Ph.D. in Information Technology
Thesis Defense

July 10th, 2024
at 15:00
Sala Conferenze Emilio Gatti – building 20

Silvia CASCIANELLI – XXXVI Cycle

Machine Learning in Oncogenomics: a key to dissecting cancer inner heterogeneity

Supervisor: Prof. Marco Masseroli
Co-supervisor: Prof. Enzo Medico (Istituto di Candiolo – IRCCS)

Abstract:

Computational oncogenomics has a pivotal role in cancer research due to the inherent complexity and heterogeneity of cancer diseases that often pose an insurmountable barrier to traditional research approaches. Comprehensive omics-based investigations and advanced computational methods are demanded to understand molecular intricacies underlying tumors, answering unsolved biological and clinical issues and contributing to precision medicine. This PhD research belongs to the computational oncogenomics area and emphasises the synergistic use of omics data processing and Data Science techniques to tackle complex clinical challenges in cancer research. It was focused on developing original workflows to provide clinically relevant insights and stratifications for different types of cancer patients. Indeed, computational oncogenomics aims not only to decipher omics landscapes of cancer but also to provide clinicians with valuable indications of patient subgroups with distinct characteristics and demanding personalized therapeutic strategies. In such a context, Machine Learning-based solutions emerged as a powerful key to dissecting cancer inner heterogeneity and unlocking robust and clinically relevant patient-centric predictions.
All the methodologies and applications of this PhD project were developed to address hot topics in biomedical research, following the typical Data Science life cycle. Great attention was therefore devoted to omics data integration and exploration, as well as to the feature engineering and selection phase. Predictive modelling and result evaluation were tailored to face the peculiar issues of oncogenomics research scenarios and provide interpretation and validation both from computational and clinical-biological perspectives. To obtain robust results, it was indeed decisive to design, implement and carefully assess suitable and fully legit Machine Learning workflows, also taking particular care of clinically and biologically validating the achieved findings. To this aim, my research proceeded with strict collaborations with experts in Medicine and Biology, which offered me a clearer view of the needs and goals to meet for each task.

The endeavour of this PhD research was directed first to the enhancement of an R/Bioconductor package for efficient investigation and integration of omics data. Then, it delved into robust Machine Learning-based cancer subtyping for reliable patient predictions, also transitioning towards multi-label classification to represent even the inner heterogeneity of many patients. Lastly, it focused on mutation-based stratifications to identify variants with therapeutic or prognostic roles in patient groups of critical clinical handling. Applications to breast and colorectal cancer proved the computational efficacy and clinical/biological relevance of the obtained results. This work achieved remarkable advances in cancer subtyping, including designing a new feature selection method to tackle unbalanced classification scenarios; investigating multi-omics, deep and semi-supervised solutions in light of the increasing omics data availability; introducing multilabel subtyping strategies, which reflect underlying cellular heterogeneity, enhance patient molecular characterization and improve the clinical value of the predictions considering both primary and secondary assignments. In addition, the integration of class discovery, transfer learning, and multi-label predictions demonstrated its efficacy in finding a more exhaustive colorectal cancer stratification, named EXA-CRIS. Its classes were also traced on a different dataset type, leveraging an adaptation strategy designed to optimize the choice of the most suitable features for the predictive task. Lastly, innovative mutation-based feature engineering and supervised frameworks, used to recognize critical subgroups of cancer patients, were combined with variant prioritisation approaches and search for actionable genes to find new potential therapeutic targets.
Thus, this research has placed its priority on implementing and meticulously assessing comprehensive Omics Data Science workflows to dissect cancer inner heterogeneity from different omics perspectives, all contributing to shaping new trajectories towards precision medicine.

PhD Committee

Prof. Pietro Pinoli, Politecnico di Milano

Prof.ssa Rosalba Giugno, Università di Verona

Prof. Giulio Caravagna, Università di Trieste