SORS/WomenInBSC: Decoupling Function from Evolutionary Conservation using protein Language Models

Abstract

Function annotation is a challenging problem in Computational Biology. Relying on evolutionary relationships is often suboptimal for function assignment. In a collaborative work, we have extensively tested various deep learning-based methods (CNNs and LMs) on full proteomes to assess their performance at the organism level. We found that transformer-based protein language models are more precise and informative than other methods for all the species tested and across the three gene ontologies studied. They also better recover functional information from transcriptomic experiments. We applied the best methods to annotate 1,000 species from the animal phyla, and have produced the FANTASIA pipeline to finding that functional recovery can effectively address experimental hypotheses.

Short Bio

Ana M. Rojas is a CSIC Research Scientist. Her main expertise focuses on protein evolution and function. A biologist by training in Madrid and the US, she specialized in Bioinformatics and Computational Biology in various labs in the USA (under a NASA-NSCORT fellowship) and in Spain (under a Marie Curie fellowship at CNB-CSIC). Later, as a staff scientist at CNIO, she established her independent group in 2009 in Badalona and moved to the Institute of Biomedicine of Seville in 2013. She subsequently relocated her group to the Andalusian Center for Developmental Biology (CABD) in Seville. She has been a Track Chair of the ISCB Function COSI since 2024, a founding member of the Spanish Society of Computational Biology (SEBIBiC) launched in 2020, and serves on the executive committee of the CSIC.BCBHub (2023), a network of CSIC computational biologists in Spain. She is a researcher at the ENIA-Chair USE-Google Spain for AI and is also very active in several outreach activities. Her current research interests focus on understanding the complex relationships among sequence, structure, and function, particularly addressing multifunctional aspects of proteins relevant to biotechnology (biosensors) and biomedicine (therapeutic drugs) using AI-based techniques.

Speakers

Speaker: Ana M. Rojas. CSIC Research Scientist
Host: Alfonso Valencia. Life Sciences Department Director, BSC
Barcelona Supercomputing Center - Centro Nacional de Supercomputación