Master 2 Internship:
Discovering relations between the structure of metabolic networks and clinical phenotypes

Supervisors: Hédi Soula[(1)], Nataliya Sokolovska[(1)]
Contact: hedi.soula@sorbonne-universite.fr,
nataliya.sokolovska@sorbonne-universite.fr

[(1)] NutriOmics, UMR S 1269, INSERM, Sorbonne Université

Context Metabolic disorders are rapidly increasing in prevalence as a consequence of the continued obesity epidemic: 6% of the French population have diabetes [de Lagasnerie et al. 2018]. Diabetes has the highest prevalence among all chronic conditions covered 100% by the French healthcare insurance. In 2018, the amount spent on the diabetic population was 19 € billion. Cardio-vascular diseases represent 13% and diabetes 8% of the total health expenses. The main motivation of this project is to discover relevant relations between the structure of metabolic graphs and environmental (i.e. lifestyle) characteristics and individuals' phenotypes.

The Limits of the State-of-the-Art Approaches

Although it is known that human gut microbiota controls factors related to human metabolism, the mechanisms are not studied yet. The most advanced methods to analyze such complex interactions are probably approaches based on graphical models and information graph theory. Standard statistical methods such as correlation and statistical tests can be used but usually they cannot model complex multivariate relationships. The state-of-the-art methods of metabolic networks reconstruction suffer from several drawbacks:

- Although metabolic networks are naturally directed graphs, they are still usually analyzed as undirected graphs, and important information is lost;
- Some parts of the reconstructed graph using tools for metabolic modeling, are missing
- Usually, the direction of graphs is fixed to one proposed by a data base used for the reconstruction, however, this direction can be not correct for particular cases.

In this project, we challenge to overcome the drawbacks of the state-of-the-art methods mentioned above. We focus on metabolite-centered representation where the nodes of a graph are metabolites.

Objectives

The main objective of the project is to go deeper than the state-of-the-art approaches proposed by [1, 2].

Our main focus will be on microbiota metabolic networks reconstruction and analysis. These metabolic networks are reconstructed from metagenomic data and describe the functional abilities of the microbiota. Since it incorporates meta-information (biochemical reactions, redundancy, etc.) it is a step forward compared to traditional differential genomics analysis (what

genes are over/under expressed). Although it is now possible to reconstruct these networks as graphs, mathematical and statistical operations on graphs are challenging and limited, especially clustering methods. Recently, the machine learning community started to develop graph embedding methods where graphs are transformed into meaningful compact representations. We are motivated to develop a new statistical method of directed networks embedding which naturally identifies communities of graphs, i.e. the sub-graphs. This task can be performed, e.g., using stochastic block models [Matias et al. 2018 ; Mehta et al., 2019] which is a technique to learn graphs containing clusters or communities, but also more general groups as, e.g., multipartite structures. Development of a metric which captures relations between the structure of metabolic graphs and its class (phenotype) is our second avenue of research. In the context of this project, we are planning, first, to validate these metrics for the human gut bacteria and its usefulness for obese patients stratification. Second, novel metrics based on the directed graph Laplacian will be proposed and tested in the context of the project. Third, we will develop a dynamic graph model to keep track of evolutions in the graph.

Although it is known that human gut microbiota controls factors related to human metabolism, the mechanisms are not studied yet. The most advanced methods to analyze such complex interactions are probably approaches based on graphical models and information graph theory. Standard statistical methods such as correlation and statistical tests can be used but usually they cannot model complex multivariate relationships.

An ideal candidate will propose and develop machine learning methods for efficient structured data analysis. It is expected that the candidate provides some theoretical foundations for the methods and also implements them in Python.

References

[1] A. Weber Zendrera, N. Sokolovska, H. Soula. Robust structure measures of metabolic networks that predict prokaryotic optimal growth temperature. *BMC Bionformatics*, 2019.

[2] A. Weber Zendrera, N. Sokolovska, H. Soula. Functional prediction of environmental variables using metabolic networks. *Scientific Reports*, 2021.