Offre de stage, 6 mois, 2023, étudiant(e) en master bioinformatique/biostatistique.

Nom et n° de l'unité de recherche d'accueil : Nom du directeur : Adresse :	UMR 1329 INRAE LABERCA Bruno Le Bizec Oniris Vétérinaire, 101 Route de Gachet, CS50707, 44307 Nantes Cedex 3
Encadrantes (LABERCA) : Encadrants (Stat-SC) : Téléphone : E-mail de contact :	Dr. German Cano-Sancho Dr. Jean-Philippe Antignac Pr. Evelyne Vigneau 0240687880 german.cano-sancho@oniris-nantes.fr

Sujet de recherche

« Étude de l'impact des variables de confusion dans le cadre de l'intégration de données exposomiques et métabolomiques en études observationnelles à l'aide de modèles multiblocs. »

Unité LABERCA (laberca.org)

Le Laboratoire d'Étude des Résidus et Contaminants dans les Aliments (LABERCA) est une Unité Mixte de Recherche de l'École Nationale Vétérinaire, Agroalimentaire et de l'Alimentation Nantes Atlantique (Oniris), et de l'Institut national de recherche pour l'agriculture, l'alimentation et l'environnement (INRAE, département AlimH). Il est par ailleurs le Laboratoire National de Référence (LNR) de la Direction Générale de l'Alimentation (DGAl) en ce qui concerne l'analyse des dioxines, des polychlorobiphényles, des hydrocarbures aromatiques polycycliques et des promoteurs de croissance interdits (dont les hormones stéroïdes) dans les denrées alimentaires d'origine animale. Du point de vue scientifique, le domaine d'activité général du laboratoire est celui de la sécurité de l'aliment, et plus précisément celui de l'étude des résidus et contaminants chimiques présents au sein de la chaîne alimentaire, dans une démarche globale d'appréciation du risque depuis l'agrofourniture jusqu'à l'Homme et sa descendance. Le LABERCA s'attache en effet à générer des données et des connaissances relatives aux sources, transferts et métabolismes des composés étudiés, afin de caractériser l'exposition à la fois externe (occurrence dans les denrées) et interne (occurrence dans les fluides et tissus biologiques) des consommateurs vis-à-vis de ces polluants chimiques. Du point de vue analytique, les deux principaux domaines de compétence et de reconnaissance du LABERCA sont d'une part le traitement des échantillons biologiques complexes en vue de l'isolement des substances étudiées présentes au sein de ces matrices à l'état de trace, et d'autre part la mesure fine de ces composés par diverses techniques et couplages basés sur la spectrométrie de masse. Le parc instrumental du laboratoire compte parmi les plus remarquables en Europe dans le domaine de la spectrométrie de masse.

Unité Statistique, Sensométrie et Chimiométrie (Stat-SC) ONIRIS-INRAE

L'unité de Statistique, Sensométrie et Chimiométrie d'Oniris est une unité sous contrat avec l'INRAE. Elle compte aujourd'hui cinq enseignants-chercheurs et deux ingénieurs de recherche. Les thématiques de recherche de l'unité relèvent de la Statistique Appliquée notamment dans le cadre de la Sensométrie et la Chimiométrie. Elles ont notamment trait à l'analyse de plusieurs tableaux de données dans un contexte non supervisé ou supervisé, la classification de variables. L'objectif général des démarches mises en œuvre sont la réduction de la dimensionalité des données, l'identification de biomarqueurs ainsi que la compréhension des relations entre les différents types d'informations. L'unité de Statistique, Sensométrie et Chimiométrie a acquis, grâce à de nombreuses collaborations, une reconnaissance nationale et internationale dans ses domaines de compétence.

Contexte scientifique

Les études de biosurveillance ont démontré que l'Homme est exposé à des mélanges complexes de substances chimiques durant toute sa durée de vie, observation traduite aujourd'hui dans le concept d''exposome'. L'exposition aux substances chimiques exogènes (par exemple les contaminants d'origine anthropique) à certaines concentrations et durant des fenêtres de sensibilité particulières (période périnatale, puberté…) peut avoir un impact sur certains processus biologiques internes et à terme contribuer à l'apparition et/ou au développement de certaines maladies. La mise en évidence des associations causales entre exposition chimique et santé dans les études observationnelles reste

toutefois un grand défi actuel. D'une part les études observationnelles sont soumises au contrôle statistique des différentes variables qui peuvent influencer les expositions et les issues de santé, connues comme variables confondantes. D'autre part la nature des études épidémiologiques est dirigée pour relever des mécanismes biologiques. A cet égard, la nouvelle génération de technologies omiques (e.g . metabolomique, epigenetique) apparait comme un levier pour révéler les potentiels liens fonctionnels entre certains marqueurs d'exposition chimique et certains marqueurs d'effet. Néanmoins les méthodes conventionnelles de régression ne sont pas adaptées pour l'intégration des données omiques. L'un des défis méthodologiques est en particulier le caractère multicorrélé (énormément de liens entre les nombreuses variables). Au cours de la dernière décennie, de nouvelles méthodes statistiques et computationnelles multi-tableaux ont été proposées pour intégrer plusieurs tableaux de données tels que les couches de données type « Omiques » (e.g. transcriptomique, protéomique) (Li et al 2012 ; Tenenhaus and Tenenhaus 2011). Les facteurs de confusion, couramment observés dans les études biologiques à haut débit, peuvent cependant affecter les performances de ces méthodes et d'autres analyses en aval (Lin et al 2016). En outre, l'intégration et les différentes stratégies de prise en compte des variables de confusion dans les approches de type multi-tableaux (ou multiblocs) restent peu explorés.

