Universitat Pompeu Fabra Barcelona | Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Marco Mariotti
Email	marco.mariotti.mm@gmail.com
Institution	Universitat de Barcelona
Website	https://www.mariottigenomicslab.com/
Group	Comparative Genomics and Recoding lab

Project

Web development & bioinformatic tools

Project Title:

Building Treedex, an interactive framework for visualization and analysis of 'omics data from multiple species

Keywords:

Software development; Programming; Data visualization; Comparative Genomics; Evolution

Summary:

High-throughput "omics" techniques such as next generation sequencing (1) and mass spectrometry (2) can yield comprehensive molecular profiles, providing informative snapshots of the genome-wide activity and regulation of cells. The magnitude of omics data is a challenge for the mind of any researcher: no human brain can truly grasp datasets comprising thousands of entities (e.g., genes), so that we necessarily rely on computational methods to make sense of data. This is true both for the actual data analysis and for its visualization, essential to summarize and distil data in forms that we can perceive and comprehend. When the data comes from a multitude of species (which we refer to as "comparative data"), there is an additional 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). In the lab, we are developing a novel framework for data visualization and analysis oriented to comparative omics, called Treedex (Tree Data explorer; see https://www.treedex.org/). 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. This will allow studying the evolution at large timescales of many fundamental molecular traits, such gene expression and metabolic regulation. • Integrate a wide set of state-of-the-art methodologies from evolutionary/comparative biology, readily available to be applied within Treedex. This "comparative omics toolkit" will mainly focus on methods of evolutionary inference, where we want to discover the hidden functional links among measured features (e.g., reconstructing the functional pathways of genes based on evolutionary patterns). This includes tools such as phylogenetic profiling, protein coevolution finding, phylogenetic regression and many others (4–8). Treedex is designed as a stand-alone application compatible with the most common operating systems. It is built using Python, R, PyQt, ETE3, PyQtGraph, and Pandas (9–12). It is at the stage of a working prototype, currently under development in a modular fashion. 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). Besides, students will utilize Treedex to analyze actual comparative data, in particular focusing on gene expression across mammals (13–15).

References:

1. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. 17, 333–51 (2016). 2. Girolamo, F., Lante, I., Muraca, M. & Putignani, L. The Role of Mass Spectrometry in the "Omics" Era. Curr. Org. Chem. 17, 2891–2905 (2013). 3. Felsenstein, J. Phylogenies and the Comparative Method. Am. Nat. 125, 1–15 (1985). 4. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc. Natl. Acad. Sci. U. S. A. 96, 4285–8 (1999). 5. 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). 6. Grafen, A. The phylogenetic regression. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 326, 119–57 (1989). 7. Marks, D. S., Hopf, T. A. & Sander, C. Protein structure prediction from sequence variation. Nat. Biotechnol. 30, 1072–1080 (2012). 8. Weinreb, C. et al. 3D RNA and Functional Interactions from Evolutionary

Couplings. Cell 165, 963–975 (2016). 9. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Mol. Biol. Evol. 33, 1635–8 (2016). 10. PyQt: python binding for Qt. Available at: https://riverbankcomputing.com/software/pyqt/intro. 11. PyQtGraph: Scientific Graphics and GUI Library for Python. Available at: http://www.pyqtgraph.org/. 12. Pandas: Python Data Analysis Library. Available at: http://pandas.pydata.org/. 13. Fushan, A. A. et al. Gene expression defines natural changes in mammalian lifespan. Aging Cell 14, 352–65 (2015). 14. Cardoso-Moreira, M. et al. Gene expression across mammalian organ development. Nature 571, 505–509 (2019). 15. Wang, Z. Y. et al. Transcriptome and translatome co-evolution in mammals. Nature 588, 642–647 (2020).

Expected skills::

Required: Python; Data visualization (i.e., plotting by coding); Basics of evolutionary biology. Useful: some experience in Pandas; PyQt (or any graphical user interface toolkit).

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed