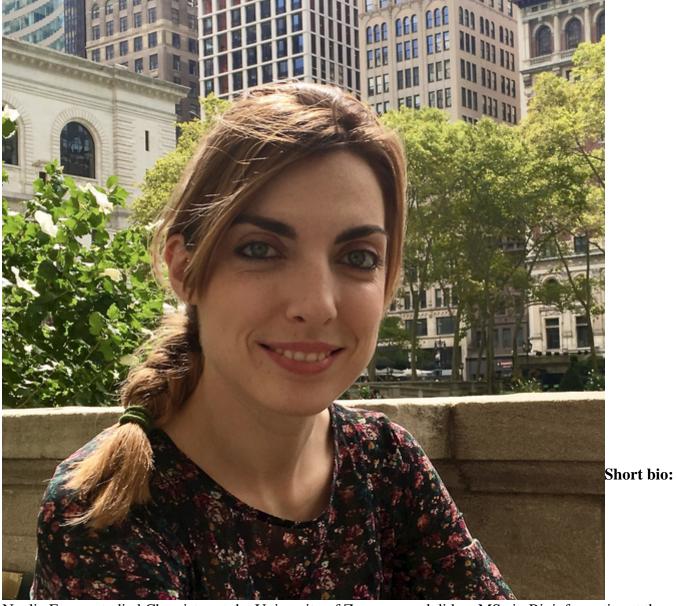# Hybrid BSC RS/Life Session/BioInfo4Women Seminar: Towards controlled protein design with deep unsupervised models

Objectives

Abstract: The design of proteins with specific functions has the potential to tackle biomedical, environmental, and industrial challenges in a biodegradable and cost-effective manner. Traditional protein design approaches have relied on finding the global energy minima of a multidimensional landscape defined with physicochemical based energy functions1. In this sense, we recently developed Fuzzle2,3, a database of reused protein fragments that Nature has reused over the course of evolution. These fragments are amenable for large-scale chimeragenesis with Protlego4, thus laying the background for designing novel functions by combining protein blocks in a Lego-like manner. Nevertheless, in recent years we are witnessing an explosion of Artificial Intelligence (AI) methods that are impacting virtually all areas of research and our daily lives. Natural Language Processing (NLP) is producing models capable of translating, understanding, and generating text with human capabilities. Given the many similarities between human languages and protein sequences5, using NLP methods for protein research opens a new unexploited door for protein design. Recently, inspired by the GPT-x language models, we trained ProtGPT2, a deep unsupervised language model that has learned the protein language upon being trained on the entire protein space6. ProtGPT2 is capable of generating protein sequences in unseen regions of the protein space while preserving natural-like properties. The inclusion of annotation tags during training will allow the directed generation of specific functions. Coupling the generation process to methods like high-throughput molecular dynamics7,8 will enable variant selection before experimentations. Recent developments in AI methods and their possible impact on protein design will be discussed.

1. Lechner, H., Ferruz, N. & Höcker, B. Strategies for designing non-natural enzymes and binders. Current Opinion in Chemical Biology vol. 47 67–76 (2018).
2. Ferruz, N. et al. Identification and Analysis of Natural Building Blocks for Evolution-Guided Fragment-Based Protein Design. J. Mol. Biol. 432, 3898–3914 (2020).
3. Ferruz, N., Michel, F., Lobos, F., Schmidt, S. & Höcker, B. Fuzzle 2.0: Ligand Binding in Natural Protein Building Blocks. Front. Mol. Biosci. 8, 805 (2021).
4. Ferruz, N., Noske, J. & Höcker, B. Protlego: A Python package for the analysis and design of chimeric proteins. Bioinformatics (2021) doi:10.1093/bioinformatics/btab253.
5. Ferruz, N. & Höcker, B. Controllable Protein design with Language Models. Nat. Mach. Int, manuscript accepted. (2022).
6. Ferruz, N., Schmidt, S. & Höcker, B. A deep unsupervised language model for protein design. bioRxiv 2022.03.09.483666 (2022) doi:10.1101/2022.03.09.483666.
7. Ferruz, N., Harvey, M. J., Mestres, J. & De Fabritiis, G. Insights from Fragment Hit Binding Assays by Molecular Simulations. J. Chem. Inf. Model. 55, 2200–2205 (2015).
8. Ferruz, N. et al. Dopamine D3 receptor antagonist reveals a cryptic pocket in aminergic GPCRs. Sci. Rep. 8, 1–10 (2018).

Short bio:

Noelia Ferruz studied Chemistry at the University of Zaragoza and did an MSc in Bioinformatics at the University Pompeu Fabra (Barcelona, Spain). She undertook her PhD in the field of Computational Biophysics at the Barcelona Biomedical Research Park (PRBB). After a short stay as a postdoctoral researcher in the neuroscience department in Pfizer, Boston, she did her postdoc in Bayreuth, Germany, in the protein design group of Prof. Birte Höcker. Since April, she has been a Beatriu de Pinós Fellow at the Institute of Informatics and Applications at the University of Girona, where she focuses on implementing neural models for protein engineering and design.

Speakers

Speaker: Noelia Ferruz, Beatriu de Pinós Fellow at the Institute of Informatics and Applications at the University of Girona
Host: Alfonso Valencia, BSC Life Sciences department director
Barcelona Supercomputing Center - Centro Nacional de Supercomputación