Universitat Pompeu Fabra Barcelona | Master in Bioinformatics for Health Sciences

Master project 2024-2025

Personal Information

Supervisor	Miguel Romero
Email	miguel.romero@bsc.es
Institution	Barcelona Supercomputing Center
Website	https://www.bsc.es/discover-bsc/organisation/scientific-structure/computational-biology
Group	Computational Biology Group

Project

Structural bioinformatics

Project Title:

Harnessing protein language models in virus research

Keywords:

Protein language models, virus research, molecular evolution, machine learning.

Summary:

Proteins are key players in most of the biological processes that take place within our cells. Studying them is essential to understanding cellular processes at a molecular level, exploring the development of diseases, and designing drugs and therapies to fight against them. In recent years, protein research has undergone a revolution thanks to advances in experimental techniques that have brought vast amounts of data and computational developments at the hardware and software level that have allowed its analysis and exploitation (Pearce and Zhang 2021; Kuhlman and Bradley 2019). The adoption of machine learning (ML) and artificial intelligence (AI) techniques is proving exceptionally fruitful (Bordin et al. 2023). Specifically, one of the most promising strategies is using language models for protein research. Protein Language Models (pLMs) trained with millions of protein sequences are used to successfully generate new protein sequences and predict features, such as their structure or function (Rives et al. 2021; Bepler and Berger 2021; Lin et al. 2022). These capabilities give pLMs enormous potential for drug development and protein engineering (Ferruz and Höcker 2022). In this project, we want to apply pLMs to the field of virus research by leveraging their capabilities to study viral proteins, facilitating our understanding of virus-host interactions and the development of therapeutic strategies. More specifically, we will be interested in: - Analyzing virus genomes to identify both conserved and variable regions susceptible to mutations. These analyses will be instrumental in developing drugs and vaccines designed to prevent virus escape mechanisms, enhancing their long-term efficacy. - Leveraging pLMs to predict the pathogenicity of virus variants. - Mapping the evolutionary trajectory of the virus. We aim to understand the evolutionary path followed by the virus up to the present day and forecast potential future paths. What you will learn: - Protocols to fine-tune and apply pLMs for studying relevant biological problems involving proteins. - Techniques to perform sequence and structural-based analysis of proteins. - Methods to study virus/host interactions at the molecular level. - State-of-the-art methods for protein design. - Collaborate in the preparation and presentation of research projects.

References:

Bepler, Tristan, and Bonnie Berger. 2021. "Learning the Protein Language: Evolution, Structure, and Function." Cell Systems 12 (6): 654-669.e3. https://doi.org/10.1016/j.cels.2021.05.017. Bordin, Nicola, Christian Dallago, Michael Heinzinger, Stephanie Kim, Maria Littmann, Clemens Rauer, Martin Steinegger, Burkhard Rost, and Christine Orengo. 2023. "Novel Machine Learning Approaches Revolutionize Protein Knowledge." Trends in Biochemical Sciences 48 (4): 345–59. https://doi.org/10.1016/j.tibs.2022.11.001. Ferruz, Noelia, and Birte Höcker. 2022. "Controllable Protein Design with Language Models." Nature Machine Intelligence 4 (6): 521–32. https://doi.org/10.1038/s42256-022-00499-z. Hie, Brian, Ellen D. Zhong, Bonnie Berger, and Bryan Bryson. 2021. "Learning the Language of Viral Evolution and Escape." Science (New York, N.Y.) 371 (6526): 284–88. https://doi.org/10.1126/science.abd7331. Kuhlman, Brian, and Philip Bradley. 2019. "Advances in Protein Structure Prediction and Design." Nature Reviews Molecular Cell Biology 20 (11): 681–97. https://doi.org/10.1038/s41580-019-0163-x. Lin, Zeming, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, et al. 2022. "Evolutionary-Scale Prediction of Atomic Level Protein Structure with a Language Model." bioRxiv. https://doi.org/10.1101/2022.07.20.500902. Marquet, Céline,

Michael Heinzinger, Tobias Olenyi, Christian Dallago, Kyra Erckert, Michael Bernhofer, Dmitrii Nechaev, and Burkhard Rost. 2022. “Embeddings from Protein Language Models Predict Conservation and Variant Effects.” Human Genetics 141 (10): 1629–47. https://doi.org/10.1007/s00439-021-02411-y. Pearce, Robin, and Yang Zhang. 2021. "Deep Learning Techniques Have Significantly Impacted Protein Structure Prediction and Protein Design." Current Opinion in Structural Biology, Protein-Carbohydrate Complexes and Glycosylation ● Sequences and Topology, 68 (June): 194–207. https://doi.org/10.1016/j.sbi.2021.01.007. Rives, Alexander, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, et al. 2021. "Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences." Proceedings of the National Academy of Sciences 118 (15): e2016239118. https://doi.org/10.1073/pnas.2016239118.

Expected skills:

1- Good programming skills (Python, Pytorch). 2 - Basic knowledge of molecular biology. 3- Strong interest in analyzing proteins at different levels (sequence, structure, function). 4- Ability to access and evaluate scientific literature.

Possibility of funding:

Yes

Possible continuity with PhD:

To be discussed