PhD Proposal

A machine learning approach to discover relations between the topology of metabolic networks and clinical phenotypes

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

Metabolic networks reflect the relationship between metabolites (biomolecules) and enzymes (proteins), and are of particular interest since they describe all chemical reactions of an organism. Metabolic networks are constructed from the genomic sequence of an organism, and graphs can be used to study fluxes through reactions, or to link the structure of graphs to some characteristics and environmental 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*, *information graph theory*, and *deep leaning* [4, 5]. 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, therefore, the reconstructed networks are noisy, potentially with missing information;
- 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 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:

1. To propose and test novel efficient *graph embedding* machine learning methods. In particular, in [1], the *Laplacian spectrum* was related to the environmental conditions, and it was shown that it incorporates important information about the networks structure. In [2], another compact embedding of the metabolic networks, called *scope* was considered. These results are promising, and there is a need to go further in this direction.
2. Novel metrics based on the directed networks will be proposed and tested in the context of the project. As mentioned above, *directed networks* and their properties are much less studied than the undirected ones. However, to preserve the precious information related to fluxes in the networks, we are motivated to focus on the directed graphs.
3. Metabolic networks are specific dynamic structures, having particular properties, e.g., the *deficiency* of a (bio)chemical reaction network is defined as a measure of its ability to support *dynamical behavior* and/or multistationarity [3, 6]. So, specific measures which better describe the topology of the metabolic networks should be considered and developed.

To apply. Please contact Hédi Soula (`hedi.soula@sorbonne-universite.fr`) and Nataliya Sokolovska (`nataliya.sokolovska@sorbonne-universite.fr`). An ideal candidate is supposed to have a Master degree in Bioinformatics/Computer Science/Mathematics/Systems Biology/Engineering. He/she will propose and develop novel machine learning methods for efficient structured data analysis. It is expected that the candidate works on theoretical foundations of the methods and also implements them (Python). We expect that the candidate is interested in biological and medical applications.

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.

[3] M. Feinberg. Foundations of Chemical Reaction Network Theory. *Springer*, 2019.

[4] G. Muzio, L. O'Bray, K. Borgwardt. Biological network analysis with deep learning. *Briefings in Bioinformatics*, 2021.

[5] H. Ali Shah, J. Liu, Z. Yang, J. Feng. Review of Machine Learning Methods for the Prediction and Reconstruction of Metabolic Pathways. Front. Mol. Biosci., 2021.

[6] D. Langary, A. Küken, Z. Nikolski. Nonstoichiometric balanced complexes: Implications on the effective deficiency of the underlying metabolic network. *bioRxiv preprint*, 2021.