

Machine-learning-based analysis of Pseudo Mass-Spectrum Images for Targeted Peptides Identification

Encadrants :

- Zied Bouyahya (zied.bouyahya@ec-lyon.fr), Centrale Lyon, LIRIS (CNRS UMR 5205).
- Léo Schneider (leo.schneider@etu.ec-lyon.fr), Centrale Lyon, LIRIS et ISA (CNRS UMR 5280).

Date butoir pour candidater : 1^{er} décembre 2024.

Contexte

Ce projet est plus particulièrement lié à un programme de recherche interdisciplinaire (financement ANR, RHU IDBIORIV - 2019-2026) incluant des médecins, des microbiologistes cliniciens et des statisticiens (Institut des Agents Infectieux, Centre International de Recherche en Infectiologie), et dont l'objectif principal est de développer de nouvelles méthodes basées sur la spectrométrie de masse et des pipelines de traitements automatiques pour réduire le délai de l'étape de diagnostic des infections du sang, des urines ou toute autre matrice biologique.

Le projet de recherche vise au développement d'algorithmes, basé sur l'apprentissage profond (*deep learning*), permettant de traiter les données brutes en mode DIA (Data-independent acquisition) [7], et capable de fournir des informations directement interprétables par les médecins infectiologues. Il s'agira en particulier d'identifier des micro-organismes pathogènes et prédire les niveaux des résistances aux antibiotiques de ces mêmes pathogènes.

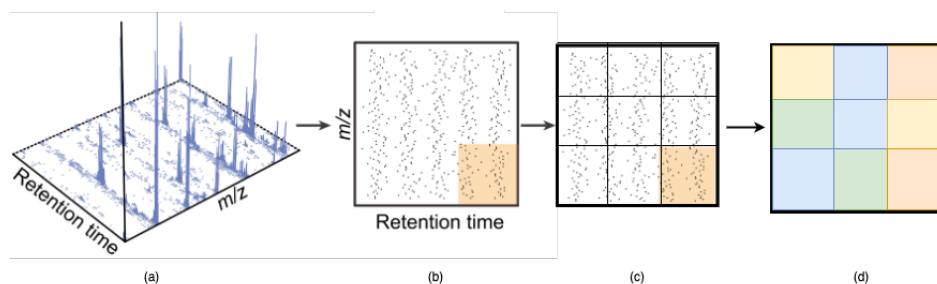


Figure 1 : Processus de transformation des données originales en image.

Sujet du stage

Les données issues d'un chromatogramme reconstitué sur la valeur m/z des ions se présentent sous la forme de séries temporelles représentant le temps de rétention (min.) pour différents rapports masses-sur-charge (m/z). Les méthodes d'analyse traditionnelles consistent en 4 étapes : (1) prétraitement des données brutes, (2) nettoyage des données, (3) identification des peptides, notamment par *peak-picking* [5] et (4) diagnostic. Dans le cas de méthodes automatiques, l'identification correspond à une comparaison entre les spectres expérimentaux et une bibliothèque de spectres théoriques. De tels spectres peuvent être générés automatiquement via des méthodes comme Prosit [6] (cf. Figure 2). L'identification automatique des peptides est une tâche particulièrement ambitieuse en raison de la complexité du bruit et des interférences liées à la matrice biologique complexe [3] ainsi que des contraintes de délais spécifiques inhérentes aux projet RHU IDBIORIV.

Dans ce projet, nous étudions également une approche différente de la méthode traditionnelle, qui consiste à transformer les gros volumes de données 1D en images monovaluées [5] sur lesquelles des méthodes de reconnaissance de formes et d'apprentissage statistique et/ou profond pourront s'appliquer (cf. Figure 2).

Les résultats seront objectivés par le calcul de performances des taux d'identification des peptides cibles (sensibilité et spécificité), en bout de chaîne, grâce à la connaissance précise des échantillons

biologiques analysés au sein des équipes de l'Institut des Sciences Analytiques (J. Lemoine) et de l'Institut des Agents Infectieux de l'Hôpital de la Croix Rousse (F. Vandenesch).

