Project Proposal

Systems Biology knowledge generation using rule induction techniques applied to Multi Omics Data

Biological systems (BSs) live immerged in a complex environment with which they interact receiving stimuli and exacerbating (pathological) phenotypical responses. The emergence of the phenotype is driven by a complex chain of interactions among different molecular layers and the comprehension of these interactions is crucial in the study of pathological conditions (Scala et Al. 2019) as well as in the prediction of the (toxical/pharmacological) effects of novel molecules.
In eukaryotes, this internal mediation process is complex and involves several molecular structures that, in turn, interact at different levels. Here, at least six molecular layers need to be considered: 1) the DNA sequence, 2) the DNA epigenetic modifications, 3) the chromatin packaging modifications, 4) the binding transcription factors, 5) the targeting miRNAs, 6) mRNA levels.
Huge collections of molecular data for several BSs in different condition are currently available in different public data sources (ENCODE, GTEx, OMICS DI, …).
Previous studies have shown the effective potential of the usage of machine learning algorithms in the prediction of particular phenotypes starting from molecular configurations (Subramanian et Al 2020). These studies have been performed considering specific BSs and molecular configurations and their outcomes often consist of *black-box* systems computing a mathematical function from molecular patterns to phenotypes. In many application fields, especially in systems biology, by following the three principles of *transparency*, *interpretability*, and *explainability*, it is extremely important and useful to also have a description of the rules governing the association between the dependent and the independent variables along with the prediction system itself (Roscher et Al. 2020). This can be accomplished by using a combination of prediction and rule induction algorithms. These latter are able to derive sets of human readable rules describing the association between the independent variable and the dependent one.

This project aims at developing a knowledge base for an Artificial Intelligence (AI) system able to represent the rules associating multi-omics molecular patterns to phenotypes for a set of selected/available human tissues.
This will be carried out by accomplishing the following research objectives:

1. The collection, preprocessing and harmonization of multi-omics data for the chosen BSs.
2. The development of a suitable knowledge base obtained by the application of rule induction algorithms.
3. The development of a system able to query, combine and summarize the rules contained in the knowledge base.

The ideal candidate for this project is expected to have a good background in computer science and machine learning techniques along with an interest in molecular biology and the ability to work in an interactive environment and to collaborate with international partners.

Supervisor(s), Lab/Group details, other additional info.

Dr. Giovanni Scala, Department of Biology, University of Naples Federico II, Italy
Prof. Barbara Majello, Department of Biology, University of Naples Federico II, Italy

References

1. Scala G, Federico A, Fortino V, Greco D, Majello B. Knowledge Generation with Rule Induction in Cancer Omics. *Int J Mol Sci*. 2019;21(1):18. Published 2019 Dec 18. doi:10.3390/ijms21010018
2. Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics Data Integration, Interpretation, and Its Application. Bioinform Biol Insights. 2020; 14:1177932219899051. Published 2020 Jan 31. doi:10.1177/1177932219899051
3. R. Roscher, B. Bohn, M. F. Duarte and J. Garcke, "Explainable Machine Learning for Scientific Insights and Discoveries," in IEEE Access, vol. 8, pp. 42200-42216, 2020, doi: 10.1109/ACCESS.2020.2976199