



Offer #2025-09116

Doctorant F/H Modélisation de données de contamination environnementale issues de méthodes d'analyse non-ciblées

The offer description below is in French

Contract type : Fixed-term contract

Level of qualifications required : Graduate degree or equivalent

Fonction : PhD Position

About the research centre or Inria department

Le centre Inria de l'Université de Rennes est l'un des huit centres d'Inria et compte plus d'une trentaine d'équipes de recherche. Le centre Inria est un acteur majeur et reconnu dans le domaine des sciences numériques. Il est au cœur d'un riche écosystème de R&D et d'innovation : PME fortement innovantes, grands groupes industriels, pôles de compétitivité, acteurs de la recherche et de l'enseignement supérieur, laboratoires d'excellence, institut de recherche technologique.

Assignment

Problématique. L'exposome représente l'ensemble des expositions auxquelles une personne est soumise tout au long de sa vie, incluant les environnements chimiques, microbiologiques, physiques, récréatifs et médicaux, ainsi que le mode de vie, l'alimentation et les infections. La grossesse (période prénatale), l'enfance et la puberté ont été identifiées comme des périodes particulièrement sensibles, durant lesquelles les expositions environnementales peuvent influencer les trajectoires de santé individuelles. L'épidémiologie au cours de la vie a besoin d'outils pour étudier les marqueurs d'exposition et leurs effets sur la santé de plus en plus complexes. Les analyses non-ciblées basées sur l'utilisation de la chromatographie liquide couplée à la spectrométrie de masse haute résolution (LC-HRMS) offrent la promesse d'identifier, voire quantifier de manière globale les polluants présents dans les matrices biologiques telles que les urines, le sang, les cheveux. Le spectromètre de masse joue le rôle de détecteur et mesure le rapport masse/charge des ions détectés dans un échantillon, ainsi que l'abondance associée. La chromatographie liquide en amont permet de séparer les composés de manière à décomplexifier un échantillon. Des données en 3 dimensions formant des pics sont ainsi obtenues (m/z , intensité, temps de rétention). Dans une approche non ciblée, nous ne nous intéressons pas à des polluants particuliers prédéfinis, mais à l'ensemble de l'empreinte chimique caractérisée par de

multiples pics correspondant à des molécules identifiées ou non. Plusieurs défis restent à relever pour exploiter de manière efficace ces données massives: les polluants d'intérêt sont peu abondants et sont masqués par les composés endogènes, ils sont donc particulièrement difficiles à détecter. Par ailleurs, tous les pics ne peuvent être décrits par la même "courbe mathématique" (i.e., gaussienne). Enfin, les techniques utilisées pour l'enregistrement de ces données sont spécifiques aux laboratoires et l'analyse conjointe des profils d'exposition produits par ces différents laboratoires est aussi un challenge non résolu.

Objectifs. L'analyse de ces données vise, comme premier objectif, à mettre en relation les pics détectés avec un événement de santé pour identifier ceux qui lui sont associés puis à les interpréter en termes de molécules en essayant de les annoter. Un deuxième objectif, non supervisé, est l'identification de profils d'expositions homogènes.

Projet

Approche existante. Cet objectif global est actuellement traité en deux grandes étapes dans la littérature. Une première étape de pré-traitement, concomitante à l'acquisition des spectres, consiste à réduire l'ensemble du spectre à une matrice position/intensité résumant l'information moléculaire de l'échantillon. Cette matrice est ensuite utilisée, dans une deuxième étape, comme entrée de modèles d'apprentissage classiques, dans un cadre supervisé ou non, pour expliquer/prédire un événement ou identifier des profils d'individus. Une telle approche présente plusieurs limites. En premier lieu, le pré-traitement des spectres par ces méthodes sont composées de plusieurs étapes. Ces différentes étapes dépendent de nombreux paramètres à spécifier et accroissent de ce fait la subjectivité liée à l'utilisateur. Un des défis est donc de chercher à réduire ce nombre de paramètres ou d'automatiser leur choix. Par ailleurs, chaque étape est source d'erreurs statistiques qui ne sont que peu quantifiées ou prises en compte dans les méthodes existantes. Il est ainsi nécessaire de quantifier l'incertitude découlant de chaque étape du processus de traitement comme un moyen d'assurer une meilleure évaluation de la qualité des données.

