# Evaluer l'impact de l'environnement sur la santé humaine : l'enjeu du traitement des données de spectrométrie de masse haute-résolution

Jade Chaker*[1], Erwann Gilles , Sophie Lefevre-Arbogast , Cécilia Samieri , Sarah Lennon , and Arthur David

[1]Univ Rennes, Inserm, EHESP, Irset (Institut de recherche en santé, environnement et travail) - UMR$_S$1085, $F-35000 Rennes, France--UnivRennes--France$

**Résumé**

La définition du concept d'exposome par Wild en 2005(1) a marqué le début d'un intérêt croissant de la communauté scientifique pour la caractérisation des liens entre les facteurs environnementaux et la santé humaine. Parmi ces facteurs, l'exposition aux contaminants chimiques environnementaux est fortement soupçonnée de contribuer à la survenue d'événements de santé défavorables, tels que le vieillissement cérébral et les pathologies associées, en particulier la maladie d'Alzheimer. Dans ce contexte, l'objectif du projet Eglantine est de caractériser l'exposome chimique humain par le biais d'échantillons sanguins de 500 individus de plus de 65 ans, avec un suivi de 100 individus sur 10 ans, pour établir des liens entre expositions chimiques environnementales et vieillissement cérébral.
Les analyses utilisées pour évaluer l'exposition humaine aux contaminants environnementaux sont, pour la plupart, ciblées. En d'autres termes, elles sont axées sur la détection et la quantification de composés préalablement définis dans un échantillon. Les méthodes ciblées sont hautement sélectives et sensibles, permettant une mesure précise des composés d'intérêt. Pour complémenter ces approches, il est possible d'avoir recours à la chromatographie liquide couplée à la spectrométrie de masse haute résolution (SMHR) pour effectuer des analyses plus exploratoires, dites non ciblées. Ces dernières visent à détecter et à caractériser tous les composés présents dans un échantillon, sans préjuger de leur identité, qui est déterminée dans un second temps à l'aide d'algorithmes d'annotation(2). Ces approches offrent une vision plus globale de la composition chimique d'un échantillon, mais produisent des données spectrales massives et complexes (jusqu'à 100,000 composés/ions détectés par analyse, avec pour chacun un rapport masse-sur-charge, un temps de rétention, et une aire pour chaque échantillon), nécessitant un traitement spécifique.

La première étape du traitement des données non-ciblées consiste à la production d'une matrice regroupant la liste des ions détectés, ainsi que l'aire du signal correspondant dans chaque échantillon. L'intégration des données spectrales issues de la SMHR est un enjeu particulièrement critique, puisque les niveaux de réponse des composés sont utilisés pour parvenir à établir des corrélations avec la survenue de l'événement de santé étudié. Il est donc crucial de s'assurer de la pertinence de l'algorithme et du paramétrage utilisé pour produire cette matrice d'ions en limitant le nombre de faux positifs (i.e. " détection " d'un signal non-existant) et surtout le nombre de faux négatifs (i.e. non-détection d'un signal existant). De plus, les échantillons sanguins sont complexes par nature, et contiennent des composés

*Intervenant

présentant des concentrations s'étendant sur jusqu'à 8 ordres de grandeur, les marqueurs d'exposition environnementale étant majoritairement parmi les moins abondants(3). Il est donc également nécessaire de s'assurer que l'algorithme de détection et d'intégration des pics soit robuste face à cette variation importante de caractéristiques de pics.

Lors de ce travail, quatre algorithmes ont été évalués : l'algorithme vendeur MarkerViewTM1.3 (AB SCIEX), l'algorithme vendeur Progenesis QI for MetabolomicsTM (Waters), l'algorithme open-source Continuous Wavelet Transformation (CWT) (implémenté dans le package R *xcms*(4) et dans le logiciel MZMine2(5)), et l'algorithme open-source Automated Data Analysis Pipeline (ADAP) (implémenté dans le logiciel MZMine2). Ce travail d'optimisation et de comparaison a été effectué en utilisant les données issues du dopage à 10 ng/mL d'échantillons de 4 échantillons de plasma et de sérum avec une solution contenant 45 molécules (4 échantillons de chaque matrice ont été analysés sans dopage pour servir de référence). Chaque algorithme a tout d'abord été optimisé individuellement, manuellement et automatiquement si possible (i.e. paramétrage automatisée de CWT dans *xcms* par IPO(6) et Autotuner(7)). Les résultats obtenus suite à ces paramétrages optimisés ont été comparés entre eux. Cette comparaison a été effectuée sur cinq critères : la fréquence de détection, le temps de calcul, la facilité d'implémentation, la répétabilité de l'intégration automatique, et la significativité de la détection (i.e. résultat du t-test comparant les aires associées aux composés dopants entre les échantillons dopés et non-dopés).