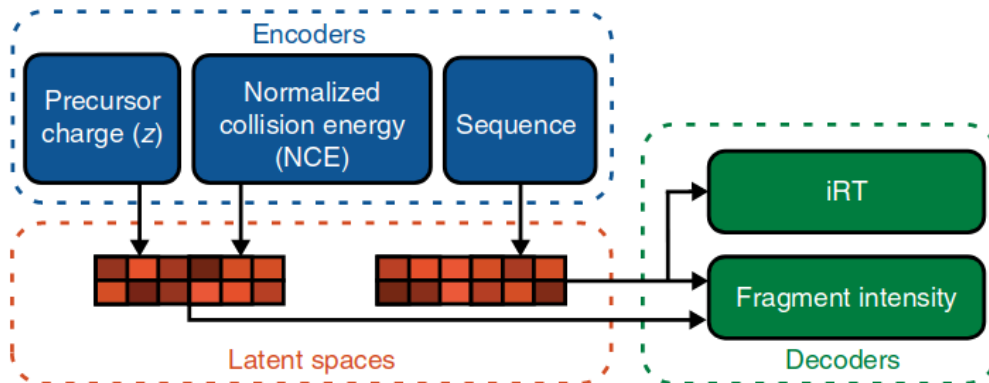


Figure 2 : Prosit [6] pipeline pour la génération de bibliothèques spectrales.

Compétences requises

L'étudiant.e devra disposer de bonnes compétences dans les domaines du *Machine Learning*, particulièrement dans les modèles de vision ou dans le traitement des séquences. Une expérience certaine de la programmation en langage Python est également requise pour implémenter les solutions envisagées (*pytorch*), les évaluer et les comparer sur le corpus de données.

Informations pratiques :

- Lieu du stage : Laboratoire LIRIS - Centrale Lyon (site Écully).
- Période de stage : A partir de Février-Mars 2025, pour une durée de 5 à 6 mois.
- Rémunéré : Oui (4.35 euros/h, 35h/semaine, soit environ 630 euros/mois).
- ZRR : L'étudiant.e engagé.e devra suivre une procédure administrative pour intégrer le LIRIS.

Pour candidater :

- Merci d'envoyer votre CV aux DEUX (2) encadrants (adresses mail ci-dessus).
- **Date butoir pour candidater : 1er décembre 2024.**
- Les étudiants présélectionnés seront invités à un entretien, et classés selon leur ordre de mérite et leurs motivations pour le sujet.

Éléments bibliographiques :

- [1] Shen X, et al. *Deep learning-based pseudo-mass spectrometry imaging analysis for precision medicine*. *Brief Bioinform.* 2022 Sep 20;23(5).
- [2] Wishart, David S., *Emerging applications of metabolomics in drug discovery and precision medicine*, *Nature reviews Drug discovery* 15.7 (2016): 473-484.
- [3] Chaleckis, Romanas, et al., *Challenges, progress and promises of metabolite annotation for LC-MS-based metabolomics*, *Current opinion in biotechnology*, 55 (2019): 44-50.
- [4] Wang, H, Yandong Y, and Zheng-Jiang Z. *Encoding LC-MS-Based Untargeted Metabolomics Data into Images toward AI-Based Clinical Diagnosis*, *Analytical Chemistry* (2023).
- [5] Guo, J, et Tao H. *Mechanistic Understanding of the Discrepancies between Common Peak Picking Algorithms in Liquid Chromatography-Mass Spectrometry-Based Metabolomics*, *Analytical Chemistry* 95, no 14 (2023): 5894-5902.
- [6] Gessulat, S., et al (2019). Prosit: Proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nature Methods*, 16(6), Article 6.
- [7] Jiapeng Liet al, Data-independent acquisition (DIA): An emerging proteomics technology for analysis of drug-metabolizing enzymes and transporters, *Drug Discovery Today: Technologies*, Volume 39, 2021, Pages 49-56, ISSN 1740-6749.