M2 internship Project: Transcriptomic Analysis using Intensive Randomization

Data Intelligence Institute of Paris (diiP) Masters-level Internship

Contact

Dorota Desaulle, MCF, UR 7537 BioSTM - Biostatistique, Traitement et Modélisation des données biologiques, Faculté of Pharmacie, Université de Paris, France (principal investigator)

email : dorota.desaulle@u-paris.fr

Project

Next-generation sequencing such as RNA-seq aims to quantify the transcriptome of biological samples and compare gene expression between different experimental conditions. The quantification of the genome alignements stemming from such technologies represent the relative measurements which cannot be directly compared between conditions without an adequate data normalization. The optimal approach to normalize such data has not reached a consensus to date (Abrams et al. 2019). Unfortunately, existing methods suffer from practical limitations and may be compromised by the presence of genes showing high expression level or strong variability. In this case a single normalization procedure can lead to erroneous results and false conclusions. Therefore, a novel statistical framework for differential analysis in transcriptomics has been proposed (Desaulle et al. 2021) which is based on intensive iterative random data normalizations and provides good control of the statistical errors. At present, it has been implemented in the R package *DArand* (Desaulle and Rozenholc 2021) and is publicly available from the Comprehensive R Archive Network. The current package is written in R language and uses only CPU parallelization. Due to the large data size and the framework based on intensive iterative randomizations, further project development requires more advance programming. More precicely, the iterative procedure uses intensive computations and may become rapidly time-consuming with respect to both the size of the transcriptomic experiment and the number of samples. Therefore, the main mission during the internship will consist in adapting the code for efficient parallel processing on a graphic processing unit (GPU) using CUDA.The computational optimization will play an important role in further methodological development. Indeed, the subsequent contribution will aim at extending the methodology from two to more biological conditions. It will be directed towards statistical analysis with more than two conditions such as differential analysis, principal component analysis (PCA) and more generally unsupervised learning tools. Here the difficulty will be to preserve an iterative structure of the procedure with data normalization and while combining results from different approaches in data analysis. The methodological aspects, the implementation and the validation will be followed by the real-data application involving the miRNA data.

Masters-level intern

The successful candidate should hold a master degree in data science or computer science with knowledge related to statistics, machine learning or AI and is also expected to interact with the researchers of the interdisciplinary teams throughout the internship.

Moreover any of the following skills will be considered as an advantage

- good programming skills including GPU computing
- strong interest for biology

- advanced level in English

Interdisciplinary teams

This transdisciplinary research project connects informatics, statistics and biology.

The internship will take place in the UR 7537 BioSTM - "Biostatistique, Traitement et Modélisation des données biologiques" at the Faculté de Pharmacie, Université de Paris. The BioSTM team gathers researchers in Data Science. A computer with a GPU unit is being acquired by BioSTM team and will be available for the internship.

The development of this project is a part of the collaboration with the "Unité de Technologies Chimiques et Biologiques pour la Santé" UTCBS CNRS UMR 8258 - INSERM U 1267 on the search of hepatic miRNAs involved in non-alcoholic steatohepatitis (NASH) disease (Hoffmann et al. 2020). The UTCBS will provide sets of transcriptomic data with more than two known conditions.

The project will benefit from the collaboration with two full-time research associates in computer science recruited by the Data Intelligence Institute of Paris.

References

Abrams, Zachary B., Travis S. Johnson, Kun Huang, Philip R. O. Payne, and Kevin Coombes. 2019. "A Protocol to Evaluate RNA Sequencing Normalization Methods." *BMC Bioinformatics* 20 (24): 679. https://doi.org/10.1186/s12859-019-3247-x.

Desaulle, Dorota, Céline Hoffmann, Bernard Hainque, and Yves Rozenholc. 2021. "Differential Analysis in Transcriptomic: The Strength of Randomly Picking 'Reference' Genes." http://arxiv.org/abs/2103.09872.

Desaulle, Dorota, and Yves Rozenholc. 2021. *DArand: Differential Analysis with Random Reference Genes.* https://CRAN.R-project.org/package=DArand.

Hoffmann, Céline, Nour El Houda Djerir, Anne Danckaert, Julien Fernandes, Pascal Roux, Christine Charrueau, Anne-Marie Lachagès, et al. 2020. "Hepatic Stellate Cell Hypertrophy Is Associated with Metabolic Liver Fibrosis." *Sci Rep* 10 (1): 3850. https://doi.org/10.1038/s41598-020-60615-0.