THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par *l'Université Toulouse III - Paul Sabatier*
Discipline ou spécialité : *Bioinformatique*

Présentée et soutenue par *David Bouyssié*
Le *31/05/2012*

Titre : *Développement de nouveaux outils bioinformatiques pour l'exploitation des données de spectrométrie de masse en protéomique haut-débit*

JURY
Dr. Pierre-Alain Binz, Institut Suisse de Bioinformatique (Genève) – Rapporteur
Dr. Myriam Ferro, Laboratoire EDyP/CEA (Grenoble) - Rapportrice
Dr. Philippe Marin, IGF/CNRS (Montpellier) - Examinateur
Dr. Bernard Monsarrat, IPBS/CNRS (Toulouse) - Directeur de thèse
Dr. Christine Gaspin, UBIA/INRA (Auzeville) – Co-directrice de thèse
Pr. Gwennaële Fichant, Université Paul Sabatier (Toulouse) - Examinatrice

---

Ecole doctorale : *Biologie - Santé - Biotechnologies*
Unité de recherche : *Institut de Pharmacologie et de Biologie Structurale (UMR5089, CNRS)*
Directeur(s) de Thèse : *Bernard Monssarat, Christine Gaspin*
Rapporteurs : *Pierre-Alain Binz, Myriam Ferro*

A ma famille,

A Aline

« If the doors of perception were cleansed everything would appear to man as it is, infinite. »
William Blake

Résumé

En biologie, la spectrométrie de masse est devenue l'outil incontournable pour l'identification des protéines. Associée à des techniques de séparation, elle est aussi utilisée pour mesurer la variation d'abondance des protéines entre plusieurs échantillons. Cependant, la très grande quantité et complexité des données liées à ce type d'analyse requièrent des programmes informatiques sophistiqués et adaptés.

Mon travail de doctorat a consisté à répondre aux différentes problématiques liées à l'exploitation des données nanoLC-MS/MS, à savoir la validation des résultats d'identification ainsi que la quantification relative des protéines pour des approches mettant en œuvre ou non un marquage isotopique. Le logiciel MFPaQ, dont deux versions sont présentées dans ce document, en est le principal résultat. La version 3 intègre des fonctionnalités telle que la validation des données Mascot, la génération de listes non-redondantes de protéines et la quantification d'analyses ICAT. La version 4, évolution majeure du logiciel, incorpore des algorithmes adaptés à l'analyse quantitative de données MS sans marquage, ainsi que la gestion des stratégies de marquage SILAC et $^{14}N/^{15}N$. Son utilisation a facilité la réalisation d'études protéomiques, dont certaines, auxquelles j'ai plus particulièrement participé, sont ici présentées. Afin de répondre aux futurs enjeux informatiques de la protéomique, j'ai entrepris dans un second temps le développement du logiciel Prosper, qui dispose d'une architecture d'organisation des données permettant de réaliser des requêtes croisées sur l'ensemble des échantillons analysés. Il constitue aussi un outil prototype pour l'élaboration de nouveaux algorithmes.

Mots-clés : Protéomique, Spectrométrie de masse, nanoLC-MS/MS, Mascot, Quantification de mélanges complexes de protéines, ICAT, SILAC, 14N/15N, label-free, carte LC-MS, logiciel.

Abstract

In biology, mass spectrometry has become an indispensable tool for protein identification. Associated with separation techniques, it can also be used to measure the variation of protein abundance between different samples. However, due to the huge quantity and complexity of the data produced by this kind of analysis, sophisticated and suitable computer programs are needed.

My PhD work was to address the different issues related to the processing of nanoLC-MS/MS data, namely the validation of the identification results, and the relative quantification of proteins using approaches based or not on isotopic labeling. The MFPaQ program, two versions of which are presented here, is the main result of this work. Version 3 includes features such as Mascot data validation, generation of non-redundant protein lists and quantification of ICAT analyses. Version 4, which represents a major upgrade of the software, incorporates additional algorithms for quantitative analysis of label-free MS data, as well as for the handling of the $^{14}N/^{15}N$ and SILAC labeling strategies. This bioinformatic tool has been used for various proteomic studies, some of which are discussed here. In order to meet future IT challenges in proteomics, I undertook later the development of the Prosper software, which is based on an optimized architecture for organizing data, and allows performing cross-queries on all analysed samples. It also constitutes a prototype tool for the development and evaluation of new algorithms.

Keywords: Proteomics, Mass spectrometry, nanoLC-MS/MS, Mascot, complex protein mixture quantification, ICAT, SILAC, 14N/15N, label-free, LC-MS map, software.

Remerciements

Ce travail de thèse a été une aventure personnelle très enrichissante et je tiens à remercier toutes les personnes qui m'ont soutenu pour la réalisation de ce projet.

Mes premiers remerciements vont à Bernard Monsarrat et Odile Schiltz : merci à vous deux pour vos encouragements ainsi que pour l'ensemble des conseils que vous m'avez donnés durant mon doctorat. Je remercie aussi vivement Christine Gaspin qui a accepté de co-diriger le déroulement de cette thèse.
Je tiens ensuite à exprimer ma sincère gratitude à Pierre-Alain-Binz, Myriam Ferro et Philippe Marin pour avoir accepté d'évaluer mon travail de thèse.

Durant six ans, j'ai collaboré avec de nombreuses personnes du laboratoire pour le développement des différents outils informatiques qui sont présentés dans ce manuscrit. Cependant, l'une d'entre elles s'est impliquée plus particulièrement dans la conception et la valorisation de ce travail. Je tiens donc à faire un remerciement particulier à Anne Gonzalez de Peredo pour son réel investissement dans ce projet, ainsi que l'ensemble de ses conseils avisés.

Mon aide quotidienne, et non des moindres, vient de mes deux collègues bio-informaticiens et colocataires de bureau. Merci à toi Emmanuelle pour les heures passées à débuguer tous ces outils et ainsi que pour tes nombreux conseils et encouragements. Un grand merci également à toi Renaud, pour qui les ordinateurs n'ont plus de secret, et qui a toujours été là pour me faciliter le travail.

Que serait devenu le logiciel MFPaQ sans son fan club ? Nul ne le sait, mais je tiens en tout cas à faire une « spécial dédikass » à ses créatrices, Mariette et Karima, ainsi qu'à l'ensemble de ses membres : Alexandre, Anne-Aurélie, Bertrand, Carine, Carole, Chrystelle, Florence, Laure, Luc, Marie-Pierre, Manue, Marlène, Mathilde, Michel, Sandrine et Violette. C'est un réel plaisir de travailler avec vous et de vous côtoyer quotidiennement, et je tenais par cette occasion à le souligner.

Je tiens à exprimer toute ma reconnaissance à mes parents qui ont toujours souhaité que j'aille le plus loin possible sans jamais m'imposer une direction particulière. Merci donc à vous deux pour m'avoir soutenu pendant toutes ces années.

Enfin, un immense merci à ma future épouse, Aline, qui m'a épaulé durant ces six années.

Et puisque deux merci valent mieux qu'un, à vous tous, à nouveau, un grand merci !

Sommaire

Liste des principales abréviations

AMT	Accurate Mass and Time tags
ARN	Acide Ribonucléique
CID	Collision-Induced Dissociation
CV	Coefficient of Variation
DDA	Data Dependent Acquisition
DRM	Detergent Resistant Membranes
ESI	Electrospray Ionization
FDR	False Discovery Rate
HPLC	High Performance Liquid Chromatography
HUVEC	Human Umbilical Vein Endothelial Cells
ICAT	Isotope-Coded Affinity Tag
LC	Liquid Chromatography
LC-MS	Liquid Chromatography coupled to Mass Spectrometry
MALDI	Matrix-Assisted Laser Desorption/Ionisation
MS	Mass Spectrometry
PAI	Protein Abundance Index
PMF	Peptide Mass Fingerprint
Q-TOF	Quadrupole Time-Of-Flight
ROC	Receiver Operator Characteristic
SDS-PAGE	Sodium Dodecyl Sulfate Poly-Acrylamide Gel Electrophoresis
SILAC	Stable Isotope Labeling by Amino acids in Cell culture
TFP	Taux de Faux Positifs
XIC	eXtracted Ion Chromatogram
XML	eXtended Markup Language

Introduction
L'analyse protéomique, applications et enjeux

Cartographier le génome, décrypter les mécanismes impliqués dans son expression, étudier la diversité, la fonction et la dynamique des protéines ainsi que de l'ensemble des métabolites d'une cellule, constituent des enjeux majeurs dans la compréhension du vivant. Cette quête d'informations qui repose sur un ensemble de techniques « –OMIQUES » (génomique, transcriptomique, protéomique et métabolomique) est non seulement essentielle pour la compréhension de processus biologiques complexes (Ideker, Galitski et al. 2001; Csete and Doyle 2002), mais aussi pour envisager, à terme, la mise en œuvre de thérapies nouvelles.

Les protéines, produits finaux des gènes, sont les principaux effecteurs des fonctions biologiques des cellules. Le « protéome » désigne l'ensemble des protéines exprimées par un génome pour une espèce, un organe, une cellule ou un compartiment cellulaire (par exemple : les protéines nucléaires, mitochondriales, membranaires et cytosoliques) et ceci, à un moment défini (Patterson and Aebersold 2003). Contrairement au génome qui est figé, le protéome est dynamique. C'est ainsi qu'un même génome peut produire des protéomes différents selon les étapes du cycle cellulaire ou de la différenciation, la réponse à des signaux biologiques ou physiques (Lottspeich 1999). La variation de l'abondance de chacune des protéines composant le protéome est déterminée par des facteurs tels que la régulation de la transcription et de la traduction, ainsi que par le niveau de dégradation de ces mêmes protéines (Gygi, Rist et al. 1999), entraînant le comportement dynamique du protéome. Par ailleurs, différents mécanismes comme l'épissage alternatif des ARN messagers, la maturation des protéines par diverses protéases, ou l'ajout de modifications post-traductionnelles (PTMs), génèrent un niveau de complexité supplémentaire et une très grande « protéo-diversité ».

Dans le domaine de la santé, il est désormais admis que le dysfonctionnement, la dégradation, l'activation ou l'inactivation, le changement du niveau d'expression, ou du niveau de modification post-traductionnelle de l'une ou de plusieurs de ces protéines peut être responsable d'une pathologie. Ainsi, l'analyse du protéome ou analyse protéomique est devenu un outil de recherche incontournable pour comprendre l'état physiopathologique d'un système biologique. Aujourd'hui, les principales investigations de la recherche dans ce secteur sont :

- la détermination du catalogue protéique des échantillons analysés,
- l'analyse des interactions entre protéines et l'étude des réseaux protéiques,
- l'analyse des modifications post-traductionnelles des protéines,
- la quantification des protéines.

Le suivi du changement de l'abondance des protéines peut non seulement se faire au niveau cellulaire, mais aussi sur les compartiments sub-cellulaires et au cours du temps. Les perfectionnements récents des méthodes d'analyse protéomique ont ainsi permis de faire émerger le concept de « protéomique dans l'espace et dans le temps », désignant des approches permettant

à la fois de localiser les protéines au sein des divers compartiments de la cellule, de suivre leur flux dynamique entre ces compartiments, et de quantifier leurs variations d'expression ou de modifications post-traductionnelles dans différents contextes.

Il est également possible d'étudier le contenu protéique extra-cellulaire tel que celui présent dans des fluides biologiques. Cette caractérisation, qui ne peut être mise en œuvre par des approches de génomique ou de transcriptomique, apporte de nouvelles perspectives thérapeutiques. En effet, de nombreuses recherches s'orientent désormais vers la recherche de « biomarqueurs protéiques », c'est-à-dire de protéines dont la présence et/ou la quantité pourraient être corrélées à une pathologie et fournir ainsi un outil de diagnostic clinique (Petricoin, Paweletz et al. 2002).

Quelle que soit l'origine d'un échantillon protéique (culture cellulaire, biopsie, fluide biologique…) son analyse est complexe et requiert l'utilisation d'un ensemble de technologies de pointe. En effet, l'étude des protéines à grande échelle a été rendue possible par le développement simultané de différentes techniques allant de la préparation de l'échantillon (fractionnement, gel d'électrophorèse, chromatographie liquide,…) aux analyses bioinformatiques (banques de données génomiques et protéiques, conception de nouveaux algorithmes, amélioration de la puissance des ordinateurs…), tout en passant par la spectrométrie de masse (amélioration de la sensibilité, la résolution, la précision des mesures de masse, la vitesse de séquençage), qui constitue la technologie au cœur de ces approches.

La réussite de l'analyse protéomique réside cependant encore aujourd'hui dans la capacité à transformer la donnée acquise, ou spectres de masse, en information (identification, caractérisation et quantification des protéines), et l'information en connaissance (réponse à une question biologique). La très grande quantité et complexité des jeux de données générés par la spectrométrie de masse nécessite de disposer de programmes informatiques sophistiqués permettant d'extraire les informations pertinentes. Dans ce contexte, l'objectif principal de mon travail de thèse a été de développer de nouveaux algorithmes et outils logiciels capables d'interpréter et d'exploiter ce type de données.

Ce document s'attache à la fois à présenter les besoins qui sont à l'origine de ces développements, les implémentations logicielles qui ont été mises en œuvre ainsi que différents éléments de conception pour des solutions novatrices à venir. Avant de décrire ces résultats issus de mon travail de doctorat je commencerai par présenter les principes de l'analyse protéomique par spectrométrie de masse avec un accent particulier sur les stratégies d'identification et de quantification de mélanges complexes de protéines.

I.

L'analyse protéomique par spectrométrie de masse

L'évolution des techniques de séparation et d'analyse des protéines, ainsi que de la bioinformatique, permet aujourd'hui l'identification systématique de milliers de protéines en un temps réduit. Depuis le développement des sources d'ionisation dites « douces », le MALDI et l'électrospray, par Tanaka et Fenn au cours des années 1980 (http://nobelprize.org) (Fenn, Mann et al. 1989), la spectrométrie de masse a connu un essor considérable dans le domaine de la biologie (Patterson and Aebersold 2003). Associée (ou non) à la séparation des protéines sur gel d'électrophorèse mono- ou bidimensionnel (1D ou 2D), elle est devenue la méthode de choix pour l'identification à haut débit et la quantification des protéines (Angel, Aryal et al. 2012). Bien que cette technique permette d'analyser la masse précise d'une protéine (si celle-ci est d'une taille relativement faible) avec une grande sensibilité, elle ne rend pas possible à elle seule l'identification d'un mélange complexe de protéines. En fait, c'est le développement conjoint de quatre domaines qui est à l'origine de l'ascension de l'analyse protéomique par spectrométrie de masse : (1) le développement instrumental de la spectrométrie de masse (2) le développement de méthodes séparatives des protéines et peptides, (3) les projets de séquençage de génomes à haut débit qui alimentent les banques de séquence, ressources essentielles pour l'identification des protéines, et (4) le développement de logiciels dédiés à l'analyse des données produites.

I-1. Principe de la spectrométrie de masse

La spectrométrie de masse est une méthode analytique qui permet de déterminer la masse moléculaire précise d'un composé chimique ou biologique. Un spectromètre de masse est classiquement composé de trois grandes parties : la source d'ionisation, l'analyseur et le détecteur. Les molécules à analyser subissent une ionisation au niveau de la source. Il y a plusieurs techniques pour ce procédé, les plus utilisés étant les sources MALDI (Matrix Assisted Laser Desorption Ionization) et ESI (Electro Spray Ionization). Au niveau de l'analyseur, les ions sont séparés en fonction de leur rapport masse sur charge (m/z), qui leur confère un comportement différent au sein d'un champ électrostatique ou électromagnétique (Domon and Aebersold 2006). Il existe différents types d'analyseurs, et plusieurs géométries d'instruments associant différents analyseurs. Enfin le détecteur collecte ces ions, quantifie leur intensité et amplifie le signal. Ces dernières étapes se déroulent dans un vide poussé afin d'éviter toute collision entre les ions et les molécules de gaz. Après la détection, un système informatique effectue le traitement des données et génère un spectre de masse (ou « MS scan ») qui reflète la variation du courant ionique observé en fonction du rapport m/z et permet de déterminer la masse moléculaire des espèces ionisées.

I-2. Méthodes d'identification des protéines

I-2.1. Approches « top-down » et « bottom-up »

La spectrométrie de masse (MS) est très sensible, mais l'obtention d'ions moléculaires stables à partir de grosses biomolécules telles que les protéines est difficile. De plus, la masse d'une protéine, même si elle est mesurée avec une grande précision, n'est pas un critère suffisant pour identifier celle-ci avec certitude, et des données supplémentaires sur la séquence de la protéine sont nécessaires. Les appareils haute-résolution introduits ces dernières années permettent d'obtenir des informations de ce type, et ouvrent la voie à des approches dites « top-down », dans lesquelles l'analyse porte sur les protéines entières, qui sont ionisées et séquencées partiellement dans le spectromètre de masse. Néanmoins, cette stratégie reste encore marginale, et la plupart des données d'identification générées jusqu'à présent sont issues d'analyses dites « bottom-up », dans lesquelles les molécules directement mesurées ne sont pas les protéines entières, mais des fragments de taille inférieure, générés au préalable par digestion de la protéine à l'aide d'une protéase séquence-spécifique. En effet, ces peptides de petite taille sont ainsi plus facilement analysables par spectrométrie de masse que les protéines dont ils sont issus. Dans les approches « bottom-up », les informations de masse ou de séquence obtenues sur ces petits fragments permettent ensuite de remonter à l'identification de la protéine parente. Ce document s'attachera à décrire uniquement les méthodes et les logiciels sur lesquels s'appuient ces stratégies d'identification « bottom-up ».

I-2.2. Identification par cartes peptidiques massiques

L'enzyme la plus couramment utilisée pour générer des peptides protéolytiques est la trypsine. C'est une protéase stable, qui clive les protéines en peptides de manière très spécifique du côté C-terminal des résidus lysine et arginine (Olsen, Ong et al. 2004). Ainsi, si la séquence d'une protéine est connue, il est possible de prédire où l'enzyme va couper. Les récentes campagnes de séquençage du génome de divers organismes, comme l'humain, ont permis de définir les séquences en acides aminés de la majeure partie des protéines de ces organismes. Ces séquences sont regroupées dans de grandes bases de données telles que SwissProt (http://www.uniprot.org/). Connaissant la séquence de ces protéines, il est possible, à l'aide de programmes informatiques, de prédire la liste des masses attendues par une digestion à la trypsine *in silico* de chacune des protéines.

Initialement, l'identification des protéines par spectrométrie de masse a été basée sur la simple mesure des masses des peptides trypsiques issus d'une protéine. L'étape suivante consiste alors à rechercher dans la banque de données une protéine qui présenterait le même profil de peptides trypsiques que la protéine analysée par spectrométrie de masse, à l'aide de logiciels adaptés. Cette approche est appelée empreinte peptidique massique (« Peptide Mass Fingerprint », PMF) ou cartographie peptidique massique, car les masses peptidiques ainsi mesurées représentent des empreintes caractéristiques de la protéine analysée.

La principale limitation de cette méthode est que les protéines analysées doivent être au préalable isolées de façon relativement pure. En effet, une empreinte peptidique massique peut discriminer efficacement une protéine uniquement si elle n'est pas mélangée à d'autres empreintes composées par les peptides de protéines contaminantes. Typiquement, les identifications par cartes peptidiques massiques ont généralement été associées à des techniques séparatives d'électrophorèse en amont,

permettant de séparer les mélanges complexes et d'isoler les protéines d'intérêt dans un spot de gel 2D, ou dans une bande de gel 1D. Néanmoins, les échantillons obtenus de cette façon sont malgré tout souvent des mélanges contenant de nombreuses protéines minoritaires, et la PMF peut alors se révéler insuffisante pour caractériser les composants du mélange.

C'est la naissance d'une autre technologie, la spectrométrie de masse en tandem, qui a permis une identification des protéines plus spécifique que la PMF (Hunt, Yates et al. 1986).

I-2.3. La spectrométrie de masse en tandem : une révolution pour l'identification des protéines

Afin d'obtenir des informations sur la séquence des acides aminés qui composent un peptide il est possible, depuis l'apparition d'instruments spécifiques, de faire appel à la spectrométrie de masse en tandem (ou MS/MS). Celle-ci repose sur un processus de fragmentation d'une ou plusieurs liaisons de la molécule étudiée. Pour cela il faut transférer aux ions stables produits lors de l'ionisation au moins l'énergie supplémentaire nécessaire à leur fragmentation. La méthode la plus utilisée permettant ce transfert d'énergie est la dissociation induite par collision (CID pour « collision-induced dissociation »). D'une manière générale, dans un premier analyseur, un ion précurseur peut être sélectionné de façon spécifique en fonction de son rapport masse sur charge (m/z) et transféré dans une cellule de collision où il percute un flux d'atomes de gaz inerte (azote, hélium ou argon). Dès lors, son énergie cinétique est transformée en partie en énergie vibrationnelle nécessaire à sa fragmentation. Dans le cas d'un ion peptidique, cette fragmentation intervient principalement au niveau de la liaison amide entre les acides aminés, générant des séries de fragments N-terminaux et C-terminaux (appelés respectivement ions b et y dans la nomenclature conventionnelle), qui seront ensuite analysés par un second analyseur, d'où le terme de spectrométrie de masse en tandem (cf figure 1).

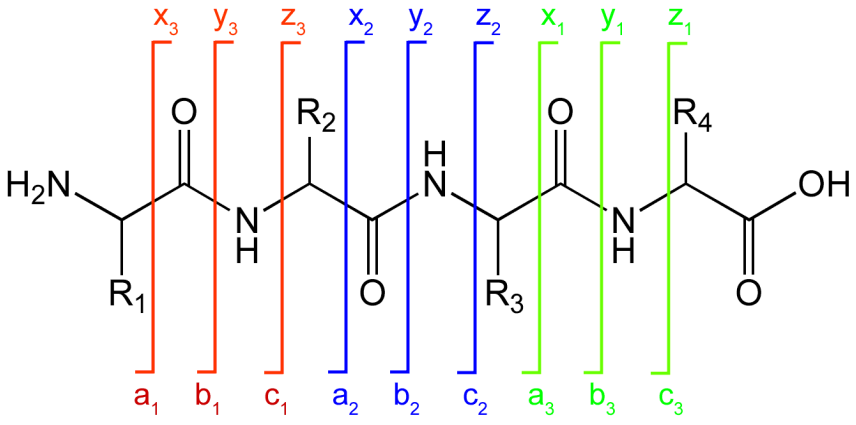


Figure 1 : nomenclature des différentes séries d'ions peptidiques qui peuvent être générées lors de la fragmentation MS/MS. Proposée initialement par Roepstorff et Fohlman (Roepstorff and Fohlman 1984), elle fut modifiée par Johnson *et al.* (Johnson, Martin et al. 1987) avant d'être définitivement adoptée. La molécule représentée en noir correspond à un peptide théorique comportant quatre résidus (la lettre R représentant la chaîne latérale). Les fragments ne seront détectés que s'ils portent au moins une charge. Si la charge est retenue du côté N-terminal alors l'ion est dit de série a, b ou c. Si la charge est retenue du côté C terminal alors l'ion est dit de série x, y ou z. Dans le cas d'une fragmentation CID on observe majoritairement des séries b et y.

Le balayage des rapports m/z des ions fragments par le détecteur génère un spectre de masse de second niveau, appelé spectre MS/MS (ou « MS2 scan »). C'est l'analyse de ce dernier qui permet de remonter à la structure de l'ion parent. En plus de la mesure de la masse des peptides trypsiques d'une protéine, la MS/MS apporte aussi des données de séquence sur ces peptides, qui sont ensuite utilisées par des logiciels dédiés afin de réaliser les recherches en banques de données protéiques. Cette approche génère donc des données plus spécifiques, et permet d'identifier les protéines de façon plus discriminante que la PMF. De plus, elle se prête bien à l'analyse de mélanges de protéines. Celles-ci n'ont pas besoin d'être isolées une à une, puisque le séquençage MS/MS sur un peptide individuel est possible même quand d'autres peptides issus de protéines différentes sont présents dans l'échantillon. Ainsi, dans le cas d'un mélange, l'échantillon peut être directement digéré à la trypsine, sans séparation préalable des protéines, et les différents peptides sont directement mesurés par le spectromètre de masse. Celui-ci détecte et sélectionne individuellement les peptides, et les séquence automatiquement. Des logiciels adaptés identifient par la suite les protéines du mélange grâce à l'ensemble des spectres MS/MS enregistrés. Par ailleurs, la multiplication des informations sur différents segments de la protéine permet non seulement de conforter l'identification, mais également d'obtenir des informations structurales en particulier sur les modifications post-traductionnelles, telles que l'addition de groupements phosphates (phosphorylation) ou de chaînes d'oligosaccharides (glycosylation). D'un point de vue fonctionnel, ces informations sont fondamentales. Les phosphorylations sont à la base de la conduction de signaux au sein de la cellule. Quant aux chaînes oligosaccharidiques, elles jouent un rôle crucial dans la modulation des propriétés chimiques de certaines protéines (glycoprotéines) et gouvernent parfois leur localisation cellulaire ou leur activité biologique.

I-3. L'analyse de mélanges protéiques complexes

I-3.1. Gamme dynamique d'analyse et fractionnement des échantillons complexes

Les échantillons biologiques sont généralement constitués de plusieurs milliers de formes protéiques différentes. L'étape de digestion enzymatique, nécessaire à l'analyse par spectrométrie de masse, augmente la complexité intrinsèque du mélange car elle génère un nombre encore plus important de molécules différentes. L'analyse simultanée de tous ces peptides excède donc en général les capacités des spectromètres de masse en termes de résolution en masse et de gamme dynamique. En effet, il est important de noter que les concentrations des différentes protéines présentes dans un protéome peuvent varier de façon extrêmement importante. De plus, les spectromètres de masse, même s'ils sont très sensibles, ne peuvent analyser des composés que sur une gamme de concentration limitée. Typiquement, un instrument récent tel que le LTQ-Orbitrap Velos présente une gamme dynamique d'environ 3 ou 4 ordres de grandeur. Lors d'une analyse par spectrométrie de masse d'un mélange trop complexe, les peptides issus des protéines très abondantes du mélange masquent le signal issu des espèces mineures, empêchant la détection de nombreuses protéines de faible abondance.

De manière à rendre l'analyse d'un échantillon biologique complexe réalisable par spectrométrie de masse, il est donc nécessaire d'utiliser en amont des stratégies de simplification du mélange, qui peuvent intervenir à différents stades de la préparation des échantillons :
- lors de la préparation de l'échantillon protéique, par fractionnement subcellulaire ou purification d'un sous-protéome d'intérêt.
- avant la digestion des protéines, en utilisant des techniques séparatives telles que les gels d'électrophorèse 1D ou 2D ou des chromatographies liquides permettant de fractionner le mélange de protéines.
- après la digestion des protéines, en utilisant des techniques séparatives dédiées au fractionnement du mélange de peptides. Typiquement, ce fractionnement est très souvent réalisé « en ligne » avec le spectromètre de masse, en couplant celui-ci à une chaîne HPLC à nano débit (ou nanoLC).

I-3.2. Principe de l'analyse nanoLC-MS/MS

La séparation d'une molécule par chromatographie repose sur l'utilisation de la différence d'affinité de celle-ci entre une phase stationnaire et une phase mobile. Il existe différentes techniques de chromatographie liquide en fonction de la phase stationnaire utilisée, chacune de ces phases stationnaires permettant de séparer les analytes en fonction de paramètres physico-chimiques particuliers. La chromatographie liquide (LC) en phase inverse, utilisée pour le couplage avec le spectromètre de masse pour l'analyse de peptides, est capable de séparer les molécules en fonction de leur hydrophobicité et de leur taille. Il est possible de modifier les caractéristiques de la phase mobile afin d'agir sur l'efficacité de la séparation. Si la composition du solvant reste identique pendant toute la durée de l'élution on parle alors de mode isocratique. Par opposition, un gradient d'élution met en jeu une variation continue de l'hydrophobicité de l'éluant lors de la séparation. Il suffit pour cela de modifier au cours du temps les proportions d'un mélange de deux solvants différents. Ce mode d'élution est utilisé pour les analyses de mélanges complexes car il sépare de façon satisfaisante les peptides. Au cours du gradient chromatographique, ces-derniers sont retenus dans la colonne pendant un certain temps. Le moment où un peptide spécifique est élué de la colonne, est appelé le temps de rétention (RT), ou temps d'élution. En sortie de la chromatographie nanoLC, les peptides élués sont ionisés dans la source électrospray, et injectés directement dans le spectromètre de masse, qui les détecte, les sélectionne individuellement, et les séquence en continu, alternant les analyses MS et MS/MS.

Le couplage nanoLC-MS/MS a deux effets positifs sur l'analyse par spectrométrie de masse des mélanges peptidiques. Tout d'abord, des peptides dits isobariques dont les signaux se seraient superposés lors d'une simple analyse ont une chance d'éluer à des temps différents et donc d'être discriminés. Deuxièmement, le nombre d'ions analysés simultanément par le spectromètre de masse est réduit. Ainsi, l'effet de suppression d'ion, où le signal d'un ion majoritaire supprime celui d'un autre ion minoritaire, est diminué (Annesley 2003), ce qui augmente la gamme dynamique observable.

I-3.3. Les résultats de l'analyse nanoLC-MS/MS : cartes LC-MS et données de séquençage MS/MS

Cette approche de couplage en ligne nanoLC-MS/MS introduit donc une dimension temporelle et dynamique dans l'analyse du mélange de départ. Ce mélange n'est pas caractérisé par un spectre de masse unique, mais par toute une série de scans MS et MS/MS enregistrés au cours du temps, tout au long du gradient chromatographique, au fur et à mesure de l'élution des peptides. Au moment où un peptide est élué de la colonne chromatographique, il est présent pendant un court intervalle de temps au niveau de la source électrospray, et son signal prend la forme d'un pic d'élution pseudo-gaussien, dont la largeur est variable en fonction de l'abondance du peptide (typiquement quelques dizaines de secondes). Il est important de souligner que la digestion d'un mélange complexe génère un grand nombre de peptides à analyser. En général, tous les peptides présents ne sont pas systématiquement fragmentés par l'instrument. En effet, de nombreux peptides sont élués simultanément, et seuls les ions les plus intenses sont alors sélectionnés pour la fragmentation (cf figure 2).

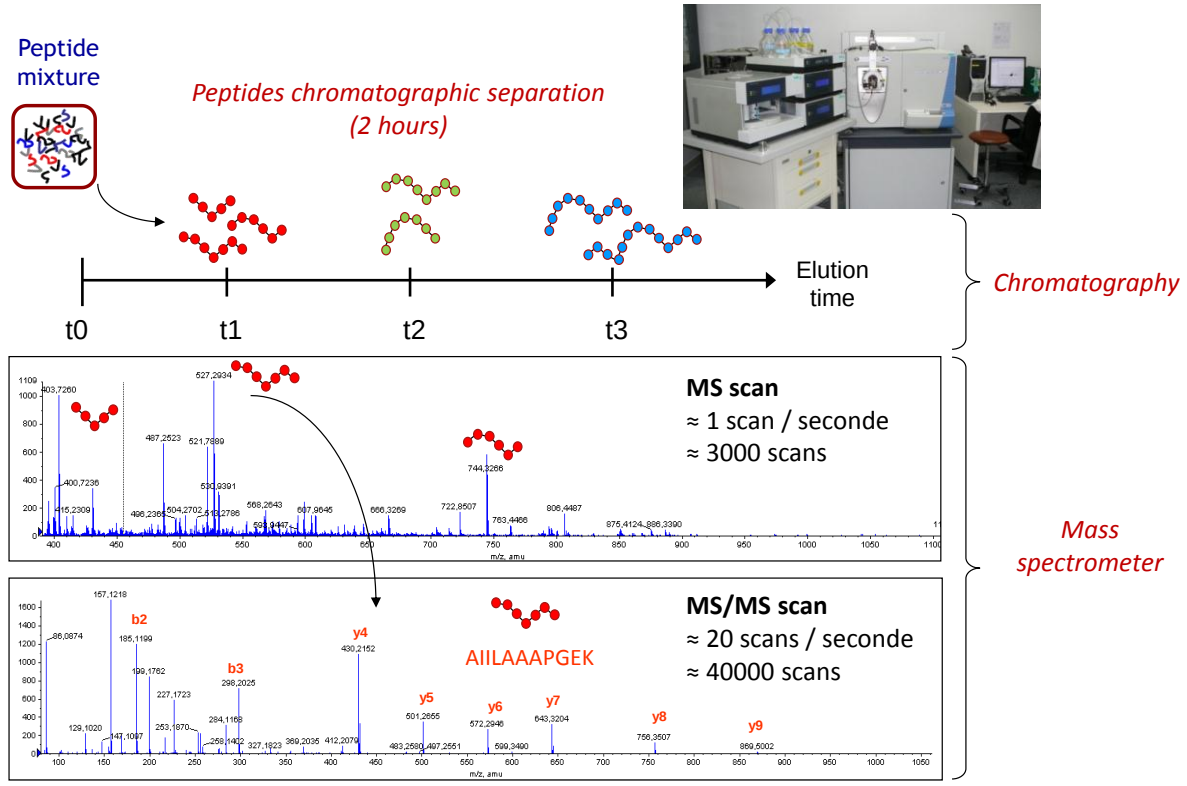


Figure 2 : vue d'ensemble de l'approche nanoLC-MS/MS. Un spectre MS représente à un instant donné l'empreinte massique de l'ensemble des peptides qui ont été détectés. Certains de ces peptides sont parallèlement sélectionnés pour la fragmentation. Ainsi, toute une série consécutive de spectres MS/MS, chacun étant étiqueté avec le rapport m/z précis du peptide correspondant et un temps de rétention unique, est automatiquement acquise. Les vitesses d'acquisition (nombre de scans par seconde) correspondent à un instrument de type LTQ-Orbitrap Velos.

L'analyse d'un mélange peptidique recouvre l'ensemble des scans enregistrés sur la période du gradient chromatographique, qui dure généralement de une à deux heures. La compilation de l'ensemble des scans MS enregistrés sur une gamme de m/z donnée, peut être visualisée sous la forme une carte en deux dimensions (cf figure 3).

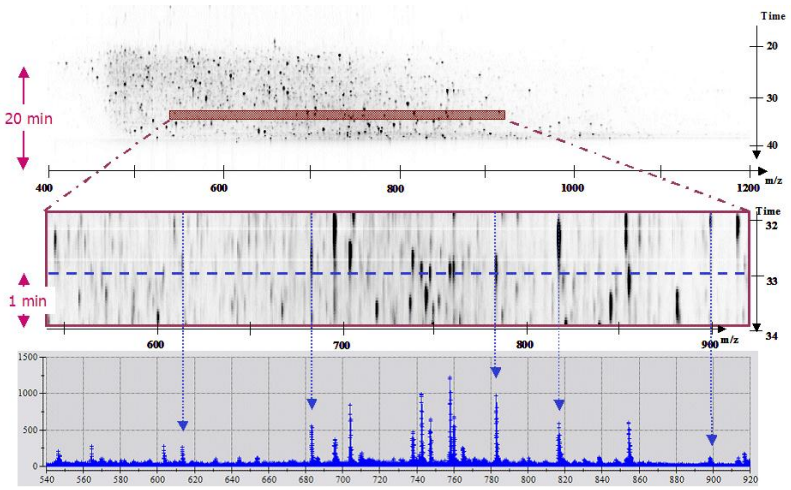


Figure 3 : le premier graphe représente une carte LC-MS en deux dimensions (t, m/z) reflétant l'ensemble des espèces peptidiques détectées tout au long de l'analyse. Le second graphe est un agrandissement d'une partie de la carte et le trait en pointillé correspond à un des spectres MS à l'origine de cette carte. Celui-ci est affiché en bleu dans le dernier panel.
Source : http://web.expasy.org/MSight/navig1.html

Le passage du mode MS au mode MS/MS est automatiquement réalisé par le logiciel de pilotage du spectromètre de masse, en fonction de critères prédéfinis au départ par l'utilisateur dans la méthode d'acquisition, et grâce à une analyse en temps réel des signaux détectés dans un scan MS à l'instant t. Par exemple, le spectromètre peut être configuré de façon à déclencher une analyse MS/MS sur les dix ions les plus intenses du scan MS en cours, s'ils dépassent un seuil de signal fixé. Ces ions seront alors successivement sélectionnés, isolés et fragmentés en MS/MS (cf figure 4). Une fois ce cycle de MS/MS (d'une durée Δt) terminé, un nouveau scan MS est enregistré afin de détecter à $t + \Delta t$ les peptides en cours d'élution de la colonne, et recommencer un cycle de MS/MS sur les ions les plus intenses. Des paramètres d'exclusion dynamique peuvent être ajustés afin d'éviter la sélection et l'acquisition répétée de multiples scans MS/MS sur le même ion. Ce mode d'acquisition des données est appelé « DDA » pour « data dependent acquisition ».

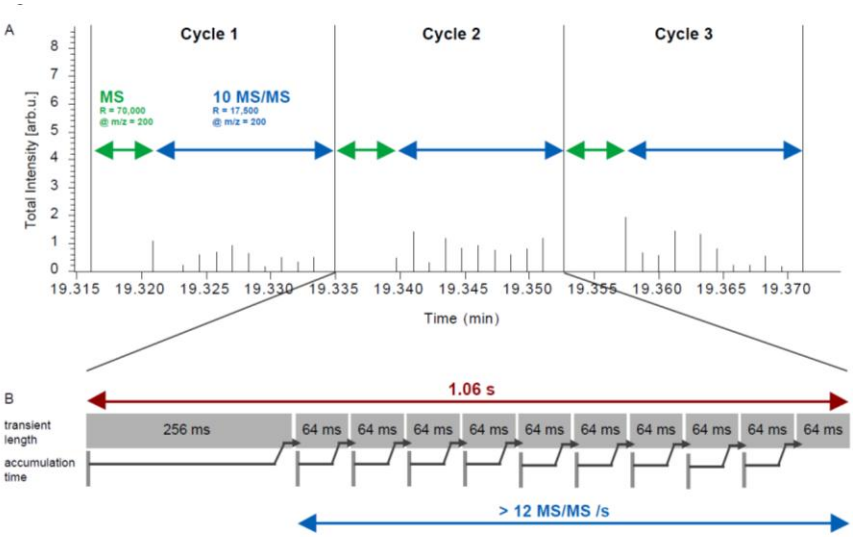


Figure 4 : représentation graphique d'une méthode d'acquisition alternant MS et MS/MS. La durée de chacune des phases de balayage peut être paramétrée par l'utilisateur via le logiciel de pilotage. Les vitesses d'acquisition correspondent à celles d'un instrument de type LTQ-Orbitrap XL.

I-3.4. Vitesse de séquençage et échantillonnage peptidique

La vitesse de séquençage MS/MS de l'instrument est un paramètre important de l'analyse par couplage nanoLC-MS/MS. Pour pouvoir caractériser au maximum un mélange complexe par cette approche, il faut que le spectromètre de masse séquence le plus de peptides possible. En effet, nous venons de voir que seule une partie du mélange peptidique détecté lors de l'analyse MS était finalement sélectionnée par le spectromètre de masse pour l'étape de fragmentation. On observe en pratique que ce sous-échantillonnage est plus ou moins stochastique, car lors de l'analyse de deux injections consécutives d'un même échantillon, on n'obtient jamais deux listes de spectres MS/MS strictement identiques. Ce problème est principalement lié au fait que la complexité des mélanges analysés excède la capacité des instruments, même les plus performants, en terme de vitesse de séquençage.

Idéalement, pour pouvoir caractériser au maximum un mélange complexe, il faut éviter qu'un nombre trop important de peptides ne soient élués simultanément de la nanoLC à un instant donné. Pour cela, il est possible de simplifier le mélange de protéines de départ, comme indiqué précédemment, ou de fractionner encore davantage le mélange de peptides, par exemple par une double chromatographie phase échangeuse d'ion/phase inverse (approche MudPIT (Washburn, Wolters et al. 2001), pour Multidimensional chromatography Peptide Identification Technology). L'utilisation de spectromètres de masse à très grande vitesse de séquençage permet également d'optimiser la couverture en spectres MS/MS des mélanges peptidiques complexes. Les instruments de dernière génération effectuent par exemple plusieurs dizaines de milliers de scans MS/MS lors d'un gradient chromatographique d'une heure sur un mélange peptidique complexe.

Le couplage nanoLC-MS/MS est désormais classiquement utilisé pour l'analyse de mélanges peptidiques, issus de la digestion de mélanges protéiques, et permet aujourd'hui d'analyser un grand nombre de peptides à partir d'un échantillon biologique dans un mode automatisé et à haut débit. Une telle approche automatisée génère une grande quantité de données MS/MS. C'est l'interprétation de chacun de ces spectres qui rend possible l'identification de l'ensemble des peptides présents dans l'échantillon. Il est cependant impossible de réaliser une telle tâche manuellement étant donné la quantité de données générées par ce type. Il a ainsi été développé durant les dix dernières années un ensemble de méthodes de calcul permettant d'automatiser l'identification de peptides et de protéines à partir des spectres MS et MS/MS. Ces méthodes, qui seront décrites dans la partie suivante, ont été implémentées dans des logiciels spécialisés appelés moteurs de recherche.

I-4. L'interprétation automatique des données de spectrométrie de masse

Durant la dernière décennie, les spectromètres de masses en tandem ont été continuellement améliorés sur les plans de la sensibilité, de la précision, de la résolution et de la vitesse de séquençage. L'ajout de nouvelles fonctions sophistiquées pour le contrôle de l'acquisition des spectres telle que l'exclusion dynamique a également contribué à l'amélioration des appareils. Ces derniers sont désormais capables de générer plus de 30000 spectres MS/MS par analyse. Au départ destinés à l'analyse de simples spectres MS (dans les approches par PMF), les moteurs de recherche

dans les banques de séquences protéiques ont également été au centre de développements actifs afin de répondre à cette montée en puissance de la spectrométrie de masse en tandem.

I-4.1. Les moteurs de recherche

Comme nous l'avons évoqué plus haut, il existe deux modes d'identification des protéines par recherche en banques de données : la cartographie peptidique ou « peptide mass fingerprinting » (PMF), et le séquençage peptidique par MS/MS. En 1993, cinq groupes différents (Henzel, Billeci et al. 1993; James, Quadroni et al. 1993; Mann, Hojrup et al. 1993; Pappin, Hojrup et al. 1993; Yates, Speicher et al. 1993) développent, à l'aide des connaissances sur la spécificité de clivage de l'enzyme trypsine, et d'une base de données de séquences protéiques, des algorithmes dédiés à l'analyse par empreinte peptidique massique. Il devient ainsi possible, à partir de données de spectrométrie de masse, en comparant l'ensemble des masses mesurées à l'ensemble des masses théoriques provenant de la digestion virtuelle (*in silico*) des protéines répertoriées dans les banques de données, d'identifier des protéines de manière univoque.

Depuis leur création, ces algorithmes ont évolués, certains d'entre eux prenant en charge les données de fragmentation MS/MS, et ont été implémentés dans de nombreux programmes appelés moteurs de recherche (Henzel, Watanabe et al. 2003), comme par exemple :
* Andromeda : http://maxquant.org/ (Cox and Mann 2008)
* Mascot : http://www.matrixscience.com/search_form_select.html (Perkins, Pappin et al. 1999)
* MOWSE : http://www.matrixscience.com/help/scoring_help.html (Pappin, Hojrup et al. 1993)
* MS-fit : http://prospector.ucsf.edu/ (Clauser, Baker et al. 1999)
* OMSSA : http://pubchem.ncbi.nlm.nih.gov/omssa/ (Geer, Markey et al. 2004)
* Phenyx : http://www.genebio.com/products/phenyx/index.html (Colinge, Masselot et al. 2003)
* Sequest : http://fields.scripps.edu/sequest/index.html (Eng, McCormack et al. 1994)
* X!Tandem : http://www.thegpm.org/tandem/ (Craig and Beavis 2004)

Ainsi, un grand nombre d'algorithmes et de programmes informatiques ont été développés pour réaliser l'identification de peptides et de protéines à partir de données MS et MS/MS. Il y a de nombreux aspects très intéressants associés à ces développements que nous pourrions évoquer mais ils vont au-delà de la portée de cette thèse. Je me contenterai donc de résumer les principes majeurs en prenant comme exemple le moteur de recherche Mascot qui a été celui principalement utilisé au cours de mon doctorat.

I-4.2. MOWSE : une première implémentation de l'analyse PMF

L'algorithme utilisé par Mascot (MOWSE) a été mis au point en 1993 par David Perkins et Darryl Pappin (Pappin, Hojrup et al. 1993). Il a été un des premiers programmes d'identification pour les analyses de PMF. Son acronyme signifie Molecular Weight Search (recherche par poids moléculaire).
La première étape de la recherche consiste à réaliser une digestion théorique des séquences de la banque protéique. Ces dernières sont digérées selon des règles de coupures identiques à celles réalisées par les enzymes utilisées pour la digestion de l'échantillon. Ensuite le moteur de recherche compare les masses peptidiques calculées pour chaque entrée protéique avec les masses peptidiques expérimentales. Chaque valeur calculée qui est suffisamment proche d'une valeur expérimentale, en tenant compte d'une certaine tolérance de masse fixée par l'utilisateur, est considérée comme une indentification possible. Plutôt que de ne compter que le nombre de séquences attribuées à une

protéine, Mowse utilise des facteurs pour pondérer de façon statistique chaque identification peptidique individuelle. La matrice de ces facteurs est calculée lors de l'indexation de la banque via un algorithme empirique. Ce dernier est basé sur la construction d'une matrice de distribution de peptides, qui se déroule en plusieurs étapes (Figure 5a). David Perkins a mis en évidence que la fréquence de présence d'une séquence peptidique dans la banque est fonction de la taille de cette séquence et de la taille de la protéine dont il provient. Ces fréquences servent à générer une matrice 3D de distribution des peptides (Figure 5b). Les tailles des peptides issus d'une protéine ne sont pas aléatoires. Elles dépendent de la position des sites de coupures (en C terminal de la lysine et de l'arginine pour la trypsine). Il est ainsi nécessaire de prendre compte la répartition non-uniforme des tailles de peptide qui résulte de la digestion par une enzyme.

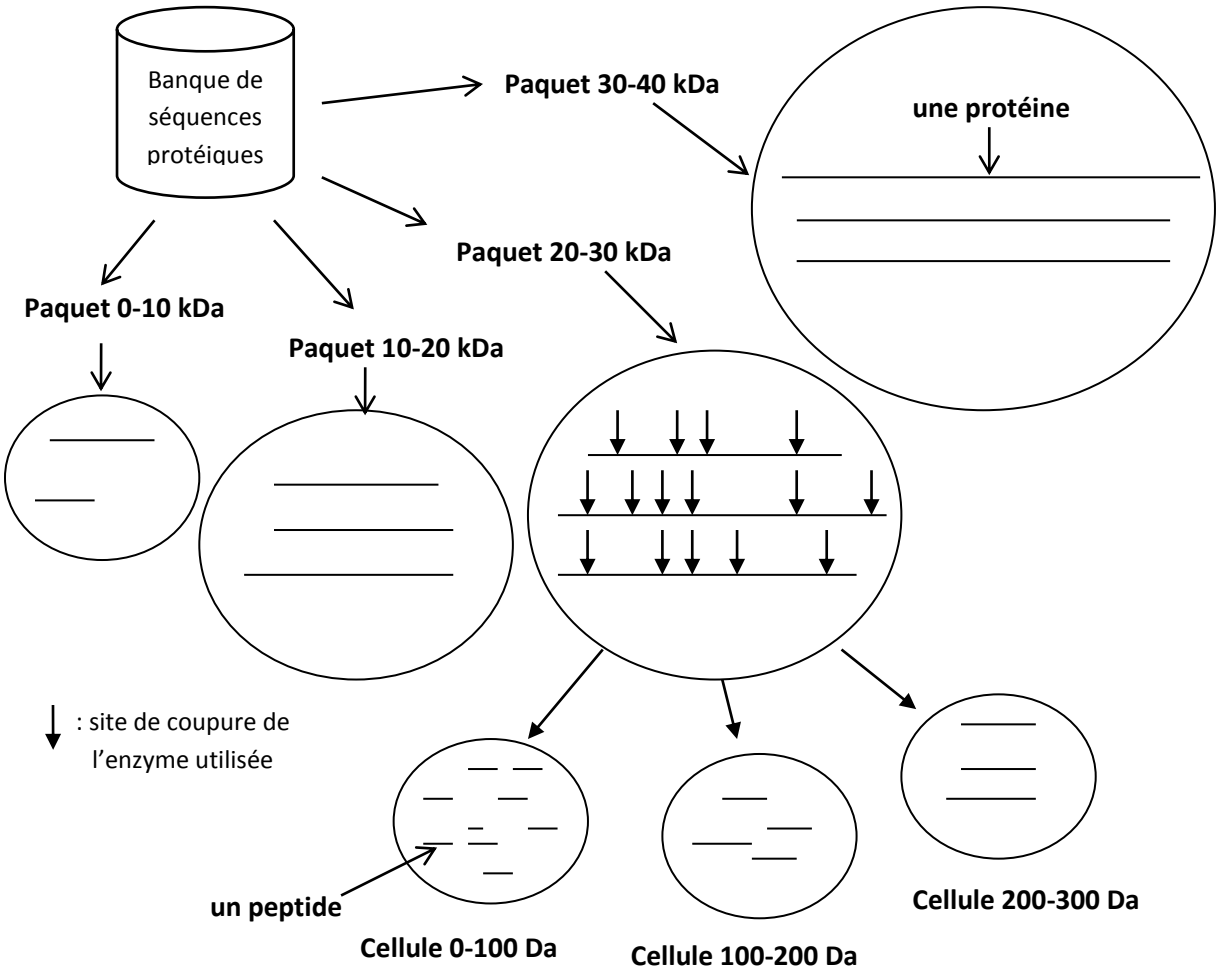


Figure 5a : constitution d'une matrice de distribution par regroupement de peptides. On part d'une liste de protéines contenues dans des bases de données. Ces protéines ont des tailles différentes. On peut les regrouper par paquets d'intervalle égal à 10 kDa. Les séquences peptidiques théoriques obtenues pour chaque paquet de protéines sont ensuite classées par cellules d'intervalle de masse égal à 100 Da. Suite à ce tri, l'algorithme MOWSE calcule la fréquence de présence d'un peptide au sein d'un paquet. On peut représenter ces fréquences dans une matrice de fréquence F où chaque ligne correspond aux peptides regroupés par intervalles de 100 Da, et chaque colonne aux protéines regroupées par intervalle de 10 kDa. Lors du parcours de la base de données chaque cellule Fi,j de la matrice est incrémenté pour accumuler la statistique de la distribution des masses peptidiques en fonction de celles de protéines. Ces fréquences sont ensuite normalisées en divisant les éléments de chaque colonne par la valeur la plus grande qui existe au sein de cette même colonne.

Pour identifier des protéines on a donc besoin de peptides qui apparaissent peu fréquemment dans les protéines. Si l'on attribue une masse à un peptide retrouvé fréquemment (peptide de petite taille), il y a de grandes chances pour que cette attribution soit due au hasard.

Une fois les fréquences peptidiques déterminées, le moteur de recherche regroupe les peptides qui mènent à l'identification d'une même protéine. C'est à partir du produit des fréquences d'apparition de l'ensemble des peptides d'une même protéine que MOWSE calcule le score protéique. Plus le produit des fréquences est faible et plus le score est élevé et par conséquent plus on a de chances que la protéine soit présente dans le mélange analysé. Le score prend également en compte le poids moléculaire de la protéine concernée, donnant ainsi lieu à la formule suivante :

$$S_{MOWSE} = \frac{50000}{MW_{prot} \times \prod_n f_{pep}}$$

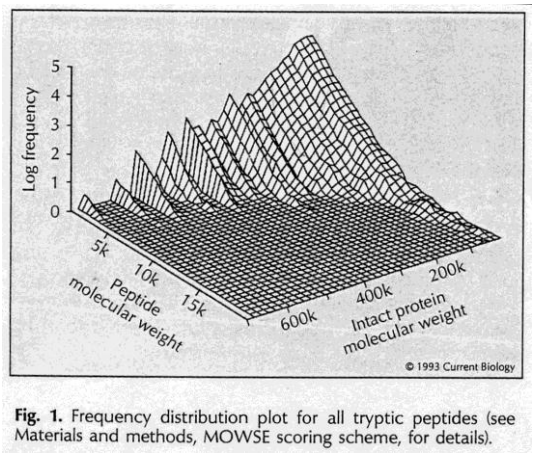


Fig. 1. Frequency distribution plot for all tryptic peptides (see Materials and methods, MOWSE scoring scheme, for details).

Figure 5b : graphique 3D de la distribution des fréquences pour tous les peptides trypsiques théoriques issus de la banque de séquences protéiques OWL (Bleasby, Akrigg et al. 1994).
Il s'agit d'une façon de visualiser la matrice des fréquences des peptides dont la construction a été présentée dans la figure 5a.
Source : (Pappin, Hojrup et al. 1993)

Ce moteur de recherche a ensuite été amélioré pour prendre en compte la séquence et la composition en acides aminés des peptides. Mais l'évolution majeure de MOWSE a été l'intégration des données de fragmentations peptidiques à partir de spectres MS/MS. En effet, cette deuxième approche d'identification permet d'obtenir des informations de séquence, ce qui augmente considérablement la spécificité et la qualité des identifications. Plusieurs algorithmes sont utilisés pour la recherche dans les banques de données à partir des spectres MS/MS (Patterson and Aebersold 2003; Steen and Mann 2004).

I-4.3. Interprétation de spectres MS/MS par recherche dans des banques de données

Un spectre MS/MS représente une empreinte caractéristique de la séquence du peptide sélectionné pour la fragmentation. Il est associé à deux types d'informations qui vont être utilisées pour identifier le peptide :
- le rapport m/z et la charge de l'ion précurseur qui a été sélectionné et fragmenté pour donner naissance à ce spectre, dont on déduit la masse du peptide entier.
- deux séries de valeurs (m/z et intensité) correspondant aux différents fragments du peptide. Ces données, éventuellement répétées pour plusieurs peptides, constituent la « peaklist » (ou liste de masses) qui peut être générée à partir des données brutes par des logiciels de détection de pics.

Il existe un grand nombre d'outils logiciels disponibles pour réaliser l'identification de peptides à partir de spectres MS/MS en utilisant différentes approches algorithmiques (Nesvizhskii 2010). Ces dernières se basent soit sur une interprétation totale ou partielle de ces spectres, soit sur l'utilisation directe des «peaklists » brutes, sans interprétation (cf figure 6).

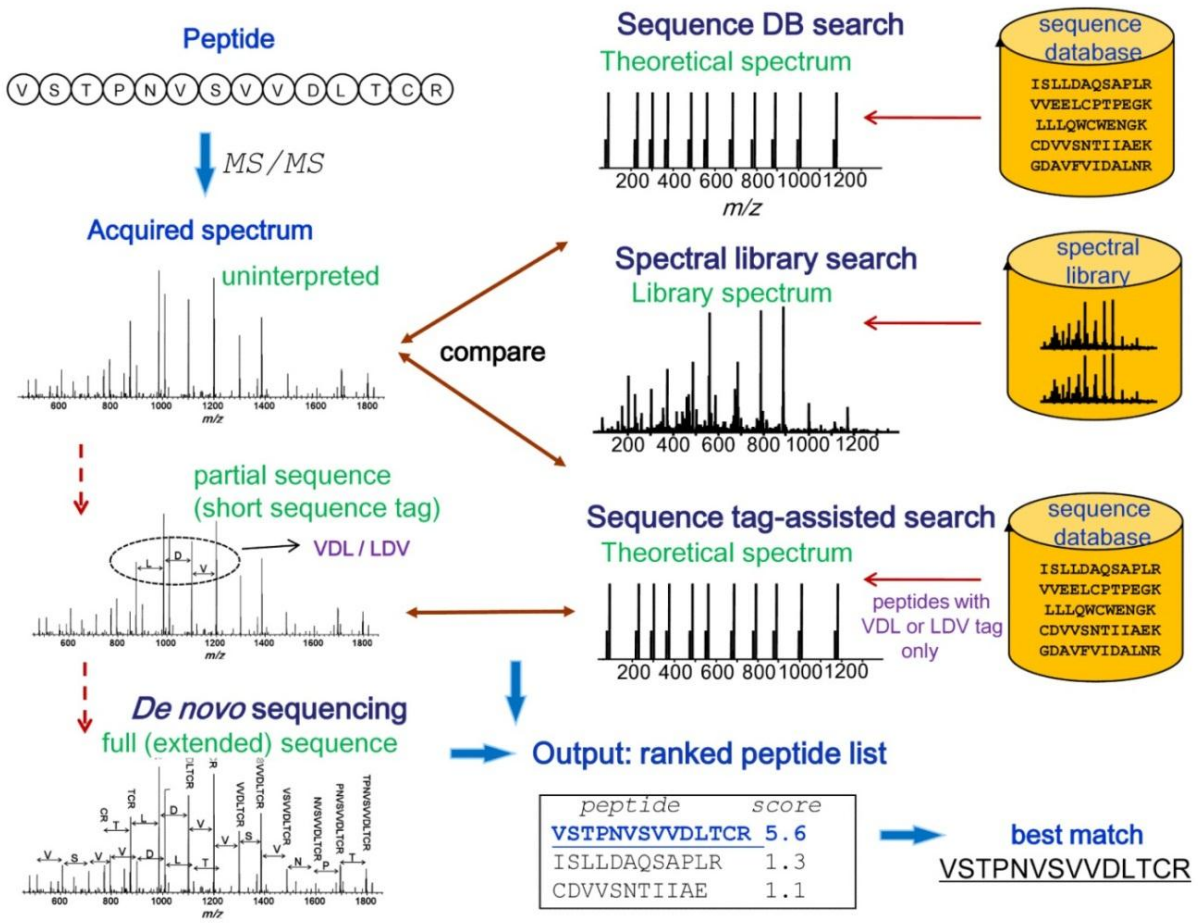


Figure 6 : aperçu général des différentes approches utilisées pour l'identification des protéines à partir de spectres MS/MS. L'approche la plus couramment utilisée est la recherche en banque de données protéiques. Lorsque des spectres expérimentaux ont été identifiés et validés dans des analyses antérieures il est aujourd'hui possible de les employer pour identifier les peptides correspondants dans de nouvelles analyses : cette stratégie est appelée « spectral library search ». Lorsque ces deux approches sont inefficaces il est également possible de mettre en œuvre des analyses *de novo* (Seidler, Zinn et al. 2010) complètes ou partielles qui consistent à déterminer la séquence du peptide en interprétant directement les données spectrales. Les informations partielles de séquences sont appelées « short sequence tag » (Mann and Wilm 1994) et sont utilisées pour effectuer une recherche dans les banques de données protéiques. Ces stratégies *de novo* fonctionnent bien si les spectres MS/MS analysés sont de bonne qualité mais fournissent dans le cas contraire des résultats ambigus ou inexploitables.
Source : (Nesvizhskii 2010)

Dans le premier cas, des logiciels sont utilisés pour interpréter les spectres MS/MS, et en extraire la séquence peptidique (séquençage de novo), ou une partie de cette séquence (« sequence tag »), en calculant les différences de masse entre les fragments mesurés. Les outils de séquençage de novo (Lutefisk, PepNovo, Peaks) identifient la séquence complète du peptide à partir du spectre MS/MS et

sont particulièrement utiles quand les protéines de départ ne sont pas présentes dans les bases de données, dans le cas notamment d'organismes non séquencés. Associés à des outils comme MS-BLAST, ils permettent de déduire l'identité des protéines analysées par comparaison avec celles d'autres organismes séquencés apparentés. Néanmoins ils demandent des spectres MS/MS de bonne qualité, un temps de calcul important et sont peu utilisés dans les analyses haut-débit portant sur des organismes bien caractérisés. Une approche intermédiaire consiste à définir uniquement à partir du spectre MS/MS des petits bouts courts de séquences peptidiques, encadrés par deux masses. Des moteurs de recherche dédiés peuvent alors utiliser ces « peptide sequence tags » pour identifier les peptides dans les bases de données. Etant donné que je n'ai pas eu l'opportunité de mettre en œuvre de telles analyses au cours de mon doctorat je ne détaillerai pas d'avantage ces différentes approches.

Dans le second cas, les moteurs de recherche effectuent directement une comparaison entre les spectres MS/MS expérimentaux bruts (la « peaklist ») et des spectres MS/MS stockés dans une bibliothèque spectrale ou bien générés *in-silico* à partir de banque de séquences protéiques. La recherche en bibliothèque spectrale (« spectral library ») est une méthode récente bien plus performante que la recherche en banque de données sur le plan de la vitesse, de la sensibilité et du taux d'erreur (Lam, Deutsch et al. 2007). Un inconvénient de ce type d'approche est qu'elle nécessite de disposer d'un enregistrement spectral correspondant à chaque peptide de l'échantillon analysé : seul les peptides identifiés au préalable peuvent donc être mis en évidence. Etant donné que ces bibliothèques sont aujourd'hui clairement incomplètes et que les logiciels ne sont pas suffisamment matures, cette approche n'est pas encore utilisée en routine dans les laboratoires. Ces derniers ont principalement recours à des moteurs de recherche pour interroger des bases de données de séquences protéiques. L'algorithme d'identification est une extension de celui utilisé pour l'analyse PMF : la première étape est identique car elle consiste en la digestion trypsique théorique de toutes les protéines présentes dans la banque interrogée. De façon analogue à l'algorithme PMF, le programme recherche dans toutes les séquences peptidiques obtenues (suite à la digestion *in silico* des protéines) celles qui correspondent au poids moléculaire des ions précurseurs fragmentés, avec une certaine tolérance d'erreur de masse (Steen and Mann 2004). Cette opération restreint seulement l'espace de recherche à un petit nombre de candidats, dépendant de l'erreur de mesure, mais ne constitue qu'une étape préliminaire. En effet, c'est l'information obtenue par l'analyse des fragments MS/MS qui permet de trouver la séquence peptidique la plus « juste ». Etant donné que le processus de fragmentation suit des règles bien particulières, l'algorithme est capable, pour chaque peptide candidat, de construire un modèle représentant la liste de fragments MS/MS théoriques. Ces derniers sont comparés avec les données expérimentales : plus la série d'ions fragments théoriques se rapproche de la série expérimentale, plus l'identification est correcte et donc plus le score du peptide est élevé. En général, la séquence du peptide correspondant à la meilleure superposition (meilleur score) est attribuée à l'ion précurseur analysé.

La principale différence entre les algorithmes existants réside dans leur fonction de corrélation (décrite pour Sequest dans (Eng, McCormack et al. 1994)), qui est utilisée pour déterminer la similarité entre un spectre MS/MS théorique et expérimental et donc la vraisemblance de l'identification. Tandis que certains outils attribuent uniquement un score de similarité ou de corrélation au spectre MS/MS, d'autres, comme Mascot, fournissent également une valeur d'espérance ou « expectation value » (E-value) (Perkins, Pappin et al. 1999) qui correspond au

nombre de peptides qui auraient par hasard un score supérieur ou égal à celui observé. En dehors du cas spécifique où cette valeur est calculée à partir des scores et des seuils probabilistiques définis dans Mascot (voir ci-dessous), le calcul de la E-value est effectué en modélisant la répartition des scores de l'ensemble des candidats peptidiques potentiels pour un spectre donné. En général ces candidats sont classés par ordre décroissant de score et le premier de la lise (« top ranking peptide match ») est retenu comme la meilleure identification du spectre. Les autres candidats peuvent être utilisés pour calculer la E-value de la meilleure identification en déterminant l'histogramme de répartition de leur score (dite distribution nulle ou H0) que l'on considère comme aléatoire. Plus le score du meilleur candidat est éloigné du cœur de cette distribution et plus il est considéré comme statistiquement correct, même si la valeur absolue de ce score n'est pas en soi très élevée. Plusieurs approches existent pour calculer cette valeur de significativité du score. Elles se distinguent par le modèle mathématique (loi de Poisson, loi hypergéométrique…) utilisé pour évaluer la distribution (Sadygov and Yates 2003; Geer, Markey et al. 2004) des fausses identifications ainsi que par la méthode, théorique ou empirique, employée pour calculer l'aire sous la queue de la distribution pour un score supérieur ou égal à celui de la meilleure identification. Cette valeur de probabilité nous renseigne sur la pertinence du meilleur score obtenu pour un spectre MS/MS donné par rapport à l'ensemble des scores des autres candidats de la banque de données. Elle est donc préférable à la valeur d'un score de similarité pour lequel un seuil arbitraire doit être défini afin de discriminer les bonnes des mauvaises identifications peptidiques.

I-4.4. Mascot : implémentation d'un modèle de score probabilistique

Au cours de mon doctorat j'ai principalement utilisé le moteur de recherche Mascot, qui constitue une référence dans le domaine de la protéomique. L'algorithme de Mascot est basé sur une approche probabilistique (« Probability-based matching ») (Perkins, Pappin et al. 1999). Il s'agit en fait d'une implémentation probabilistique de l'algorithme MOWSE. Une des spécificités de Mascot est sa fonction de calcul de score. Ce dernier est défini par la formule :

$$S = -10 \times LOG10(P)$$

où P représente la probabilité absolue qu'une identification donnée soit un événement dû au hasard. Une probabilité de 1/10000 donne donc un score de 40. Ainsi, à l'inverse de la p-value, le score est proportionnel à la « véracité » de l'identification et représente donc une valeur plus intuitive pour l'utilisateur.

Le score fourni, qui reflète la similarité entre des données théoriques et expérimentale, est calculé à partir du nombre de fragments expérimentaux qui sont corrélés aux fragments théoriques. L'algorithme de calcul est un processus itératif qui s'apparente à une recherche de maximum de corrélation. Chaque spectre MS/MS est traité de la façon suivante :

1) Les intensités des pics sont normalisées afin d'obtenir des valeurs d'intensité relatives.
2) Mascot commence par sélectionner un petit nombre de pics expérimentaux correspondants aux intensités les plus élevées et calcule une valeur initiale de score. La fonction qui fournit un score pour un nombre donné de fragments corrélés, et pour un certain nombre de paramètres de recherche, n'est pas décrite.
3) Toujours par intensité décroissante un autre pic est ajouté à la sélection.

4) Un nouveau score est calculé et comparé avec le score précédent. Si le score ne diminue pas l'étape 3 est réitérée.
5) Mascot fournit le meilleur score qu'il a déterminé et qui devrait donc correspondre à la sélection de pics optimale.

On peut remarquer que Mascot n'essaye pas de trouver toutes les correspondances possibles dans le spectre mais essaye plutôt de réaliser une discrimination sur la base de l'intensité afin de prendre en compte principalement des pics « réels » et laisser de côté le bruit de fond.

I-4.5. Des peptides aux protéines

Quel que soit le logiciel utilisé pour l'identification des peptides, la dernière étape du processus informatique consiste toujours à regrouper les peptides qui pointent vers une même séquence protéique. Cette étape n'est cependant pas triviale car les relations qui existent entre ces deux entités sont multiples : d'une part une protéine peut être décrite par un ensemble de séquences peptidiques, et d'autre part un peptide peut correspondre à plusieurs séquences protéiques. Cette problématique, inhérente à l'analyse protéomique, a été largement décrite dans la littérature sous le terme de « protein inference » (Nesvizhskii and Aebersold 2005; Alves, Arnold et al. 2007). Elle nécessite des algorithmes spécifiques et les moteurs de recherche utilisent en général une méthode dite « parcimonieuse » qui consiste à trouver la liste de protéines la plus petite capable d'expliquer l'ensemble des séquences peptidiques identifiées.

De nombreuses méthodes ont été décrites ces 10 dernières années pour gérer cette correspondance multiple entre peptides et protéines (Nesvizhskii, Keller et al. 2003; Nesvizhskii and Aebersold 2005; Weatherly, Atwood et al. 2005; Alves, Arnold et al. 2007; Price, Lucitt et al. 2007; Zhang, Chambers et al. 2007; Li, Arnold et al. 2009; Ma, Dasari et al. 2009; Gerster, Qeli et al. 2010; Meyer-Arendt, Old et al. 2011). Certaines approches déterministes se basent sur la comparaison des ensembles de séquences peptidiques qui ont été attribués aux différentes séquences protéiques de la banque de données. En pratique, chacune des protéines de l'espace de recherche envisagé est associée à un groupe de séquences peptidiques qui ont pu être déduites de l'analyse des spectres MS/MS. L'étape suivante consiste à créer une hiérarchie de ces groupes peptidiques. Pour chaque protéine identifiée par un certain groupe de peptides, on peut par comparaison avec toutes les autres protéines distinguer les différents cas suivants (cf figure 7) :

- « distinct sets » : protéines présentant des peptides distincts de toutes les autres protéines.
- « differentiable sets » ou « overlapping sets » : protéines qui présentent une correspondance partielle de séquences communes et en plus des séquences spécifiques.
- « indistinguishable sets » ou « samesets » : protéines qui présentent exactement les mêmes séquences peptidiques que d'autres protéines identifiées.
- « subsets » : protéines dont tous les peptides sont également identifiés dans une autre protéine dite « overset » car présentant des séquences peptidiques spécifiques.
- « subsumable sets » : ensemble contenant une protéine dont les peptides peuvent être expliqués par ceux d'autres protéines identifiées.
- « shared peptides sets » : protéines qui ne possèdent aucun peptide spécifique et qui ne peuvent pas être considérées comme un « subset » (il faut pour cela un « ovserset » présentant des séquences spécifiques).

Dans une approche parcimonieuse, telle que celle effectuée par Mascot, seuls les cas « samesets » et « subsets » sont réellement utilisés pour fournir la liste de groupes de protéines définitive. Ainsi, tous les « samesets » sont regroupés et seuls les groupes de protéines qui ne sont pas des « subsets » sont conservés car ils sont supposés expliquer davantage l'information peptidique obtenue.

D'autres méthodes plus sophistiquées essayent d'appliquer des poids aux peptides et aux protéines en utilisant des approches probabilistes : compensation par la longueur de la séquence protéique, poids calculé sur la base de la fréquence d'apparition d'un peptide dans différentes séquences protéiques. Elles restent cependant plutôt marginales et l'approche que nous venons de décrire est en général la plus utilisée.

