AAP Contrats doctoraux en Intelligence artificielle

Cofinancé par l'ANR

PhD open position: Machine learning to predict new gene regulatory variants involved in immunological diseases

Contact: **Aitor Gonzalez (aitor.gonzalez@univ-amu.fr)**

1. DESCRIPTION OF THE PHD THESIS PROJECT

1.1 OBJECTIVES OF THE PROJECT BASED ON THE CURRENT STATE OF THE ART

Complex diseases are influenced by genetic variants. In particular, it is expected that cancer is associated with variants altering on the one hand the proliferation of malignant cells and their ability to invade the host, and on the other hand the immunosurveillance mechanisms [1-3]. Genome-wide association study (GWAS) is a key approach to link genetic factors to human phenotypes. Many SNPs identified by GWAS fall in non-coding regions and are likely gene regulatory variants. The TAGC laboratory is strongly interested in multifactorial diseases such as cancer and diseases related to the immune system [1; 4-8]. The TAGC laboratory hosts a sequencing platform and researchers that generate new genomics data related to gene regulatory regions. Furthermore, there are bioinformaticiens that analyze these data and develop new bioinformatics tools [9-13].

Machine learning is a very powerful technique to integrate multivariate data and make new predictions. Supervised machine learning models take particular sets of regulatory variants or regions and molecular data such eQTLs, transcription factor binding sites and motifs to predict functional regions [14]. In the case of regulatory variants, most supervised models are trained with rare variants showing large effects and do not focus on particular diseases [15; 16]. However several international initiatives exist to associate common genetic and phenotypic variation such as the INSERM Genomic Variability 2018, where the TAGC lab participates (https://bit.ly/2TIfSeo). Recently we have developed a new method to train a supervised model with common regulatory variants associated to complex diseases that we have run on intergenic and intronic regions. Interestingly, our method has a good prediction performance for diseases related to the immune system [11].

In the present project, we intend to develop an approach to specialize the model for a specific disease. More specifically, we plan to use frequent variants associated to diseases related to the immune system to train the model. This model will be then used to prioritize regulatory variants arising in acute myeloid (AML) and T-cell acute lymphoblastic leukemia (ALL), which are studied in the TAGC laboratory. Both leukemias have strong

genetic predisposition [17; 18]. We hypothesize that such models would predict causal variants involved in tumorigenesis or anticancer immunosurveillance.

1.2 METHODOLOGY

Regulatory variants involved in immune system diseases will be selected in non-coding regions from datasets such as the GWAS catalog. These regulatory variants will be prioritized using different published tools as benchmark for new predictive models. The SNPs will be annotated with molecular data relevant to gene regulation such as active chromatin markers, transcription factor binding sites and transcriptions motifs. Analysis of these annotations will be carried out to define the most predicitive annotations. Selected annotations will be used to create predictive models. These models will be used to score every position of the human genome and prioritize SNPs where the immune system is involved. The model will be also used to look for pathogenic mutations in regulatory regions of cancer samples from in-house data or public data repositories such as the TCGA database.

1.3 WORK PLAN

Objective	Beginning	End
Collect regulatory variants and annotations	11/20	08/21
Priorisation tool benchmark for regulatory variants	03/21	12/21
Data analysis of regulatory variant annotations	07/21	04/22
Training and testing of the model	11/21	08/22
Analysis of genomics datasets of leukemias and other tumors	03/22	12/22
Web portal and/or software package to share predictions	07/22	04/23
Write and submit PhD thesis and paper	11/22	08/23

1.4 SUPERVISOR AND RESEARCH GROUP DESCRIPTION

The TAGC has developed an strong interest and expertise in high-throughput and bioinformatics methods [9; 10; 12; 13]. The TAGC laboratory is interested in T-cell Acute Lymphoblastic Leukaemia (T-ALL) and and acute myeloid leukemia (AML) [7; 8]. The TAGC laboratory hosts a sequencing platform and researchers that generate new genomics data related to gene regulatory regions. Furthermore, there are bioinformaticiens that analyze these data and develop new bioinformatics tools.

