Choisissez l'institut ou l'initiative :

Intitulé du Projet de Recherche Doctoral : Graph embedding and clustering : new algorithms for personalized medicine

Directeur de Thèse porteur du projet (titulaire d'une HDR) :

NOM : **ROQUAIN** Prénom : **Etienne**
Titre : Maître de Conférences des Universités ou

e-mail : etienne.roquain@upmc.fr
Adresse professionnelle : Campus Jussieu
(site, adresse, bât., bureau) Tour 15-16, 213
Case courrier 158
4 place Jussieu
75252 Paris Cedex 05

Unité de Recherche :
Intitulé : Laboratoire de Probabilités, Statistique et Modélisation (LPSM)
Code *(ex. UMR xxxx)* : UMR 8001

ED386-Sciences Mathématiques Paris Centre

Ecole Doctorale de rattachement de l'équipe & d'inscription du doctorant :

Doctorants actuellement encadrés par le directeur de thèse (préciser le nombre de doctorants, leur année de 1ere inscription et la quotité d'encadrement) : 0

Co-encadrant :

NOM : **SOKOLOVSKA** Prénom : **Nataliya**
Titre : Maître de Conférences des Universités ou HDR ☒

e-mail : nataliya.sokolovska@gmail.com

Unité de Recherche :
Intitulé : NutriOmics
Code *(ex. UMR xxxx)* : UMR-S 1269

Ecole Doctorale de rattachement : **ED394-Physiologie,Physiopathologie & Thérapeutique**
Ou si ED non Alliance SU :

Doctorants actuellement encadrés par le co-directeur de thèse (préciser le nombre de doctorants, leur année de 1ere inscription et la quotité d'encadrement) : 1 doctorant depuis septembre 2018, encadrement à 50%

Cotutelle internationale : ☒ Non ☐ Oui, précisez Pays et Université :

Description du projet de recherche doctoral *(en français ou en anglais)*

3 pages maximum – interligne simple – Ce texte sera diffusé en ligne

Détailler le contexte, l'objectif scientifique, la justification de l'approche scientifique ainsi que l'adéquation à l'initiative/l'Institut.

Le cas échéant, préciser le rôle de chaque encadrant ainsi que les compétences scientifiques apportées. Indiquer les publications/productions des encadrants en lien avec le projet.
Préciser le profil d'étudiant(e) recherché.

Voir le document pdf joint.

Merci de nommer votre fichier pdf :
«ACRONYME de l'institut/initiative_2_NOM Porteur Projet_2020 »

à envoyer simultanément par e-mail à l'ED de rattachement et au programme :
cd_instituts_et_initiatives@listes.upmc.fr avant le 30 mars.

Graph embedding and clustering : new algorithms for personalized medicine

Directeurs de thèse: Etienne Roquain[(1)], Nataliya Sokolovska[(2)]
Co-encadrants: Tabea Rebafka[(1)], Hédi Soula[(2)]

[(1)] LPSM, Sorbonne Université
[(2)] NutriOmics, INSERM, Hôpital La Pitié, Sorbonne Université

Context Finding relationships between individual features (e.g., composition of the human gut microbiota, food habits, way of life) and a state of illness (e.g., obesity stage) is a very important topic in medical research nowadays [1, 11]. This project aims at developing **new algorithms to efficiently stratify patients** into groups reflecting their clinical status, in the view of designing **personalized treatments**. It will involve novel methods in machine learning. Our proposal is **interdisciplinary**, at the interface of computational statistics, fundamental biology and medicine. The NutriOmics team disposes of huge data sets of various forms and our primary focus is on the analysis of **metabolic networks**. The latter reflect the relationships between metabolites and proteins, and can be derived from genome sequences of an organism. Since these medical data are large scale, we will pay particular attention to developing fast and efficient methods and algorithms.

Scientific aims Our proposal is composed of the four following intertwined parts:

1. development of a quality graph embedding that is interpretable, which provides deeper insights on the understanding of medical data and that comes with a fast inference procedure;
2. development of a novel efficient machine learning algorithm for patient stratification based on neuronal networks to make individual treatment recommendations;
3. development of a graph clustering procedure that improves on the interpretability of patient stratification;
4. development of an outlier detection procedure to identify atypical profiles and increase the reliability of the stratification rules.

From a methodological point of view, the most challenging part of the project is that metabolic networks are **graphs** which are by their complex nature hard to analyze and compare. Recent mathematical tools have to be suitably adapted to accommodate such graph data and hence our work will also contribute to the more general field of network analysis.

Justification of the scientific approach To reduce the complexity of graphs, much current research activity in machine learning is concerned with finding **graph embeddings**, that is vector representations of the graphs. While many such graph embeddings have been investigated recently [8, 3], they are in general difficult to interpret. However, since medical data are extremely heterogeneous and metabolic graph representing reactions in living organisms are huge, developing graph embedding with a **high degree of interpretability** is crucial. In the statistical literature, the stochastic block model provides a small meta-graph representing the general graph topology and a node embedding [5]. While this model increases interpretability, it suffers from high computational costs. Our first task will be to take the best of the two worlds, by bringing together the stochastic block model and neural networks in form of a sparse variational autoencoder and a scalable inference algorithm to compute node embeddings. For this, we will adapt ideas of [6] to clinical graph-structured data, so that the embeddings themselves can give insights into the underlying medical or biological phenomenon. This will already bring a deeper understanding of illnesses or treatment effects.

