Ph.D. in Information Technology
Thesis Defense

April 19th, 2024
at 14:30
Sala Seminari Nicola Schiavoni

Tommaso ALFONSI – XXXVI Cycle

Methods and Tools for Data Integration and Knowledge Discovery in Viral Genomics

Supervisor: Prof. Stefano Ceri

Abstract:

Viral genomics is a branch of science positioned at the intersection of biological research and computer science. It holds a pivotal role in understanding and predicting the behavior of viruses. Challenges in viral genomics encompass the development and application of computational methodologies to analyze vast datasets of viral genomes. The overarching goal is to unravel complex genomic architectures, delineate evolutionary dynamics, and employ computational techniques for predicting virulence factors and potential therapeutic targets. However, the field faces significant challenges, particularly due to the inherent variability of viruses. Each virus exhibits diversity in genome structures, sizes, and features. Additionally, viruses can evolve, generate thousands of variants, leading to rapid dissemination and potentially leap to new host species.

The exponential growth in data size since the introduction of Next Generation Sequencing (NGS) technology in 2008, followed by Nanopore Sequencing, has transformed the landscape of genomic data; in less than a decade, genomic data moved from scarce and costly to relatively cheap and vastly available.

Despite its abundance, viral genomic data is dispersed across numerous laboratories and organizations, often lacking a common agreement on formats and standards.

This thesis addresses these complexities by developing a data integration pipeline that resulted in the creation of ViruSurf, an integrated viral sequence and metadata repository. This repository provides a unified representation of viral data, independent of source or viral species. It also supports applications such as EpiSurf, an epitope repository and analysis tool, and VirusViz, a sequence analysis tool.

Right after the pandemic's start, numerous research studies focusing on COVID-19 were released. However, the scattered nature of the information, expressed in natural language and duplicated or conflincting at times, impeded the practical application of their results.

To support domain experts and health authorities in effectively using and organizing this knowledge, a knowledge model named CoV2K has been introduced. CoV2K integrates diverse information fields and is supported by a (knowledge) database, accessible through an API. What sets CoV2K apart is its ability to connect disparate knowledge domains, such as variants, epidemiological and clinical effects, facilitating the discovery of relationships among data entities. This model enhances the understanding of the complex nature of genomic data and supports the integration of data and knowledge across different scientific domains related to SARS-CoV-2.

Building upon these foundations, the application of CoV2K for automated knowledge discovery in the context of viral genomics is demonstrated. This approach aligns with artificial reasoning

methodologies, fostering increased collaboration between the artificial intelligence and biological sciences communities. The utilization of CoV2K enables the dynamic adaptation and evolution of the knowledge base, unlocking the potential for extracting meaningful insights from complex viral genomic data.
In the later stages of the pandemic, the emergence of recombinant variants in SARS-CoV-2 prompted a focused effort on the rapid detection of recombination events across sequences of single RNA viruses. The proposed method demonstrates superior diagnostic accuracy and detection speed compared to manual approaches and existing software methods, offering a valuable tool for public health responses to potential novel threats.
In conclusion, this thesis contributes significantly to the field of viral genomics. It addresses challenges related to data integration, knowledge modeling, automated reasoning, and the detection of viral recombinations. The applications and methodologies presented provide valuable insights for public health preparedness against evolving viral threats. Future work is hinted at to extend the novel recombination detection approach to other viruses, ensuring continuous advancements in viral surveillance and research.

PhD Committee

Prof. **Marco Masseroli, Politecnico di Milano**

Prof.ssa **Manuela Sironi, IRCCS**

Prof. **Paolo Missier, University of Newcastle**