Avis de Soutenance

Madame Agnès DUPAS

Chimie

Soutiendra publiquement ses travaux de thèse intitulés
Intégration de la variabilité génétique bactérienne dans le traitement de données protéomiques en DIA (LC-MS/MS) : application à Legionella pneumophila.

Travaux dirigés par Monsieur Jerôme LEMOINE

Soutenance prévue le ***vendredi 05 juin 2026*** à 13h30
Lieu : Université Lyon 1, bâtiment Irène Joliot Curie - amphithéâtre des thèses au 3 rue Enrico Fermi à Villeurbanne

Composition du jury proposé

M. Jerôme LEMOINE	Professeur des universités	Université Lyon 1	Directeur de thèse
M. Jean ARMENGAUD	Directeur de recherche	CEA Marcoule	Rapporteur
Mme Virginie BRUN	Directeur de recherche	CEA Leti	Rapporteure
Mme Carole CHAIX	Directrice de recherche	CNRS Lyon	Examinatrice
Mme Sophie JARRAUD	Professeure des universités - praticienne hospitalière	Université Lyon 1	Examinatrice
M. Yannick CHARRETIER	Chercheur associé	BioMérieux - Marcy l'étoile	Examinateur
M. Abdelrahim ZOUED	CNRS Lyon	Invité	

Mots-clés : Protéomique, Spectrométrie de masse, DIA, Variabilité allélique, Legionella pneumophila

Résumé :
La protéomique par acquisition indépendante des données (DIA) en spectrométrie de masse est aujourd'hui largement utilisée pour l'analyse de protéomes bactériens entiers. Les avancées dans le traitement de données DIA permettent d'obtenir des identifications fiables de protéines et d'améliorer la couverture des protéomes. Cependant, certaines limitations persistent lorsqu'il s'agit d'intégrer la variabilité génétique, source majeure de diversité phénotypique. L'étude des protéomes bactériens nécessite donc des approches capables de prendre en compte cette diversité. Deux problématiques principales ont été mises en évidence. La première concerne la définition de l'homologie de séquence, afin de déterminer dans quelle mesure deux séquences protéiques mutées peuvent être considérées comme une même protéine ou une même fonction biologique. La seconde porte sur le traitement des données DIA et plus précisément sur l'ambiguïté des peptides, qui

peuvent être communs à plusieurs séquences, en particulier dans un contexte riche en variants. Dans ce cadre, la protéospécificité des peptides constitue un élément clé pour améliorer la fiabilité de l'inférence protéique. L'objectif de cette thèse est de développer une approche de traitement de données permettant d'intégrer la variabilité génétique bactérienne et de réduire l'ambiguïté des peptides, afin d'améliorer l'inférence protéique à partir de données DIA. Pour cela, une stratégie de regroupement des séquences protéiques a été mise en place, basée sur des alignements et des critères de couverture et d'identité de séquence, en amont de la construction de la bibliothèque spectrale. Ces regroupements ont permis de définir deux niveaux de spécificité des peptides : des peptides communs à plusieurs séquences d'un même groupe, considérés spécifiques d'une protéine canonique, et des peptides spécifiques d'une séquence variante donnée. Cette reformulation de la protéospécificité a permis d'identifier les protéines canoniques, tout en attribuant, lorsque possible, l'information du variant en particulier. Le flux a été développé et évalué sur 15 isolats de Legionella pneumophila, et sa capacité à réaliser un protéotypage a été validée. Ce travail propose ainsi une méthodologie pour intégrer la variabilité génétique dans l'analyse protéomique de bactéries en DIA et ouvre des perspectives pour la comparaison de cohortes de souches bactériennes sur la base de leur protéome, au plus proche de leur phénotype.

Summary:

Data-independent acquisition (DIA) mass spectrometry-based proteomics is now widely used for the analysis of whole bacterial proteomes. Advances in DIA data processing enable reliable protein identification and improved proteome coverage. However, limitations remain when it comes to integrating genetic variability, a major source of phenotypic diversity. The study of bacterial proteomes therefore requires approaches capable of accounting for this diversity. Two main challenges were identified. The first concerns the definition of sequence homology, in order to determine to what extent mutated protein sequences can be considered as the same protein or biological function. The second relates to DIA data processing, and more specifically to peptide ambiguity, as peptides may be shared across multiple sequences, particularly in variant-rich contexts. In this framework, peptide proteospecificity is a key factor for improving the reliability of protein inference. The objective of this thesis is to develop a data processing approach that integrates bacterial genetic variability and reduces peptide ambiguity, in order to improve protein inference from DIA data. To this end, a strategy based on protein sequence grouping was implemented, relying on sequence alignments and coverage and identity criteria prior to spectral library construction. These groupings allowed the definition of two levels of peptide specificity: peptides shared among sequences within a group, considered specific to a canonical protein, and peptides specific to a given variant sequence. This reformulation of proteospecificity enables the identification of canonical proteins while assigning, when possible, the corresponding variant information. The workflow was developed and evaluated on 15 Legionella pneumophila isolates, and its ability to perform proteotyping was validated. This work thus provides a methodology to integrate genetic variability into bacterial DIA proteomics and opens perspectives for comparing cohorts of bacterial strains based on their proteome, bringing analyses closer to their phenotype.