exploration of files within the project's file structure, allowing the user to compute, view, or download data (the appropriate microservices will be called).

3. **Design a client application (frontend)** - The frontend should communicate with the backend, through the API of each individual microservice:

   - Multiple microservices may be called for a singular action from the user.
   - Optimization of the tree is to be done in the frontend, because of latency. For this reason, a way of showing, loading, and computing the data partly needs to be implemented.

The security and data privacy measures will be implemented through modern authentication methods.

In addition to this, testing and validation must be prioritized to ensure that the platform is robust and reliable, covering unit and integration tests.

Optionally, modifications to the existing prototypes could be done in order to improve the efficiency of some algorithms.

The project report and the wiki are updated regularly with the progress of the project.

3 Schedule

Date	Duration (weeks)	Assignments
20/02/2023	2	- Study and analyze PHYLOViZ, PHYLOViZ Online and other related tools (phyloDB, phylolib and FLOWViZ)
28/02/2023	2	- Write project proposal - Set up development environment - Design the architecture
13/03/2023	1	- Finish and deliver the project proposal - Finish the architecture design
20/03/2023	3	- Implement the application backend (microservices and other components of the server-side)
10/04/2023	1	- Test the application backend modules - Improve project documentation
17/04/2023	1	- Prepare progress presentation
24/04/2023	1	- Progress presentation
01/05/2023	3	- Implement the application frontend
22/05/2023	2	- Test the application frontend modules - Prepare Beta version for delivery
05/06/2023	1	- Deliver Beta version and finish the report
15/06/2023	1	- Final delivery

References

[1] A. P. Francisco, C. Vaz, P. T. Monteiro, J. Melo-Cristino, M. Ramirez, and J. A. Carriço, "Phyloviz: phylogenetic inference and data visualization for sequence based typing methods," *BMC bioinformatics*, vol. 13, pp. 1–10, 2012.

[2] M. Nascimento, A. Sousa, M. Ramirez, A. P. Francisco, J. A. Carriço, and C. Vaz, "Phyloviz 2.0: providing scalable data integration and visualization for multiple phylogenetic inference methods," *Bioinformatics*, vol. 33, no. 1, pp. 128–129, 2017.

[3] B. Ribeiro-Gonçalves, A. P. Francisco, C. Vaz, M. Ramirez, and J. A. Carriço, "Phyloviz online: web-based tool for visualization, phylogenetic inference, analysis and sharing of minimum spanning trees," *Nucleic acids research*, vol. 44, no. W1, pp. W246–W251, 2016.

[4] B. Lourenço, "A framework for large scale phylogenetic analysis," *arXiv:2012.13363 [q-bio]*, 01 2021. [Online]. Available: https://arxiv.org/abs/2012.13363

[5] M. Luis and C. Vaz, "Flowviz: Framework for phylogenetic processing," *arXiv:2211.15282 [cs]*, 11 2022. [Online]. Available: https://arxiv.org/abs/2211.15282