Internship offer:

Development and application of gene regulatory network inference based on ensemble learning.

Host Laboratory
UMR INRA/INSA de Lyon 203 BF2I (Biologie Fonctionnelle, Insectes et Interactions)
INSA Bâtiment Louis Pasteur
20 avenue Albert Einstein
69621 Villeurbanne cedex
http://bf2i.insa-lyon.fr/

Internship supervisors
Sergio Peignier
Mail : sergio.peignier@insa-lyon.fr

Nicolas Parisot
Mail : nicolas.parisot@insa-lyon.fr

Duration: 3-4 months

Applications: Contact the supervisors by email (CV attached to the email)

Application deadline: 31 December 2018

Scientific context

Gene regulatory networks describe the complex interactions between specialized genes such as transcription factors and other genes. Such interactions are to a large extent, responsible for the regulation of gene expression, and their adaptation to different conditions. Understanding such networks, and analysing their organization and their dynamics, are important steps towards the comprehension of complex mechanisms that shape living organisms.

Inferring gene regulatory networks from high-throughput data (e.g., RNAseq, Microarrays) is a complex problem, and differents methods tackling this task have been proposed so far in the literature. These techniques can be classified in five major families, namely *Correlation* based methods, *Mutual Information* based techniques, *regression* based methods, *bayesian* inference algorithms, and *model based* techniques (e.g. based on ODEs). Each family presents its own advantages and biases, and each method may be more prone to detect particular kinds of regulatory interactions from the

data. This phenomenon has been characterized by Marbach et al., 2012 [1]. Interestingly, in this same study, the authors have also shown that combining different methods leads to better and more robust results. This approach, known in the machine learning field as *ensemble learning* [2], has also been used successfully to deal with complex real world problems, from different domains.

The ensemble-based approach proposed by Marbach et al., 2012 [1], has used algorithms specifically designed for gene regulatory network inference. Nevertheless, there exist in practice a large number of generalist machine learning algorithms (e.g., regression methods, feature selection methods) that can be used to infer gene regulatory networks. The goal of this internship is to contribute to the development, the assessment and the application of a new machine learning ensemble approach, for gene regulatory network inference.

Internship project

During this internship the student will get familiarized with the literature on gene regulatory network inference. The student will also learn how to apply general machine learning algorithm (e.g., regression methods, feature selection methods) to infer regulatory frameworks.

The student will contribute to the development of a Machine Learning ensemble approach, for gene regulatory network inference. The software will be developed in Python, using implementations of machine learning algorithms from the scikit-learn data mining library, and these algorithms will be combined in an ensemble learning framework using the pystacknet library.

Moreover, the student will assess this methodology, and compare it with existing methods (e.g. PyPanda [3], Arboretum [4]), on benchmark datasets, made available by the DREAM project, the Modencode and the Encode project. In this context the student could also contribute to the construction of new benchmarks from RNAseq and CHIPseq datasets.

Finally the student will be able to apply this methodology to infer the gene regulatory network of the pea aphid organism, from RNAseq and Microarray datasets. Then the student will use the NetworkX library to apply graph theory algorithms to characterize the network, to study its topology, and analyse possible functional roles of network modules.

Context

This internship would take place at the BF2I lab, within the SymT team. The aim of the lab is to study the complex biological interactions between insects and their symbiotic bacteria. The SymT team focuses on the trophic interactions between the pea aphid organism and its symbiotic bacteria. The student will be under the supervision of Sergio Peignier (for the Machine Learning and graph theory side) and Nicolas Parisot (for topics mostly related to gene expression data, transcriptomics, RNAseq, CHIPseq, ...).

References

[1] Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., Allison, K. R., DREAM5 Consortium, Kellis, M., Collins, J. J., … Stolovitzky, G. (2012). Wisdom of crowds for robust gene network inference. *Nature methods*.

[2] Zhang, C., & Ma, Y. (Eds.). (2012). *Ensemble machine learning: methods and applications*. Springer Science & Business Media.

[3] Van IJzendoorn, D. G., Glass, K., Quackenbush, J., & Kuijjer, M. L. (2016). PyPanda: a Python package for gene regulatory network reconstruction. *Bioinformatics*.

[4] Thomas Moerman, Sara Aibar, Carmen Bravo González-Blas, Jaak Simm, Yves Moreau, Jan Aerts, Stein Aerts; GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks, *Bioinformatics*