Internship + PhD thesis proposal

Updates at `http://www.laurent-duval.eu/lcd-2018-intern-phd-epigenetics-omics-graph-processing.html`

1 Description: Bioinformatics graph processing for multi-omics networks

1.1 Context

Micro-organisms used for bio-based chemistry are studied with the aim of conceiving potential microbial factories. They are piloted by their genome expression, with very diverse mechanisms acting at various biological scales, sensitive to external conditions (environment, temperature, nutrients). The eruption of novel high-throughput experimental technologies has demultiplied the available *omics* data and means of understanding for the studied systems. Their handling however increasingly requires advanced bioinformatic tools based on optimization, graphs and machine learning strategies for instance, as well as theoretical frameworks to provide insights into the multi-level causation [NDT15].

The global objective of the whole project is to develop innovative analysis methods for such highly integrated data modeled as networks. They should include genomic, transcriptomic and epigenetic data, for multiple microorganisms strains. The methodology would inherit from a wealth of techniques developed over graphs for scattered data, social networks, etc. and built upon biologically-related *a priori* as recently developed at IFP Energies nouvelles [PCB$^+$15, PCDP17].

1.2 PhD subject

The main objective of the PhD is to decipher the panoply of gene regulation levels to improve the understanding regarding specific protein and enzyme production. From model organisms to *Trichoderma reseei*, such a panoply will arise through data integration encoding different biological mechanisms and strains, combining genome, transcriptome and epigenome. The chosen path is that of graph models combined with network optimization, allowing the integration of different natures of data. The deployment of clustering and source separation methods (in the line of BRANE Cut and BRANE Clust algorithms) is targeted to the identification of *omics* data features driving enzyme production. Attention will be paid to novel evaluation metrics, as their standardization remains a crucial stake in bioinformatics.

1.3 Internship subject

As a preparation to the above subject, the trainee will focus on the statistical integration [AC14] of different transcriptomic data for multiple microorganisms strains sharing the same genealogy. Feature learning for multilayer networks[1] and the exploration of the resulting higher-order networks [BGL16] will evaluated in the context on this study.

2 Supervision, required skills, duration, location

- CentraleSupélec: Fragkiskos D. Malliaros, Jean-Christophe Pesquet
- IFP Energies nouvelles (IFPEN): Aurélie Pirayre, Frédérique Bidard-Michelot, Laurent Duval
- Skills: applied mathematics, optimization, graph analysis, machine learning, data science, bioinformatics, minimal biological knowledge ; languages: R, Matlab, Python
- Duration (internship): 4 to 6 months, Paris-Saclay/Rueil-Malmaison

References

[AC14] C. Angelini and V. Costa. Understanding gene regulatory mechanisms by integrating ChIP-seq and RNA-seq data: statistical solutions to biological problems. *Front. Cell Dev. Biol.*, 2, Sep. 2014.

[BGL16] A. R. Benson, D. F. Gleich, and J. Leskovec. Higher-order organization of complex networks. *Science*, 353:163–166, 2016.

[NDT15] C. Nardini, J. Dent, and P. Tieri. Editorial: Multi-omic data integration. *Front. Cell Dev. Biol.*, 3, Jul. 2015.

[PCB$^+$15] A. Pirayre, C. Couprie, F. Bidard, L. Duval, and J.-C. Pesquet. BRANE Cut: biologically-related a priori network enhancement with graph cuts for gene regulatory network inference. *BMC Bioinformatics*, 16(1):369, Dec. 2015.

[PCDP17] A. Pirayre, C. Couprie, L. Duval, and J.-C. Pesquet. BRANE Clust: cluster-assisted gene regulatory network inference refinement. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 2017.

[1] Feature Learning in Multi-Layer Networks: `http://snap.stanford.edu/ohmnet/`