Dans un premier temps, il a été démontré que l'utilisation d'outils automatisés de paramétrage des algorithmes de détection et intégration des pics, initialement développés pour la métabolomique (i.e. caractérisation de marqueurs biologiques endogènes, souvent plus abondants), n'était pas adaptée aux applications en exposomique (i.e. caractérisation de marqueurs issus d'exposition environnementale, souvent peu abondants). Ainsi, le paramétrage suggéré par IPO, basé sur les pics jugés " fiables " a résulté en une largeur de pic trop élevée (30.7 s), car il base sa notion de fiabilité sur les composés les plus abondants. Ce paramétrage a mené à une détection de moins de 30% des composés dopés dans les deux matrices. A l'inverse, l'outil Autotuner a suggéré une largeur de pic trop faible ($< 10$ s), menant ainsi à une scission excessive des pics détectés, et donc à une mauvaise performance en répétabilité ($<$ 20% des composés avec une répétabilité satisfaisante). L'optimisation manuelle a donc été préférée dans le cadre de l'application considérée. Il a, dans un second temps, été constaté que l'optimisation individuelle des outils permettait d'augmenter la fréquence de détection des composés de 25 à 85% (*xcms*). En effet, certains paramètres comme la largeur de pic et le niveau de bruit généralement proposés par défaut ne sont pas optimisés pour les pics moins abondants, et doivent donc être réduits. De plus, bien que les outils open source permettent d'avoir plus de visibilité et de maîtrise sur le choix des algorithmes et des paramètres utilisés, ils nécessitent une meilleure connaissance technique et présentent des temps de calcul 4 à 16 fois plus long que les logiciels vendeurs. L'algorithme ADAP, décrit comme étant particulièrement efficace pour éliminer les faux positifs comparé à l'algorithme CWT(8), a en effet permis d'obtenir les meilleurs résultats en fréquence de détection des composés dopés (98%), répétabilité de l'intégration automatique (98%), et significativité de la détection (p-value $<$ 10-4), mais le moins bon résultat en temps de calcul (18h contre une dizaine de minutes pour Markerview).

En conclusion, tous les algorithmes ont permis d'obtenir des performances satisfaisantes en termes de fréquence de détection, de répétabilité et de significativité de détection après optimisation. Il demeure cependant nécessaire de continuer à optimiser ces outils de traitement des données issues de la SMHR, car aucun algorithme n'a permis de détecter tous les composés dopés identifiés manuellement. La compréhension des limites de l'étape de production de la matrice d'ions est cruciale pour envisager d'appliquer les méthodes non-ciblées à large échelle comme c'est le cas pour le projet Eglantine, pour parvenir à une quantification relative fiable des composés corrélés au vieillissement cérébral.

(1) C. P. Wild, " Complementing the Genome with an "Exposome": The Outstanding Challenge of Environmental Exposure Measurement in Molecular Epidemiology ", *Cancer Epidemiology, Biomarkers & Prevention*, vol. 14, no 8, p. 1847-1850, août 2005, doi: 10.1158/1055-9965.EPI-05-0456.

(2) J. Chaker, E. Gilles, C. Monfort, C. Chevrier, S. Lennon, et A. David, " Scannotation: A Suspect Screening Tool for the Rapid Pre-Annotation of the Human LC-HRMS-Based Chemical Exposome ", *Environ. Sci. Technol.*, nov. 2023, doi: 10.1021/acs.est.3c04764.

(3) A. David *et al.*, " Towards a comprehensive characterisation of the human internal chemical exposome: Challenges and perspectives ", *Environment International*, vol. 156, p. 106630, nov. 2021, doi: 10.1016/j.envint.2021.106630.

(4) C. A. Smith, E. J. Want, G. O'Maille, R. Abagyan, et G. Siuzdak, " XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification ", *Anal. Chem.*, vol. 78, no 3, p. 779-787, févr. 2006, doi: 10.1021/ac051437y.

(5) T. Pluskal, S. Castillo, A. Villar-Briones, et M. Orešič, " MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data ", *BMC Bioinformatics*, vol. 11, no 1, p. 395, juill. 2010, doi: 10.1186/1471-2105-11-395.

(6) G. Libiseller *et al.*, " IPO: a tool for automated optimization of XCMS parameters ", *BMC Bioinformatics*, vol. 16, no 1, p. 118, avr. 2015, doi: 10.1186/s12859-015-0562-8.

(7) C. McLean et E. B. Kujawinski, " AutoTuner: High Fidelity and Robust Parameter Selection for Metabolomics Data Processing ", *Anal. Chem.*, vol. 92, no 8, p. 5724-5732, avr. 2020, doi: 10.1021/acs.analchem.9b04804.
(8) X. Du, A. Smirnov, T. Pluskal, W. Jia, et S. Sumner, " Metabolomics Data Preprocessing Using ADAP and MZmine 2 ", *Methods Mol Biol*, vol. 2104, p. 25-48, 2020, doi: 10.1007/978-1-0716-0239-3_3.