Machine learning to predict new gene regulatory variants involved in immunological diseases

A 3-year PhD position is available from September-December 2020 at the Inserm U1090/TAGC and the I2M, UMR 7373 unit in Marseilles, France funded by the ANR and AMU's Institute for Cancer and Immunology. The topic of this project is the use of "Machine learning to predict new gene regulatory variants involved in immunological diseases". The successful candidate will work under the joint supervision of Dr Aitor Gonzalez (TAGC), Dr Badih Ghattas (I2M) and Prof Pascal Rihet (TAGC).

How to apply

Interested candidates should send these documents to Aitor Gonzalez (aitor.gonzalez@univ-amu.fr) **before June 26th 2020.**

- CV

- Cover letter summarizing previous experience and motivations for this project, future plans, ...

- Two recommendation letters or contacts for these letters.

- Grades and transcripts after high school

Description of the PhD thesis project

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; 2] . 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.
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 [3].
In the present project, we expect the PhD student 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 different tumors and immune system diseases such as acute myeloid (AML) and T-cell acute lymphoblastic leukemia (ALL). It is hypothesized hypothesize that such models would predict causal variants involved in tumorigenesis or anticancer immunosurveillance.

Supervisor and research group description

The TAGC has developed an strong interest and expertise in high-throughput and bioinformatics methods . The TAGC laboratory is interested in in T-ALL and and AML. This PhD project will be jointly supervised by Aitor Gonzalez, Badih Ghattas, Pascal Rihet. Aitor González (TAGC) is a bionformaticien that uses machine learning to analyze the non-coding regions and variants of the genome [3; 4]. 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 [5; 6]. 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 [7; 8].

Expected profile of the candidate

The candidate should have a Master's degree in an area related to Bioinformatics, Biophysics, Population Genetics or Computer Science 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, teaching in Bioinformatics to Bachelor and Master students might be also possible.

Bibliography

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

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

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

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

[5] **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.

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

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

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