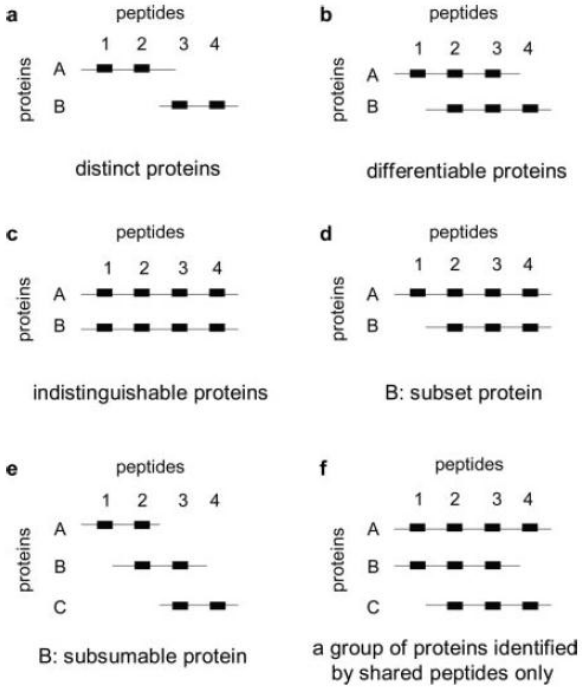


Figure 7 : les différents cas possibles de regroupement des peptides.
Source : (Kall, Storey et al. 2008) Nesvizhskii and Aebersold 2005

Nous avons détaillé dans les paragraphes précédents les méthodes utilisées pour identifier les protéines à partir de spectres MS/MS. Il est important de noter les limites de l'analyse effectuée par les moteurs de recherche tels que Mascot. Comme toute approche statistique, les scores peptidiques basés sur des calculs probabilistes dépendent de prérequis et de modèles. Un de ces prérequis est que les entrées des banques de données protéiques sont des séquences aléatoires. Un autre prérequis est que les masses expérimentales mesurées sont des observations indépendantes. L'approche par recherche en base de données suppose par ailleurs que les protéines présentes dans l'échantillon soient réellement recensées dans les banques, et que l'espace de recherche défini par l'utilisateur (lié par exemple à la tolérance de masse et aux modifications post-traductionnelles autorisées pour réaliser la recherche) reflète la « réalité biologique » de l'échantillon analysé. Tous ces éléments ne sont en pratique pas nécessairement respectés et le résultat fourni par le moteur de recherche est donc forcément biaisé. Nous allons voir maintenant que de nombreuses méthodes plus ou moins sophistiquées ont été développées ces dernières années afin de contrôler et de quantifier les erreurs présentes dans les résultats de recherches en banque de données.

I-5. Méthodes empiriques et statistiques pour la validation des résultats d'identification

L'analyse *in silico* des spectres MS/MS présente de nombreuses sources d'erreurs. La robustesse de la fonction de score et du modèle probabiliste, la qualité des spectres MS/MS, la pertinence des paramètres de recherche, l'exhaustivité de la banque de séquences protéiques sont autant de facteurs qui peuvent fausser l'identification des protéines de l'échantillon. La validation des résultats d'analyse protéomique a donc naturellement fait l'objet d'une prise de conscience durant ces dix dernières années (Carr, Aebersold et al. 2004). Des directives pour la publication de données sur l'identification de peptides et de protéines sont désormais définies (Taylor, Paton et al. 2007). Les principaux problèmes sont la présence d'un nombre plus ou moins élevé de faux positifs dans les résultats fournis, et le fait que chaque logiciel possède son propre algorithme de score. Etant donné l'absence de standards, il était important de mettre en place des normes pour assurer la qualité des informations publiées. D'une manière générale, les chercheurs doivent maintenant justifier leurs résultats en apportant des informations sur les méthodes, les seuils et les moteurs de recherche utilisés.

En parallèle de la définition de ces directives, le développement de méthodes pour la validation des résultats a également été au centre des préoccupations de la communauté protéomique. En effet, l'analyse nanoLC-MS/MS fonctionnant aujourd'hui de manière automatique, la capacité à générer des données dépasse largement celle à les analyser. Si l'on considère le nombre très élevé de spectres MS/MS pouvant être acquis lors d'une analyse protéomique de type « shotgun », il devient alors impossible d'imaginer une validation manuelle des assignements peptidiques obtenus. Les méthodes statistiques, qui ont été développées de manière intensive ces dernières années (Nesvizhskii and Aebersold 2004), sont aujourd'hui capables d'estimer des taux de faux positifs présents dans les jeux de données fournis par les moteurs de recherche. Les techniques les plus avancées permettent également d'augmenter la discrimination entre vraies et fausses identifications.

I-5.1. PeptideProphet et « Posterior Error Probability »

Une première approche empirique fut celle implémentée dans le programme PeptideProphet (Keller, Nesvizhskii et al. 2002). L'objectif initial était de convertir des valeurs de score de corrélation fournies par le moteur de recherche Sequest en valeurs de probabilité. Pour cela, PeptideProphet construit un modèle (dit modèle de mélange ou « mixture model ») de discrimination entre les vraies et les fausses identifications peptidiques. Ce modèle est construit à partir de deux distributions de scores. Les valeurs de score des « top ranking peptides » servent à construire le modèle des identifications correctes alors que les scores des autres séquences (rang supérieur à 1) sont utilisés pour construire le modèle des identifications incorrectes (cf figure 8). Les valeurs de scores de corrélation fournies par Sequest ne sont pas optimales pour bien discriminer les deux populations que nous venons de décrire. Ainsi PeptideProphet calcule un nouveau score dit de discrimination qui prend en compte la valeur de départ (Xcorr) ainsi qu'un certain nombre de facteurs supplémentaires, définis de la façon suivante :

- deltaCn: la différence relative entre le premier et le second score Xcorr pour tous les peptides candidats du spectre MS/MS.
- SpRank: le classement du score obtenu pour ce peptide par rapport aux autres peptides candidats.

- dM: la différence de masse absolue entre la valeur observée pour l'ion précurseur et celle calculée pour le peptide en question.

A partir d'un jeu de donnée d'apprentissage, l'algorithme détermine la combinaison de ces facteurs qui discrimine au mieux les vraies des fausses identifications : cette approche est appelée « Expectation Maximisation algorithm ». De façon plus imagée chaque sous-score se voit attribuer un poids plus ou moins élevé et le logiciel optimise l'ensemble des poids afin de maximiser la discrimination entre les vraies et les fausses identifications.

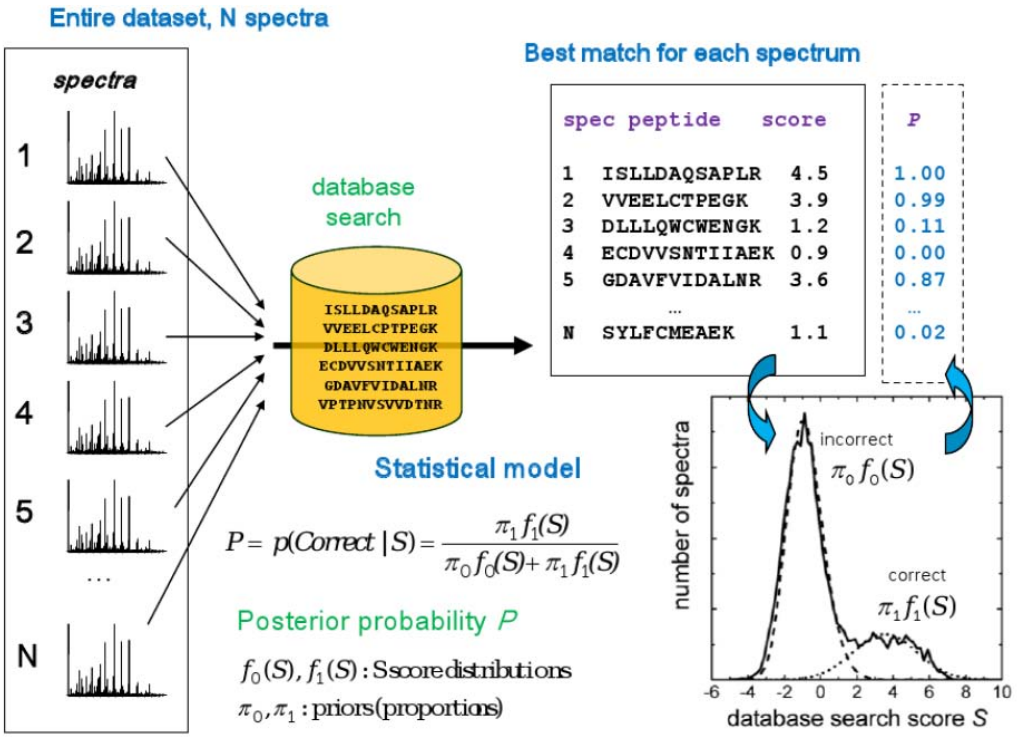


Figure 8 : illustration de l'analyse statistique réalisée par PeptideProphet.
Source : (Nesvizhskii 2010)

En effectuant une nouvelle recherche, l'algorithme est ensuite capable de déterminer les distributions (correctes et incorrectes) des scores discriminants pour le jeu de donné en question et d'en déduire la probabilité (« Posterior Error Probability » ou PEP) qu'une identification peptidique soit correcte. Il devient ainsi possible de filtrer les résultats sur la base de cette valeur de probabilité. Le principal inconvénient de cette approche est qu'elle repose sur un mécanisme d'apprentissage. Ainsi les valeurs de probabilité fournies reflèteront le TFP (taux de faux positifs) si le jeu de donné analysé et les paramètres de recherche utilisés sont proches de ceux mis en place dans la phase d'apprentissage.

I-5.2. Stratégie « target-decoy »

Bien que PeptideProphet ait apporté une solution au problème de la validation des résultats fournis par les moteurs de recherche, de nouvelles méthodologies permettant de contrôler le TFP associé au jeu de données analysé ont été développées en parallèle. Parmi ces différentes méthodes, la stratégie dite « target-decoy » a été largement plébiscitée par la communauté internationale. En effet, celle-ci est simple à mettre en œuvre car elle ne nécessite pas d'étape d'apprentissage et elle est en plus applicable à tout type de moteur de recherche. Contrairement à l'analyse statistique que

nous venons de détailler elle ne donne pas accès à des valeurs de confiance pour chaque identification mais à une estimation globale du TFP du jeu donné qui aura été filtré au préalable selon les critères de score du moteur de recherche. Cette mesure est appelée « False Discovery Rate » dans la littérature d'analyse statistique. Dans ce domaine le concept général du contrôle du FDR (Benjamini and Hochberg 1995) a comme objectif de corriger les valeurs de p-value issues de l'utilisation répétée d'un même test statistique (« multiple testing »). Il consiste à trier les p-value calculées par ordre décroissant puis à multiplier chaque p-value par le nombre de tests réalisés et à diviser le tout par le rang qu'occupe la p-value dans la liste ordonnée (la valeur résultante étant le FDR). Ainsi la p-value la plus faible est celle qui est la plus corrigée alors que la p-value la plus élevée reste inchangée. Il s'agit d'une amélioration de la correction dite de Bonferroni où l'on se contente de multiplier chaque p-value par le nombre de tests réalisés (méthode plus simple à mettre en place mais très stringente car elle génère beaucoup de faux négatifs). Même si la terminologie utilisée pour la validation des résultats d'identifications par spectrométrie de masse est identique nous allons voir que la mise œuvre du contrôle du taux faux positifs est cependant très différente.

L'estimation du TFP pour l'identification de spectres MS/MS nécessite une procédure d'analyse particulière. Celle-ci consiste d'une manière générale à inclure dans la recherche des versions aléatoires (ou « decoy ») des séquences protéiques présentes dans la banque de données cible (ou « target ») (Peng, Elias et al. 2003). Puisque l'on s'attend à n'obtenir que des fausses identifications dans la banque aléatoire, on peut comparer le nombre d'identifications issues de cette banque, après avoir appliqué des critères de validation, avec celui que l'on a obtenu dans la banque cible et ainsi estimer un TFP (Elias, Haas et al. 2005). Différents algorithmes existent pour la création de banques aléatoires : ceux générant des séquences inversées et ceux générant des séquences aléatoires. Dans les banques inversées, les séquences des entrées sont lues du C-ter vers le N-ter et dans les banques « aléatoires », les acides aminés sont mélangés mais le nombre d'entrées est conservé. Il a cependant été démontré que le type de banque aléatoire utilisé avait peu d'influence sur la valeur du TFP calculé (Elias and Gygi 2007). Par contre la façon dont on utilise les banques aléatoires peut modifier la fonction de calcul dont la formulation générale est :

$$FDR = FP_{obs} / (FP_{obs} + VP_{obs})$$

avec FP_{obs} représentant le nombre de faux positifs observés et VP_{obs} le nombre de vrais positifs observés. On distingue en effet deux méthodologies pour leur utilisation (cf figure 9) : on peut soit effectuer deux recherches séparées dans les banques cibles et aléatoires, soit effectuer une seule recherche dans une banque concaténée qui contient les deux types de séquences. Dans cette deuxième approche les séquences peptidiques générées *in-silico* sont en compétition pour l'assignation à un spectre MS/MS. En effet, on ne retient en général qu'une seule séquence pour un spectre donné, le « top ranking peptide », celle-ci pouvant donc provenir dans ce cas d'une séquence protéique target ou decoy. Elias et Gygi (Elias and Gygi 2007) ont démontré que les faux positifs target et decoy sont présents en quantité équivalente, et comme il sont en compétition dans une recherche en banque concaténée on s'attend à avoir une équiprobabilité d'observer une fausse identification target ou decoy. On peut donc en déduire l'égalité suivante : $FP_{target} \approx FP_{decoy} = 0.5\ FP_{obs}$. Etant donné que seuls les faux positifs decoy sont clairement identifiables, le nombre de faux positifs observés ne peut être déterminé qu'à partir des identifications provenant de la banque aléatoire. Si l'on considère que l'ensemble des hits decoy (P_{decoy}) ne contient que des faux positifs alors on en déduit que : $FP_{obs} \approx 2 \times FP_{decoy} \approx 2 \times P_{decoy}$. En ce qui concerne le calcul du dénominateur de la formule

générale du FDR (FP_{obs} + VP_{obs}) il est également nécessaire de prendre en compte cet effet de compétition qui peut aussi s'écrire : $FP_{obs} = FP_{target} + FP_{decoy}$. Comme le nombre total d'observations positives target (P_{target}) vérifie l'égalité $P_{target} = FP_{target} + VP_{target}$ et que $VP_{obs} = VP_{target}$, on peut alors déduire que :

$$\left.\begin{aligned} & FP_{obs} + VP_{obs} \\ &= FP_{target} + FP_{decoy} + VP_{target} \\ &= P_{target} + FP_{decoy} \\ &= P_{target} + P_{decoy} \end{aligned}\right\} \text{Ce qui nous permet d'établir au final que : } FDR_{concat} = \frac{FP_{obs}}{FP_{obs} + VP_{obs}} = \frac{2\,x\,P_{decoy}}{P_{target} + P_{decoy}}$$

L'inconvénient principal de cette approche est qu'elle double l'espace de recherche ce qui peut avoir un effet sur certaines valeurs retournées par les moteurs de recherche. Par exemple les valeurs de seuil d'identité et d'homologie de Mascot sont plus élevées avec ce type de recherche mais étant donné que l'ensemble des identifications sont impactées de façon similaire cela n'a pas d'effet notable sur le résultat final (Elias and Gygi 2007).

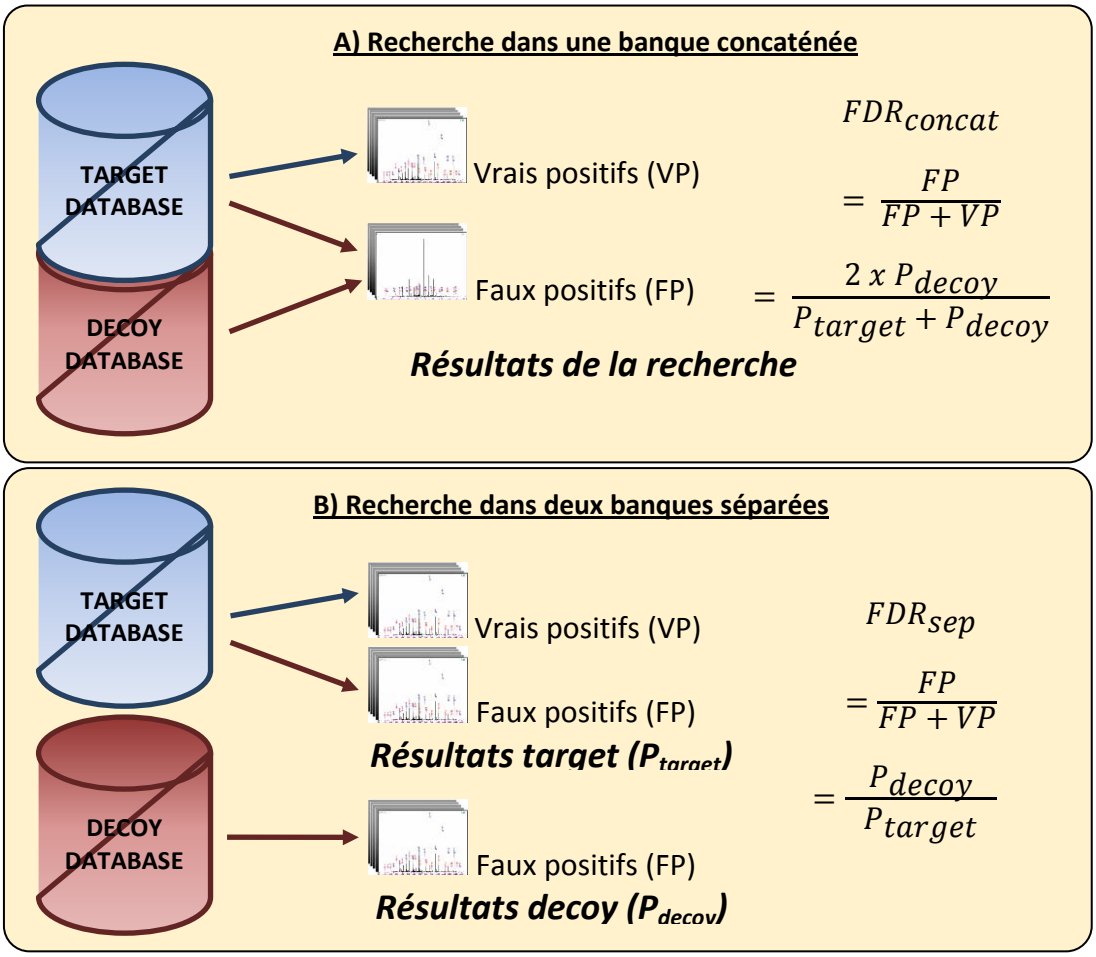


Figure 9 : modes de recherche en banques target/decoy. Deux approches sont possibles : A) une unique recherche à partir d'une banque contenant les séquences target et decoy. B) deux recherches séparées dans une banque contenant les séquences target et une autre contenant les decoy. Le calcul du FDR est différent selon le mode de recherche employé.

La deuxième manière de procéder pour réaliser ce type d'analyse est de rechercher les séquences target et decoy séparément. Cela simplifie d'ailleurs le calcul du TFP qui devient :

$$FDR_{sep} = \frac{FP}{FP + VP} = \frac{P_{decoy}}{P_{target}}$$

Les valeurs de P_{target} et P_{decoy} ne sont pas biaisées par l'effet de compétition qui existe dans la recherche en banques concatanées et peuvent donc être utilisées directement pour le calcul du FDR. En contrepartie, le nombre de faux positifs observés est surestimé car les spectres MS/MS donnant une identification dans la recherche en banque decoy sont toujours comptabilisés comme des faux positifs, même dans les cas où ces spectres ont également donné une identification de meilleur score dans la banque target. Pour répondre à ce problème Navarro *et al.* (Navarro and Vazquez 2009) ont mis au point une nouvelle stratégie d'analyse des résultats issus de recherches en banques séparées. Celle-ci consiste à simuler l'effet de compétition par analyse conjointe des scores obtenus dans les deux recherches pour chaque spectre MS/MS. Cette approche sera présentée plus en détail dans la partie IV de ce document. Les auteurs font également un bon résumé des inconvénients des analyses « target-decoy » séparées et concaténées : "*we cannot conclusively affirm that an FDR estimate in the reference population is worse than a more accurate FDR estimate in an inflated population*". Bien qu'imparfaite, l'approche target-decoy permet cependant de fournir un critère objectif reflétant la qualité d'un jeu de données obtenu après filtrage des résultats du moteur de recherche, élément essentiel pour la diffusion et la publication d'études reposant sur l'analyse de spectres MS/MS.

I-5.3. Stratégies hybrides

Contrairement à l'approche paramétrique de PeptideProphet, l'approche « target/decoy » ne nécessite pas de faire des hypothèses sur la distribution des scores obtenus par le moteur de recherche. En revanche elle ne donne accès qu'au taux d'erreur global de l'identification des peptides et non à la probabilité qu'un peptide individuel soit correct. Kaëll *et al.* ont proposé un nouvelle stratégie qui permet d'estimer le FDR local (autre nom pour la « posterior error probability » ou PEP) à partir de recherches target-decoy séparées (Kall, Storey et al. 2008).

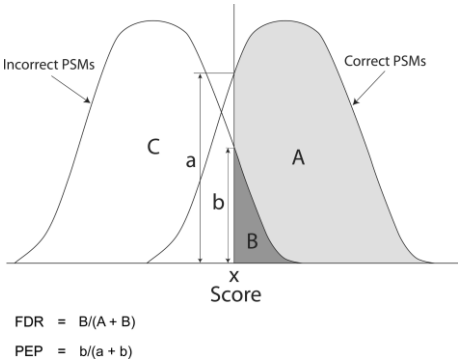


Figure 10 : illustration de l'utilisation des distributions de score des « peptide spectrum matches » (PSMs) corrects et incorrects afin d'estimer l'erreur associée à une recherche en banque de données.
La valeur de FDR correspond au ratio de l'aire B (nombre de PSMs incorrects) divisée par l'aire (A + B) c'est-à-dire le nombre total de PSMs validés.
Le PEP correspond quant à lui au ratio du nombre de PSMs ayant un score égal à une valeur « x » pour la distribution incorrecte divisée par le nombre total de ces mêmes observations pour les deux distributions.

Si l'on admet que les fausses identifications peptidiques suivent la même distribution de score dans les deux banques de données, le nombre de faux positifs peut être alors corrélé à une valeur de score donnée. Les auteurs ont utilisé une méthode semi-supervisée capable d'estimer la distribution des scores des peptides « decoy » à partir du jeu de données analysé. Cette méthode non-paramétrique (Kall, Storey et al. 2008), basée sur une régression logistique, ne nécessite donc pas un apprentissage

préalable qui devrait être adapté au type de configuration instrumentale utilisé (instrument et analyseur). De plus, contrairement à la version initiale de PeptideProphet, cette méthode n'est pas dépendante d'une fonction de score particulière et peut donc être appliquée en théorie à n'importe quel moteur de recherche. Depuis sa création, PeptideProphet a cependant évolué et peut aujourd'hui modéliser la distribution des identifications incorrectes à partir de résultats decoy rendant ainsi l'approche semi-supervisée (Choi and Nesvizhskii 2008) et permettant désormais au logiciel d'analyser les résultats d'un plus grand nombre de moteurs de recherche.

I-5.4. Validation des protéines

La validation des séquences peptidiques, dont nous venons de détailler les possibilités de mise en œuvre, est une étape nécessaire mais pas suffisante dans les études où l'on s'intéresse à caractériser le protéome global d'un échantillon donné. En effet, dans ce cas on s'intéresse principalement au TFP au niveau protéique et non peptidique. Il est important de noter que si l'on se contente de filtrer la liste des peptides validés pour un TFP donné et que l'on conserve uniquement les protéines qui contiennent au moins un de ces peptides validés, alors le résultat obtenu aura un taux d'erreur au niveau protéique supérieur à celui fixé au niveau peptidique. Ce problème est principalement lié au fait que les peptides corrects appartenant à une même protéine se regroupent de façon non aléatoire, contrairement aux peptides incorrects (decoy) qui ont tendance à ne pas être identifiés sur les mêmes séquences protéiques. Ainsi même si le TFP peptidique est faible il est possible d'obtenir un TFP protéique élevé avec le mode filtrage énoncé.

Dans une étude de 2003 (Peng, Elias et al. 2003), les auteurs ont estimé qu'un TFP de 26 % pour les protéines identifiées avec un peptide unique est réduit à 1,4 % pour des identifications avec 2 peptides et devient nul pour des identifications avec 3 ou plus de peptides. Selon les auteurs d'une autre étude (Steen and Mann 2004), les identifications de protéines avec un seul peptide doivent être limitées à des cas exceptionnels. Partant de ce constat de nombreuses études ont été publiées en mettant de côté les identifications protéiques à 1 peptide, généralisant ainsi une règle appelée « the two-peptides rule ». Par la suite, d'autres études (Gupta and Pevzner 2009) ont remis en question cette dernière et ont démontré que pour un pour même FDR, il est possible d'obtenir un plus grand nombre d'identifications protéiques si l'on comptabilise les « one-hit-wonders », i.e. les protéines identifiées avec un seul peptide, en les validant de façon plus stringente.

Une méthode alternative de validation des protéines est de ne pas filtrer les identifications peptidiques mais de cumuler leur score. Il s'agit de l'approche initiale de Mascot qui peut soit calculer un score standard (somme des scores de peptides) ou score plus évolué appelé MudPIT (cf partie I-5.5). ProteinProphet utilise également une approche similaire en combinant les « posterior probabilities » des peptides. Les développeurs de ce programme se sont cependant heurtés à un problème de surestimation des probabilités pour les protéines ne comportant qu'un unique peptide. Ils ont essayé de le contourner en corrigeant les probabilités peptidiques, de manière à pénaliser les « one-hit-wonders » et à favoriser au contraire les « multi-hit-wonders ». L'ajustement effectué dépend principalement de la complexité de l'échantillon (nombre de spectres MS/MS) et est déterminé automatiquement pour chaque jeu de données par l'intermédiaire d'une procédure itérative. Il a cependant été démontré récemment (Claassen, Reiter et al. 2011) que cette correction n'était pas vraiment optimale. Les auteurs expliquent que l'approche sophistiquée et probabilistique de ProteinProphet a tendance à favoriser des protéines présentant de multiples peptides dont les

spectres sont de qualité moyenne ou faible, et que cet effet est d'autant plus prononcé que l'échantillon est complexe. Ils définissent également la notion de « schémas de sélection » qui consiste à valider les protéines de façon plus ou moins sévère selon le nombre de peptides identifiés. La conclusion de leur étude est la suivante : "*For all considered cases, processing of the data with simple protein inference approaches and keeping all the spectral evidence achieves competitive proteome coverage*". Le concept formalisé ici n'est pas réellement nouveau car différents logiciels développés tels que MFPaQ (Bouyssie, Gonzalez de Peredo et al. 2007) et IRMa (Dupierris, Masselon et al. 2009) ont implémenté des routines de validation basés sur des « schémas de sélection » (appelés règles ou filtres de validation). Nous verrons d'ailleurs dans la partie II le détail de la procédure de validation réalisée par le logiciel MFPaQ.

I-5.5. Validation des résultats Mascot

Une information intéressante dont on dispose dans Mascot pour valider les résultats fournis est celle du calcul du seuil d'identité ou « identity threshold ». Cette valeur est propre à chaque identification peptidique. Elle est dérivée du modèle probabiliste utilisé par Mascot et peut être calculée via la formule :

$$IT = -10 \times LOG10(\frac{p}{N})$$

où p est le seuil de significativité (p-value) et N correspond au nombre de peptides candidats de la banque de données qui ont une masse compatible avec la masse expérimentale analysée (i.e. sont dans la fenêtre de masse calculée à partir d'une valeur de tolérance). Ainsi, le seuil d'identité reflète à la fois le niveau de significativité que l'on veut se fixer (p=0.05 par exemple) et également la taille de l'espace de recherche. Celle-ci est proportionnelle aux paramètres de recherche : taille de la banque de données, spécificité de l'enzyme, nombre de modifications post-traductionnelles recherchées, valeur de la tolérance d'erreur de masse de l'ion précurseur. Il est important de noter que la variation de la taille de l'espace de recherche modifie la valeur du seuil d'identité mais qu'elle n'a pas d'influence sur la valeur du score Mascot. En fait, le seul paramètre de recherche qui a une influence réelle sur le score peptidique est la tolérance d'erreur de masse appliquée sur les fragments MS/MS car elle affecte le nombre de pics expérimentaux pouvant être corrélés aux données théoriques.
Ce seuil d'identité est utilisé par Mascot pour calculer une E-value du peptide, qui reflète le nombre de fois où l'on peut s'attendre à obtenir par chance un score identique ou meilleur. La E-value est calculée suivant la formule suivante :

$$\text{E-value} = p \times 10^{((\text{Identity threshold} - \text{Ion score}) / 10))}$$

Ainsi cette E-value n'est pas définie comme décrit dans la section précédente (I-4.3) par la détermination de l'histogramme de répartition des scores de tous les candidats peptidiques pour un spectre donné. En effet, ici Mascot ne prend en compte que le nombre de candidats potentiels pour le spectre MS/MS analysé mais pas les valeurs de score obtenues pour chacun de ces candidats. Cependant, lorsque le spectre MS/MS a été attribué à un nombre de peptides candidats suffisamment élevé, Mascot est également capable de calculer un seuil dit d'homologie. Le calcul de ce dernier est peu décrit mais il s'apparente au calcul de la queue de la distribution de scores aléatoires (déterminée à partir des peptides dont le rang est supérieur à 1). Ainsi en remplaçant la

valeur du seuil d'identité par celle du seuil d'homologie dans la formule de calcul de la E-value on obtient une valeur qui se rapproche d'avantage de la définition que nous avions faite précédemment. Toutefois, le seuil d'homologie n'est pas toujours disponible et par conséquent certains spectres MS/MS disposent uniquement d'une valeur de seuil d'identité. Quel que soit le seuil disponible il est important de le mettre en vis-à-vis de la valeur du score du peptide afin de valider avec une plus grande spécificité les identifications peptidiques. C'est d'ailleurs la méthode que Mascot a introduit en 2004 dans sa version 2.0 pour déterminer le score des protéines. Cette nouvelle fonction appelée « MudPIT » dispose d'un algorithme plus sophistiquée que la fonction de score « standard » qui calcule la somme des scores des peptides non redondants appartenant à une même protéine :

```
Protein score = 0

For each peptide match {
  If there is a homology threshold and ions score > homology threshold {
    Protein score += peptide score - homology threshold
  } else if ions score > identity threshold {
    Protein score += peptide score - identity threshold
  }
}
Protein score += 1 * average of all the subtracted thresholds
```

Avant la version 2.2 de Mascot (fin 2006) les seuils étaient calculés à partir d'une valeur fixe de probabilité (p-value = 0.05) alors qu'il est aujourd'hui possible de spécifier cette valeur de probabilité depuis l'interface du logiciel. L'introduction de ce nouveau score protéique permet au moteur de recherche de gagner en spécificité en éliminant les protéines qui présentent de nombreux peptides de faibles scores. Il est en effet préférable de disposer d'une bonne identification peptidique plutôt que de nombreuses mauvaises. Une approche alternative pour valider les résultats d'identification Mascot consiste à calculer une valeur de seuil moyennée sur l'ensemble des peptides identifiés puis à utiliser ce seuil pour tous les peptides. Cette valeur est calculée et affichée par Mascot et a été, de par sa facilité d'accès, très employée par les utilisateurs pour valider les résultats du moteur de recherche.

Comme nous l'avons vu précédemment, les méthodes de validation des résultats d'identification ont beaucoup évolué durant ces dix dernières années et les moteurs de recherche ont suivi la tendance. En effet, dans cette optique de validation des résultats et d'estimation du TFP, de nouvelles fonctionnalités ont été intégrées dans le logiciel Mascot. Celui-ci est capable aujourd'hui de générer à la volée des versions decoy des séquences classiquement présentes dans les banques. Ces séquences decoy ont la même longueur que les séquences target mais possèdent une composition aléatoire. L'analyse s'apparente ensuite à une recherche dans deux banques séparées, le fichier résultat Mascot (.dat) contenant au final les deux types de résultat. Enfin, au niveau de l'affichage il est possible de visualiser le TFP des peptides pour une certaine valeur de p-value fixée (0.05 par défaut).

La généralisation de l'approche target-decoy au niveau de moteurs de recherche a réellement simplifié la validation des résultats d'identification, mais au-delà du calcul du TFP, il est également nécessaire de disposer de méthodes de validation capables de maximiser le nombre d'identifications pour un TFP donné.

I-6. L'analyse quantitative de données LC-MS/MS

Nous avons évoqué dans les paragraphes précédents comment les données de spectrométrie de masse pouvaient être utilisées pour identifier à haut-débit les peptides obtenus à partir de mélanges complexes de protéines. Nous n'avons cependant pas encore abordé les stratégies quantitatives qui peuvent être mises en œuvre sur ce même type de données (Xie, Liu et al. 2011). Les abondances peptidiques et protéiques au sein d'un échantillon, ou les variations d'abondance entre plusieurs échantillons, sont des informations cruciales pour la compréhension de phénomènes biologiques. La quantification de données de spectrométrie de masse implique la mise en œuvre d'une série d'étapes de calcul pouvant exploiter les données MS et/ou MS/MS. L'analyse de la grande quantité d'information cachée dans ces couches de données ne peut évidemment pas être appréhendée par une approche manuelle, et les méthodes de calcul informatique sont donc devenues des outils incontournables pour répondre à cette problématique. Le chapitre suivant présente les principes des stratégies quantitatives protéomiques (avec ou sans marquage) ainsi que qu'un court historique sur l'apparition des outils bioinformatiques dédiés à ce type d'analyse.

I-6.1. Les stratégies quantitatives protéomiques

Ces dernières années, différentes approches de transcriptomique, de protéomique et de métabolomique ont été développées afin d'analyser les variations de systèmes biologiques après un changement dans l'environnement. Les approches basées sur la protéomique quantitative consistent soit à déterminer la quantité absolue de protéines dans un ou plusieurs échantillons, soit le plus fréquemment, à comparer les quantités relatives de protéines présentes dans différentes conditions (par exemple malade/sain ; +/- traitement ;…) en observant l'apparition, la disparition ou les variations des signaux associés aux protéines (Ong and Mann 2005). Les méthodes pour la quantification absolue des protéines font le plus souvent intervenir l'utilisation de peptides de référence isotopiquement alourdis. Ces derniers sont ajoutés à l'échantillon à analyser, servant ainsi de standards internes pour la quantification des peptides légers. En conséquence, la quantification absolue est souvent associée à des stratégies de protéomique ciblée, dans lesquelles un nombre limité de peptide d'intérêt sont spécifiquement mesurés, ainsi que leurs standards internes associés, à l'aide de spectromètres de masse dédiés de type triple-quadripôle. Dans les approches globales, en revanche, où le but est de d'identifier et de quantifier l'ensemble (ou une grande partie) d'un protéome, ce sont le plus souvent des approches de quantification relative qui sont mises en œuvre (cf figure 11).

Bien que l'analyse de gels 2D a été précurseur dans ce domaine, elle est peu à peu remplacée par des analyses nanoLC-MS/MS qui permettent aujourd'hui de quantifier un plus grand nombre de protéines tout en donnant accès de façon immédiate à l'identité de ces mêmes protéines. La quantification porte alors sur plusieurs milliers d'espèces, et nécessite des logiciels adaptés pour le traitement de données complexes. C'est sur ce type d'analyse quantitative globale à grande échelle qu'a principalement porté ce travail de thèse.

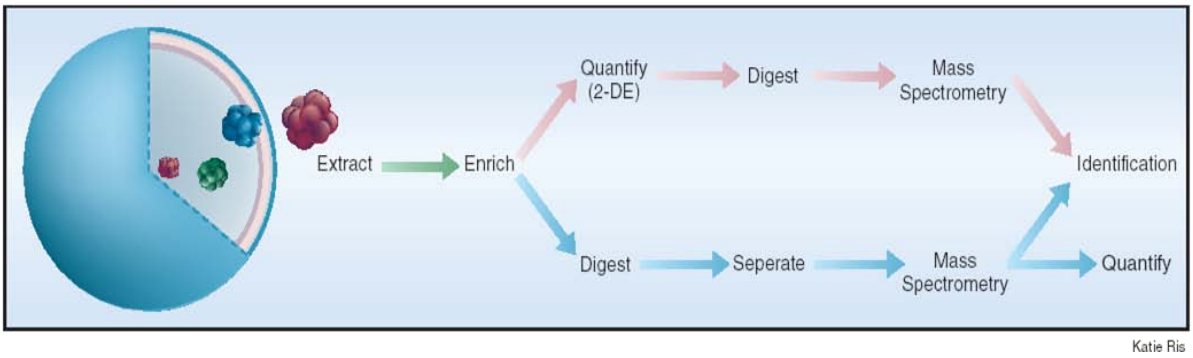


Figure 11 : méthodes de quantification relative des protéines. La comparaison d'images de gels d'électrophorèse bi-dimensionnels et l'analyse par nanoLC-MS/MS sont les deux grandes méthodes qui peuvent être mises en œuvre afin de réaliser une étude protéomique différentielle et quantitative.
Source : Patterson, S.D. & Aebersold, R.H (2003) Nature Genetics 33 (suppl.), 311-323

Pour réaliser cette quantification relative nanoLC-MS/MS sur des mélanges protéiques complexes, il existe deux grands types de stratégies : celles basées sur le marquage isotopique des peptides ou des protéines dans une des conditions comparées, et celles réalisées sans marquage. Le diagramme suivant donne un aperçu des différentes approches quantitatives existantes qui sont basées sur ces deux grandes stratégies.

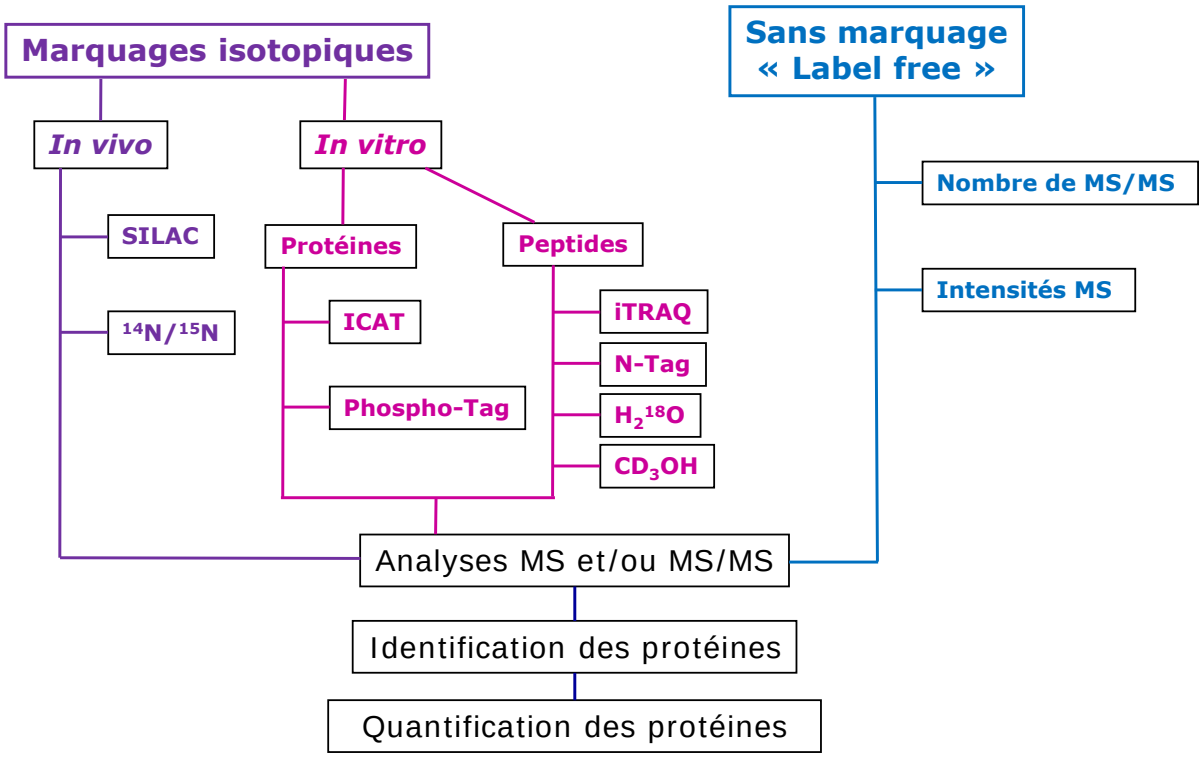


Figure 12 : diagramme des différentes approches protéomiques utilisées pour l'analyse quantitative par nanoLC-MS/MS.

Les stratégies mettant en œuvre un marquage ont été les premières utilisées dans le cadre d'analyses quantitatives par nanoLC-MS/MS. L'introduction d'un marqueur sur les molécules rend possible la quantification relative de deux conditions au sein de la même acquisition nanoLC-MS/MS.

En effet, la théorie de dilution des isotopes stables nous dit qu'un peptide marqué par un isotope stable est chimiquement identique à son homologue non marqué. Par conséquent, les deux peptides se comportent de manière identique lors de la séparation chromatographique ainsi que pendant l'analyse par spectrométrie de masse (de l'ionisation à la détection). Comme il est possible de mesurer précisément la différence de masse entre le peptide marqué et non marqué en spectrométrie de masse, la quantification peut être réalisée en intégrant et comparant les intensités du signal des peptides marqués et non marqués (cf figure 13).

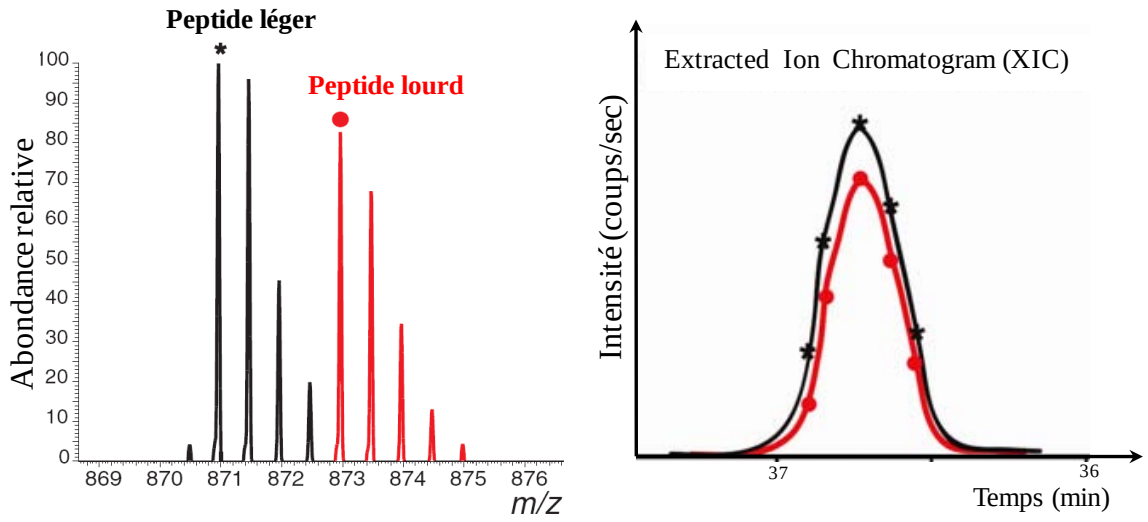


Figure 13 : extraction des données quantitatives à partir d'un spectre de masse.
A gauche, après acquisition du spectre de masse, visualisation des massifs isotopiques pour chaque peptide, marqué (rouge) ou non marqué (noir). A droite, au cours de l'élution chromatographique, le signal du peptide est suivi par la courbe dont l'aire correspond au courant mesuré (XIC), une mesure proportionnelle à l'abondance du peptide. La mesure des aires indique que le peptide rouge est présent à 85% du peptide noir.

De fait, ces approches consistent donc à marquer un des deux échantillons (échantillon lourd) avec un isotope stable (D, ^{13}C, ^{15}N), puis à le rassembler avec l'échantillon léger, les deux échantillons étant ensuite traités et analysés simultanément. Les paires peptidiques ainsi formées se différencient uniquement par un écart de masse Δm au sein des spectres de masse, et permettent de quantifier de façon relative les peptides correspondants.

Les marqueurs isotopiques peuvent être introduits de différentes façons (marquage chimique, enzymatique, métabolique), et à différentes étapes du processus analytique (cf figure 14). Plus le marquage est introduit en amont de ce processus, plus les échantillons peuvent être rassemblés précocement, éliminant les différentes sources de variabilité liées au traitement parallèle des échantillons.

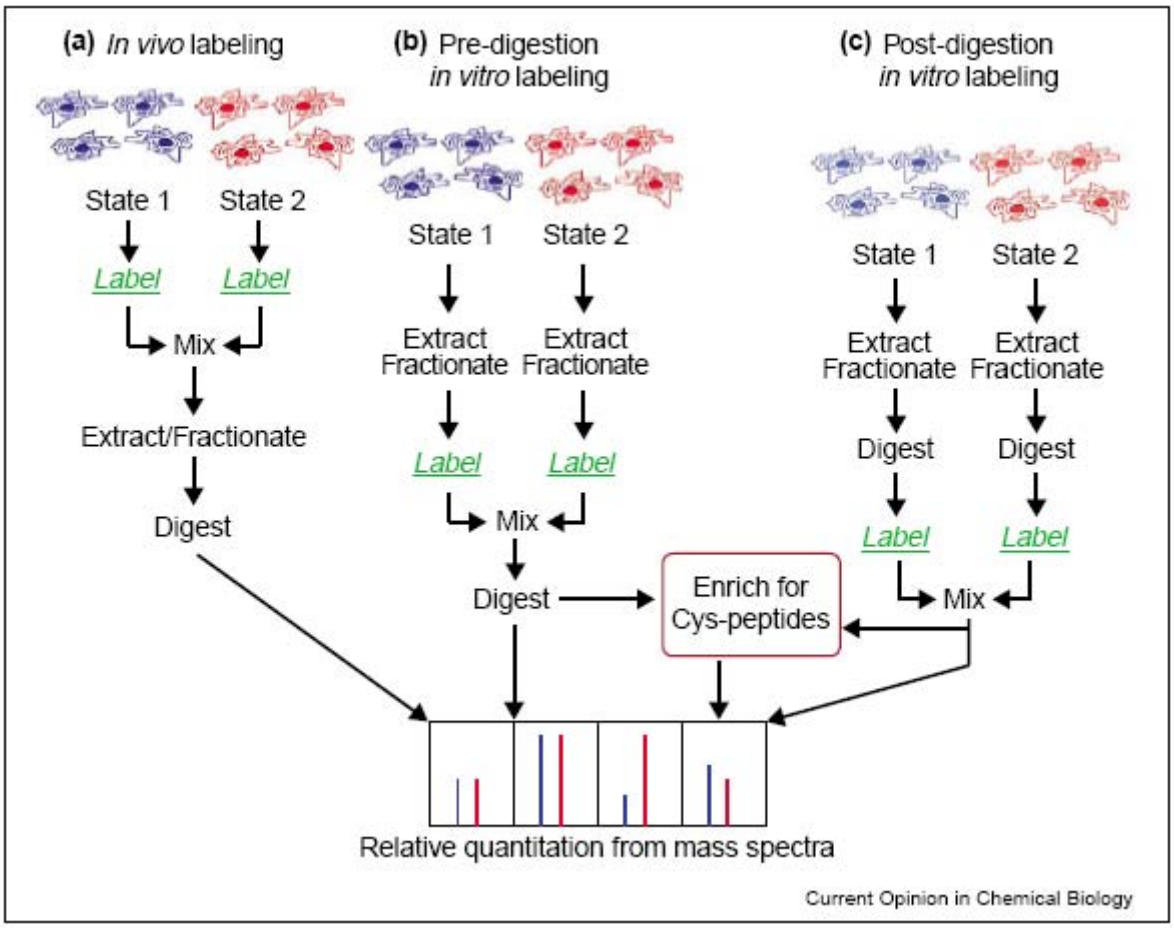


Figure 14 : comparaison des protocoles de traitement des échantillons pour les différentes stratégies de marquage isotopique. Le marquage métabolique (a) est réalisé dès la culture des cellules tandis que les marquages chimiques sont réalisés soit avant la digestion des protéines (b) soit sur le mélange peptidique (c). Suivant le type de marquage utilisé (notamment l'ICAT) le mélange de peptides marqués peut subir une étape d'enrichissement.
Source : Sechi & Oda (2003) Current Opinion in Chemical Biology, 7, 70-77

Au cours de mon doctorat j'ai principalement travaillé sur le traitement de données issues d'un marquage chimique de type ICAT ou d'un marquage métabolique de type SILAC ou ^{15}N. Ainsi je me contenterai de détailler le principe de ces trois techniques même si nous avons vu dans la figure 12 qu'il en existe plusieurs autres. Je présenterai ensuite les techniques sans marquage qui ont également représenté une partie importante de ce travail de thèse.

Le marquage chimique ICAT

Le premier marquage chimique utilisé pour la quantification relative des protéines a été décrit par le groupe de Ruedi Aebersold en 1999, il s'agit de la technique ICAT (Isotope-Coded Affinity Tag). Dans cette approche, les résidus cystéines sont marqués avec un réactif comprenant soit des ^{12}C, soit des ^{13}C qui introduisent une différence de masse de 9 Da entre le peptide léger et lourd (Hansen, Schmitt-Ulms et al. 2003) (cf figure 15). Il s'agit d'un marquage réalisé sur les protéines, après l'étape d'extraction à partir des échantillons biologiques, et avant la digestion trypsique. Le marqueur ICAT étant par ailleurs biotinylé, il est possible par la suite de purifier sélectivement les peptides marqués, ce qui simplifie les mélanges à analyser et facilite la détection de protéines minoritaires. En revanche, seuls les peptides contenant une cystéine peuvent être isolés et quantifiés dans cette approche. La cystéine étant un acide aminé présent dans seulement 80% des protéines, une partie du protéome n'est intrinsèquement pas détectable par cette méthode.

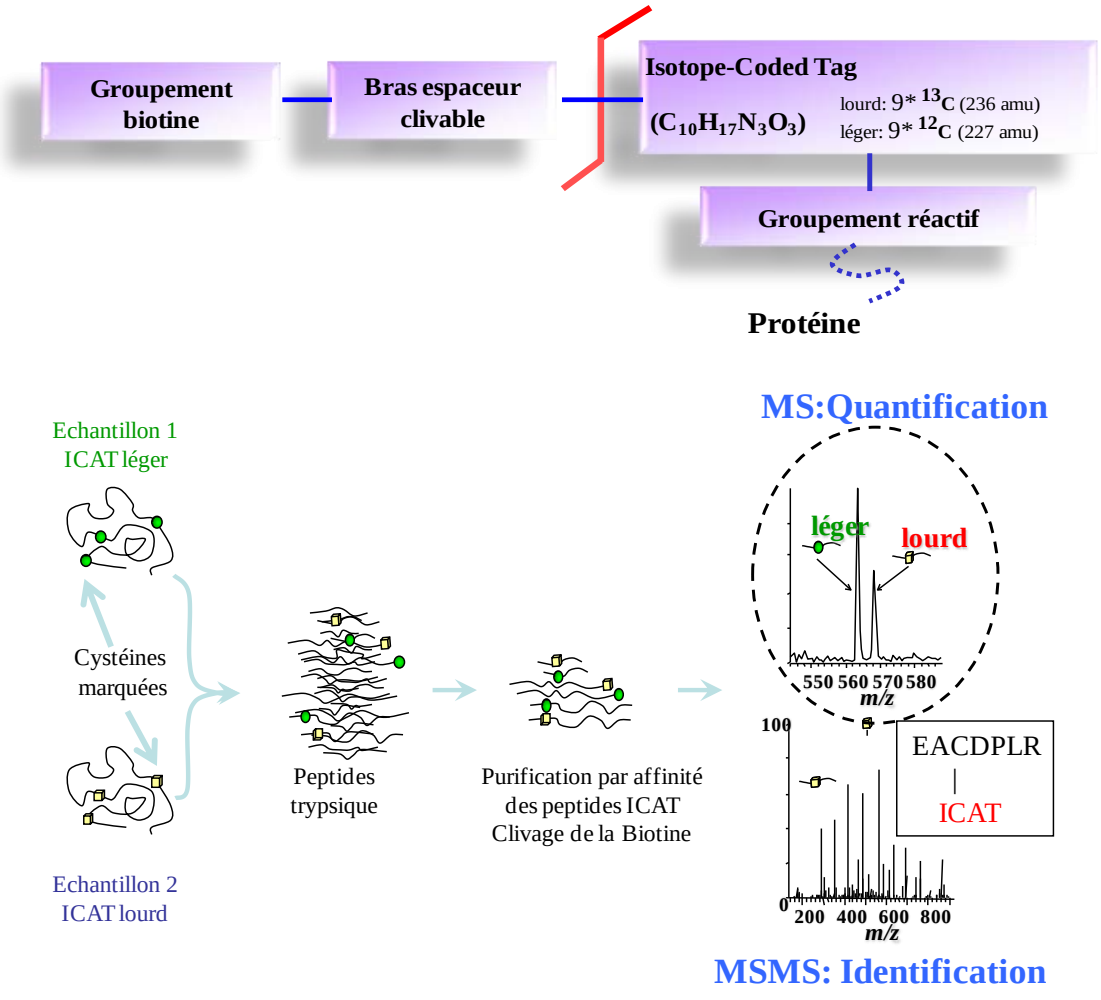


Figure 15 : principe de la stratégie ICAT. En haut est présentée la structure du réactif ICAT. En bas, le schéma illustre la mise en œuvre du marquage. Deux lots de protéines sont marqués avec le réactif léger et le réactif lourd respectivement. Ils sont ensuite mélangés et subissent une digestion trypsique. Les peptides marqués par les ICAT (lourd et léger) sont enrichis par chromatographie d'affinité sur une colonne d'avidine. Le groupement biotine est éliminé grâce au bras clivable en milieu acide avant l'analyse par spectrométrie de masse. Les couples peptide lourd et léger sont repérés par MS, identifiés par MS/MS puis quantifiés sur les spectres MS.

Si les réactions mises en œuvre par ces techniques de marquage chimique sont incomplètes, elles peuvent introduire un biais dans l'analyse quantitative qui en découle. Ainsi les marquages chimiques doivent être très spécifiques, totaux, et impliquer une manipulation minimale de l'échantillon.

Les marquages métaboliques ^{15}N et SILAC

Un des premiers marquages métaboliques des protéines décrit dans littérature est celui mettant en œuvre le marquage total de bactéries en utilisant un milieu de culture enrichi en azote ^{15}N (Oda, Huang et al. 1999). Nous verrons dans la partie III-6 que ce marquage est également applicable sur les plantes de petite taille. Peu après une autre technique a été développée par le groupe de Mathias Mann en 2002 (Ong, Blagoev et al. 2002) : le marquage SILAC (« Stable Isotope Labeling by Amino acids in Cell culture »). Le principe de ces deux techniques est très similaire et consiste à ajouter dans le milieu de culture un élément isotopiquement alourdi qui sera ensuite intégré au niveau des protéines par la cellule.

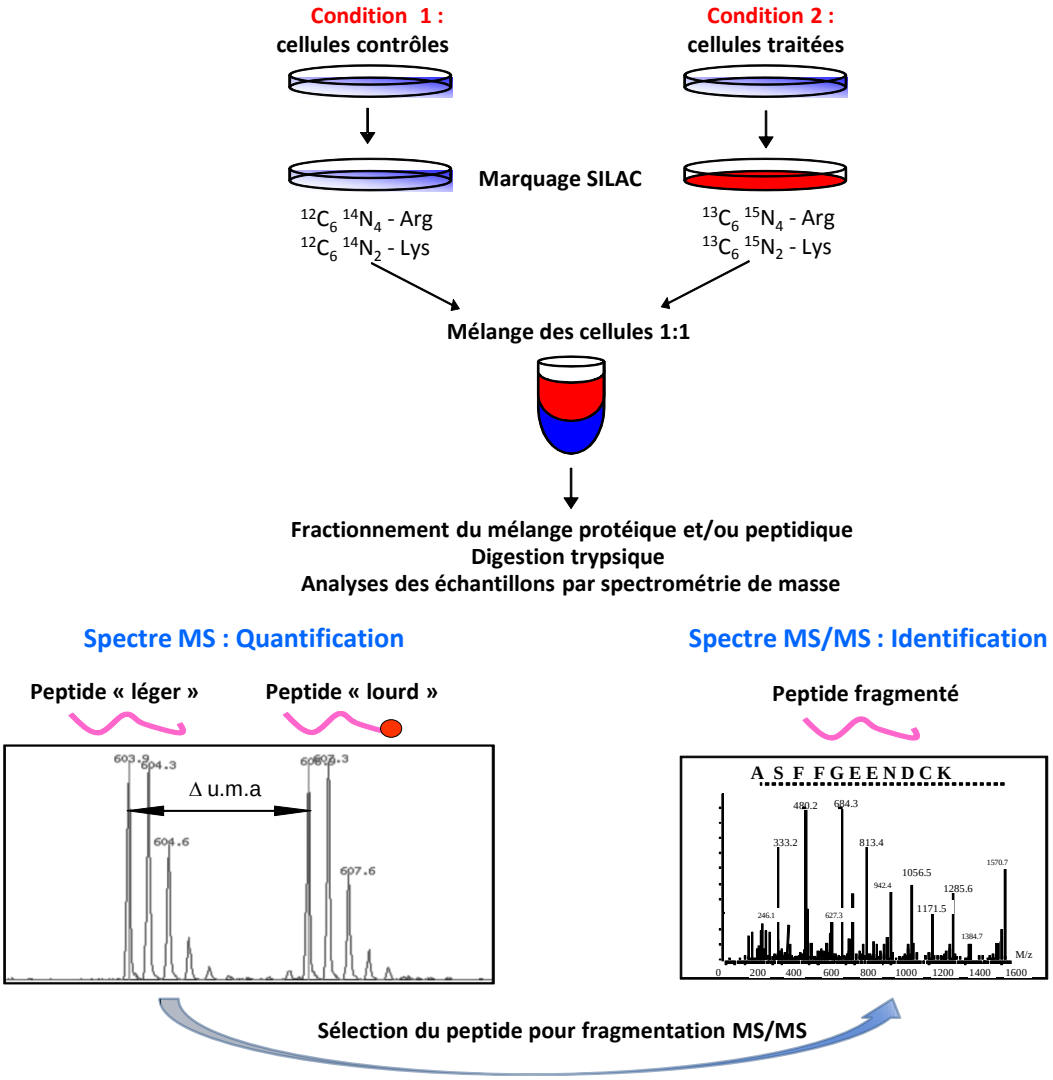


Figure 16 : principe de la stratégie de marquage SILAC.
Deux lots de cellules sont cultivés en présence d'acides aminés lourds ou légers. L'un des lots à comparer est traité (stimulation, stress, siRNA…) et les cellules sont mélangées dans un rapport 1 :1. Les protéines ainsi marquées sont fractionnées, puis digérées à la trypsine avant l'analyse par spectrométrie de masse. Les couples de peptides lourds et légers sont quantifiés par extraction du signal MS et identifiés par interprétation des données MS/MS.

Dans le cas du SILAC (cf figure 16), cet élément correspond à un ou plusieurs acides aminés pour lesquels certains atomes ont été remplacés par une forme plus lourde mais stable. Ainsi, au cours de la croissance cellulaire et du « turn-over » protéique, les acides aminés marqués sont incorporés dans toutes les chaînes de protéines nouvellement synthétisées par la cellule. En pratique on utilise généralement un double marquage des lysines et des arginines (Graumann, Hubner et al. 2008), assurant ainsi à l'ensemble des produits issus du clivage trypsique d'une protéine au moins un acide aminé marqué. L'écart de masse entre la forme légère et lourde d'un peptide donné est fonction de l'incrément de masse du marqueur présent dans sa séquence (+6Da pour une lysine marquée $^{13}C_6{}^{15}N_2$ et +10 Da pour une arginine marquée $^{13}C_6{}^{15}N_4$) et du nombre total de marqueurs présents, nombre qui est normalement égal à 1 si le peptide ne présente pas de coupure manquée. En utilisant des marquages distincts, il est possible par cette technique de comparer près de trois conditions différentes (de Godoy, Olsen et al. 2006).
La quantification relative est réalisée en comparant les intensités des massifs isotopiques des paires de peptides dans l'ensemble du spectre MS. Contrairement au marquage métabolique ^{15}N, le nombre de marqueurs incorporés par cette méthode SILAC est défini et ne dépend pas de la longueur du peptide, ce qui facilite largement l'analyse des données.

Le principal avantage des stratégies de marquage métabolique réside dans le fait que les échantillons traités différentiellement sont combinés en amont des traitements biochimiques nécessaires pour réaliser l'analyse protéomique. Ceci exclut donc de nombreux biais introduits par ces différentes étapes et permet d'obtenir des données quantitatives présentant un plus faible taux d'erreur. Cette technique est particulièrement bien adaptée aux cellules immortalisées dont la majorité arrive à incorporer plus de 97% du marquage (Ong and Mann 2006) après 5 cycles de division cellulaire. Elle ne permet cependant pas de quantifier des tissus cliniques ou des fluides biologiques.

L'approche SILAC a remporté un franc succès ces dernière années et de nombreux travaux très intéressants sont retrouvés dans la littérature (Graumann, Hubner et al. 2008; Hanke, Besir et al. 2008; Kruger, Moser et al. 2008). Cette technique a été utilisée pour détecter de faibles variations quantitatives ainsi que pour la caractérisation et la dynamique des modifications post traductionnelles (Blagoev, Kratchmarova et al. 2003). Cette stratégie est parfaitement adaptée à l'étude quantitative des profils d'expression protéiques à grande échelle, mais elle est désormais également utilisée avec succès pour l'étude d'interactomes afin de distinguer les vrais partenaires protéiques, des faux positifs (Selbach and Mann 2006; Vinther, Hedegaard et al. 2006; Dengjel, Kristensen et al. 2008; Trinkle-Mulcahy, Boulon et al. 2008).

Les techniques sans marquage

Bien que très performantes les approches utilisant un marquage isotopique présentent plusieurs limitations, comme le nombre maximal d'échantillon pouvant être comparés (au mieux 8 échantillons pour un marquage iTRAQ 8plex), leur coût, les contraintes liées à l'introduction du marqueur. Certaines études, telle que l'analyse de grandes séries d'échantillons cliniques, ne peuvent pas être menées avec ce type de méthodologie. Ainsi, très récemment des approches alternatives ont émergées, elles sont réalisées sans marquage ou « label free » et sont devenues une nécessité pour réaliser des études quantitatives à grande échelle.

L'identification des protéines par spectrométrie de masse ne renseigne pas directement sur leur abondance dans un mélange. En effet, les scores d'identification fournis par les divers logiciels après recherche dans les banques de données ne sont pas directement corrélés à la quantité de protéine analysée (Ong and Mann 2006). Cela est principalement lié au fait que le score des peptides fournis par les algorithmes des banques de données est basé sur la similarité entre les données de fragmentation expérimentales et celles générées *in silico*. Par conséquent, l'identification d'un peptide de faible abondance présentant un signal faible mais donnant naissance à un spectre MS/MS de qualité peut conduire à une identification sans ambiguïté de sa séquence lui attribuant ainsi un score important (Steen and Mann 2004). A ce titre, l'utilisation des scores d'identification des protéines comme mesure de la quantité et de l'abondance des protéines n'est pas envisageable. D'autres paramètres plus appropriés relatifs au principe de l'analyse par spectrométrie de masse en mode DDA (cf partie I-3.3) peuvent donner accès à des informations quantitatives. En effet, dans ce mode d'acquisition, le spectromètre sélectionne automatiquement les espèces détectées les plus intenses pour les séquencer en MS/MS. Il a ainsi été démontré qu'il existe une corrélation entre l'abondance d'une protéine au sein d'un échantillon et le nombre d'événements MS/MS que l'on peut observer pour cette même protéine (Liu, Sadygov et al. 2004). Ainsi une quantification relative peut être effectuée en comparant le nombre de ces spectres MS/MS acquis pour une protéine donnée entre deux séries d'expériences. Il est cependant important de noter qu'une protéine de grande taille produira un nombre plus élevé de spectres MS/MS qu'une petite protéine. Il est possible d'appliquer un facteur pour corriger ce biais et ainsi obtenir une valeur semi-quantitative absolue (Florens, Carozza et al. 2006) mais cette correction n'a pas d'utilité dans le cas d'une quantification relative.

L'analyse par « spectral counting », même si elle constitue une méthode simple pour réaliser une quantification sans marquage, souffre cependant d'un certain nombre de limitations :

- elle n'est valide que pour une gamme de concentration limitée,
- les protéines présentes en faible quantité dans un échantillon ne sont pas systématiquement séquencées et les plus minoritaires peuvent ne pas être du tout séquencées.

Cette méthode ne fournit donc qu'une réponse partielle par rapport à celle que l'on pourrait potentiellement obtenir à partir des données acquises. Pour creuser en profondeur la dynamique du protéome il est nécessaire d'utiliser une stratégie pouvant tirer parti de l'ensemble de l'information disponible, c'est-à-dire basée sur l'analyse des signaux MS présents dans les fichiers bruts. Cependant, la complexité des données LC-MS et la faible résolution des appareils de type ion-trap ne permettait pas à l'époque de réaliser des comparaisons sans marquage avec une fiabilité suffisante (un trop grand nombre de faux positifs pouvant être généré lors de la comparaison de masses peptidiques). Le développement d'instruments à haute-résolution de type LTQ-Orbitrap a constitué une étape nécessaire à la mise au point de cette nouvelle méthodologie sans marquage ou « label free ». Celle-ci suscite un grand intérêt pour sa facilité de mise en œuvre car elle ne nécessite pas de modifier les échantillons, et elle permet d'obtenir une quantification fiable des protéines contenues dans un mélange, tout en réduisant considérablement les coûts de l'analyse (Olsen, Nielsen et al. 2007). Elle est de plus en plus utilisée et s'applique à de nombreux domaines en biologie (Bodenmiller, Wanka et al. 2010; Luber, Cox et al. 2010; Bildl, Haupt et al. 2012).

Une acquisition LC-MS peut être vue comme une carte constituée de l'ensemble des spectres MS générés par l'instrument. Cette carte dite LC-MS correspond à un espace comprenant trois dimensions (cf figure 17) : le temps d'élution (x), le rapport m/z (y) et l'intensité mesurée (z).

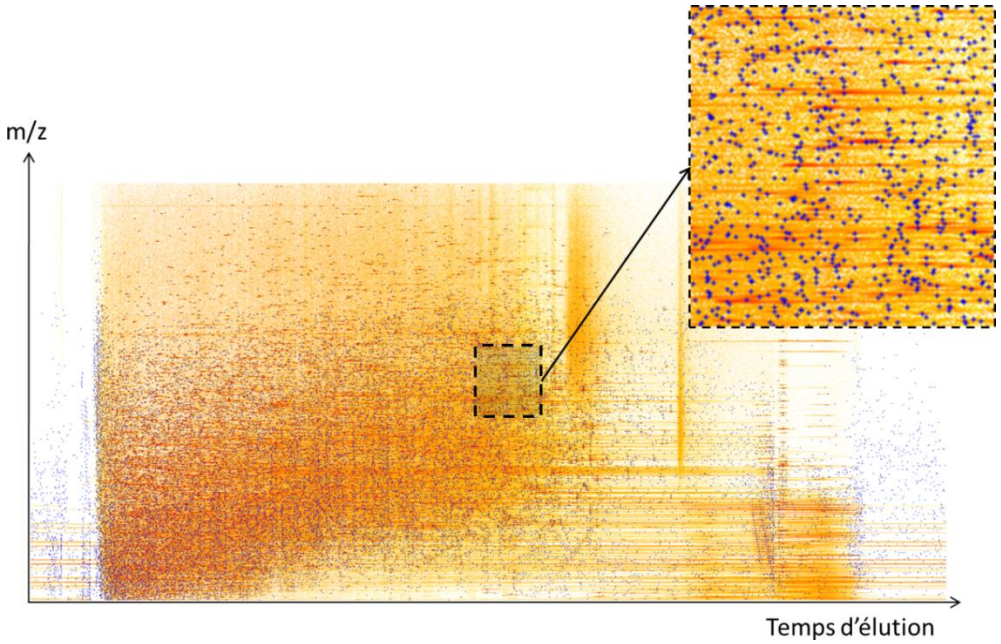


Figure 17 : image générée par MsInspect représentant une carte LC-MS. Le carré en pointillé en haut à droite montre avec un agrandissement supérieur un aperçu de la complexité des données de la carte. Les points en bleu sont censés correspondre à la masse monoisotopique des ions peptidiques.

Plusieurs approches sont possibles pour réaliser l'analyse des données MS :

- approche non supervisée : il s'agit d'essayer de détecter des signaux peptidiques à partir d'une carte LC-MS (cf figure ci-dessus). Cette reconnaissance met en œuvre dans un premier temps des algorithmes de détection de pics (« peak picking »), puis les différents pics pouvant correspondre à un même ion peptidique sont regroupés, à la fois sur l'échelle des m/z (différents isotopes d'un massif peptique, pour chaque état de charge d'un peptide) et sur l'échelle des temps (massifs isotopiques détectés sur différents spectres MS consécutifs, tout au long de l'élution d'un peptide). Ce processus est basé sur la comparaison des données expérimentales avec les modèles théoriques connus de distribution isotopique et d'élution chromatographique des peptides. Le but de l'analyse est de fournir une liste d'empreintes (ou « features »), correspondant à l'ensemble des signaux regroupés pour un même ion peptidique, avec les coordonnées correspondantes. L'identification des peptides peut être déterminée en réalisant une identification ciblée lors d'une seconde acquisition ou bien à partir d'une banque de données décrivant pour un ensemble de peptides identifiés au préalable des informations telles que la séquence, la masse et le temps d'élution. Cette deuxième méthode est décrite dans la littérature sous le terme de « Accurate Mass and Time tags » ou AMT (Smith, Anderson et al. 2002).
- approche supervisée : on connaît (ou suppose) les coordonnées (x,y) des signaux peptidiques à extraire. Dans une expérience LC-MS, l'intensité du signal MS d'un peptide qui élue de la colonne chromatographique peut être suivie au cours du temps (cf figure 18). L'aire sous la courbe du pic chromatographique est le courant ionique extrait (XIC, eXtracted Ion Current ou encore eXtracted Ion Chromatogram) et elle est proportionnelle à l'abondance du peptide dans

l'échantillon. Il a en effet été démontré que le XIC était linéairement dépendant de la quantité du peptide (Ong and Mann 2005). L'analyse du signal consiste donc à extraire l'intensité du signal situé à un emplacement précis sur la carte LC-MS et à fournir le XIC correspondant.

