Ph.D. in Information Technology
Thesis Defense

April 22nd, 2024
at 9:30
Room Alessandra Alario – building 21

Ruba AL KHALAF – XXXVI Cycle

Knowledge modeling and data analysis methods for understanding the viral genome evolution

Supervisor: Prof. Stefano Ceri

Abstract:

Embarking on an exploration of unparalleled depth into the intricacies of SARS-CoV-2 during the COVID-19 pandemic, this research represents a significant contribution to global health comprehension through the lens of two overarching fields of study: I) Viral data modeling and management and II) Data-driven large-scale viral sequences analysis. In Part I, the main focus is on knowledge modeling, emphasizing the knowledge extraction from scientific literature (based on deep learning) and its ontological representation. A careful study of the domain was performed to address three critical issues that arose during the pandemic and to suggest proper solutions for such issues: 1) Information quality concerns, tackled through the development of the CoV2K model, which extracts information from multiple resources, transforms it, and makes it accessible via a RESTful API. 2) The overwhelming volume of data in the scientific literature, handled through CoVEffect, which predicts the effects of mutations and variants in ~ 7,000 scientific publications' abstracts using a GPT-2 prediction model specifically trained to solve such a complex task. 3) The domain's complexity, addressed by the development of OntoEffect, an OntoUML-based ontology explaining the effects of SARS-CoV-2 variants, which offers clear and precise domain explanations provided via ontological unpacking methods. In Part II, the main focus is on understanding different aspects of viral evolution using data-driven large-scale genomic analysis. Specifically, I first employed robust statistical methods to identify different genetic phenomena, e.g., co-occurring and mutually exclusive pairs of mutations. Then, I extended the study to the evolutionary events within viral lineages, shedding light on the virus's adaptive mechanisms. Second, I studied the conservancy of specific small regions of viral proteins called epitopes, which can provoke the host immune system response. A mutation on an epitope range might affect its recognition, possibly empowering an immune escape variant. To investigate this aspect, I presented several use cases employing the EpiSurf tool, a metadata-driven search server for analyzing mutations on epitopes of viral species. Moreover, I performed a database-wide study on Omicron – the most complete "escapee" variant of SARS-CoV-2 – as a case study to reveal critical insights into its subvariants, their characteristic mutations, and their potential implications on immune evasion, enhancing our comprehension of this important variant. All in all, these research outcomes provide valuable resources and knowledge for the ongoing battle against COVID-19, paving the way to their extension for fighting future pandemics caused by the same or other pathogens.

Keywords: SARS-CoV-2; COVID-19, Knowledge modeling, Data analysis

PhD Committee

Prof. **Pietro Pinoli, Politecnico di Milano**

Prof. **Daniele Focosi, Università di Pisa**

Prof**. Giancarlo Guizzardi, University of Twente**