Objectifs

Dans ce contexte, le projet proposé a pour objet global de développer et de mettre en œuvre une stratégie d'intégration de l'effet des variables de confusion afin de progresser dans l'étude du lien environnement-santé.

Les principaux objectifs de ce projet sont :

(1) de réaliser un état de l'art concernant les méthodes et démarches actuellement disponibles pour considérer des variables de confusion et/ou supprimer la variabilité non désirable ;

(2) d'évaluer l'influence des différentes démarches permettant de gérer les variables de confusion dans l'intégration de données exposomiques et métabolomiques à travers deux cas d'étude.

Méthodologie envisagée

1. **État de l'art** concernant l'étude de l'impact des variables de confusion lors de l'utilisation de modèles dit multiblocs lorsqu'il s'agit d'étudier des données multi-omiques dans le but d'identifier d'éventuels chemins d'actions menant à un impact sur la santé humaine. Une bibliographie approfondie devra permettre à la fois de comprendre l'enjeu de considérer les variables de confusion dans le cadre susmentionné, mais également d'identifier des méthodes permettant de les considérer, qui seront étudiées par la suite.
2. **Évaluation de l'impact des démarches permettant de gérer les variables de confusion** dans l'application de modèles multiblocs, préalablement choisis et étudiés, avec l'objectif d'identifier des données métabolomiques et exposomiques dans deux cas d'études :
    - **Cas d'étude 1, intégration de données multi-plateforme analytique autour du lait maternel**. Des données de phénotypage moléculaire (exposomique, nutrimetabolomique) de lait maternel ont été générées dans le cadre du Projet régional LactOMICS pour 60 échantillons prélevés chez des mères d'enfants nés prématurés (Cano-Sancho et al., 2020).
    - **Cas d'étude 2, intégration de données de biomarqueurs endogènes inflammatoires (oxylipines et cytokines) et exposomiques en lien avec l'âge gestationnel**. Ce cas d'étude s'appuie sur les données publiées par Aung et al. (2019). Dans ce cas d'étude, le but est d'identifier les cascades d'effets allant de l''exposome' chimique, à la santé humaine (âge gestationnel), en passant par la dysfonction métabolique indicatrice des profils inflammatoires.

    Cette partie fera notamment appel à la mise en œuvre de démarches de comparaison des résultats lorsque l'on fait varier des variables potentiellement confondantes. Des solutions mettant en œuvre aussi bien des connaissances théoriques sur les effets de ces types de variables, mais également une capacité à produire un protocole et un code sous R permettant l'évaluation de leurs impacts dans notre cas d'étude, seront attendues. Les outils logiciels utilisés seront différents packages sous environnement R.

Résultats attendus

- Catalogue des méthodes existent dans la littérature pour prendre en compte et mesurer l'impact des variables de confusion dans un cadre multiblocs appliqué à l'intégration de données -omiques en études de santé, et en particulier environnementale.
- Identification des impacts des variables confondantes en fonction de leurs caractéristiques à l'aide d'un protocole d'évaluation de cet impact.

Références

Aung MT, Song Y, Ferguson KK, Cantonwine DE, Zeng L, McElrath TF, Pennathur S, Meeker JD, Mukherjee B. Application of an analytical framework for multivariate mediation analysis of environmental data. Nat Commun. 2020 Nov 6;11(1):5624.

Cano-Sancho G, Alexandre-Gouabau MC, Moyon T, Royer AL, Guitton Y, Billard H, Darmaun D, Rozé JC, Boquien CY, Le Bizec B, Antignac JP. Simultaneous exploration of nutrients and pollutants in human milk and their impact on preterm infant growth: An integrative cross-platform approach. Environ Res. 2020 Mar;182:109018. doi: 10.1016/j.envres.2019.109018. Epub 2019 Dec 13. PMID: 31863943.

Lin Z, Yang C, Zhu Y, Duchi J, Fu Y, Wang Y, Jiang B, Zamanighomi M, Xu X, Li M, Sestan N, Zhao H, Wong WH. Simultaneous dimension reduction and adjustment for confounding variation. Proc Natl Acad Sci U S A. 2016 Dec 20;113(51):14662-14667. doi: 10.1073/pnas.1617317113. Epub 2016 Dec 7. PMID: 27930330; PMCID: PMC5187682.

Tenenhaus, A., & Tenenhaus, M. (2011). Regularized Generalized Canonical Correlation Analysis. Psychometrika, 76(2), 257-284. doi:10.1007/s11336-011-9206-8

Li, W., Zhang, S., Liu, C. C., & Zhou, X. J. (2012). Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. *Bioinformatics, 28*(19), 2458-2466.

Compétences

Solides connaissances en statistiques et analyse de données (Master 2 en statistiques / bioinformatique). Maîtrise du langage de programmation R. Intérêt pour le traitement de données de santé.

Lieu de stage

Laboratoire d'Etude des Résidus et Contaminants dans les Aliments (LABERCA) UMR INRAE 1329 ONIRIS Site de la Chantrerie, 101 Route de Gachet, La Chantrerie CS50707, 44307 Nantes Cedex 3

Unité Stat-SC ONIRIS Site de Géraudière, Rue de la Géraudière, CS 82225, 44322 Nantes

Rémunération

- Environ 554,40 €/mois.