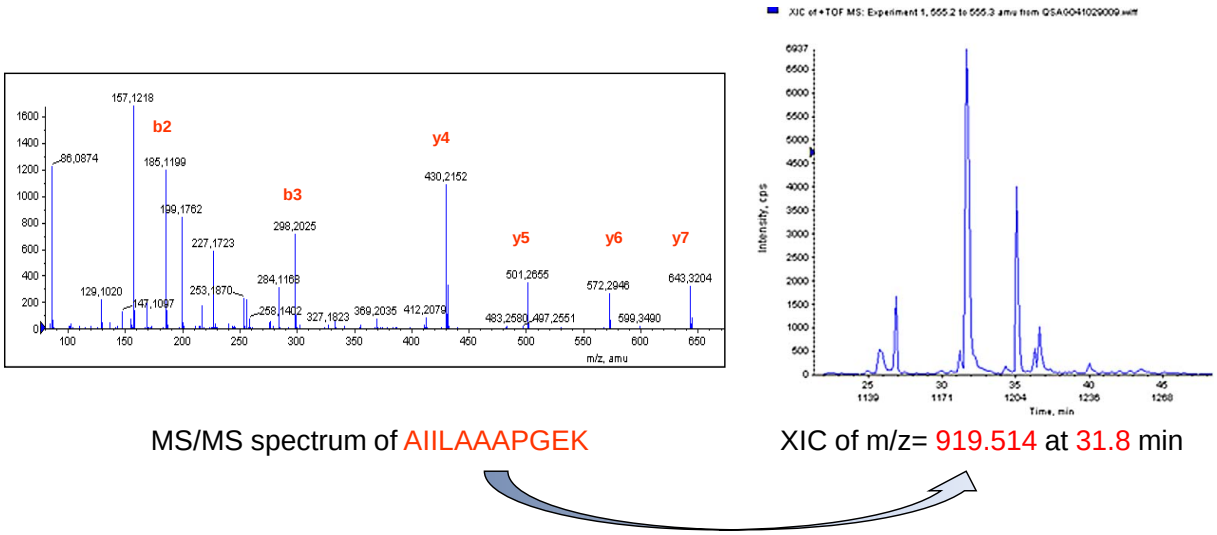


Figure 18 : extraction du signal MS d'un peptide qui a été identifié au préalable par un moteur de recherche.

La première approche est plus exhaustive que la seconde car elle peut fournir des informations quantitatives pour des espèces peptidiques qui n'ont pas été nécessairement fragmentées par le spectromètre de masse. En ce qui concerne la deuxième approche, on peut supposer que la connaissance précise de la masse monoisotopique des peptides peut diminuer le risque d'erreur de quantification, mais à l'heure actuelle et à ma connaissance aucune étude ne l'a encore démontré.

Quelle que soit la stratégie employée pour réaliser l'extraction du signal il est nécessaire d'apparier les signaux extraits dans le contexte d'une analyse quantitative comparative (cf figure 19). Pour cela, les cartes LC-MS doivent être au préalable alignées de manière à corriger les variabilités d'élution chromatographique des peptides. En effet, le temps d'élution d'un peptide donné peut varier de plusieurs dizaines de secondes d'une analyse LC-MS à une autre. Même si la masse du peptide peut être mesurée avec une très grande précision il arrive que des peptides de masses très proches soit élués dans un même laps de temps. La figure 17 montre à quel point la densité des mesures est importante. Par conséquent la comparaison de cartes LC-MS dont l'échelle de temps n'aurait pas été corrigée au préalable génèrerait un grand nombre de mauvais appariements.

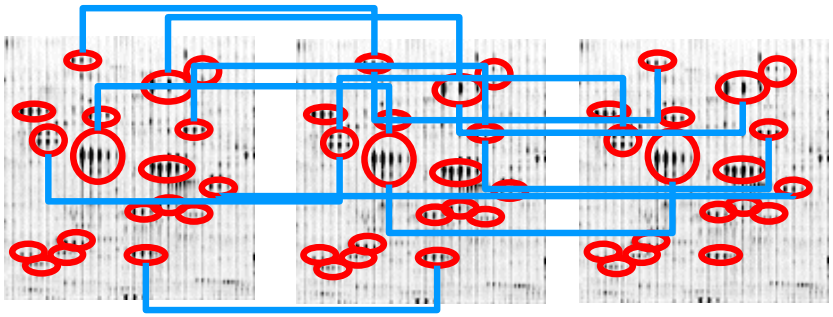


Figure 19 : mise en correspondance des espèces détectées sur plusieurs cartes LC-MS.

Différents algorithmes existent pour réaliser cette correction et ils sont en général optimisés pour le type d'approche employée. En effet, la méthode supervisée bénéficie d'une information de séquence qui permet d'identifier précisément les molécules et donc de réaliser un alignement avec un taux d'erreur réduit. Je ne détaillerai pas l'ensemble des méthodes qui ont été décrites dans la littérature mais j'apporterai dans les parties suivantes des explications précises sur les algorithmes d'alignement que j'ai été amené à développer au cours de mon doctorat.

En résumé, l'analyse des données LC-MS/MS dont nous venons de détailler les principes peut être divisée en trois grandes approches (cf figure 20) :

- l'extraction d'une paire de signaux MS détectés au sein d'une même analyse dans le cadre des stratégies mettant en œuvre un marquage isotopique,
- le comptage et la comparaison du nombre de spectres de fragmentation (MS/MS) des peptides d'une protéine donnée détectée dans des analyses parallèles (« spectral counting-based label-free quantification »),
- l'extraction, l'alignement et la comparaison des intensités de signal MS d'un même peptide détectée dans des analyses parallèles («MS intensity-based label free quantification »).

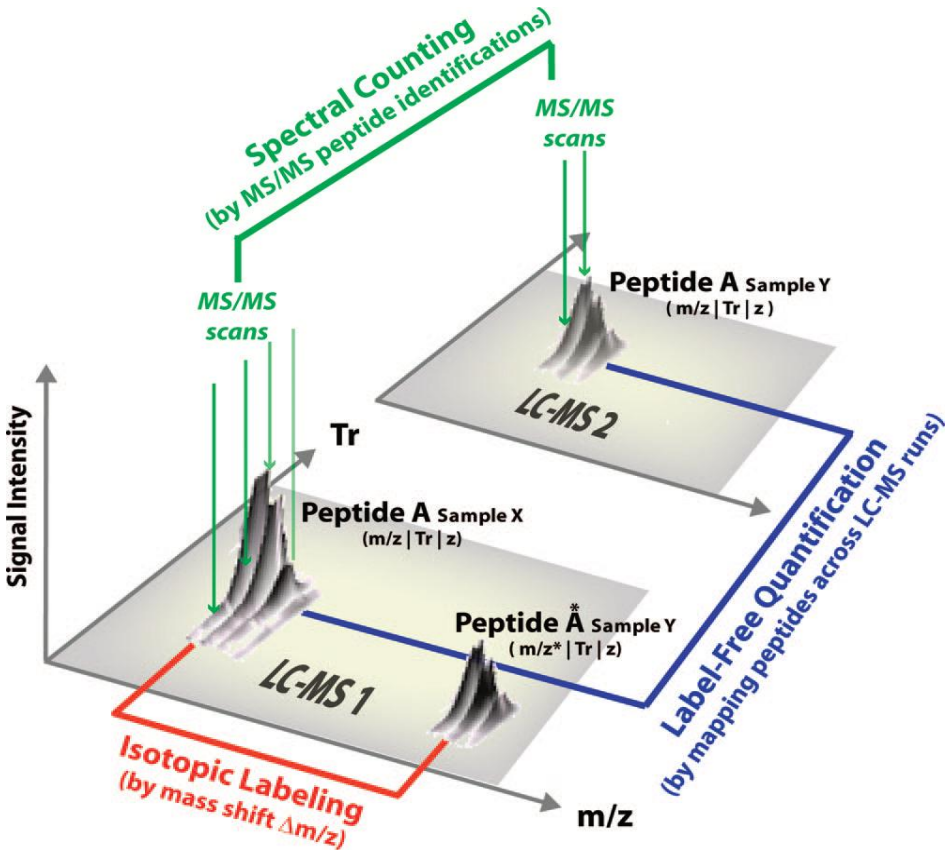


Figure 20 : vue d'ensemble de différentes approches d'analyse quantitative par LC-MS/MS.
D'après (Mueller, Brusniak et al. 2008)

Le principe du traitement de données issues d'expériences de quantification dans des approches de marquage isotopique est quasiment identique quelle que soit la stratégie utilisée (ICAT, SILAC…). En effet, l'évaluation de l'abondance relative des peptides est réalisée en comparant les signaux MS des deux formes légères et lourdes d'un même peptide. La différence majeure entre ces stratégies réside

dans la valeur de l'écart de masse existant entre les peptides marqués. Le traitement informatique consiste alors en l'extraction des signaux correspondant aux paires isotopiques, les peptides de chaque doublet étant repérés par leur différence de masse caractéristique.
Les approches sans marquage présentent l'avantage d'être moins fastidieuses à mettre en œuvre en amont car elles ne requièrent justement pas d'étape de marquage. De plus elles peuvent s'appliquer à tout type d'échantillon. Toutefois elles comportent quelques biais car des erreurs de quantification peuvent être introduites lors des différentes étapes de préparation nécessaires à l'analyse par spectrométrie de masse, réalisées en parallèle sur les échantillons. Par ailleurs, l'analyse par « spectral counting » n'est applicable en pratique que pour une partie de l'échantillon, constituée par les protéines les plus abondantes. Enfin, dans le cas de l'analyse « label free » des signaux MS, un retraitement important des données après acquisition est requis. Il est en effet nécessaire de corriger les données extraites des erreurs systématiques et des biais liés à l'injection séparée des peptides à comparer et aux performances intrinsèques de la méthode analytique : reproductibilité de la séparation en LC (variabilité sur le temps d'élution), de la réponse en signal MS du spectromètre de masse (variabilité sur l'intensité), de sa calibration et de sa résolution (variabilité sur la masse mesurée). Ces corrections incluent donc la normalisation des signaux, l'alignement en temps des cartes LC-MS, ou encore la calibration en masse des espèces détectées. De nombreux outils bioinformatiques sophistiqués ont donc été développés ces dernières années pour réaliser ce type d'opérations et permettre ainsi l'analyse de ces jeux de données très complexes.

I-6.2. L'apparition d'outils informatiques dédiés à l'analyse des signaux MS

L'exploitation des données de protéomique quantitative, au même titre que la validation des résultats d'identification, a nécessité le développement de logiciels adaptés.
Historiquement, ce sont les méthodes quantitatives MS basées sur un marquage isotopique qui ont été les premières disponibles car elles étaient compatibles avec des mesures de faible résolution. L'automatisation de la quantification a été rendue possible via certains logiciels tels que XPRESS (Han, Eng et al. 2001) et ASAPRatio (Li, Barshick et al. 2003) qui se basent sur l'identification de peptides par MS/MS pour reconstituer des XIC des deux peptides lourd et léger. Ces outils étaient cependant dépendants au départ du moteur de recherche Sequest et il a fallu attendre le développement du logiciel msQuant (Schulze and Mann 2004) pour disposer d'un outil prenant en charge les données d'identification Mascot.
Ces logiciels de première génération étaient pour la plupart difficiles à prendre en main et pouvaient présenter des problèmes de compatibilité à différents niveaux : format des données brutes (différents suivant le type d'instrument utilisé), système d'exploitation utilisé (windows, Linux…), moteurs de recherche employés (Mascot, Phenyx, Sequest...) ou encore des méthodes de marquage utilisées (ICAT, SILAC, iTRaQ…). Comme exposé dans la suite de ce document, c'est dans ce contexte qu'a débuté le développement du logiciel MFPaQ, avec pour objectif d'offrir une solution efficace et interactive pour la validation et la quantification des analyses protéomiques basées sur des marquages isotopiques.
Peu de temps après, l'apparition des instruments haute-résolution et à haute vitesse de séquençage s'est accompagnée d'efforts de plus en plus importants pour fournir des outils bioinformatiques adaptés au traitement de données de plus en plus volumineuses, mais également de meilleure qualité. La haute résolution a notamment été l'élément déclencheur de la mise au point d'algorithmes réalisant un traitement non supervisée des données MS, donnant ainsi accès à la quantification d'espèces non séquencées et donc à une plus grande profondeur du protéome étudié.

Ce type d'algorithme appliqué à l'analyse des données par marquage isotopique a été implémenté dans le logiciel MaxQuant (Cox and Mann 2008). Celui-ci a été développé par le groupe de Mathias Mann à l'institut MaxPlanck et il constitue la référence dans le domaine au moment de la rédaction de ce manuscrit. On peut cependant noter qu'il n'est compatible qu'avec les instruments vendus par le constructeur Thermo Scientific, ce qui limite clairement son utilisation.

En parallèle, et comme nous l'avons vu, les méthodes sans marquage d'analyse des signaux MS ont émergé donnant naissance à tout un panel d'outils informatiques dédiés à ce type d'analyse (http://www.ms-utils.org). La plupart d'entre eux sont d'ailleurs basés sur une approche non supervisée qui consiste à générer puis à comparer des cartes LC-MS. On peut citer parmi les plus connus SuperHirn (Mueller, Rinner et al. 2007), MSInspect (Bellew, Coram et al. 2006), OpenMS (Sturm, Bertsch et al. 2008), Decon2LS (Jaitly, Mayampurath et al. 2009), ou le logiciel commercial Progenesis LC-MS. Bien qu'ils offrent une solution attractive pour l'analyse exhaustive des données disponibles, ces algorithmes basés sur la détection d'empreintes peptidiques et l'alignement de cartes LC-MS nécessitent des temps de calcul importants, ce qui rend encore pour l'instant difficile l'analyse de grandes séries de fichiers. Par ailleurs, la mise en relation des empreintes peptidiques détectées avec les données d'identification est plus ou moins bien gérée en fonction des logiciels. Enfin, dans la mesure où les cartes LC-MS sont souvent générées de façon individuelle, en utilisant des valeurs seuil pour la reconnaissance des empreintes peptidiques, les signaux de faible intensité sont souvent détectés de façon non-reproductible dans différents échantillons, ce qui entraîne la présence de nombreuses valeurs manquantes et complique l'analyse statistique des données.

D'un autre côté, il est intéressant de noter que certains autres outils ont continué à implémenter une approche supervisée pour la quantification des données « label-free » (comme celle utilisée à l'origine dans XPRESS), en partant des données d'identification et en extrayant les valeurs de XIC pour chaque pic d'élution peptidique. Cette méthode, qui est en principe plus simple et rapide, est par exemple utilisée dans Serac (Old, Meyer-Arendt et al. 2005), Quoil (Hoffert, Wang et al. 2007), Ideal-Q (Tsou, Tsai et al. 2010), ou dans le logiciel commercial PeakView (AB Sciex), et c'est également celle que nous avons choisi de mettre en œuvre dans MFPaQ, comme présenté dans la partie III de ce document.

La description de l'ensemble des logiciels existants dépasse le cadre de ce manuscrit et la plupart d'entre eux ont déjà été recensés et présentés dans plusieurs revues (Mueller, Brusniak et al. 2008; Zhu, Smith et al. 2010). On peut cependant résumer en disant qu'ils diffèrent en fonction de leur compatibilité avec les systèmes d'exploitation (Windows, Apple OS X ou Linux), des formats de fichiers bruts (RAW, Wiff, mzXML), des moteurs de recherche (Sequest, Mascot, Phenyx, X!Tandem), du type de marquage (iTraQ, ICAT, SILAC, label free) et de l'approche d'analyse du signal MS. Dans la partie IV de ce manuscrit, l'évaluation de certains de ces outils est présentée, via l'utilisation du logiciel Prosper qui permet d'intégrer et de comparer les résultats de quantification issus de plusieurs logiciels.

MFPaQ version 3 : développement d'un outil pour la validation, l'organisation, et la quantification des données de nanoLC-MS/MS

Nous allons décrire dans cette partie le logiciel MFPaQ version 3 dont le développement a été l'un des premiers enjeux de mon travail de thèse. Nous détaillerons notamment les besoins associés à cet outil ainsi que quelques éléments de conception. Cette partie se terminera par la publication de référence de ce logiciel (Bouyssie, Gonzalez de Peredo et al. 2007), contenant une application à l'étude de la dynamique du protéome membranaire de cellules endothéliales humaines suite à l'effet de facteurs pro-inflammatoires.

II-1. Validation des résultats d'identification

L'objectif principal du développement de ce logiciel était d'automatiser le processus de validation des résultats d'identification du moteur de recherche Mascot. Comme nous l'avons vu la quantité de données générées par ce type de logiciel ne rend pas possible une vérification manuelle de l'ensemble des interprétations de spectres MS/MS. C'est d'ailleurs pour cette raison qu'un ensemble de méthodes de validation des résultats, que nous avons évoquées précédemment, a été développé au sein de différents laboratoires au cours des dix dernières années. Parmi les travaux précurseurs dans le domaine, on compte notamment les méthodes implémentées dans les outils PeptideProphet en 2002 puis ProteinProphet en 2003, dédiés à la validation des résultats du moteur de recherche Sequest. Entre 2003 et 2006, face à la montée en puissance des études protéomiques à grande échelle, se développe également une préoccupation grandissante concernant la qualité des jeux de données publiés, et les méthodes de validation appliquées pour les générer. Ces questions feront l'objet de nombreux débats au sein de la communauté protéomique, aboutissant à la création de groupes de discussions chargés de définir des standards en termes de processus expérimentaux et de formats de résultats, et à la mise en place en 2006 de « guidelines » relativement strictes pour la publication des données d'identification dans les journaux de protéomique. Cependant, au début de ma thèse en 2006, les méthodes de validation existantes ne disposaient pas encore d'implémentations utilisables sur la plupart des plateformes logicielles, et d'autres approches n'avaient pas encore été décrites (comme les méthodes target-decoy réellement démocratisées en 2007). C'est dans ce contexte que j'ai donc commencé par concevoir et implémenter une méthode à la fois simple et efficace pour résoudre cette problématique, le logiciel MFPaQ (Mascot File Parsing and Quantification).

Mon premier travail a consisté à mettre un point un lecteur des données du moteur de recherche Mascot. Pour y parvenir j'ai utilisé la bibliothèque « msparser » qui est fournie avec sa documentation par Matrix Science. Ce module permet d'accéder facilement au contenu des fichiers .dat du moteur de recherche. Ces derniers contiennent l'ensemble des résultats d'identifications :

masses, séquences, scores peptidiques et protéiques, ainsi que toutes les autres informations qui sont affichées dans l'interface de Mascot.

Afin de pouvoir effectuer plusieurs validations sur un même jeu de donnée des résultats dans un temps raisonnable il m'est apparu judicieux d'extraire les informations contenues dans un fichier .dat et de le stocker dans un autre format qui pourrait être chargé plus rapidement. Le format de fichier XML (eXtensible Markup Language) qui permet de définir un schéma structuré et hiérarchisé d'un ensemble d'informations sous forme de document me paraissait particulièrement bien adapté à cette problématique. Ainsi, après avoir établi la liste des informations Mascot nécessaires à l'étape de validation des données, j'ai alors réalisé un schéma XML définissant la structure du fichier de données. Le langage XML a d'ailleurs par la suite été largement plébiscité par la communauté protéomique. En effet le groupe PSI (Proteomics Standards Initiative, http://www.psidev.info/) l'a choisi pour implémenter l'ensemble de ses standards de fichiers protéomiques. Dans la version la plus récente du logiciel, le support de représentation des données a finalement été remplacé par des fichiers au format SQLite afin d'améliorer les performances de l'outil.

Une fois la structure des données définie, j'ai concentré mon travail sur le développement d'un ensemble de règles empiriques pour filtrer les résultats d'identification. L'établissement de ces règles a été effectué en collaboration avec les ingénieurs et chercheurs du laboratoire à partir des connaissances et des constats suivants :

- le score peptidique de Mascot est un bon indicateur de la qualité de l'interprétation du spectre MS/MS,
- à partir d'une valeur seuil de probabilité il est possible de demander au moteur de recherche de fournir une valeur de seuil de score correspondante,
- chaque spectre MS/MS se voit attribuer un certain nombre d'interprétations de séquences peptidiques différentes classées par ordre de score décroissant. Il est ainsi possible de savoir si une séquence donnée correspond ou non à la meilleure interprétation du spectre correspondant,
- une séquence peptidique de petite taille (<7 AA) a plus de chance de correspondre à un faux positif,
- il y a une corrélation directe entre le nombre de peptides attribués à une protéine et la confiance que l'on peut avoir sur l'identification de cette dernière.

Ces différents éléments ont donc servi à élaborer des règles pour la validation des identifications protéiques : une protéine peut être validée si elle comporte au moins **P** peptides de longueur de séquence supérieure ou égale à **L**, de score supérieur ou égal à **S**, et au plus de rang **R.** Le logiciel est capable de gérer une combinaison maximale de 3 différentes règles (pour **P**=1, 2 ou 3) et par défaut le logiciel est configuré pour fonctionner avec une combinaison de 2 règles (cf figure 21).

Il est aujourd'hui admis dans la littérature qu'une taille minimale de séquence peptidique comprise entre 6 et 8 acides aminés donne les meilleurs résultats (Elias and Gygi 2007). En ce qui concerne le seuil de score, celui-ci peut être déterminé de façon plus ou moins stringente en fonction d'un seuil de p-value fourni au moteur de recherche (valeur seuil moyenne pour l'ensemble des peptides identifiés). Au sein du laboratoire, nous avons déterminé de manière empirique au début de ma thèse que nous obtenions des validations satisfaisantes pour des p-value p1=0.001 et p2=0.05 (permettant de définir respectivement les scores Mascot appliqués pour les règles à 1 et 2 peptides).

Ces valeurs ont été validées plus tard lorsque les méthodes de recherche en banques inverses (cf partie I-5.2 sur la stratégie « target-decoy ») furent démocratisées. En effet, nous avons pu vérifier qu'un FDR (False Discovery Rate) autour de 1% était obtenu lors de l'utilisation de ces paramètres dans de nombreux projets d'analyse de mélanges complexes de protéines. Ainsi, l'approche multi-règles simple et intuitive implémentée au départ dans MFPaQ (si une protéine est identifiée avec peu de peptides, les seuils de score appliqués sur ces peptides doivent être plus élevés), a pu trouver une validation expérimentale grâce aux méthodes target-decoy.

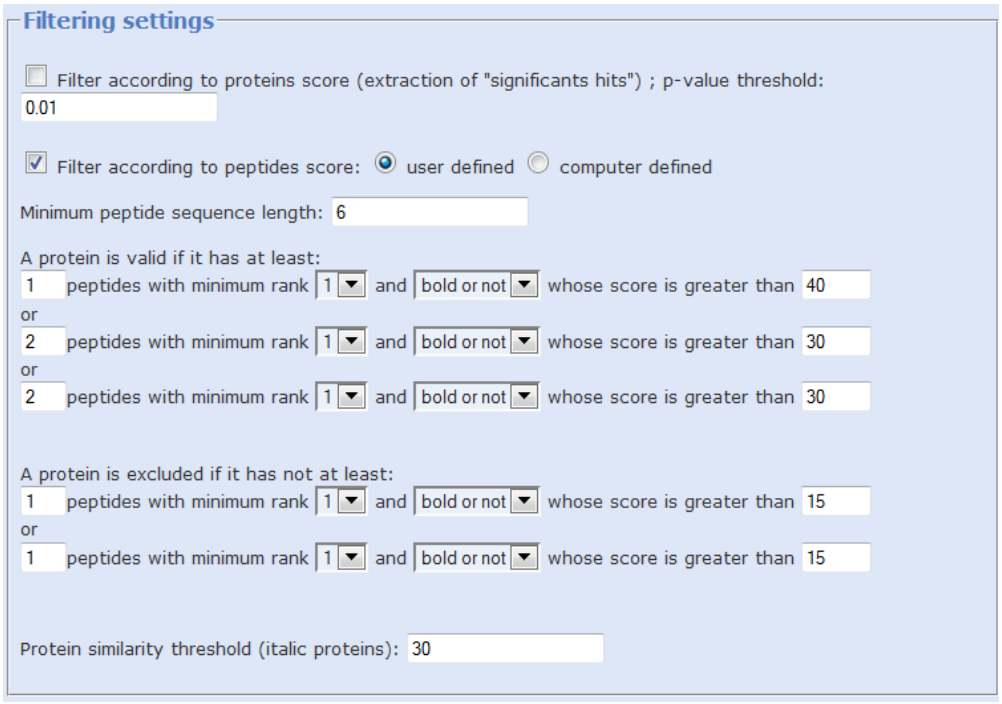


Figure 21 : capture d'écran de la fenêtre MFPaQ qui permet de régler les paramètres de validation.

Dès le début 2008, le logiciel a été en effet capable de tirer parti des analyses en banque inverse grâce à l'implémentation d'un algorithme qui détermine de façon automatique les critères des règles permettant d'obtenir un FDR voulu. Celui-ci fonctionne de la manière suivante :

1) initialisation d'une p-value de départ (par défaut 0.1)
2) calcul des filtres correspondant à la p-value actuelle
3) validation des protéines à partir des filtres actuels
4) calcul du FDR = nb_protéines_decoy/nb_protéines_target
5) si FDR > FDR voulu => étape 2 avec p_value = 80% ancienne_p_value ; sinon fin du processus

Cette validation peut être réalisée sur un fichier unique ou sur un ensemble de fichiers .dat regroupés en un jeu de données appelé « expérience ». Dans ce cas, le FDR estimé est une moyenne calculée sur l'ensemble des fichiers. La réalisation de ce type d'opérations est un besoin lié au traitement de données issues d'analyses de type « shotgun », où l'échantillon subit un fractionnement avant son analyse en spectrométrie de masse. La partie suivante détaille l'architecture qui a été développée afin d'effectuer des traitements par lot sur les jeux données.

II-2. Gestion des expériences « shotgun »

Afin de maximiser la couverture d'identification des protéines identifiées pour un échantillon donné la séparation de ce dernier en amont de son analyse par nanoLC-MS/MS s'est avérée être une méthode efficace. Cependant, ce fractionnement engendre une fragmentation des jeux de données produits. En effet, chaque fraction correspond à un run LC-MS/MS, qui est à l'origine d'un ensemble de spectres MS/MS (peaklist), donnant lui-même naissance à un fichier résultat à l'issue de la recherche en banque de données.

Une possibilité pour simplifier l'analyse informatique de multiples peaklists est de produire une peaklist unique pour l'ensemble des acquisitions LC-MS/MS. Cette opération de concaténation est en général intégrée aux outils de gestion des tâches des moteurs de recherche (Mascot Daemon par exemple). La peaklist ainsi générée ne produit plus qu'un seul fichier résultat qui peut ensuite être traité facilement par des outils de validation. Cette approche présente deux principaux inconvénients dans le cas d'un fractionnement par gel SDS-page. Premièrement, il est possible de perdre de l'information liée à la migration de la protéine sur le gel. En effet une protéine donnée peut exister sous plusieurs formes (variants de maturation, avec ou sans glycosylation…) pouvant être séparées sur différentes bandes du gel. Cette information peut être obtenue si l'on analyse les différentes fractions individuellement mais pas dans le cas d'une concaténation des spectres MS/MS. Deuxièmement, l'obtention d'un unique fichier résultat peut compliquer la quantification des données LC-MS, car il est plus difficile dans ce cas de relier un spectre MS/MS donné au fichier d'acquisition dont il est issu. Pour y parvenir, il est impératif que cette information soit présente dans l'en-tête de chacun des spectres de la peaklist unique. Il est aussi nécessaire d'extraire cette information selon le format de cet en-tête, qui peut varier d'un logiciel de génération de peaklist à un autre.

Ainsi, pour ne perdre aucune information le logiciel MFPaQ a été développé pour être capable de gérer la notion de fractionnement de l'échantillon et même d'en tirer parti. Il est par exemple possible de connaître les bandes du gel dans lesquelles une protéine donnée a été identifiée mais également quelle est la bande où elle est le plus représentée (meilleur score). Cette information est aussi accessible au niveau du module de quantification du logiciel (cf partie II-3). L'intégration des résultats d'identification depuis différentes fractions d'un même échantillon a augmenté la complexité de l'architecture des données au sein du logiciel. Il a été nécessaire d'introduire deux structures de données spécifiques à cette problématique :

- Un jeu de données appelé « expérience » qui représente un ensemble de fichiers résultats relatifs au même échantillon (piste de gel, fractions SCX, spots de gel 2D). Chaque expérience peut ensuite être manipulée facilement au sein de l'interface du logiciel pour effectuer des traitements par lot sur l'ensemble des fractions correspondantes (validation des résultats, statistiques, exports…).
- Un jeu de données appelé « liste de protéines » qui correspond au regroupement non-redondant des résultats sur un ensemble de fichiers donnés. Une des fonctions du logiciel est la possibilité de produire une telle liste à partir d'une expérience donnée. Ces listes permettent d'avoir une vision globale du protéome associé à un échantillon donné tout en conservant l'information relative au fractionnement utilisé.

Comme nous l'avons décrit dans la partie I-4.5, plusieurs méthodes sont possibles pour effectuer le regroupement d'un ensemble de protéines identifiées dans différents fichiers. En général, les moteurs de recherche utilisent des algorithmes qui suivent une approche dite « parcimonieuse », celle-ci étant aujourd'hui approuvée par l'ensemble de la communauté protéomique. Cependant si on applique ce type d'algorithme pour regrouper des identifications issues d'une séparation protéique, une information importante ne sera pas prise en compte : le fait qu'une protéine donnée peut exister sous différentes formes distinctes (séquence protéique plus ou moins longue par exemple). Ainsi pour maximiser la conservation de l'ensemble de l'information obtenue lors d'une analyse « shotgun » j'ai décidé de mettre au point un autre algorithme pour le regroupement des protéines identifiées. Cet algorithme se base sur la création de « clusters » de groupes protéines à partir d'une fonction de « clusterisation ». Cette fonction recherche les groupes de protéines provenant des différents fichiers résultats qui présentent au moins une séquence protéique en commun. L'inconvénient lié à cette approche est la possibilité de produire « une clusterisation en chaîne » pouvant aboutir au regroupement de deux groupes distincts (i.e. n'ayant aucune séquence protéique en commun) mais ayant au moins un point commun avec un même troisième groupe. A la fin de l'opération on obtient une liste de « clusters » protéiques faisant référence aux groupes de protéines des fichiers d'identification individuels. L'ensemble des « clusters » constitue ce que nous avons dénommé une liste de protéines non redondante (cf figure 22).

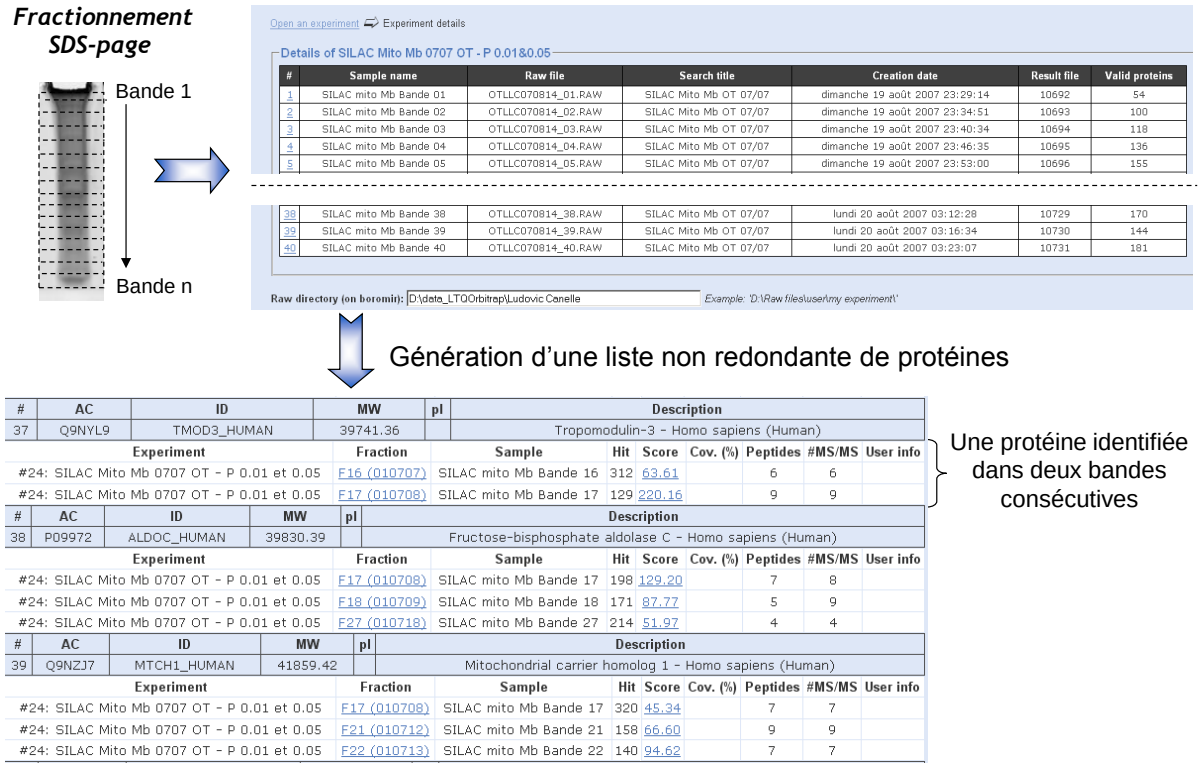


Figure 22 : génération d'une liste non redondante de protéines à partir des résultats d'identification de plusieurs fractions d'un même échantillon.

Il est possible d'effectuer différentes opérations sur ces listes en utilisant le même algorithme de « clusterisation » que nous venons de décrire. On peut par exemple rassembler deux listes de protéines qui proviennent de deux échantillons différents. Une autre fonctionnalité intéressante est

la possibilité de comparer ces listes. Dans le cas d'une comparaison de deux listes de protéines, le logiciel génère trois nouvelles listes :

- une première pour les protéines communes des deux listes comparées,
- une seconde pour les protéines spécifiques de la première liste,
- une troisième pour les protéines spécifiques de la deuxième liste.

Le logiciel supporte la comparaison de 5 listes au maximum. Le nombre de listes générées à l'issu d'une comparaison augmente très rapidement suivant la fonction $2^n - 1$. Ainsi pour une comparaison de 5 listes, le logiciel peut générer jusqu'à 31 listes de protéines ($2^5 - 1 = 32 - 1 = 31$).

Le logiciel MFPaQ nous permet ainsi de valider et d'intégrer l'ensemble des résultats issus de différentes fractions d'un même échantillon, et de comparer efficacement les listes de protéines non redondantes ainsi générées. Cette fonctionnalité a été très utilisée au sein du laboratoire afin d'établir le protéome d'échantillons d'intérêt. Certains travaux réalisés dans le cadre de projets de recherche ont d'ailleurs donné lieu à la publication de protéomes de référence. On peut citer notamment la publication du protéome des globules rouges suite à un traitement d'égalisation du contenu protéique avec les billes ProteoMiner™ (Roux-Dalvai, Gonzalez de Peredo et al. 2008), ou encore celui du protéome des kératinosomes (Raymond, Gonzalez de Peredo et al. 2008), organelles sécrétées au niveau de l'épiderme par les kératinocytes.

II-3. Quantification basée sur l'utilisation d'un marquage isotopique

Comme décrit dans l'introduction, les approches quantitatives en protéomique ont longtemps été basées sur le marquage isotopique des échantillons à comparer. Les premières approches de ce type ont été celles basées sur les réactifs ICAT et le marquage à l'azote ^{15}N décrites en 1999, suivies au début des années 2000 par le développement de plusieurs autres méthodes de marquage isotopique (SILAC, iTRAQ, O18….). Cependant, entre la description de ces approches expérimentales biochimiques, et la possibilité de les utiliser en pratique pour la quantification de protéines identifiées à grande échelle, il a existé pendant plusieurs années un fossé, lié à l'absence de logiciels permettant d'extraire de façon systématique, pour chaque protéine, l'information quantitative associée à l'intensité des peptides formant les paires isotopiques. Alors que les processus bioinformatiques liés à l'exploitation des spectres MS/MS (moteurs de recherche en bases de données, méthodes de validation des identifications…) étaient déjà avancés, le lien entre les données d'identification et les données quantitatives contenues dans les spectres MS est resté longtemps non géré par les logiciels d'analyse protéomique. Pour faire ce lien, il est nécessaire de pouvoir lire de façon systématique l'ensemble des spectres MS enregistrés au cours d'une analyse nanoLC-MS/MS, contenus dans les différents formats de fichiers bruts générés par les divers types de spectromètres de masse. Paradoxalement, alors que les constructeurs ont toujours fourni avec leurs appareils des macros informatiques permettant d'extraire des fichiers bruts l'ensemble des informations liés à l'acquisition MS/MS (les « peaklists », ou liste des masses de fragments détectés pour chaque ion précurseur fragmenté), l'accès aux données MS s'est longtemps révélé moins facile. Ce n'est qu'avec la définition des formats standards de type mzXML ou mzML que les convertisseurs permettant de transformer les fichiers bruts des constructeurs se sont généralisés.

Au sein du laboratoire les ingénieurs et chercheurs travaillaient en 2004 sur l'utilisation du marqueur ICAT pour réaliser des expériences d'analyses différentielles quantitatives. Les seuls outils développés à l'époque (Xtract, ASAPRatio) ne fonctionnaient que sous des environnements Linux, ce qui était incompatible avec notre parc informatique et logiciel, fonctionnant eux sous Microsoft Windows. Il étaient par ailleurs associés à la quantification de protéines identifiées avec le moteur de recherche Sequest, à partir de fichiers bruts de type .raw acquis sur des spectromètres de masse à trappe ionique du constructeur ThermoFinnigan. Nous avons donc décidé de développer notre propre logiciel de quantification MS dans le cadre d'expériences mettant en œuvre le marquage ICAT, analysées avec un spectromètre de masse de type Q-TOF (Qstar, Applied Biosystems) et le moteur de recherche Mascot. Cet outil a été mis au point en tant que nouveau module du logiciel MFPaQ. Les fonctionnalités exprimées par les utilisateurs étaient les suivantes :

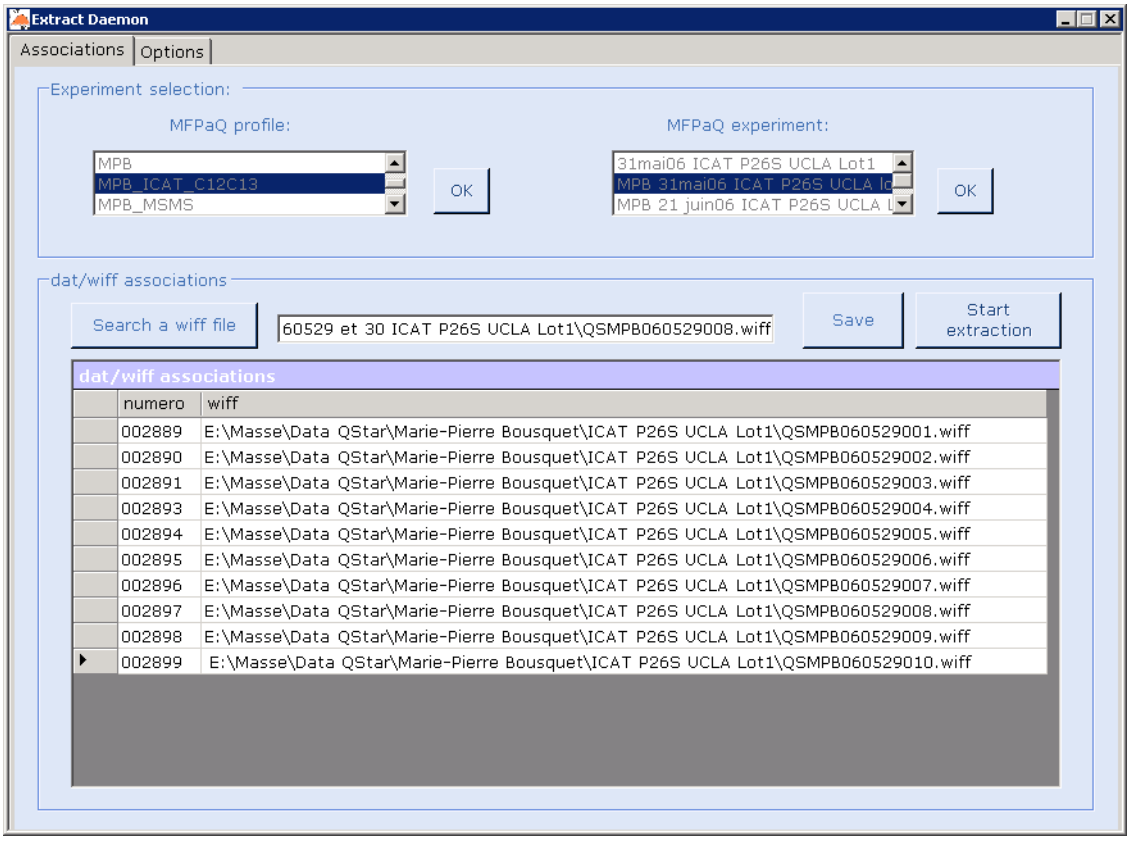
- un système performant d'extraction du signal (reconnaissance du massif isotopique, détection de problèmes de co-élution entre les peptides d'une même paire, détection du bruit de fond),
- la prise en compte de la validation des résultats d'identification,
- une analyse statistique au niveau peptidique et protéique (variabilité des ratios, détection des valeurs extrêmes),
- une interface conviviale pour l'évaluation de la qualité des signaux extraits, avec visualisation des spectres MS,
- la possibilité d'exporter facilement les résultats de quantification dans Excel.

Ce système a été développé avec le langage Visual Basic .Net. Ce dernier a été choisi car il permettait d'utiliser facilement les bibliothèques de code du constructeur Applied Biosystems pour l'accès aux données contenues dans les fichiers d'acquisition du spectromètre de masse de type QStar utilisé au laboratoire (format .wiff).

L'exécution du processus de quantification de l'outil ainsi développé, nécessite de créer un fichier de configuration qui associe les fichiers d'acquisition à quantifier aux fichiers Mascot correspondants, et fournit les paramètres de quantification souhaités : type de marquage, filtres des peptides à quantifier, paramètres pour l'extraction du signal. Pour faciliter l'élaboration de ce fichier de configuration une interface graphique a été mise au point (cf figure 23).

Une fois le fichier de configuration créé l'extraction du signal est alors exécutée selon ces étapes :

1) récupération de la liste des peptides correspondant à des protéines validées pour le fichier .dat en cours de traitement
2) filtrage des peptides selon le score et le rang (définis dans le fichier de configuration)
3) appariement des peptides léger/lourd identifiés par MS/MS
4) calcul de la masse théorique du peptide manquant dans le cas où un seul des deux peptides de la paire isotopique a été identifié par MS/MS
5) extraction du signal au temps d'élution correspondant au scan MS/MS ayant donné le meilleur score Mascot
6) enregistrement des « eXtracted Ion Chromatograms » pour un affichage ultérieur dans l'interface web du logiciel MFPaQ
7) calcul des ratios peptidiques
8) calcul des ratios protéiques, exclusion des ratios peptidiques extrêmes (« outliers ») et calcul de valeurs statistiques (coefficient de variation, intervalle de confiance à 95%).



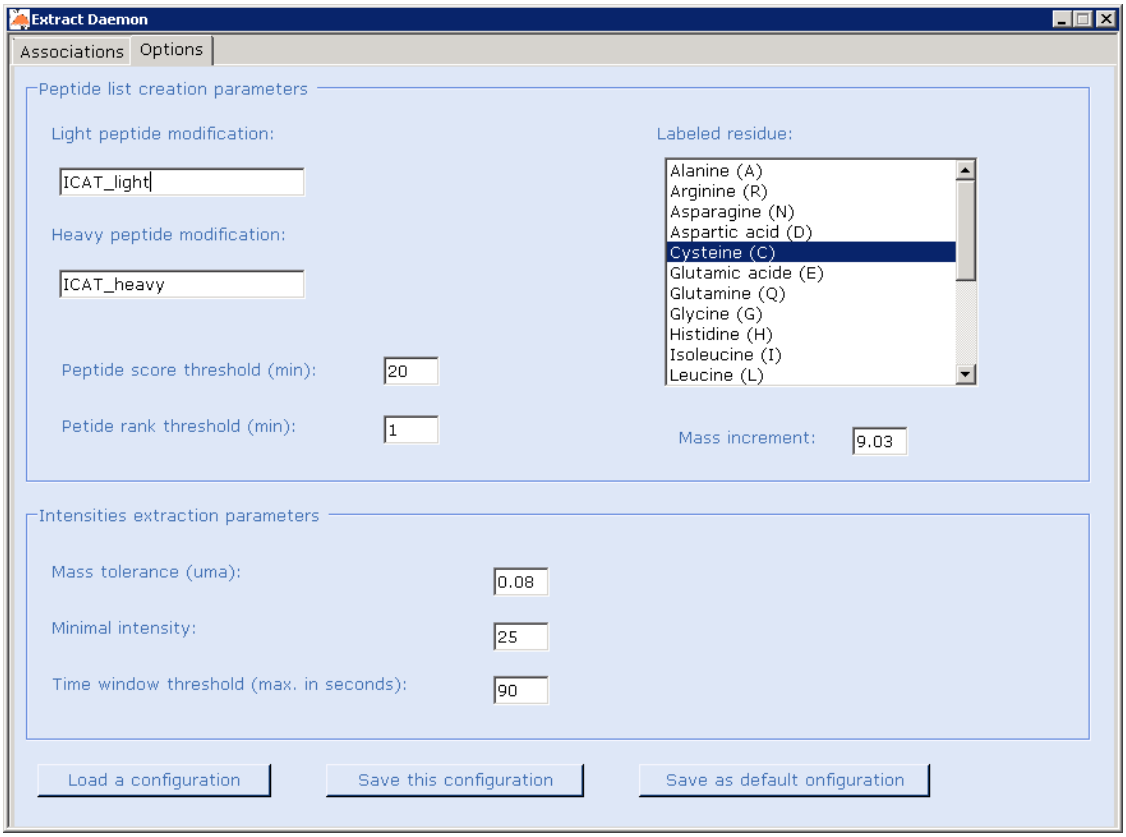
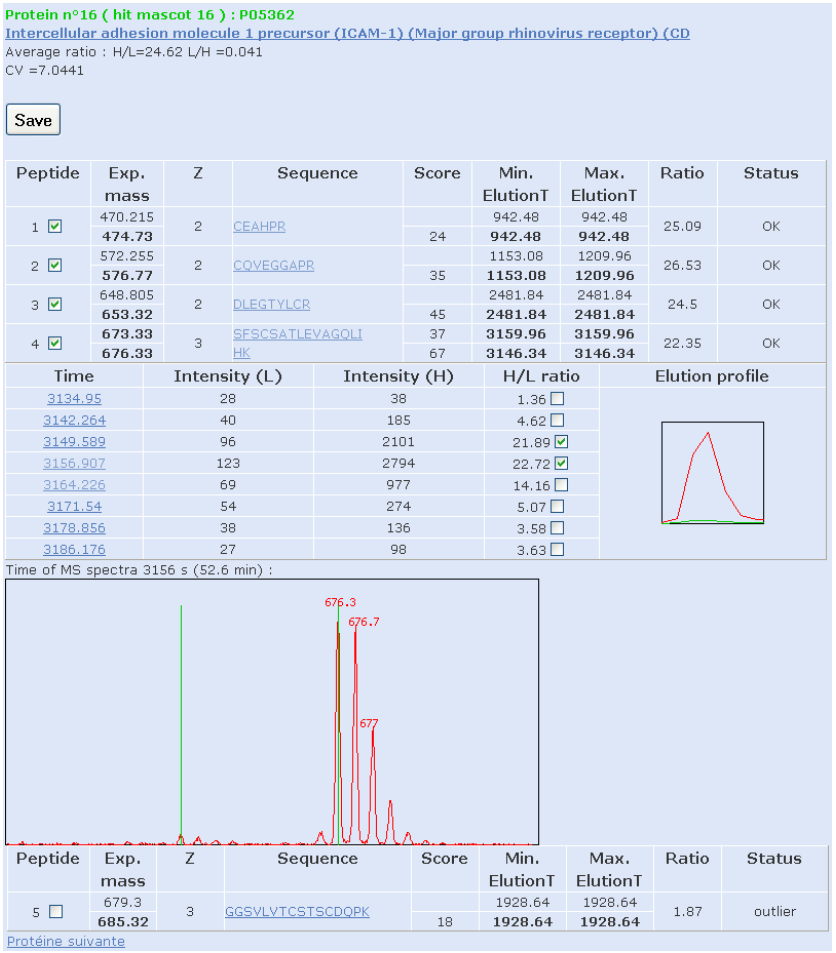


Figure 23 : captures d'écran de l'Extract Daemon. Cette interface graphique permet de lier les données Mascot aux fichiers d'acquisition et de lancer le processus de quantification.

Dans le cas d'une quantification à partir de plusieurs fractions d'un même échantillon, le logiciel est capable dans un deuxième temps de regrouper l'ensemble des résultats obtenus. Ainsi de nouveaux ratios protéiques sont générés en réalisant la moyenne des ratios calculés dans chacune des fractions. De nouvelles valeurs statistiques sont également déterminées et elles rendent compte de la variabilité de la quantification entre les différentes fractions analysées. Une grande variabilité peut avoir comme origine la présence de deux formes de la même protéine ayant des abondances relatives différentes (par exemple une protéine avec ou sans glycosylation).

Enfin, les résultats de quantification calculés lors des étapes précédemment décrites, peuvent être parcourus via une interface web conviviale, comme le montre la capture d'écran ci-dessous.



Cet outil a été utilisé pour plusieurs projets au sein du laboratoire et notamment dans le cas de l'étude de la dynamique du protéome membranaire de cellules endothéliales humaines suite à l'effet de facteurs pro-inflammatoires. C'est d'ailleurs dans le cadre de la publication de cette étude que le logiciel MFPaQ, alors dans sa version 3, a été décrit pour la première fois.

PUBLICATION

« Mascot file parsing and quantification (MFPaQ), a new software to parse, validate, and quantify proteomics data generated by ICAT and SILAC mass spectrometric analyses: application to the proteomics study of membrane proteins from primary human endothelial cells. »

Bouyssié D, Gonzalez de Peredo A, Mouton E, Albigot R, Roussel L, Ortega N, Cayrol C, Burlet-Schiltz O, Girard JP, Monsarrat B

Mol Cell Proteomics. **2007** Sep;6(9):1621-37.

Mascot File Parsing and Quantification (MFPaQ), a New Software to Parse, Validate, and Quantify Proteomics Data Generated by ICAT and SILAC Mass Spectrometric Analyses

APPLICATION TO THE PROTEOMICS STUDY OF MEMBRANE PROTEINS FROM PRIMARY HUMAN ENDOTHELIAL CELLS*[S]

David Bouyssié‡§, Anne Gonzalez de Peredo‡§¶, Emmanuelle Mouton‡, Renaud Albigot‡, Lucie Roussel‖, Nathalie Ortega‖, Corinne Cayrol‖, Odile Burlet-Schiltz‡, Jean-Philippe Girard‖, and Bernard Monsarrat‡

Proteomics strategies based on nanoflow (nano-) LC-MS/MS allow the identification of hundreds to thousands of proteins in complex mixtures. When combined with protein isotopic labeling, quantitative comparison of the proteome from different samples can be achieved using these approaches. However, bioinformatics analysis of the data remains a bottleneck in large scale quantitative proteomics studies. Here we present a new software named Mascot File Parsing and Quantification (MFPaQ) that easily processes the results of the Mascot search engine and performs protein quantification in the case of isotopic labeling experiments using either the ICAT or SILAC (stable isotope labeling with amino acids in cell culture) method. This new tool provides a convenient interface to retrieve Mascot protein lists; sort them according to Mascot scoring or to user-defined criteria based on the number, the score, and the rank of identified peptides; and to validate the results. Moreover the software extracts quantitative data from raw files obtained by nano-LC-MS/MS, calculates peptide ratios, and generates a non-redundant list of proteins identified in a multisearch experiment with their calculated averaged and normalized ratio. Here we apply this software to the proteomics analysis of membrane proteins from primary human endothelial cells (ECs), a cell type involved in many physiological and pathological processes including chronic inflammatory diseases such as rheumatoid arthritis. We analyzed the EC membrane proteome and set up methods for quantitative analysis of this proteome by ICAT labeling. EC microsomal proteins were fractionated and analyzed by nano-LC-MS/MS, and database searches were performed with Mascot. Data validation and clustering of proteins were performed with MFPaQ, which allowed identification of more than 600 unique proteins. The software was also successfully used in a quantitative differential proteomics analysis of the EC membrane proteome after stimulation with a combination of proinflammatory mediators (tumor necrosis factor-α, interferon-γ, and lymphotoxin α/β) that resulted in the identification of a full spectrum of EC membrane proteins regulated by inflammation. ***Molecular & Cellular Proteomics 6:1621–1637, 2007.***

In recent years, nanoflow (nano-)[1] LC-MS/MS has emerged as an efficient alternative to two-dimensional electrophoresis in the field of proteomics. This technology has proved to be a powerful method for the identification of proteins in complex mixtures and has been applied to characterize the proteome of several organisms, organelles, and multiprotein complexes. Moreover many developments have been made to use nano-LC-MS/MS-based strategies in differential proteomics studies to compare the proteome of two or more samples in a quantitative or semiquantitative way. Although recent approaches use direct comparison of the MS peptide signals from independent nano-LC-MS/MS runs (1–5), most studies up to now have used quantitative methods based on isotopic

From the ‡Laboratoire de Protéomique et Spectrométrie de Masse des Biomolécules and ‖Laboratoire de Biologie Vasculaire, Equipe Labellisée "Ligue 2006," Institut de Pharmacologie et de Biologie Structurale, CNRS UMR 5089, 205 route de Narbonne, 31077 Toulouse, France

Received, December 21, 2006, and in revised form, April 11, 2007
Published, MCP Papers in Press, May 28, 2007, DOI 10.1074/mcp.T600069-MCP200

[1] The abbreviations used are: nano-, nanoflow; c-ICAT, cleavable ICAT; EC, endothelial cell; HUVEC, human umbilical vein endothelial cell; IFNγ, interferon γ; MFP, Mascot File Parser; MFPaQ, Mascot File Parsing and Quantification; MudPIT, multidimensional protein identification technology; SILAC, stable isotope labeling with amino acids in cell culture; TNF-α, tumor necrosis factor-α; XML, extensible markup language; 1D, one-dimensional; cps, counts/s; H, heavy; L, light; ALCAM, activated leukocyte cell adhesion molecule; ICAM, intercellular cell adhesion molecule; VCAM-1, vascular cell adhesion molecule-1; PECAM-1, platelet/endothelial cell adhesion molecule-1; iTRAQ, isobaric tags for relative and absolute quantification; HLA, human leukocyte antigen; STEM, strategic extractor for Mascot's results.

labeling of proteins or peptides combined with nano-LC-MS/MS analyses (6, 7). In these approaches, light and heavy isotopic labels are introduced into the proteins from the different samples to be compared. The samples are then mixed together, and a single nano-LC-MS/MS analysis is run. The relative abundance of a given protein can then be deduced from the ion signal intensity ratio calculated for light/heavy peptide pairs from this protein. This leads to a more accurate relative quantification of the proteins from the samples to be compared because the samples are analyzed simultaneously in a single nano-LC-MS/MS run. In the ICAT method, proteins are chemically labeled on cysteines with a biotinylated heavy or light reagent (8, 9), whereas in the SILAC (stable isotope labeling with amino acids in cell culture) method, the label is introduced during protein synthesis by growing cells in a medium containing a heavy or light amino acid (10). In these strategies, systematic identification of as many proteins as possible is usually performed by nano-LC-MS/MS analysis with shotgun approaches involving prefractionation of the proteins or the peptides, and correctly assigned proteins can be quantified afterward on the basis of the MS signal of the corresponding peptides. This in turn leads to the production of a huge amount of MS/MS and MS spectra that must be handled for identification and quantification, thus necessitating appropriate bioinformatics tools.

Data analysis and validation of the results from MS/MS searches have become major issues of mass spectrometry-based proteomics, and a lot of efforts are made to provide efficient tools for evaluating and organizing data. Although Mascot (11) and Sequest (12) remain the two reference softwares that are widely used for protein identification from MS/MS data, the protein matching lists that they return still contain false positives and skip some false negatives. To improve the reliability of results, new search engines and scoring techniques were recently developed. These include the S-score (13), the softwares PeptideProphet and ProteinProphet (14–16), and the new Phenyx search engine based on the OLAV algorithm (17). Other tools and methods aiming at facilitating the validation and handling of Mascot results include MSQuant (18) and STEM (19). Here we describe a new program named Mascot File Parsing and Quantification (MFPaQ) that allows fast and user-friendly verification of Mascot result files as well as data quantification from an experiment performed by isotopic labeling using either ICAT or SILAC methods.

This software provides an interactive interface with Mascot results. It is based on three modules, the Mascot File Parser module, the quantification module, and a third module designed for differential analysis in which validated protein lists are compared.

The potentialities of the MFPaQ software are illustrated by the analysis of the results from a nano-LC-MS/MS proteomics study of membrane proteins from primary human endothelial cells (ECs). ECs, which form a monolayer lining all blood vessels, play a key role in diverse physiological and pathological processes, including chronic inflammatory diseases such as rheumatoid arthritis, in which they are involved in the regulation of leukocyte extravasation, angiogenesis, cytokine production, protease and extracellular matrix synthesis, antigen presentation, vasodilatation, and blood vessel permeability (20, 21). In this study, we tried to better characterize the membrane proteome of human ECs and to set up methods for quantitative analysis of this proteome by ICAT labeling. Microsomes from ECs were fractionated by 1D SDS-PAGE, resulting gel slices were analyzed by nano-LC-MS/MS, and database searches were performed using Mascot. Data validation and clustering of proteins were performed with MFPaQ, which allowed the identification of more than 600 unique proteins. The software was then successfully used to perform quantification of proteins from a 1:1 heavy/light c-ICAT labeling test experiment. Finally we stimulated human ECs with a combination of key proinflammatory cytokines, TNF-α, IFNγ, and lymphotoxin α/β, and performed a differential proteomics analysis using the ICAT method. The validated results obtained using MFPaQ software allowed the identification of 44 EC membrane proteins regulated by inflammation.

MATERIALS AND METHODS

MFPaQ Details and System Requirements—MFPaQ is a Web-based application that runs on a server on which Mascot Server 2.1 and Perl 5.8 must be installed as well. It functions with an Internet Information Services Web Server under Windows XP Pro edition and Windows 2003 Server. Scripts are written in Perl language and use the modules XML-Simple, Spreadsheet-WriteExcel, and GD. The user interface is accessible via a Web browser: Microsoft Internet Explorer and Mozilla Firefox are currently compatible with the application. Proteomics data (protein and peptide identifications, validated protein lists, and quantification results) are stored in the XML file format. To perform quantification, an external module called "Extract Daemon" has been developed for extracting intensity values from raw data. This module was developed in Visual Basic.Net and works at the moment with ".wiff" files acquired on a QStar XL or QStarElite instrument (Applied Biosystems, Foster City, CA). It must be installed on the same server as MFPaQ on which Analyst QS 1.1 or Analyst QS 2.0 should be installed as well. Two versions of the application, compatible with these corresponding versions of Analyst QS, are freely available at mfpaq.sourceforge.net. Although Mascot 2.1 and Analyst QS are necessary to process and quantify new data with MFPaQ, the application can be installed alone and is able to display all detailed protein lists and peptide information presented in the results section. MS/MS spectra for all assigned peptide sequences can be viewed if Mascot has been installed on the same computer.

EC Culture and Cytokine Stimulation—Primary human umbilical vein endothelial cells (HUVECs) were isolated from fresh human umbilical cords and further purified with CD105 microbeads (Miltenyi Biotec, Auburn, CA) as described previously (22). HUVECs were grown in endothelial cell growth medium (Promocell, Heidelberg, Germany) and used after four passages for proteomics analyses. Cytokine treatment was performed by incubating the ECs for 12 h in Opti-MEM (Invitrogen) with a combination of TNF-α (25 ng/ml, R&D Systems), IFNγ (50 ng/ml, R&D Systems), and lymphotoxin α/β (200 ng/ml, R&D Systems).

Purification of Microsomes—Cells were washed with PBS and collected with a cell scraper in 0.25 M sucrose, 10 mM Hepes, 2 mM

$MgCl_2$, pH 7.6, supplemented with protease inhibitors (Complete, Roche Applied Science). Cell lysis was performed with an Ultraturax homogenizer, and the resulting homogenate was centrifuged for 10 min at 800 × *g* to remove nuclei and cell debris. The postnuclear supernatant was centrifuged for 10 min at 10,000 × *g*, resulting in a pellet enriched in mitochondria that was not analyzed. The supernatant was centrifuged at 200,000 × *g* for 45 min, and the microsomal pellet was washed by resuspension in 100 mM Na_2CO_3, pH 12, to remove soluble contaminants and centrifuged again at 200,000 × *g* for 45 min. The washed pellet was solubilized in 50 mM Tris, 6 M urea, 0.5% SDS, pH 8.3. Protein concentration was determined with the reductant compatible-detergent compatible assay (Bio-Rad).

c-ICAT Labeling—Microsomal proteins (96 μg) in 50 mM Tris, 6 M urea, 0.5% SDS, pH 8.3, were reduced with tris(2-carboxyethyl)phosphine HCl (0.1 mmol) for 2 h at room temperature and labeled with one unit of heavy or light c-ICAT (Applied Biosystems) for 3 h at room temperature in the dark. The reaction was stopped by adding Laemmli buffer to the samples, resulting in a 25 mM DTT final concentration. Samples labeled with the heavy or light reagent were then mixed and loaded on a 1D SDS-PAGE gel to fractionate the protein mixture and eliminate excess ICAT reagent.

Analysis by 1D Gel/Nano-LC-MS/MS—Microsomal proteins were fractionated on a 1D SDS-PAGE gel (1.5 mm × 8 cm), the gel was briefly stained with Coomassie Blue, and the entire migration lane was cut into 20 homogeneous gel slices. Gel slices were washed and digested with modified sequencing grade trypsin (Promega, Madison, WI), and resulting peptides were extracted. For unlabeled proteins, extracted peptides were directly analyzed by nano-LC-MS/MS. For the ICAT labeling experiment, labeled peptides were purified by affinity chromatography on a monomeric avidin cartridge according to the manufacturer's protocol (Applied Biosystems, Framingham, MA). Peptides were eluted from the cartridge with 30% ACN, 0.4% TFA in H_2O and dried down in a SpeedVac, and the cleavable biotin moiety of the labeling reagent was then submitted to acid hydrolysis according to the manufacturer's protocol. Resulting peptides were analyzed by nano-LC-MS/MS using an LC Packings system (Dionex, Amsterdam, The Netherlands) coupled to a QStar XL mass spectrometer (Applied Biosystems). Dried peptides were reconstituted in 12 μl of solvent A′ (5% ACN, 0.05% TFA in HPLC-grade water), and 6 μl were loaded onto a precolumn (300-μm inner diameter × 5 mm) using the Switchos unit of the LC Packings system, delivering a flow rate of 20 μl/min solvent A′. After desalting for 7 min, the precolumn was switched on line with the analytical column (75-μm inner diameter × 15-cm PepMap C_{18}) equilibrated in 95% solvent A (5% ACN, 0.1% formic acid in HPLC-grade water) and 5% solvent B (95% ACN, 0.1% formic acid in HPLC-grade water). Peptides were eluted from the precolumn to the analytical column and then to the mass spectrometer with a gradient from 5 to 50% solvent B (during either 60 or 80 min) at a flow rate of 200 nl/min delivered by the Ultimate pump. The QStar XL was operated in information-dependant acquisition mode with the Analyst QS 1.1 software. MS and MS/MS data were recorded continuously with a 5-s cycle time. Within each cycle, MS data were accumulated for 1 s over the mass range *m/z* 300–2000 followed by two MS/MS acquisitions of 2 s each on the two most abundant ions over the mass range *m/z* 80–2000. Dynamic exclusion was used within 60 s to prevent repetitive selection of the same ions. Collision energies were automatically adjusted according to the charge state and mass value of the precursor ions. The MS to MS/MS switch threshold was set to 10 cps.

Database Searching—The Mascot Daemon software (version 2.1.6) was used to automatically extract peak lists from Analyst QS .wiff files and to perform database searches in batch mode with all the .wiff files acquired on each gel slice. For creation of the peak lists, the default charge state was set to 2+, 3+, and 4+. MS and MS/MS centroid parameters were set to 50% height percentage and a merge distance of 0.1 amu. All peaks in MS/MS spectra were conserved (threshold intensity set to 0% of highest peak). For MS/MS grouping, the following averaging parameters were selected: spectra with fewer than five peaks or precursor ions with less than 5 cps or more than 10,000 cps were rejected, the precursor mass tolerance for grouping was set to 0.1 Da, the maximum number of cycles per group was set to 10, and the minimum number of cycles per group was set to 1. MS/MS data were searched against all entries in the public database UniProt version 8.1, which consists of Swiss-Prot Protein Knowledgebase Release 50.1 and TrEMBL Protein Database Release 33.1 (3,192,898 entries in total), using the Mascot search engine (Mascot Daemon, version 2.1.6; Matrix Science, London, UK). To evaluate the false positive rate in these large scale experiments, we repeated the searches using identical search parameters and validation criteria against a random database. The database was the compilation of UniProt Swiss-Prot and UniProt TrEMBL databases (same versions described above) in which the sequences have been reversed. Oxidation of methionine was set as a variable modification for all Mascot searches, and for ICAT labeling experiments, alkylation of cysteine with light ^{12}C c-ICAT and with heavy ^{13}C c-ICAT also was set as a variable modification. Specificity of trypsin digestion was set for cleavage after Lys or Arg, and two missed trypsin cleavage sites were allowed. The peptide MS and MS/MS tolerances were set to 0.15 and 0.25 Da, respectively.

RESULTS

MFPaQ Features

MFPaQ is a software tool that facilitates organization, mining, and validation of Mascot results and offers different functionalities to work on validated protein lists. A schematic overview of the program is given in Fig. 1. The software is organized around a core module, the Mascot File Parser ("MFP") module that extracts data from Mascot result files (.dat) and allows the user to browse, validate, and cluster the results. The MFP module stores protein and peptide lists in .xml files that can be used by the "differential analysis" module to compare the lists of proteins from two or more experiments and by the "quantification" module to compute the ratios of the proteins in an isotopic labeling experiment (ICAT or SILAC). The software is a Web-based application that runs on a server (where Mascot Server is installed and the .dat result files are generated). It can be accessed by different users via a Web browser. Each user can create his own profile by defining several criteria that will be used by the MFP module to validate the proteins extracted from Mascot files. User profiles and criteria can be modified and saved at any time to perform another extraction using different criteria. Each user works under a personal session in which he can create and store experiments.

The MFP Module for Validation and Classification

Description of the MFP Module

A first module, the Mascot File Parser, performs validation and classification of the proteins from a result data file according to Mascot scoring or according to user-defined criteria based on the number, score, and rank of identified

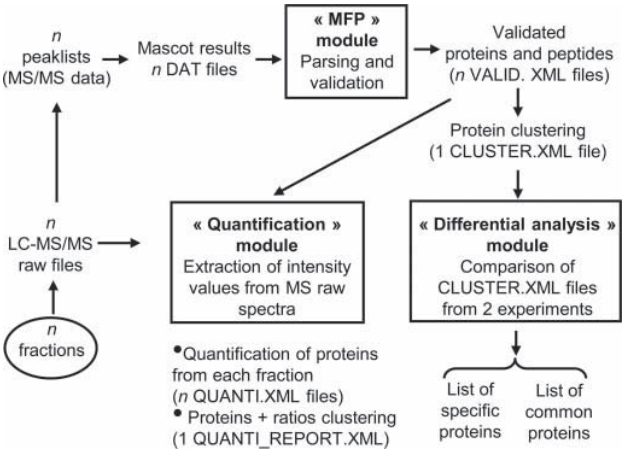


FIG. 1. **General scheme of MFPaQ.** The MFP module extracts data from Mascot result files (.dat files) and generates lists of protein groups and associated lists of peptide matches stored in the VALID.xml files. Several of these files can be gathered to create a non-redundant list of protein groups (CLUSTER.xml file), and concatenated lists from different experiments can be compared in the differential analysis module. The quantification module uses peptide data contained in the VALID.xml files (ion *m/z* and peptide retention time) to extract the intensity value for each peptide pair (ICAT or SILAC labeling) in the survey scan of the corresponding raw file, calculates the ratios of the peptides and of the proteins for each nano-LC-MS/MS run (QUANTI.xml files), and generates a global quantification report (QUANTI_REPORT.xml) by gathering the protein identifications and averaging quantification data from multiple nano-LC-MS/MS runs.

peptides. This module offers to the user a convenient interface to manually validate or reject ambiguous identifications. It can also group identical or highly homologous proteins from several result data files to eliminate redundancy and to provide a global and relevant list of the proteins present in the sample. The use of the MFP module consists in three main steps detailed below corresponding to the extraction of Mascot files in batch mode, protein validation, and generation of protein lists.

Extraction of Mascot Files in Batch Mode—The MFP module offers the possibility to create an "experiment" corresponding to the extraction of one or several Mascot result files (.dat files). Depending on how the shotgun analysis of a protein sample is conducted, it may be relevant to perform either a single Mascot search or several searches for this sample. For example, if a whole complex protein mixture is enzymatically digested and the resulting peptides are fractionated using chromatography (*e.g.* on a strong cation exchange column), each peptide fraction will then be analyzed by nano-LC-MS/MS, and different peptides belonging to the same protein will be analyzed in several of these nano-LC-MS/MS runs. In this case, making a unique peak list from all the MS/MS scans acquired in all the runs will be necessary to identify efficiently the proteins in a single Mascot database search. Conversely if the protein mixture is fractionated first (*e.g.* in a series of 1D gel slices) and each protein fraction is digested, then peptides from each fraction will be analyzed by nano-LC-MS/MS, and all the peptides from a protein will be analyzed in the same run. In that case, several Mascot database searches should be performed with the different peak lists obtained from the nano-LC-MS/MS runs, and the different protein lists obtained should be gathered afterward to avoid erroneous assignments of MS/MS spectra acquired in one fraction to a protein present in another fraction. In this way, no information is lost in the identification process, and the physicochemical properties of the proteins that were used to perform fractionation in the first step (*e.g.* molecular weight in the case of 1D SDS-PAGE separation) may represent an additional parameter of interest for the validation of protein identification. For example, in the case of an ambiguous identification by Mascot, a strong discrepancy between the theoretical molecular weight of the predicted protein and the experimental molecular weight corresponding to the gel slice on which the analysis was performed can be used as a criterion by the user to reject the identification. In both cases, MFPaQ provides a clear interface for visualizing, mining, and organizing the results of a multisearch experiment. The software extracts in batch mode the data contained in a series of Mascot .dat files specified by the user under an experiment and displays a table with links to a validation window for each of these searches as illustrated in Fig. 2*A*.

