Proposition de stage de Master 2/dernière année d'école Ingénieur en Biostatistique, Apprentissage automatique ou domaine similaire

Etablissement d'accueil : Equipe SISTM (Statistics In System biology and Translational Medicine), commune à l'INRIA et l'INSERM, Bordeaux, France

Encadrement : Marta Avalos, enseignante-chercheur en Biostatistique, Institut de Santé Publique de l'Université de Bordeaux, Equipe de recherche SISTM

En collaboration avec :

- **Equipe SISTM, INRIA Bordeaux Sud-Ouest** (Rodolphe Thiébaut, PU-PH, Chef d'équipe, Perrine Soret, doctorante)
- **Service de Parasitologie-Mycologie, CHU de Bordeaux** et équipe "Remodelage bronchique" de l'Inserm U1045 (Laurence Delhaes, Médecin, PU-PH, Chef de service)
- **Machine Learning Research Group, Data61, CSIRO, Australia** (Cheng Soon Ong, Principal Researcher, Richard Nock, Senior Principal Researcher)

Sujet : **High-dimensional compositional microbiota data: exploring new methods and software implementations**

Au cours des dernières années, un intérêt croissant a été porté au *microbiote* (l'ensemble des bactéries, champignons et autres micro-organismes de l'organisme humain). En effet, les progrès en bioinformatique et en séquençage de la dernière décennie ont permis des avancées importantes dans la compréhension du rôle du microbiote sur la santé. Le nombre de publications scientifiques sur le sujet a augmenté exponentiellement. En particulier, le microbiote a été associé à certaines pathologies, telles que des maladies inflammatoires de l'intestin, diabètes, cancers, ou encore des pathologies respiratoires chroniques. Pour ces dernières, pourtant, pendant très longtemps, la présence de bactéries dans le poumon était associée à une maladie car, chez l'homme non malade, les voies aériennes inférieures étaient considérées comme stériles. De multiples questions concernant le rôle du microbiote pulmonaire, ses interactions avec l'hôte ou les interactions entre les différentes communautés le constituant (bactérienne, fongique, virale,…) restent néanmoins en suspens et représente un axe de recherche novateur au sein du Service de Parasitologie-Mycologie du CHU de Bordeaux et l'équipe Inserm "Remodelage bronchique". Le projet de recherche multicentrique MucoFong, coordonné par Laurence Delhaes, porte sur l'évaluation et prise en charge du risque fongique chez les patients atteints de mucoviscidose.

Les données microbiote sont usuellement mesurées en tant qu'abondance relative des espèces, c'est-à-dire, il s'agit de données compositionnelles (CoDa pour *Compositional Data*), dont la somme vaut 1. Puisqu'une composante peut être déterminée à partir de la somme du reste de la

composition, les composantes sont mathématiquement et statistiquement dépendantes. Cette structure complique l'analyse et ne permet pas d'effectuer des inférences valides à partir d'analyses statistiques standard. Aitchison, 1982 et Egozcue et collègues, 2003, entre autres, ont fourni un cadre pour analyser des CoDa en projetant les données de l'espace simplex contraint à l'espace euclidien en utilisant des transformées non linéaires telles que la log-cote ou le rapport log-isométrique. Ces développements ont été initialement motivés par des études en géologie et économie. Des transformations non linéaires adaptées à la nature phylogénétique des données microbiote ont récemment été proposées (Silverman et al. 2017). Des niveaux de difficulté supplémentaires sont rajoutés à l'étude des données microbiote, lorsqu'on tient compte de la grande dimension des données provenant de séquençage à haut débit, d'une part, ou de l'hétérogénéité des données intégrant le microbiote (communautés bactérienne et fongique, notamment), d'autre part.

Un exemple de la difficulté du traitement de ces données est l'étude des corrélations entre les composantes bactérienne et fongique du microbiote pulmonaire. Les résultats de ces analyses seront intimement liés à la transformation non linéaire préalable (aucune, rapport log-isométrique, rapport log-isométrique plus information sur la phylogénie) et à la distance entre les deux blocs de données choisie (distance euclidienne, Bray-Curtis, Unifrac,…). Un cadre théorique général qui permet d'étudier le choix de distances adéquates est celui des distances de Bregman, qui unifient la distance euclidienne et font le lien avec les approches probabilistes (famille de distributions exponentielle). L'idée est d'intégrer l'apprentissage à partir des données de la fonction de distance optimale pour le jeu des données disponible (parmi les options jugées pertinentes). L'adaptation des distances de Bregman aux données compositionnelles fait actuellement l'objet de développements théoriques et algorithmiques dans l'équipe australienne de *machine learning* Data61.

L'**objectif principal du stage** est l'analyse des relations entre les différents composants du microbiote pulmonaire (bactérien, fongique), ainsi que l'identification de composantes du microbiote pulmonaire (notamment les composantes fongiques, moins étudiées dans la littérature) ayant un effet sur la sévérité de la mucoviscidose. Pour cela, des approches récentes pour l'analyse et la modélisation de données microbiote seront explorées. Notamment, les développements en cours au sein de l'équipe Data61 constitueront une piste privilégiée.

Références

1. Aitchison J. The Statistical Analysis of Compositional Data. Journal of the Royal Statistical Society. Series B, 44(2):139-177, 1982.

2. Banerjee A, Merugu S, Dhillon IS, Ghosh J. Clustering with Bregman Divergences. Journal of Machine Learning Research, 6:1705-1749, 2005.

3. Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barcelo-Vidal C. Isometric logratio transformations for compositional data analysis. Mathematical Geology, 35(3):279–300, 2003.

4. Nguyen LD, Viscogliosi E, Delhaes L. The lung mycobiome: an emerging field of the human respiratory microbiome. Front Microbiol., 13;6:89, 2015.

5. Nock R, Menon AK, Ong CS. A scaled Bregman Theorem with Applications. In: NIPS, Barcelona, December 2016.

6. Silverman JD, Washburne, AD, Mukherjee S, David LA. A phylogenetic transform enhances analysis of compositional microbiota data. eLife, 6:e21887, 2017.

7. Soret P, Avalos M, Ong CS, Thiébaut R. High-dimensional compositional microbiota data: state-of-the-art of methods and software implementations. In : GdR Statistique et Santé, Bordeaux, Octobre 2017.

8. Yang L, Jin R. Distance metric learning: A comprehensive survey. Technical report, Department of Computer Science and Engineering, Michigan State University, 2006.

Profil attendu du stagiaire :

Des connaissances de base des méthodes d'apprentissage statistique sont nécessaires.

Le (la) stagiaire devra manipuler aisément les outils informatiques. Les quelques peu d'algorithmes qui sont aujourd'hui disponibles sont développés en utilisant différents langages de programmation (python, R, Matlab). Sans avoir besoin d'une connaissance approfondie de chacun de ces langages, il faut de bonnes connaissances en programmation, permettant de jongler selon les besoins.

Dialoguer avec les cliniciens et comprendre leurs problématiques demande un certain investissement et de bonnes capacités d'interprétation.

Les échanges avec l'équipe australienne (oral et écrit) se feront en anglais.

Durée de stage : Entre 4 et 6 mois pendant l'année académique 2017-2018

Gratification de stage : Selon les taux en vigueur (~550 €/mois)

Adresse et contact :

Marta Avalos (marta.avalos-fernandez@inria.fr)

Institut de Santé Publique, Epidémiologie et Développement (ISPED) - Université de Bordeaux

146, Rue Léo-Saignat ; 33076 Bordeaux Cedex

Tél. : 05 57 57 15 34