Prediction of Genes associated With Autism Spectrum Disorder Using Graph Embedding methods

Place of work: Instituto Nacional de Saúde Doutor RIcardo Jorge

Supervisors: Hugo Martiniano (hugo.martiniano@insa.min-saude.pt); Astrid Vicente (astrid.vicente@insa.min-saude.pt);

In recent years machine learning methods designed to work with graph data have emerged. Many biological datasets are naturally structured in this way, such as protein-protein interaction (PPI) networks or Biological Pathways so the application of these algorithms to data in this domain is often straightforward.

In this project we propose the application of these methods to networks of interactions between genes, based on protein-protein interactions, as well as to other types of biological networks, obtained from publicly-accessible databases, such as gene-gene similarity networks derived from the Gene Ontology.

The ultimate aim is the identification of risk genes associated with Autism Spectrum DIsorder (ASD), a prototypical complex disorder with a strong genetic component. More specific aims are the comparison of the performance of several graph embedding methods and different data sources on the task of predicting the association of genes to ASD.

The main task are:

1 - Apply in-house Machine learning pipelines to perform binary classification of genes
2 - Extend and adapt existing pipelines for new networks

Knowledge of Python and common Python Machine Learning and processing frameworks (scikit-learn, pandas), Bash and the linux command line utilities would be useful.