Validation of Proteins—The MFP module extracts protein entries from Mascot files and can rank them according to either the Mascot "Standard scoring" or "MudPIT scoring." To facilitate manual validation, the software applies to the proteins of the list a two-color code related to filtering rules defined by the user under its configuration profile. Proteins that passed the "validation criteria" are displayed in green. They can be considered as confident hits that do not need further verification and will automatically be checked in the validation window. Proteins that meet the "exclusion criteria" are discarded and are not displayed in the list. All other proteins, which are considered as ambiguous identifications, appear in red and can be manually verified by the user. The filtering rules used for the classification of a protein in green and red are based either on the protein score defined in Mascot or on multiple criteria related to the peptide matches (sequence interpretation of an MS/MS spectrum) assigned to this protein. In the first case, the software basically displays in green color the "significant hits" list given in the Mascot Peptide summary report. Mascot uses the probability-based Mowse algorithm to calculate ion scores, defined as $-10 \times \log(p)$ where p is the probability that the observed match for this ion is a random event. Protein scores are derived from ion scores as a non-probabilistic basis for ranking protein hits and are computed differently in Standard scoring and MudPIT scoring. The significant hits list given by Mascot contains the proteins with total scores higher than the significance threshold, which depends on the database size and is calculated by

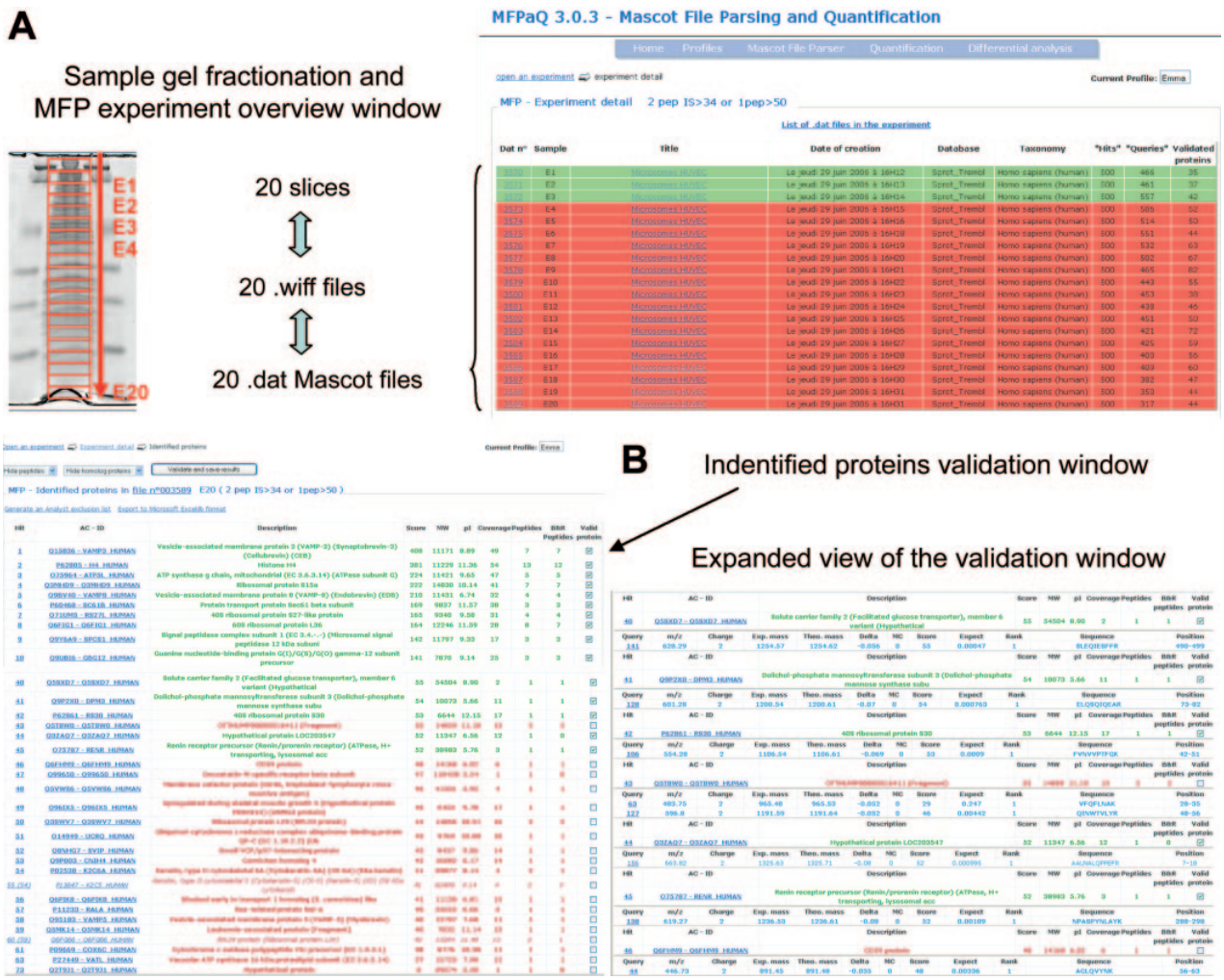


FIG. 2. **Visualization and parsing of Mascot results in the MFP module.** *A*, an experiment is created in MFP from a series of Mascot .dat files, corresponding for example to several nano-LC-MS/MS runs performed on consecutive gel slices of a 1D gel. Extracted Mascot results are automatically validated according to user-defined criteria. However, the user keeps a trace of the fraction (.dat) manual validation process: before being verified, modified, and saved, they appear in *red*, and afterward they are displayed in *green*. *B*, protein validation is also color-coded: *green* proteins fulfill user-defined criteria, whereas ambiguous identifications are displayed in *red*. The proteins are ranked according to Mascot scoring. In the *expanded view* of the window, peptide information is available, and links to MS/MS spectra allow verification of the peptide sequence assignment.

default with the probability for a match to occur at random with a probability of less than 5% ($p < 0.05$). However, some false positives are clearly present in this list, and some false negatives are missing. Although Mascot still appears to be at the moment one of the most efficient search engines, this lack of specificity has prompted a lot of efforts from several groups to set up more reliable scoring systems. Although promising, these systems will need further validation and are not yet of general use. Therefore, manual validation is still often performed by many users at least for borderline proteins around the Mascot threshold. The MFP module is very helpful for this process because it can classify the proteins according to more or less stringent criteria based on the number, the rank, and the score of the peptide matches assigned to a protein. The protein displayed in MFPaQ will still be ranked according to Mascot scoring, but proteins possessing for example at least two bold and red peptide matches in Mascot, with scores higher than 40, will appear in green and will be automatically validated. Proteins that do not fulfill these criteria, although being in the significant list of Mascot, will appear in red and will have to be verified manually. To that aim, all the information given by Mascot for a protein hit is also available in the MFP window: protein mass, pI, total score, list of assigned queries with the corresponding peptide sequence, theoretical and experimental masses of the peptide matches, delta value between these two masses, score and rank of the

peptide matches, and E-value for the assignment. Links to the Mascot "Peptide view" window containing MS/MS centroided spectra are available as well and allow rapid verification by the user of ambiguous proteins displayed in red (Fig. 2*B*). It has to be noted that in MS/MS strategies identical peptides can often be mapped to different protein sequences present in a database, corresponding either to redundant sequences, amino acid variants, splice isoforms, different protein fragments, or protein homologs. The Mascot software automatically groups together protein sequences matching exactly the same set of peptides. Under the MFP module, it is possible to display a concise list of the proteins identified where only one member of each group appears but also a detailed list containing all the members of each group of proteins sharing the same set of peptides. Moreover MFP is able to detect protein homologs or protein fragments related to another protein ranked higher in the list. These proteins are usually identified with a subset of shared peptides (displayed as red and non-bold peptides in Mascot) but are not grouped together with the previous hit because additional, specific peptide sequences are also assigned to them. The MFP module displays these proteins in italic if these supplemental sequences are low scoring peptide matches (score lower than 30) that do not allow their identification as specific hits (in that case, these proteins or protein groups were not validated in the following results section). However, it displays them as real specific hits if they have at least one high scoring (score higher than 30) red and bold specific peptide match. These features, and the interactive validation window, enable the user to save a lot of time by browsing and easily validating the results.

Saving Protein Lists—Once the verification has been performed, validated proteins (or protein groups), including all associated peptide information, are saved in XML files and can be exported into Excel. Another important feature of the MFP module is the possibility of generating exclusion lists for further nano-LC-MS/MS experiments. Such lists can be used to perform a second nano-LC-MS/MS run of the same sample in which intense ions that were already assigned to a validated protein in the first run will not be selected again for MS/MS, potentially giving the mass spectrometer more time to sequence less abundant peptides. Finally the MFP module can also generate a unique, non-redundant list of proteins from all the validated result files of a multisearch experiment. This unique feature from MFPaQ is particularly useful when protein fractionation is performed because the same protein can be identified several times in adjacent gel slices. The software compares proteins or protein groups (composed of all the proteins matching the same set of peptides) and creates clusters from protein groups found in different gel slices if they have one common member. This feature allows the editing of a global list of unique proteins (or clusters) representing the entire sample analyzed in the experiment.

Application of the MFP Module to the Identification of Membrane Proteins from Primary Human ECs

EC microsomes were prepared, washed with sodium carbonate at high pH to enrich the mixture in integral membrane or membrane-anchored proteins, fractionated by 1D SDS-PAGE, and analyzed by nano-LC-MS/MS. Analysis of highly hydrophobic proteins is often difficult because the classical buffers used in many protein separation techniques (two-dimensional electrophoresis and liquid chromatography) and the conditions compatible with enzymatic digestion are often not efficient enough to solubilize them, leading to protein aggregation and precipitation. 1D SDS-PAGE is a well suited approach for the separation of highly hydrophobic proteins because they can be efficiently solubilized in Laemmli buffer and fractionated. The enzymatic digestion step can then be easily performed in gel once the proteins have been fixed in the gel and the SDS has been washed out of the bands. Twenty gel slices were cut all along the migration lane, digested with trypsin, and analyzed by nano-LC-MS/MS with a 60-min-long gradient. Mascot results obtained for all of them were filtered out and validated with the MFP module of MFPaQ. Table I presents the number of proteins or protein groups identified in each gel slice when using different criteria for protein validation: either Mascot scoring (Standard or MudPIT) or criteria based on the number, the rank, and the score of the peptide matches. Validation based on the Mascot Standard scoring and a protein score higher than 34 ($p < 0.05$) resulted in a final non-redundant list of 1477 protein groups (data not shown), whereas validation based on Mascot MudPIT scoring, with the same threshold, gave a final non-redundant list of 855 protein groups (Table I, column 1). Performing a random database search and applying the later criterion for validation (Mascot MudPIT score higher than 34) led to the identification of 101 protein groups, indicating a false positive rate on the previous list of about 11%. Using the same procedure with the Standard scoring we obtained a false positive rate of 16%. When the stringency of filtering was increased by validating only proteins with at least two reliable peptide matches (rank 1 and individual score higher than 34), the number of validated proteins went down to 491. A random database search with the same criteria led to the validation of only two proteins, indicating a false positive rate of 0.4% (Table I, column 2). Thus, although this list appeared to be much more reliable according to the estimated false positive rate, it was also much more restrictive and potentially omitted a large number of false negatives. Among them, many proteins were identified on the basis of only one peptide match, and we thus tested several criteria of validation to rescue some of these proteins while maintaining an acceptable level of false positive rate. In addition to the proteins identified with more than two peptide matches with individual score higher than 34, we allowed automatic validation of proteins identified with a single peptide match. When the minimal score of these

TABLE I

Number of automatically validated protein groups identified in each gel slice and in the final non-redundant list after proteomics analysis of EC microsomes

Database searches were performed using Mascot for each nano-LC-MS/MS run, and result files were parsed with different criteria for protein validation selected under the MFP module. For each gel slice fraction, the number of protein groups (proteins matching the same set of peptides) is shown. The number indicated for the non-redundant final protein lists refers to a number of unique protein groups obtained after clustering by the software of the different lists of protein groups in the consecutive fractions. Random database searches were performed for the same nano-LC-MS/MS runs using similar parameters and parsed using the same criteria to evaluate the rate of false positive hits after automatic validation.

1D gel slices	1. Mascot validation (total protein MudPIT score >34)		2. Two peptides with individual scores >34		3. Two peptides with individual scores >34 or one peptide with individual score higher than 41		4. Two peptides with individual scores >34 or one peptide with individual score higher than 50		
	Search Swiss-Prot-TrEMBL	Search reverse database	Search Swiss-Prot-TrEMBL	Search reverse database	Search Swiss-Prot-TrEMBL	Search reverse database	Search Swiss-Prot-TrEMBL	Search reverse database	Search Swiss-Prot-TrEMBL + manual validation
E1	48	2	27	0	39	0	35	0	42
E2	58	8	32	0	44	2	37	0	44
E3	63	7	30	0	53	4	43	1	50
E4	69	0	38	0	59	0	49	0	61
E5	60	3	34	0	50	2	50	1	55
E6	62	5	33	0	50	0	41	0	56
E7	75	6	37	0	60	1	54	0	59
E8	83	7	37	1	62	3	55	1	63
E9	97	4	50	0	70	1	63	0	69
E10	73	5	34	0	57	1	50	1	61
E11	57	3	29	0	49	1	38	0	54
E12	57	5	28	0	51	1	45	0	56
E13	64	8	34	0	49	2	46	0	56
E14	81	9	41	1	74	3	69	1	75
E15	72	6	44	0	60	3	57	0	63
E16	72	2	38	0	60	0	50	0	61
E17	72	4	36	0	64	1	57	0	67
E18	60	9	34	0	52	0	46	0	56
E19	61	11	27	0	51	2	43	0	57
E20	65	5	27	0	54	4	44	0	58
Total number of validated proteins (non-redundant list)	855	101	491	2	706	29	626	4	733
False positive percentage (%)	11.80		0.40		4.10		0.63		

single peptide hits was set to 41 ($p < 0.01$), the list of validated protein groups significantly increased to 706, but the false positive rate went up to 4% (Table I, column 3). Finally intermediate criteria were selected by setting the minimal score for these single peptide match hits to 50, which gave a final non-redundant list of 626 protein groups and a false positive rate of 0.6% (Table I, column 4). We chose to use this criteria for automatic validation with MFP (green proteins; see Fig. 2). A manual check was additionally performed on ambiguous proteins that did not fulfill these criteria (displayed in red in the MFP window) and that potentially still contained false negatives. Manual verification of the MS/MS spectra allowed the rescue of 107 positive hits (Table I, column 4) when the fragmentation data were of high quality and strongly indicative of the peptide sequence (at least four consecutive y ions and a delta mass between measured and theoretical peptide molecular mass lower than 0.1 Da). The list of 626 proteins automatically validated by MFP with the above mentioned criteria is provided in Supplemental Data 1. For more clarity, only one member of each protein group (proteins matching the same set of peptides) is displayed. The lists of protein groups identified in each gel slice fraction with all peptide information associated are also provided (Supplemental Data 2) as well as annotated MS/MS spectra in the case of single peptide-based matches (Supplemental Data 9). The complete database search results with detailed protein groups and peptide assignments can be viewed and browsed over by downloading the MFPaQ software and associated data files at mfpaq.sourceforge.net. Automatic classification of the protein list according to Gene Ontology annotations was then performed with the GoMiner software (discover.nci.nih.gov/gominer/) and indicated that, of 450 proteins annotated in terms of subcellular localization, 254 proteins are membrane proteins, and 90 proteins appear to be localized at

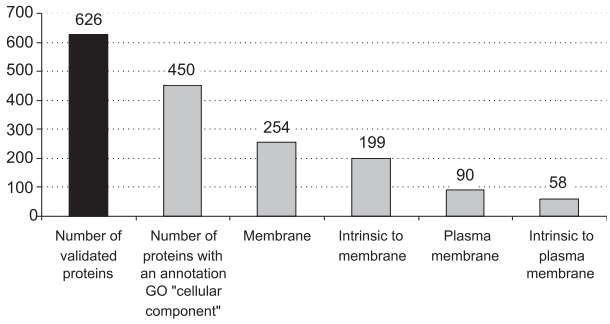


FIG. 3. **Classification of the proteins identified in EC microsomes according to their subcellular localization.** Automatic classification into different categories was performed using the GoMiner software (discover.nci.nih.gov/gominer/) according to the Gene Ontology (*GO*) annotations available for each protein.

the plasma membrane (Fig. 3). The list of proteins identified comprises at least 41 known EC surface markers (Supplemental Data 1) including classical endothelial markers CD31 (PECAM-1), VE-cadherin (Cadherin-5), CD105 (Endoglin), CD146 (MUC18/melanoma cell adhesion molecule), podocalyxin, tyrosine kinase receptor Tie-2, intercellular cell adhesion molecule (ICAM)-2, endothelial cell-selective adhesion molecule, aminopeptidase N (CD13), angiotensin-converting enzyme (CD143), dipeptidyl-peptidase IV (CD26), and endothelin-converting enzyme.

The MFPaQ Quantification Module for the Relative Quantification of Isotopically Labeled Proteins

Description of the MFPaQ Quantification Module

An important feature of MFPaQ is a quantification module, which extracts quantitative data from raw files obtained by nano-LC-MS/MS when using either ICAT or SILAC labeling techniques. The software allows the verification of the calculated ratios and the manual deselection of some peptide pairs or some MS scans in case of aberrant ratio calculation (co-elution with other peptides, weak signal, etc.). After validation of the proteins identified, the quantification module uses the peptide lists generated by the MFP module to select the peptides containing an isotopic modification specified by the user (*e.g.* a cysteine modified by a c-ICAT reagent or a peptide containing an arginine in the case of a SILAC labeling with heavy arginine). To this aim, each validated result file must be associated with the corresponding raw data file (.wiff files). Intensities of peptide pairs are then extracted from the MS Survey scans of a series of raw data files in batch mode, and heavy/light ratios are computed for each peptide pair. The ratios of all validated peptide matches are averaged for each protein in a gel slice, and a coefficient of variation is calculated for the ratio of the proteins that have been quantified with several peptide matches (Fig. 4*A*). When a protein is identified and quantified several times in consecutive gel slices, a final protein ratio is computed by averaging the different ratios found for this protein in the different fractions, and a global coefficient of variation is calculated. Proteins or protein groups identified and quantified in different fractions are also clustered to generate a final non-redundant list of protein groups, with their normalized protein ratio and the associated global coefficient of variation, presented in the "Quantification report." To check and validate the quantification results, direct links are provided for each protein ratio to a "Quanti Viewer" window showing all data used for quantification of an individual protein. These include the list of isotopically labeled peptide pairs identified for this protein with peptide score, mass, and elution time; the list of MS scans used to extract peptide intensities; and the corresponding MS spectra of the peptide pairs (Fig. 4*B*). The program automatically selects the MS scans of good quality to reconstitute the elution peaks for each member of the peptide pair. Then it computes the elution profile intensities and the corresponding ratio. Another feature of MFPaQ is to manually deselect some MS scans or directly deselect some peptide pairs in the case of aberrant ratio calculation (co-elution with another peptide, weak signal, etc.). The ratios are then automatically recalculated and updated in the quantification report.

Validation of the MFPaQ Quantification Module Using a 1:1 Heavy/Light c-ICAT Ratio of Labeled Proteins Extracted from EC Microsomes

To test the efficiency of ICAT labeling of EC microsomal proteins as well as the efficiency of quantification by the MFPaQ software, we performed a 1:1 heavy/light test labeling experiment using 100 μg of microsomes. Equal amounts of material were labeled with either light or heavy c-ICAT and mixed together. Proteins were solubilized in 6 M urea and 0.5% SDS to improve protein denaturation and labeling efficiency (23) and were then fractionated by 1D SDS-PAGE. Twenty gel slices were cut all along the migration lane and digested with trypsin. For each fraction, c-ICAT labeled peptides were enriched by monomeric avidin chromatography, the biotin moiety of the tag was then submitted to acidic cleavage, and the resulting peptides were analyzed by nano-LC-MS/MS with a 60-min-long gradient. Proteins identified by Mascot in each fraction were extracted and validated with the MFP module of MFPaQ using the optimized criteria described above (*i.e.* at least two peptide matches of rank 1 with score higher than 34 or one peptide match of rank 1 of score higher than 50). The MFPaQ software allowed the validation of 164 unique protein groups (Supplemental Data 3) from which 155 were assigned at least one ICAT labeled peptide match of score higher than 20 (threshold applied on validated peptide matches for MS data intensity extraction). Peptide information associated with each protein group in the different gel slice fractions are shown in Supplemental Data 4, and annotated MS/MS spectra are provided in the case of single peptide-based matches (Supplemental Data 9). Quantification was then performed on the validated protein groups using the

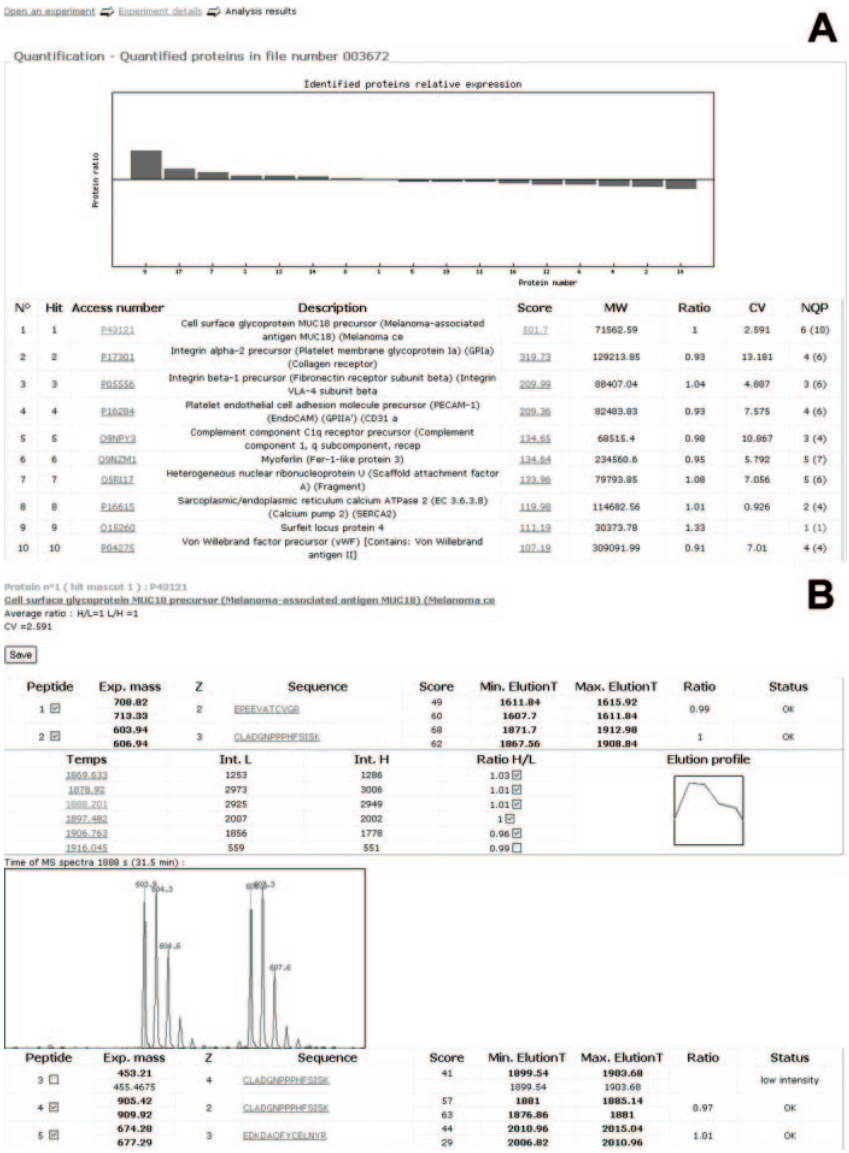


FIG. 4. **Visualization of the peptide and protein ratios from the 1:1 ICAT labeling test experiment using EC microsomes in the quantification module.** *A*, for each validated protein list corresponding to one protein fraction (*i.e.* gel slice), protein ratios are calculated and displayed in an individual window along with the protein scores, numbers of peptides quantified per protein, and coefficient of variation (*CV*) of the protein ratios for this gel slice. *NQP* is the number of quantified peptide pairs. The *Ratio column* refers to the H/L ratios. *B*, detailed results for quantification of a particular protein in one gel slice can be inspected in a separate window showing *m/z*, scores, elution times of the ions used for quantification, and the different MS scans used for quantification for each peptide pair can be inspected. Peptide pairs and MS scans can be manually selected or deselected for a new calculation of the ratio. *Min.*, minimum; *Max.*, maximum; *Exp.*, experimental; *Int.*, intensity; *Temps*, time.

MFPaQ quantification module. After calculation of an average ratio for each protein group identified, the software applies to all of them a normalization factor defined as the median ratio of the protein population. This compensates for a possible bias introduced during labeling if slightly different total protein amounts of the two samples to be compared are taken. In the test experiment, the software calculated a normalization factor of 0.98. This value was expected because the test experiment compared two aliquots of the same sample. Similarly it is also expected that all the normalized ratios for the identified proteins are very close to 1 because no differential expression of proteins occurs. The histogram in Fig. 5*A* presents the normalized ratios computed by the software for all quantified proteins (either heavy/light or light/heavy ratios are represented to always obtain a final value >1). Of the 155 proteins possessing at least one ICAT labeled peptide match, 152 were successfully quantified (Supplemental Data 5). The calculated H/L and L/H ratios for this population vary between 1 and 1.22, which is in good agreement with the classical ~20% accuracy attributed to the c-ICAT labeling method associated with mass spectrometry analysis (8, 24, 25). Standard deviation from the median value of 1 is only 6%. Thus, the results of this test experiment indicate that the ICAT labeling method associated with the analytical mass spectrometry procedure described, database search result filtering using stringent criteria, and quantification with the MFPaQ software may be able to measure changes in ratios in a statistically significant way.

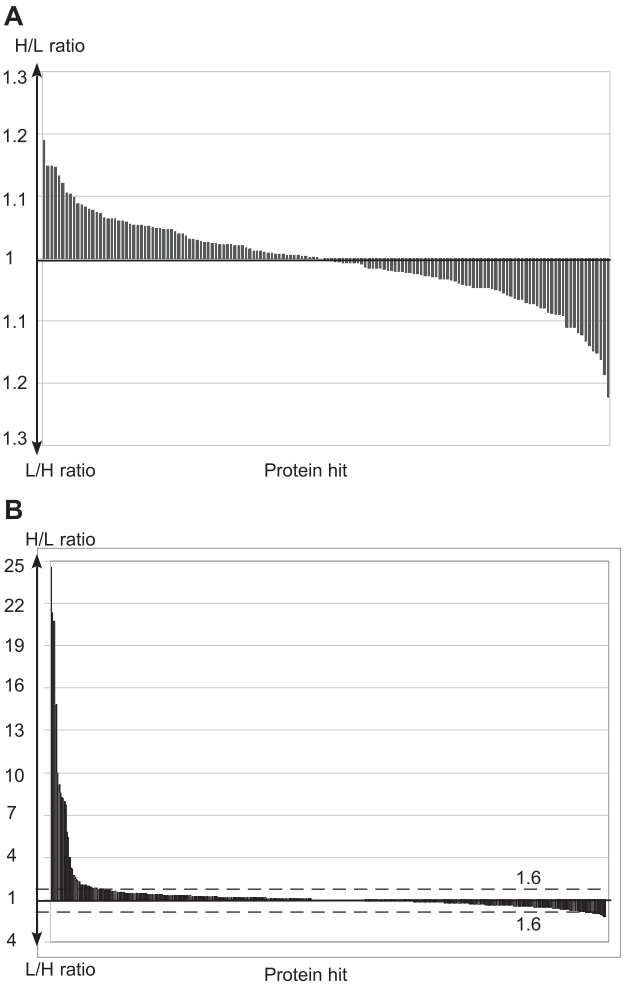


FIG. 5. **Protein relative expression ratios of the final experiment quantification reports.** The *left part* of these histograms displays proteins with heavy/light ratios >1, whereas the *right part* displays proteins with light/heavy ratios >1. *A*, quantification results from the ICAT labeling 1:1 test experiment using EC microsomes. *B*, quantification results from the differential proteomics study following treatment of ECs with proinflammatory cytokines and c-ICAT labeling.

Quantitative Study of EC Membrane Proteins Regulated by Inflammatory Cytokines

ECs in secondary lymphoid organs and chronically inflamed tissues are found in a microenvironment rich in proinflammatory cytokines (21, 26). Therefore, in an effort to mimic the inflammatory microenvironment found *in vivo*, cultured ECs were pretreated with a combination of potent proinflammatory cytokines before ICAT labeling of microsomal proteins. For this differential proteomics study, about 60 μg of microsomal proteins from untreated ECs were labeled with light c-ICAT reagent, and the same amount of microsomal proteins from ECs stimulated with TNF-α, IFNγ, and lymphotoxin-α/β were labeled with heavy c-ICAT reagent. The samples were mixed and fractionated by 1D SDS-PAGE into 17 gel slices, which were digested with trypsin. After enrichment of c-ICAT peptides on a monomeric avidin cartridge and acidic cleavage of the tag, analysis of the peptides was performed by nano-LC-MS/MS with an 80-min-long gradient. The gradient time was increased to improve MS/MS coverage of the peptidic mixture and to maximize the number of proteins identified. Application of the MFPaQ software using the same database search parameters and protein extraction criteria as described above resulted in a final non-redundant list of 229 identified proteins. To maximize the number of quantified proteins in the experiment we then applied less stringent filtering criteria for automatic validation with the MFP module (at least one peptide match of rank 1 with ion score higher than 35, corresponding to $p < 0.05$) and manually checked all ambiguous proteins by close inspection of MS/MS spectra. Criteria for the manual validation of proteins were the following: at least one c-ICAT labeled peptide of rank 1 with relevant MS/MS fragmentation pattern (at least four consecutive y ions) and a good correlation between the theoretical molecular weight of the protein hit and the corresponding molecular weight of the gel slice number. In that way, we obtained a final list of 475 validated unique protein groups (Supplemental Data 6). Peptide information and annotated MS/MS spectra in the case of single peptide-based matches are shown, respectively, in Supplemental Data 7 and 9. From the 475 validated protein groups, 452 had at least one c-ICAT labeled peptide match of score higher than 20. Of them, the MFPaQ software could successfully quantify 415 protein groups (Supplemental Data 8). The normalization factor applied to all protein ratios for this experiment was 0.911, reflecting a 10% error in protein concentration measurement. In the final quantification report, 44 proteins are overexpressed under cytokine treatment with heavy/light ratios between 1.6 and 24.6 (Fig. 5*B*, Table II, and Supplemental Data 8), and on the other hand, 39 proteins are underexpressed with ratios light/heavy between 1.6 and 2.2. The most induced proteins are ICAM-1 (ratio of 25), vascular cell adhesion molecule-1 (VCAM-1; ratio of 21), and E-selectin, which represent major EC proteins involved in inflammation and TNF-α response. ICAM-1, which mediates firm adhesion of leukocytes to the vascular endothelium via interaction with lymphocyte function-associated antigen-1, is well known to be up-regulated on endothelium upon inflammation and has been shown to be important for transendothelial migration of lymphocytes (27, 28). VCAM-1, which is induced on ECs at sites of inflammation, is one of the most important cell adhesion molecules involved in recruitment of monocytes via interaction with monocyte integrin VLA-4 (29, 30). E-selectin mediates leukocyte rolling and is not constitutively expressed on ECs but is up-regulated upon inflammatory stimulation (31, 32). Another cell adhesion protein, the activated leukocyte cell adhesion molecule (ALCAM)/CD166, which localizes to EC junctions and plays a role in monocyte transendothelial migration (33), was also shown to be up-regulated although with a lower ratio (2-fold change). Many

TABLE II

Proteins overexpressed in EC microsomes in response to treatment with proinflammatory cytokines

Kin of IRRE-like protein, kin of irregular chiasm-like protein 1 precursor.

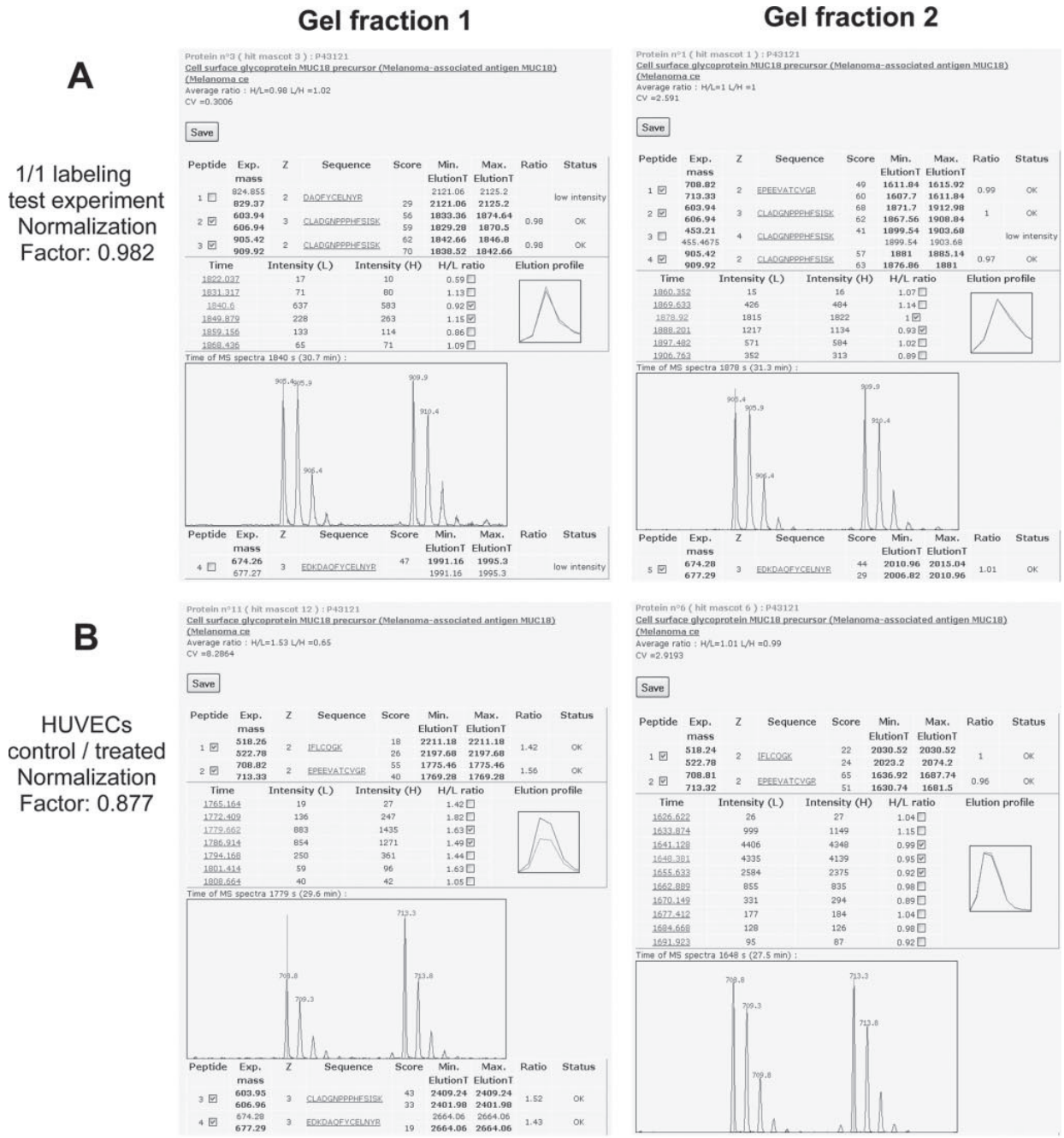
UniProt accession number	Protein name	Protein score[a]	Number of ICAT peptide pairs[a]	Protein ratio[a]	CV[a]	Final average normalized protein ratio[b]	Global CV[b]
					%		%
P05362	Intercellular adhesion molecule-1 precursor (ICAM-1) (CD54)	169	3 (5)	21.2	5.4	24.6	13.4
P19320	Vascular cell adhesion protein-1 precursor (VCAM-1) (CD106)	132	6 (8)	29.5	31.2	21.4	39.8
P16581	E-selectin precursor (endothelial leukocyte adhesion molecule-1)	79	4 (5)	19.2	18.4	20.7	7.9
P20591	Interferon-induced GTP-binding protein Mx1	71	4 (4)	16.5	16.0	14.9	37.2
Q96PP9	Guanylate-binding protein 4	48	1 (1)	17.5		14.8	49.1
P29728	Enoyl-CoA hydratase, mitochondrial precursor	44	2 (2)	8.7	11.1	10.0	
Q5D1D5	Guanylate-binding protein 1	59	2 (3)	8.0	14.8	9.2	
P32455	Interferon-induced guanylate-binding protein 1	37	2 (3)	7.5	6.0	8.6	
P30447	HLA class I histocompatibility antigen, A-23 α chain precursor	172	5 (5)	7.8	10.8	8.3	26.4
Q2A689	MHC[c] class I antigen	121	2 (5)	8.4	0.1	8.2	25.4
P23381	Tryptophanyl-tRNA synthetase (tryptophan-tRNA ligase)	68	4 (4)	7.4	15.9	8.0	8.1
P13747	HLA class I histocompatibility antigen, α chain E precursor	51	1 (1)	6.8		7.7	
P02794	Ferritin heavy chain (ferroxidase, EC 1.16.3.1) (ferritin H subunit)	0[d]	1 (1)	5.1		5.8	
O15162	Phospholipid scramblase 1 (PL scramblase 1)	31	1 (1)	4.7		5.4	
P28838	2′-5′-Oligoadenylate synthetase 2 ((2–5′)oligo(A) synthetase 2)	51	3 (3)	3.5	13.7	3.9	
P48735	Isocitrate dehydrogenase (NADP), mitochondrial precursor	54	1 (1)	2.9		3.3	0.2
Q86YK5	Tumor necrosis factor receptor superfamily member 5 (fragment)	29	2 (2)	2.8	4.8	3.2	
P62745	Rho-related GTP-binding protein RhoB precursor (H6)	118	4 (4)	2.1	21.4	2.7	17.2
P02792	Ferritin light chain (ferritin L subunit)	36	1 (1)	2.3		2.6	
P40261	Nicotinamide *N*-methyltransferase (EC 2.1.1.1)	0[d]	1 (1)	2.1		2.4	
P01130	Low density lipoprotein receptor precursor (LDL receptor)	48	3 (3)	2.1	6.5	2.4	
P10515	Pyruvate dehydrogenase complex E2 subunit	152	4 (4)	1.7	6.6	2.3	20.8
Q1HGM8	Activated leukocyte cell adhesion molecule variant 2	121	4 (4)	1.8	8.7	2.1	
P13473	Lysosome-associated membrane glycoprotein 2 precursor	61	1 (1)	1.8		2.0	
P07996	Thrombospondin-1 precursor	123	7 (8)	1.8	9.4	2.0	
O00330	Pyruvate dehydrogenase protein X component	30	1 (1)	1.8		2.0	
P61224	Ras-related protein Rap-1b precursor (GTP-binding protein smg p21B)	40	1 (1)	1.8		2.0	
Q06210	Glucosamine-fructose-6-phosphate aminotransferase	83	3 (3)	1.7	27.1	2.0	
Q96J84	Kin of IRRE-like protein 1 precursor	52	1 (1)	1.7		2.0	
P30084	HLA class I histocompatibility antigen, A-1 α chain precursor	40	1 (1)	1.6		1.9	
P52597	Heterogeneous nuclear ribonucleoprotein F (hnRNP F)	21	1 (1)	1.6		1.9	
Q38L19	Heat shock protein 60	169	2 (3)	1.8	1.5	1.8	14.5
P31689	DnaJ homolog subfamily A member 1 (heat shock 40-kDa protein 4)	26	1 (1)	1.6		1.8	
P51149	Ras-related protein Rab-7	71	1 (1)	1.6		1.8	2.6
Q8IUW5	Similar to expressed sequence AA536743	43	1 (1)	1.6		1.8	
Q14258	Tripartite motif protein 25 (zinc finger protein 147)	151	2 (2)	1.5	6.8	1.8	5.1
Q5T653	39 S ribosomal protein L2	0[d]	1 (1)	1.5		1.8	
Q9ULA0	Aspartyl aminopeptidase (EC 3.4.11.21)	48	1 (1)	1.5		1.7	
P09543	2′,3′-Cyclic-nucleotide 3′-phosphodiesterase	32	2 (3)	1.5	12.8	1.7	
P13489	Ribonuclease inhibitor (ribonuclease/angiogenin inhibitor 1)	58	4 (4)	1.5	7.2	1.7	
P62820	Ras-related protein Rab-1A (YPT1-related protein)	37	1 (1)	1.5		1.7	
Q7Z457	Poliovirus receptor-related 2 (fragment)	142	1 (1)	1.5		1.7	
Q14764	Major vault protein (MVP) (lung resistance-related protein)	183	5 (5)	1.5	8.6	1.7	1.3
Q6FHV5	RAB8A protein	0[d]	1 (1)	1.4		1.6	

[a] Data related to the protein in the 1D gel slice where it was identified with the best Mascot protein score (major gel slices): Mascot MudPIT protein score, number of ICAT peptides pairs used by the software to quantify the protein in this particular gel slice (the number in parentheses corresponds to the total number of quantified ICAT peptide pairs), ratio computed by the software, and its associated coefficient of variation (CV) in percent (calculated if the number of ICAT peptides pairs used for quantification is higher than 1).

[b] Final protein ratio computed for the protein in the whole experiment after averaging the different ratios found for the protein if it was quantified in different consecutive gel slices and correcting by the normalization factor. A global coefficient of variation (CV; percentage) of this ratio is also calculated if the protein was quantified in different gel slices.

[c] Major histocompatibility complex.

[d] Mascot MudPIT scoring generates protein scores of 0.



		Gel fraction 1	Gel fraction 2	Gel fraction 3	Normalized average protein ratio in the experiment	CV
1/1 test labeling experiment	Number of ICAT pairs quantified	10	4	5	1.027	2.94
	Average ratio of the protein in the fraction	0.999	0.98	1.038		
	CV	2.591	0.301	4.448		
	Normalized average protein ratio in the fraction	1.02	1	1.06		
differential study: HUVECs control vs treated	Number of ICAT pairs quantified	5	6	N.D.	1.463	29.02
	Average ratio of the protein in the fraction	1.53	1.009			
	CV	8.286	2.919			
	Normalized average protein ratio in the fraction	1.76	1.16			

small GTPases (*i.e.* RhoB, Rap1b, Rab7, Rab28, and Rab1a) were also found to be induced by inflammatory cytokines in human primary ECs as well as large GTPases guanylate-binding proteins 1 and 4 and GTP-binding protein Mx1, which have been shown previously to be up-regulated by IFNγ (34, 35). Other known interferon-induced proteins identified (Table II) included HLA class I molecules, 2′-5′-oligoadenylate synthetase (36, 37), tryptophanyl-tRNA synthetase (38), and phospholipid scramblase 1 (39, 40). Finally expression of cell adhesion molecules CD31/PECAM-1 and ICAM-2 as well as many other cell surface proteins (Supplemental Data 8) remained unchanged after TNF-α, IFNγ, and lymphotoxin α/β stimulation of primary human ECs.

Differential Induction of CD146 Isoforms in ECs Treated with Proinflammatory Cytokines

Very often a protein can be identified in several consecutive gel slices, particularly in the case of very abundant species, which will for example show significant tailing on a 1D gel lane. In this case, the ratios calculated for this protein in the different gel slices should be similar, and a low coefficient of variation on the final global protein ratio will indicate a good accuracy in quantification of the protein. However, different isoforms or fragments of a protein can also be identified in different gel slices, and although they will eventually belong to the same protein group as they will match the same set of peptides, the ratios calculated for each of them may actually differ. In that case, the final global ratio calculated for the protein group will be associated to a high coefficient of variation value indicating a discrepancy between individual gel slice calculated ratios, but this may reflect biologically relevant information. An example of such a case is given in Fig. 6 for the MUC18/CD146 protein, an immunoglobulin superfamily adhesion molecule and component of EC junctions involved in cell-cell cohesion and angiogenesis (41–43). In the 1:1 test labeling experiment, this protein is quantified with ratios close to 1 in each of the three gel slices where it was identified, leading to a final protein ratio of 1.02 with a low global coefficient of variation of about 3%. On the other hand, in the differential proteomics study following cytokine stimulation, the ratio calculated for this protein in gel slice 1 (high molecular weight fraction) is 1.76 (after correction with the normalization factor), whereas it is only 1.16 in gel slice 2 (low molecular weight fraction), leading to a final protein ratio of 1.46 with a high global coefficient of variation of about 29%. The coefficients of variation computed for the ratios in each of the two fractions are quite low, reflecting a good correlation between the values obtained for all the peptide pairs used for calculation of these ratios. Thus, the discrepancy between the protein ratios obtained in the two fractions may reflect a real difference in regulation of two distinct protein isoforms following cytokine treatment rather than a bad quantification. It could be assumed for example that a highly glycosylated form of the MUC18/CD146 is specifically induced in response to inflammatory signals.

TABLE III
Comparison of software features for bioinformatics analysis of proteomics data

	Protein identification	Result validation	Result grouping	Quantification
Mascot	+	−	−	−
Sequest	+	−	−	−
Phenyx	+	+	−	−
TPP[a]	−	+	−	+
MSQuant	−	+	−	+
STEM	−	+	+[b]	+
MFPaQ	−	+	+	+

[a] Trans Proteomic Pipeline with ProteinProphet and PeptideProphet included.
[b] Grouping for validation results and not for quantification results.

DISCUSSION

In this study, we developed a new software tool, designated MFPaQ, that proved to be efficient for data validation and quantification after ICAT labeling, protein fractionation, analysis of consecutive fractions by several nano-LC-MS/MS runs, and multisearch with the Mascot engine. First this software greatly facilitated the sorting of protein lists and the verification of Mascot result files. Indeed although several search engines like Mascot, Sequest, or Phenyx are usually considered to be very efficient for protein identification, false protein assignments are clearly not avoided. In the case of the Mascot search engine, improvements were obtained in the 2.0 version with the introduction of the MudPIT scoring mode. Our study of the membrane proteome from ECs shows that application of this scoring yielded much fewer false positive hits than the Standard scoring. However, even with this new scoring, a validation step involving parsing of the results, either manually by the user or by application of automatic filters, still appears to be necessary. Convenient tools are not always available inside the identification softwares themselves to perform this task (Table III). For example, a unique filtering option can be selected in Mascot 2.0 and Mascot 2.1 to retain in the final list of proteins only those that have at least one bold and red peptide match. The filtering rules available in MFPaQ are more comprehensive because the number, the

Fig. 6. **Quantification of the MUC18/CD146 protein.** *A*, 1:1 labeling test experiment. *B*, differential proteomics study following treatment of ECs with proinflammatory cytokines. Shown are individual quantification windows showing detailed results for the protein MUC18/CD146 in two different gel slices for each experiment: *m/z*, scores, and elution time of the ions used for quantification and the ratio calculated for each peptide pair. MS scans corresponding to ions at *m/z* 905.42 (peptide CLADGNPPPHFSISK, 2+) or at *m/z* 708.82 (peptide EPEEVATCVGR, 2+) are displayed. *Min.*, minimum; *Max.*, maximum; *Exp.*, experimental; *CV*, coefficient of variation.

rank, and the score of the peptide matches can be specified, and different filtering rules can be applied to validate the proteins. By performing a second Mascot search with the MS/MS data in a reversed database and by applying to the results the same filtering rules, the user can obtain a rough evaluation of the percentage of false positives associated to the automatic validation step. Thus, it is possible to adapt the stringency of the filtering rules to minimize the number of false positives while retaining a maximum of identified proteins.

Other bioinformatics tools can perform proteomics data validation (Table III), like the Trans Proteomic Pipeline, which is based on the softwares PeptideProphet and ProteinProphet (16). These two softwares are powerful validation tools that were initially designed to sort, filter, and analyze the results of the Sequest search engine. They first assign measures of confidence to peptide sequences returned by Sequest, via a statistical data modeling algorithm, and then to the proteins from which they were likely derived, thus estimating the accuracy of peptide and protein identifications made. They have been very efficiently applied in large scale shotgun proteomics studies based on peptide fractionation and MS/MS data analysis using Sequest (44), but they do not appear to be suited for validation of Mascot results. Other programs like MSQuant (18) and STEM (19) offer functionalities to validate Mascot data files (Table III). However, one particular advantage of MFPaQ is that it provides a synthetic view of the identifications that can be obtained when using a shotgun strategy based on protein fractionation. Indeed several Mascot result files can be automatically parsed in batch mode with the MFP module and grouped afterward to generate a global concatenated, non-redundant list of identified protein groups.

Finally an important feature of MFPaQ is the quantification module, which provides data on protein relative expression following isotopic labeling and identification with Mascot. Some recently released commercial softwares offer the possibility to perform quantification for isotope labeling methods, like ProteinPilot from Applied Biosystems and ProteinScape from Bruker Daltonics. However, they are not always of generic use and run under a specific environment. ProteinPilot, for example, offers new functionalities both for parsing and quantifying the data from .wiff files in ICAT, iTRAQ, and SILAC but is based mainly on the results of the Paragon search engine and not on Mascot results. ProteinScape, for its part, can process the MS/MS data with several search engines, among which is Mascot, but is only designed to perform quantification on Bruker Daltonics raw files. The very latest version of Mascot, Mascot 2.2, now seems able to perform quantification but only based on the data contained in the MS/MS peak lists (*e.g.* iTRAQ quantification or semiquantitative label-free strategies based on peptide match counting). To perform MS-based quantification (*e.g.* ICAT or SILAC strategies), the intensity values for the peptides should be extracted from the raw data by another commercial program, Mascot Distiller. Although this application indeed seems promising to handle a wide range of mass spectrometer data file formats, the quantitation features are not yet implemented. It will be achieved using the Mascot Distiller Quantitation Toolbox, a program able to perform quantitation based on the relative intensities of extracted precursor ion chromatograms. The open source software MSQuant can do that and now works with a variety of MS data files formats but is specifically designed for SILAC analyses. MFPaQ has the advantage to process either SILAC or ICAT data. Moreover the MFP module and the quantification module of MFPaQ are well suited to easily manage an experiment constituted of multiple Mascot search result data files, corresponding for example to several protein fractions. This is an important feature because protein fractionation is very often performed in differential proteomics studies based on isotopic labeling. Indeed such studies usually proceed in two steps: first, as many proteins as possible are identified by MS/MS and database searching; and second, some of these proteins can be quantified if they were identified with a peptide bearing the isotopic modification by extracting the intensities of the peptide pair from raw MS data. This means that in such approaches only proteins that were identified first can potentially be quantified afterward. Thus, although very good quantification may be achieved on major protein components of a complex mixture, variation of expression of minor protein components may well be missed because these species will not be identified. This represents a major drawback, particularly if changes are expected to occur on low abundance species. It is thus critical in these strategies to extensively characterize the sample and to identify as many proteins as possible to track variations on a maximum number of species. A classical way for that is to perform a shotgun analysis of the sample, for example by protein fractionation, which currently seems to constitute the most efficient method to maximize the analytical coverage of a highly complex protein mixture. Thus, it is important that bioinformatics tools for data quantification can process and integrate data obtained after protein fractionation. The MFPaQ software is particularly useful for that in contrast to other quantification programs (Table III). Indeed it can generate a global quantification report to integrate and synthesize all the data obtained for a protein in the different fractions in which it could be identified and quantified. Display of coefficients of variation for protein ratios in each fraction makes it possible to track potential errors in quantification due for example to erroneous calculation of a specific peptide pair ratio and to exclude them from quantification to improve the final result. Moreover display of the global coefficient of variation on the final averaged protein ratio allows the evaluation of the statistical significance of the final value calculated by the software.

Here we applied the MFPaQ software to characterize the membrane proteome of human ECs and the variation of protein expression profile in response to cytokine stimulation. The MFP module of the MFPaQ software proved to be very

useful for the proteomics analysis of EC membrane proteins. More than 600 proteins were identified after fractionation of the crude microsomal fraction from primary human ECs by 1D SDS-PAGE and nano-LC-MS/MS analysis (Supplemental Data 1). More than 55% of these proteins are membrane proteins according to automatic bioinformatics classification; this represents a relatively good enrichment of the membrane proteome compared with similar studies (45). The list of identified proteins comprises at least 41 known endothelial cell surface markers (Supplemental Data 1), including classical endothelial markers such as CD31/PECAM-1, VE-cadherin, ICAM-2, Tie-2, CD146/MUC18, podocalyxin, endothelial cell-selective adhesion molecule, angiotensin-converting enzyme/CD143, endothelin-converting enzyme, dipeptidyl-peptidase IV/CD26, and ALCAM/CD166. Strikingly although all these classical EC markers were identified in a proteomics study of luminal EC plasma membrane proteins freshly isolated from rat lungs *in vivo*, they were not found in EC surface proteins purified from cultured rat lung microvascular ECs (46). Therefore, our results indicate that cultured primary human ECs (HUVECs), a widely used *in vitro* EC model first described in 1973 (47), retain a cell surface phenotype closer to the *in vivo* EC phenotype than cultured rat lung ECs. In addition, our results suggest that the number of EC membrane proteins that differ between ECs *in vivo* and *in vitro*, previously suggested to be ~50% (46), may have been overestimated.

To the best of our knowledge, this is the first proteomics analysis of EC membrane proteins regulated by inflammatory cytokines. We used a combination of key proinflammatory mediators (TNF-α, IFNγ, and lymphotoxin α/β) to mimic the inflammatory microenvironment and performed a differential quantitative proteomics study using the ICAT method and the quantification module of the MFPaQ software. Our results revealed that ICAM-1, VCAM-1, and E-selectin, three critical cell adhesion molecules for leukocyte-endothelium interactions in inflammation (27), are the major EC membrane proteins up-regulated by inflammatory stimuli and the only ones induced more than 15-fold. These proteomics results are fully consistent with previous microarray data showing that ICAM-1 (-fold change, 111.9), E-selectin (-fold change, 48.0), and VCAM-1 (-fold change, 31.7) mRNAs were the most significantly induced after TNF-α treatment of human primary ECs (48). E-selectin, ICAM-1, and VCAM-1 mediate the initial rolling and arrest steps of leukocyte-EC interactions (27), which are followed by leukocyte transendothelial migration through EC junctions. Interestingly we identified two components of EC junctions that are regulated by proinflammatory mediators in human primary ECs. These two molecules, CD146 high molecular weight isoform and ALCAM/CD166, belong to the same protein family of immunoglobulin cell adhesion molecules, consisting of five extracellular immunoglobulin domains, a single transmembrane domain, and a short cytoplasmic tail, and may function in cell-cell cohesion (33, 41). The up-regulation of these proteins in ECs treated with proinflammatory cytokines may therefore play an important role in the response of ECs to inflammation at the level of EC junctions and leukocyte transendothelial migration.

In conclusion, the present work validates the use of a new software tool for fast and efficient parsing of proteomics results obtained from several Mascot files and extraction of quantitative data from raw MS files in isotopic labeling strategies using either the ICAT or SILAC technique. Developments are in progress to adapt this software to additional types of labeling strategies including iTRAQ labeling and ^{15}N metabolic labeling. Moreover a clear perspective of development for the application is to improve its compatibility with different MS platforms. The first module of the software, the Mascot File Parser, runs independently of the MS acquisition software and is thus of general use for proteomics platforms equipped with various instruments. The current quantification module, for its part, is dedicated to process .wiff data files generated on QStar instruments by the Analyst QS software. Future versions of this module will be compatible with data files acquired with different types of mass spectrometers and with different MS acquisition softwares. The MFPaQ software, as well as all validated proteomics data associated with this study, are freely available at mfpaq.sourceforge.net.

Acknowledgments—We are grateful to L. Canelle, K. Chaoui, M. Evain, C. Froment, M. Matondo, A. Stella, M. P. Bousquet-Dubouch, S. Uttenweiler-Joseph, and C. Gaspin for β testing of the MFPaQ software and fruitful discussions.

* This work was supported in part by grants from Région Midi-Pyrénées, MAIN European Network of Excellence Grant FP6-502935, and the Génopole Toulouse Midi-Pyrénées. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[S] The on-line version of this article (available at http://www.mcponline.org) contains supplemental material.

§ Both authors contributed equally to this work.

¶ To whom correspondence should be addressed: Inst. de Pharmacologie et de Biologie Structurale, CNRS UMR 5089, 205 route de Narbonne, 31077 Toulouse, France. Tel.: 33-5-61-17-55-41; Fax: 33-5-61-17-59-94; E-mail: anne.gonzalez-de-peredo@ipbs.fr.

REFERENCES

1. Forner, F., Foster, L. J., Campanaro, S., Valle, G., and Mann, M. (2006) Quantitative proteomic comparison of rat mitochondria from muscle, heart, and liver. *Mol. Cell. Proteomics* **5,** 608–619
2. Wang, G., Wu, W. W., Zeng, W., Chou, C. L., and Shen, R. F. (2006) Label-free protein quantification using LC-coupled ion trap or FT mass spectrometry: reproducibility, linearity, and application with complex proteomes. *J. Proteome Res.* **5,** 1214–1223
3. Listgarten, J., and Emili, A. (2005) Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Mol. Cell. Proteomics* **4,** 419–434
4. Radulovic, D., Jelveh, S., Ryu, S., Hamilton, T. G., Foss, E., Mao, Y., and Emili, A. (2004) Informatics platform for global proteomic profiling and biomarker discovery using liquid chromatography-tandem mass spectrometry. *Mol. Cell. Proteomics* **3,** 984–997
5. Prakash, A., Mallick, P., Whiteaker, J., Zhang, H., Paulovich, A., Flory, M., Lee, H., Aebersold, R., and Schwikowski, B. (2006) Signal maps for mass spectrometry-based comparative proteomics. *Mol. Cell. Proteomics* **5,** 423–432

6. Ong, S. E., Foster, L. J., and Mann, M. (2003) Mass spectrometric-based approaches in quantitative proteomics. *Methods* **29,** 124–130
7. Heck, A. J., and Krijgsveld, J. (2004) Mass spectrometry-based quantitative proteomics. *Expert Rev. Proteomics* **1,** 317–326
8. Yi, E. C., Li, X. J., Cooke, K., Lee, H., Raught, B., Page, A., Aneliunas, V., Hieter, P., Goodlett, D. R., and Aebersold, R. (2005) Increased quantitative proteome coverage with $^{13}C/^{12}C$-based, acid-cleavable isotope-coded affinity tag reagent and modified data acquisition scheme. *Proteomics* **5,** 380–387
9. Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **17,** 994–999
10. Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., and Mann, M. (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **1,** 376–386
11. Pappin, D. J., Hojrup, P., and Bleasby, A. J. (1993) Rapid identification of proteins by peptide-mass fingerprinting. *Curr. Biol.* **3,** 327–332
12. Tabb, D. L., Eng, J. K., and Yates, J. R., III (2000) Protein identification by SEQUEST, in *Proteome Research: Mass Spectrometry* (James, P., ed) pp. 125–142, Springer-Verlag, New York
13. Savitski, M. M., Nielsen, M. L., and Zubarev, R. A. (2005) New data base-independent, sequence tag-based scoring of peptide MS/MS data validates Mowse scores, recovers below threshold data, singles out modified peptides, and assesses the quality of MS/MS techniques. *Mol. Cell. Proteomics* **4,** 1180–1188
14. Nesvizhskii, A. I., Keller, A., Kolker, E., and Aebersold, R. (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **75,** 4646–4658
15. Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74,** 5383–5392
16. von Haller, P. D., Yi, E., Donohoe, S., Vaughn, K., Keller, A., Nesvizhskii, A. I., Eng, J., Li, X. J., Goodlett, D. R., Aebersold, R., and Watts, J. D. (2003) The application of new software tools to quantitative protein profiling via isotope-coded affinity tag (ICAT) and tandem mass spectrometry: II. Evaluation of tandem mass spectrometry methodologies for large-scale protein analysis, and the application of statistical tools for data analysis and interpretation. *Mol. Cell. Proteomics* **2,** 428–442
17. Colinge, J., Masselot, A., Giron, M., Dessingy, T., and Magnin, J. (2003) OLAV: towards high-throughput tandem mass spectrometry data identification. *Proteomics* **3,** 1454–1463
18. Andersen, J. S., Wilkinson, C. J., Mayor, T., Mortensen, P., Nigg, E. A., and Mann, M. (2003) Proteomic characterization of the human centrosome by protein correlation profiling. *Nature* **426,** 570–574
19. Shinkawa, T., Taoka, M., Yamauchi, Y., Ichimura, T., Kaji, H., Takahashi, N., and Isobe, T. (2005) STEM: a software tool for large-scale proteomic data analyses. *J. Proteome Res.* **4,** 1826–1831
20. Cines, D. B., Pollak, E. S., Buck, C. A., Loscalzo, J., Zimmerman, G. A., McEver, R. P., Pober, J. S., Wick, T. M., Konkle, B. A., Schwartz, B. S., Barnathan, E. S., McCrae, K. R., Hug, B. A., Schmidt, A. M., and Stern, D. M. (1998) Endothelial cells in physiology and in the pathophysiology of vascular disorders. *Blood* **91,** 3527–3561
21. Middleton, J., Americh, L., Gayon, R., Julien, D., Aguilar, L., Amalric, F., and Girard, J. P. (2004) Endothelial cell phenotypes in the rheumatoid synovium: activated, angiogenic, apoptotic and leaky. *Arthritis Res. Ther.* **6,** 60–72
22. Lacorre, D. A., Baekkevold, E. S., Garrido, I., Brandtzaeg, P., Haraldsen, G., Amalric, F., and Girard, J. P. (2004) Plasticity of endothelial cells: rapid dedifferentiation of freshly isolated high endothelial venule endothelial cells outside the lymphoid tissue microenvironment. *Blood* **103,** 4164–4172
23. Ramus, C., Gonzalez de Peredo, A., Dahout, C., Gallagher, M., and Garin, J. (2006) An optimized strategy for ICAT quantification of membrane proteins. *Mol. Cell. Proteomics* **5,** 68–78
24. Chou, J., Choudhary, P. K., and Goodman, S. R. (2006) Protein profiling of sickle cell versus control RBC core membrane skeletons by ICAT technology and tandem mass spectrometry. *Cell. Mol. Biol. Lett.* **11,** 326–337
25. Molloy, M. P., Donohoe, S., Brzezinski, E. E., Kilby, G. W., Stevenson, T. I., Baker, J. D., Goodlett, D. R., and Gage, D. A. (2005) Large-scale evaluation of quantitative reproducibility and proteome coverage using acid cleavable isotope coded affinity tag mass spectrometry for proteomic profiling. *Proteomics* **5,** 1204–1208
26. Girard, J. P., and Springer, T. A. (1995) High endothelial venules (HEVs): specialized endothelium for lymphocyte migration. *Immunol. Today* **16,** 449–457
27. Springer, T. A. (1994) Traffic signals for lymphocyte recirculation and leukocyte emigration: the multistep paradigm. *Cell* **76,** 301–314
28. Staunton, D. E., Marlin, S. D., Stratowa, C., Dustin, M. L., and Springer, T. A. (1988) Primary structure of ICAM-1 demonstrates interaction between members of the immunoglobulin and integrin supergene families. *Cell* **52,** 925–933
29. Davies, M. J., Gordon, J. L., Gearing, A. J., Pigott, R., Woolf, N., Katz, D., and Kyriakopoulos, A. (1993) The expression of the adhesion molecules ICAM-1, VCAM-1, PECAM, and E-selectin in human atherosclerosis. *J. Pathol.* **171,** 223–229
30. Elices, M. J., Osborn, L., Takada, Y., Crouse, C., Luhowskyj, S., Hemler, M. E., and Lobb, R. R. (1990) VCAM-1 on activated endothelium interacts with the leukocyte integrin VLA-4 at a site distinct from the VLA-4/fibronectin binding site. *Cell* **60,** 577–584
31. Bevilacqua, M. P., Stengelin, S., Gimbrone, M. A., Jr., and Seed, B. (1989) Endothelial leukocyte adhesion molecule 1: an inducible receptor for neutrophils related to complement regulatory proteins and lectins. *Science* **243**, 1160–1165
32. Bevilacqua, M. P. (1993) Endothelial-leukocyte adhesion molecules. *Annu. Rev. Immunol.* **11,** 767–804
33. Masedunskas, A., King, J. A., Tan, F., Cochran, R., Stevens, T., Sviridov, D., and Ofori-Acquah, S. F (2006) Activated leukocyte cell adhesion molecule is a component of the endothelial junction involved in transendothelial monocyte migration. *FEBS Lett.* **580**, 2637–2645
34. Naschberger, E., Bauer, M., and Sturzl, M. (2005) Human guanylate binding protein-1 (hGBP-1) characterizes and establishes a non-angiogenic endothelial cell activation phenotype in inflammatory diseases. *Adv. Enzyme Regul.* **45,** 215–227
35. Sahni, G., and Samuel, C. E. (1986) Mechanism of interferon action. Expression of vesicular stomatitis virus G gene in transfected COS cells is inhibited by interferon at the level of protein synthesis. *J. Biol. Chem.* **261,** 16764–16768
36. Roberts, W. K., Hovanessian, A., Brown, R. E., Clemens, M. J., and Kerr, I. M. (1976) Interferon-mediated protein kinase and low-molecular-weight inhibitor of protein synthesis. *Nature* **264,** 477–480
37. Rebouillat, D., and Hovanessian, A. G. (1999) The human 2′,5′-oligoadenylate synthetase family: interferon-induced proteins with unique enzymatic properties. *J. Interferon Cytokine Res.* **19,** 295–308
38. Rubin, B. Y., Anderson, S. L., Xing, L., Powell, R. J., and Tate, W. P. (1991) Interferon induces tryptophanyl-tRNA synthetase expression in human fibroblasts. *J. Biol. Chem.* **266,** 24245–24248
39. Der, S. D., Zhou, A., Williams, B. R., and Silverman, R. H. (1998) Identification of genes differentially regulated by interferon α, β, or γ using oligonucleotide arrays. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 15623–15628
40. Dong, B., Zhou, Q., Zhao, J., Zhou, A., Harty, R. N., Bose, S., Banerjee, A., Slee, R., Guenther, J., Williams, B. R., Wiedmer, T., Sims, P. J., and Silverman, R. H. (2004) Phospholipid scramblase 1 potentiates the antiviral activity of interferon. *J. Virol.* **78,** 8983–8993
41. Bardin, N., Anfosso, F., Masse, J. M., Cramer, E., Sabatier, F., Le Bivic, A., Sampol, J., and Dignat-George, F. (2001) Identification of CD146 as a component of the endothelial junction involved in the control of cell-cell cohesion. *Blood* **98,** 3677–3684
42. Xie, S., Luca, M., Huang, S., Gutman, M., Reich, R., Johnson, J. P., and Bar-Eli, M. (1997) Expression of MCAM/MUC18 by human melanoma cells leads to increased tumor growth and metastasis. *Cancer Res.* **57,** 2295–2303
43. Yan, X., Lin, Y., Yang, D., Shen, Y., Yuan, M., Zhang, Z., Li, P., Xia, H., Li, L., Luo, D., Liu, Q., Mann, K., and Bader, B. L. (2003) A novel anti-CD146 monoclonal antibody, AA98, inhibits angiogenesis and tumor growth. *Blood* **102,** 184–191
44. von Haller, P. D., Yi, E., Donohoe, S., Vaughn, K., Keller, A., Nesvizhskii, A. I., Eng, J., Li, X. J., Goodlett, D. R., Aebersold, R., and Watts, J. D. (2003) The application of new software tools to quantitative protein profiling via isotope-coded affinity tag (ICAT) and tandem mass spectrometry: I. Statistically annotated datasets for peptide sequences and proteins identified via the application of ICAT and tandem mass spectrometry to proteins copurifying with T cell lipid rafts. *Mol. Cell. Proteom-*

ics **2,** 426–427
45. Nielsen, P. A., Olsen, J. V., Podtelejnikov, A. V., Andersen, J. R., Mann, M., and Wisniewski, J. R. (2005) Proteomic mapping of brain plasma membrane proteins. *Mol. Cell. Proteomics* **4,** 402–408
46. Durr, E., Yu, J., Krasinska, K. M., Carver, L. A., Yates, J. R., Testa, J. E., Oh, P., and Schnitzer, J. E. (2004) Direct proteomic mapping of the lung microvascular endothelial cell surface in vivo and in cell culture. *Nat. Biotechnol.* **22,** 985–992
47. Jaffe, E. A., Nachman, R. L., Becker, C. G., and Minick, C. R. (1973) Culture of human endothelial cells derived from umbilical veins. Identification by morphologic and immunologic criteria. *J. Clin. Investig.* **52,** 2745–2756
48. Murakami, T., Mataki, C., Nagao, C., Umetani, M., Wada, Y., Ishii, M., Tsutsumi, S., Kohro, T., Saiura, A., Aburatani, H., Hamakubo, T., and Kodama, T. (2000) The gene expression profile of human umbilical vein endothelial cells stimulated by tumor necrosis factor α using DNA microarray analysis. *J. Atheroscler. Thromb.* **7,** 39–44

MFPaQ version 4 : évolution du logiciel pour la gestion des données haute-résolution

III-1. Introduction

Depuis le début de ma thèse en 2006 de nouvelles technologies ont été développées au niveau instrumental. L'évolution majeure de ces dernières années fut la création d'une nouvelle génération d'instrument, le LTQ-Orbitrap, capable de réaliser des analyses MS à haute résolution (grâce à l'analyseur Orbitrap) et disposant par ailleurs d'une vitesse de séquençage élevée (grâce à la trappe linéaire LTQ). Il existait déjà depuis quelques années des appareils capables de réaliser des acquisitions MS à haute résolution tels que le FT-ICR. Cependant ces derniers nécessitaient une infrastructure particulière (liée à la technologie magnétique mise en œuvre) et réalisaient des temps de cycles (MS et MS/MS) plus longs que sur ces appareils de nouvelle génération.

La démocratisation de la haute-résolution a accompagné le développement de méthodes d'analyse quantitative basées sur le traitement du signal MS sans utilisation de marquage isotopique. Cette nouvelle approche dite « label-free » est en effet plus facile à mettre en œuvre sur des données à haute-résolution, permettant d'obtenir une très grande précision de masse sur les ions peptidiques, que sur des données issues d'appareils à basse résolution, où la comparaison entre différentes analyses peut générer un grand nombre de faux appariements de signaux peptidiques.

