Ph.D. in Information Technology
Thesis Defense

March 30th, 2022
at 9:30
Room Alpha and online by Teams

Michele LEONE – XXXIII Cycle

Genomic metadata integration and data processing methods for the analysis of chromatin behaviour in different biological conditions

Supervisor: Prof. **Marco Masseroli**

Abstract:

The information necessary for the development and proper functioning of most living organisms is contained in the DNA, usually associated with proteins. In the last decades, researchers have started cataloguing chromatin proteins and their modifications. This has led to the identification of several chromatin modifications or "marks" and the discovery of many regulatory elements throughout the genome. Many studies have been carried out with the aim of simplifying chromatin complexity by dividing it into a certain number of chromatin-states, to capture known classes of genomic elements. This thesis project aims to extend the concept of chromatin states, considering all the different types of genomic features, to create a framework that, starting from a set of functional elements, identifies the corresponding samples available in the most important web resources, integrates metadata information about tissue type and possible pathological conditions through the automatic extraction of controlled semantic terms from certain biomedical ontologies with innovative machine learning approaches, and finds combinations of these genomic features. Rather than treating the extraction of useful metadata from free-text descriptions as classification or named entity recognition, it was modelled as machine translation, leveraging state-of-the-art sequence-to-sequence models to directly map unstructured input into a structured text format. As consequence, an active learning framework to receive feedback from the users and improve the metadata prediction was designed, leading to the development of a technique to interpret the predictions of the model and apply this interpretation mechanism in a web interface to help the user give a correct feedback. Once chromatin states have been identified, a data-driven analysis is achieved through the clustering of regions and samples, the identification of genome clusters, and the gene-set enrichment analysis to associate these clusters to gene functional categories.

PhD Committee
Prof. **Rosario Piro,** Politecnico di Milano
Prof. **Mario Cannataro,** Universita' degli Studi Magna Graecia
Prof. **Rosalba Giugno,** Universita' di Verona