Un algorithme de regroupement pour gérer l'identification des protéines dans les analyses de protéomique à haut-débit

Benoit Valot[1] and Olivier Langella[1]

UMR de GÉNÉTIQUE VÉGÉTALE, 91190 Gif-sur-Yvette, France
Olivier.Langella@moulon.inra.fr

Contents

Abstract

L'identification des protéines par les moteurs de recherche est une étape nécessaire aux analyses de protéomique *bottom-up* à haut débit. Elle est basée sur l'identification des peptides et sur leur assemblage le long des séquences protéiques. En général les moteurs de recherche se limitent au classement des protéines identifiées par ordre de score ou de probabilité. Lorsque plusieurs protéines contiennent des peptides identiques, un post-traitement est nécessaire pour lever les ambiguïtés. Classiquement il est basé sur l'élimination par regroupement des protéines identifiées uniquement par des peptides communs à d'autres protéines. Mais cela ne suffit pas pour gérer les cas d'intersections entre groupes de peptides communs.

Nous présentons ici un algorithme original basé sur un deuxième niveau de regroupement, qui réunit les protéines contenant au moins un peptide en commun. Il permet de détecter des protéines redondantes non éliminées par le premier regroupement, de connaître précisément le nombre de peptides communs et spécifiques à chaque protéine et d'estimer le nombre d'isoformes identifiées dans l'échantillon. Par ailleurs nous avons vérifié qu'à de rares exceptions près, ce niveau de classement regroupe les protéines par fonction.

1 Difficultés de l'identification des protéines par analyse *bottom-up*

L'identification des protéines par les moteurs de recherche à partir d'analyse LC-MS/MS est devenue une procédure standard dans les analyses protéomiques à haut-débit. Ce type de

données nécessite une validation statistique (filtre probabilistique, FDR, ...), mais aussi de déterminer quelles protéines sont réellement présentes à partir des peptides identifiés.

Les protéines, qui sont des molécules de taille importante sont rarement analysées directement dans des approches à haut-débit. Classiquement, les protéines vont être préalablement digérées en peptides par des enzymes protéolytiques (trypsine). Ce sont ces derniers qui vont être analysés par spectrométrie de masse. Donc, l'identification se fait au niveau peptidique, et par reconstruction on accède aux protéines.

Il est assez aisé à partir des bases de données de connaître le lien entre peptides et protéines. Mais, les familles multigéniques et les domaines conservés sont autant de cas pour lesquels un peptide va correspondre à plusieurs protéines différentes. À cela s'ajoutent les problèmes d'unicité dans les bases de données, plusieurs accessions pouvant faire référence au même gène. On voit qu'il faut alors pouvoir déterminer de façon fiable et reproductible quelles protéines sont présentes à partir des peptides identifiés.

2 Présentation d'un algorithme de regroupement

C'est pour répondre à ces problèmes de redondance et de peptides communs que nous avons développé un nouvel algorithme de regroupement. Il s'effectue en 3 étapes : (i) reconstruction de l'ensemble des protéines à partir des peptides; (ii) regroupement des protéines ayant le même jeu de peptides (sous-groupes) et élimination des protéines ayant un sous ensemble. (iii) regroupement des *sous-groupes* partageant au moins un peptide en commun pour former un *groupe* de protéines. La troisième étape permet de détecter des protéines redondantes non éliminées par la deuxième étape et de connaître précisément le nombre de peptides communs et spécifiques à chaque protéine.

2.1 Reconstruction des protéines

Cette étape consiste uniquement à répertorier pour chaque protéine les spectres identifiés (figure 1). Ce sont les spectres et non les séquences peptidiques qui sont répertoriés. En effet, les données expérimentales sont les spectres MS/MS et un spectre peut identifier plusieurs peptides (ex: spectre 201, figure 1).

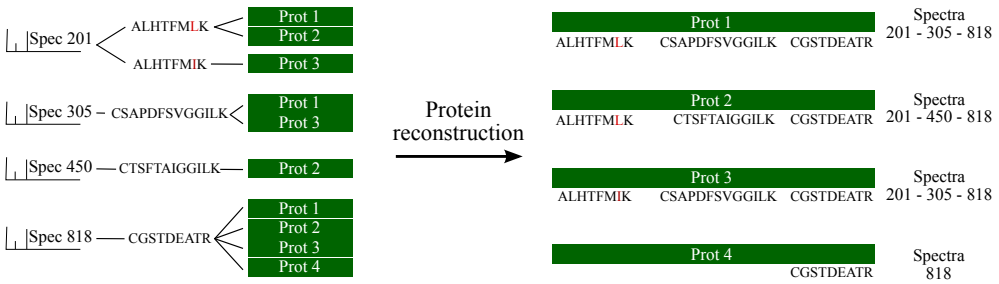

Figure 1: Reconstruction des protéines

A la fin de cette première étape, on obtient la liste complète des protéines incluant la redondance qu'il faut maintenant éliminer.

2.2 Création des sous-groupes

Cette deuxième étape permet d'éliminer la majorité des erreurs d'identification (figure 2).