Ces améliorations instrumentales ont également eu un impact positif sur les analyses mettant en œuvre un marquage isotopique. En effet, ce type d'approche multiplie par 2 la complexité des scans MS car chaque peptide apparaît sous la forme d'un double massif isotopique. L'accès à une lecture plus précise des signaux a permis de diminuer le nombre d'interférences entre les signaux (appelés aussi épaulement de pics) et donc d'obtenir une quantification plus fiable.

L'inconvénient principal associé à cette évolution technologique fut l'augmentation considérable du nombre de spectres acquis et par conséquent celle du volume de données associé (environ 50 Mo pour un fichier QStar et 500 Mo pour un fichier LTQ-Orbitrap XL). Le deuxième inconvénient a été celui du format du fichier d'acquisition. J'avais développé un module de quantification destiné au format .wiff d'Applied Biosystems. Ce module était par conséquent incompatible avec le format .raw de ThermoFinnigan qui nécessite d'utiliser des bibliothèques de code spécifiques pour accéder à son contenu. J'ai donc dû revoir entièrement le code permettant d'accéder aux signaux MS présents dans ce nouveau type de fichier.

III-2. Nouveau module de Quantification

3.2.1 Création d'une nouvelle bibliothèque pour l'accès aux signaux MS

Peu avant le développement de cette nouvelle génération instrumentale, le format mzXML (Pedrioli, Eng et al. 2004) a été proposé par l'ISB comme standard ouvert pour le stockage des données brutes.

L'objectif était de disposer d'un format de fichier facilement accessible (le format XML étant pris en charge par la plupart des langages de programmation) et également d'un convertisseur pour chaque format propre à chaque instrument. Ce nouveau standard apparaissait donc comme une solution intéressante pour l'écriture de la nouvelle version du module de quantification. La transformation d'un fichier brut au format mzXML s'accompagne cependant de deux inconvénients. Le premier est celui du temps de conversion qui était à l'époque de 30 minutes en moyenne pour un fichier LTQ-Orbitrap XL. Dans le cas d'analyses de gel 1D comportant par exemple une quarantaine de fractions cela pouvant donc représenter jusqu'à 20 heures de calcul. Des techniques de parallélisation peuvent bien sûr être mises en œuvre pour diminuer cette durée mais elles nécessitent de disposer d'un parc informatique adapté (cluster ou grille d'ordinateurs). Ainsi, en plus du support du format mzXML, il m'est apparu judicieux de considérer également la possibilité de lire le fichier brut directement. Deux options d'accès aux données étaient alors envisageables. La première était semblable à celle que j'avais choisi pour le format .wiff c'est-à-dire utiliser les bibliothèques de code du constructeur. Je n'étais cependant pas tout à fait satisfait par cette solution car elle ne permet pas de manipuler les spectres depuis le langage principal de l'application (i.e. Perl) et elle limite la portabilité du logiciel à la plateforme Windows. J'ai donc opté pour une seconde solution qui a consisté à décoder la structure du fichier brut produit par les LTQ-Orbitrap (format .RAW).

Pendant toute la durée du gradient chromatographique, le logiciel d'acquisition enregistre les spectres de façon séquentielle au sein d'un fichier. Le format de ce dernier, dans le cas d'une analyse Orbitrap, repose sur une structure binaire contenant une suite d'objets correspondant aux données acquises en série. Afin de décoder cette structure mon travail a consisté à identifier ces objets ainsi que le nombre, le type et la dimension des variables associées. J'ai réussi à identifier les informations majeures nécessaires à l'extraction du signal à savoir :

- la position et l'identifiant de chaque spectre dans le fichier,
- le temps d'élution associé à chacun de ces spectres,
- le rapport m/z et l'intensité de chaque point présent dans un spectre.

La compréhension de la structure des fichier .RAW LTQ-Orbitrap m'a permis de développer une bibliothèque Perl capable d'accéder au contenu des spectres présents dans le fichier. Cette bibliothèque est donc capable de lire à la fois des fichiers .mzXML et .RAW en utilisant les mêmes instructions pour l'accès aux données (on parle de « Application Programming Interface » ou API). Elle a donc servi d'élément de base pour la construction du nouveau module de quantification. Ce dernier était ainsi assuré de disposer d'une longévité plus importante que celui développé spécifiquement pour les fichiers du QStar.

La nature des données produites par ce nouveau type d'instruments a également été source de d'une refonte des algorithmes liés à l'extraction du signal ou « peak picking ». En effet, contrairement au QStar, le LTQ-Orbitrap génère des données MS en haute résolution rendant ainsi incompatible les stratégies d'analyses que j'avais employées auparavant. A partir de la bibliothèque d'accès aux données précédemment décrite j'ai donc développé un nouveau module de quantification dédié à l'analyse de données haute-résolution. Les pics étant mieux résolus, leur largeur à mi-hauteur est nettement plus faible et le risque d'épaulement est également diminué. Ainsi au lieu de chercher à intégrer l'aire sous le pic, l'utilisation de la valeur maximale du pic comme

valeur quantitative devient tout aussi efficace. En effet, dans un modèle gaussien théorique il existe une proportionnalité directe entre la hauteur d'un pic (H) et l'aire (A) du pic :

$$A = \sqrt{2\,\pi}\,\sigma H$$

Avec sigma (σ) correspondant à la déviation standard et pouvant être déduit de la largeur à mi-hauteur (FWHM) via l'équation suivante :

$$FWHM = 2\sqrt{2\ln(2)}\,\sigma \approx 2.3548\,\sigma$$

Si l'on considère que la variation de sigma est négligeable (même largeur à mi-hauteur) avec des appareils de haute résolution alors on peut approximer l'aire du pic uniquement à partir sa hauteur.
En pratique c'est cette valeur d'intensité maximale qui est stockée pour chacun des pics dans le fichier .raw issu d'un LTQ-Orbitrap. Ainsi l'extraction du signal est accélérée car l'étape d'intégration du signal n'est plus nécessaire.
J'ai cependant pu tirer parti de mes précédents travaux pour un certain nombre de concepts qui restaient inchangés. C'est le cas notamment de l'étape préliminaire du calcul des possibilités d'appariement de peptides marqués/non marqués.

3.2.2 Développement d'une interface graphique « riche »

Le nouveau module étant écrit entièrement en Perl il est désormais possible de réaliser toutes les étapes du processus de quantification au niveau de l'interface web. En effet, dans l'ancienne version il était nécessaire d'utiliser une application annexe pour réaliser l'extraction des signaux (cf. partie II-3). La saisie du plan d'expérience se fait désormais en remplissant un simple formulaire web (cf figure 24).

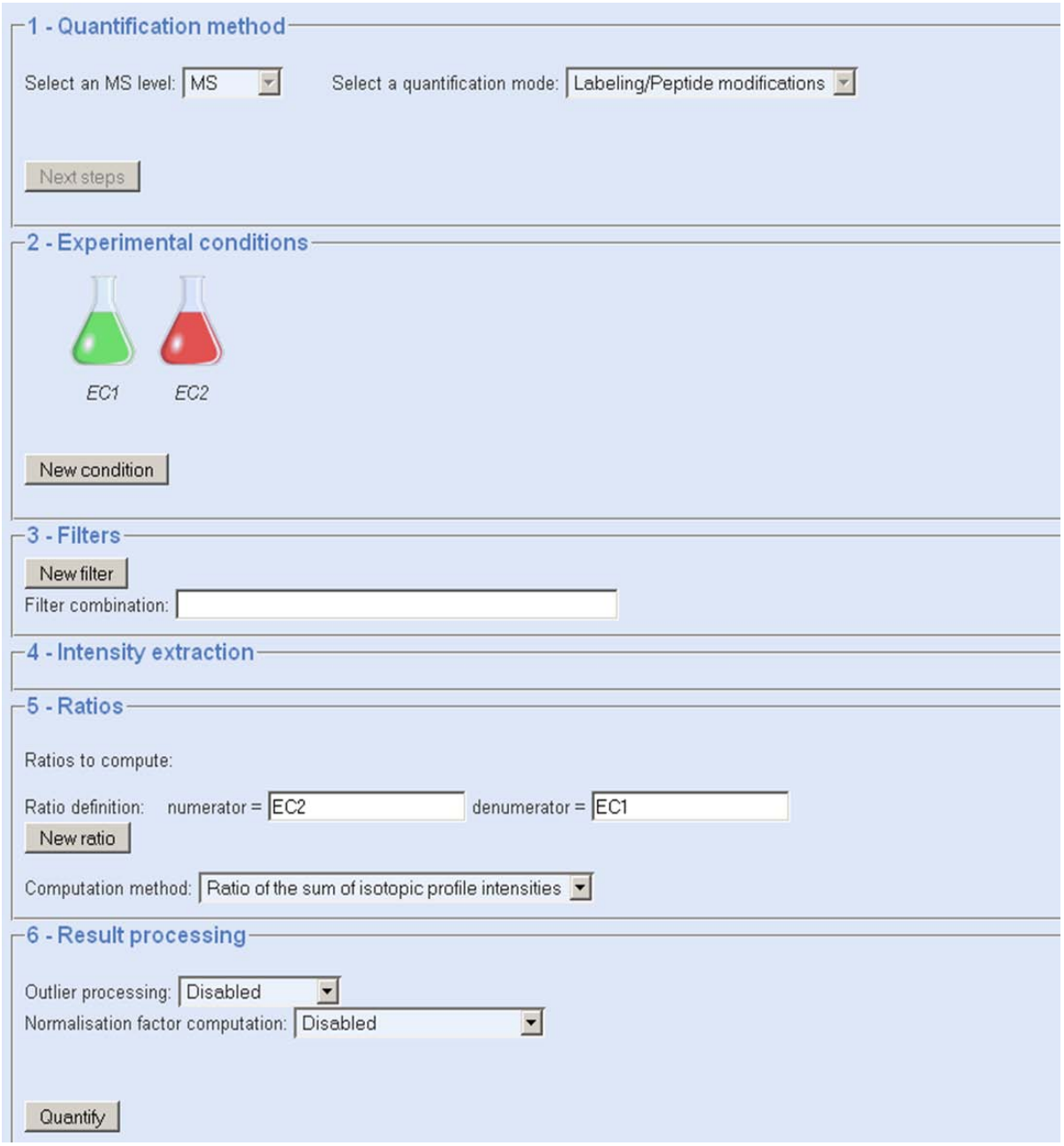


Figure 24 : formulaire utilisé pour soumettre une analyse quantitative. Dans cet exemple deux conditions expérimentales sont définies et un ratio EC2/EC1 permet de les comparer.

J'ai également essayé d'améliorer l'interaction avec les données de façon à ce que l'utilisateur ait un contrôle plus important sur les valeurs de quantification. Ainsi l'interface de visualisation des résultats apporte les nouveautés suivantes (cf figure 25) :

- L' « eXtracted Ion Chromatogram » (XIC) devient interactif. Cela signifie que l'utilisateur peut d'un simple clic modifier la zone sur laquelle l'intégration su signal est réalisée.
- Les spectres MS sont accessibles depuis la même interface. En cliquant sur un des points du XIC le spectre MS correspondant s'affiche et l'utilisateur peut avoir un regard critique sur le massif isotopique extrait et l'environnement autour de ce signal.
- Pour chaque peptide quantifié, il est possible de savoir s'il correspond à une séquence spécifique de la protéine affichée ou s'il s'agit d'une séquence commune à d'autres protéines. Ce deuxième cas peut être la source d'un biais important dans les valeurs de

quantification calculées par le logiciel et il est donc devenu nécessaire d'avoir accès à cette information. Afin d'évaluer facilement l'importance du problème, les protéines comportant un peptide commun sont directement accessibles depuis l'interface via l'utilisation de liens hypertextes.

Figure 25 : capture d'écran du module de quantification de MFPaQ v4. Cette vue illustre la richesse de l'interface graphique des résultats de quantification. Nous visualisons ici les données quantitatives de tous les peptides d'une même protéine. La table du haut nous renseigne sur le résultat de quantification pour chacun des peptides. Au premier plan la fenêtre « Other proteins » affiche la liste des protéines (sous forme de liens hypertextes) pour lesquelles la séquence non spécifique (VEGDLKGPEVDIK) est retrouvée. Cette fenêtre est active lorsque l'utilisateur clique sur le lien « NO (3) », ce qui signifie ici que le peptide est non spécifique et qu'il est retrouvé dans 3 autres séquences protéiques. Enfin chaque lien affiché peut être cliqué pour visualiser les résultats de quantification des protéines correspondantes.
Le panel en bas à gauche représente le XIC pour le peptide sélectionné. Le panel en bas à droite permet de visualiser les massifs isotopiques des formes légères et lourdes du peptide pour le spectre MS correspondant à l'apex du XIC (intensité maximale). Ces deux panels sont interactifs (fonctions de zoom et de sélection).

Le nouveau module offre donc une aide à la décision plus importante et permet de valider rapidement un grand nombre de protéines (via un système de sélection/désélection manuelle).

La quantification d'un mélange complexe de protéines mettant en œuvre une séparation sur gel et l'utilisation d'un instrument tel qu'un LTQ-Orbitrap XL génère un grand nombre d'identifications protéiques (environ 5000 protéines pour 40 fractions issues d'un échantillon de cellules humaines). Pour ce type de projet il est donc vite devenu inconcevable de vérifier manuellement l'ensemble des protéines quantifiées. Afin de résoudre ce problème j'ai implémenté un ensemble de macros capable de réaliser des désélections automatiques de peptides suivant certains critères. L'utilisateur peut exécuter ces scripts depuis l'interface web et valider ainsi très rapidement l'ensemble de ses données.

Voici une liste non exhaustive de ces macros :

- « RemoveOutliers » : désélectionne les peptides qui ont un ratio très différent des autres peptides de la protéine. La méthode de détection des valeurs extrêmes que j'ai utilisée correspond à celle de la boîte à moustache : min = Q1 - 1.5*IQR et max = Q3 + 1.5*IQR (Q1 et Q3 correspondent aux quartiles 1 et 3 ; IQR est la valeur d'interquartile soit Q3 - Q1).
- « DeselectPepWithMC » : désélectionne les peptides qui présentent au moins une coupure manquée ou « Missed Cleavage ».
- « DeselectPepWithMeth » : désélectionne les peptides qui contiennent une méthionine dans leur séquence. Le résidu méthionine est susceptible de subir une oxydation, diminuant ainsi la reproductibilité de la mesure quantitative. En retirant les peptides à méthionine, on peut donc parfois améliorer le résultat de quantification.
- « SpecificPepSelector » : désélectionne les peptides non spécifiques (identifiés dans plusieurs groupes de protéines). Là aussi ces peptides peuvent introduire un biais dans l'analyse et il est donc important de pouvoir les retirer.

III-3. Prise en charge des marquages isotopiques SILAC et ^{15}N

L'autre modification importante dans ce nouveau module de quantification a été la prise en charge de nouvelles méthodes marquage isotopique, et notamment les marquages SILAC et ^{14}N/^{15}N.

Au début de mes développements bioinformatiques aucun des logiciels disponibles ne permettait de mener à bien le traitement des données de spectrométrie de masse issues d'analyses protéomiques quantitatives basées sur un double marquage isotopique de type SILAC (Lysine +6 et Arginine +10) sur LTQ-Orbitrap. A l'époque, la réalisation d'un module de quantification prenant en charge ce double marquage était un enjeu important au niveau de l'équipe dans le cadre de plusieurs projets de recherche utilisant l'approche SILAC. On peut citer notamment un projet à large échelle visant à étudier la différence de sensibilité d'inhibiteurs du protéasome entre des lignées cellulaires de leucémies aigües myéloïdes matures et immatures. L'implémentation de ce nouveau module de quantification a permis d'exploiter les résultats de cette étude et de mettre en évidence plusieurs protéines différentiellement régulées (Mariette Mantondo, manuscrit de thèse intitulé « Inhibition du protéasome dans des cellules de leucémies aiguës myeloïdes : apport des approches protéomiques »).

Le marquage $^{14}N/^{15}N$ a également nécessité des développements spécifiques, car il complexifie l'étape d'appariement informatique des peptides légers et lourds. En effet, dans le cas d'un marquage isotopique qui met en œuvre l'incorporation métabolique d'un acide aminé modifié (Arg ou Lys par exemple dans le cas du SILAC) ou la modification chimique d'un des acides aminés (Cys dans le cas de l'ICAT), l'écart de masse obtenu pour les peptides d'une paire isotopique est uniquement fonction du nombre de modifications que comporte le peptide marqué, spécifiquement sur l'acide aminé ciblé. Cette information est facilement accessible depuis les résultats du moteur recherche car celui-ci fournit la séquence et la liste des modifications de chaque peptide identifié. Cependant, dans le cas d'un marquage au niveau d'un certain type d'atome (comme l'azote dans le cas du ^{15}N) il est nécessaire de calculer le nombre de ces atomes présents dans chacun des peptides (cf figure 26). L'obtention de cette valeur permet d'apparier les formes légères et lourdes des peptides mais surtout de calculer la masse d'une des deux formes qui n'aurait pas été identifiée (car non sélectionnée pour la MS/MS). Une fois cette opération effectuée la suite du traitement des données est identique à celui effectué dans le cas d'un marquage isotopique des acides aminés (cf partie II-3).

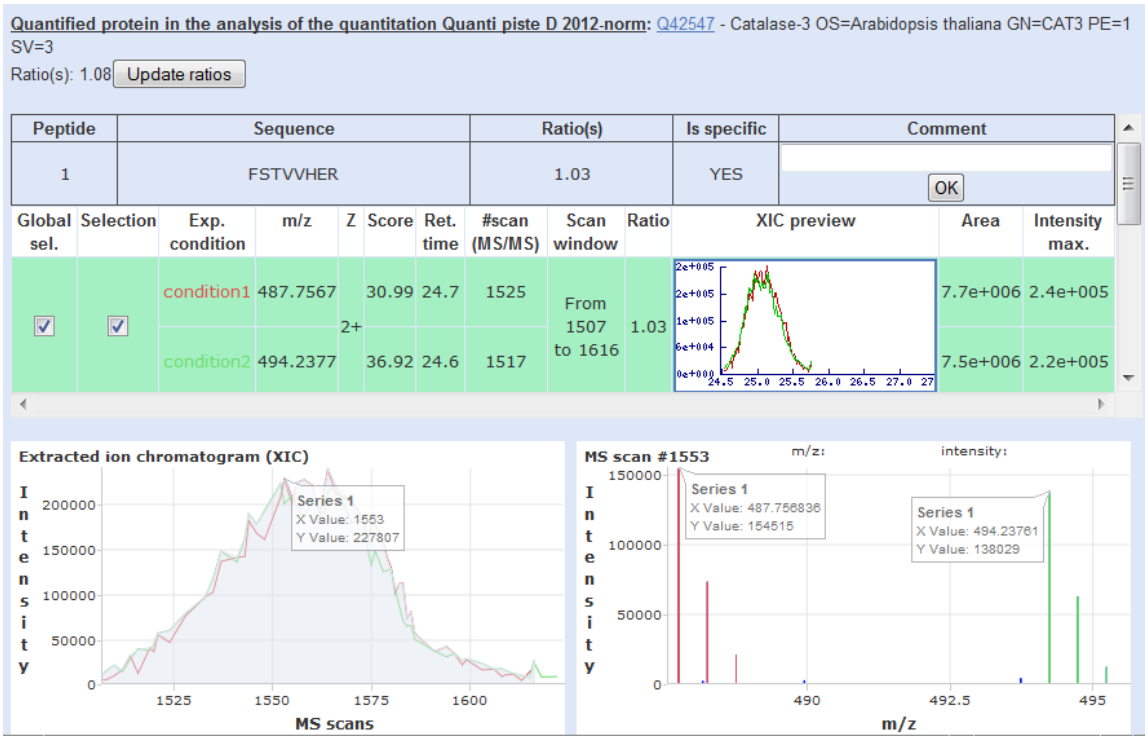


Figure 26 : quantification d'un peptide par marquage ^{15}N. Le nombre d'atomes d'azote sur le peptide permet de déterminer l'écart de masse entre la forme légère et la formule lourde. Ainsi pour le peptide de séquence FSTVVHER : Ntot = N(F) + N(S) + N(T) + N(V) + N(V) + N(H) + N(E) + N(R) = 1 + 1 + 1 + 1 + 1 + 3 + 1 + 4 = 13. Comme z = 2 alors $\Delta m/z = 13 * (M_{15N} - M_{14N}) / 2 = 13 * 0.997 / 2 = 6.481$. On peut vérifier que cet écart est le bon d'après les mesures expérimentales : 494.2377 - 487.7567 = 6.481.

Ce marquage métabolique (au même titre que le SILAC) diminue grandement les biais qui peuvent être introduits lors du traitement des échantillons et permet ainsi d'obtenir des données quantitatives plus précises. De plus, étant donné qu'il peut être facilement administré aux plantes de petite taille, il est utilisé dans un nombre croissant de projets de recherche relatifs à l'étude des végétaux. En collaboration avec l'UMR Plante-Microbe-Environnement (INRA Dijon) nous avons mis

en œuvre cette méthode de quantification afin d'étudier l'effet d'une drogue (la cryptogéine) sur le contenu protéique des microdomaines (aussi appelés DRM pour « Detergent-Resistant Membranes », ou radeaux lipidiques) dans des cellules de feuilles de tabac. Le doctorant avec lequel j'ai collaboré sur ce projet a réalisé de nombreux tests manuels (appels d'ion), ce qui a permis de valider les données fournies par le logiciel mais également d'optimiser les procédures d'extraction du signal. Les résultats de ces travaux ont été publiés dans le journal « Molecular and Cellular Proteomics » : (Stanislas, Bouyssie et al. 2009).

La disponibilité d'un module dédié à l'analyse de données issues de marquage $^{14}N/^{15}N$ a été essentielle pour la réalisation de cette étude car à l'époque il n'existait pas d'outil informatique permettant d'effectuer de telles analyses à partir d'acquisitions produites par un LTQ-Orbitrap et également à partir de fractionnement sur gel SDS-Page. Au-delà de ces considérations techniques, l'analyse des données a mis en évidence que la quantité de cinq protéines, présentes au niveau des radeaux lipidiques, était modifiée sous l'effet de la cryptogéine. Quatre de ces cinq protéines, impliquées dans le trafic intra-cellulaire, ont été observées en quantité plus faible alors qu'une protéine impliquée dans la signalisation (appartenant à la famille des 14-3-3) était plus abondante. Des résultats similaires avaient déjà été obtenus dans des études réalisées sur des cellules animales mais c'est la première fois que cela a été démontré chez une plante.

PUBLICATION

« Quantitative proteomics reveals a dynamic association of proteins to detergent-resistant membranes upon elicitor signalling in tobacco. »

Stanislas T*, Bouyssie D*, Rossignol M, Vesa S, Fromentin J, Morel J, Pichereaux C, Monsarrat B, Simon-Plas F

Mol Cell Proteomics. **2009** Sep;8(9):2186-98.

* contribution équivalente des deux auteurs.

Quantitative Proteomics Reveals a Dynamic Association of Proteins to Detergent-resistant Membranes upon Elicitor Signaling in Tobacco*

Thomas Stanislas‡§, David Bouyssie§¶‖, Michel Rossignol¶‖, Simona Vesa‡, Jérôme Fromentin‡, Johanne Morel‡, Carole Pichereaux¶‖**, Bernard Monsarrat¶‖‡‡, and Françoise Simon-Plas‡§§**

A large body of evidence from the past decade supports the existence, in membrane from animal and yeast cells, of functional microdomains playing important roles in protein sorting, signal transduction, or infection by pathogens. In plants, as previously observed for animal microdomains, detergent-resistant fractions, enriched in sphingolipids and sterols, were isolated from plasma membrane. A characterization of their proteic content revealed their enrichment in proteins involved in signaling and response to biotic and abiotic stress and cell trafficking suggesting that these domains were likely to be involved in such physiological processes. In the present study, we used $^{14}N/^{15}N$ metabolic labeling to compare, using a global quantitative proteomics approach, the content of tobacco detergent-resistant membranes extracted from cells treated or not with cryptogein, an elicitor of defense reaction. To analyze the data, we developed a software allowing an automatic quantification of the proteins identified. The results obtained indicate that, although the association to detergent-resistant membranes of most proteins remained unchanged upon cryptogein treatment, five proteins had their relative abundance modified. Four proteins related to cell trafficking (four dynamins) were less abundant in the detergent-resistant membrane fraction after cryptogein treatment, whereas one signaling protein (a 14-3-3 protein) was enriched. This analysis indicates that plant microdomains could, like their animal counterpart, play a role in the early signaling process underlying the setup of defense reaction. Furthermore proteins identified as differentially associated to tobacco detergent-resistant membranes after cryptogein challenge are involved in signaling and vesicular trafficking as already observed in similar studies performed in animal cells upon biological stimuli. This suggests that the ways by which the dynamic association of proteins to microdomains could participate in the regulation of the signaling process may be conserved between plant and animals. ***Molecular & Cellular Proteomics 8:2186–2198, 2009.***

The plasma membrane of eukaryotes delineates the interface between the cell and the environment. Thus it is particularly involved in environmental signal recognition and their transduction into intracellular responses, playing a crucial role in many essential functions such as cell nutrition (involving transport of solutes in and out of the cell) or response to environmental modifications (including defense against pathogens).

Over the last 10 years, a new aspect of the plasma membrane organization has arisen from biophysical and biochemical studies performed with animal cells. Evidence has been given that the various types of lipids forming this membrane are not uniformly distributed inside the bilayer but rather spatially organized (1). This leads in particular to the formation of specialized phase domains, also called lipid rafts (2, 3). Recently a consensus emerged on the characteristics of these domains. Both proteins and lipids contribute to the formation and the stability of membrane domains that should be called "membrane rafts" and are envisaged as small (10–200-nm), heterogeneous, highly dynamic, sterol- and sphingolipid-enriched domains that compartmentalize cellular processes (4). Small rafts can sometimes be stabilized to form larger platforms through protein-protein and protein-lipid interactions (5). Because of their particular lipidic composition (enrichment in sterol, sphingolipids, and saturated fatty acids), these domains form a liquid ordered phase inside the membrane. This structural characteristic renders them resistant to solubilization by non-ionic detergents, and this property has been widely used to isolate lipid rafts as detergent-resistant mem-

From the ‡Institut National de la Recherche Agronomique (INRA), Unité Mixte de Recherche (UMR) Plante Microbe Environnement 1088/CNRS 5184/Université de Bourgogne, 17 Rue Sully, BP 86510 F-21000 Dijon, France, ¶Institut de Pharmacologie et de Biologie Structurale (IPBS), CNRS, 205 route de Narbonne, F-31077 Toulouse, France, ‖IPBS, Université Paul Sabatier, Université de Toulouse, F-31077 Toulouse, France, and **IPBS, Institut Fédératif de Recherche 40 Plateforme Protéomique, 205 route de Narbonne, F-31077 Toulouse, France

Received, February 19, 2009, and in revised form, June 2, 2009

Published, MCP Papers in Press, June 13, 2009, DOI 10.1074/mcp.M900090-MCP200

branes (DRMs)[1] for further analysis (1). The most important hypothesis to explain the function of these domains is that they provide for lateral compartmentalization of membrane proteins and thereby create a dynamic scaffold to organize certain cellular processes (5). This ability to temporally and spatially organize protein complexes while excluding others conceivably allows for efficiency and specificity of cellular responses. In yeasts and animal cells, the association of particular proteins with these specialized microdomains has emerged as an important regulator of crucial physiological processes such as signal transduction, polarized secretion, cytoskeletal organization, generation of cell polarity, and entry of infectious organisms in living cells (6). Much of the early evidence for a functional role of lipid rafts came from studies of hematopoietic cells in which multichain immune receptors including the high affinity IgE receptor (FcεRI), the T cell receptor, and the B cell receptor (BCR) translocate to lipid rafts upon cross-linking (7). Moreover this signaling involves the relocalization of several proteins; for instance the ligation of the B cell antigen receptor with antigen induced a dissociation of the adaptor protein ezrin from lipid rafts (8). This release of ezrin acts as a critical trigger that regulates lipid raft dynamics during BCR signaling.

In plants, the investigations of the presence of such microdomains are very recent and limited to a reduced number of publications (for a review, see Ref. 9). A few years ago, Peskan *et al.* (10) reported for the first time the isolation of Triton X-100-insoluble fractions from tobacco plasma membrane. Mongrand *et al.* (11) provided a detailed analysis of the lipidic composition of such a detergent-resistant fraction indicating that it was highly enriched in a particular species of sphingolipid (glycosylceramide) and in several phytosterols (stigmasterol, sitosterol, 24-methylcholesterol, and cholesterol) compared with the whole plasma membrane from which it originates. Similar results were then obtained with DRMs prepared from *Arabidopsis thaliana* cell cultures (12) and from *Medicago truncatula* roots (13). So the presence in plant plasma membrane of domains sharing with animal rafts a particular lipidic composition, namely strong enrichment in sphingolipids together with free sterols and sterol conjugates, the latter being specific to the plant kingdom (11, 13), now seems established. In plant only a few evidences suggest *in vivo* the role of dynamic clustering of plasma membrane proteins, and they refer to plant-pathogen interaction. A cell biology study reported the pathogen-triggered focal accumulation of components of the plant defense pathway in the plasma membrane (PM), a process reminiscent of lipid rafts (14). Consistently a proteomics study of tobacco DRMs led to the identification of 145 proteins among which a high proportion were linked to signaling in response to biotic stress, cellular trafficking, and cell wall metabolism (15). This suggests that these domains are likely to constitute, as in animal cells, signaling platforms involved in such physiological functions.

Cryptogein belongs to a family of low molecular weight proteins secreted by many species of the oomycete *Phytophthora* named elicitins that induce a hypersensitivity-like response and an acquired resistance in tobacco (16). To understand molecular processes triggered by cryptogein, its effects on tobacco cell suspensions have been studied for several years. Early events following cryptogein treatment include fixation of a sterol molecule (17, 18); binding of the elicitor to a high affinity site located on the plasma membrane (19); alkalinization of the extracellular medium (20); efflux of potassium, chloride, and nitrate (20, 21); fast influx of calcium (22); mitogen-activated protein kinases activation (23, 24); nitric oxide production (25, 26); and development of an oxidative burst (27, 28). We previously identified NtrbohD, an NADPH oxidase located on the plasma membrane, as responsible for the reactive oxygen species (ROS) production occurring a few minutes after challenging tobacco Bright Yellow 2 (BY-2) cells with cryptogein (29). The fact that most of these very early events involve proteins located on the plasma membrane and that one of them, NtrbohD, has been demonstrated as exclusively associated to DRMs in a sterol-dependent manner (30) prompted us to analyze the modifications of DRM proteome after cryptogein treatment. In the present study, we aimed to confirm the hypothesis that, as observed in animal cells, the dynamic association to or exclusion of proteins from lipid rafts could participate in the signaling process occurring during biotic stress in plants.

To achieve this goal, we had to set up a quantitative assay allowing a precise comparison of the amounts of each protein in DRMs extracted from either control or cells treated with cryptogein. Among several technologies, we excluded DIGE (31), recently used to analyze whole cell proteome variations in plants (32–34), because membrane proteins are poorly soluble in the detergents used for two-dimensional electrophoresis; this limitation is all the more marked for proteins selected on the basis of their insolubility in non-ionic detergent, the criteria for DRMs isolation. Stable isotope labeling of proteins or peptides combined with MS analysis represents alternative strategies for accurate, relative quantification of proteins on a global scale (35, 36). In these approaches, proteins or peptides of two different samples are differentially labeled with stable isotopes, combined in an equal ratio, and then jointly processed for subsequent MS analysis. Relative quantification of proteins is based on the comparison of signal intensities or peak areas of isotope-coded peptide pairs extracted from the respective mass spectra. Stable isotopes can

[1] The abbreviations used are: DRM, detergent-resistant membrane; BY-2, Bright Yellow 2; DRP, dynamin-related protein; PM, plasma membrane; BCR, B cell receptor; ROS, reactive oxygen species; nano-LC-MS/MS, microcapillary high performance LC-MS/MS; MFPaQ, Mascot file parsing and quantification; LTQ, linear trap quadrupole; SGN, SOL Genomic Networks; FDR, false discovery rate; Cry, cryptogein; SILAC, stable isotope labeling with amino acids in cell culture; GFP, green fluorescent protein.

be introduced either chemically into proteins/peptides via derivatization of distinct functional groups of amino acids or metabolically during protein biosynthesis (for a review, see Ref. 37). Metabolic labeling strategies are based on the *in vivo* incorporation of stable isotopes during growth of organisms. Nutrients or amino acids in a defined medium are replaced by their isotopically labeled (^{15}N, ^{13}C, or ^{2}H) counterparts eventually resulting in uniform labeling of proteins during the processes of cell growth and protein turnover (38). As a consequence, differentially labeled cells or organisms can be combined directly after harvesting. This minimizes experimental variations due to separate sample handling and thus allows a relative protein quantification of high accuracy.

$^{14}N/^{15}N$ labeling has been recently proved to be suitable for comparative experiments performed with whole plants (39–42) and in plant suspension cells where the level of incorporation is equal to the isotopic purity of the salt precursor (43, 44). It has been used successfully to analyze some variations induced in *A. thaliana* plasma membrane proteome following heat shock (45) or cadmium exposure (46) and to compare phosphorylation levels of plasma membrane proteins after challenge of *Arabidopsis* cells with elicitors of defense reaction (47). In the present study, we used a mineral $^{14}N/^{15}N$ metabolic labeling of tobacco BY-2 cells before treatment with cryptogein and subsequent isolation of DRMs. The DRM proteins were further analyzed by one-dimensional SDS-PAGE and digested by trypsin, and peptides were subjected to microcapillary high performance LC-MS/MS (nano-LC-MS/MS). This metabolic method allowed a complete labeling of the proteome, and consequently a major drawback of this method is probably the difficulty to perform an exhaustive analysis of the very large amount of data generated. To solve this problem, a new quantification module of the MFPaQ software (48) was developed, allowing the automatic quantification of the identified peptides. The results derived from the program were validated through a comparison with manual quantification. Thus, we achieved the complete analysis of the DRM proteome variation and identified four proteins whose abundance in DRMs was decreased and one that was enriched in DRMs upon elicitation. The biological relevance of these results, which indicate that, in plant as in animals, the dynamic association of proteins to membrane domains is part of a signaling pathway, will be further discussed.

EXPERIMENTAL PROCEDURES

Materials—BY-2 cells (*Nicotiana tabacum* cv. Bright Yellow 2) were grown in Murashige and Skoog medium, pH 5.6, containing Murashige and Skoog salt (49), 1 mg/liter thiamine-HCl, 0.2 mg/liter 2,4-dichlorophenylacetic acid, 100 mg/liter *myo*-inositol, 30 g/liter sucrose, 200 mg/liter KH_2PO_4, and 2 g/liter MES. Cells were maintained by weekly dilution (2:80) into fresh medium.

Cell Labeling—For quantitative proteomics experiments, the labeling was achieved by substituting $^{15}NH_4{}^{15}NO_3$ (98% ^{15}N) and $K^{15}NO_3$ (99% ^{15}N) to the equivalent concentration of these salts (20 mM) in the culture medium for at least four passages over 4 weeks.

ROS Determination—Cells were harvested 6 days after subculture, filtered, and resuspended (1 g for 10 ml) in a 2 mM MES buffer, pH 5.90, containing 175 mM mannitol, 0.5 mM $CaCl_2$, and 0.5 mM K_2SO_4. After a 3-h equilibration on a rotary shaker (150 rpm) at 25 °C, cells were treated with 50 nM cryptogein. The production of ROS was determined by chemiluminescence using luminol and a luminometer (BCL Book). Every 10 min, a 250-μl aliquot of the cell suspension was added to 50 μl of 0.3 mM luminol and 300 μl of the assay buffer (175 mM mannitol, 0.5 mM $CaCl_2$, 0.5 mM K_2SO_4, and 50 mM MES, pH 6.5).

Preparation and Purity of Tobacco Plasma Membrane—All steps were performed at 4 °C. Cells were collected by filtration, frozen in liquid N_2, and homogenized with a Waring Blendor in grinding medium (50 mM Tris-MES, pH 8.0, 500 mM sucrose, 20 mM EDTA, 10 mM DTT, and 1 mM PMSF). The homogenate was centrifuged at 16,000 × *g* for 20 min. After centrifugation, supernatants were collected, filtered through two successive screens (63 and 38 μm), and centrifuged at 96,000 × *g* for 35 min. This microsomal fraction was purified by partitioning in an aqueous two-phase system (polyethylene glycol 3350/dextran T-500; 6.6% each) to obtain the plasma membrane fraction (50). Marker activities used to evaluate the contamination of the plasma membrane fraction were as follows: azide-sensitive ATPase activity at pH 9 for mitochondria, nitrate-sensitive ATPase activity at pH 6 for tonoplast, antimycin-insensitive NADH cytochrome *c* reductase for endoplasmic reticulum, and analysis of lipid monogalactosyldiacylglycerol contents for chloroplasts (11).

Isolation of Detergent-resistant Membranes—Plasma membranes were resuspended in a buffer A containing 50 mM Tris-HCl, pH 7.4, 3 mM EDTA, and 1 mM 1,4-dithiothreitol and treated with 1% Triton X-100 (w/v) for 30 min on ice with very gentle shaking every 10 min. Solubilized membranes were placed at the bottom of a centrifuge tube and mixed with 60% sucrose (w/w) to reach a final concentration of 48% (w/w) and overlaid with a discontinuous sucrose gradient (40, 35, 30, and 20%, w/w). After a 20-h centrifugation at 100,000 × *g*, a ring of Triton X-100-insoluble membranes was recovered at the 30–35% interface, diluted in buffer A, and centrifuged for 4 h at 100,000 × *g*. The pellet was resuspended in buffer A, and protein concentrations were determined using the Bradford reagent with BSA as the standard.

Protein Separation by SDS-PAGE—DRM proteins were solubilized in a buffer consisting of 6 M urea, 2.2 M thiourea, 5 mM EDTA, 0.1% SDS, 2% *N*-octyl glucoside, and 50 mM Tris-HCl. Samples were first incubated at room temperature for 15 min and then in a sonic bath for another 15 min. After a centrifugation at 16,000 × *g* for 15 min, no pellet was observed. An aliquot of solubilized proteins (15 μg) was added to Laemmli buffer before being deposited on an 8% acrylamide gel and separated by SDS-PAGE. Proteins were visualized by Coomassie Blue staining. Each lane was systematically cut into 20 bands of similar volume for MS/MS protein identification.

Protein Digestion—Each band was incubated in 25 mM ammonium bicarbonate and 50% ACN until destaining. Gel pieces were dried in a vacuum SpeedVac (45 °C), further rehydrated with 30 μl of a trypsin solution (10 ng/μl in 50 mM NH_4HCO_3), and finally incubated overnight at 37 °C. The resulting peptides were extracted from the gel as described previously (51). The trypsin digests were dried in a vacuum SpeedVac and stored at −20 °C before LC-MS/MS analysis.

Nano-LC-MS/MS Analysis—The trypsin digests were separated and analyzed by nano-LC-MS/MS using an Ultimate 3000 system (Dionex, Amsterdam, the Netherlands) coupled to an LTQ-Orbitrap mass spectrometer (Thermo Fisher Scientific, Bremen, Germany). The peptide mixture was loaded on a C_{18} precolumn (300-μm-inner diameter × 15 cm PepMap C_{18}, Dionex) equilibrated in 95% solvent A (5% acetonitrile and 0.2% formic acid) and 5% solvent B (80%

acetonitrile and 0.2% formic acid). Peptides were eluted using a 5–50% gradient of solvent B during 80 min at 300 nl/min flow rate. The LTQ-Orbitrap was operated in data-dependent acquisition mode with the Xcalibur software (version 2.0.6, Thermo Fisher Scientific). Survey scan MS spectra were acquired in the Orbitrap on the 300–2000 *m/z* range with the resolution set to a value of 60,000. The five most intense ions per survey scan were selected for CID fragmentation, and the resulting fragments were analyzed in the linear trap (LTQ). Dynamic exclusion was used within 60 s to prevent repetitive selection of the same peptide. To automatically extract peak lists from Xcalibur raw files, the ExtractMSN macro provided with Xcalibur was used through the Mascot Daemon interface (version 2.2.0.3, Matrix Science, London, UK). The following parameters were set for creation of the peak lists: parent ions in the mass range 400–4500, no grouping of MS/MS scans, and threshold at 1000. A peak list was created for each fraction analyzed (*i.e.* gel slice), and individual Mascot searches were performed for each fraction.

Database Search—MS/MS spectra were processed by Mascot against a subset of SOL Genomic Networks (SGN; download November 2007), which is an integrated genome database for Solanacae (52). The Unigene subset of this database is built by assembling together in contigs expressed sequence tag sequences that are ostensibly fragments of the same gene. We put together the tomato (*Lycopersicum esculentum*) and tobacco (*N. tabacum*) Unigene sequences in our database subset (60,227 entries). Tomato expressed sequence tags were selected for two reasons: because of the phylogenetic closeness between the genus *Nicotiana* and the genus *Lycopersicum* on the one hand and for the great number of entries in the database (34,829) on the other hand.

The following search parameters were applied: trypsin as cleaving enzyme, "ESI-Trap" as instrument, peptide mass tolerance of 10 ppm, MS/MS tolerance of 0.5 Da, and one missed cleavage allowed. Methionine oxidation and asparagine and glutamine deamination were chosen as variable modifications. ^{15}N metabolic labeling was chosen as a quantitative method for Mascot database searching, allowing identification of labeled and unlabeled peptides within the same database search.

Bioinformatics Analysis—We used the MFPaQ program (48) version 4 to validate the data. This software is a Web application that allows fast and user-friendly verification of Mascot result files as well as data quantification using either isotopic labeling or label-free methods. It provides an interactive interface with Mascot results. It is based on three modules, the Mascot File Parser module, the quantification module, and a third module designed for differential analysis in which validated protein lists are compared. Version 4 of the program has a new data storage system (based on SQLite) and an improved quantification module that now handles the ^{15}N labeling approach. The module is compatible with LTQ-Orbitrap .raw files and .mzXML files. Because the software is written in the Perl programming language, it may be installed on a wide range of operating systems (it has only been tested under Windows XP/2003 and Linux Ubuntu). An automated installer (for the Windows platform) and the source code of this release are available for download from the MFPaQ Website.

The protein validation was performed according to user-defined criteria based on the number, score, and rank of identified peptides. A protein has been validated if it has at least one top ranking peptide with a score greater than 48 (p value <0.001), two top ranking peptides with a score greater than 35 (p value <0.03), or three top ranking peptides with a score greater than 31 (p value <0.1). The above criteria were adjusted to obtain a false positive rate of about 1% at the protein level. To evaluate the false positive rate in these large scale experiments, all the initial database searches were performed using the "decoy" option of Mascot, *i.e.* the data were searched against a combined database containing the real specified protein sequences (target database) and the corresponding reversed protein sequences (decoy database). MFPaQ used the same criteria to validate decoy and target hits, computed the false discovery rate (FDR = number of validated decoy hits/(number of validated target hits + number of validated decoy hits) × 100) for each gel slice analyzed, and calculated the average of FDR for all slices belonging to the same gel lane (*i.e.* to the same sample). The FDRs found for the analysis of samples A and B were 1 and 1.2%, respectively.

The MFPaQ software is able to detect highly homologous Mascot protein hits, *i.e.* proteins identified with some top ranking MS/MS queries also assigned to another protein hit of higher score (*i.e.* red, non-bold peptides). These homologous protein hits have been validated only if they have been additionally assigned a specific top ranking (red and bold) peptide of score higher than 48.

The MFPaQ quantification module has been improved to handle the ^{15}N labeling approach. The MS data processing and the results obtained using this quantification module are presented under "Results."

RESULTS

Characterization of Biological Material—Metabolic ^{15}N labeling was achieved by growing BY-2 cells on a basal salt medium containing ^{15}N (or ^{14}N) as a sole nitrogen source. After four subcultures of 7 days, the incorporation rates of isotopes were maximum (98% ^{15}N) (supplemental Data S1).

Treatment of these cells with the fungal elicitor cryptogein was performed in independent experiments carried out to create an inverse labeling: on the one hand ^{15}N-labeled cells were treated with cryptogein, whereas ^{14}N-labeled cells were untreated ([^{15}N]Cry/^{14}N); on the other hand ^{14}N-labeled cells were treated with cryptogein, whereas ^{15}N-labeled cells remained untreated ([^{14}N]Cry/^{15}N).

To control that the physiological responses were not affected by the isotope labeling and were comparable in the different experiments, a typical marker of elicitation, the production of ROS, was monitored during 90 min on aliquots of cells. As indicated in Fig. 1, both kinetics and intensity of ROS production were similar in the two experiments particularly in the early phase (0–30 min). Our purpose was to test the dynamics of protein association/dissociation to DRMs during cryptogein-induced signal transduction. As this kind of event typically occurs upstream of the signaling pathways, cell metabolism was stopped by freezing after 5 min of cryptogein treatment, which corresponds to the onset on ROS production (Fig. 1). Control and treated cells of each experiment were then mixed at equal weight, leading to two samples: [^{15}N]Cry/^{14}N-control (sample A) and [^{14}N]Cry/^{15}N-control (sample B). Each of these samples was submitted to subcellular fractionation, leading to DRM isolation according to the procedure indicated in Fig. 2. We selected the conventional phase partition procedure to obtain a highly enriched PM fraction. Typical preparations were enriched by a factor of 7–8 in the PM marker vanadate-sensitive ATPase activity compared with the starting material consisting of a crude microsomal fraction. Biochemical characterization of this PM fraction revealed that it was virtually free of mitochondrial contamination. Endoplasmic reticulum marker enzyme was depleted by a factor 8, and tonoplast was depleted by a factor 20 between microsomes and PM. These

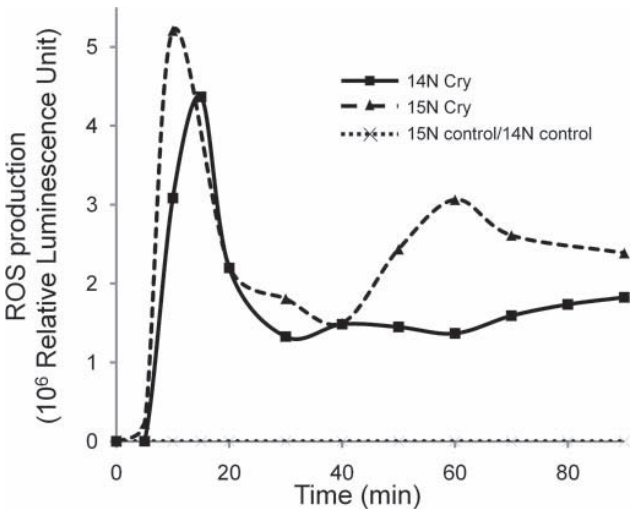


FIG. 1. **Chemiluminescence detection of ROS accumulation triggered by cryptogein in BY-2 cells.** At zero time, cryptogein (50 nM) was added to the cell suspension. Every 10 min (from 0 to 120 min), ROS accumulation was determined by chemiluminescence with luminol as described under "Experimental Procedures." Control values are very close for both ^{14}N and ^{15}N BY-2 cells, so only one curve is presented in the graph. ^{14}N (*squares*) and ^{15}N (*triangles*) BY-2 cells show a similar accumulation of H_2O_2 from 5 min after elicitation with cryptogein. Values are expressed in arbitrary units of chemiluminescence.

results indicate that the contamination by other endomembranes was low, and thus the PM fraction constitutes a suitable starting material for extraction of DRMs and for further proteomics analysis. It is well known that, in biological membranes, the formation of microdomains correlates with resistance to solubilization by non-ionic detergents, and this property has been widely used for biochemical characterization of these domains (1). Here plasma membranes were incubated at 4 °C with Triton X-100 (final concentration, 1%) at a detergent/protein ratio of 15, previously established as the most suitable for DRM extraction from this material (11). In these conditions, the amount of proteins recovered in the DRM fraction was around 5% (w/w) of the initial quantity present in the plasma membrane fraction.

Identification and Quantification of BY-2 DRM Proteins by Mass Spectrometry—Tobacco plasma membrane DRMs were found to be soluble in a buffer consisting of both non-ionic (*N*-octyl glucoside) and ionic (SDS) detergents and high concentration of chaotropic agents. As sample complexity rendered a prefractionation necessary, we then chose to separate the protein mixture by one-dimensional electrophoresis (SDS-PAGE) (Fig. 2). This separation also made samples compatible with subsequent LC-MS/MS analysis. The electrophoresis lane was cut into 20 pieces, and digests obtained from each piece after trypsin addition were separately analyzed by nano-HPLC coupled to an LTQ-Orbitrap mass spectrometer. The raw data were searched by the Mascot software using the SGN database subset of tobacco and tomato contigs (called Unigenes) chosen to improve the number of matching peptides. The redundancy between MS/MS spectra corresponding to the same peptide is managed by the search engine (Mascot). The mass spectrometer thus allowed the identification of 11,540 total peptides among which 2015 were unique (1719 unique peptides for sample A and 1788 for sample B). This led to the identification of 748 Unigenes for sample A and 728 for sample B.

In this study, we used the MFPaQ software to characterize DRM proteins and the variation of protein expression profile in response to cryptogein stimulation. The quantification module of the MFPaQ software (48) has been improved to manage the ^{15}N labeling strategy. We briefly describe here the algorithm implemented in the module. First, the software extracts Mascot search results obtained using the ^{15}N metabolic labeling quantitative method. Then labeled and unlabeled peptides are grouped into peptide pairs. If only one of the forms is present, the software automatically computes the number of nitrogens for the corresponding peptide sequence and then predicts the monoisotopic mass of the missing form. Intensities of peptide pairs are then extracted from the MS survey scans of raw data files in batch mode, and heavy/light ratios are computed for each peptide pair. The ratios of all validated peptides are averaged for each protein within a gel slice, and a coefficient of variation is calculated for the ratio of the proteins that have been quantified with several peptides. Peptide ratio outliers for a given protein are automatically excluded from the protein average ratio. The outlier detection method is based on the box plot analysis and used if a minimum of four peptide pairs has been quantified. When a protein is identified and quantified several times in consecutive gel slices, a final protein ratio is computed by averaging the several ratios found for this protein in the different fractions, and a global coefficient of variation is calculated. Finally the program is able to export all of the quantification results including peptide and protein information in Excel spreadsheets.

Automatic analysis using MFPaQ was performed on the peptides identified in the two samples. The distribution of their intensity ratio in treated/control sample (*i.e.* $^{15}N/^{14}N$ for sample A and $^{14}N/^{15}N$ for sample B) is presented in Fig. 3. The median value of the ratio is 1.07 for sample A and 1.10 for sample B.

To validate the results obtained with the software, a manual quantification was performed on a subset of the identification results (585 peptides). The ion currents resulting from ^{14}N and ^{15}N monoisotopic peaks of these peptides were measured using the Xcalibur software (version 2.0.6). The peptides were selected if the score Mascot for one of the two isotopes was valid (score >48). The evaluation of ion current of the other isotope was performed for the exact mass (nearly 5 ppm) and an identical retention time. The quantification was made by calculating, for each peptide, the ratio of peak areas corresponding to the two samples (treated *versus* control). This manual analysis performed on 585 peptides led to the distri-

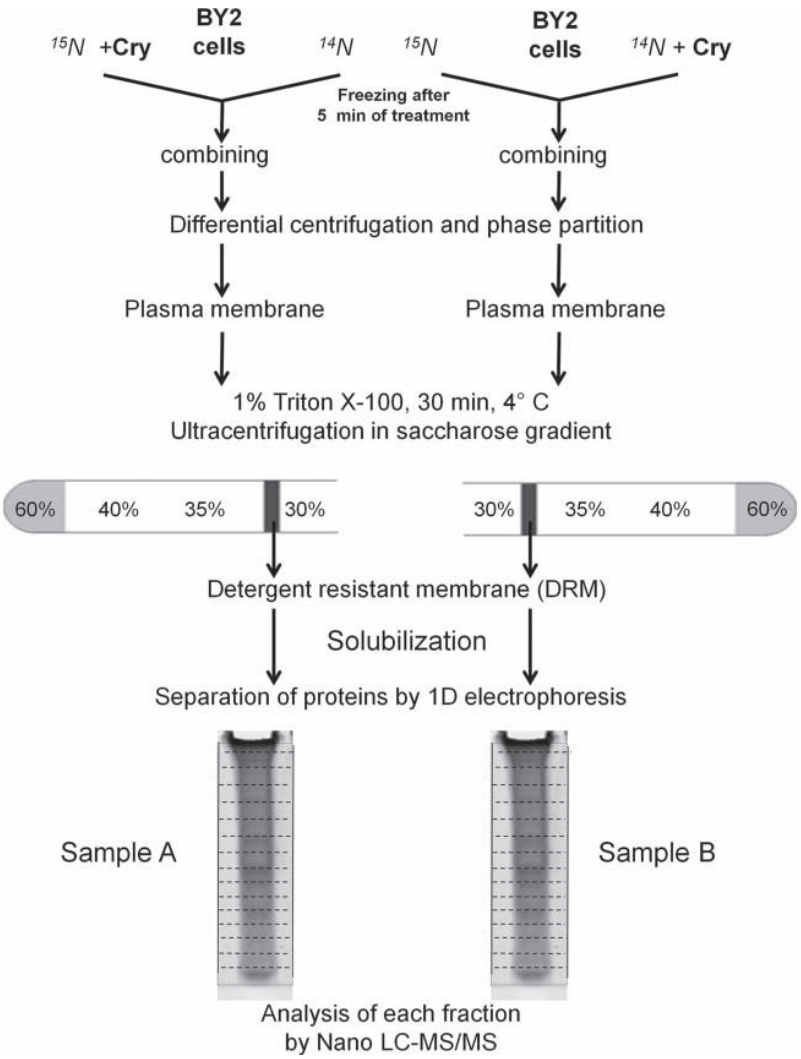


FIG. 2. **Work flow of the experimental labeling experiments.** Two experiments were carried out in parallel. In one case, ^{15}N cells were subjected to cryptogein treatment, and ^{14}N cells were used as control. In the second case, ^{14}N cells were subjected to cryptogein treatment, and ^{15}N cells were used as control. For each experiment, control and treated cells were frozen after 5 min of treatment, combined at equal weight after treatment, and subjected to the DRM isolation procedure described under "Experimental Procedures." Proteins were separated by one-dimensional (*1D*) electrophoresis, and each lane was cut into 20 bands of equal volume for MS/MS protein identification.

bution exposed in Fig. 4. The median value of this distribution is 1.13 giving a difference of 0.06 compared with the value obtained with the software results. It has to be noted that the median value, using automatic quantification results, is 1.15 if it is based on the same peptide subset as the one used for the manual procedure. Thus, both techniques indicate a global medium ratio of intensity in treated/control cells DRMs slightly above 1 and a good coincidence between global data obtained by manual and automatic quantification.

The correlation between the ratios obtained with the two techniques for a single peptide was then examined, calculating the relative difference between the values obtained for 383 peptides common to samples A and B. The result, presented in Fig. 5, is identical for the two samples and indicated that for 76% of these peptides the relative difference between the two methods is below 10% (and for 90% of them is it below 20%).

We further analyzed the coefficient of variation among the treated/control ratios, calculated by MFPaQ, of the different peptides identifying a single Unigene. As indicated on Fig. 6, this coefficient of variation is inferior to 0.1 for 60% of the Unigenes and below 0.2 for 85% of them. The mean value of these coefficients for all the Unigenes is 0.13.

In the database used, SGN numbers correspond to contigs that may belong to the same protein. To eliminate this redundancy, we used the "bulk search" tool available on the SGN Website that allows mapping between SGN identifiers and National Center for Biotechnology Information (NCBI) identifying GI numbers. Unigenes corresponding to the same GI number were grouped under the same protein identification. In this way, the final protein lists contain some proteins from other species than those used to build the SGN database subset (*N. tabacum* and *L. esculentum*). By applying the rules of elimination of the redundancy described above, we obtained a list of 350 proteins identified and quantified that corresponds to the smallest set of proteins explaining the identified peptides presence (supplemental Data S2 and S3).

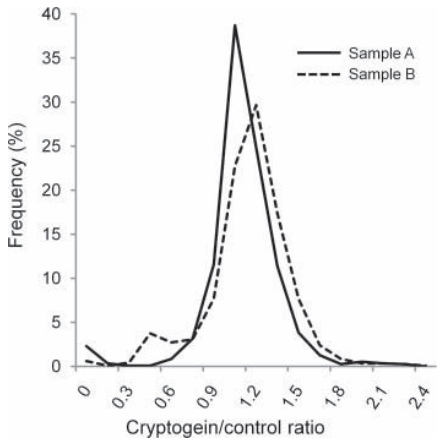


FIG. 3. **Distribution of peptide abundance ratios calculated automatically.** All peptides common to the two samples (1183) were automatically quantified by the module developed in MFPaQ. The ratios represented correspond to (the intensity of a given peptide in DRMs extracted from elicited cells)/(its intensity in DRMs extracted from control cells) (*i.e.* $^{15}N/^{14}N$ for sample A and $^{14}N/^{15}N$ for sample B).

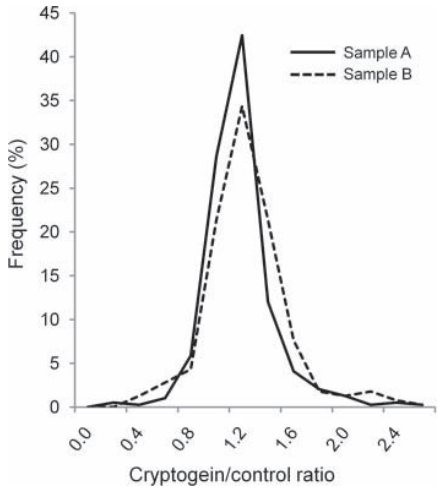


FIG. 4. **Distribution of peptide abundance ratios calculated manually.** For 583 peptides chosen randomly, a manual quantification was performed by calculating the area of the corresponding peaks. The ratios represented correspond to (the intensity of a given peptide in DRMs extracted from elicited cells)/(its intensity in DRMs extracted from control cells) (*i.e.* $^{15}N/^{14}N$ for sample A and $^{14}N/^{15}N$ for sample B).

Only 61% of proteins have been identified and quantified in both experiments A and B. The main reason is the well known bias of LC-MS/MS experiments ("shotgun" analysis) of a peptide sample issued from a complex mixture of proteins with a dynamic range of concentrations comprising several orders of magnitude. If the number of ionized peptides is greater than the mass spectrometer can analyze (because of its sequencing speed) then the peptides in small quantities are not systematically identified. Quite similar results indicating 60% of identified plant DRM proteins found in two independent analyses have been published recently (44).

The fact that our biological material corresponds to a plant species, the genome of which is not fully sequenced, led us to use the SGN database to increase the amount of identifications. However, this added a supplemental step (described above) of Unigene grouping to reach the final protein identification. We thus further analyzed the suitability of the MFPaQ software by processing our MS/MS spectra directly against the UniProt database (downloaded on October 2008). As expected, the number of identified and quantified proteins was lower (291), but the mean value of the coefficient of variations of the treated/control ratios for the different peptides identifying a single protein was quite similar to the one described above for the Unigenes (data not shown). Thus, all the validation steps indicate the suitability of the MFPaQ software to analyze the data from a quantitative proteomics experiment using $^{14}N/^{15}N$ labeling.

Cryptogein-induced Modification of Protein Association to Tobacco DRMs—These results described above prompted us to use the automatic quantification procedure to analyze the relative abundance of the proteins identified in DRMs extracted from control or elicited cells in the two reciprocal experiments.

Quantification of Unigenes using MFPaQ led to a similar distribution of individual ratios treated/control in the two samples; the median value is 1.07 for sample A and 1.04 for sample B (Fig. 7). A similar analysis performed on proteins identified using the UniProt database, thus giving direct protein quantification, yielded quite comparable results (1.075 for sample A and 1.03 for sample B).

In the two samples, 80% of the Unigenes quantified using MFPaQ exhibited a ratio of treated/control close to the median value of the experiment (between 0.8 and 1.4), indicating that the association of most of the proteins to DRMs is not significantly modified by cryptogein. Furthermore a control experiment mixing equally control ^{15}N- and control ^{14}N-labeled cells led to a quite similar dispersion of the $^{15}N/^{14}N$ ratios; 80% of them were between 0.9 and 1.3.

We thus analyzed the 20% of Unigenes that were out of this set in the two experiments: they were equally distributed in the two samples above (10%) and below (10%) the median value, leading to threshold values that were identical in the two experiments (0.8 for exclusion and 1.4 for enrichment). The Unigenes exhibiting a similar modification of their association to DRMs after cryptogein treatment in the two experiments were then selected and submitted to the procedure of grouping using the "bulk search tool" to eliminate redundancy between Unigenes corresponding to a single protein as described above.

This led to the identification of five proteins: the abundance of four of them decreased in DRMs extracted from cryptogein treated cells, whereas one of them was more abundant in DRMs after elicitation (Table I). Interestingly the four proteins excluded from DRMs after cryptogein treatment are linked to vesicular trafficking (Dynamin-1A, Dynamin-1E, Dynamin-2A,

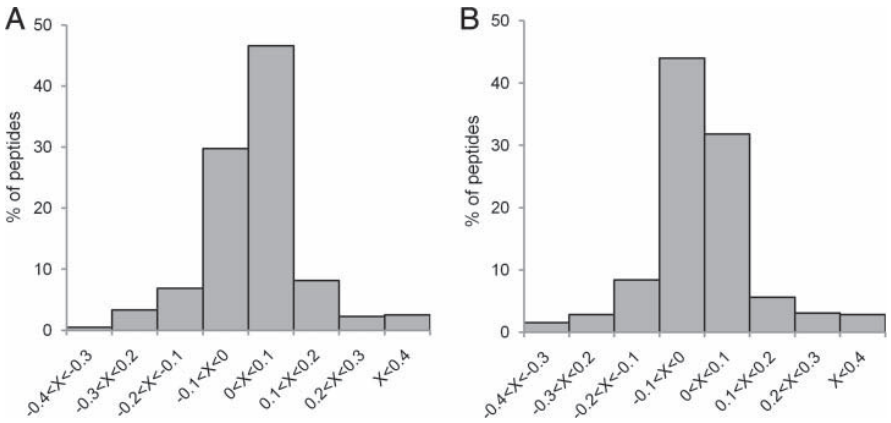


FIG. 5. **Distribution of relative distance between manual and automatic peptide quantifications.** Shown is the comparison of peptide intensity ratios as determined by MFPaQ and by manual inspection is expressed by the relative difference: (Automatic quantification − Manual quantification)/Manual quantification. *A*, sample A; *B*, sample B.

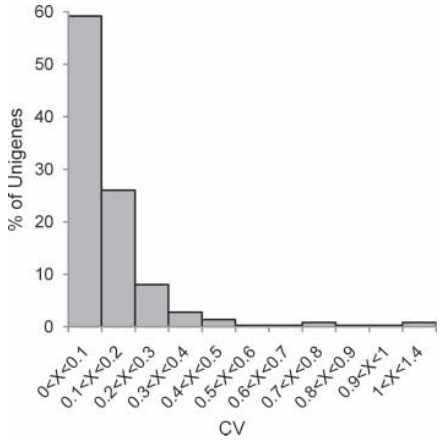


FIG. 6. **Distribution of coefficients of variation (*CV*) among cryptogein/control ratios of peptides identifying a single protein.** For all proteins identified with at least two peptides this coefficient of variation was obtained by dividing the standard deviation of the ratios of intensity (cryptogein/control) of the different identifying peptides by the mean value of these ratios.

and Dynamin-2B), whereas the protein that was more abundant in DRMs after elicitation is a signaling proteins (a 14-3-3 protein).

DISCUSSION

Automatic Procedure for Quantitative Analysis of Protein Abundance following $^{15}N/^{14}N$ Labeling—The number of proteins identified in the present study (350) is higher than the number found in our previous work (15). This can be explained by the fact that we used a more powerful mass spectrometer (LTQ-Orbitrap) here than in the previously published study (nanospray LCQ Deca XP Plus ion trap mass spectrometer) and that we performed the database search against a Solanacae-specific database allowing a higher level of identification. However, both the identities of proteins and their functional groups are consistent between the two studies; the proteins involved in signaling and response to stress were in both cases the most represented (supplemental data S3). In counterpart, the nano-LC-MS/MS analysis performed on $^{14}N/^{15}N$-labeled samples generated a large amount of data. This in turn necessitated appropriate bioinformatics tools for data analysis. Here we used the MFPaQ software (48) that was initially developed to perform quantification of ICAT and SILAC experiments. The quantification module has been extended to manage ^{15}N labeling experiments, and the results obtained with the software were validated by manual assessment on a subset of peptides (Fig. 5). The results obtained indicate a close correlation between manual and automatic peptide quantification and a low deviation of ratios for peptides identifying a single protein.

A few other tools exist for performing quantitative analysis of ^{15}N labeling experiments. Andreev *et al.* (53) developed the Quantitation of N-15/14 (QN) algorithm based on the Trans-Proteomic Pipeline (54) for $^{14}N/^{15}N$ quantification of identification results obtained with the Sequest search engine and therefore not compatible with Mascot. Moreover some modules of this program are written in MATLAB and thus require the installation of this package. Another module has been developed for the Trans-Proteomic Pipeline by Palmblad *et al.* (55). It uses Mascot identification results as input and any type of raw files converted to the mzXML format. The Mascot search must be run twice, once against ^{14}N masses and once against ^{15}N masses, resulting in two Mascot result files for each raw file that must be subsequently renamed using an external script. Finally the MSQuant program (56), often used for SILAC experiments, can also perform $^{14}N/^{15}N$ quantification starting from Mascot results and LTQ-Orbitrap raw files and was recently described for such applications (44). However, this standalone software also needs extra steps (Mascot results export and conversion using manual configuration and execution of scripts). The MFPaQ software offers a very user-friendly alternative for integrated protein validation and $^{14}N/^{15}N$ quantification through a Web browser interface. It can read directly non-processed RAW files from an LTQ-Orbitrap or standard mzXML files and is also well integrated with the Mascot search engine (direct access to the identification results on the Mascot server). MFPaQ provides an interactive interface allowing the user to organize and compile results

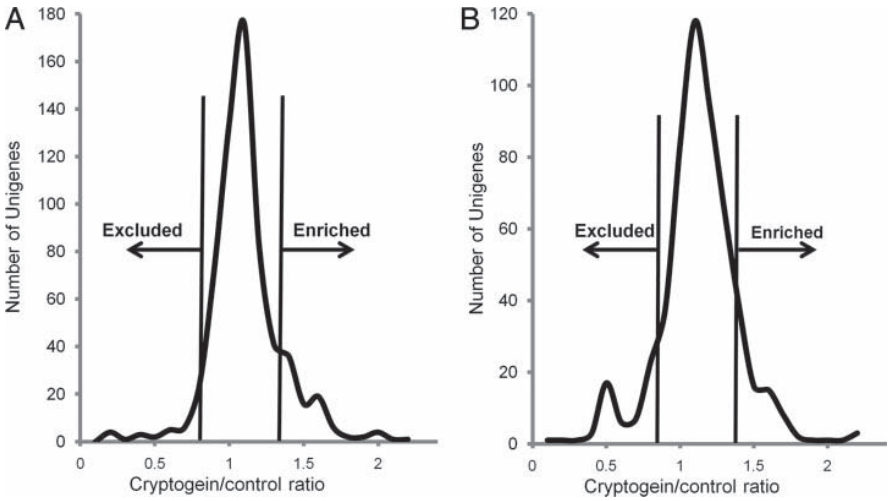


FIG. 7. **Distribution of cryptogein/control protein ratios in samples A and B.** In the two samples, ratios were calculated for each protein as the mean value of the ratios of its identifying peptides. *Vertical lines* indicate the thresholds corresponding to the 10% of proteins exhibiting the highest (enriched) or the lowest (excluded) ratios in each experiment. The values of these thresholds are identical in the two experiments: 0.8 for exclusion and 1.4 for enrichment.

TABLE I

Proteins differentially quantified in DRMs extracted from either control or elicited cells

Among the 10% of proteins exhibiting the highest or the lowest cryptogein/control intensity ratios, those indicated in this table behaved similarly in the two experiments. Sp., plant species abbreviated as follows: *N.t.*, *N. tabacum*; *A.t.*, *A. thaliana*. Acc. no., accession number in GenBank™ database. CV, coefficient of variation among the treated/control ratios of peptides identifying a single protein.

Acc. no.	Sp.	Protein name	Number of Unigenes	Sample A			Sample B		
				Global ratio	Number of peptides	CV	Global ratio	Number of peptides	CV
Vesicular trafficking									
GI:5931765	*N.t.*	Dynamin-1A	2	0.41	13	3.13	0.68	20	10.7
GI:18411520	*A.t.*	Dynamin-1E	3	0.45	15	10.7	0.71	12	5.6
GI:15218486	*A.t.*	Dynamin-2A	1	0.394	5	16	0.79	4	16
GI:15218837	*A.t.*	Dynamin-2B	1	0.524	4	4.5	0.73	2	4.7
Signaling									
GI:3023189	*N.t.*	14-3-3 C	1	1.43	5	5.1	2.15	4	3.4

from shotgun experiments and proven to be efficient for data validation and quantification after ^{15}N labeling, protein fractionation, analysis of consecutive fractions by several nano-LC-MS/MS runs, and multisearch with the Mascot engine. It was used successfully in this study to point out variations of protein abundances in the DRM compartment of plant cells following stimulation with cryptogein.

Quantitative Analysis of the Tobacco Cell DRM Proteome: Effect of Cryptogein Stimulation—In animal cells, a few studies report a quantitative analysis of compositional changes in DRM protein content upon biological stimulus. MacLellan *et al.* (57), using ICAT, reported an increase of 23 proteins in DRMs isolated from human smooth muscle cells after stimulation with platelet-derived growth factor. Using the same technique of chemical labeling, Gupta *et al.* (8) found three proteins enriched and one depleted in DRMs isolated from B cells following ligation of the BCR with antigens. Using the two-dimensional DIGE system Kobayashi *et al.* (58) identified 20 proteins, the abundance of which increased in DRMs following T cell receptor stimulation in T lymphocytes. Very recently, a quantitative analysis of macrophage DRM proteome using stable isotope labeling of amino acids indicated an enrichment of about 15 proteins in response to lipopolysaccharide (59). Thus, although a deep modification in the composition of lipid raft proteome has been proposed following T cell antigen receptor triggering (60), the above cited studies are in favor of a limited modification of protein association to DRMs upon biological stimulus. This is consistent with our results indicating one protein enriched and four depleted in tobacco DRMs following cryptogein treatment. Moreover enrichment or depletion factors (ranging in our study from 0.4 to 2.1) are in agreement with studies performed in animal cells indicating factors ranging from 0.8 to 1.85 in macrophage DRMs in response to lipopolysaccharides (59), from 0.37 to 3.5 in DRMs from B cells (8), or from 1.3 to 1.8 in DRMs from human smooth muscles challenged with growth factor (57). Finally it has to be noted that times of treatment with the biological stimulus in these studies (5 min (8, 59) or 15 min (57)) are quite comparable to those used in this study (5 min of cryptogein treatment).

