Ph.D. in Information Technology: Thesis Defense

January 28th, 2021

online by Teams – at 11.30

Luca NANNI – XXXIII Cycle

"Computational inference of DNA folding principles: from data management to machine learning"

Advisor: Prof. **Stefano Ceri**
Co-Advisor: **Prof. Colin Logie,** Radboud University Nijmegen

Abstract:

DNA is the molecular basis of life and consists of approximately three billion base pairs, which would total about three meters if linearly untangled. To fit in the cell nucleus at the micrometer scale, DNA has, therefore, to fold itself several times. Understanding the mechanisms of genome folding is a major biological research problem.

In this thesis, I first present the design and implementation of the PyGMQL software for interactive and scalable data exploration for genomic data. PyGMQL is based on the big data genomic engine GMQL and presents itself as a easy-to-use Python library, interacting seamlessly with other data analysis packages.

I then apply my software to the study of chromatin conformation data. I discover a set of spatial rules based on the insulator protein CTCF and its orientation which correlate with genome topology, highlighting the existence of a "grammar of genome folding".

I finally focus on the relationship between chromatin conformation and gene expression. I use graph representation learning to encode chromatin topological features of genes. The learnt gene embeddings are then used to predict if two genes are co-expressed or not, given independent gene expression data. The results indicate a correlation between chromatin topology and co-expression, shedding a new light on this debated topic and providing a novel computational framework for the study of co-expression networks.

PhD Committee

Prof. **Rosario Michael Piro**, DEIB
Prof. **Mario Cannataro**, Universita' di Catanzaro
Prof. **Vladimir B Teif,** University of Essex