

THÈSE DE DOCTORAT DE

NANTES UNIVERSITÉ

ÉCOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies
de l'Information et de la Communication
Spécialité : *Informatique*

Par

Albane LYSIAK

Développement de méthodes informatiques pour l'évaluation et l'amélioration de l'identification par spectrométrie de masse des peptides modifiés

Thèse présentée et soutenue à Nantes, le 6 décembre 2022
Unité de recherche : LS2N (Université de Nantes), BIA (INRAE Nantes)

Rapporteurs avant soutenance :

Thierry LECROQ Professeur des universités, LITIS/Université de Rouen Normandie
Christine CARAPITO Directrice de recherche CNRS, IPHC/Université de Strasbourg

Composition du Jury :

Président :	Thomas BURGER	Directeur de recherche CNRS, EDyp/Université Grenoble-Alpes
Examineurs :	Thierry LECROQ	Professeur des universités, LITIS/Université de Rouen Normandie
	Christine CARAPITO	Directrice de recherche CNRS, IPHC/Université de Strasbourg
	Martial REY	Chargé de recherche CNRS, Institut Pasteur (Paris)
Dir. de thèse :	Guillaume FERTIN	Professeur des universités, LS2N/Université de Nantes
Co-dir. de thèse :	Dominique TESSIER	Ingénieure de recherche, BIA/INRAE Nantes
Co-enc. de thèse :	Géraldine JEAN	Enseignant-chercheur, LS2N/Université de Nantes

Invité(s) :

Hélène ROGNIAUX Ingénieure de recherche, BIA/INRAE Nantes

Titre : Développement de méthodes informatiques pour l'évaluation et l'amélioration de l'identification par spectrométrie de masse des peptides modifiés

Mot clés : Bioinformatique, protéomique, spectrométrie de masse, peptide, OMS

Résumé : La spectrométrie de masse (MS) est l'une des méthodes privilégiées pour identifier les protéines. Elles sont habituellement identifiées à partir de leurs peptides. Pour cela, les spectres obtenus à partir des peptides sont comparés à une base de données de spectres théoriques à l'aide d'un score de similarité, et les PSMs (*Peptide-Spectrum Matches*) produits aident à l'identification des spectres. Cependant, la plupart des spectres générés ne peuvent pas être correctement identifiés. L'une des raisons est que les protéines, et donc les peptides résultants, portent des modifications, ce qui complexifie l'identification des spectres issus de ces peptides. Pour résoudre ce problème, les approches OMS (*Open Mass Search*) offrent des éléments

prometteurs. Cependant ces méthodes sont encore confrontées à certains obstacles, surtout lorsque le peptide comporte des modifications inconnues. De plus, l'évaluation de leurs résultats est complexe. Dans le cadre de cette thèse, j'ai d'abord évalué des stratégies OMS à l'aide de spectres théoriques jouant le rôle de spectres expérimentaux, ce qui a permis de développer de nouveaux critères pour évaluer les PSMs. À la suite de ce travail, j'ai développé `SpecGlob`, un algorithme qui repose sur la programmation dynamique pour aligner les masses contenues dans un PSM, afin d'identifier plusieurs modifications inconnues qui séparent le peptide qui a généré le spectre à identifier du peptide candidat.

Title: Development of computational methods to evaluate and improve the identification by mass spectrometry of modified peptides

Keywords: Bioinformatics, proteomics, mass spectrometry, peptide, OMS

Abstract: Mass spectrometry (MS) is one of the main methods used to identify proteins. They are usually identified from their peptides. To this end, spectra obtained from their peptides are compared to a database of theoretical spectra using a similarity score, and resulting PSMs (*Peptide-Spectrum Matches*) are used to identify spectra. Nevertheless, most of the spectra that are generated cannot be properly identified. One of the reasons is that proteins, and thus resulting peptides, carry modifications that complexify the identification of corresponding spectra. To solve this problem, OMS (*Open Mass Search*) methods

offer promising elements. But these methods still face obstacles, especially when the peptide carries unknown modifications. Moreover, the evaluation of their results is complex. During my PhD, I first evaluated OMS strategies with theoretical spectra playing the role of experimental spectra, which enabled the development of new criteria to evaluate PSMs. Following this work, I developed `SpecGlob`, an algorithm that relies on dynamic programming to align masses of spectra of a PSM, aiming at identifying several unknown modifications separating the peptide that generated the spectrum from the candidate peptide.