PASTEUR PARIS UNIVERSITE (PPU) INTERNATIONAL DOCTORAL PROGRAM 2020

@IMAGINE INSTITUTE

PROJECT

- **FILE #11**
- **ACRONYM:** AI-GraphDiag
- **TITLE:** Graph representation learning on clinical & multi-omics data for improved rare-diseases diagnosis and patient-stratification

LABORATORY

- **SURNAME, FIRST NAME:** RAUSELL, Antonio
- **IP DEPARTEMENT:** IHU Imagine, INSERM UMR 1163
- **DOCTORAL SCHOOL:** Doctoral School Bio Sorbonne Paris Cité (BioSPC)
- **UNIVERSITY:** Université de Paris
- **FUNCTION:** Group Leader
- **TEL:** +33(0)142754575
- **E-MAIL:** antonio.rausell@institutimagine.org

LABORATORY PRESENTATION AND RESEARCH TOPICS

- **SUPERVISOR HDR:** RAUSELL, Antonio, antonio.rausell@institutimagine.org

- **SPECIFY THE TEAM NAME:** Clinical Bioinformatics Lab, Imagine Institute

- **WEBSITE OF THE TEAM:** https://www.institutimagine.org/en/antonio-rausell-161

DESCRIPTION OF THE PROPOSED PROJECT

- **KEYWORDS:** Artificial Intelligence, Multi-omics, Rare diseases, Genetic diagnosis, Patient stratification, Graph-representation learning

- **ABSTRACT**

This PhD project aims at developing novel computational methods for the improvement of (i) current diagnostic rates of pediatric rare diseases and (ii) patient stratification for tailored clinical follow up. To that aim, we propose the implementation of an Artificial Intelligence-based computational framework mining heterogeneous large-scale data including clinical data and genetic and multi-omics profiling of patient's samples. Graph-representation learning approaches capable to handle multi-modal and multi-layered networks will be developed. Here we aim at performing (i) node classification, (ii) subgraph classification and (iii) edge prediction, within and across layers. The previous approaches will be applied to collaborative studies currently ongoing at the Imagine Institute on specific rare disease cohorts. The methods and software developed through the PhD project will be integrated in comprehensive bioinformatics workflows applicable in Precision Medicine at the Imagine Institute and Necker-Enfants Malades Hospital.

- **DESCRIPTION OF THE PROJECT**

- **Context**

The Clinical Bioinformatics Laboratory of the Imagine Institute, under the direction of Dr. Antonio Rausell, develops statistical methods, machine-learning algorithms and bioinformatics pipelines with a clinical focus. A main interest of the laboratory is the computational assessment of human rare genetic variants with a potential clinical impact and its application to the study of specific rare diseases. Emphasis is dedicated to non-coding variants with a potential regulatory role. Ongoing research integrates heterogeneous large-scale data including clinical records, cell phenotyping, high-dimensional genome/exome sequencing and transcriptomes (bulk and single-cell RNA-seq analysis) from both in-house biobanks and public sources. Main goals are identifying cell markers with a value for diagnosis, prognosis and treatment, and providing software to help decision-making at the clinics.

- **Aims**

This PhD project aims at developing novel computational methods for the improvement of (i) current diagnostic rates of pediatric rare diseases and (ii) patient stratification for tailored clinical follow up. To that aim, we propose the implementation of an Artificial Intelligence-based computational framework mining heterogeneous large-scale data including clinical data and genetic and multi-omics profiling of patient's samples. Clinical data typically involve electronic health records (in text format), diverse clinical tests and imaging. Genetic and multi-omics profiling may include exome and genome sequencing as well as transcriptomics, proteomics, epigenomics and metabolomics. Moreover, recent developments allow molecular profiling of samples at single-cell level for one or several of the previous multi-omics approaches. Such *big data* covers a comprehensive set of biological complexity scales spanning the genetic, epigenetic, molecular, cellular, tissue, organism and physiological aspects of an individual. Complex biological systems are often modelled through graph-structured data where the different entities and their relations are represented, respectively, as nodes and edges in a network. However, heterogeneous and longitudinal data types require multi-modal and multi-layered networks simultaneously representing multiple entities (e.g. genes, proteins, cells, organs, individuals, families, clinical signs, diseases, drugs, etc.) and multiple relationships (e.g. gene regulation, co-expression, molecular interactions, extracellular signaling, phenotypic similarities, clinical syndromes, comorbidities, clinical paths, etc.).