Although metabolic labeling and quantitative proteomics were used to analyze the sterol dependence of protein association to *A. thaliana* DRMs (44), only one very recent study has so far been performed to determine the variation of pro-

tein content of plant DRMs upon a biological stimulus (61). However, this study has been performed comparing DRMs from control plants and plants acclimated to low temperatures for several days and thus does not refer to an early signaling process. The data presented here indicate that a process of discrete modification of DRM protein content seems to occur very early in a similar manner in plant and animal cells in a context of biological stimulation. It has to be noted that, in many cases, functional studies have confirmed the biological relevance of the data obtained by quantitative proteomics studies. For instance, the protein ezrin had been shown using quantitative proteomics to be excluded from DRMs of B cells upon ligation of antigen receptor: functional studies using cells transformed with dominant positive mutants of this protein demonstrated that it was an inhibitor of lipid rafts coalescence, indicating that the release of ezrin from lipid rafts regulates their dynamics during signaling (8). In a similar manner, histochemical analysis confirmed that some proteins enriched in T cell DRMs after activation were indeed redistributed from cytosol to microdomains of the membrane during this process. All these data seem to confirm the hypothesis that plant microdomains could play, like their animal counterpart, a key role in the signaling process leading to the establishment of cell responses.

Biological Significance of Variations of Protein Content Observed in DRMs of Cryptogein-elicited Tobacco Cells—The proteins, the abundance of which is modified in DRMs after elicitation of tobacco cells with cryptogein, belong to two functional families: vesicular trafficking and signaling. Similar results have already been obtained in animal cells: upon BCR engagement, apart from proteins corresponding to the BCR complex, all proteins that exhibited modifications of their association to DRMs were linked to the cytoskeleton or vesicle trafficking (8). Consistently proteins undergoing modification of their association to human smooth muscle cell DRMs after treatment with growth factor are essentially cytoskeletal proteins and endocytosis-related proteins (57). These two examples come from quantitative proteomics studies of the DRM fraction; however, a huge amount of the data probing the association to DRMs of particular proteins using immunological tools clearly indicates that signal transduction and membrane trafficking are two physiological process essentially linked to DRMs (62–65). In tobacco the functional grouping of DRM and plasma membrane proteins indicated that proteins involved in signaling and cell trafficking undergo an increase of their relative importance in DRMs compared with plasma membrane (15). This suggested that, like their animal counterpart, plant DRMs could be involved in the regulation of such a physiological function, which is confirmed by the results presented in this study.

The fungal elicitor cryptogein has been proven in tobacco cells to trigger numerous physiological events, obviously corresponding to a signaling cascade. Thus, the identification in this study of a DRM-associated protein related to signaling is quite relevant. This protein is a 14-3-3 protein enriched in DRMs after cryptogein treatment. Studies over the past 20 years have proven 14-3-3 to be ubiquitous; it is found in most eukaryotic organisms and tissues (66). In animals, these proteins play a central role in the regulation of many cellular processes such as control of the cell cycle, differentiation, apoptosis, targeting of proteins to different cellular locations, and coordination of multiple signal transduction pathways (67–70). These proteins could achieve such functions by directly regulating the activity of proteins involved in a signal transduction cascade, promoting the formation of multiprotein complexes, or modulating the expression of particular genes by regulating the activity or localization of transcription factors. In a previous study, we used a two-hybrid screen to find proteins able to specifically interact with NtrbohD, the tobacco oxidase that has been proven to be responsible for ROS production in tobacco cells elicited with cryptogein (29). These experiments led to the isolation of a cDNA encoding a protein belonging to the family of 14-3-3 proteins (Nth14-3-3, AJ309008). When BY-2 cell lines were transformed with antisense constructs of *Nth14-3-3*, the expression of the antisense transgene was correlated with a strong inhibition of ROS production following elicitation with cryptogein (72). This demonstrated the involvement of a 14-3-3 protein in the regulation of ROS production. The 14-3-3 protein identified in the present study is extremely close to Nth14-3-3 (89% identity and 94% homology), making quite plausible a functional redundancy between these two isoforms (72). Indeed the question of the specificity of function of this isoform remains because the tobacco genome comprises 17 isoforms of 14-3-3 proteins, and studies on the binding of different isoforms of 14-3-3 to different isoforms of H^+-ATPase in *A. thaliana* indicated no absolute specificity (73). However, it has to be noted that among the different isoforms of 14-3-3 identified in this study several exhibited a treated/control ratio above 1.4 in one experiment and slightly below in the other. Thus although we only considered one isoform as strictly above the threshold in the two experiments, the hypothesis of an association of several 14-3-3 proteins to DRMs upon cryptogein treatment cannot be excluded. NtrbohD has been proven to be exclusively associated to the tobacco DRM fraction in a sterol-dependent manner (30), and the enrichment in this fraction of a 14-3-3 protein able to act as a positive regulator of this oxidase at a timing corresponding to the onset of ROS production would be particularly biologically relevant.

Recently our group demonstrated that among the very early events triggered by cryptogein a clathrin-mediated endocytosis process occurred 5 min after stimulation (74). It is thus noteworthy that the four proteins identified as involved in cell trafficking belong to the dynamin-related protein (DRP) family. Dynamins are high molecular weight GTPases that play a central role in dynamics of membrane biogenesis and maintenance in eukaryotic cells that require a constant turnover of membrane constituents mediated by the processes of endo-

cytosis and exocytosis (for a review, see Ref. 75). The basic feature of this group of proteins is that they form helical structures able to wrap around the membranes and either tubulate them or pinch them off of larger membrane sheets. In *A. thaliana*, DRP1A expressed as a functional DRP1A-GFP fusion protein under the control of its native promoter formed discrete foci at the plasma membrane of root epidermal cells (76) consistent with a localization in membrane domains of this protein. The same study indicated that DRP1A-GFP dynamics are perturbed upon treatment with fenpropimorph, a pharmacological inhibitor of the sterol biosynthetic pathway in plants, or with compounds, such as tyrphostin A23, known to block clathrin-mediated endocytosis. Moreover a null mutant for DRP1A exhibited reduced endocytosis, indicating the involvement of this protein in such a process. DRP1A and DRP1E from *A. thaliana* share 80% homology, and no clear data are available concerning the physiological role of DRP1E. The DRP2 subfamily is characterized by the presence of a pleckstrin homology domain believed to bind to phosphoinositides with a broad range of specificity and affinity (77). In recent years, both phosphoinositides and phosphoinositide-binding proteins have been reported to display a restricted, rather than a uniform, distribution across intracellular membranes (78). A significant enrichment of BY-2 cell DRMs in phosphoinositides species has been observed.[2] Moreover from a functional point of view, recent results obtained in plants indicate a crucial role of phosphatidylinositol in the setup of endocytosis triggered either by salt stress (79) or *Rhizobium* infection (80). Finally DRP2A is involved in clathrin-coated vesicle trafficking (81), and a colocalization of this protein with DRP2B has been reported (71). All these data are consistent both with the localization in tobacco DRMs of these four dynamin-related proteins and with their putative involvement in an endocytotic process already demonstrated in our model at a timing (5 min) in total agreement with the proteomics analysis conducted here. Finally the decrease observed for these proteins in their association to DRMs is, in this context, quite coherent with their internalization concomitant to the vesicle formation. We are currently analyzing further the role of these candidate proteins in the signaling process triggered by the elicitor using reverse genetics.

Conclusion—The development and validation of an automatic procedure for analyzing the data generated after *in vivo* labeling of tobacco cells allowed a quantitative analysis of the variation of a whole subcellular proteome after stimulation by an elicitor of defense reaction. The discrete variation of association to DRMs of some proteins upon cryptogein treatment indicate that plant microdomains could, like their animal counterpart, play a role in the signaling process underlying the setup of defense reaction in plants. Furthermore proteins identified as differentially associated to tobacco DRMs after cryptogein challenge are involved in signaling and vesicular trafficking as already observed in similar studies performed in animal cells upon biological stimuli. This suggests that the ways by which the dynamic association of proteins to microdomains could participate in the regulation of signaling process may be conserved between plant and animals. Finally this study led to the identification of five putative new actors of the cryptogein signaling pathway. Future investigation will have to determine the position of this dynamics of association to microdomains in the signaling cascade triggered by the elicitor and the precise role of these proteins in this process in link with physiological events already identified.

Acknowledgments—We are grateful to M. Ponchet and B. Industrie for the gift of cryptogein and to N. Leborgne-Castel and Sebastien Mongrand for critical reading of the manuscript.

* This work was supported in part by the Plant Health division of INRA (Ph.D. grant to T. S.), by the Conseil Régional de Bourgogne (Ph.D. grant to T. S. and J. M. and postdoctoral grant to S. V.), and by Agence Nationale de la Recherche (ANR) Grant ANR-05-JCJC-0209.

S The on-line version of this article (available at http://www.mcponline.org) contains supplemental material.

§ Both authors contributed equally to this work.

‡‡ Supported by grants from the "Fondation pour la Recherche Médicale" (FRM-contrat "Grands Equipements"), the Génopole Toulouse Midi-Pyrénées, and the Région Midi-Pyrénées and by ANR Grant ANR-Plates-Formes Technologiques du Vivant (PFTV).

§§ To whom correspondence should be addressed: UMR Plante Microbe Environnement INRA 1088/CNRS 5184/Université de Bourgogne, 17 Rue Sully, BP 86510 21065 Dijon Cedex, France. Tel.: 33-3-80-69-32-75; Fax: 33-3-80-69-37-53; E-mail: simon@dijon.inra.fr.

[2] S. Mongrand, personal communication.

REFERENCES

1. Brown, D. A., and London, E. (1998) Structure and origin of ordered lipid domains in biological membranes. *J. Membr. Biol.* **164,** 103–114
2. Rietveld, A., and Simons, K. (1998) The differential miscibility of lipids as the basis for the formation of functional membrane rafts. *Biochim. Biophys. Acta* **1376,** 467–479
3. Simons, K., and Ikonen, E. (1997) Functional rafts in cell membranes. *Nature* **387,** 569–572
4. Pike, L. J. (2006) Rafts defined: a report on the Keystone Symposium on Lipid Rafts and Cell Function. *J. Lipid Res.* **47,** 1597–1598
5. Simons, K., and Toomre, D. (2000) Lipid rafts and signal transduction. *Nat. Rev. Mol. Cell Biol.* **1,** 31–39
6. Rajendran, L., and Simons, K. (2005) Lipid rafts and membrane dynamics. *J. Cell Sci.* **118,** 1099–1102
7. Sedwick, C. E., and Altman, A. (2002) Ordered just so: lipid rafts and lymphocyte function. *Sci. STKE* **2002,** RE2
8. Gupta, N., Wollscheid, B., Watts, J. D., Scheer, B., Aebersold, R., and DeFranco, A. L. (2006) Quantitative proteomics analysis of B cell lipid rafts reveals that ezrin regulates antigen receptor-mediated lipid rafts dynamics. *Nat. Immunol.* **7,** 625–633
9. Zappel, N. F., and Panstruga, R. (2008) Heterogeneity and lateral compartmentalization of plant plasma membranes. *Curr. Opin. Plant Biol.* **11,** 632–640
10. Peskan, T., Westermann, M., and Oelmüller, R. (2000) Identification of low-density Triton X-100-insoluble plasma membrane microdomains in higher plants. *Eur. J. Biochem.* **267,** 6989–6995
11. Mongrand, S., Morel, J., Laroche, J., Claverol, S., Carde, J. P., Hartmann, M. A., Bonneu, M., Simon-Plas, F., Lessire, R., and Bessoule, J. J. (2004) Lipid rafts in higher plant cells: purification and characterization of Triton X-100-insoluble microdomains from tobacco plasma membrane. *J. Biol. Chem.* **279,** 36277–36286
12. Borner, G. H., Sherrier, D. J., Weimar, T., Michaelson, L. V., Hawkins, N. D.,

Macaskill, A., Napier, J. A., Beale, M. H., Lilley, K. S., and Dupree, P. (2005) Analysis of detergent-resistant membranes in Arabidopsis. Evidence for plasma membrane lipid rafts. *Plant Physiol.* **137,** 104–116
13. Lefebvre, B., Furt, F., Hartmann, M. A., Michaelson, L. V., Carde, J. P., Sargueil-Boiron, F., Rossignol, M., Napier, J. A., Cullimore, J., Bessoule, J. J., and Mongrand, S. (2007) Characterization of lipid rafts from Medicago truncatula root plasma membranes: a proteomic study reveals the presence of a raft-associated redox system. *Plant Physiol.* **144,** 402–418
14. Bhat, R. A., Miklis, M., Schmelzer, E., Schulze-Lefert, P., and Panstruga, R. (2005) Recruitment and interaction dynamics of plant penetration resistance components in a plasma membrane microdomain. *Proc. Natl. Acad. Sci. U.S.A.* **102,** 3135–3140
15. Morel, J., Claverol, S., Mongrand, S., Furt, F., Fromentin, J., Bessoule, J. J., Blein, J. P., and Simon-Plas, F. (2006) Proteomics of plant detergent-resistant membranes. *Mol. Cell. Proteomics* **5,** 1396–1411
16. Ricci, P. (1997) in *Plant-Microbe Interactions* (Stacey, G., and Keen, N. T., eds) pp. 53–75, Chapman and Hall, New York
17. Osman, H., Vauthrin, S., Mikes, V., Milat, M. L., Panabières, F., Marais, A., Brunie, S., Maume, B., Ponchet, M., and Blein, J. P. (2001) Mediation of elicitin activity on tobacco is assumed by elicitin-sterol complexes. *Mol. Biol. Cell.* **12,** 2825–2834
18. Vauthrin, S., Mikes, V., Milat, M. L., Ponchet, M., Maume, B., Osman, H., and Blein, J. P. (1999) Elicitins trap and transfer sterols from micelles, liposomes and plant plasma membranes. *Biochim. Biophys. Acta* **1419,** 335–342
19. Wendehenne, D., Binet, M. N., Blein, J. P., Ricci, P., and Pugin, A. (1995) Evidence for specific, high-affinity binding sites for a proteinaceous elicitor in tobacco plasma membrane. *FEBS Lett.* **374,** 203–207
20. Blein, J. P., Milat, M. L., and Ricci, P. (1991) Response of cultured tobacco cells to cryptogein, a proteinaceous elicitor from *Phytophthora cryptogea.* Possible plasmalemma involvement. *Plant Physiol.* **95,** 486–491
21. Wendehenne, D., Lamotte, O., Frachisse, J. M., Barbier-Brygoo, H., and Pugin, A. (2002) Nitrate efflux is an essential component of the cryptogein signaling pathway leading to defense responses and hypersensitive cell death in tobacco. *Plant Cell* **14,** 1937–1951
22. Lecourieux, D., Mazars, C., Pauly, N., Ranjeva, R., and Pugin, A. (2002) Analysis and effects of cytosolic free calcium increases in response to elicitors in Nicotiana plumbaginifolia cells. *Plant Cell* **14,** 2627–2641
23. Lebrun-Garcia, A., Ouaked, F., Chiltz, A., and Pugin, A. (1998) Activation of MAPK homologues by elicitors in tobacco cells. *Plant J.* **15,** 773–781
24. Zhang, S., Du, H., and Klessig, D. F. (1998) Activation of the tobacco SIP kinase by both a cell wall-derived carbohydrate elicitor and purified proteinaceous elicitins from *Phytophthora* spp. *Plant Cell* **10,** 435–450
25. Gould, K. S., Lamotte, O., Klinguer, A., Pugin, A., and Wendehenne, D. (2003) Nitric oxide production in tobacco leaf cells: a generalized stress response? *Plant Cell Environ.* **26,** 1851–1862
26. Lamotte, O., Gould, K., Lecourieux, D., Sequeira-Legrand, A., Lebrun-Garcia, A., Durner, J., Pugin, A., and Wendehenne, D. (2004) Analysis of nitric oxide signaling functions in tobacco cells challenged by the elicitor cryptogein. *Plant Physiol.* **135,** 516–529
27. Rusterucci, C., Stallaert, V., Milat, M. L., Pugin, A., Ricci, P., and Blein, J. P. (1996) Relationship between active oxygen species, lipid peroxidation, necrosis, and phytoalexin production induced by elicitins in *Nicotiana. Plant Physiol.* **111,** 885–891
28. Simon-Plas, F., Rustérucci, C., Milat, M. L., Humbert, C., Montillet, J. L., and Blein, J. P. (1997) Active oxygen species production in tobacco cells elicited by cryptogein. *Plant Cell Environ.* **20,** 1573–1579
29. Simon-Plas, F., Elmayan, T., and Blein, J. P. (2002) The plasma membrane oxidase NtrbohD is responsible for AOS production in elicited tobacco cells. *Plant J.* **31,** 137–147
30. Roche, Y., Gerbeau-Pissot, P., Buhot, B., Thomas, D., Bonneau, L., Gresti, J., Mongrand, S., Perrier-Cornet, J. M., and Simon-Plas, F. (2008) Depletion of phytosterols from the plant plasma membrane provides evidence for disruption of lipid rafts. *FASEB J.* **22,** 3980–3991
31. Tonge, R., Shaw, J., Middleton, B., Rowlinson, R., Rayner, S., Young, J., Pognan, F., Hawkins, E., Currie, I., and Davison, M. (2001) Validation and development of fluorescence two-dimensional differential gel electrophoresis proteomics technology. *Proteomics* **1,** 377–396
32. Amme, S., Matros, A., Schlesier, B., and Mock, H. P. (2006) Proteome analysis of cold stress response in *Arabidopsis thaliana* using DIGE-technology. *J. Exp. Bot.* **57,** 1537–1546
33. Bohler, S., Bagard, M., Oufir, M., Planchon, S., Hoffmann, L., Jolivet, Y., Hausman, J. F., Dizengremel, P., and Renaut, J. (2007) A DIGE analysis of developing poplar leaves subjected to ozone reveals major changes in carbon metabolism. *Proteomics* **7,** 1584–1599
34. Casati, P., Zhang, X., Burlingame, A. L., and Walbot, V. (2005) Analysis of leaf proteome after UV-B irradiation in maize lines differing in sensitivity. *Mol. Cell. Proteomics* **4,** 1673–1685
35. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* **422,** 198–207
36. Gingras, A. C., Gstaiger, M., Raught, B., and Aebersold, R. (2007) Analysis of protein complexes using mass spectrometry. *Nat. Rev. Mol. Cell Biol.* **8,** 645–654
37. Thelen, J. J., and Peck, S. C. (2007) Quantitative proteomics in plants: choices in abundance. *Plant Cell* **19,** 3339–3346
38. Beynon, R. J., and Pratt, J. M. (2005) Metabolic labeling of proteins for proteomics. *Mol. Cell. Proteomics.* **4,** 857–872
39. Bindschedler, L. V., Palmblad, M., and Cramer, R. (2008) Hydroponic isotope labelling of entire plants (HILEP) for quantitative plant proteomics: an oxidative case study. *Phytochemistry* **69,** 1962–1972
40. Hebeler, R., Oeljeklaus, S., Reidegeld, K. A., Eisenacher, M., Stephan, C., Sitek, B., Stühler, K., Meyer, H. E., Sturre, M. J., Dijkwel, P. P., and Warscheid, B. (2008) Study of early leaf senescence in *Arabidopsis thaliana* by quantitative proteomics using reciprocal $^{14}N/^{15}N$ labeling and difference gel electrophoresis. *Mol. Cell. Proteomics* **7,** 108–120
41. Huttlin, E. L., Hegeman, A. D., Harms, A. C., and Sussman, M. R. (2007) Comparison of full versus partial metabolic labeling for quantitative proteomics analysis in *Arabidopsis thaliana. Mol. Cell. Proteomics* **6,** 860–881
42. Nelson, C. J., Huttlin, E. L., Hegeman, A. D., Harms, A. C., and Sussman, M. R. (2007) Implication of ^{15}N-metabolic labeling for automated peptide identification in *Arabidopsis thaliana. Proteomics* **7,** 1279–1292
43. Engelsberger, W. R., Erban, A., Kopka, J., and Schulze, W. X. (2006) Metabolic labeling of plant cell cultures with K(15)NO3 as a tool for quantitative analysis of proteins and metabolites. *Plant Methods* **2,** 1–11
44. Kierszniowska, S., Seiwert, B., and Schulze, W. X. (2009) Definition of Arabidopsis sterol-rich membrane microdomains by differential treatment with methyl-beta-cyclodextrin and quantitative proteomics. *Mol. Cell. Proteomics* **8,** 612–623
45. Palmblad, M., Mills, D. J., and Bindschedler, L. V. (2008) Heat-shock response in *Arabidopsis thaliana* explored by multiplexed quantitative proteomics using differential metabolic labelling. *J. Proteome Res.* **7,** 780–785
46. Lanquar, V., Kuhn, L., Lelièvre, F., Khafif, M., Espagne, C., Bruley, C., Barbier-Brygoo, H., Garin, J., and Thomine, S. (2007) 15N-metabolic labeling for comparative plasma membrane proteomics in *Arabidopsis* cells. *Proteomics* **7,** 750–754
47. Benschop, J. J., Mohammed, S., O'Flaherty, M., Heck, A. J., Slijper, M., and Menke, F. L. (2007) Quantitative phosphoproteomics of early elicitor signaling in *Arabidopsis. Mol. Cell. Proteomics* **6,** 1198–1214
48. Bouyssié, D., Gonzalez de Peredo, A., Mouton, E., Albigot, R., Roussel, L., Ortega, N., Cayrol, C., Burlet-Schiltz, O., Girard, J. P., and Monsarrat, B. (2007) Mascot file parsing and quantification (MFPaQ), a new software to parse, validate and quantify proteomics data generated by ICAT and SILAC mass spectrometric analyses: application to the proteomics study of membrane proteins from primary human endothelial cells. *Mol. Cell. Proteomics* **6,** 1621–1637
49. Murashige, T., and Skoog, F. (1962) A revised method for rapid growth and bioassays with tobacco tissue cultures. *Physiol. Plant* **15,** 473–497
50. Larsson, C., Sommarin, M., and Widell, S. (1994) in *Aqueous Two-Phase Systems* (Walter, H., and Johansson, G., eds) Vol. 228, pp. 451–459, Academic Press Inc., San Diego, CA
51. Borderies, G., Jamet, E., Lafitte, C., Rossignol, M., Jauneau, A., Boudart, G., Monsarrat, B., Esquerré-Tugayé, M. T., Boudet, A., and Pont-Lezica, R. (2003) Proteomics of loosely bound cell wall proteins of *Arabidopsis thaliana* cell suspension cultures: a critical analysis. *Electrophoresis* **24,** 3421–3432
52. Mueller, L. A., Solow, T. H., Taylor, N., Skwarecki, B., Buels, R., Binns, J., Lin, C., Wright, M. H., Ahrens, R., Wang, Y., Herbst, E. V., Keyder, E. R., Menda, N., Zamir, D., and Tanksley, S. D. (2005) The SOL genomics network. A comparative resource for Solanacae biology and beyond. *Plant Physiol.* **138,** 1310–1317

53. Andreev, V. P., Li, L., Rejtar, T., Li, Q., Ferry, J. G., and Karger, B. L. (2006) New algorithm for $^{15}N/^{14}N$ quantitation with LC-ESI-MS using an LTQ-FT mass spectrometer. *J. Proteome Res.* **5,** 2039–2045
54. Keller, A., Eng, J., Zhang, N., Li, X. J., and Aebersold, R. (2005) A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol. Syst. Biol.* **1,** 2005.0017
55. Palmblad, M., Bindschedler, L. V., and Cramer, R. (2007) Quantitative proteomics using uniform ^{15}N-labeling, MASCOT, and the trans-proteomic pipeline. *Proteomics* **7,** 3462–3469
56. Andersen, J. S., Wilkinson, C. J., Mayor, T., Mortensen, P., Nigg, E. A., and Mann, M. (2003) Proteomic characterization of the human centrosome by protein correlation profiling. *Nature* **426,** 570–574
57. MacLellan, D. L., Steen, H., Adam, R. M., Garlick, M., Zurakowski, D., Gygi, S. P., Freeman, M. R., and Solomon, K. R. (2005) A quantitative proteomic analysis of growth factor-induced compositional changes in lipid rafts of human smooth muscle cells. *Proteomics* **5,** 4733–4742
58. Kobayashi, M., Katagiri, T., Kosako, H., Iida, N., and Hattori, S. (2007) Global analysis of dynamic changes in lipid raft proteins during T-cell activation. *Electrophoresis* **28,** 2035–2043
59. Dhungana, S., Merrick, B. A., Tomer, K. B., and Fessler, M. B. (2009) Quantitative proteomics analysis of macrophage rafts reveals compartmentalized activation of the proteasome and proteasome-mediated ERK activation in response to lipopolysaccharide. *Mol. Cell. Proteomics* **8,** 201–213
60. Bini, L., Pacini, S., Liberatori, S., Valensin, S., Pellegrini, M., Raggiaschi, R., Pallini, V., and Baldari, C. T. (2003) Extensive temporally regulated reorganization of the lipid raft proteome following T-cell antigen receptor triggering. *Biochem. J.* **369,** 301–309
61. Minami, A., Fujiwara, M., Furuto, A., Fukao, Y., Yamashita, T., Kamo, M., Kawamura, Y., and Uemura, M. (2009) Alterations in detergent-resistant plasma membrane microdomains in *Arabidopsis thaliana* during cold acclimatation. *Plant Cell Physiol.* **50,** 341–359
62. Alonso, M. A., and Millán, J. (2001) The role of lipid rafts in signalling and membrane trafficking in T-lymphocytes. *J. Cell Sci.* **114,** 3957–3965
63. Dykstra, M., Cherukuri, A., Sohn, H. W., Tzeng, S. J., and Pierce, S. K. (2003) Location is everything: lipid rafts and immune cell signaling. *Annu. Rev. Immunol.* **21,** 457–481
64. Holowka, D., Gosse, J. A., Hammond, A. T., Han, X., Sengupta, P., Smith, N. L., Wagenknecht-Wiesner, A., Wu, M., Young, R. M., and Baird, B. (2005) Lipid segregation and IgE receptor signaling: a decade of progress. *Biochim. Biophys. Acta* **1746,** 252–259
65. Pike, L. J. (2003) Lipid rafts: bringing order to chaos. *J. Lipid Res.* **44,** 655–667
66. DeLille, J. M., Sehnke, P. C., and Ferl, R. J. (2001) The Arabidopsis 14-3-3 family of signaling regulators. *Plant Physiol.* **126,** 35–38
67. Finnie, C., Borch, J., and Collinge, D. B. (1999) 14-3-3s: eukaryotic regulatory proteins with many functions. *Plant. Mol. Biol.* **40,** 545–554
68. Fu, H., Subramanian, R. R., and Masters, S. C. (2000) 14-3-3s: eukaryotic regulatory proteins with many functions. *Annu. Rev. Pharmacol. Toxicol.* **40,** 617–647
69. Palmgren, M. G., Fuglsgang, A. T., and Jahn, T. (1998) Deciphering the role of 14-3-3s. *Exp. Biol. Online* **3**
70. Roberts, M. R. (2000) Regulatory 14-3-3 protein-protein interactions in plant cells. *Curr. Opin. Plant Biol.* **3,** 400–405
71. Fujimoto, M., Arimura, S., Nakazono, M., and Tsutsumi, N. (2008) Arabidopsis dynamin-related protein DRP2B is co-localized with DRP1A on the leading edge of the forming cell plate. *Plant Cell Rep.* **27,** 1581–1586
72. Elmayan, T., Fromentin, J., Riondet, C., Alcaraz, G., Blein, J. P., and Simon-Plas, F. (2007) Regulation of reactive oxygen species production by a 14-3-3 protein in elicited tobacco cells. *Plant Cell Environ.* **30,** 722–732
73. Alsterfjord, M., Sehnke, P. C., Arkell, A., Larsson, H., Svennelid, F., Rosenquist, M., Ferl, R. J., Sommarin, M., and Larsson, C. (2004) Plasma membrane H+-ATPase and 14-3-3 isoforms of Arabidopsis leaves: evidence for isoform specificity in the 14-3-3/H+-ATPase interaction. *Plant Cell Physiol.* **45,** 1202–1210
74. Leborgne-Castel, N., Lherminier, J., Der, C., Fromentin, J., Houot, V., and Simon-Plas, F. (2008) The plant defense elicitor cryptogein stimulates clathrin-mediated endocytosis correlated with reactive oxygen species production in Bright Yellow-2 tobacco cells. *Plant Physiol.* **146,** 1255–1266
75. Verma, D. P., and Hong, Z. (2005) The ins and outs in membrane dynamics: tubulation and vesiculation. *Trends Plant Sci.* **10,** 159–165
76. Konopka, C. A., and Bednarek, S. Y. (2008) Comparison of the dynamics and functional redundancy of the Arabidopsis dynamin-related isoforms DRP1A and DRP1C during plant development. *Plant Physiol.* **147,** 1590–1602
77. Salim, K., Bottomley, M. J., Querfurth, E., Zvelebil, M. J., Gout, I., Scaife, R., Margolis, R. L., Gigg, R., Smith, C. I., Driscoll, P. C., Waterfield, M. D., and Panayotou, G. (1996) Distinct specificity in the recognition of phosphoinositides by the pleckstrin homology domains of dynamin and Bruton's tyrosine kinase. *EMBO J.* **15,** 6241–6250
78. Carlton, J. G., and Cullen, P. J. (2005) Coincidence detection in phosphoinositide signaling. *Trends Cell Biol.* **15,** 540–547
79. König, S., Ischebeck, T., Lerche, J., Stenzel, I., and Heilmann, I. (2008) Salt-stress-induced association of phosphatidylinositol 4,5-bisphosphate with clathrin-coated vesicles in plants. *Biochem. J.* **415,** 387–399
80. Peleg-Grossman, S., Volpin, H., and Levine, A. (2007) Root hair curling and Rhizobium infection in Medicago truncatula are mediated by phosphatidylinositide-regulated endocytosis and reactive oxygen species. *J. Exp. Bot.* **58,** 1637–1649
81. Hong, Z., Bednarek, S. Y., Blumwald, E., Hwang, I., Jurgens, G., Menzel, D., Osteryoung, K. W., Raikhel, N. V., Shinozaki, K., Tsutsumi, N., and Verma, D. P. (2003) A unified nomenclature for Arabidopsis dynamin-related large GTPases based on homology and possible functions. *Plant Mol. Biol.* **53,** 261–265

III-4. Quantification sans marquage

L'analyse différentielle d'échantillons protéiques est devenue un outil important pour la compréhension de mécanismes biologiques intra et extra cellulaires. Le développement de nouvelles générations d'instruments a permis de démocratiser les approches de quantification sans marquage. Comme décrit dans l'introduction, ces approches peuvent être divisées en deux grandes familles (« Spectral counting », basée sur un comptage d'événements de séquençage MS/MS associés à une protéine donnée, ou « MS based », basée sur l'analyse et l'extraction du signal MS), et ce chapitre décrit la façon dont ces deux approches ont été implémentées dans le logiciel MFPaQ.

3.4.1 Spectral-counting

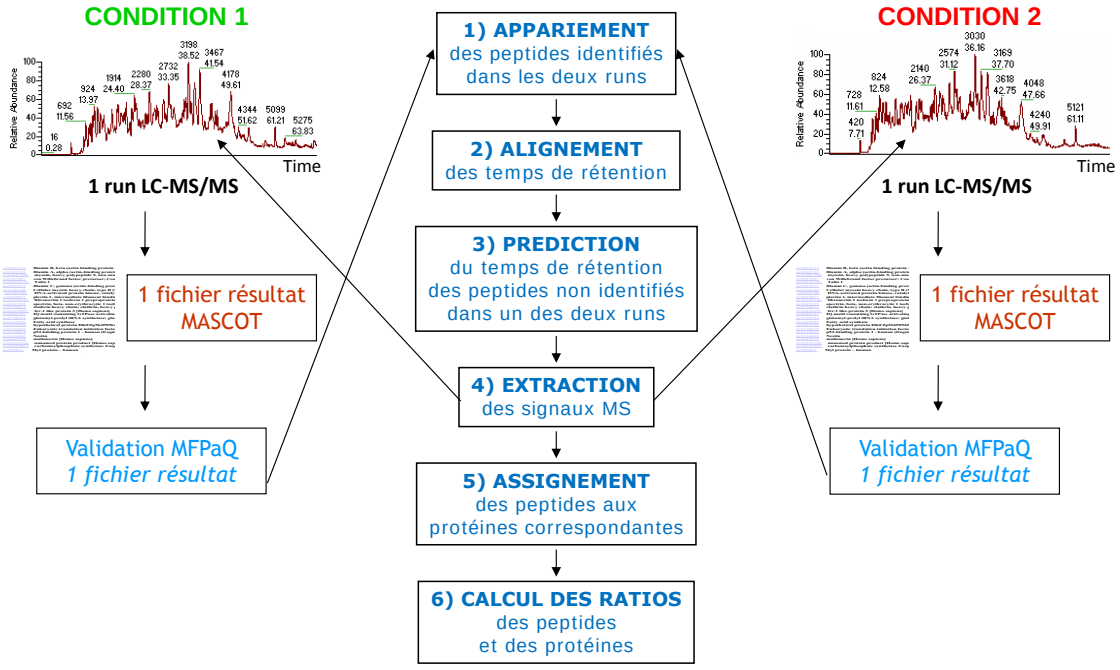
En 2004, Liu *et al.* (Liu, Sadygov et al. 2004) démontrent qu'il existe une forte corrélation entre le niveau d'abondance relative d'une protéine et le nombre de spectres MS/MS qui en sont issus. Cette valeur étant facilement accessible depuis les résultats du moteur de recherche, cette approche a eu un vif succès et de nombreuses méthodes statistiques ont été développées dans le but de fournir des résultats plus fiables. Cette simplicité d'utilisation a également permis une implémentation rapide dans les logiciels. Certains moteurs de recherche ont également étendu leurs fonctionnalités en intégrant ces nouvelles méthodes statistiques. Dans sa version 2.2 (disponible début 2007), Mascot a par exemple ajouté pour chaque protéine identifiée une valeur dérivée du « spectral counting » appelée emPAI (« exponentially modified Protein Abundance Index ») (Ishihama, Oda et al. 2005). Cet index est censé pondérer le nombre d'événements MS/MS observés par le nombre de peptides qui devraient être observés pour une protéine donnée (nombre fortement corrélé à la taille de la protéine). Ainsi l'index obtenu correspond à une valeur semi-quantitative qui permet d'une part d'obtenir une quantification relative (comparaison de l'abondance d'une même protéine dans différents échantillons), mais aussi d'approximer la quantité absolue de chaque protéine au sein d'un mélange, et de trier les protéines selon leur importance relative. Cependant dans le 1er cas (analyse différentielle de la même protéine entre deux échantillons) cet index n'a pas d'avantage par rapport à un comptage classique du nombre de MS/MS, dans la mesure où l'élément pondérant le nombre de MS/MS s'annule lorsque l'on calcule le quotient de deux index de la même protéine.

L'implémentation dans MFPaQ du « spectral counting » a été triviale puisqu'il a suffi d'ajouter dans les fichiers exportés par le logiciel le nombre de MS/MS pouvant être associés à chacune des protéines. Cette stratégie de quantification a été utilisée dans de nombreux projets, dont certains ont été publiés (Bousquet-Dubouch, Nguen et al. 2009; Burande, Heuze et al. 2009), et s'est avérée particulièrement efficace dans des études de complexes protéiques mettant en œuvre des techniques d'immuno-précipitation, où le complexe immunopurifié est généralement comparé à un contrôle afin de discriminer les partenaires spécifiques des contaminants.

3.4.2 Analyse des signaux MS

Dans MFPaQ nous avons choisi d'implémenter une approche supervisée (cf partie I-6.1 sur les techniques sans marquage) en réalisant une extraction de signal pour chaque peptide identifié par Mascot. L'élément principal qui a déterminé ce choix a été celui de la simplicité de l'approche. En effet la méthode non supervisée nécessite la mise en œuvre d'algorithmes d'analyse du signal bien plus complexes.

Comme nous l'avons vu ces deux méthodes possèdent des avantages et des limitations. L'approche non supervisée est plus exhaustive car l'ensemble des signaux observés est pris en compte, contrairement à la deuxième approche où l'on ne quantifie que les peptides identifiés par MS/MS. L'algorithme de quantification que j'ai implémenté dans MFPaQ comporte les étapes suivantes :



1) APPARIEMENT

Le logiciel commence par lister et associer l'ensemble des peptides identifiés dans les acquisitions LC-MS/MS à comparer. Si un peptide a été identifié avec plusieurs états de charge, ces derniers sont tous considérés. Au final on obtient une liste d'ions peptidiques où chaque espèce a été au moins identifiée dans une des acquisitions à comparer.

2) ALIGNEMENT

Le logiciel sélectionne les ions peptidiques qui sont identifiés dans les toutes les acquisitions à comparer et détermine leur temps d'élution dans chacune des acquisitions. Pour cela il recherche les spectres MS/MS correspondant à chaque ion qu'il a précédemment sélectionné. Pour chacun de ces spectres l'ion précurseur le plus intense est pris comme référence de temps d'élution du peptide.

Voici sous forme de pseudo code l'algorithme correspondant à cette étape :

```
FOREACH peptideIon IN peptideIons
  CALL peptideIon.getMs2Spectra RETURNING ms2Spectra
  SET precursorIons TO array
  FOREACH ms2Spectrum IN ms2Spectra
    CALL ms2Spectrum.getPrecursorIon RETURNING precursorIon
    precursorIons.add(precursorIon)
  ENDFOREACH
  CALL getMostIntensePrecursorIon WITH precursorIons RETURNING bestPrecursorIon
  peptideIon.elutionTime = bestPrecursorIon.elutionTime
ENDFOREACH
```

Le résultat correspond à une liste d'ions peptidiques identifiés dans toutes les acquisitions et pour lesquels on connaît le temps d'élution : on parle de « landmarks ». A partir de ces données il est possible de générer une matrice des temps d'élution. Le logiciel choisit un « run » (acquisition) de référence au hasard et calcule pour chaque « landmark » l'écart de temps entre chaque « run » et la référence choisie :

Séquence peptidique	Charge	Temps élution RUN n°1 (minutes)	Temps élution RUN n°2 (minutes)	Ecart de temps R2-R1
AFSGYLGTDQSK	2+	24,31	26,24	1.94
HRPQVAIICGSGLGGLTDK	3+	36,26	39,96	3.7
HRPQVAIICGSGLGGLTDK	2+	36,25	39,95	3.7
ADLINNLGTIAK	2+	48,01	52,07	4.06
GFGGIGGILR	2+	52,28	55,91	3.63
ELEQVCNPIISGLYQGAGGPGPGGFGAQGPK	3+	82,02	83,29	1.26
VYSPHVLNLTLIDLPGITK	3+	94,01	93,91	-0.10
…				

3) PREDICTION

La liste obtenue après l'étape n°1 est incomplète car elle ne comporte des valeurs de temps d'élution que dans les runs où le peptide a été séquencé. L'obtention de la matrice d'alignements des runs permet cependant de prédire ces valeurs dans les runs où le peptide n'a pas été identifié mais où il peut potentiellement être présent. Pour prédire ces valeurs le logiciel réalise les étapes présentées dans le pseudo-code suivant :

```
// Initialisation
CALL getEtMatrix RETURNING etMatrix
CALL getRandomRunNumber RETURNING refNumber

// Boucle sur l'ensemble des ions peptidiques identifies dans un moins un run
// L'entité coPeptideIons correspond à un groupe d'ions peptidiques
// de même espèce mais appartenant à différents runs LC-MS/MS
FOREACH coPeptideIons IN coPeptideIonsList
  CALL coPeptideIons.getCoPeptideIonWithHighestScore RETURNING bestCoPeptideIon
  CALL bestCoPeptideIon.getElutionTimeInReference( refNumber ) RETURNING refET
  CALL coPeptideIons.getCoPeptideIonsWithoutET RETURNING coPeptideIonsWithoutET
  FOREACH coPeptideIon IN coPeptideIonsWithoutET
    // Prédiction du temps d'élution de l'ion peptidique
    coPeptideIon.computeElutionTime(refNumber, refET, etMatrix)
  ENDFOREACH
ENDFOREACH
```

La fonction computeElutionTime utilise une interpolation linéaire pour prédire la valeur du temps d'élution. Cela consiste à rechercher tout d'abord les deux « landmarks » les plus proches qui encadrent la valeur de temps dans le run de référence. Dans l'espace deltaET = f(ET) le programme calcule alors l'équation d'une droite qui passe par les deux points correspondants aux landmarks

sélectionnés. L'interpolation consiste alors à utiliser l'équation de cette droite pour calculer le temps d'élution du peptide à partir de la valeur précise de temps du bestCoPeptideIon.

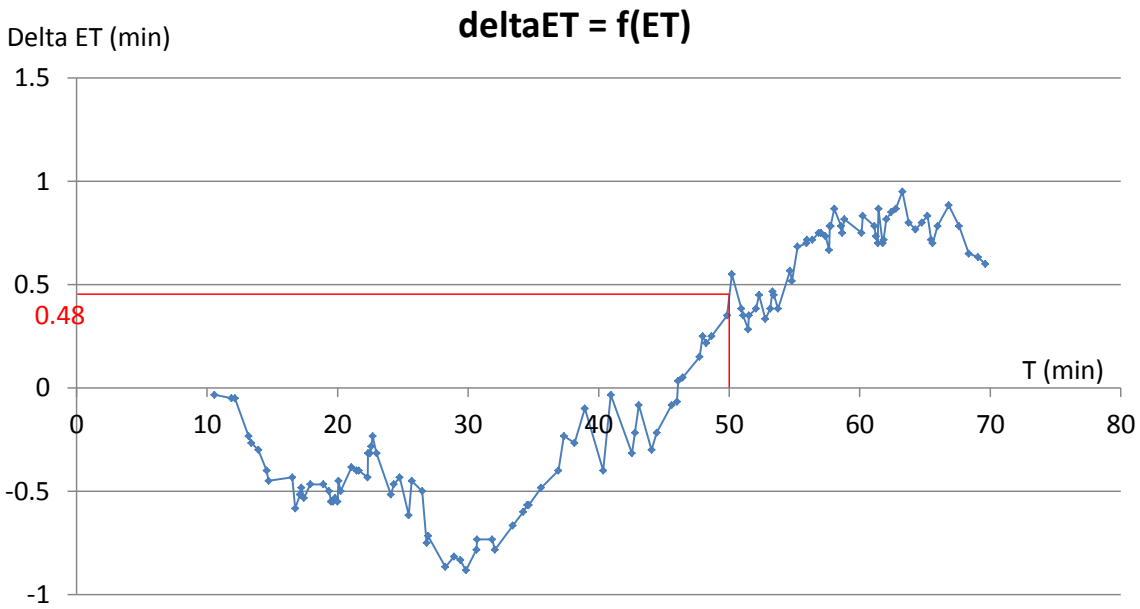


Figure 27 : exemple de résultat de l'alignement d'un run avec le run de référence. L'axe des abscisses représente le temps d'élution en minutes dans le run de référence alors que l'axe des ordonnées représente l'écart de temps entre le run aligné et le run de référence. Chaque point correspond à un des « landmarks » sélectionné par le logiciel. Par exemple, pour un peptide qui élue à 50 minutes dans le run de référence on peut ainsi déterminer par interpolation linéaire (représentée par les traits rouges) que le peptide devrait être trouvé environ 30 secondes plus tard (0,48 minutes) dans le run aligné.

4) EXTRACTION

L'étape précédente fournit une liste d'ions peptidiques dans laquelle chacun possède une valeur de m/z et de temps de rétention pour chacune des élutions chromatographique. La disponibilité d'une mesure à haute-résolution de la masse de l'ion permet de réaliser cette extraction sur une plage de m/z restreinte et donc de diminuer les conflits avec d'éventuels peptides de masse proche et qui élueraient au même moment.

Les données accumulées sont ensuite utilisées par le logiciel afin de procéder à l'extraction du signal dans des intervalles de masse et de temps proches de ceux déterminés expérimentalement. L'intervalle de temps peut être déduit soit de la valeur associée à l'identification du spectre MS/MS soit de la prédiction via l'interpolation linéaire. Dans le cas où le temps d'élution est prédit, le logiciel n'a pas la certitude de trouver du signal pour la valeur de temps dont il dispose. Le programme enclenche donc une recherche du signal pour trouver un point d'ancrage (cf figure 28).

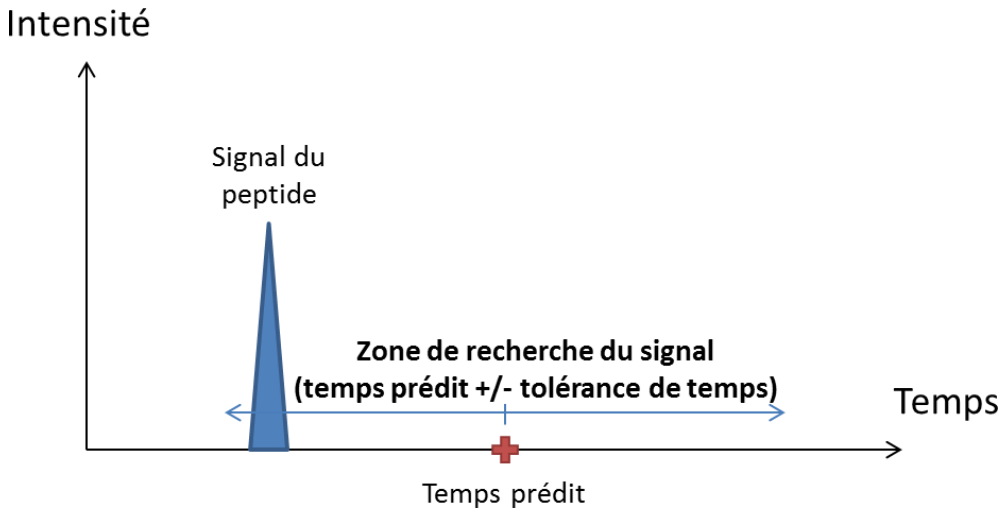


Figure 28 : recherche du signal peptidique à partir d'une masse précise et d'une valeur prédite du temps d'élution. A partir du rapport m/z et du temps prédit pour un certain ion peptidique, le programme réalise une extraction du signal brut sur une certain fenêtre de masse (entre 5 et 10ppm sur un LTQ-Orbitrap) et sur une certaine fenêtre de temps (+/- 90 secondes par défaut). Le signal le plus proche du temps prédit est sélectionné prioritairement.

Si le signal peptidique a été détecté en dehors de cette fenêtre de temps il ne sera pas pris en compte par le logiciel pouvant ainsi donner lieu à une valeur manquante. Les étapes d'alignement et de prédiction du temps d'élution sont donc primordiales pour assurer une bonne qualité d'extraction des données.

Une fois le point d'ancrage trouvé les étapes suivantes sont identiques à celles qui sont exécutées pour les ions où le temps est connu expérimentalement :

- recherche de la pente (gauche ou droite) présentant la plus grande dérivée positive,
- parcours des données suivant ce sens et recherche d'un maximum local (apex),
- extraction du signal autour de l'apex.

Cette dernière étape d'extraction comprend plusieurs conditions d'arrêt d'extraction :

- l'intensité est inférieure à une valeur seuil absolue et définie par l'utilisateur (0 par défaut)
- l'intensité est inférieure à une valeur seuil relative par rapport à l'apex et définie par l'utilisateur (1% par défaut)
- le massif isotopique extrait présente une allure qui n'est pas cohérente avec la masse du peptide.

5) REGROUPEMENT

Le regroupement des peptides en protéines est similaire à celui qui peut être réalisé à partir de résultats d'identification et suit ainsi le principe de parcimonie (cf partie I-4.5). Cependant, dans le contexte de la quantification des protéines, la connaissance des peptides dits protéotypiques (spécifique d'une séquence protéique) est une information qui peut être essentielle pour l'interprétation des données quantitatives. En effet, un peptide non spécifique, i.e. appartenant à plusieurs séquences protéiques, peut représenter une donnée quantitative biaisée correspondant à la moyenne pondérée des abondances des protéines donnant naissance à ces peptides.

Le logiciel étiquette chaque peptide quantifié du nombre de groupes de protéines auxquelles il se rapporte. Ce nombre est ensuite visible au niveau de l'interface graphique aidant ainsi à différencier les peptides sur la base de leur spécificité. Une macro citée dans le dernier paragraphe de la partie III-2.2 permet notamment de désélectionner les espèces non spécifiques.

6) CALCUL DES RATIOS

Pour chaque peptide quantifié, le programme calcule ensuite l'ensemble des ratios défini par l'utilisateur. Le calcul des ratios au niveau de la protéine correspond simplement à la moyenne géométrique des ratios peptidiques précédemment déterminés.

En parallèle, le programme calcule pour chaque protéine une valeur d'indice de l'abondance protéique (« Protein Abundance Index » ou PAI) :

Séquence	Charge	Aire condition 1	Aire condition 2	Ratio	Aire Max
SYTITGLQPGTDYK	**2+**	**2.70E+08**	**8.20E+08**	**3.04**	**8.20E+08**
LGVRPSQGGEAPR	**2+**	**2.40E+08**	**7.80E+08**	**3.25**	**7.80E+08**
LGVRPSQGGEAPR	3+	1.90E+08	6.60E+08	3.47	6.60E+08
SEPLIGR	**2+**	**1.30E+08**	**3.80E+08**	**2.92**	**3.80E+08**
DSMIWDCTCIGAGR + Oxidation (M)	2+	3.60E+07	1.20E+08	3.33	1.20E+08
VRVTPK	2+	6.30E+06	1.70E+07	2.70	1.70E+07
DSMIWDCTCIGAGR	2+	4.60E+06	8.80E+06	1.91	8.80E+06
PAI		2.13E+08	6.60E+08	3.09	

Figure 29 : exemple de calcul du PAI pour la fibronectine humaine à partir de ses peptides quantifiés. Pour chaque peptide une valeur maximale est déterminée pour l'ensemble de conditions comparées. Les peptides sont ensuite classés par ordre décroissant d'aire maximale et seuls les 3 peptides les plus intenses sont conservés. Si deux états de charge du même peptide (par exemple LGVRPSQGGEAPR) ont été quantifiés seul le plus intense est pris en compte. Le PAI correspond à la moyenne des intensités de ces 3 ions peptidiques sélectionnés (dernière ligne du tableau).

Le calcul du PAI permet d'obtenir des valeurs quantitatives protéiques plus fiables car les peptides les plus intenses présentent en général une meilleure reproductibilité de signal. De plus, lorsque des réplicats analytiques sont mis en œuvre il est possible de calculer des valeurs statistiques au niveau des protéines. On peut ainsi parfaitement réaliser un test de Student sur les logarithmes des valeurs de PAI (transformation nécessaire pour avoir des valeurs qui suivent une distribution normale). Il devient également possible de suivre le profil d'abondance d'une protéine donnée au travers de plusieurs échantillons. Le PAI représente donc une métrique d'abondance protéique.

Dans le contexte d'analyses cliniques ce type d'approche est d'une importance capitale pour la mise en évidence de biomarqueurs protéiques. Afin de démontrer la faisabilité de telles analyses sur des fluides biologiques nous avons d'ailleurs réalisé une étude protéomique différentielle sur du liquide céphalo-rachidien (LCR). Au-delà de ces considérations techniques liées à la quantification, ce type d'échantillon présente une autre contrainte qu'il est nécessaire de contourner pour mener à bien une analyse de qualité : la grande gamme dynamique des protéines de l'échantillon. Pour y remédier les ingénieurs et chercheurs du laboratoire ont évalué la méthode ProteoMiner™ (Bio-Rad), qui permet d'égaliser le contenu protéique de l'échantillon à l'aide de librairies de ligands peptidiques fixés sur billes. Dans la perspective d'études cliniques sur le LCR, il était cependant nécessaire d'évaluer la reproductibilité de cette technique pour une analyse quantitative. La publication qui suit

illustre l'utilisation du module de quantification sans marquage du logiciel MFPaQ pour la quantification des protéines du LCR. Les résultats obtenus montrent :

- l'efficacité de la quantification réalisée par MFPaQ sur des réplicats d'injection nanoLC-MS/MS. La figure 4A présente la distribution des CVs (coefficients de variation) des intensités extraites à partir des différents réplicats. On peut observer que la majorité des espèces ont un CV inférieur à 20%, avec une forte densité autour de 5-10%.
- qu'il est possible d'augmenter le nombre de protéines quantifiées en utilisant une approche basée sur une bibliothèque d'identifications MS/MS. Contrairement à la stratégie de type AMT (« Accurate Mass and Time tags », (Smith, Anderson et al. 2002)) où la bibliothèque est en général construite sur un grand nombre d'analyses nanoLC-MS/MS espacées dans le temps, l'approche que nous avons employé consiste à fractionner et analyser le contenu d'un échantillon de référence, puis à rechercher l'ensemble des peptides ainsi identifiés dans les fichiers que l'on souhaite comparer. Ainsi après le fractionnement d'un échantillon de LCR sur gel SDS page 1D et l'analyse nanoLC-MS/MS des 20 bandes de gel, une banque de données contenant l'ensemble des résultats d'identification a été créée. Comme le montre la figure 6 de l'article, l'utilisation de cette dernière pour la quantification d'échantillons de LCR non fractionnés augmente considérablement le nombre de protéines quantifiées tout en conservant un niveau de reproductibilité raisonnable.

PUBLICATION

« In-depth exploration of cerebrospinal fluid by combining peptide ligand library treatment and label-free protein quantification. »

Mouton-Barbosa E, Roux-Dalvai F, Bouyssié D, Berger F, Schmidt E, Righetti PG, Guerrier L, Boschetti E, Burlet-Schiltz O, Monsarrat B, Gonzalez de Peredo A

Mol Cell Proteomics. **2010** May;9(5):1006-21.

In-depth Exploration of Cerebrospinal Fluid by Combining Peptide Ligand Library Treatment and Label-free Protein Quantification*[S]

Emmanuelle Mouton-Barbosa,[a,b,c] Florence Roux-Dalvai,[a,b,c] David Bouyssié,[a,b] François Berger,[d] Eric Schmidt,[e] Pier Giorgio Righetti,[f,g] Luc Guerrier,[h] Egisto Boschetti,[h] Odile Burlet-Schiltz,[a,b] Bernard Monsarrat,[a,b,i] and Anne Gonzalez de Peredo[a,b,j]

Cerebrospinal fluid (CSF) is the biological fluid in closest contact with the brain and thus contains proteins of neural cell origin. Hence, CSF is a biochemical window into the brain and is particularly attractive for the search for biomarkers of neurological diseases. However, as in the case of other biological fluids, one of the main analytical challenges in proteomic characterization of the CSF is the very wide concentration range of proteins, largely exceeding the dynamic range of current analytical approaches. Here, we used the combinatorial peptide ligand library technology (ProteoMiner) to reduce the dynamic range of protein concentration in CSF and unmask previously undetected proteins by nano-LC-MS/MS analysis on an LTQ-Orbitrap mass spectrometer. This method was first applied on a large pool of CSF from different sources with the aim to better characterize the protein content of this fluid, especially for the low abundance components. We were able to identify 1212 proteins in CSF, and among these, 745 were only detected after peptide library treatment. However, additional difficulties for clinical studies of CSF are the low protein concentration of this fluid and the low volumes typically obtained after lumbar puncture, precluding the conventional use of ProteoMiner with large volume columns for treatment of patient samples. The method has thus been optimized to be compatible with low volume samples. We could show that the treatment is still efficient with this miniaturized protocol and that the dynamic range of protein concentration is actually reduced even with small amounts of beads, leading to an increase of more than 100% of the number of identified proteins in one LC-MS/MS run. Moreover, using a dedicated bioinformatics analytical work flow, we found that the method is reproducible and applicable for label-free quantification of series of samples processed in parallel. ***Molecular & Cellular Proteomics 9:1006–1021, 2010.***

The identification of biological markers heralding a pathological condition represents a major challenge in clinical proteomics. In this context, body fluids are a particularly interesting material source because their collection is minimally invasive compared with tissue biopsies. Indeed, in the discovery phase, the selection of a fluid in close proximity to a diseased organ may increase the probability of finding a biomarker originating from a pathological tissue. In that respect, cerebrospinal fluid (CSF),[1] which is in closest contact with the brain, represents a useful reservoir of potential clinically relevant biomarkers for neurological diseases. Although acquisition of CSF via lumbar puncture is an invasive procedure, it is less invasive and more readily obtainable than a brain biopsy.

CSF has several functions, including buoyancy, protection of the brain against pressure gradients, transport of active biological substances for normal maintenance of the brain, and excretion of toxic and waste substances (1, 2). It surrounds the exterior of the central nervous system, filling also the four large ventricular cavities inside the brain, the spinal canal, and subarachnoid spaces. CSF is produced mainly in the choroid plexus, a structure present in the four ventricles, consisting of a single layer of epithelial cells surrounding a core of connective tissue and blood capillaries. These epithelial cells have tight junctions on the side facing the ventricle that prevent the majority of blood substances from crossing the cell layer into the CSF and form what is known as the blood-CSF barrier. They produce CSF through a secretory

From the [a]Institut de Pharmacologie et de Biologie Structurale (IPBS), CNRS, 205 route de Narbonne, 31077 Toulouse, France, [b]Université de Toulouse, Université Paul Sabatier, 31077 Toulouse, France, [d]INSERM U 836, Grenoble Institut des Neurosciences, Université Joseph Fourier, 38041 Grenoble, France, [e]Department of Neurosurgery, University Hospital Purpan, 31059 Toulouse, France, [f]Politecnico di Milano, Milan 20133, Italy, and [h]Bio-Rad, Commissariat à l'Energie Atomique-Saclay, 91191 Gif-sur-Yvette, France

Received, October 29, 2009, and in revised form, January 8, 2010

Published, MCP Papers in Press, January 21, 2010, DOI 10.1074/mcp.M900513-MCP200

[1] The abbreviations used are: CSF, cerebrospinal fluid; ECF, extracellular fluid; HOS, hydro-organic solution; PAI, protein abundance index; FDR, false discovery rate; XIC, extracted ion chromatogram; RT, retention time; 2D, two-dimensional; 1D, one-dimensional; nano-LC, liquid nanochromatography; LTQ, linear trap quadrupole; IPI, International Protein Index; MFPaQ, Mascot File Parsing and Quantification; SLITRK, SLIT and NTRK-like proteins; NTRK, Neurotrophic Tyrosine Kinase Receptor.

process based on active ion transport mechanisms that allow maintenance of a constant ionic composition in the CSF and thus create a stable ionic environment for the brain independent of plasma ion concentration that can fluctuate significantly (3). The protein content of the CSF produced in the choroid plexus originates mainly from blood proteins, which slowly cross the epithelial layer by a diffusion mechanism dependent on molecular size (1, 4). Because of the presence of this barrier, however, the final protein concentration in CSF is around 200 times lower than in plasma. Some CSF proteins, on the other hand, are known to be synthesized and actively secreted by the epithelial cells of the choroid plexus, such as transthyretin or cystatin C (5).

In addition, a small amount of the CSF is not produced by the choroid plexus but originates from the extracellular fluid (ECF) that fills the extracellular space of the brain. This ECF is connected with the CFS via the perivascular space also known as Virchow-Robin space. Indeed, CSF also extends into the sulci and the depth of the cerebral cortex along the blood vessels in these perivascular spaces in which small molecules diffuse freely. ECF also flows into the ventricles by crossing the ependyma cell barrier between brain and CSF. ECF carries proteins derived directly from the brain that have been estimated to represent about 20% of the final CSF protein content (6). Therefore, there is a continuous movement of metabolites from deep parenchyma to cortical subarachnoid space and ventricular system. In that respect, CSF is also a biochemical window into the brain and constitutes an interesting source of potential neurological biomarkers.

Protein analysis in CSF has been actively performed during the past years to discover species relevant to a disease. Many validated clinical assays, each directed toward a particular protein and mostly based on antibody detection methods, have been developed to investigate protein candidates as potential markers of several pathologies. These focused measurements (for a review, see Ref. 1) allowed a precise quantification of the selected proteins in CSF and have shown that the dynamic range of protein concentrations spans at least 8 orders of magnitude between the most abundant protein (albumin at 130–350 mg/liter, representing alone around 45% of CSF protein content) and the lowest detected proteins in such assays in the ng/liter range. This very large dynamic range represents, as for other body fluids, a major analytical challenge and has largely hampered comprehensive studies aiming at global CSF proteomic characterization. The first studies were based on 2D gels, which provided a useful tool to establish CSF maps, analyze protein modifications, and perform differential profiling (7–10). However, the relatively low dynamic range of the technique has usually limited the total number of detectable unique proteins to around 40 abundant species. Liquid nanochromatography (nano-LC)-MS-based technologies allowed a significant increase of this number during the past few years, and different studies reported several hundreds of unique proteins confidently identified in CSF (11–14). To reach this number, extensive sample fractionation was usually applied at the protein or peptide level to lower the concentration dynamic range of each individual fraction as well as to increase the MS/MS sequencing time and analytical coverage of the sample. Most of these studies, however, remained largely qualitative or restricted to the quantitative differential analysis of a small number of isotopically labeled samples. Only recently, the development of label-free approaches, based on the direct comparison of MS peptide signals across different nano-LC-MS runs using appropriate bioinformatics tools, has opened the way to the quantitative profiling of large series of patient samples for clinical studies with mass spectrometry techniques (15). Thanks to the high resolution of recent mass spectrometers, such strategies seem now more accurate and promising for fluids proteomic profiling. However, a compromise has yet to be found between a labor-intensive prefractionation of the samples and the number of patients included in the study, which should be ideally high for statistical significance.

Here, we performed a proteomic characterization of CSF using the combinatorial peptide ligand library technology, which has been described recently as an efficient approach to decrease the concentration dynamic range of complex protein mixtures (16–19). This method is based on the treatment of the sample with combinatorial libraries of peptide ligands bound to porous beads. Each bead contains a unique hexapeptide ligand distributed throughout its porous structure, and any protein in the starting material can theoretically interact with one or a few beads among the wide diversity ($10–20^6$) of ligand beads from the library. Once the most abundant protein species have saturated their binding sites, the remaining molecules are washed away in the flow-through, whereas minor protein species are progressively enriched on their corresponding beads. This strategy has been efficiently applied to capture a very large population of previously undetected proteins in several types of samples, such as urine (20), serum (21), and platelets (22), and we used it recently for the extensive characterization of the red blood cell proteome (23). Using a large pool of CSF from different sources, this method, associated with fractionation and nano-LC-MS/MS analysis, allowed here the detection of more than a thousand proteins in CSF. Moreover, we evaluated several technical features relevant for its use in clinical studies, such as the conservation of relative quantitative information among parallel sample treatments, and the applicability on low CSF volumes typically obtained from lumbar puncture. We found that the miniaturized protocol is efficient and reproducible and allows a significant increase of the number of quantified proteins in the peptide library-treated, but unfractionated sample. It could thus offer an interesting alternative for the rapid and in-depth profiling of large series of CSF patient samples, avoiding the use of extensive prefractionation operations. It provides one of the first clinical compatible proteomics strat-

egies targeting the deep proteome for CSF biomarker discovery, which should be a crucial biomedical issue for the exploration and management of brain diseases.

EXPERIMENTAL PROCEDURES

Materials—The solid-phase combinatorial peptide library called ProteoMiner™ (NH_2-Library) and its carboxylated version (COOH-Library) are both from Bio-Rad. The former was purchased as a commercial product; the second is not commercially available and was a gift of the company. Complete protease inhibitor mixture tablets were from Roche Diagnostics. Sequencing grade trypsin was from Promega (Madison, WI). *N*-Ethylmaleimide, urea, thiourea, CHAPS, isopropanol, acetonitrile, trifluoroacetic acid, and sodium dodecyl sulfate were all from Sigma-Aldrich. All other chemicals were also from Aldrich and were of analytical grade. C_{18} precolumns and analytical columns were from Dionex (Amsterdam, The Netherlands). Equipment and reagents for mono- and bidimensional electrophoresis were both from Bio-Rad.

CSF Collection—Human CSF used in this study was a pool obtained from various samples collected in two different hospital centers. Some patients were from the University Hospital of Grenoble, France where collection followed the French regulation and was approved at the national level (molecular neurosurgery bank) with approval number AC-2007-23, and others were from the University Hospital Purpan, Toulouse, France where collection was approved by the local ethical review board (Comité de Protection des Personnes Sud Ouest II) and declared at the national level (Direction Générale de la Recherche et de l'Innovation) under the reference DC-2008-463.

For the large CSF study (treatment on 1-ml ligand bead columns), 1290 ml of CSF were gathered from a pool of 250 patients (Pool 1). Most of the patients underwent a lumbar puncture in the context of headache and/or suspicion of a neurological disease not confirmed by magnetic resonance imaging and CSF study ($n = 199$). Volumes of samples obtained from these patients were 0.2 ml ($n = 100$) or 0.5 ml ($n = 99$). Another group of patients suffered from hydrocephalus, *i.e.* an excess of CSF ($n = 49$), and a tap test was performed to withdraw CSF via lumbar puncture. Volumes of the samples obtained in this way were 0.5 ($n = 1$), 1 ($n = 8$), 3 ($n = 30$), 6 ($n = 5$), and 12 ml ($n = 5$). In addition, large volumes were obtained for samples collected using an overnight CSF drainage protocol for two patients suffering from hydrocephalus (108, 150, 425, and 350 ml). All samples came from the CSF discarded after biochemical evaluation. For the assays on low volumes of beads, another pool of CSF samples was used (Pool 2) from 40 patients suffering from headache and investigated by lumbar puncture for a potential neurological disease not confirmed by magnetic resonance imaging and CSF study. The volume of samples collected in this way was 0.5 ml per individual. In every case, we checked the absence of cellular abnormalities in CSF. We also checked the absence of an increased total proteomic content to eliminate pathologies with inflammatory reaction as well as blood-brain barrier abnormalities. Infectious diseases as well as tumor formations and meningitis were eliminated. All samples were taken with the informed consent of the patients and were frozen immediately after clinical laboratory analysis.

Treatment of CSF Proteins with Hexapeptide Ligand Libraries—About 1290 ml of frozen pooled human CSF (CSF Pool 1; 0.42 mg/ml protein concentration) was thawed at 4 °C, added with two tablets of protease inhibitor mixture, and dialyzed against 50 mM ammonium bicarbonate overnight at +4 °C. The dialyzed material was then centrifuged at 5000 × *g* at 4 °C for 30 min to obtain a clear protein solution and lyophilized. CSF proteins were solubilized in 70 ml of physiological PBS and then loaded on a column containing 1 ml (6.6-mm inner diameter × 32 mm in length) of NH_2-Library at a flow rate of 0.25 ml/min. The column effluent was directly injected in a second column of the same dimensions packed with COOH-Library. The columns connected in series were then washed with PBS until UV base line at 280 nm of the effluent of the second column. After the wash, each individual column was independently subjected to three distinct elutions of captured proteins using, respectively, TUC solution (2 M thiourea, 7 M urea, 2% CHAPS), UCA solution (9 M urea, citric acid up to pH 3.3) and a hydro-organic solution (HOS) composed of 6% (v/v) acetonitrile, 12% (v/v) isopropanol, 10% (v/v) ammonia at 20%, and 72% (v/v) water. The six eluates were immediately neutralized (second and third eluates), dialyzed against 20 mM ammonium bicarbonate (cutoff of dialysis membrane was 1000 Da), and then lyophilized. Dry protein samples were then stored at −20 °C waiting for further analysis. Small aliquots were taken for protein assay. The amount of protein obtained from each fraction was: 2352, 7342, and 440 μg from NH_2-Library for TUC, acidic urea, and hydro-organic eluates, respectively, and 1422, 3572, and 88 μg from COOH-Library for TUC, acidic urea, and hydro-organic eluates, respectively.

1D SDS-PAGE Fractionation and Nano-LC-MS/MS Analysis of CSF Proteins—One hundred and fifty micrograms of each elution fraction from the two library columns as well as 150 μg of the initial non-treated human CSF and 150 μg of the flow-through were diluted in Laemmli buffer and boiled for 5 min before being separated on a 12% acrylamide SDS-PAGE gel. Proteins were visualized by Coomassie Blue staining. Each lane was cut into 20 homogenous slices that were washed in 100 mM ammonium bicarbonate for 15 min at 37 °C followed by a second wash in 100 mM ammonium bicarbonate, acetonitrile (1:1) for 15 min at 37 °C. Reduction and alkylation of cysteine residues were performed by mixing the gel pieces in 10 mM DTT for 35 min at 56 °C followed by 55 mM iodoacetamide for 30 min at room temperature in the dark. An additional cycle of washes in ammonium bicarbonate and ammonium bicarbonate/acetonitrile was then performed. Proteins were digested by incubating each gel slice with 0.6 μg of modified sequencing grade trypsin in 50 mM ammonium bicarbonate overnight at 37 °C. The resulting peptides were extracted from the gel by three steps: a first incubation in 50 mM ammonium bicarbonate for 15 min at 37 °C and two incubations in 10% formic acid, acetonitrile (1:1) for 15 min at 37 °C. The three collected extractions were pooled with the initial digestion supernatant, dried in a Speed-Vac, and resuspended with 14 μl of 5% acetonitrile, 0.05% trifluoroacetic acid.

The peptides mixtures were analyzed by nano-LC-MS/MS using an Ultimate3000 system (Dionex) coupled to an LTQ-Orbitrap mass spectrometer (Thermo Fisher Scientific, Bremen, Germany). Five microliters of each sample were loaded on a C_{18} precolumn (300-μm inner diameter × 5 mm; Dionex) at 20 μl/min in 5% acetonitrile, 0.05% trifluoroacetic acid. After 5 min of desalting, the precolumn was switched on line with the analytical C_{18} column (75-μm inner diameter × 15 cm; PepMap C_{18}, Dionex) equilibrated in 95% solvent A (5% acetonitrile, 0.2% formic acid) and 5% solvent B (80% acetonitrile, 0.2% formic acid). Peptides were eluted using a 5–50% gradient of solvent B during 80 min at a 300 nl/min flow rate. The LTQ-Orbitrap was operated in data-dependent acquisition mode with the Xcalibur software. Survey scan MS spectra were acquired in the Orbitrap on the 300–2000 *m/z* range with the resolution set to a value of 60,000. The five most intense ions per survey scan were selected for CID fragmentation, and the resulting fragments were analyzed in the linear trap (LTQ). Dynamic exclusion was used within 60 s to prevent repetitive selection of the same peptide.