The TAGC laboratory is using machine learning to integrate data related to gene regulatory regions and predict new gene regulatory regions [11; 14]. To develop these models, the TAGC laboratory is collaborating with other mathematics and computer science laboratories such as the Institute de Mathématiques de Marseille (I2M) and the Laboratoire d'Informatique et Systèmes (LIS) [10].

2. RECENT PUBLICATIONS

Chèneby, J.; Ménétrier, Z.; Mestdagh, M.; Rosnet, T.; Douida, A.; Rhalloussi, W.; Bergon, A.; Lopez, F. and Ballester, B. (**2020**). *ReMap 2020: a database of regulatory regions from an integrative analysis of Human and Arabidopsis DNA-binding sequencing experiments.*, Nucleic acids research 48 : D180-D188.

González, A.; Artufel, M. and Rihet, P. (**2019**). *TAGOOS: genome-wide supervised learning of non-coding loci associated to complex phenotypes.*, Nucleic acids research 47 : e79.

Nguyen, N. T. T.; Contreras-Moreira, B.; Castro-Mondragon, J. A.; Santana-Garcia, W.; Ossio, R.; Robles-Espinoza, C. D.; Bahin, M.; Collombet, S.; Vincens, P.; Thieffry, D.; van Helden, J.; Medina-Rivera, A. and Thomas-Chollier, M. (**2018**). *RSAT 2018: regulatory sequence analysis tools 20th anniversary.*, Nucleic acids research 46 : W209-W214.

Dao, L. T. M.; Galindo-Albarrán, A. O.; Castro-Mondragon, J. A.; Andrieu-Soler, C.; Medina-Rivera, A.; Souaid, C.; Charbonnier, G.; Griffon, A.; Vanhille, L.; Stephen, T.; Alomairi, J.; Martin, D.; Torres, M.; Fernandez, N.; Soler, E.; van Helden, J.; Puthier, D. and Spicuglia, S. (**2017**). *Genome-wide characterization of mammalian promoters with distal enhancer functions.*, Nature genetics 49 : 1073-1081.

Seyres, D.; Darbo, E.; Perrin, L.; Herrmann, C. and González, A. (**2016**). *LedPred: an R/bioconductor package to predict regulatory sequences using support vector machines.*, Bioinformatics (Oxford, England) 32 : 1091-1093.

3. EXPECTED PROFILE OF THE CANDIDATE

The candidate should have a Master's degree in an area related to Bioinformatics, Biophysics, Computer Science or Mathematics with a good background in statistics, data analysis and machine learning . He should be interested in a project that includes machine learning and human genetics. He should feel confortable with software programming with a preference for Python and data analysis. He should also have some knowledge of molecular biology and/or genetics. The expected start of this PhD position is between September and December 2020 and the funding runs for three years. In addition, Bioinformatics teaching in French to life science students might by possible if desired by the PhD student.

4. SUPERVISORS' PROFILE

This PhD project will be co-supervised by Aitor González, Badih Ghattas and Pascal Rihet. Aitor González (TAGC) is a bionformaticien that uses machine learning to analyze the non-coding regions and variants of the genome [11; 14]. Pascal Rihet (TAGC) uses quantitative genetics methods and experimental validation to find genetic markers of complex diseases such as malaria and autoimmunological diseases with an emphasis in gene regulatory regions [19-21]. Badih Ghattas (I2M) is a mathematicien with expertise in statistical modeling and prediction using machine and deep learning with a large previous experience of collaboration with biologists [10; 22].

VISA DU RESPONSABLE DE L'INSTITUT ET DU DIRECTEUR DE LABORATOIRE CONCERNÉS

Visa du responsable de l'institut,

NOM Prénom

Fait à Marseille, le

Signature

Visa du directeur du laboratoire,

NOM Prénom

Fait à Marseille,

Signature

Bibliography

[1] **Farnault et al.** (2012). Hematological malignancies escape from NK cell innate immune surveillance: mechanisms and therapeutic implications. *Clinical & developmental immunology* **2012**, 421702.