Recently, we have implemented *Tiresias*, a comprehensive computational framework for the supervised learning of node classes on multiplex networks: https://github.com/RausellLab/Tiresias (Dritsa et al, in preparation). State-of-the-art approaches covered in *Tiresias* include Random Walks-based diffusion algorithms, graph embeddings and graph neural networks including Convolutional Neural Networks (GCN). In this PhD project we propose the implementation of novel algorithms that will allow us to expand such graph-representation learning framework to mine simultaneously multi-modal and multi-layered networks. Here we aim at performing (i) node classification, (ii) subgraph classification (both for within and across layer subgraphs), and (iii) edge prediction (again within and across layers). Both supervised and unsupervised learning will be addressed. Algorithms will be adapted to cope with the specific challenges of the different data types, accounting for ascertainment and technological biases, biological and technical noise, missing data as well as population structure and environmental and life-style confounding factors. Special attention will be paid to computational efficiency, to avoid overfitting in the training phase and to reach generalizable models through independent testing across different cohorts and medical sites.

The previous approaches will be applied to collaborative studies currently ongoing at the Imagine Institute on specific rare disease cohorts (e.g. ciliopathies, autoinflammatory diseases, developmental disorders) where WGS is being combined with multi-omics approaches and large-scale clinical data monitoring as described above. The methods and software developed through the PhD project will be integrated in comprehensive bioinformatics workflows applicable in Precision Medicine at the Imagine Institute and *Necker-Enfants Malades* Hospital. The project will require the development of research skills navigating across disciplines including bioinformatics, medical informatics, machine learning and human genomics. Research will be actually performed in a such a highly interdisciplinary environment and in close interaction with clinical and experimental teams.

- **REFERENCES**

Google Scholar: https://scholar.google.com/citations?user=W7YCTs4AAAAJ&hl=en

Cortal A, Matignetti L, Six E, Rausell A. Cell-ID: gene signature extraction and cell identity recognition at individual cell level. BioRxiv 2020 doi: https://doi.org/10.1101/2020.07.23.215525

Rausell A*, Luo Y, Lopez M, Seeleuthner Y, Rapaport F, Favier A, Stenson PD, Cooper DN, Patin E, Casanova JL*, Quintana-Murci LL, Abel L*. Common homozygosity for predicted loss-of-function variants reveals both redundant and advantageous effects of dispensable human genes. Proc Natl Acad Sci USA (2020) 117 (24) 13626-13636; **Corresponding authors*

Caron B, Luo Y, Rausell A. NCBoost classifies pathogenic non-coding variants in Mendelian diseases through supervised learning on purifying selection signals in humans. Genome Biology (2019) 20:32

Rato S*, Rausell A*, Munoz M, Telenti A, Ciuffi A. Single-cell analysis identifies cellular markers of the HIV permissive cell. PLOS Pathogens (2017), 13(10):e1006678. *Cofirst authorship

Fischer A, Rausell A. Primary immunodeficiencies suggest redundancy within the human immune system. Science immunology (2016) Vol. 1, Issue 6, doi: 10.1126/sciimmunol.aah5861

Juliá M, Telenti A, Rausell A*. Sincell: an R/Bioconductor package for statistical assessment of cell-state hierarchies from single-cell RNA-seq. Bioinformatics (2015), doi: 10.1093/bioinformatics/btv368. *Corresponding author.

Bartha I*, Rausell A*, McLaren P, Tardaguila M, Mohammadi P, Fellay J, Telenti A. The Characteristics of Heterozygous Protein Truncating Variants in the Human Genome. Plos Computational Biology (2015), 11(12):e1004647. *Cofirst authorship.

Rausell A, Mohammadi P, McLaren PJ, Bartha I, Xenarios I, Fellay J, Telenti A. Analysis of stop-gain and frameshift variants in human innate immunity genes. Plos Computational Biology (2014), 10 (7), e1003757.

EXPECTED PROFILE OF THE CANDIDATE

- **EXPERIENCE REQUIRED**
    - Candidates should hold a Master degree in Bioinformatics, Computational Biology, Computer Science, Biostatistics, Statistical Genetics, Applied Mathematics or related fields.
    - Programming skills in R, Python, and UNIX Bash
    - Former experience (e.g. Master internships) in complex systems modelling and strong data analyses skills will be a plus.
    - Ability to work independently
    - Languages: English