Ph.D. in Information Technology: Thesis Defense

February 4th, 2021

online by Zoom – at 10.30

Anna BERNASCONI – XXXII Cycle

Model, Integrate, Search... Repeat: a Sound Approach to Building Integrated Repositories of Genomic Data

Supervisor: Prof. **Stefano Ceri**
Co-Supervisor: Dr. **Alessandro Campi**

Abstract:

The integration of genomic data and of their describing metadata is, at the same time, an important, difficult, and well-recognized challenge. It is important because a wealth of public data repositories is available to drive biological and clinical research; combining information from various heterogeneous and widely dispersed sources is paramount to a number of biological discoveries. It is difficult because the domain is complex and there is no agreement among the various data formats, data models, and metadata definitions, which refer to different vocabularies and ontologies. It is well-recognized in the bioinformatics community because, in the common practice, repositories are accessed one-by-one, learning their specific metadata definitions as result of long and tedious efforts, and such practice is error-prone; moreover, downloaded datasets need considerable efforts prior to insertion in analysis pipelines.

Within the context of the European project data-driven Genomic Computing (GeCo), which supports genomic research by proposing bioinformatics abstractions and tools, this PhD thesis focuses on the data integration problem, sharing the motivations and methodologies of the project and addressing one of its objectives.

We propose a conceptual model of metadata and an extended architecture for integrating datasets, retrieved from a variety of genomic data sources, based upon a structured transformation process; we then provide a user-friendly search system providing access to the consolidated repository of metadata attributes, enriched by a multi-ontology knowledge base.

Inspired by our work on genomic data integration, during the outbreak of the COVID-19 pandemic we successfully re-applied the model-build-search paradigm used for human genomics, building on the analogies among the two domains of human and viral genomic.

The availability of conceptual models, related databases and search systems for both humans and viruses' genomics will provide important opportunities for genomic and clinical research, especially if virus data will be connected to its host, provider of genomic and phenotype information.

PhD Committee
Prof. **Marco Masseroli**, DEIB
Prof. **Carlo Batini**, Universita' di Milano Bicocca
Prof. **Oscar Pastor**, Universitat Politècnica de València