Ce projet de thèse, en collaboration avec l'IRSET, vise à développer une approche plus globale afin de réduire les étapes de prétraitement et l'incertitude découlant des erreurs propagées par les étapes successives. Pour ce faire, nous proposons une modélisation fonctionnelle du spectre à l'aide de bases de fonctions flexibles et adaptées aux caractéristiques des spectres acquis. Parmi les difficultés liées aux spectres, une première est que les pics observés de ces données de LC-HRMS pour les différents individus ne sont pas correctement alignés, nous pourrions intégrer dans nos modèles une étape d'alignement basé sur le transport optimal et la distance de Wassertein. Par ailleurs, les polluants présents dans les échantillons biologiques correspondent généralement à des pics de petite taille dont l'intensité est proche du niveau du bruit, notre modèle devra donc en tenir compte afin de séparer les pics associés à des molécules réelles de ceux correspondant à du bruit. Enfin, les différentes variabilités, telles que celles dues aux différentes techniques des laboratoires, ou structures de groupes seront prises en compte dans le modèle final à l'aide d'effets mixtes. Nous définirons également un terme de pénalité spécifiquement adapté à la sélection de portions de courbes.

Cette modélisation nous permettra d'identifier, sans a priori, les polluants dont l'effet est le plus significatif sur un événement de santé et pourra être adaptée au cas où la variable d'intérêt est une durée de vie telle que le décès ou l'apparition d'un cancer.

Main activities

- Développer un modèle fonctionnel pour l'analyse des données LC-HRMS en intégrant une étape d'alignement des courbes
- Etudier les performances du modèle sur des données simulées et des données réelles
- Développement d'un package R ou Python
- Diffuser les travaux via des publications et des exposés

Skills

Les candidats doivent être titulaires d'un master (ou équivalent) en mathématiques appliquées ou en statistiques. Ils doivent manifester un fort intérêt pour les applications à l'exposome et à la santé environnementale.

Benefits package

- Restauration subventionnée
- Transports publics remboursés partiellement
- Congés: 7 semaines de congés annuels + 10 jours de RTT (base temps plein) + possibilité d'autorisations d'absence exceptionnelle (ex : enfants malades, déménagement)
- Possibilité de télétravail (après 6 mois d'ancienneté) et aménagement du temps de travail
- Équipements professionnels à disposition (visioconférence, prêts de matériels informatiques, etc.)
- Prestations sociales, culturelles et sportives (Association de gestion des œuvres sociales d'Inria)
- Accès à la formation professionnelle
- Sécurité sociale

Remuneration

Salaire brut : 2200€

General Information

- **Town/city** : Rennes
- **Inria Center** : [Centre Inria de l'Université de Rennes](#)
- **Starting date** : 2025-10-01
- **Duration of contract** : 3 years
- **Deadline to apply** : 2025-09-22

Contacts

- **Inria Team** : AT-REN AE

- **PhD Supervisor :**
Gares Valerie / valerie.gares@inria.fr

About Inria

Inria is the French national research institute dedicated to digital science and technology. It employs 2,600 people. Its 200 agile project teams, generally run jointly with academic partners, include more than 3,500 scientists and engineers working to meet the challenges of digital technology, often at the interface with other disciplines. The Institute also employs numerous talents in over forty different professions. 900 research support staff contribute to the preparation and development of scientific and entrepreneurial projects that have a worldwide impact.

The keys to success

Vous pouvez donner là, un portrait à "gros traits" du (de la) collaborateur(trice) attendu(e) : ce que vous voyez comme nécessaire et suffisant et qui peut associer :

- goûts et appétences,
- domaine d'excellence,
- éléments de personnalité ou de caractère,
- savoir et savoir faire transversaux...

Cette rubrique permet de compléter et alléger (réduire) la liste plus formelle des compétences :

- "Se sentir à l'aise dans un environnement de dynamique scientifique, aimer apprendre et écouter sont des qualités essentielles pour réussir cette mission."
- " Passionné(e) par l'innovation, avec une expertise dans le développement Ruby on Rail et une grande capacité de conviction. Une thèse dans le domaine *** constitue un réel atout."

Warning : you must enter your e-mail address in order to save your application to Inria. Applications must be submitted online on the Inria website. Processing of applications sent from other channels is not guaranteed.

Instruction to apply

Déposer en ligne CV et lettre de motivation

Defence Security :

This position is likely to be situated in a restricted area (ZRR), as defined in Decree No. 2011-1425 relating to the protection of national scientific and technical potential (PPST). Authorisation to enter an area is granted by the director of the unit, following a favourable Ministerial decision, as defined in the decree of 3 July 2012 relating to the PPST. An unfavourable Ministerial decision in respect of a position situated in a ZRR would result in the cancellation of the appointment.

Recruitment Policy :

As part of its diversity policy, all Inria positions are accessible to people with disabilities.