Progetto: "Development of computational tools for functional characterization of human health data"

Proponente: Dr. Castrense Savojardo

Progetto di ricerca: base di partenza scientifica ed obiettivi

Despite the significant improvements in both speed and accuracy of sequencing and expression characterization techniques (e.g. transcriptomics, proteomics), functional experimental characterization of proteins remains a challenging issue. Gene function knowledge is limited even for the human organism, as a large fraction of coding regions is completely missing any annotation. Moreover, despite the unprecedented amount of genetic variation data available in public databases, our ability to interpret these data in terms of the impact on phenotypes is still hampered by the lack of proper tools for organizing and analyzing them.
The project proposed here will be developed in the context of PRIN2017 project (CUP J34I19001610005) entitled "Protein bioinformatics for human health" whose general aim is to tackle protein functional characterization by an entirely computational approach for improving the quality and interoperability of the available software and expand current functional knowledge on "dark" regions. One of the objectives of the project include the development of novel methods for functional characterization of human proteins and their variations with respect to different phenotypes, including changes in protein localization and interactions as well as disease onset. Proteins are primary players in the cell and their functional characterization is indispensable to understand life at the molecular and mechanistic level. With high-throughput Omics experiments producing ever expanding quantities of raw data, their translation into meaningful biological information is becoming increasingly complicated.
The aim of the proposed project, within the general objectives of the PRIN project, is to develop and test novel tools for the annotation of protein function and subcellular localization. Specifically, the annotation of protein subcellular compartment from sequence requires the integration of different sources of information and the adoption of cutting-edge artificial intelligence approaches.
The research will build on top of existing tools developed by the Bologna Biocomputing Group for protein function and localization annotation, such as the Bologna Annotation Resource (BAR3) [1] and BUSCA [2]. These tools will be updated and revised in the context of the proposed research with the goal of improving the functional characterization of proteins. Moreover, taking advantage of new data deposited in public databases such as UniProtKB, new computational tools will be developed for extending the ability of predictors to previously uncovered subcellular compartments and to characterize important cellular targeting processes driven by protein modifications such as protein myristoylation and/or prenylation.

Articolazione del progetto e tempi di realizzazione

The proposed is articulated in two main tasks:

Task 1) Updating and improving existing function and subcellular localization prediction tools (months 1-6)

New, high-quality datasets will be extracted from public sequence databases such as UniProtKB. These datasets will serve as training and independent testing data for updated and novel function and localization annotation tools. Different machine-learning approaches will be tested to implement tool for extending prediction capabilities to previously uncovered cellular compartments e.g. at the sub-nuclear level. Tools developed will be integrated into existing pipelines such as BUSCA (https://busca.biocomp.unibo.it) for large-scale annotation of protein function and subcellular

localization. The new methods will be adopted for the genome-wide analysis of the full proteome of *Neisseria meningitidis* (UP000000425) and results integrated into the PRIN2017 project repository.

Task 2) Development of computational tools for protein lipidation prediction (months 6-12)

Protein lipidations have a role in many biological processes including signal transduction, apoptosis and pathological processes induced by viruses and fungi. They represent about 2% of the proteins in eukaryotic proteomes. Several experimental methods are available to study lipidation *in vivo* or *in vitro*. In parallel to experimental methods, several *in silico* approaches have been developed, tackling the problem of lipidated proteins detection. The aim of this task is to take advantage of new datasets available in public database and from proteomics studies for the development of a new computational tool for protein lipidation detection and characterization.

Programma formativo

The researcher dedicated to this project must have experience in implementing and optimizing computational systems for biological data analysis and annotation.

Specific knowledge is required about computer programming with particular focus on technologies and software frameworks for the definition, querying and maintenance of databases.

Her/his work will be mainly devoted on the curation of tools for large-scale prediction/annotation of protein function and subcellular localization.

In particular, the researcher will be involved in the following activities:

1) Updating and improving existing function and subcellular localization prediction tools

The annotation of the function of the proteins encoded by a genome, including their subcellular localization, is essential for understanding their role in biological processes and for inferring the potential role of variations detected in specific individuals. In the context of this task, the researcher will leverage on previously developed tools for protein localization prediction such as BaCelLo [3], SChloro [4], DeepMito [5] and BetAware-Deep [6]. These tools will be updated and possibly retrained on new data extracted from public databases such as UniProtKB. This will allow to i) improve tool prediction performance over the state-of-the-art and ii) to extend their predictive capabilities to previously uncovered compartments. Different machine-learning approaches will be evaluated for the development of novel and improved tools. Novel and updated tools will be all integrated into BUSCA (http://busca.biocomp.unibo.it), an existing pipeline developed and maintained at the Biocomputing Group for large-scale annotation of protein localization. Moreover, the updated BUSCA pipeline will be used to complement the Bologna Annotation Resource (BAR, https://bar.biocomp.unibo.it/bar3/) in protein function prediction, for the annotation of GO terms in the Cellular Component aspect. All developed pipelines will be evaluated on real datasets of interest, including those extracted from international community-wide benchmarking experiments like the Critical Assessment of protein Function Annotation algorithms (CAFA, https://www.biofunctionprediction.org/cafa/). Genome-wide analysis will be performed on the full proteome of *Neisseria meningitidis* (UP000000425), and results integrated into the PRIN2017 project repository.

2) Development of computational tools for protein myristoylation and prenylation prediction

In the context of this task, the researcher will collect experimental data about lipidation (including myristoylation, prenylation and others) from public sequence databases such as UniProtKB. These data will be complemented with data published in the context of large-scale proteomic studies. Different machine-learning approaches, including Support Vector Machines and Deep Artificial Neural Networks, will be evaluated for the development of novel and improved tools for *in-silico* protein lipidation detection. The tool developed will be integrated into annotation workflows already developed at the Biocomputing Group.

The proposed research will deliver:

- Updated tools and pipelines for automatic functional annotation of proteins
- Scientific papers describing the adopted procedures and the results obtained with the new systems.

The researcher will present results at major international conferences in the area of Computational Biology.

References

1. Profiti G, Martelli PL, Casadio R. The Bologna Annotation Resource (BAR 3.0): improving protein functional annotation. Nucleic Acids Res. 2017;45:W285–90.

2. Savojardo C, Martelli PL, Fariselli P, Profiti G, Casadio R. BUSCA: an integrative web server to predict subcellular localization of proteins. Nucleic Acids Res. 2018;46:W459–66.

3. Pierleoni A, Martelli PL, Fariselli P, Casadio R. BaCelLo: a balanced subcellular localization predictor. Bioinforma Oxf Engl. 2006;22:e408-416.

4. Savojardo C, Martelli PL, Fariselli P, Casadio R. SChloro: directing Viridiplantae proteins to six chloroplastic sub-compartments. Bioinforma Oxf Engl. 2017;33:347–53.

5. Savojardo C, Bruciaferri N, Tartari G, Martelli PL, Casadio R. DeepMito: accurate prediction of protein sub-mitochondrial localization using convolutional neural networks. Bioinformatics. 2020;36(1):56-64.

6. Madeo G, Savojardo C, Martelli PL, Casadio R. BetAware-Deep: An Accurate Web Server for Discrimination and Topology Prediction of Prokaryotic Transmembrane β-barrel Proteins. Journal of Molecular Biology. 2020; 433(11):166729.