Nom et Prénom : EL JADID SARA
Date de soutenance : 12/12/2020
Directeur de Thèse : R. TOUAHNI

Sujet de Thèse :

Développement de Nouvelles Méthodes pour l'Analyse des Données Protéomiques

Résumé :

Grâce aux avancées récentes de la protéomique basée sur la spectrométrie de masse, la compréhension des mécanismes cellulaires, de la pathogenèse et de la relation entre génotype et phénotype a considérablement progressé. Le spectromètre mesure la masse des fragments de protéines et génère des spectres expérimentaux qui représentent des séries de pics indiquant la présence des différents fragments. En analysant ces spectres, nous essayons d'identifier la protéine d'origine. Deux méthodologies sont communément utilisées pour pouvoir identifier les protéines présentes dans l'échantillon d'origine, la recherche dans les bases de données, et le séquençage de novo de peptide. Cependant, la complexité de données liées à ce type d'analyse exige des outils informatiques sophistiqués afin de produire des résultats plus précis et des interprétations adéquates.

En se basant sur les deux méthodologies citées ci-dessus, l'objectif de cette thèse est de proposer de nouvelles méthodes alliant une meilleure sensibilité et spécificité pour une identification plus précise des protéines. Nous proposons tout d'abord un nouveau workflow d'analyse des données de spectrométrie de masse en tandem couplée à la chromatographie liquide à haute performance (HPLC-MS/MS) pour l'identification des protéines à travers la recherche dans les bases de données. Le workflow est basé sur une combinaison d'approches statistiques qui ont montré un meilleur compromis de sensibilité-spécificité. Nous avons validé les résultats de notre approche sur une étude portante sur la physiopathologie d'incontinence urinaire de stress (SUI).

Dans une deuxième phase, nous avons developpé IsoUniNovoR, un outil permettant le séquençage de novo des peptides en incorporant l'information isotopique pour remédier au souci de faible précision de masse qui biaise le résultat de l'identification. Nous avons choisi d'implementer notre outil sous R, car c'est l'environnement par excellence de la communauté des chercheurs en biologie et biotechnologie.

MOTS-CLES:
Protéomique, Spectrométrie de Masse, Identification des Protéines, Isotopes, Moteurs de Recherche dans les Bases de Données, Séquençage De novo, Calcul à Haute Performance, Faux Positifs.

Abstract :

Recent advances in proteomics based on Mass Spectrometry (MS) have led to considerable progress in the understanding of cellular mechanisms, pathogenesis and the relationship between genotype and phenotype. MS is an essential technique in proteomics for the identification of unknown proteins. The spectrometer measures the mass of protein fragments and generates experimental spectra which represent a series of peaks indicating the presence of the different fragments. Through the analysis of these spectra, we attempt to identify the original protein. Two methodologies are commonly used to identify the proteins present in the original sample, database searching, and de novo peptide sequencing. However, the complexity of data related to this type of analysis requires sophisticated computational tools to produce more accurate results and suitable interpretations.

Based on the two methodologies mentioned above, this thesis aims to propose new methods combining a better sensitivity and specificity for more precise protein identification. We first propose a new workflow for the analysis of High Performance Liquid Chromatography-tandem Mass Spectrometry data for protein identification through database searching. The workflow is based on a combination of statistical approaches that have shown a better sensitivity and

specificity trade-off. We validated the results of our approach to a study on the pathophysiology of Stress Urinary Incontinence.

In a second phase, we have developed IsoUniNovoR software, which allows de novo sequencing of peptides by incorporating isotopic information to address the low mass accuracy issue that biases the identification result. We chose to implement our tool under R because it is the environment of choice for the biology and biotechnology research community where, until now, no de novo sequencing algorithms have been available..

KEY WORDS:
Proteomics, Mass Spectrometry, Protein Identication, Isotopes, Database Search Engine, De Novo Sequencing, High-Performance Computing, False Discovery Rate.