Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

# Master project 2024-2025

## Personal Information


| | |
|---|---|
| **Supervisor** | Marco Mariotti |
| **Email** | marco.mariotti@ub.edu |
| **Institution** | Universitat de Barcelona |
| **Website** | https://www.mariottigenomicslab.com/ |
| **Group** | Comparative Genomics of Gene Expression |


## Project

# Web development & bioinformatic tools

**Project Title:**

Development of computational tools for bioinformatics: PyRanges and Treedex


**Keywords:**

Software development; Programming; Data science; Pandas; Data visualization


**Summary:**


Our lab is developing two tools intended for a broad bioinformatics audience: PyRanges and Treedex. They are both in python and share some requirements in expertise. The student may participate in the development of one, the other, or both. 1. PyRanges Genomic annotations are a key tool in the field of computational, molecular, and evolutionary biology. They provide detailed information about the structure and function of genes, as well as their regulatory elements. Annotations are essential for many bioinformatics fields, e.g. RNAseq quantification, variant effect prediction, orthology assignment, and many others. PyRanges is a recently developed python library1 to handle genomic annotations and any similarly-shaped data, i.e. tables of intervals whose coordinates refer to a biological sequence. PyRanges is built on top of Pandas2 and allows to efficiently compute subsequences, overlaps, and other commonly used operations. Though PyRanges has been originally developed by another lab, our group is now part of the core development team. We are pursuing development in several directions, including new functionalities as well as a new graphical interface to visualize annotations. 2. Treedex High-throughput “omics” techniques such as next generation sequencing and mass spectrometry can yield comprehensive molecular profiles, providing informative snapshots of the genome-wide activity and regulation of cells. When the data comes from a multitude of species (which we refer to as “comparative data”), there is an important complication: the phylogenetic dimension, i.e., the fact that all species are related by a specific tree-like structure called phylogeny. In any analysis of comparative data, phylogeny must be taken into account at all times, since it dictates the fundamental architecture of what we measure (3). We are developing a novel framework for data visualization and analysis oriented to comparative omics, called Treedex (Tree Data explorer). This tool has two main objectives: • Facilitate the interactive exploration of comparative data of any magnitude and type, creating an intuitive link between the features under consideration and the phylogeny of species. • Integrate a state-of-the-art methodologies from evolutionary/comparative biology to be readily available within Treedex. This “comparative omics toolkit” will mainly focus on methods of evolutionary inference such as phylogenetic profiling, where we want to discover the hidden functional links among measured features (e.g., reconstructing the functional pathways of genes based on evolutionary patterns 4). Treedex is being developed as a module of the ETE4 framework (http://etetoolkit.org/), combined with Plotly/Dash and Pandas (2,5,6). The students taking part in this project will participate by creating either back-end functionalities (i.e., how evolutionary methods are run under the hood) or working on the front-end (i.e., how plots look like and how user-interaction is implemented).


**References:**


1. Stovner, E. B. & Sætrom, P. PyRanges: efficient comparison of genomic intervals in Python. Bioinformatics 36, 918–919 (2020). 2. Pandas: Python Data Analysis Library. http://pandas.pydata.org/. 3. Felsenstein, J. Phylogenies and the Comparative Method. Am Nat 125, 1–15 (1985). 4. Kensche, P. R., van Noort, V., Dutilh, B. E. & Huynen, M. A. Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution. J R Soc Interface 5, 151–70 (2008). 5. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Mol Biol Evol 33, 1635–8 (2016). 6. Plotly Technologies Inc. Collaborative data science. Plotly Technologies Inc.

**Expected skills:**

Python; Pandas; git / github; required for Treedex: basics of evolutionary biology.

**Possibility of funding:**

To be discussed

**Possible continuity with PhD:**

To be discussed