[2] **Zitvogel et al.** (2016). Mouse models in oncoimmunology *Nature Reviews Cancer* **16**, 759-773.

[3] **Fridman et al.** (2017). The immune contexture in cancer prognosis and treatment. *Nature reviews. Clinical oncology* **14**, 717-734.

[4] **Fauriat et al.** (2007). Deficient expression of NCR in NK cells from acute myeloid leukemia: Evolution during leukemia treatment and impact of leukemia cells in NCRdull phenotype induction. *Blood* **109**, 323-330.

[5] **Delahaye et al.** (2011). Alternatively spliced NKp30 isoforms affect the prognosis of gastrointestinal stromal tumors. *Nature medicine* **17**, 700-707.

[6] **Renou et al.** (2017). Homeobox protein TLX3 activates miR-125b expression to promote T-cell acute lymphoblastic leukemia. *Blood advances* **1**, 733-747.

[7] **Kermezli et al.** (2019). A comprehensive catalog of LncRNAs expressed in T-cell acute lymphoblastic leukemia. *Leukemia & lymphoma* **60**, 2002-2014.

[8] **Perez-Alea et al.** (2019). Dual Targeting of ALDH1 and ALDH3: A Promising Therapeutic Approach in Acute Myeloid Leukemia *Blood* **134**, 3364-3364.

[9] **Nguyen et al.** (2018). RSAT 2018: regulatory sequence analysis tools 20th anniversary. *Nucleic acids research* **46**, W209-W214.

[10] **Ferré et al.** (2019). OLOGRAM: Determining significance of total overlap length between genomic regions sets. *Bioinformatics (Oxford, England)*.

[11] **González et al.** (2019). TAGOOS: genome-wide supervised learning of non-coding loci associated to complex phenotypes. *Nucleic acids research* **47**, e79.

[12] **Lopez et al.** (2019). Explore, edit and leverage genomic annotations using Python GTF toolkit. *Bioinformatics (Oxford, England)* **35**, 3487-3488.

[13] **Chèneby et al.** (2020). ReMap 2020: a database of regulatory regions from an integrative analysis of Human and Arabidopsis DNA-binding sequencing experiments. *Nucleic acids research* **48**, D180-D188.

[14] **Seyres et al.** (2016). LedPred: an R/bioconductor package to predict regulatory sequences using support vector machines. *Bioinformatics (Oxford, England)* **32**, 1091-1093.

[15] **Amlie-Wolf et al.** (2018). INFERNO: inferring the molecular mechanisms of noncoding genetic variants. *Nucleic acids research* **46**, 8740-8753.

[16] **Wang et al.** (2018). IW-Scoring: an Integrative Weighted Scoring framework for annotating and prioritizing genetic variations in the noncoding genome *Nucleic Acids Research* **46**, e47-e47.

[17] **Papaemmanuil et al.** (2009). Loci on 7p12.2, 10q21.2 and 14q11.2 are associated with risk of childhood acute lymphoblastic leukemia *Nature Genetics* **41**, 1006-1010.

[18] **Charrot et al.** (2019). AML through the prism of molecular genetics *British Journal of Haematology* **188**, 49-62.

[19] **Baaklini et al.** (2017). Beyond genome-wide scan: Association of a cis-regulatory NCR3 variant with mild malaria in a population living in the Republic of Congo *PLOS ONE* **12**, e0187818.

[20] **Labiad et al.** (2018). A transcriptomic signature predicting septic outcome in patients undergoing autologous stem cell transplantation. *Experimental hematology* **65**, 49-56.

[21] **Thiam et al.** (2018). NCR3 polymorphism, haematological parameters, and severe malaria in Senegalese patients *PeerJ* **6**, e6048.

[22] **Ghattas et al.** (2019). Assessing variable importance in clustering: a new method based on unsupervised binary decision trees *Computational Statistics* **34**, 301-321.