SCIENCES & BIOENGINEERING SCIENCES

The Research Group

Artificial Intelligence lab

has the honor to invite you to the public defense of the PhD thesis of

Sofia Papadimitriou

to obtain the degree of Doctor of Sciences

Title of the PhD thesis:

Towards multivariant pathogenicity predictions: using machine-learning to directly predict and explore disease-causing oligogenic variant combinations

Promotor:

Prof. dr. Ann Nowé

Co-promotor:

Prof. dr. Tom Lenaerts

The defense will take place on

September 15, 2020 at 17h

in Forum A, ULB La Pleine Campus

If you would like to attend in person, please contact sofia.papadimitriou@vub.be . An online link for live-streaming will be soon added here: https://bit.ly/2Z4z1Zs

Members of the jury

Prof. dr. Coen De Roover (VUB, chair)
Prof. dr. Sonia Van Dooren (VUB, secretary)
Prof. dr. Gianluca Bontempi (ULB)
Prof. dr. Guillaume Smits (ULB)
Prof. dr. Elfride de Baere (UGent)
Prof. dr. Christian Gilissen (Radboud University, NL)

Curriculum vitae

2008 - 2013: BSc in Biology, *Aristotle University of Thessaloniki, Greece*

2014-2016: MSc in Bioinformatics, *Wageningen University, the Netherlands*

2016 - Present: PhD candidate, *Vrije Universiteit Brussel (Artificial Intelligence lab) and Université Libre de Bruxelles (Machine Learning Group)*

Abstract of the PhD research

The emergence of statistical and predictive methods able to analyse genomic data has revolutionised the field of medical genetics, allowing the identification of disease-causing gene variants (i.e. mutations) for several human genetic diseases. Although these approaches have greatly improved our understanding of Mendelian «one gene - one phenotype» genetic models, studying diseases related to more intricate models that involve causative variants in several genes (i.e. oligogenic diseases) still remains a challenge, either due to the lack of sufficient methodologies and disease-specific cohorts to study or the phenotypic complexity associated with such diseases. This situation makes it difficult to not only understand the genetic mechanisms of the disease, but to also offer proper counseling and support to the patient. Until recently, no specialized predictive methods existed to directly predict causative variant combinations for oligogenic diseases. However, with the advent of data on variant combinations in gene pairs (i.e. bilocus variant combinations) leading to disease, collected at the Digenic Diseases Database (DIDA), we hypothesized that the transition from single to variant combination pathogenicity predictors is now possible.

To investigate this hypothesis, we organised our research on two main routes. At first, we developed an interpretable variant combination pathogenicity predictor, called VarCoPP, for gene pairs. For this goal, we trained multiple Random Forest algorithms on pathogenic bilocus variant combinations from DIDA against neutral data from the 1000 Genomes Project and investigated the contribution of the incorporated variant, gene and gene pair features to the prediction outcome. In the second part, we explored the usefulness of different gene pair burden scores based on this novel predictive method, in discovering oligogenic signatures in neurodevelopmental diseases, which involve a spectrum of monogenic to polygenic cases. We performed a preliminary analysis on the Deciphering Developmental Diseases (DDD) project containing exome data of 4195 families and assessed the capability of our scores in supporting already diagnosed monogenic cases, discovering significant pairs compared to control cases and linking patients in communities based on the genetic burden they share, using the Leiden community detection algorithm.

The performance of VarCoPP shows that it is possible to predict disease-causing bilocus variant combinations with good accuracy both during cross-validation and when testing on new cases. We also show its relevance for disease-related gene panels, and enhanced its clinical applicability by defining confidence zones that guarantee with 95\% or 99\% probability that a prediction is indeed a true positive, guiding clinical researchers towards the most relevant results. This method and additional biological annotations are incorporated in an online platform called ORVAL that allows the prediction and exploration of candidate disease-causing oligogenic variant combinations with predicted gene networks, based on patient variant data. Our preliminary analysis on the DDD cohort shows that - although all bi-locus burden scores show advantages, disadvantages and certain types of biases - taking the maximum pathogenicity score present inside a gene pair seems to provide, at the moment, the most unbiased results. We also show that our predictive methods enable us to detect patient communities inside DDD, based exclusively on the shared pathogenic bi-locus burden between patients, with more than half of these communities containing enriched phenotypic and molecular pathway information. Our predictive method is also able to bring to the surface genes not officially known to be involved in disease, but nevertheless, with a biological relevance, as well as a few examples of potential oligogenicity inside the network, paving the way for further exploration of oligogenic signatures for neurodevelopmental diseases.