Multi-omics transfer learning to extend proteomics coverage beyond mass spectrometry quantitation limits

Open PhD position in data science

September 2020

1 Context

Distinct genes are expressed in different cell types and under different conditions, yielding different proteins from cell to cell. Precisely measuring the dynamics of proteins (the 'atoms of life') would provide an unrivaled characterization of biological states. However, methodological obstacles currently impede robust and accurate estimation of protein abundance. On the one hand, the core technology of proteomics (namely mass spectrometry) is hampered by a complex missing data problem [1], with peptides (*i.e.* protein fragments) being missed at random, while others are below the detection threshold. On the other hand, RNA-seq allows to robustly measure abundance of the whole transcriptome, with few missing data, but RNA abundance sometimes lacks correlation with protein abundance.

2 Objectives

Considering, we propose to integrate RNA-seq and mass spectrometry based proteomics. More precisely, and knowing transcription levels do not always reflect protein concentrations, **the goal of this project will be to assess how well transcriptomic can help imputing quantitative proteomics data when peptides fall below the detection limit of the instrument**.

3 Methodology

To achieve this goal, we propose the following roadmap:

1. **Exploratory analysis of paired transcriptomic and proteomic samples.** Preliminary analysis of datasets using standard pipelines and assessment of correlation levels [2] between the two sets of data. Discrepancies between RNA and protein abundances have different sources: (1) not all RNAs are translated into proteins; (2) proteins and RNA have different half-lives; (3) some proteins are transported from other cell-types.

2. **Develop a novel method to estimate protein abundance using jointly transcriptomic and proteomic data.** Leverage the high quality information provided by the transcriptomic data to build a new predictor of protein abundance through the transfer learning / domain adaptation framework [3].

3. **Facilitate reproducible and open science** by sharing the method in a high quality open-source package.

4 Scientific environment

- Within the Fundamental Research division of CEA Grenoble, the lab Exploring the Dynamics of Proteomes (EDyP – `http://www.edyp.fr/web/`) gathers multiple scientific areas of expertise (ranging from biology to applied mathematics) with the aim to develop analytical and computational methods that improve the proteome coverage of complex biological samples.

- The TIMC-IMAG – `https://www-timc.imag.fr/en/` lab gathers scientists and clinicians towards the use of computer science and applied mathematics for understanding and controlling normal and pathological processes in biology and healthcare. Within the lab, the team BCM (Biologie Computationelle et Mathématique) focuses on developing data-driven and modeling methods for biology, living systems, and to better support our healthcare system.

- This project will be financially supported by the *artificial intelligence for high throughput biomedical investigations* program of the Grenoble Multidisciplinary Institute for Artificial Intelligence (MIAI – `https://miai.univ-grenoble-alpes.fr/`), which fosters academic collaborations between Grenoble hospital, academic labs (among which TIMC-IMAG and EDyP), and artificial intelligence industry.

5 Profile

The profile sought is that of a graduate student (Master degree or equivalent) in Computer Science (Major in Artificial Intelligence, Data Science, or Bioinformatics) or in Applied Mathematics (Major in Signal Processing or Statistics) who has a strong interest in interdisciplinary work in biology. They must have programming skills (R or Python) and be fluent in either French or English. Applicants must send their CV to:

- Nelle Varoquaux, CNRS researcher, TIMC-IMAG (https://www-timc.imag.fr/):
  - `nelle.varoquaux@univ-grenoble-alpes.fr`
  - `https://nellev.github.io/`
- Thomas Burger, CNRS researcher, EDyP-lab (`http://www.edyp.fr/web/`):
  - `thomas.burger@cea.fr`
  - `https://sites.google.com/site/thomasburgerswebpage/`

References

[1] Cosmin Lazar, Laurent Gatto, Myriam Ferro, Christophe Bruley, and Thomas Burger. Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies. *Journal of proteome research*, 15(4):1116–1125, 2016.

[2] Malik Tiomoko and Romain Couillet. Random matrix-improved estimation of the wasserstein distance between two centered gaussian distributions. In *2019 27th European Signal Processing Conference (EUSIPCO)*, pages 1–5. IEEE, 2019.

[3] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016.