Sujet de stage : Modélisation de données de contamination environnementale issues de méthodes d'analyse non-ciblées

November 13, 2024

Contexte

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 [1, 2]. 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, concomittante 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 [3]. 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 nom-

bre 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.

Approche proposée. Notre projet est d'adopter une approche plus globale afin de réduire les étapes de prétraitement et l'incertitude découlant des erreurs propagées par les étapes successives [4]. 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 pourrons intégrer dans nos modèles une étape d'alignement basé sur le transport optimal et la distance de Wassertein [5]. 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.

Ce stage est susceptible de déboucher sur une thèse.

References

[1] J. Chaker, D. M. Kristensen, T. I. Halldorsson, S. F. Olsen, C. Monfort, C. Chevrier, B. Jégou, and A. David. Comprehensive evaluation of blood plasma and serum sample preparations for hrms-based chemical exposomics: overlaps and specificities. *Analytical Chemistry*, 94(2):866–874, 2022.

[2] A. David, J. Chaker, E. J. Price, V. Bessonneau, A. J. Chetwynd, C. M. Vitale, J. Klánová, D. I. Walker, J.-P. Antignac, R. Barouki, et al. Towards a comprehensive characterisation of the human internal chemical exposome: Challenges and perspectives. *Environment International*, 156:106630, 2021.

[3] G. Renner and M. Reuschenbach. Critical review on data processing algorithms in non-target screening: challenges and opportunities to improve result comparability. *Analytical and Bioanalytical Chemistry*, 415(18):4111–4123, 2023.

[4] Alejandro Sánchez Brotons, Jonatan O. Eriksson, Marcel Kwiatkowski, Justina C. Wolters, Ido P. Kema, Andrei Barcaru, Folkert Kuipers, Stephan J. L. Bakker, Rainer Bischoff, Frank Suits, and Péter Horvatovich. Pipelines and systems for threshold-avoiding quantification of lc–ms/ms data. *Analytical Chemistry*, 93(32):11215–11224, 2021. PMID: 34355890.

[5] M. Thorpe, S. Park, S. Kolouri, G. K. Rohde, and D. Slepčev. A transportation l^p distance for signal analysis. *Journal of mathematical imaging and vision*, 59:187–210, 2017.

Compétences requises

- Etudiant 5ème année ou Master 2
- Modélisation statistique
- Maîtrise du logiciel R

Encadrement

- **Valérie Garès** `valerie.gares@insa-rennes.fr` (Chercheuse à l'INRIA Rennes),
- **Madison Giacofci** `joyce.giacofci@univ-rennes2.fr` (Maîtresse de Conférences à Rennes 2),
- **Ioana Gavra** `ioana.gavra@univ-rennes1.fr` (Maîtresse de Conférences à Rennes 2)
- et **Sarah Lennon** `sarah.lennon@univ-rennes.fr` (Maitresse de conférence, Université de Rennes 1)

Candidature

Pour candidater, vous pouvez envoyer **un cv et une lettre de motivation avant le** 31 décembre à valerie.gares@insa-rennes.fr, joyce.giacofci@univ-rennes2.fr, ioana.gavra@univ-rennes1.fr et sarah.lennon@univ-rennes.fr.