Quantitative Measurements of Proteins Spiked in Serum after Peptide Library Treatment—Human serum samples (5 ml) were spiked with four proteins (bovine β-lactoglobulin, bovine β-casein, bovine κ-casein, and rabbit phosphorylase *b*, all purchased from Sigma-Aldrich) at four different final concentrations (20 nM, 200 nM, 2 μM, and 20 μM). This was done in triplicate for each spiked protein amount. Each of the 12 samples was then loaded on a spin column containing

0.5 ml of ProteoMiner beads and incubated with rotation for 2 h. The beads were then washed three times with PBS. After the wash, each individual column was independently subjected to elution with 1.25 ml of 2× Laemmli sample buffer (80 mM Tris-HCl, pH 6.8, 20% glycerol, 4% SDS, 50 mM DTT). Aliquots of 50 μl of each sample were loaded on a gel for 1D SDS-PAGE and migrated through the stacking gel until the top of the separating gel, and proteins were concentrated into one gel band. Tryptic peptides were prepared and analyzed in one nano-LC-MS/MS run as described above. To quantify the spiked proteins after peptide library treatment, the extracted ion chromatogram (XIC) signal of some well characterized tryptic peptide ions from these proteins was manually extracted from the MS survey of nano-LC-MS/MS raw files using the Xcalibur software. XIC areas were integrated in Xcalibur under the QualBrowser interface using the ICIS algorithm. Mean values and S.D. were calculated for triplicate measurements.

Treatment of CSF Samples on Small Volumes of Peptide Library Beads—A pooled CSF sample of 20 ml (CSF Pool 2) was divided into 2-ml aliquots that were added to either 5 (four replicates) or 2 μl (four replicates) of PBS-equilibrated ProteoMiner beads in 2-ml centrifugation tubes. The CSF-bead suspensions were gently stirred at 4 °C for 5 h, and then the beads were collected by centrifugation and washed twice with PBS. The proteins captured by the beads were eluted by addition of 15 μl of Laemmli sample buffer for electrophoresis and boiling 2 min at 95 °C. Proteins were loaded on a gel for 1D SDS-PAGE, and the whole protein mixture was concentrated into a unique gel band. Tryptic peptides were prepared and analyzed in one nano-LC-MS/MS run as described above. For comparison, an aliquot of crude CSF (10 μl; Pool 2) was prepared, digested, and analyzed in the same way.

Database Search and Data Analysis—The Mascot Daemon software (version 2.2.0, Matrix Science, London, UK) was used to perform database searches in batch mode with all the raw files acquired on each sample. To automatically extract peak lists from Xcalibur raw files, the Extract_msn.exe macro provided with Xcalibur (version 2.0 SR2, Thermo Fisher Scientific) was used through the Mascot Daemon interface. The following parameters were set for creation of the peak lists: parent ions in the mass range 400–4500, no grouping of MS/MS scans, and threshold at 1000. A peak list was created for each analyzed fraction (*i.e.* gel slice), and individual Mascot searches were performed for each fraction. Data were searched against all entries in the IPI human v3.61 protein database (82,631 sequences). Carbamidomethylation of cysteines was set as a fixed modification, and oxidation of methionine and protein N-terminal acetylation were set as a variable modifications for all Mascot searches. Specificity of trypsin digestion was set for cleavage after Lys or Arg except before Pro, and one missed trypsin cleavage site was allowed. The mass tolerances in MS and MS/MS were set to 5 ppm and 0.6 Da, respectively, and the instrument setting was specified as "ESI-Trap." Mascot results were parsed with the in-house developed software Mascot File Parsing and Quantification (MFPaQ) version 4.0 (24), and protein hits were automatically validated if they satisfied one of the following criteria: identification with at least one top ranking peptide of a minimal length of 5 amino acids and with a Mascot score higher than the identity threshold at $p = 0.001$ (99.9% probability) or at least two top ranking peptides each of a minimal length of 5 amino acids and with a Mascot score higher than the identity threshold at $p = 0.05$ (95% probability). All the annotated MS/MS spectra corresponding to proteins identified with a unique peptide are shown in supplemental Data 5 and 6. To calculate the false discovery rate (FDR), the search was performed using the "decoy" option in Mascot, and MFPaQ used the same criteria to validate decoy and target hits. The FDR was calculated at the protein level (FDR = number of validated decoy hits/(number of validated target hits + number of validated decoy hits) × 100), and using the specified validation criteria, it ranked between 0 and 0.8% for all the samples analyzed with an average value of 0.4%. When several proteins matched exactly the same set of peptides, only one member of the protein group was reported in final protein lists in supplemental Data 2 and 3 for more clarity (the one returned by Mascot in the protein summary list), but detailed protein groups are displayed in an additional sheet. Moreover, in each Mascot result file, the MFPaQ software detected highly homologous Mascot protein hits, *i.e.* proteins identified with top ranking MS/MS queries also assigned to another protein hit of higher score. These homologous protein hits were validated and included in the final list only if they were additionally assigned a specific top ranking peptide with Mascot score higher than the identity threshold at $p = 0.001$. In the case of sample fractionation by 1D SDS-PAGE, MFPaQ was used to create a unique non-redundant protein list from the identification results of each fraction (*i.e.* gel slice). Clustering of proteins was performed based on peptide sharing by grouping together all protein sequences matching the same set of peptides (only top ranking peptides with a Mascot score higher than the identity threshold at $p = 0.05$ were considered). To merge or compare lists of protein groups from different samples, clustering was performed based on accession number sharing (clusters of protein groups are created if they have one common member).

Label-free Quantification—Quantification of proteins was performed using the label-free module implemented in the MFPaQ v4.0.0 software (SourceForge). For each sample, the software uses the validated identification results (using the validation criteria described above) and XICs of the identified peptides in the corresponding raw nano-LC-MS files based on their experimentally measured retention time (RT) and monoisotopic *m/z* values. Only top ranking peptides with a Mascot score higher than the identity threshold at $p = 0.05$ were selected for quantification. If some peptide ions were sequenced by MS/MS and validated only in some of the samples to be compared, their XIC signal was extracted in the nano-LC-MS raw file of the other samples using a predicted RT value calculated as follows. The software selects all the peptide ions that are identified by MS/MS in all the LC-MS/MS runs that are to be aligned. It extracts the XIC signal of each of these peptide ions in each run (within a time interval around the MS/MS sequencing time), defines their retention time as the apex of their elution peak, and stores all these time values in a calibration matrix. Then, for all peptide ions that were identified by MS/MS in only some of the runs, this matrix can be used to predict their retention time in the other runs by a linear interpolation method. Quantification of peptide ions was performed based on calculated XIC areas values. For pairwise comparison, the software computed peptide ratios (defined as the average of the ratios from peptide ions of different charge state) and protein ratios (defined as the average of the ratios of the peptides assigned to the protein). These mean values were calculated from an arithmetic averaging of the ratios, after applying a transformation to the asymmetric distribution of ratio values, to make it symmetric and centered around 0: $x - 1$ for ratios higher than 1 and $1 - (1/x)$ for ratios smaller than 1. The arithmetic average was then computed, and the final mean ratio was obtained by applying the reverse transformation to this value. Systematic experimental errors were corrected by dividing each protein ratio by the median of all the ratios. When multiple replicate analyses were performed, the coefficient of variation of peptide ion XIC areas was calculated over the different replicates after normalization of the area value for each nano-LC-MS run (normalization was based on the total sum of all the XIC areas extracted by the software in the run). To compare abundances of different proteins or to represent the abundance profile of one protein in different samples, a protein abundance index was calculated, defined as the average of XIC area values for three intense reference tryptic peptides identified for this protein (the three peptides exhibiting the highest intensities across the different replicates are

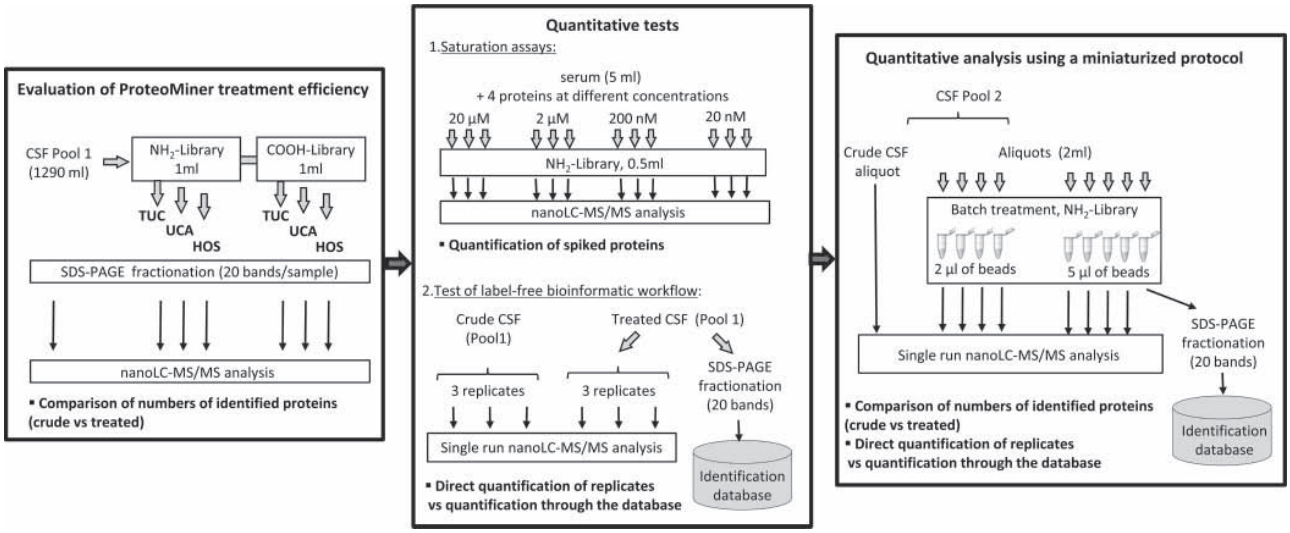


FIG. 1. **Experimental work flow of study.** First, an extensive proteomics analysis of CSF was performed using a large volume of CSF (Pool 1) and two different peptide ligand library columns of 1 ml to provide a high level of ligand diversity and optimize the number of identified proteins. We assessed the efficiency of the treatment by comparing the number of identified proteins before and after CSF treatment. Then, quantitative tests were performed to (i) evaluate whether quantification of proteins was compatible with the method (exogenous proteins were spiked into serum at various concentrations and manually quantified after peptide library treatment) and (ii) set up and evaluate the accuracy of a bioinformatics label-free quantitative work flow in MFPaQ. Finally, the peptide library treatment was scaled down, and the efficiency of the miniaturized protocol was checked by comparing the number of identified proteins before and after CSF treatment, whereas its reproducibility was assessed using the label-free quantitative work flow.

selected as reference peptides and used to compute the protein abundance index (PAI) of this protein in each replicate).

To optimize the number of quantified proteins in replicate nano-LC-MS/MS analyses, quantification was also performed with the use of previously built identification databases, containing *m/z* and RT values associated with peptide sequences, which were subsequently used to extract MS signals in individual runs. For each quantitative experiment, a small database was thus created from the same sample that was to be analyzed and quantified using the label-free module of MFPaQ. This module was first tested on replicate nano-LC-MS runs of the first eluate from the NH_2-Library (obtained from treatment of CSF Pool 1), and a database was created from an SDS-PAGE shotgun analysis of this sample (150 μg fractionated into 20 gel slices). To this aim, the fractions were analyzed by nano-LC-MS/MS as described above, the sequenced and validated peptide ions were stored in MFPaQ, and this database was used to quantify proteins identified in triplicate nano-LC-MS/MS analysis of 4 μg of either crude or treated CSF (TUC eluate from NH_2-Library). Another database was built in the same way to check the reproducibility of the miniaturized protocol and quantify proteins detected in nano-LC-MS runs of replicate samples from CSF Pool 2 treated on small volumes of beads. This was done by shotgun analysis of a 2-ml CSF aliquot (Pool 2) treated on 5 μl of beads and fractionated into 20 gel slices.

RESULTS

Treatment of Large Volume of Pooled CSF Samples with Peptide Libraries—To assess the efficiency of ProteoMiner treatment on CSF, an extensive proteomics study was performed using an experimental scheme previously described for other biological fluids (22, 23) (Fig. 1). The sample was loaded on a sequence of two columns containing two different peptide ligand libraries, one composed of an N-terminal collection of hexapeptides synthesized by combinatorial chemistry (NH_2-Library), and the other one containing the same peptides in which the N terminus has been modified to a carboxylate group (COOH-Library). The use of this second library has been shown before to be beneficial for capturing different species present in the sample compared with the NH_2-Library and for finally increasing the number of identified proteins (22, 23). Moreover, to optimize the number of protein species captured by both combinatorial libraries, a high diversity of hexapeptide ligands is typically used to treat the samples. Thus, a volume of 1 ml of ligand beads was used here in each column as described in previous reports. As a large overloading of total bead capacity has to be performed to saturate abundant proteins and obtain a sufficient enrichment of low abundance species, large protein amounts are generally necessary when using this classical protocol. This was achieved in the case of CSF by pooling samples from different sources to yield a total amount of 770 mg of CSF proteins. The captured proteins were then collected by eluting each column sequentially with TUC, UCA, and HOS buffers. The six resulting eluates were analyzed by SDS-PAGE as were the non-treated CSF and the flow-through from the columns (150 μg from each sample). As shown in Fig. 2*A*, a very intense band of albumin is observed in the latter two samples. Very different patterns were obtained for the column eluates, which showed a large decrease of the albumin band and the apparition of many new protein bands. This was confirmed by analysis of the samples using two-

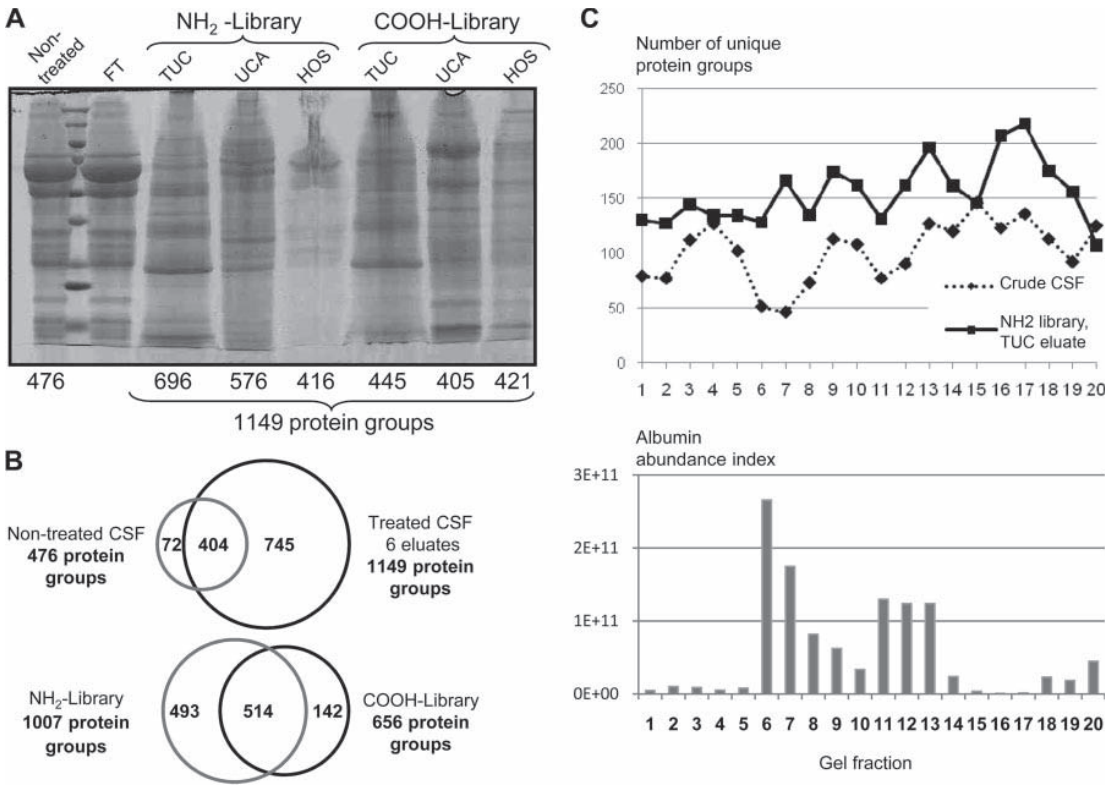


FIG. 2. **CSF treatment on two peptide ligand libraries and nano-LC-MS/MS analysis.** A large pool of CSF samples was treated with 1 ml of beads from the NH_2-Library and COOH-Library, and both columns were sequentially eluted with TUC, UCA, and HOS buffers. *A*, analysis of non-treated CSF, flow-through (*FT*), and elution fractions by monodimensional SDS-PAGE. The number of non-redundant protein groups identified by nano-LC-MS/MS in each gel lane is indicated. *B*, Venn diagrams giving the extent of overlap of protein groups found by LC-MS/MS analysis between non-treated CSF and all the eluates from both libraries and between all the eluates from the NH_2-Library and all the eluates from the COOH-Library. *C*, number of unique protein groups identified in each 1D gel fraction over the lanes of either crude CSF (*dotted line*) or the first eluate from the NH_2-Library (*solid line*) and the albumin abundance index (calculated as the mean XIC area value for the three most intense peptides of the protein) in each 1D gel fraction over the lane of crude CSF.

dimensional electrophoresis. Although the 2D pattern of initial untreated CSF shows a relatively low number of protein spots (mainly albumin as well as light and heavy chains of immunoglobulins), a high number of new protein spots emerge in the patterns from fractions of library-treated samples (supplemental Data 1).

The lanes of the 1D gel corresponding to the six elution fractions and to the crude CSF were then each cut into 20 gel slices, digested with trypsin, and analyzed by nano-LC-MS/MS on an LTQ-Orbitrap mass spectrometer. Proteins identified by Mascot were validated using stringent criteria to yield a false discovery rate around 0.4% at the protein level, and concatenated lists of unique protein groups were generated using the MFPaQ software for each gel lane. Fig. 2*A* shows the numbers of protein groups identified for the six eluates and for the untreated CSF. There is a significant effect of the peptide library treatment as the number of unique species identified in the richest eluate (696 unique protein groups in the TUC elution from the NH_2-Library) is increased by 46% compared with the crude material (476 protein groups in non-treated CSF). Moreover, the analysis of the six fractions from the two libraries yielded a global number of 1149 unique protein groups identified in CSF after ProteoMiner treatment. More than 80% of the proteins identified in crude CSF were also found in the bead eluates where 745 new proteins could be additionally identified (Fig. 2*B*). Most of the 1149 protein groups identified after ProteoMiner treatment were found in both peptide ligand libraries with a major contribution of the NH_2-Library. However, the analysis of the fractions from the COOH-Library allowed detection of about 12% additional protein species (Fig. 2*B*). Detailed concatenated lists corresponding to proteins identified in either untreated CSF or peptide library-treated CSF as well as a global list containing all proteins identified in this study (non-treated CSF and all eluate fractions), *i.e.* 1212 non-redundant protein groups, are given in supplemental Data 2. These results indicate that treatment of the sample with the two peptide libraries significantly enhances the number of proteins identified by nano-LC-MS/MS. Of course, we compare the results obtained on 150 μg of non-treated CSF with those obtained on

TABLE I
Proteins associated with gene ontology term neurogenesis

The columns display the best Mascot protein score (if the protein was identified in various 1D gel fractions from either crude CSF or ProteoMiner eluates), total number of unique peptides assigned to the protein in the 1D gel fraction where it showed the best Mascot score, and total number of MS/MS queries over all the 1D gel fractions analyzed. Proteins displayed in bold were identified only in the ProteoMiner eluates and not in crude CSF. NSF, *N*-ethylmaleimide-sensitive factor; NRCAM, neuronal cell adhesion molecule.

ID	Description	Best score	Peptides	Total MS/MS
IPI00021842	Apolipoprotein E	2269	38	3317
IPI00291262	Clusterin	1499	26	2366
IPI00006114	Pigment epithelium-derived factor	1470	25	880
IPI00032220	Angiotensinogen	1402	16	877
IPI00022246	Azurocidin	864	14	361
IPI00220642	14-3-3 protein γ	882	19	219
IPI00216319	**14-3-3 protein η**	**530**	**19**	**106**
IPI00333778	NRCAM protein	549	26	85
IPI00374563	Agrin	622	28	78
IPI00299059	Neural cell adhesion molecule L1-like protein	565	27	65
IPI00027166	Metalloproteinase inhibitor 2	214	10	41
IPI00024966	Contactin-2	355	21	39
IPI00013976	**Laminin subunit β-1**	**313**	**12**	**39**
IPI00442299	**Neurexin-1α**	**200**	**10**	**33**
IPI00160552	**Tenascin-R**	**158**	**6**	**26**
IPI00027463	Protein S100-A6	225	3	25
IPI00029693	**Neuropilin-2**	**122**	**8**	**25**
IPI00179330	Ubiquitin and ribosomal protein S27a	105	6	18
IPI00384225	**Meteorin**	**245**	**6**	**18**
IPI00387168	**Proprotein convertase subtilisin/kexin type 9**	**290**	**7**	**16**
IPI00012283	**Semaphorin-3B**	**243**	**7**	**13**
IPI00102543	**SLIT and NTRK-like protein 1**	**57**	**8**	**12**
IPI00165438	**Muscle type neuropilin 1**	**73**	**2**	**10**
IPI00299399	**Protein S100-B**	**165**	**4**	**7**
IPI00179589	Myotrophin	80	2	6
IPI00333140	**Delta and Notch-like epidermal growth factor-related receptor**	**101**	**2**	**6**
IPI00217507	**Neurofilament medium polypeptide**	**121**	**5**	**5**
IPI00034558	**Neurexin-3β**	**36**	**4**	**4**
IPI00305975	**Spondin-2**	**74**	**4**	**4**
IPI00002412	**Palmitoyl-protein thioesterase 1**	**60**	**1**	**2**
IPI00217146	**SLIT and NTRK-like protein 4**	**77**	**1**	**2**
IPI00009253	**α-Soluble NSF attachment protein**	**65**	**1**	**1**

150 μg of the elution fractions from the library columns, but it must be noted that a much larger amount of starting CSF material was necessary to perform the treatment. Indeed, using the peptide ligand library strategy, the more material that is loaded, the more low abundance proteins get progressively enriched, whereas major species are not. Conversely, a straightforward fractionation of the crude sample by SDS-PAGE is limited by the capacity of the 1D gel. Even a massive overloading of the 1D gel provides a relatively small increase in the number of identified proteins (23) as it mainly results in loading more of the major proteins without providing a real enrichment of minor species On the contrary, such proteins can actually be enriched by overloading the peptide ligand beads, which results in a reduction of protein dynamic range and a better MS/MS sampling of the ions detected by nano-LC-MS/MS. This effect is particularly strong for the gel fractions containing mainly albumin in the crude CSF: as this protein is largely decreased in the corresponding gel slices of the ProteoMiner eluates, many new proteins are unmasked in these bands and largely account for the number of newly identified proteins after treatment (Fig. 2*C*).

To evaluate the origin and function of the proteins revealed by the treatment, we classified the global list of protein groups identified in the study according to their gene ontology terms using the GoMiner software. An example of a subclass of proteins of probable neural origin, associated with the gene ontology term "neurogenesis," is shown in Table I. Proteins are ranked by decreasing abundance (reflected by their total number of MS/MS counts), and those identified only after ProteoMiner treatment are displayed in bold (no protein specific to the crude material was present in this category). It clearly appears that the newly identified species are mostly low abundance proteins, and many of them are involved in very specific neuronal functions such as control of axonal extension and dendrite outgrowth. This is for example the case for Semaphorin-3B, Meteorin, and proteins of the SLITRK family, which have been shown before to be overexpressed in some astrocytic brain tumors (25). These data

indicate that the treatment of CSF with the peptide libraries is able to unmask minor components originating from brain that could constitute biologically relevant biomarkers and that its use could be interesting for clinical studies on CSF.

Assessment of Saturation of Peptide Library Beads for Highly Concentrated Proteins—As the very principle of this strategy is to modify the concentration of the proteins in the starting material, its potential use for quantitative proteomic profiling studies on CSF remained to be assessed. We have previously performed quantitative experiments using red blood cell lysate samples spiked with growing amounts of a yeast protein and treated in triplicate on peptide library beads. We found (i) a good reproducibility of the MS signal of the spiked protein between replicate experiments and (ii) a linear response of the MS signal with increasing concentrations of this protein (23). We concluded that relative quantitative comparison of a protein in different samples was still possible after the treatment as long as the protein did not saturate the beads. Indeed, high abundance proteins that tend to saturate their binding sites during the capture stage would be impossible to quantify because they asymptotically approach a certain value of saturation that does not significantly change with an additional load. For such proteins, any differential expression ratios between samples would probably not be practically exploitable after treatment. To evaluate this saturation effect and roughly assess at which order of concentration it would occur, we performed a test on a panel of proteins spiked at very high amounts in serum samples (up to 20 μM; *i.e.* at the level of some major serum proteins) that were afterward treated in triplicate on ProteoMiner beads (Fig. 1). Fig. 3 shows the mean of the MS signal response (XIC area) of peptides originating from these proteins as a function of the final concentration of the protein spiked in the sample. Interestingly, we obtained a constantly growing signal over a wide range of concentrations, globally linear at intermediate concentrations and flattening only between 2 and 20 μM, probably due to the progressive saturation of the protein baits on the ligand library. These data suggest that quantification of proteins after the treatment may be possible over a concentration range spanning at least 3 orders of magnitude albeit with a compression of differential ratios for highly abundant proteins.

Label-free Quantification of Proteins in CSF Samples Using MFPaQ Software—To evaluate the quantitative reproducibility of the peptide library treatment on a more global scale, a bioinformatics tool able to automatically extract the MS signal of all detected peptides was necessary. We previously described the MFPaQ software, which was designed to parse and validate protein identifications from Mascot result files and quantify the identified proteins when using isotopic labeling strategies such as ICAT, stable isotope labeling with amino acids in cell culture, or $^{14}N/^{15}N$ labeling (24, 26). The quantification module of this software starts from validated peptides lists and uses the *m/z* value and retention time associated to peptide ions (retrieved from Mascot result files)

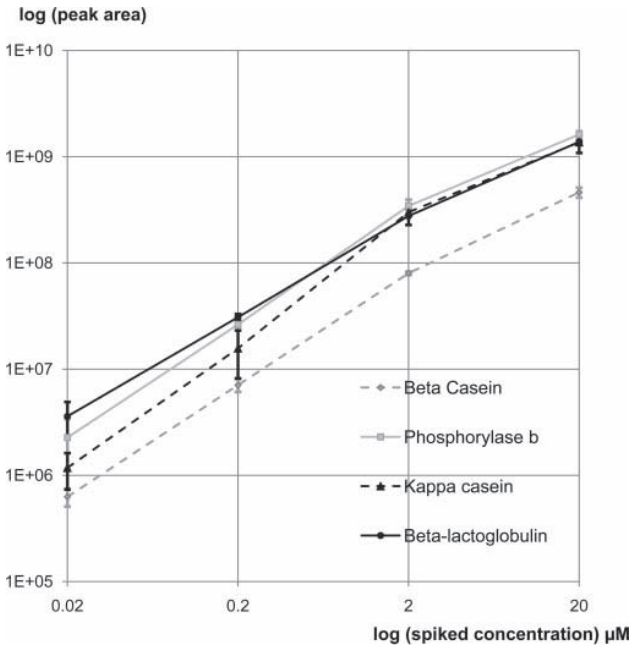


FIG. 3. **Quantitative measurements after ProteoMiner treatment of proteins spiked in serum.** Peak areas correspond to the average of XIC area values for the best identified peptide assigned to a spiked protein calculated over three nano-LC-MS runs (replicate ProteoMiner experiments). This mean value was plotted in a logarithmic scale *versus* the concentration of the spiked protein. *Error bars* correspond to S.D. of the value calculated over the three replicates.

to extract the XIC signal of these ions in the corresponding raw files. In the case of isotopic labeling strategies, if only one member of the light/heavy peptide pair is identified by MS/MS, the software calculates the theoretical *m/z* value of its co-peptide and extracts its XIC signal in the same MS survey scans as the two peptides are expected to co-elute in the same run. In MFPaQ version 4.0, the quantification module was upgraded so that it could handle label-free quantification using a similar approach. Thus, it was used here to extract the XIC signal for the same *m/z* value in two or several LC-MS/MS runs, corresponding to the same unlabeled peptide detected in different parallel analyses.

To assess the accuracy of the bioinformatics label-free quantification and the technical reproducibility of the MS measurement, replicate analyses were performed on either crude CSF or ProteoMiner-treated CSF (first eluate from the NH_2-Library), and identified proteins were quantified using this tool (Fig. 1). For these assays, the samples were not fractionated, and the whole protein mixture was digested and analyzed in one LC-MS/MS run. As shown in Fig. 4*A*, the reproducibility of the nano-LC-MS runs on the Orbitrap mass spectrometer was found to be fairly good between replicate injections of either crude or treated CSF. The peptides XIC area values showed a coefficient of variation for triplicate nano-LC-MS runs typically around 5–10%. The vast majority of the protein population was correctly quantified by the soft-

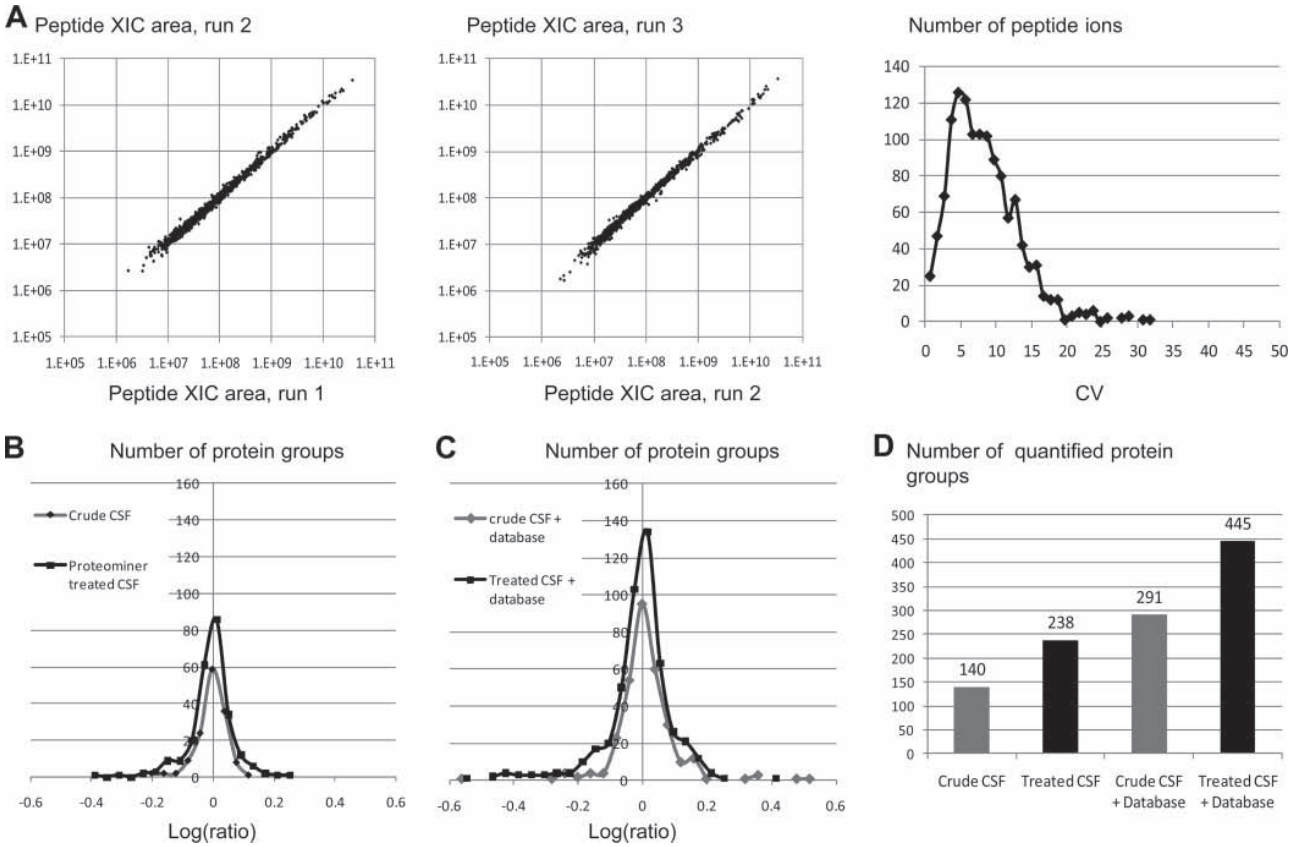


FIG. 4. **Label-free quantification of CSF proteins using MFPaQ.** *A*, reproducibility and accuracy of peptide quantification over triplicate LC-MS/MS analyses of crude CSF sample tryptic digests: correlation of XIC area values in pairwise LC-MS run comparisons (*left panels*) and distribution of peptide ions according to their coefficient of variation (*CV*) for the three replicate measurements (*right panel*). *B*, *C*, and *D*, quantification of proteins over two replicate LC-MS/MS runs of either crude or ProteoMiner-treated CSF. The distribution of protein groups was plotted according to their relative abundance ratio calculated by MFPaQ either by direct comparison of the two LC-MS/MS runs (*B*) or using a previously created identification database (*C*). The total numbers of correctly quantified protein groups are displayed for both samples using the two methods (*D*).

ware with a protein ratio showing little deviation to the expected value of 1 between replicate injections and few outlier values due to incorrect signal extraction (Fig. 4*B*). As expected, the number of proteins identified and quantified in treated CSF was higher than in crude CSF (238 *versus* 140; Fig. 4*D*). These numbers indicate that the treatment with the combinatorial ligand library is even more beneficial when the sample is analyzed in only one LC-MS/MS run without fractionation: indeed, in that case, the huge prevalence of albumin in the global mixture hampers the identification of all others proteins in crude CSF, whereas this effect is much reduced in the treated sample, leading to a final increase of 70% of the number of identified proteins.

Contrary to pattern-based strategies in which LC-MS features have to be defined from the analysis of peptide elution and isotopic profiles in LC-MS maps, the approach used in MFPaQ, based on extracted ion chromatograms of identified peptides, is driven by experimentally measured RT and by monoisotopic *m/z* values validated from MS/MS sequencing. Thus, it allows performing peak detection in a quick and accurate way. A drawback of this method, however, is that only peptides that have been identified by MS/MS in at least one of the LC-MS/MS runs can be quantified. Although the peptide library treatment allows an increase of the number of identified proteins in CSF, this number remains relatively small in a one-run analysis because of MS/MS undersampling of the highly complex peptide mixture. To circumvent this problem, we implemented a strategy based on an identification database containing sequences of previously identified peptides along with their *m/z* and retention time associated values. After fractionation of treated CSF on a 1D SDS gel and LC-MS/MS analysis of 20 gel slices, a database containing 5482 peptide ions, matching to 720 protein groups, was built. Extraction of XICs for all the peptide ions of the database was then performed in replicate analytical runs of the unfractionated sample (either crude or treated CSF). Because of the limited dynamic range of the instrument, not all of the peptides from this database could be retrieved in the individual

runs. However, using such an approach, the MS/MS undersampling problem could be largely overcome, and the number of proteins correctly quantified in replicate runs of the individual samples significantly increased (Fig. 4*C*). Best results were obtained with treated CSF in which up to 445 proteins could be quantified in one run (Fig. 4*D*). Thus, the bioinformatics work flow developed in the MFPaQ software provided an efficient quantitative solution for profiling several hundreds of proteins in treated CSF samples.

Efficiency and Reproducibility of ProteoMiner Treatment on Low CSF Volumes—The experiment described above, designed to provide an extensive and qualitative characterization of the CSF proteome, was performed by treating a large volume of biological fluid on 1-ml peptide bead columns containing a very large panel of the two peptide ligand libraries. Moreover, differential elutions were performed as well as extensive fractionation of each eluate on a 1D gel to reach a better analytical coverage of the sample. Clearly, such an experimental scheme cannot be applied for routine clinical studies. A lumbar puncture typically yields only 1–2 ml of CSF, and because of the very low protein concentration in this fluid, this corresponds to less than 1 mg of material. With such amounts, the bead volume must be decreased to only a few microliters to keep a reasonable overloading ratio of starting material *versus* bead capacity (about 10 mg/ml). Moreover, if the aim of the study is to profile in a reproducible way large numbers of CSF patient samples, fractionation will be very difficult to implement. Thus, we miniaturized the treatment protocol to be compatible with such a clinical application model. A pooled CSF sample was prepared from several lumbar punctures (CSF Pool 2), and 2-ml aliquots were treated in replicate experiments by incubation on either 2- or 5-μl batch volumes of NH_2-Library beads (Fig. 1). Elution was performed by boiling the beads in Laemmli buffer (28), samples were loaded on a gel for 1D SDS-PAGE, and the whole protein mixture was concentrated in one gel band. After tryptic digestion, the resulting peptides were analyzed in one LC-MS/MS run. With such low volumes of beads, both the efficiency and the reproducibility of the treatment had to be assessed. Fig. 5 shows the statistics for sequencing events and identification numbers in LC-MS/MS analytical runs corresponding to either untreated CSF or replicate treated CSF using a small volume of peptide library beads. Although the number of MS/MS queries for all samples is fairly similar, many more peptides and proteins are confidently identified in treated CSF than in initial untreated CSF. Slightly better results are obtained with a 5-μl than with a 2-μl volume of beads: the final number of protein groups identified increases by ~120 and 100% with these two respective treatments. Thus, this miniaturized protocol is still efficient and significantly improves the number of proteins identified in single run LC-MS/MS analysis of CSF samples.

To check the reproducibility of the protocol, label-free quantification of the proteins identified in four replicate experiments was performed using MFPaQ. To evaluate the abundance level of each protein in the samples, we plotted a PAI calculated as the average of signal intensity for the three most abundant tryptic peptides of each protein. In the four LC-MS/MS analyses of the replicate ProteoMiner experiments, the global number of unique protein groups identified and quantified by the software was 283 (5-μl treatment) and 250 (2-μl treatment) when the four runs were compared directly (without database). Although not all of these proteins were identified by MS/MS in every LC-MS/MS run, MS signal could be extracted for all of them in the four replicates for both conditions. As shown in Fig. 6*A*, the PAI profiles for the 283 proteins identified after 5-μl bead treatment were relatively constant over the four replicates. The median of the coefficients of variation for all peptide intensity values over four replicates was 9.2% (5-μl treatment) and 11.7% (2-μl treatment) (Fig. 6*B*), showing a low variability of the experimental protocol. To increase the number of quantified proteins, we applied the bioinformatics work flow based on an identification database. This was done by nano-LC-MS/MS shotgun analysis of one sample from CSF Pool 2 treated on 5 μl of beads fractionated by SDS-PAGE into 20 gel bands, which allowed building a database containing 5015 peptide ions,

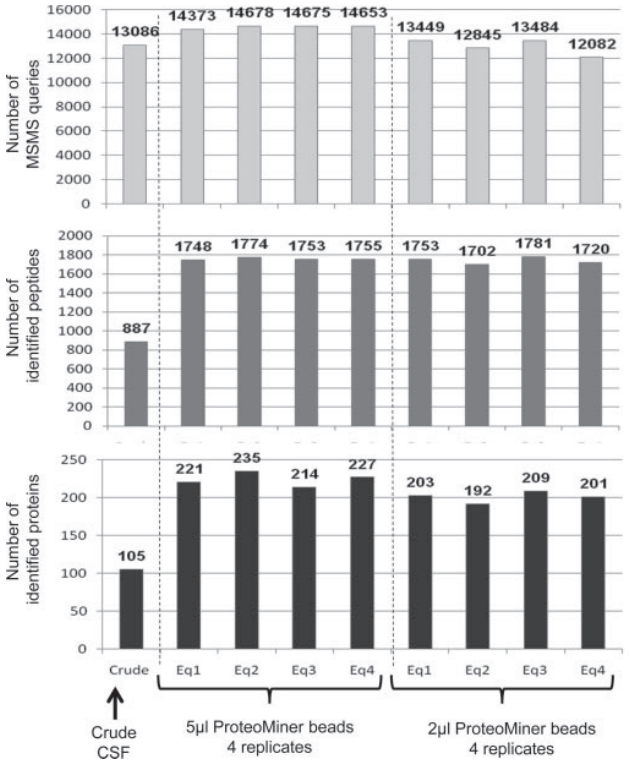


FIG. 5. **Efficiency of ProteoMiner treatment of low volumes of CSF.** Histograms show the number of MS/MS queries, unique identified peptides, and non-redundant identified protein groups in nano-LC-MS/MS analysis of a tryptic digest of either crude CSF or four replicates of 2-ml CSF aliquots treated with low volumes of ProteoMiner beads (5 or 2 μl).

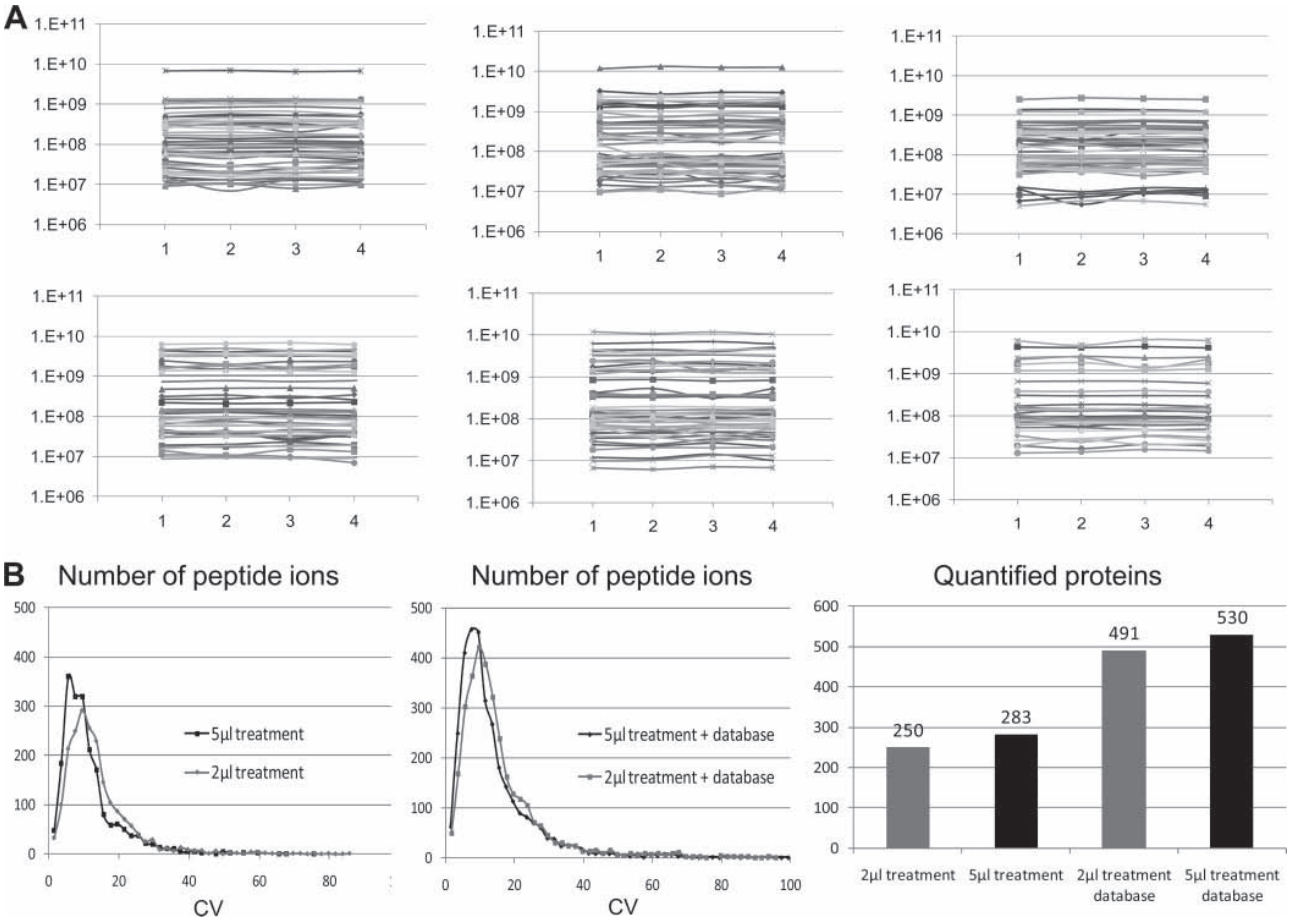


FIG. 6. **Reproducibility of ProteoMiner treatment of low volumes of CSF.** *A*, the six *upper panels* represent the PAI profile for each of the 283 proteins identified in CSF treated with 5 µl of ProteoMiner beads over four replicate experiments. *B*, distribution of the population of quantified peptides as a function of the coefficient of variation (*CV*) of their XIC area calculated over four replicate ProteoMiner experiments with either 5 or 2 µl of beads. Quantification was performed either by direct extraction of the signal of peptides identified in the four nano-LC-MS/MS runs (*left panel*) or by performing signal extraction using *m/z* and retention time values of peptides stored in a previously built identification database (*middle panel*). The final number of proteins quantified with both strategies (with or without database) in the four replicate experiments is represented in the *right panel* for treatment with either 5 or 2 µl of beads.

matching to 568 protein groups. In this way, 530 proteins could be quantified with reproducible PAI values over the four replicates for the 5-µl treatment, and 491 proteins could be quantified for the 2-µl treatment (Fig. 6*B* and supplemental Data 3). Coefficients of variation for the quantified peptides on this population of proteins were only slightly higher (10.4 and 12.4%, respectively, for the 5- and 2-µl treatments), indicating that the quantification remained fairly accurate even on lower abundance proteins. Together, these data indicate that the miniaturized protocol is efficient and reproducible and that several hundred proteins can be accurately quantified in 2-h-long analytical runs by using this treatment on low CSF volumes.

DISCUSSION

In this study, we tried to evaluate the benefits of combinatorial peptide libraries for proteomic characterization of CSF and their potential use for clinical profiling studies by nano-LC-MS. In the first place, an experiment was performed on a large pool of CSF samples with two columns (NH_2- and COOH-Library) each containing 1 ml of beads and sequentially eluted with different buffers; each eluate was then analyzed by 1D gel fractionation and nano-LC-MS/MS. The aim of this experimental design was to optimize the number of proteins detected in CSF by (i) the use of the maximal diversity of hexapeptide ligands to favor the binding of many different species and (ii) the extensive fractionation of the proteins recovered from the columns to improve the analytical coverage of the sample. By using this work flow, we could confidently identify 1149 unique protein groups in the elution fractions from the libraries *versus* 476 protein groups in non-treated CSF. Clearly, there was a positive effect of peptide library beads for the reduction of the dynamic range of the protein mixture by strongly cutting down the high abundance

proteins while enriching the low abundance species. This effect was particularly visible in the gel bands containing albumin in which many more proteins were identified in the treated sample than in the starting material. Indeed, as this protein alone represents about 45% of the total protein amount, it is largely responsible of the masking of low abundance species in CSF as in other fluids such as serum or urine. Moreover, the increased number of proteins detectable in the treated sample was noticeable in almost all regions of the 1D gel migration lane, indicating that the enrichment of minor species was effective on the global protein population. This was also clearly visible from the 2D maps of the untreated CSF and the eluates from the libraries. As reported previously for other fluids (20, 22, 23), a few proteins detected in the starting material were not found in the treated sample (66 proteins in the case of CSF), corresponding to proteins that may not find a bait ligand to form a stable complex. Despite this drawback, the benefit of using the ligand libraries was quite substantial with 745 newly identified proteins.

Together, this study allowed the identification of 1212 unique protein groups in CSF, which represents an in-depth characterization of this fluid by nano-LC-MS/MS. Other data sets of CSF proteins have been published in the past. For instance, using 1D gel fractionation and LC-MS/MS analysis on high resolution mass spectrometers (Orbitrap or FTICR), Zougman *et al.* (13) reported the identification of 798 protein groups by analyzing individual or pooled CSF samples. In a previous report, Pan *et al.* (11) gathered all the data of several CSF studies performed in their group using a variety of protein fractionation techniques (1D SDS-PAGE, ACN precipitation, glycoprotein affinity chromatography, and ICAT labeling), peptide separation methodologies (strong cation exchange and reverse-phase HPLC), and mass spectrometric platforms (ion trap, FTICR, and MALDI-TOF/TOF) (12, 14, 29–31) and merged the results of all these analytical strategies to yield a list of 2594 protein accession numbers. When clustered at 98% homology and compared using the Protein Center software (Proxeon), the three protein lists (this study, 1148 protein groups; Zougman *et al.* (13), 761 protein groups; Pan *et al.* (11), 2093 protein groups) showed a certain extent of overlap with roughly around 500 protein groups shared between two lists in pairwise comparisons and a core set of about 400 protein groups found in common between all three data sets (supplemental Data 4). However, each list showed a relatively high number of specific proteins not found in the two other lists. In the case of the present study, 564 protein groups not reported in the previous publications were identified, most of them found thanks to the sample treatment with peptide libraries. Many reasons may account for the significant difference in protein identifications between these investigations, related either to the analytical techniques used or to the CSF sample itself. Indeed, because of the depth and diversity of the CSF proteome, none of the lists are probably exhaustive, and different fractionation methods or sample treatments may give access to separate categories of proteins. In this respect, the application of selective affinity purification methods (*e.g.* glycoprotein capture or ICAT labeling) *versus* more general enrichment strategies such as combinatorial peptide ligand libraries likely resulted in the detection of different low abundance species. On the other hand, the molecular composition of CSF is highly dynamic and may vary significantly due to individual phenotype fluctuations, age of the patients, sample collection (lumbar puncture *versus* ventricular withdrawal), volume of CSF extracted, pathologies, and CSF bulk flow rate. In particular, the level of blood-derived proteins in CSF is highly dependent on blood-CSF barrier function and dysfunction. Moving from the concept of a morphological blood-CSF "leakage" model, this barrier function is now more widely interpreted using a model connecting molecular flux of blood proteins with CSF flow rate (4, 32). Basically, the CSF flow rate is considered as the main factor modulating protein concentration in the fluid, and a reduced flow rate is for example sufficient to explain the non-linear increase of blood-derived CSF protein concentrations measured in neurological diseases (33). The variation of CSF flow rate with age and pathologies of the individuals will thus largely influence the relative concentrations of blood- or brain-derived proteins in the sample analyzed and may largely account for the heterogeneity of CSF proteomics reports so far.

In an attempt to determine the portion of the CSF proteome originating from blood, Zougman *et al.* (13) compared the list of proteins identified in their study with the Human Proteome Organisation plasma proteome (data set of 889 high confidence plasma proteins) and found that the overlap was quite small as only 24% of the proteins identified in their CSF study were found in the plasma list. When we performed the same kind of comparison with our data, we found a slightly higher overlap when the list of proteins identified in crude CSF was used (39% of the proteins found in the plasma list), which may well reflect the different nature of the samples analyzed (large pool of CSF from disease patients in our study *versus* five diagnostically normal individuals). However, when the comparison was performed with the more extensive list obtained from analysis of ProteoMiner-treated CSF, only 19% of the identified proteins were shared with the plasma data set. Similarly, 18% of the CSF data set from Pan *et al.* (11) overlapped with the plasma list. Thus, as discussed previously, it appears from these numbers that a very large portion of the CSF proteome is constituted by intrinsic species that may not be derived from blood but rather from brain. On the other hand, it is usually described that only about 20% of the protein content of CSF is predominantly brain-derived (1, 6, 34). However, this percentage relates to total protein amounts, whereas the different qualitative proteomics studies evaluate the diversity of the CSF composition only in terms of protein numbers. To get a more precise description of the CSF proteome, we quantified the proteins identified in our study by extracting MS intensity signal for each identified

peptide and calculated for each protein a PAI defined as the average MS signal response for its three most intense tryptic peptides (35). When summed together, the PAI of the 175 proteins shared between the crude CSF list and the plasma list represents around 75% of the total PAI sum, a value close to what is generally described.

Despite the high abundance of blood-derived proteins in CSF, the peptide library technology used here was found to be an efficient approach to unravel low abundance brain species that could not be detected before sample treatment. Interestingly, in CSF from patients without any neurological disease, it permits the detection of proteins related to different cell compartments inside the brain and to different crucial metabolic pathways, sometimes known to be overexpressed in particular diseases. For example, among the proteins involved in neurogenesis displayed in Table I, many were identified only after CSF treatment. This is the case for Semaphorin3B and neuropilin-1 and -2, which play a role in axonal guidance and cell migration (36–38). Detection of these proteins may be of particular interest when studying tumoral brain pathologies such as glioma as they have been proposed to be involved in local invasion and migration of tumor cells, which are pivotal mechanisms in glioma progression (39, 40). Similarly, the 14-3-3 η isoform and the SLIT and NTRK-like protein 1, identified only in treated CSF, have also been shown to be expressed in astrocytoma and glioma cells (25, 41). Other classical brain markers were detected after peptide library treatment, such as neuron-specific enolase (γ-enolase) and the astrocytic protein S100B, which are released in the CSF following brain cell degradation and are often used to detect pathologies involving brain damage such as Creutzfeldt-Jakob disease (42) or to predict the severity of central nervous system injury in cases of cerebral hypoxia, brain infarction, or trauma (43–45). The detection of other minor astrocytic proteins such as the phosphoprotein PEA-15 clearly opens a window on the astrocyte biology (46), whereas detection of IQGAP1 could be indicative of the neural progenitor activity inside the brain (47). Pentraxin-1, a secreted protein involved in synapse remodeling (48), and the Cerebellin-3 neuropeptide (49) are potential indicators of the neurochemical activity of neuronal cells. Neurotrophin receptors were also detected in treated CSF, such as TrkB (receptor of the brain-derived neurotrophic factor), which plays a role in cognition, learning, and memory formation by modulating synapse plasticity and neuronal survival and thus represents a critical molecule in neurodegenerative diseases such as Alzheimer disease (50, 51). Similarly, the detection of the PARK7/DJ-1 protein, one of the Parkinson disease-associated proteins (52), as well as different actors of the proteasome and protease families is interesting in the perspective of biomarker investigation in the context of neurodegenerative diseases (53). Finally, several enzymes involved in major metabolisms were detected, such as the oxidative mitochondrial metabolism (pyruvate kinase isozymes M1/M2, isocitrate dehydrogenase, dihydrolipoyl dehydrogenase, acetyl-CoA acetyltransferase, and malate dehydrogenase), which is important for the exploration of several neurological diseases (54).

Although the peptide ligand libraries appear to be a powerful tool to qualitatively map the CSF proteome, its use for clinical studies was yet to be evaluated. The quantitative aspects and reproducibility of the method were important features to be checked to validate its use for differential proteomics studies. The present work shows quantitative data obtained by measuring, after nano-LC-MS analysis, the signal intensities of peptides derived from growing amounts of exogenous proteins spiked in serum samples subsequently treated with ProteoMiner beads. These spike tests were performed over a wide range of concentrations, up to high final values, to evaluate the saturation effect and check whether relative quantification was still possible even for high or medium abundance proteins. We observed for all proteins a linear MS response up to the μM range, indicating that relative quantification of proteins should be accurate up to such concentration. This is in agreement with previous observations (23) and with a recent study in which quantitative aspects of peptide library treatment were evaluated by 2D electrophoresis (55). Indeed, protein baits captured by the ligand library behave under the law of mass action where relative concentration of species and their affinity constant for a given peptide play the most important role. Binding of a protein to a given ligand takes place proportionally to the concentration of the protein, but as a limited number of molecules of this peptide ligand are available at the surface of the bead, the protein tends to saturate the bead at high concentrations. This saturation effect was visible at μM concentrations for most of the proteins measured. However, no clear plateau was reached probably because complex mechanisms involving multiple affinity equilibria take place at the surface of the beads: one protein does not interact only with one peptide ligand but with several ligands with different affinity constants, and competition between protein species for a given ligand takes place proportionally to their respective concentrations. Thus, a protein may tend to saturate with higher loads, but a growing MS signal is still measurable, indicating that differential expression of even relatively highly abundant proteins could potentially be measured but with an underestimation of the real ratio.

In addition to these spike experiments, reproducibility of the technology was checked on the global population of proteins detected by nano-LC-MS analysis using the label-free quantification module of the MFPaQ software. Reproducibility assays were performed with a miniaturized protocol applicable to real clinical CSF samples. Indeed, due to the high dilution of CSF proteins and to the low volumes of fluid usually available after lumbar puncture, the total protein amount contained in such samples is quite low, which necessitates performing the treatment on very low volumes of beads. The use of small

volumes of peptide ligand beads raises an important concern, which is to assess the effect of a large undersampling of the peptide library on both the efficiency and the reproducibility of the method. Indeed, the library is a pool of beads, each bearing a unique, specific ligand, and because of the diameter of the beads, the total volume of resin necessary to have a statistical representation of almost all the hexapeptide ligands has been estimated to be around a few milliliters.[2] Thus, when very small volumes of beads are taken out of the bulk, the final number of peptide ligands is much smaller than the total number of library diversomers, and in addition, the population of ligands sampled out for each experiment is variable. We show here that this had no major effect on the efficiency of the treatment. When small volume CSF samples were incubated with very low volumes of beads, the number of proteins detected by nano-LC-MS/MS was increased by more than 100% when the analysis was performed in one run, a result even better than what was obtained in the initial experiment using 1 ml of beads. Moreover, label-free quantification on the global population of peptides and proteins identified in replicate treatment experiments showed that, despite the sampling of the library, all protein species were captured with comparable efficiency in all replicate experiments. Coefficients of variation calculated on MS intensity values were centered around 9–12%, a value corresponding roughly to the technical variability of the nano-LC-MS measurement. This can be explained by the fact that the statistical number of peptidomers covering the interaction needs is much reduced compared with combinatorial calculations. Basically, in terms of functionality, peptides that differ from each other because one glycine is replaced by an alanine or because isoleucine is replaced by valine have probably very similar capturing properties for a given protein. In addition, it was shown that sequences of three amino acids seemed already enough for the capture of most proteins from a crude extract (56). Actually, although with hexapeptides obtained with for example 15 amino acids the number of diversomers is 11.4 millions, with tripeptides the number of diversomers is reduced to 3375. Translated into a volume of beads, around 2 μl would be sufficient to cover 90% of tripeptides corresponding to the distal part of the ligands. This approach supports the reproducibility of the data obtained on CSF using 2 and 5 μl of beads. Naturally, the larger the number of beads, the better the coverage of the library, which seems also to be consistent with the fact that a little better reproducibility is obtained with 5 μl of beads compared with 2 μl of beads.

Ideally, a proteomics analytical method aiming at biomarker discovery should allow (i) the detection of a large number of proteins, (ii) their quantification with a good accuracy, and (iii) the analysis of a large number of patient samples to provide statistically significant results. However, a compromise has to be found between these requirements. For example, although extensive 1D gel fractionation of a sample represents a useful way to identify a high number of species by MS/MS, it may not be an ideal solution to quantitatively profile large series of samples in a reproducible way. Conversely, although SELDI-TOF is a well suited tool to perform this later task, it may only give access to the profiling of a limited number of low mass proteins. Nano-LC-MS profiling of enzymatically digested proteins using label-free quantification on high resolution mass spectrometers may provide alternative solutions. These strategies have already been used to identify new biomarkers in CSF with good results (15, 27). In such approaches, very limited or no fractionation at all is usually performed on the sample because of the large number of analyses that must be performed (number of patients and technical replicates) and the intrinsic variability that may be introduced by the fractionation procedure itself. In such an experimental scheme where the protein mixture is analyzed in one run, the treatment of CSF samples on a peptide ligand library was shown to be particularly useful to lower the dynamic range of the protein mixture and increase the number of detected and quantified proteins without fractionating the sample. Moreover, the MFPaQ software (SourceForge) was shown to be an efficient tool to accurately quantify proteins in CSF. With the use of an identification database, previously built from the analysis of 20 fractions of the sample separated by 1D SDS-PAGE, around 500 protein species could be mapped and successfully quantified in replicate nano-LC-MS/MS analyses of CSF aliquots treated on low volumes of beads. Clearly, the analytical coverage of the sample was not as large as in the first large scale, extensive proteomics analysis, which yielded a final list of 1212 proteins. However, it was obtained with an analytical time reduced by a factor 7, and the analysis was reproducible and potentially applicable to many more replicates. Moreover, the list of proteins detected using this strategy still contains many biologically relevant species derived from brain cells (*e.g.* the brain injury markers S100B and neuronal-specific enolase, glial fibrillary acidic protein, pentraxin-1 and -2, neuropilin-1 and -2, PARK7/DJ-1, SLITRK1, etc.), providing potentially interesting candidates in the perspective of biomarker discovery studies.

In conclusion, the analytical work flow presented here using the miniaturized ProteoMiner protocol and the bioinformatics data processing method developed in MFPaQ allowed profiling of the CSF proteome with a reasonable depth (about 500 proteins quantified) in short analytical times (less than 2 h per sample) and with good accuracy (coefficients of variation typically around 10% for replicate measurements). It could thus represent a useful approach for future clinical studies on CSF.

[2] L. Li, C. J. Sun, S. Freeby, D. Yee, S. Kieffer-Jaquinod, L. Guerrier, E. Boschetti, and L. Lomas, submitted manuscript.

* This work was supported by grants from the Institut National du Cancer (INCa), Agence Nationale de la Recherche (ANR/INCa Programme Plates-formes technologiques du vivant), Fondation pour la Recherche Médicale (Programme Grands Equipements), Cancéropole Grand Sud-Ouest, Génopole, and Région Midi-Pyrénées.

[S] This article contains supplemental Data 1–6.

[c] Both authors contributed equally to this work.

[g] Supported by a grant from Fondazione Cariplo (Milan).

[i] To whom correspondence may be addressed: Inst. de Pharmacologie et de Biologie Structurale, 205 route de Narbonne, 31077 Toulouse Cedex 4, France. E-mail: bernard.monsarrat@ipbs.fr.

[j] To whom correspondence may be addressed: Inst. de Pharmacologie et de Biologie Structurale, 205 route de Narbonne, 31077 Toulouse Cedex 4, France. E-mail: gonzalez@ipbs.fr.