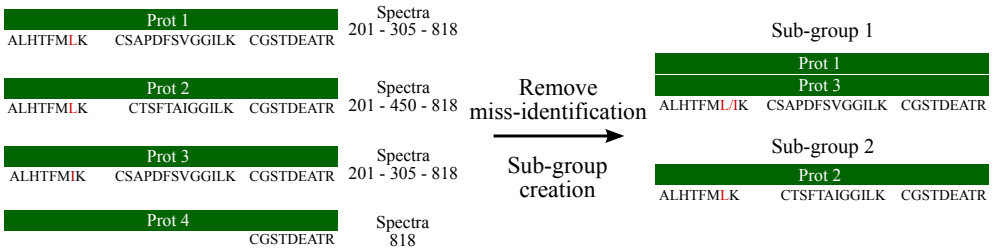

Figure 2: Création des sous-groupes

L'ensemble des protéines ayant un même jeu de spectres sont regroupées au sein d'un *sous-groupe*. Ces protéines sont indiscernables car on n'a aucune information pour déterminer si elles sont toutes présentes ou seulement l'une d'entre elles (ex: protéines 1 et 3 dans le *sous-groupe* 1). Les protéines identifiées avec un sous-ensemble de spectres sont éliminées car aucun peptide spéficique ne permet d'affirmer qu'elles sont présentes dans l'échantillon (ex: protéine 4).

2.3 Création des groupes

Un *groupe* est composé d'un ensemble de *sous-groupes* partageant au moins une séquence peptidique (figure 3). Cette troisième étape permet d'éliminer les protéines qui ne contiennent qu'une combinaison de peptides déjà identifiés dans d'autres protéines. A la suite, l'accès aux peptides spécifiques et communs de chaque *sous-groupe* est facilité.

Par ailleurs, nous avons vérifié qu'à de rares exceptions près, les protéines d'un même *groupe* appartiennent à une seule famille fonctionnelle. Cela permet d'accéder rapidement à la quantité de protéines identifiées (nombre de *sous-groupes*) et de familles de protéines (nombre de *groupes*).

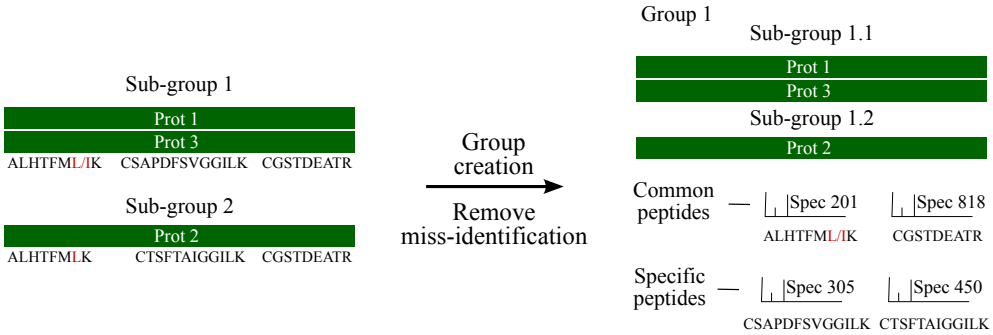

Figure 3: Création des groupes

3 Présentation de X!TandemPipeline

X!TandemPipeline propose une interface graphique pour effectuer le filtrage, la visualisation et la manipulation des résultats d'identification de protéines par spectrométrie de masse (MS/MS). Ces derniers peuvent être chargés à partir de fichiers Mascot .dat ou X!Tandem .xml.

3.1 Fonctions

Outre les filtres classiques à appliquer sur les résultats d'identification (valeurs seuil de FDR, e-value peptide/proteins, nombre de peptides identifiés...) au moment du chargement des données, X!TandemPipeline permet d'appliquer notre algorithme de grouping. 3 modes de chargement sont proposés : (i) **Individual**: Chaque analyse est filtrée, puis groupée séparément (Spots 2D,...). (ii) **Combined**: Les résultats d'identification des différentes analyses sont rassemblées. Le filtre et le regroupement s'effectuent sur l'ensemble (SDS-PAGE, Shotgun...). (iii) **Phosphopeptides**: Seul les peptides phosphorylés sont conservés et filtrés. Afin d'éliminer les incertitudes sur les positions, les peptides sont regroupés par un algorithme particulier non présenté ici.

X!TandemPipeline offre une vue complète des résultats d'identification, leur édition et leur export en format tableur. L'export en formats XML est également possible pour permettre de nouveaux traitements en utilisant éventuellement d'autres logiciels développés par la plateforme PAPPSO.

3.2 Performances

X!TandemPipeline a été développé pour pouvoir tourner sur une configuration basique. A titre d'exemple, le traitement d'une analyse classique de protéomique à haut-débit (43 analyses LC-MS/MS d'échantillons de protéines totales de levure) sur un ordinateur classique (1,4Go RAM, 3Ghz CPU) a duré 53s pour la lecture des fichiers résultats et 28s pour le filtrage et le regroupement des protéines. Au final, sur les 1728 protéines identifiées pour un total de 292737 spectres MS/MS assignés, le regroupement a donné 1353 *groupes* de protéines pour un total de 1490 *sous-groupes* de protéines distinctes.

3.3 Disponibilité

X!TandemPipeline est un logiciel libre (licence GPL), disponible au téléchargement sur notre site web [1] pour Windows, Linux, MacOs. Le dépôt subversion utilisé est public [2]. Chacun peut contribuer, utiliser tout ou partie du code dans un projet tiers s'il le désir, à la seule condition de respecter la licence d'utilisation.

4 Conclusions

La confiance dans les listes de protéines identifiées est une étape majeure du traitement des données de protéomique. Elle passe par une validation statistique des résultats d'identification. Le regroupement des protéines en *groupe* et *sous-groupe* permet à la fois d'éliminer les fausses identifications dues seulement à l'éxistence de peptides partagés et de classer efficacement les peptides selon leur spécifité ou non.

[1] `http://pappso.inra.fr/bioinfo`
[2] `https://sourcesup.renater.fr/projects/xtandempipeline/`