As a by-product, this embedding will then be used as a data-transformation to enable **supervised classification via neuronal networks**. While a first implementation can easily be done via the open-source neural-network library `keras`, we anticipate that the difficult part will come from the calibration of the training data. Indeed, the response variables should be chosen according to the possible treatments,

but not all treatment combinations are present in the data-base. Solving this issue will lead to new decision rules for patient stratification that can then be used to make individual predictions of the success of a given treatment for new patients.

To increase the interpretability of the result and to gain further insights into biological aspects, a complementary strategy will be to build a non-supervised machine learning rules, which will learn the patient groups without the clinical status, but only via the metabolic network. Doing so, less constrained classes of patients are expected to be identified, which will lead to new treatment strategies. From a methodological point of view, the task consists in developing a **clustering procedure for a collection of graphs** and we propose to use stochastical block models by elaborating upon the work in [9].

Finally, we will develop **outlier detection rules** to identify atypical metabolic networks, which is particularly important to determine patients that are significantly different from the majority and that are to be removed and analyzed apart. The challenge is that our outlier removal procedures will come with a false discovery rate guarantee. Since recent work in this area is built in the case where each observation is a single number and not a graph [2, 10], suitable generalizations will be considered. A direction is to use the new vectorized graph embeddings described above and then generalize the existing tools for multivariate objects.

Fit with the Institute Our interdisciplinary team has complementary skills that both fit into the themes of the Institute of Computing and Data sciences (ISCD) and will allow to cover the various expertise areas needed for the project:

- Nataliya Sokolovska (N.S.): methodological aspects of machine learning;
- Etienne Roquain (E.R.): mathematical and statistical aspects of large scale data analysis;
- Tabea Rebafka (T.R.): algorithmic and computational aspects of network analysis;
- Hédi Soula (H.S.): all areas of system biology, including metabolic network analysis.

Members of this team have already achieved a number of fruitful collaborations in various contexts: N.S and T.R.: two PEPS CNRS projects on random matrices and dimensionality reduction methods, see [4]; E.R and T.R.: new method for graph inference (both from algorithmic and mathematical point of view), see [9]; N.S. and H.S.: supervision of a PhD and collaboration on the metabolic networks reconstruction, see [12]. A tight collaboration between all four members of the team is expected to create stimulating interactions and a synergy that will end up with new breakthrough methods. The PhD project will start in the continuity of the Défi Santé Numérique project (CNRS/INSERM) *Modelling metabolism of intestinal microbiome by multi-omics statistical data integration*, accepted in 2019.

Candidate profile The candidate should have a solid background in mathematics, algorithmics, computational statistics and an interest in medical applications will be highly appreciated. The PhD student will be jointly supervised by the two following research units of Sorbonne Université: Laboratoire de Probabilités, Statistique et Modélisation (LPSM) and NutriOmics.

References

[1] Aron-Wisniewski, J. *et al.* (2019). Major microbiota dysbiosis in severe obesity: fate after bariatric surgery. *Gut*, 68(1):70–82.

[2] Carpentier, A., Delattre, S., Roquain, E., and Verzelen, N. (2020). Estimating minimum effect with outlier selection. *Annals of Statistics*, to appear.

[3] Grover, A. and Leskovec, J. (2016). node2vec: Scalable feature learning for networks. *Proceedings of the 22nd ACM SIGKDD International conference on Knowledge Discovery and Data mining.*

[4] **Kharouf, Malika and Rebafka, Tabea and Sokolovska, Nataliya (2018) Consistent Spectral Methods for Dimensionality Reduction *IEEE, 26th European Signal Processing Conference (EUSIPCO)*, 286–290.**

[5] Matias, C. and Robin, S. (2014). Modeling heterogeneity in random graphs through latent space models: a selective review. *Esaim Proc. & Surveys*, 47:55–74.

[6] Mehta, N., Duke, L. C., and Rai, P. (2019). Stochastic blockmodels meet graph neural networks. *Proceedings of the 36th International Conference on Machine Learning.*

[7] Pavlopoulos, G. A., Secrier, M., Moschopoulos, C. N., Soldatos, T. G., Kossida, S., Aerts, J., Schneider, R., Bagos, P.G. (2011). Using graph theory to analyze biological networks. *BioData Min. 24(1):10.*

[8] Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). Deepwalk. *Proceedings of the 20th ACM SIGKDD International conference on Knowledge Discovery and Data mining.*

[9] **Rebafka, T. and Roquain, E. and Villers, F. (2019). Graph inference with clustering and false discovery rate control. *arXiv:1907.10176.***

[10] Roquain, E. and Verzelen, N. (2020). On using empirical null distributions in Benjamini-Hochberg procedure. *arXiv:1912.03109.*

[11] Selber-Hnatiw, S. *et al.* (2020). Metabolic networks of the human gut microbiota. *Microbiology*, 166(2):96-119.

[12] **Zendrera, Adèle Weber and Sokolovska, Nataliya and Soula, Hédi A. (2019) Robust structure measures of metabolic networks that predict prokaryotic optimal growth temperature *BMC bioinformatics*, 20(1):499.**

(The publications in bold are the publications of the supervisors linked to the proposal)