REFERENCES

1. Hühmer, A. F., Biringer, R. G., Amato, H., Fonteh, A. N., and Harrington, M. G. (2006) Protein analysis in human cerebrospinal fluid: physiological aspects, current progress and future challenges. *Dis. Markers* **22,** 3–26
2. Johanson, C. E., Duncan, J. A., 3rd, Klinge, P. M., Brinker, T., Stopa, E. G., and Silverberg, G. D. (2008) Multiplicity of cerebrospinal fluid functions: new challenges in health and disease. *Cerebrospinal Fluid Res.* **5,** 10
3. Speake, T., Whitwell, C., Kajita, H., Majid, A., and Brown, P. D. (2001) Mechanisms of CSF secretion by the choroid plexus. *Microsc. Res. Tech.* **52,** 49–59
4. Reiber, H. (1994) Flow rate of cerebrospinal fluid (CSF)—a concept common to normal blood-CSF barrier function and to dysfunction in neurological diseases. *J. Neurol. Sci.* **122,** 189–203
5. Tu, G. F., Cole, T., Southwell, B. R., and Schreiber, G. (1990) Expression of the genes for transthyretin, cystatin C and beta A4 amyloid precursor protein in sheep choroid plexus during development. *Brain Res. Dev. Brain Res.* **55,** 203–208
6. Reiber, H. (2001) Dynamics of brain-derived proteins in cerebrospinal fluid. *Clin. Chim. Acta* **310,** 173–186
7. Goldman, D., Merril, C. R., and Ebert, M. H. (1980) Two-dimensional gel electrophoresis of cerebrospinal fluid proteins. *Clin. Chem.* **26,** 1317–1322
8. Guillaume, E., Zimmermann, C., Burkhard, P. R., Hochstrasser, D. F., and Sanchez, J. C. (2003) A potential cerebrospinal fluid and plasmatic marker for the diagnosis of Creutzfeldt-Jakob disease. *Proteomics* **3,** 1495–1499
9. Sickmann, A., Dormeyer, W., Wortelkamp, S., Woitalla, D., Kuhn, W., and Meyer, H. E. (2000) Identification of proteins from human cerebrospinal fluid, separated by two-dimensional polyacrylamide gel electrophoresis. *Electrophoresis* **21,** 2721–2728
10. Terry, D. E., and Desiderio, D. M. (2003) Between-gel reproducibility of the human cerebrospinal fluid proteome. *Proteomics* **3,** 1962–1979
11. Pan, S., Zhu, D., Quinn, J. F., Peskind, E. R., Montine, T. J., Lin, B., Goodlett, D. R., Taylor, G., Eng, J., and Zhang, J. (2007) A combined dataset of human cerebrospinal fluid proteins identified by multi-dimensional chromatography and tandem mass spectrometry. *Proteomics* **7,** 469–473
12. Abdi, F., Quinn, J. F., Jankovic, J., McIntosh, M., Leverenz, J. B., Peskind, E., Nixon, R., Nutt, J., Chung, K., Zabetian, C., Samii, A., Lin, M., Hattan, S., Pan, C., Wang, Y., Jin, J., Zhu, D., Li, G. J., Liu, Y., Waichunas, D., Montine, T. J., and Zhang, J. (2006) Detection of biomarkers with a multiplex quantitative proteomic platform in cerebrospinal fluid of patients with neurodegenerative disorders. *J. Alzheimers Dis.* **9,** 293–348
13. Zougman, A., Pilch, B., Podtelejnikov, A., Kiehntopf, M., Schnabel, C., Kumar, C., and Mann, M. (2008) Integrated analysis of the cerebrospinal fluid peptidome and proteome. *J. Proteome Res.* **7,** 386–399
14. Zhang, J., Goodlett, D. R., Quinn, J. F., Peskind, E., Kaye, J. A., Zhou, Y., Pan, C., Yi, E., Eng, J., Wang, Q., Aebersold, R. H., and Montine, T. J. (2005) Quantitative proteomics of cerebrospinal fluid from patients with Alzheimer disease. *J. Alzheimers Dis.* **7,** 125–133; discussion 173–180
15. Roy, S., Josephson, S. A., Fridlyand, J., Karch, J., Kadoch, C., Karrim, J., Damon, L., Treseler, P., Kunwar, S., Shuman, M. A., Jones, T., Becker, C. H., Schulman, H., and Rubenstein, J. L. (2008) Protein biomarker identification in the CSF of patients with CNS lymphoma. *J. Clin. Oncol.* **26,** 96–105
16. Guerrier, L., Thulasiraman, V., Castagna, A., Fortis, F., Lin, S., Lomas, L., Righetti, P. G., and Boschetti, E. (2006) Reducing protein concentration range of biological samples using solid-phase ligand libraries. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* **833,** 33–40
17. Righetti, P. G., Boschetti, E., Lomas, L., and Citterio, A. (2006) Protein Equalizer Technology: the quest for a "democratic proteome". *Proteomics* **6,** 3980–3992
18. Righetti, P. G., and Boschetti, E. (2008) The ProteoMiner and the FortyNiners: searching for gold nuggets in the proteomic arena. *Mass Spectrom. Rev.* **27,** 596–608
19. Boschetti, E., and Righetti, P. G. (2009) The art of observing rare protein species in proteomes with peptide ligand libraries. *Proteomics* **9,** 1492–1510
20. Castagna, A., Cecconi, D., Sennels, L., Rappsilber, J., Guerrier, L., Fortis, F., Boschetti, E., Lomas, L., and Righetti, P. G. (2005) Exploring the hidden human urinary proteome via ligand library beads. *J. Proteome Res.* **4,** 1917–1930
21. Sennels, L., Salek, M., Lomas, L., Boschetti, E., Righetti, P. G., and Rappsilber, J. (2007) Proteomic analysis of human blood serum using peptide library beads. *J. Proteome Res.* **6,** 4055–4062
22. Guerrier, L., Claverol, S., Fortis, F., Rinalducci, S., Timperio, A. M., Antonioli, P., Jandrot-Perrus, M., Boschetti, E., and Righetti, P. G. (2007) Exploring the platelet proteome via combinatorial, hexapeptide ligand libraries. *J. Proteome Res.* **6,** 4290–4303
23. Roux-Dalvai, F., Gonzalez de Peredo, A., Simó, C., Guerrier, L., Bouyssié, D., Zanella, A., Citterio, A., Burlet-Schiltz, O., Boschetti, E., Righetti, P. G., and Monsarrat, B. (2008) Extensive analysis of the cytoplasmic proteome of human erythrocytes using the peptide ligand library technology and advanced mass spectrometry. *Mol. Cell. Proteomics* **7,** 2254–2269
24. Bouyssié, D., Gonzalezde Peredo, A., Mouton, E., Albigot, R., Roussel, L., Ortega, N., Cayrol, C., Burlet-Schiltz, O., Girard, J. P., and Monsarrat, B. (2007) Mascot file parsing and quantification (MFPaQ), a new software to parse, validate, and quantify proteomics data generated by ICAT and SILAC mass spectrometric analyses: application to the proteomics study of membrane proteins from primary human endothelial cells. *Mol. Cell. Proteomics* **6,** 1621–1637
25. Aruga, J., Yokota, N., and Mikoshiba, K. (2003) Human SLITRK family genes: genomic organization and expression profiling in normal brain and brain tumor tissue. *Gene* **315,** 87–94
26. Stanislas, T., Bouyssie, D., Rossignol, M., Vesa, S., Fromentin, J., Morel, J., Pichereaux, C., Monsarrat, B., and Simon-Plas, F. (2009) Quantitative proteomics reveals a dynamic association of proteins to detergent-resistant membranes upon elicitor signaling in tobacco. *Mol. Cell. Proteomics* **8,** 2186–2198
27. Fang, Q., Strand, A., Law, W., Faca, V. M., Fitzgibbon, M. P., Hamel, N., Houle, B., Liu, X., May, D. H., Poschmann, G., Roy, L., Stühler, K., Ying, W., Zhang, J., Zheng, Z., Bergeron, J. J., Hanash, S., He, F., Leavitt, B. R., Meyer, H. E., Qian, X., and McIntosh, M. W. (2009) Brain-specific proteins decline in the cerebrospinal fluid of humans with Huntington disease. *Mol. Cell. Proteomics* **8,** 451–466
28. Candiano, G., Dimuccio, V., Bruschi, M., Santucci, L., Gusmano, R., Boschetti, E., Righetti, P. G., and Ghiggeri, G. M. (2009) Combinatorial peptide ligand libraries for urine proteome analysis: investigation of different elution systems. *Electrophoresis* **30,** 2405–2411
29. Pan, S., Wang, Y., Quinn, J. F., Peskind, E. R., Waichunas, D., Wimberger, J. T., Jin, J., Li, J. G., Zhu, D., Pan, C., and Zhang, J. (2006) Identification of glycoproteins in human cerebrospinal fluid with a complementary proteomic approach. *J. Proteome Res.* **5,** 2769–2779
30. Zhang, J., Goodlett, D. R., Peskind, E. R., Quinn, J. F., Zhou, Y., Wang, Q., Pan, C., Yi, E., Eng, J., Aebersold, R. H., and Montine, T. J. (2005) Quantitative proteomic analysis of age-related changes in human cerebrospinal fluid. *Neurobiol. Aging* **26,** 207–227
31. Xu, J., Chen, J., Peskind, E. R., Jin, J., Eng, J., Pan, C., Montine, T. J., Goodlett, D. R., and Zhang, J. (2006) Characterization of proteome of human cerebrospinal fluid. *Int. Rev. Neurobiol.* **73,** 29–98
32. Reiber, H. (2003) Proteins in cerebrospinal fluid and blood: barriers, CSF flow rate and source-related dynamics. *Restor. Neurol. Neurosci.* **21,** 79–96
33. Reiber, H., and Peter, J. B. (2001) Cerebrospinal fluid analysis: disease-

related data patterns and evaluation programs. *J. Neurol. Sci.* **184,** 101–122
34. Thompson, E. J. (1988) *The CSF Proteins: a Biochemical Approach*, Elsevier, Amsterdam
35. Silva, J. C., Gorenstein, M. V., Li, G. Z., Vissers, J. P., and Geromanos, S. J. (2006) Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol. Cell. Proteomics* **5,** 144–156
36. Yazdani, U., and Terman, J. R. (2006) The semaphorins. *Genome Biol.* **7,** 211
37. Kruger, R. P., Aurandt, J., and Guan, K. L. (2005) Semaphorins command cells to move. *Nat. Rev. Mol. Cell Biol.* **6,** 789–800
38. Geretti, E., and Klagsbrun, M. (2007) Neuropilins: novel targets for anti-angiogenesis therapies. *Cell Adh. Migr.* **1,** 56–61
39. Karayan-Tapon, L., Wager, M., Guilhot, J., Levillain, P., Marquant, C., Clarhaut, J., Potiron, V., and Roche, J. (2008) Semaphorin, neuropilin and VEGF expression in glial tumours: SEMA3G, a prognostic marker? *Br. J. Cancer* **99,** 1153–1160
40. Rich, J. N., Hans, C., Jones, B., Iversen, E. S., McLendon, R. E., Rasheed, B. K., Dobra, A., Dressman, H. K., Bigner, D. D., Nevins, J. R., and West, M. (2005) Gene expression profiling and genetic markers in glioblastoma survival. *Cancer Res.* **65,** 4051–4058
41. Yang, X., Cao, W., Lin, H., Zhang, W., Lin, W., Cao, L., Zhen, H., Huo, J., and Zhang, X. (2009) Isoform-specific expression of 14-3-3 proteins in human astrocytoma. *J. Neurol. Sci.* **276,** 54–59
42. Pennington, C., Chohan, G., Mackenzie, J., Andrews, M., Will, R., Knight, R., and Green, A. (2009) The role of cerebrospinal fluid proteins as early diagnostic markers for sporadic Creutzfeldt-Jakob disease. *Neurosci. Lett.* **455,** 56–59
43. Undén, J., Strandberg, K., Malm, J., Campbell, E., Rosengren, L., Stenflo, J., Norrving, B., Romner, B., Lindgren, A., and Andsberg, G. (2009) Explorative investigation of biomarkers of brain damage and coagulation system activation in clinical stroke differentiation. *J. Neurol.* **256,** 72–77
44. Bloomfield, S. M., McKinney, J., Smith, L., and Brisman, J. (2007) Reliability of S100B in predicting severity of central nervous system injury. *Neurocrit. Care* **6,** 121–138
45. Gonçalves, C. A., Leite, M. C., and Nardin, P. (2008) Biological and methodological features of the measurement of S100B, a putative marker of brain injury. *Clin. Biochem.* **41,** 755–763
46. Sharif, A., Canton, B., Junier, M. P., and Chneiweiss, H. (2003) PEA-15 modulates TNFalpha intracellular signaling in astrocytes. *Ann. N.Y. Acad. Sci.* **1010,** 43–50
47. Balenci, L., Saoudi, Y., Grunwald, D., Deloulme, J. C., Bouron, A., Bernards, A., and Baudier, J. (2007) IQGAP1 regulates adult neural progenitors in vivo and vascular endothelial growth factor-triggered neural progenitor migration in vitro. *J. Neurosci.* **27,** 4716–4724
48. Tsui, C. C., Copeland, N. G., Gilbert, D. J., Jenkins, N. A., Barnes, C., and Worley, P. F. (1996) Narp, a novel member of the pentraxin family, promotes neurite outgrowth and is dynamically regulated by neuronal activity. *J. Neurosci.* **16,** 2463–2478
49. Pang, Z., Zuo, J., and Morgan, J. I. (2000) Cbln3, a novel member of the precerebellin family that binds specifically to Cbln1. *J. Neurosci.* **20,** 6333–6339
50. Patapoutian, A., and Reichardt, L. F. (2001) Trk receptors: mediators of neurotrophin action. *Curr. Opin. Neurobiol.* **11,** 272–280
51. Schindowski, K., Belarbi, K., and Buée, L. (2008) Neurotrophic factors in Alzheimer's disease: role of axonal transport. *Genes Brain Behav.* **7,** Suppl. 1, 43–56
52. da Costa, C. A. (2007) DJ-1: a newcomer in Parkinson's disease pathology. *Curr. Mol. Med.* **7,** 650–657
53. Ross, C. A., and Pickart, C. M. (2004) The ubiquitin-proteasome pathway in Parkinson's disease and other neurodegenerative diseases. *Trends Cell Biol.* **14,** 703–711
54. Mattson, M. P., Gleichmann, M., and Cheng, A. (2008) Mitochondria in neuroplasticity and neurological disorders. *Neuron* **60,** 748–766
55. Hartwig, S., Czibere, A., Kotzka, J., Passlack, W., Haas, R., Eckel, J., and Lehr, S. (2009) Combinatorial hexapeptide ligand libraries (ProteoMiner): an innovative fractionation tool for differential quantitative clinical proteomics. *Arch. Physiol. Biochem.* **115,** 155–160
56. Simó, C., Bachi, A., Cattaneo, A., Guerrier, L., Fortis, F., Boschetti, E., Podtelejnikov, A., and Righetti, P. G. (2008) Performance of combinatorial peptide libraries in capturing the low-abundance proteome of red blood cells. 1. Behavior of mono- to hexapeptides. *Anal. Chem.* **80,** 3547–3556

3.4.2 Quantification d'échantillons fractionnés

L'utilisation de la banque d'identification pour la quantification d'échantillons protéiques non fractionnés peut s'avérer pratique dans le contexte d'analyses cliniques où l'on doit analyser des grandes séries d'échantillons. Elle permet en effet de maximiser le nombre de protéines pour lesquelles on peut fournir à la fois des données d'identification et de quantification. Cependant, lorsque l'on s'intéresse à des plans d'expériences plus simples, mettant en œuvre par exemple la comparaison de seulement deux conditions expérimentales, cette stratégie ne représente pas la solution optimale. L'analyse de grandes séries d'échantillons ne nous permet pas de réaliser un fractionnement du mélange protéique étudié car le temps d'analyse requis serait trop important. Par contre, lorsque l'on compare seulement deux échantillons, l'utilisation d'un fractionnement de type SDS-page permet d'augmenter considérablement la gamme dynamique de la quantification tout en nécessitant un temps d'analyse raisonnable. On peut ainsi espérer quantifier des espèces très minoritaires qui seraient restées invisibles dans le cas de l'injection de l'échantillon non fractionné, avec toutefois un plus faible niveau de reproductibilité de la quantification réalisée.

Ce compromis entre exhaustivité et reproductibilité de l'analyse quantitative est discuté dans un article récemment soumis pour publication (cf annexe) concernant une étude réalisée au sein du laboratoire sur la caractérisation protéomique extensive de cellules endothéliales humaines primaires au cours de la réponse inflammatoire. En effet, depuis sa version 4, MFPaQ est capable de gérer des analyses SDS-PAGE comparatives pour des échantillons non marqués. En pratique chaque piste du gel correspond à une des conditions, ou à un réplicat de fractionnement d'une des conditions. Suite à la découpe systématique de toutes les pistes du gel et à l'analyse par nanoLC-MS/MS, l'analyse quantitative réalisée par le logiciel consiste alors à comparer l'abondance des peptides et des protéines entre les différentes conditions et pour un même niveau de découpe. Les optimisations apportées à MFPaQ pour la gestion de ce type d'analyse ont donc consisté à organiser les jeux de données suivant le fractionnement mis en œuvre, et à appliquer automatiquement l'ensemble des algorithmes de quantification sur chaque niveau de découpe. Comme le montre la figure 5a du manuscrit en annexe, une normalisation effectuée suivant cette procédure donne de meilleurs résultats qu'une normalisation globale où un facteur correctif est déterminé uniquement à partir du signal total mesuré pour chacune des pistes comparées. En effet, dans le premier cas le logiciel détermine automatiquement un facteur de normalisation différent pour chacun des niveaux de découpe, issu de la comparaison du signal pour des analyses proches dans le temps, correspondant à des bandes de gels ayant un contenu protéique similaire.

Par ailleurs, des procédures permettant d'intégrer l'information quantitative sur l'ensemble de la piste SDS-PAGE ont également été mises en œuvre. Le PAI tel qu'il a été décrit précédemment est calculé pour l'ensemble des protéines identifiées et quantifiées au sein de chaque niveau de découpe. Une fois que toutes les fractions ont été analysées, le logiciel calcule alors une liste quantitative globale appelée « quantification overview ». Celle-ci s'apparente à la liste de protéines que le logiciel détermine à partir de données d'identifications dans le sens où elle représente un ensemble de groupes de protéines non-redondant. Ce dernier comporte les informations quantitatives recensées sur l'ensemble des fractions et il est possible de l'exporter au format Excel pour effectuer des traitements ultérieurs comme des analyses statistiques (test de Student, ANOVA, ACP…). Le fichier de sortie du « quantification overview » contient une valeur de PAI totale pour chacune des pistes analysées. Cette valeur totale est calculée à partir de la somme des 3 bandes de

gel consécutive s, de part et d'autre de la bande présentant une intensité maximale. Cette recherche de maximum est effectuée de la manière suivante pour chacun des groupes de protéines :

1) recherche du fichier d'acquisition (fichier brut) présentant la valeur de PAI la plus élevée sur l'ensemble des pistes et des bandes analysées,
2) mémorisation du niveau de découpe correspondant et recherche du PAI dans les niveaux précédents et suivants pour les différentes conditions,
3) somme des PAI pour chacune des conditions sur les 3 niveaux de découpe sélectionnés.

L'intégration du PAI sur 3 bandes consécutives permet de diminuer le risque d'erreur lié à la variabilité de la migration des protéines sur un gel SDS-PAGE, et à la variabilité de la découpe par l'expérimentateur des bandes de gels sur les différentes pistes. Comme le montre la table 1 de l'article fourni en annexe, l'intensité maximale des ions précurseurs quantifiés ne correspond pas toujours à un même niveau de découpe : environ 4% des signaux ont des valeurs maximales détecté sur 2 ou 3 niveaux de découpes différents pour l'ensemble des conditions comparées. Le bénéfice lié à cette méthode de calcul du PAI total est illustré dans le deuxième graphique de la figure 5.

Au final, la reproductibilité quantitative a été estimée pour des échantillons ayant subi ou non un fractionnement SDS-PAGE (cf figure 1 du manuscrit). Dans le cas d'échantillons fractionnés, les procédures de normalisation et d'intégration décrites ci-dessus ont été appliquées. Comme le montre la figure 4 du manuscrit, la mise en œuvre d'un fractionnement permet d'augmenter le nombre de protéines quantifiées de façon non négligeable. Que ce soit pour des réplicats d'injection ou de gel, les coefficients de variation observés pour ce mode d'analyse sont légèrement supérieurs à ceux obtenus pour des échantillons non fractionnés mais ils restent cependant tout à fait raisonnables (CV médian de 7%), et cette approche quantitative a donc pu être utilisée efficacement pour l'analyse de cellules endothéliales stimulées par différentes cytokines pro-inflammatoires.

Prosper : nouveaux enjeux et développements

Le développement du logiciel MFPaQ a permis de répondre à un grand nombre de besoins fonctionnels du laboratoire tant au niveau de la validation des résultats d'identification que de la quantification des acquisitions LC-MS/MS. Cependant, l'architecture du logiciel a fini par en limiter les usages. Le logiciel MFPaQ ne peut pas par exemple gérer un très grand nombre d'échantillons au niveau du module de quantification, ou encore accéder aux séquences des protéines identifiées. Il n'est pas non plus possible de récupérer l'ensemble des spectres MS/MS correspondant à une séquence peptidique donnée qui aurait été validée par MFPaQ à partir des analyses de différents échantillons. Ce type de requête transversale sur l'ensemble des données traitées est pourtant en passe de devenir une fonctionnalité essentielle pour des analyses protéomiques ciblées. En effet, pour réaliser ce type d'analyse il est nécessaire de connaître les informations de fragmentation relatives à un peptide d'intérêt dont on souhaiterait suivre l'abondance au sein d'un ensemble d'échantillons à comparer. Pour l'instant il est nécessaire d'utiliser des bases de données publiques telles que Peptide Atlas pour accéder à ce type d'information, alors qu'elle pourrait être également recherchée à partir de l'ensemble des données expérimentales acquises sur la plateforme protéomique. Dans le cadre de notre participation au projet HPP (Human Proteome Project), nous avons été sollicités pour fournir la liste de l'ensemble des protéines du chromosome 14 que nous avions identifiées ces dernières années. Nous avons dû mettre en place un moteur de recherche spécifique des données stockées par MFPaQ, alors que la disponibilité d'un outil nous permettant de réaliser des requêtes croisées sur l'ensemble des jeux de données traités nous aurait grandement facilité la tâche.

En parallèle de mes travaux, le laboratoire grenoblois EDyP (Etude de la dynamique des protéomes, CEA) dirigé par Christophe Bruley avait commencé la conception d'une base donnée relationnelle pour le stockage et l'exploitation de données d'identification par spectrométrie de masse : la MSIdb pour « Mass Spectrometry Identification database ». Dans la perspective de créer une infrastructure protéomique nationale (ProFI pour « Proteomics French Infrastructure ») nous avions commencé une étroite collaboration entre nos deux laboratoires et le LSMBO (Laboratoire de Spectrométrie de Masse Bio-Organique, Institut Hubert Curien) de Strasbourg dirigé par Alain Van Dorsselaer. L'objectif de l'infrastructure est de promouvoir le développement de méthodologies pour la protéomique, et en particulier de nouveaux outils bioinformatiques robustes et efficaces pour l'analyse des données à très haut débit. La première étape de notre collaboration sur le plan informatique a consisté à concevoir un modèle de base de données qui pourrait être commun aux différents sites de façon à homogénéiser la représentation des données traitées.

Au-delà de cette collaboration, j'ai entrepris l'implémentation d'autres schémas de bases de données ainsi que la réalisation d'un prototype appelé « Prosper » pour la gestion et l'exploitation des données stockées dans ces différentes bases.

IV-1. Une architecture plus élaborée

La base de données MSIdb ne répondait pas à elle seule à l'ensemble des besoins de notre laboratoire. J'ai donc pris l'initiative de développer d'autres schémas de bases de données relationnelles dans le but d'obtenir une couverture complète de ces besoins fonctionnels :

- la LCMSdb pour « Liquid Chromatography Mass Spectrometry database » dont le but est de servir d'entrepôt de cartes LC-MS. La base a été conçue de telle manière que ces cartes peuvent provenir de différents logiciels de « peak picking » (comme OpenMS, Progenesis LCMS ou Decon2LS) permettant ainsi de les comparer mais également d'avoir une grande flexibilité au niveau du traitement des données MS.
- la PDIdb pour « Protein Database Index » qui stocke l'ensemble des séquences protéiques analysées par les moteurs de recherche et indexe les fichiers d'annotation tels que ceux fournis par Uniprot. La base de données recense également l'historique de tous les numéros d'accessions relatifs à une protéique permettant ainsi de retrouver une protéine facilement même si son identifiant est obsolète.
- l'UDSdb pour « User Data Sets database » qui est la colonne vertébrale du système d'information car elle réalise l'organisation des jeux de données des utilisateurs en projets mais fédère aussi les données de la MSIdb et de la LCMSdb.

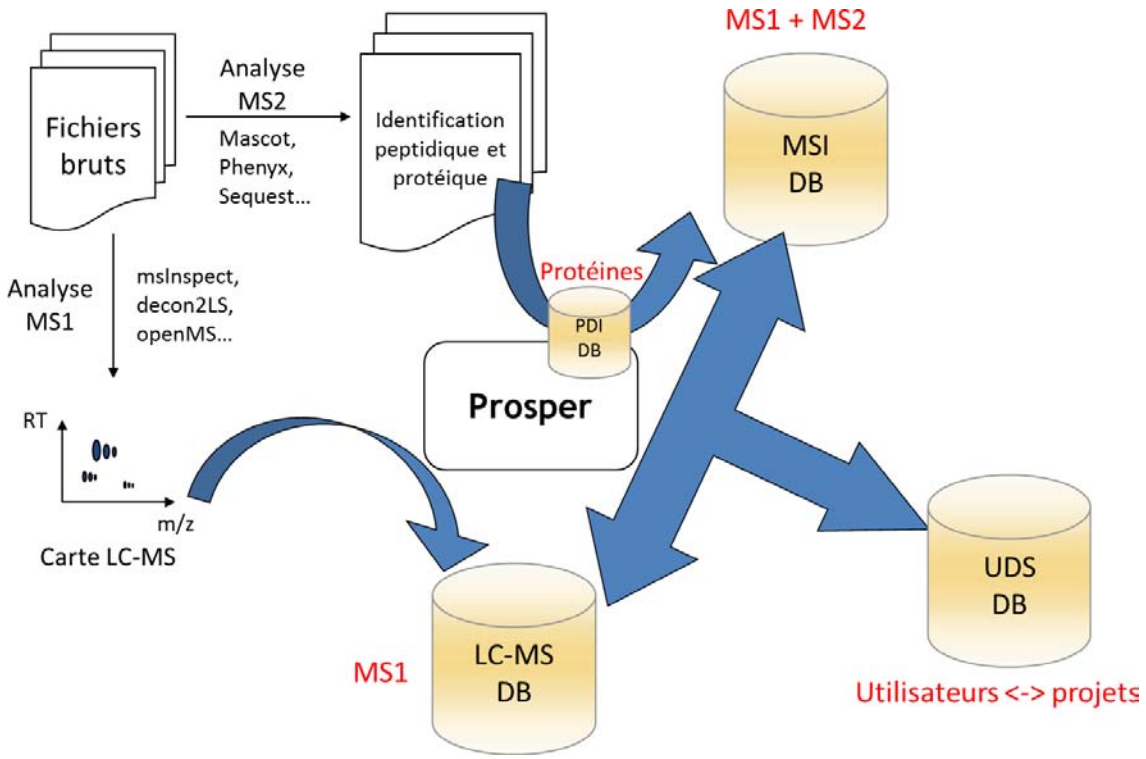


Figure 30 : architecture du système d'information de Prosper. Les fichiers bruts génèrent à la fois des données d'identification sous forme de fichiers résultats des moteurs de recherche, et de quantification sous forme généralement de cartes LC-MS. Le système d'information de Prosper permet d'intégrer ces deux niveaux d'information dans les bases de données MSIdb et LCMSdb ainsi que d'organiser les jeux de données utilisateurs via l'UDSdb.

Cette nouvelle architecture (cf figure 30) offre la flexibilité nécessaire pour stocker les résultats des traitements des jeux de données protéomiques mais également pour réaliser des requêtes complexes sur l'ensemble des données en un temps réduit.

Le développement d'un système d'information est un processus de modélisation long et complexe. Même avec une connaissance théorique du domaine et une expérience de développement logiciel tel que MFPaQ il est difficile de concevoir le modèle idéal sans l'avoir préalablement testé. Dans cette optique j'ai mis au point le logiciel Prosper en partenariat avec Stéphanie Bolot, ingénieur recrutée au sein du laboratoire pendant un an et demi. Je souhaitais disposer d'un prototype qui pourrait à la fois mettre à l'épreuve cette nouvelle architecture de stockage des données et également tester différents algorithmes de traitement des données protéomiques. Il s'agit d'ailleurs d'un travail cyclique car le modèle alimente les algorithmes qui à leur tour permettent de vérifier si le modèle est pertinent. En effet, le système d'information fournit une persistance des données nécessaire pour stocker les données traitées par les algorithmes ainsi que les résultats obtenus. Par contre, si le modèle n'est pas suffisamment exhaustif il est possible qu'un algorithme donné ne puisse pas stocker le résultat de son analyse au sein du système d'information et il faut alors changer le modèle. La modification d'un modèle est une opération triviale tant qu'elle n'est que théorique. En effet, si une base de données a été créée en utilisant un modèle de première génération et que l'on souhaite ensuite utiliser un modèle de deuxième génération, il est alors nécessaire de réaliser la migration des données de la base du premier vers le second modèle. Si les modifications sont nombreuses cette opération peut être délicate surtout si la volumétrie des données stockées est importante. La réalisation du prototype Prosper, dans le langage Perl, avait donc également comme objectif d'éviter des migrations trop fréquentes des bases de données qui seraient utilisées en routine sur les différents sites.

IV-2. MSIdb et algorithmes de validation

4.2.1 Stratégie de validation avancée des données d'identification

Comme nous l'avons vu dans l'introduction, les stratégies de validation des résultats d'identification ont grandement évolué ces dernières années. Une des méthodes qui s'est clairement démarquée par sa simplicité de mise œuvre ainsi que par sa flexibilité (adaptable à tout type de moteur de recherche, d'instrument…) est l'approche « target-decoy ». Nous avons vu que celle-ci présente cependant des inconvénients et notamment la dualité du mode de recherche (banques séparées ou banques concaténées), ceci ayant pour conséquence d'aboutir à des calculs de FDRs différents selon le mode utilisé et donc à des résultats de validation différents.

Nous avons cependant évoqué brièvement l'existence d'une méthode unifiant le calcul du FDR. En effet la stratégie proposée dans (Navarro and Vazquez 2009) consiste à forcer la compétition entre les identifications target et decoy lorsque la recherche est effectuée en banques séparées. Ainsi de façon similaire à une recherche concaténée l'effet de compétition élimine certains hits decoy qui sont mieux interprétés dans la banque target, mais par contre l'espace de recherche n'est pas doublé évitant ainsi d'introduire un biais dans le calcul du FDR. La mise en compétition repose sur l'utilisation d'une « joint-table » où les scores des peptides sont comparés pour les deux types d'analyses target et decoy (cf figure 31). C'est l'utilisation du résumé statistique de cette table qui permet d'accéder à la valeur du FDR.

MS query	Target Score	Decoy Score	Class
1	50	35	Target better (tb)
2	35	40	Decoy better (db)
3	70	20	Target only (to)
4	15	35	Decoy only (do)
5	25	10	Target under (tu)
6	10	25	Decoy under (du)

Note : « Score threshold » = 30

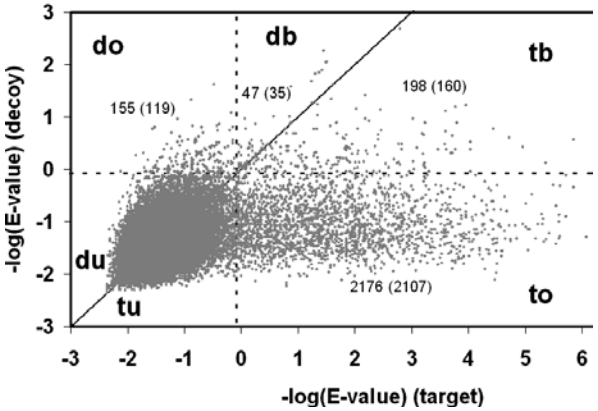


Figure 31 : mise en compétition des résultats d'identification issus d'analyses « target-decoy » en banques séparées. La table est une illustration de la « joint table » utilisée dans l'approche décrite dans (Navarro and Vazquez 2009). Elle montre les différents cas de figure (colonne « Class ») qui peuvent exister lorsque l'on compare les scores de peptides identifiés par une approche target/decoy à partir des mêmes spectres MS/MS. Le comptage des peptides appartenant à chacune des classes est utilisé comme base pour la détermination du taux de faux positifs calculé via la formule : FDR = (2db + do) / (db + tb + to). Le graphe de droite est tiré de la publication de cette même étude et illustre d'une façon globale sous la forme d'un nuage de points le résultat de cette mise en compétition (recherche Mascot réalisées sur 40000 spectres MS/MS d'un protéome total de cellules Jurkat obtenus à l'aide d'un instrument de type LCQ-DECA XP (Lopez-Ferrer, Martinez-Bartolome et al. 2004)). On perçoit clairement une symétrie des données entre les hits target et les hits decoy (zone « du/tu »), correspondant essentiellement aux fausses identifications. Les vrais positifs sont majoritairement représentés dans les zones tb et to. Comme le montre la figure, il est possible d'utiliser le logarithme de la E-value à la place du score pour générer la « joint table ».

J'ai implémenté cet algorithme dans le logiciel Prosper en espérant ainsi disposer d'une méthode quasi universelle pour l'estimation du taux de faux positifs des résultats d'identification. Au lieu d'utiliser les valeurs de score ou d'E-value fournies par Mascot j'ai opté pour l'utilisation d'une E-value ajustée pour chacun des peptides. Comme nous l'avons vu dans l'introduction cette valeur est classiquement calculée à partir de l' « identity threshold ». On peut cependant obtenir un meilleur pouvoir si l'on calcule cette E-value à partir de l' « homology threshold » lorsqu'il est inférieur au seuil d'identité. En pratique le seuil d'homologie n'est pas toujours défini et on ne peut donc pas calculer cette valeur probabilistique pour tous les peptides. Cependant on peut créer une fonction de calcul hybride qui choisit le seuil le plus faible parmi les seuils qui sont définis pour un peptide donné, et calcule ensuite la E-value ajustée à partir de ce seuil. En utilisant cette méthode dans le cas d'un mélange complexe de protéines, et pour un FDR peptidique de 5%, nous observons en général un gain d'environ 30% en identification de spectres MS/MS et un gain d'environ 10% en identification de protéines. Cette approche est d'ailleurs similaire à celle que Mascot emploie pour le calcul du score protéique MudPIT. Une fois cette nouvelle E-value calculée pour tous les peptides, ces derniers sont ensuite validés en faisant varier de manière itérative un seuil de E-value jusqu'à obtenir une valeur de FDR souhaité (5% par exemple). A chaque étape la valeur de FDR est calculée via l'approche de Navarro décrite précédemment.

La validation des peptides ne constitue que la première étape du processus de validation réalisé par Prosper. L'ensemble des peptides validés sont ensuite regroupés en protéines en suivant le principe de « parcimonie ». Les groupes de protéines définis comme des « oversets » (cf partie I-4.5) subissent

à leur tour une étape de validation avec un contrôle du FDR. Pour cela, les données « target » et « decoy » ont été traitées de façon analogue (même algorithme de regroupement). Le FDR est alors calculé simplement de la façon suivante :

FDR = 100 x nb_protéines_decoy / nb_protéines_target

Comme nous l'avons vu dans la partie I-5.4 les vraies identifications peptidiques se regroupent de façon non aléatoire en protéines contrairement aux fausses identifications. En pratique, on constate en effet que les protéines « target » sont en général identifiées par plusieurs peptides alors que les protéines « decoy » sont peu fréquemment représentées par plusieurs peptides. Par conséquent, le FDR protéique est toujours plus élevé que le FDR peptidique si la validation n'est réalisée que sur les peptides. La validation au niveau peptidique n'est donc pas suffisante si l'on souhaite réellement contrôler le taux d'erreur d'identifications protéiques. Pour atteindre un FDR protéique souhaité, il est nécessaire de filtrer les protéines identifiées à partir de critères spécifiques. J'ai implémenté deux stratégies de validation des protéines dans Prosper. La première, identique à celle utilisée dans MFPaQ, met en œuvre des jeux de critères peptidiques (seuil de score, longueur de séquence) plus ou moins stringents en fonction du nombre de peptides identifiés (1 et 2 par défaut) pour la protéine considérée (cf partie II-2). Bien que cette approche assez simple se soit révélée très satisfaisante, elle ne fournit pas la liste de protéines optimale pour un FDR donné. Afin d'illustrer les biais introduits par cette stratégie, prenons l'exemple de deux protéines A et B :

- soit A une protéine identifié avec un peptide de score 25 et autre de score 37,
- soit B une protéine identifiée avec un peptide de score 27 et autre de score 28.

Si nous établissons les règles de validation stipulant qu'une protéine est validée si elle possède :

- soit 1 peptide de score supérieur ou égal à 40,
- soit 2 peptides de score supérieur ou égal à 26.

Alors la protéine B sera validée mais pas la protéine A. Pourtant la présence d'un peptide de score 37 (proche du seuil de score pour les identifications protéiques à un peptide unique) nous indique la présence d'une bonne identification peptidique. De façon intuitive, nous souhaiterions que le score de 37 puisse « compenser » celui de 25. Une façon simpliste de mettre en œuvre cette compensation est de sommer les scores peptidiques, ce qui nous donne un score de 62 pour la protéine A et de 55 pour la protéine B. Ce calcul correspond à la fonction standard de calcul de score protéique de Mascot. Comme nous l'avons vu dans la partie 1.4.5 elle présente le défaut de prendre en compte les identifications peptidiques avec des scores très faibles. La fonction de score MudPIT a pour but de corriger ce défaut et également de prendre en compte les statistiques de seuil d'identité et d'homologie pour chaque peptide individuel. La deuxième stratégie de validation de Prosper s'inspire d'ailleurs de cette fonction de score. Cependant, contrairement à la fonction de Mascot, celle de Prosper n'ajoute pas à la fin la somme des seuils soustraits. Voici l'algorithme de cette fonction :

```
Protein score = 0

For each peptide match {
  If there is a homology threshold and ions score > homology threshold {
    Protein score += peptide score - homology threshold
  } else if ions score > identity threshold {
    Protein score += peptide score - identity threshold
  }
}
Protein score += 1 * average of all the subtracted thresholds
```

Ainsi les scores protéiques de Prosper ont une valeur minimale qui est toujours proche de zéro alors que le score MudPIT définit une valeur minimale variable selon le jeu de donnée considéré (la somme des seuils soustraits pouvant changer d'un fichier à un autre).

Afin d'atteindre la valeur de FDR souhaité au niveau protéique les seuils d'identité et d'homologie sont modulés en faisant varier la valeur de p-value qui permet de les définir. Le programme s'initialise avec une p-value élevée (1 par défaut) afin de démarrer l'algorithme avec une validation peu stringente. Il réalise ensuite une suite d'itérations afin de maximiser le nombre de protéines validées pour un FDR souhaité. Les différentes étapes de l'algorithme sont résumées ci-dessous :

p-value_départ = 1
1) p-value = 0.9 * ancienne_p_value
2) calcul des seuils peptidiques d'identité et d'homologie
3) calcul des scores protéiques avec la fonction définie précédemment
4) recherche du FDR souhaité en fonction d'un seuil de score protéique (0 initialement)
4.1) calcul du nombre de protéines validées et du FDR associé
4.2) augmentation du seuil de score protéique, réitération de l'étape 4.1 jusqu'au FDR attendu
5) fin lorsque le seuil de score protéique optimal est trouvé (maximum de protéines validées pour le FDR attendu)

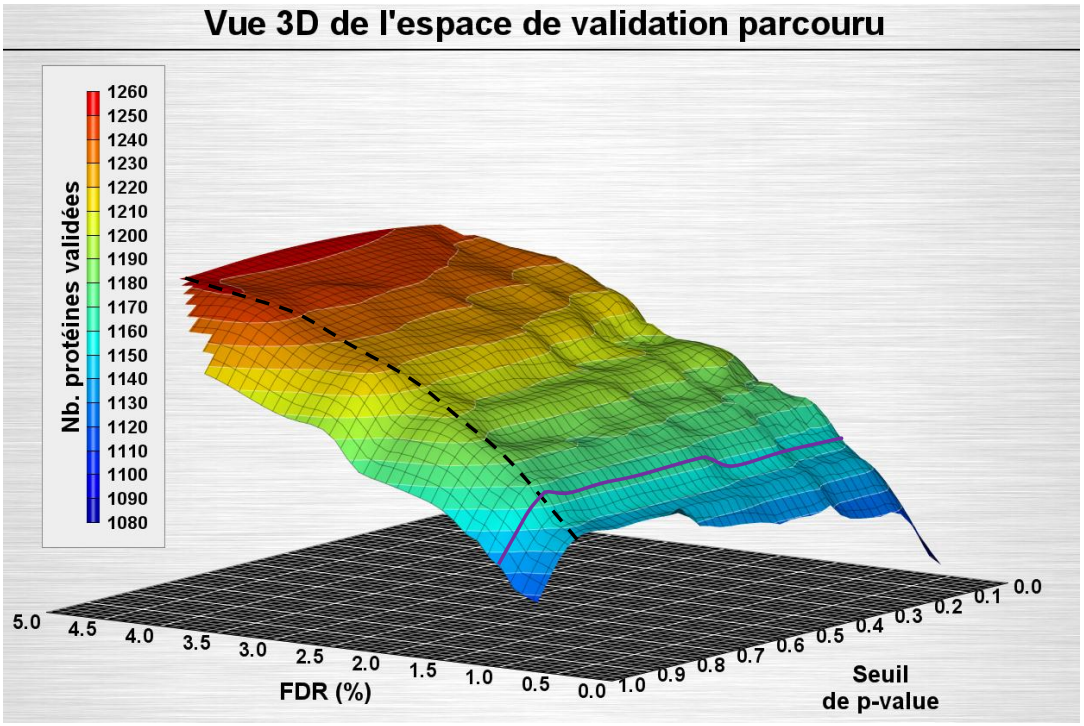


Figure 32 : ce graphique en trois dimensions résume les valeurs calculées par le programme sur l'ensemble des itérations de validation protéique d'un jeu donné choisi à titre d'exemple (protéome total de cellules HUVEC analysé par un LTQ-Orbitrap Velos). Le trait en pointillé et noir correspond au nombre de protéines validées en fonction de différents FDR protéiques pour une même valeur de p-value. Le trait mauve correspond à tous les cas de p-value où un FDR de 1% a été obtenu. L'intersection de ces deux courbes est la valeur recherchée par l'algorithme. Elle est obtenue dans cet exemple pour une p-value de 0.9 et un seuil de score protéique égal à 21 (non visible sur le graphe).

Afin d'illustrer cette procédure, la figure 32 montre la représentation en 3D des résultats obtenus à partir d'un fichier résultat Mascot, après validation des peptides avec un FDR de 5%, et application de l'algorithme ci-dessus. Cet exemple montre que l'on obtient des résultats de validation plutôt similaires pour une majorité des p-values utilisées pour calculer les seuils peptidiques (notamment entre 0.9 et 0.1). L'optimisation effectuée permet cependant de «gagner» une vingtaine de protéines supplémentaires : 1162 validées avec une p-value de 0.9 pour un FDR de 1% (correspondant à un seuil de score protéique de 21) contre 1143 pour une valeur plus classique de 0.05 et un FDR de 1% (correspondant à un seuil de score protéique de 9.2).

4.2.2 Comparaison entre les méthodes de validation de MFPaQ et Prosper

Afin de comparer les performances relatives d'une validation protéique basée sur des règles de score de peptides et d'une validation basée sur le calcul d'un score protéique, j'ai appliqué au jeu de donné utilisé dans l'exemple précédent ces deux différentes stratégies. J'ai ensuite calculé une courbe ROC (« Receiver Operator Characteristic ») pour chacune des stratégies (cf figure 33). Ces courbes montrent l'évolution de la sensibilité, i.e. le nombre protéines validées, en fonction de la spécificité, représentée ici par le FDR.

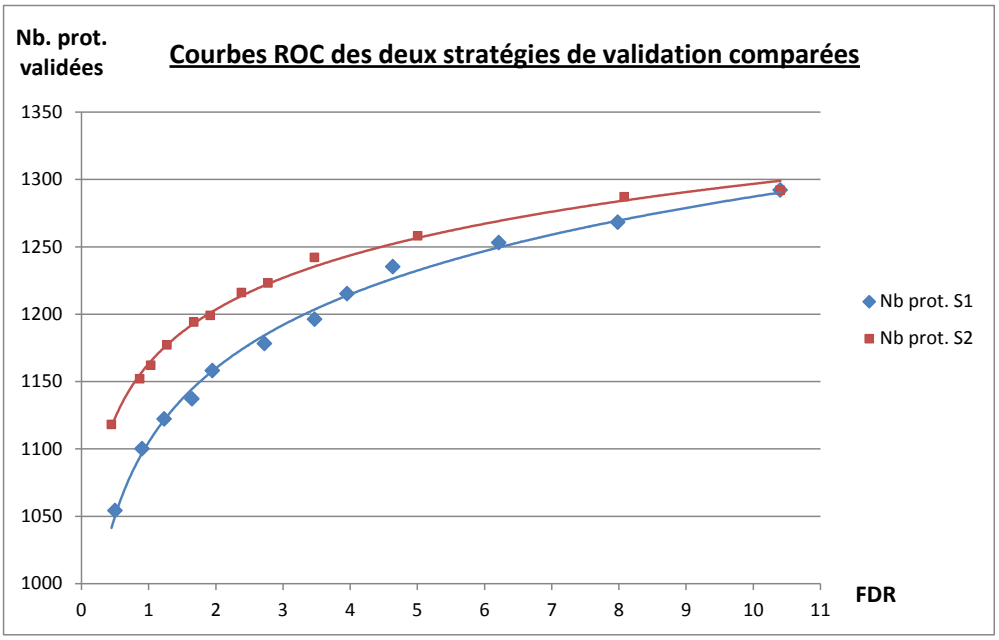


Figure 33 : courbes ROC pour les deux stratégies de validation des protéines. La stratégie S1 en bleu correspond à celle qui avait été implémentée à l'origine dans MFPaQ (règles basées sur les scores des peptides) mais mise en œuvre ici dans Prosper après validation des peptides avec un FDR de 5%. La stratégie S2 correspond à la nouvelle validation de Prosper basée sur le calcul d'un score protéique. On constate une légère supériorité de la nouvelle stratégie par rapport à l'ancienne : pour un même FDR on valide d'avantage de protéines avec la seconde stratégie.

La courbe en rouge correspond en fait à la courbe en pointillé dans la figure 32, c'est-à-dire la meilleure courbe ROC obtenu pour l'ensemble de l'espace de validation parcouru. Cette étape d'optimisation et le calcul d'un score protéique ont permis, comme on peut le voir dans la figure ci-dessus, d'améliorer le ratio sensibilité/spécificité par rapport à l'ancienne méthode utilisée dans MFPaQ. Je pense néanmoins qu'il est encore possible d'améliorer cette méthodologie de validation des protéines en combinant les deux approches présentées. La fonction de score des protéines

utilisée dans la stratégie 2 est identique pour toutes les protéines quel que soit le nombre de peptides identifiés. Il est possible qu'en définissant plusieurs fonctions de score optimisées pour le nombre de peptides attribués à une protéine donnée on puisse améliorer notre rapport sensibilité/spécificité. On pourrait imaginer par exemple une fonction de score spécifique des « one-hit-wonders » et une autre spécifique des « multi-hit-wonders ». On aurait donc au final une combinaison de règles non plus basées sur des scores de peptides mais des scores de protéines. L'implémentation de cette stratégie sera sûrement réalisée dans une version ultérieure de Prosper.

En conclusion, les optimisations apportées dans Prosper, que nous avons détaillées dans les paragraphes précédents (validation peptidique avec E-values ajustées en fonction des seuils d'identité/homologie, calcul du FDR peptidique sur la base de la méthode de Navarro, et validation des protéines sur la base de la fonction de score protéique), ont permis d'améliorer de façon significative l'étape de validation. On observe généralement une augmentation de 15 à 25% le nombre de protéines identifiées par rapport à MFPaQ, pour un même FDR protéique de 1%.

IV-3. LCMSdb et algorithmes de quantification

4.3.1 Importation

Nous avons vu dans la partie introductive qu'un grand nombre d'outils existe pour générer des cartes LC-MS. Cependant, jusqu'à ce jour on ne disposait pas de solution logicielle permettant d'intégrer et de comparer des données générées par des outils différents. Il était donc très fastidieux d'évaluer les performances relatives des solutions de « peak picking » disponibles. Prosper a été conçu dès le départ pour gérer ce type de problématique fournissant ainsi une grande flexibilité dans les « workflows » d'analyse de données LC-MS. En effet, même si toute la chaîne de traitement (alignement, normalisation, comparaison) des cartes fait partie intégrante du logiciel il est néanmoins possible de choisir la source des données à utiliser (cf figure 34).

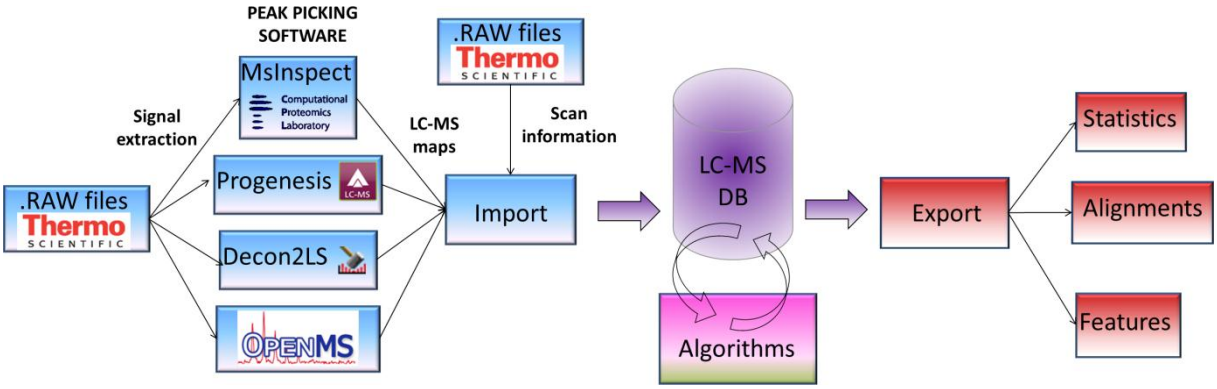


Figure 34 : vue d'ensemble du « workflow » LC-MS de Prosper. Des sources hétérogènes de cartes LC-MS peuvent être importées dans la base de données. Une fois les cartes chargées, celles-ci sont prises en charge par une série d'algorithmes qui effectuent le traitement et la comparaison des données. Le résultat de l'analyse peut enfin être exporté sous différents types de fichiers.

4.3.2 Clustering

Les cartes générées par les logiciels de « peak picking » ne sont pas toujours fiables à 100% et présentent fréquemment des signaux redondants, c'est-à-dire correspondants au même composé.

L'opération de « clustering » permet de supprimer cette redondance (cf figure 35). Il ne faut pas confondre cette étape avec celle de déconvolution consistant à regrouper tous les états de charge détectés pour une même molécule donnée.

En sortie de cet algorithme, les cartes sont plus « propres » permettant ainsi de réduire des problèmes d'ambiguïté au niveau des étapes d'alignement et de comparaison des cartes.

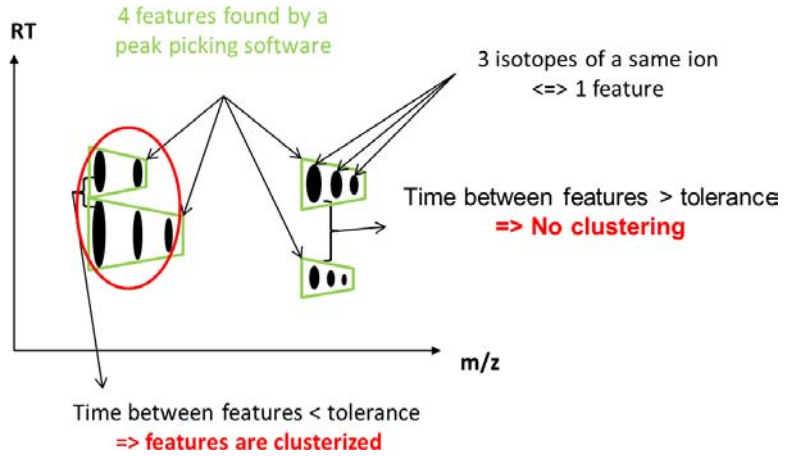


Figure 35 : regroupement des signaux redondants. Les signaux (« features ») avec le même état de charge, le même rapport m/z et qui sont à proximité dans l'échelle du temps (RT) sont groupés en une même espèce (« feature cluster »). Les signaux trop éloignées en temps sont conservés tels quels.

4.3.2 Alignement

Pour corriger les défauts de reproductibilité des séparations chromatographiques il est nécessaire d'aligner les cartes LC-MS à comparer. L'algorithme d'alignement de Prosper consiste à sélectionner une carte de référence au hasard puis à comparer chaque carte par rapport à cette carte de référence. Ensuite pour chaque paire analysée le programme détermine toutes les correspondances possibles entre les espèces détectées sur ces deux cartes pour des fenêtres de temps et de masse données (600 secondes et 10 ppm par défaut). Le résultat de cette procédure peut être visualisé sous la forme d'un nuage de points (cf figure 36) mais ne constitue pas l'alignement définitif. En effet, la dernière étape consiste à trouver le chemin qui parcourt les régions les plus denses de ce nuage de point. Cette dernière étape a été implémentée sous la forme d'un lissage (« smoothing ») basé sur l'utilisation d'une médiane mobile.

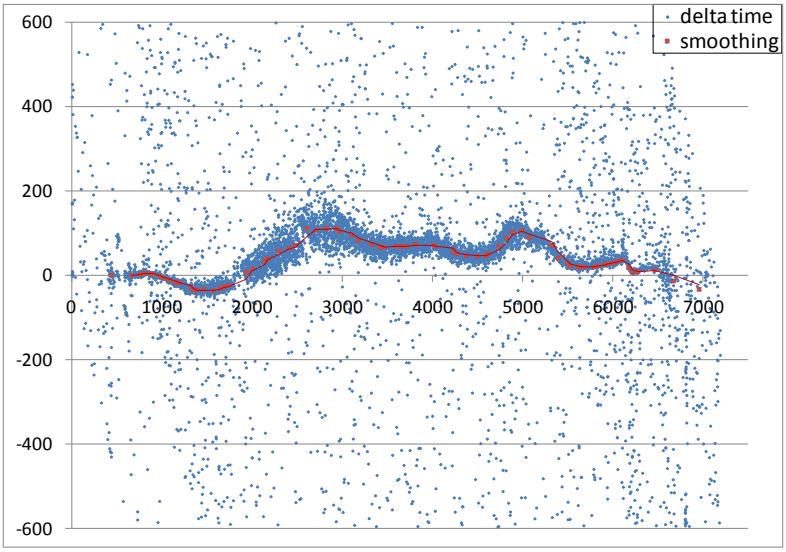


Figure 36 : alignement d'une carte par rapport à une carte de référence. Le nuage de point représente la variation temporelle (en secondes) de plusieurs points de référence ou « landmark » (entre la carte comparée et la carte de référence) en fonction du temps observé (en secondes) dans la carte de référence.

Le programme réalise cette opération d'alignement plusieurs fois en changeant de façon aléatoire la carte de référence. Enfin il résume l'information accumulée sur l'ensemble des alignements en calculant la somme des distances absolues pour chaque carte par rapports aux différentes références testées. La carte présentant la somme la plus faible est considérée comme étant la « plus proche » des autres cartes, ou autrement dit, la carte médiane. Celle-ci est considérée comme la carte de référence définitive et sera utilisée comme telle pour la suite des opérations.

Nous avons évalué le gain apporté par les alignements des cartes en comparant les écart-types des temps d'élution bruts et corrigés pour trois injections du même échantillon. Pour cela deux cartes consensus ont été créées à partir des cartes LC-MS correspondant à chaque injection (cartes générées avec le logiciel MsInspect). La première a été créée sans réaliser la procédure d'alignement, contrairement à la seconde. Pour chaque « feature » de chaque « master map » l'écart-type des temps d'élution des occurrences observées sur les différentes cartes a été calculé. Les résultats de cette comparaison sont présentés dans la figure suivante :

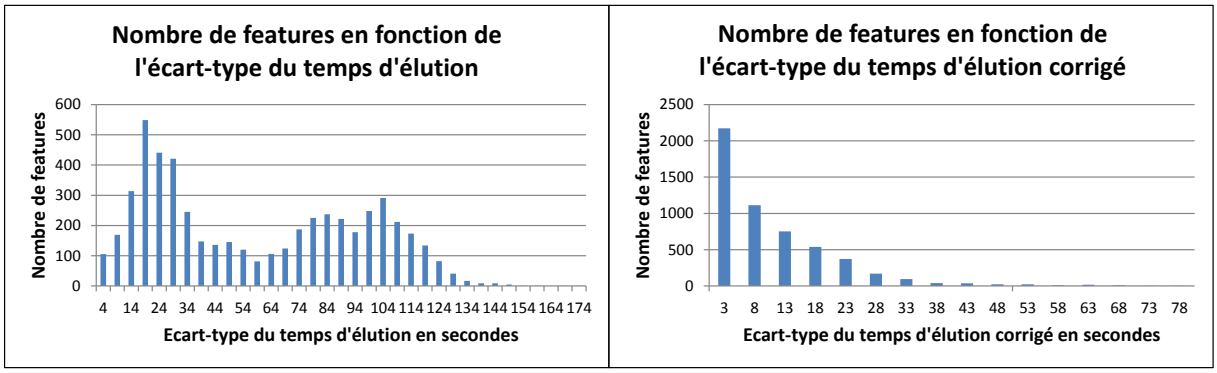


Figure 37 : comparaison de l'écart-type des temps d'élution avant (à gauche) et après correction (à droite). Après correction il n'y a pas d'espèce appariée avec un écart-type supérieur à 80 secondes et l'on remarque également que la majorité des espèces ont un écart-type associé inférieur à 30 secondes.

4.3.2 Normalisation

La deuxième source d'erreur lors de la comparaison de cartes LC-MS est la variabilité des mesures des signaux MS réalisées par l'instrument. On peut distinguer deux grandes catégories de variabilité : biologique et technique. Sur le plan technique, les variations observées pour les signaux MS entre deux analyses peuvent être dépendantes de la quantité de matériel chargée, de la reproductibilité de la configuration instrumentale utilisée mais également du programme informatique réalisant le traitement du signal. Les biais systématiques observés sur les mesures d'intensité entre deux acquisitions successives d'échantillons similaires sont cependant dus principalement à des erreurs sur les quantités totales de matériel injecté dans chaque cas, ou à l'instabilité du système instrumental nanoLC-MS qui peut présenter des performances variables au cours d'une série d'analyse, occasionnant une réponse différente au niveau du signal MS enregistré pour des ions peptidiques d'abondance identique. Lorsque ces différences sont trop importantes les données peuvent ne pas être exploitables et il est d'ailleurs conseillé d'effectuer un contrôle qualité des acquisitions pour s'assurer de la faisabilité de l'analyse informatique. Cependant, ces biais systématiques sont inhérents à toute mesure analytique, et dans la plupart des cas la mise en œuvre d'une normalisation des signaux permet de corriger suffisamment ce type de biais.

De nombreuses méthodes de normalisation ont été décrites dans la littérature, chacune d'elle utilisant une méthode mathématique différente pour égaliser l'information (Christin, Bischoff et al. 2011). On distingue principalement des méthodes de calcul non linéaires et linéaires, et il a été démontré que dans la plupart des cas, les méthodes linéaires corrigent suffisamment les biais systématiques (Callister, Barry et al. 2006). Dans Prosper, nous avons réalisé 3 implémentations différentes de correction linéaire, en calculant les facteurs de normalisation comme le ratio de la somme des intensités, le ratio de la médiane des intensités, ou comme la médiane des ratios des intensités. Cette dernière stratégie qui avait été publié en 2006 (Dieterle, Ross et al. 2006) est celle qui nous a donné les meilleurs résultats. Elle consiste à calculer les ratios des intensités entre deux cartes à comparer puis à fixer le facteur de normalisation comme l'inverse de la médiane de ces ratios (cf figure 38). Prosper réalise ce calcul pour chaque appariement avec la carte de référence et dispose ainsi d'un facteur de correction pour chacune des cartes, le facteur de la carte de référence étant égal à 1.

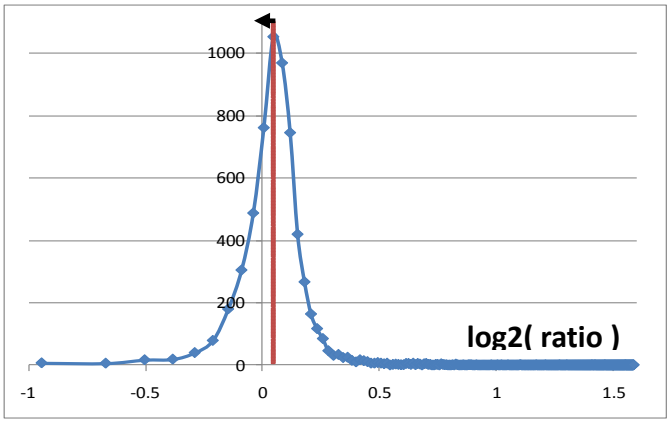


Figure 38 : distribution des ratios transformés en log2 et calculés à partir des intensités d'espèces observées sur deux cartes LC-MS. La barre verticale rouge montre que la médiane de la distribution est légèrement décentrée. La valeur du facteur de normalisation est égale à l'inverse de cette valeur médiane. La flèche en noir illustre le processus de normalisation qui consiste à recentrer la distribution des ratios sur 0.

4.3.3 Comparaison des cartes : création d'une « master map »

Une fois que les cartes ont été corrigées et alignées, la dernière étape consiste à créer une carte consensus ou « master map ». Pour cela, le programme essaye de trouver la meilleure correspondance possible entre les espèces détectées sur les différentes cartes. La carte consensus peut être assimilée à une représentation non redondante de l'ensemble des espèces détectées sur les cartes comparées (cf figure 39).

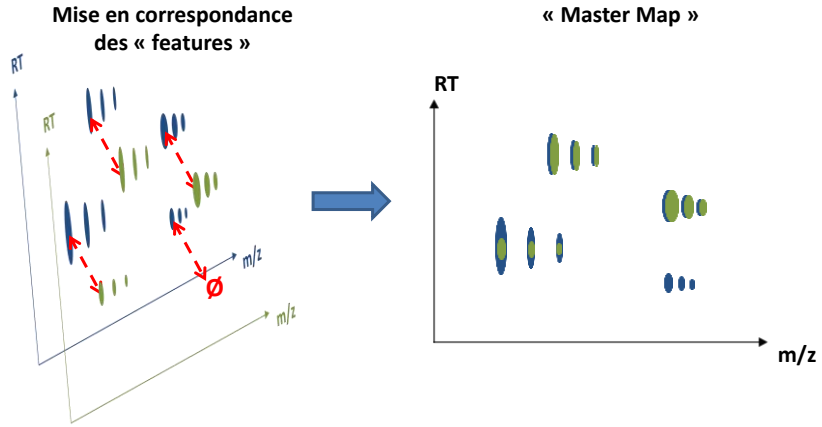


Figure 39 : création de la « master map » par une mise en correspondance des « features » détectées sur deux cartes LC-MS différentes. Les temps d'élution utilisés sont ceux qui ont été corrigés à partir du résultat de l'alignement des cartes. On remarque que certaines espèces peuvent voir leur intensité varier d'une carte à l'autre et même parfois n'être présentes que dans une des cartes.

Dans le but d'éviter d'inclure du bruit de fond dans la carte consensus, le logiciel réalise d'abord une correspondance entre les espèces les plus intenses (avec une intensité supérieure à un certain seuil), puis introduit les valeurs les plus faibles uniquement si elles correspondent à une espèce de haute intensité dans une des cartes comparées (cf figure 40).

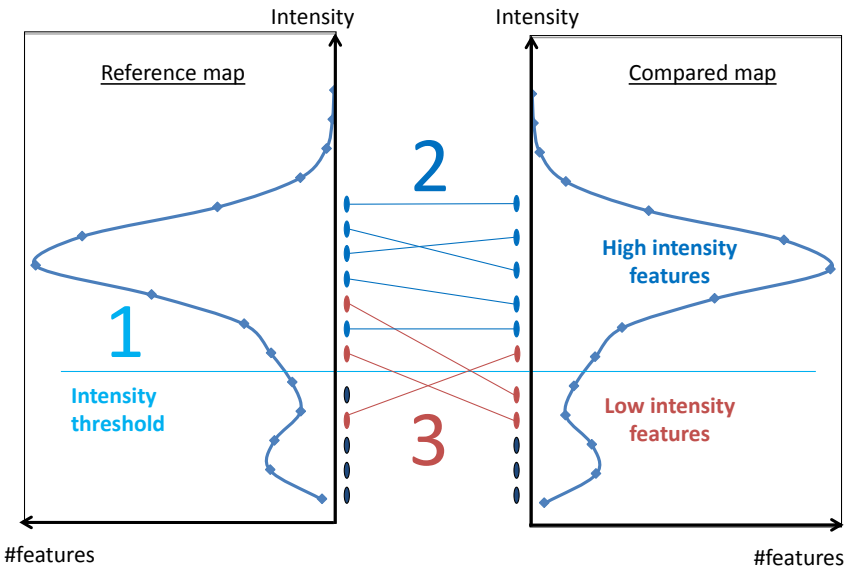


Figure 40 : distribution des intensités des cartes utilisées pour construire une « master map ». La construction est réalisée en 3 étapes : 1) élimination des espèces qui ont une intensité normalisée inférieure à une valeur seuil 2) mise en correspondance des espèces les plus intenses 3) les espèces pour lesquelles il manque des attributions (valeurs manquantes) dans au moins une des cartes sont à nouveau comparées avec les espèces de faible intensité, mises de côté lors de la première étape.

4.3.4 Application : évaluation et comparaison d'outils de « peak picking »

Dans le but de comparer quatre logiciels d'extraction de pics (MsInspect, Progenesis LC-MS, Decon2LS and OpenMS), nous avons réalisé 3 réplicats d'injection LC-MS/MS d'un lysat total de cellules HUVEC (Human Umbilical Vein Endothelial Cells) sur un LTQ-Orbitrap Velos. Les fichiers d'acquisition ont été analysés par chacun des outils afin d'obtenir des cartes LC-MS. La base de données LCMSdb ainsi que les algorithmes présentés ci-dessus ont permis d'évaluer les performances de ces différents logiciels sur les plans de l'extraction du signal ainsi que sur celui de la reproductibilité de quantification. Les calculs ont été exécutés en utilisant les mêmes paramètres pour chacun des logiciels :

- « feature clustering » : tolérance m/z = 10 ppm et tolérance de temps = 60 secondes
- alignement : tolérance m/z = 10 ppm et tolérance de temps = 600 secondes
- « feature mapping » : seuil d'intensité = 35% de l'intensité médiane, tolérance m/z = 10 ppm et tolérance de temps = 60 secondes

Nous avons tout d'abord réalisé des statistiques liées à la détection des espèces pour chacun des outils considérés (cf figure 41).

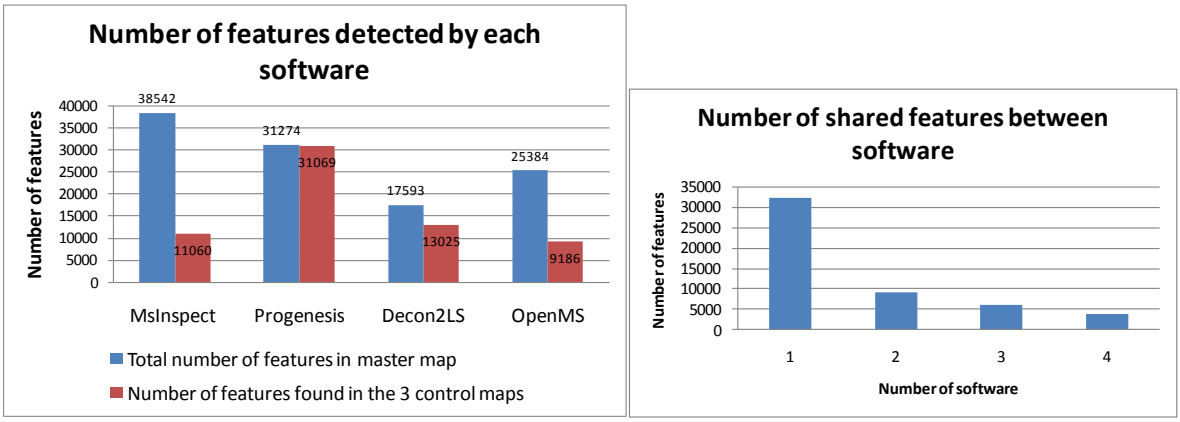


Figure 41 : comparaison du nombre d'espèces détectées pour les différents logiciels évalués.
L'histogramme de gauche montre les statistiques déterminées pour chaque outil individuel. Progenesis est capable d'avoir une correspondance dans chacune des cartes pour quasiment toutes les espèces de la « master map ». Par contre les autres logiciels génèrent un nombre plus ou moins élevé de valeurs manquantes, MsInspect étant l'exemple le plus frappant.
A droite est présentée une statistique croisée sur les différents logiciels. On constate que la plupart des espèces (63%) sont spécifiques d'un logiciel donné et que relativement peu d'espèces (moins de 5000) sont retrouvées dans les quatre logiciels.

L'analyse comparative des cartes générées par ces différents outils montre qu'il existe une grande hétérogénéité des propriétés (m/z, RT, charge) attribuées aux espèces détectées. Nous avons par ailleurs cherché à comparer la distribution des états de charge attribués aux espèces détectées.

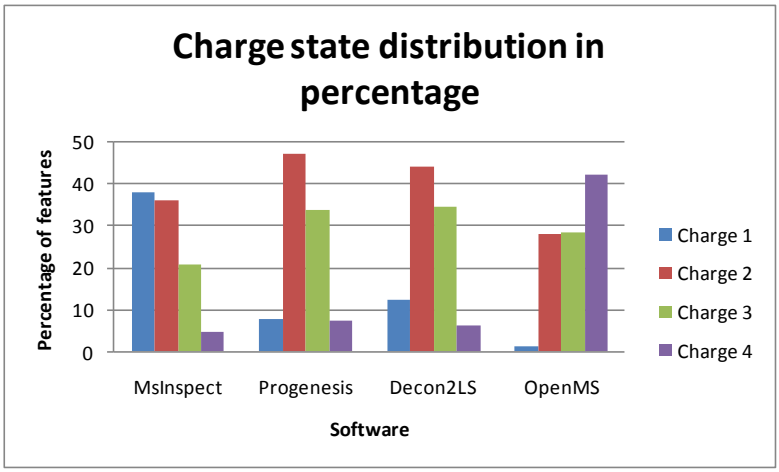


Figure 42 : distribution relative des états de charge des espèces détectées avec chaque outil de « peak picking ». Progenesis et Decon2LS génèrent des distributions assez similaires. MsInspect affecte la majeure partie de ses espèces à des charges 1+ et 2+ tandis qu'OpenMS assigne une majorité de 4+.

Nous constatons une fois de plus le caractère hétérogène du contenu des cartes LC-MS générées par ces différents outils. Les résultats actuels ne reflètent cependant pas la qualité et la reproductibilité quantitative des données. Afin d'en avoir une idée plus précise nous avons comparé les intensités extraites sur les réplicats d'injection par chacun des logiciels. Nous avons ensuite calculé le coefficient de variation des intensités présentes au niveau de chaque « master map ». L'analyse de la distribution de ces coefficients de variations est présentée dans les figures ci-dessous.

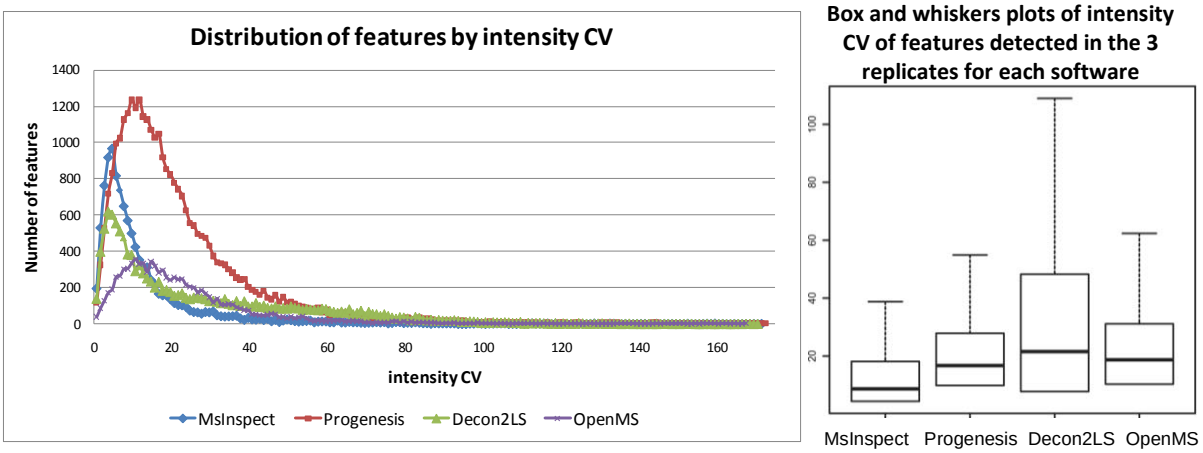


Figure 43 : les distributions des coefficients de variation (CVs) reflètent la reproductibilité de quantification de chaque logiciel. MsInspect et Decon2LS se classent respectivement premier et dernier de cette comparaison. Progenesis et OpenMS présentent des résultats équivalents et se situent entre les deux autres sur le plan de la reproductibilité.

En conclusion, Prosper est capable de comparer des données LC-MS générées par différents outils de « peak picking » ce qui était jusqu'à aujourd'hui une opération non triviale. Chaque logiciel diffère des autres d'un point de vue qualitatif au niveau de la nature des espèces détectées sur les cartes LC-MS, mais également au niveau de la reproductibilité quantitative des intensités extraites. Les résultats présentés ne permettent pas de définir l'outil idéal mais on peut tout de même constater qu'en fonction du logiciel utilisé le résultat final de quantification est très différent.

L'analyse de la spécificité et de la sensibilité de ces outils (détection et classification correcte des vrais positifs *versus* faux positifs dans le cadre d'une analyse différentielle) n'a pas été abordée ici mais elle pourrait apporter une aide précieuse en ce qui concerne la détermination de l'outil de quantification idéal, l'optimisation de leur paramétrage, ou bien permettre la mise au point d'un algorithme d'extraction du signal optimal.

Perspectives

La complexité et la quantité des données générées en utilisant les spectromètres de masse modernes sont beaucoup plus importantes qu'auparavant, rendant ainsi plus difficile le développement de logiciels efficaces et robustes. De nombreuses étapes informatiques sont nécessaires pour réduire l'énorme volume de données brutes en informations pertinentes. Ces opérations de transformation et d'analyse des données reposent sur des algorithmes complexes exécutés en série et mettant en œuvre différentes méthodes mathématiques et statistiques (analyse du signal, modélisation des distributions des observations, calcul de probabilités, algorithmes de classification et de regroupement, analyse de variances…). Chacune de ces étapes repose sur des approximations (propres aux algorithmes ou liées aux paramètres définis par l'utilisateur), introduisant des biais systématiques dans le processus d'interprétation des données. En conséquence, les résultats fournis par les outils bioinformatiques existants ne correspondent pas à l'interprétation optimale qui pourrait être réalisée. Pour y parvenir il est encore nécessaire de mettre au point de nouvelles méthodes mais également d'affiner et d'améliorer les méthodes existantes.

La version actuelle du logiciel Prosper a été très utile pour tester, modifier et valider les modèles de bases de données que nous avons mis au point et a également servi de laboratoire d'essai pour l'écriture de nouveaux algorithmes. Le développement de ce prototype nous a également permis de mettre en évidence de nouvelles problématiques liées au traitement des données, notamment en ce qui concerne l'analyse des cartes LC-MS. Un des principaux problèmes des logiciels de « peak picking » existants est lié à la présence de données manquantes lors de la comparaison de cartes LC-MS. En effet, les logiciels « open-source » cités ci-dessus effectuent tous une analyse individuelle sur chacune des acquisitions réalisées, puis essayent ensuite de regrouper les espèces quantifiées dans une liste non redondante constituant une carte consensus. Cependant, de nombreux facteurs peuvent induire une variabilité du signal MS observé : fluctuation de la sensibilité de l'instrument, modification de l'environnement du signal MS pour un peptide donné, surtout lors de la comparaison d'échantillons différents. De plus chaque logiciel génère les cartes LC-MS avec certains paramètres de seuil afin d'exclure le bruit de fond. Ainsi un peptide qui a pu être détecté dans un premier échantillon peut ne plus l'être dans un second échantillon si son intensité chute légèrement en dessous de ce seuil. Bien que certaines méthodes statistiques (Clough, Braun et al. 2011) puissent être mises en œuvre pour remplacer ces données manquantes par des valeurs estimées (lorsque le même échantillon a été analysé plusieurs fois), il est cependant préférable de minimiser ce problème en amont lorsque cela est possible. Le logiciel commercial Progenesis LCMS apporte une solution à ce problème en considérant l'ensemble des échantillons analysés au moment de l'extraction du signal. Il réalise ainsi une superposition et un alignement des données brutes, puis réalise son processus d'extraction de façon transversale sur l'ensemble des acquisitions LC-MS ainsi superposées, évitant de cette manière la présence de valeurs manquantes.

Un autre défi pour l'analyse quantitative de données MS est la conversion des abondances peptidiques en abondances protéiques. Premièrement, il est tout d'abord nécessaire de relier correctement les données d'identification aux données de quantification. Deuxièmement, il faut

également disposer d'une méthode de calcul des abondances protéiques suffisamment robuste c'est-à-dire diminuant au mieux la variabilité quantitative lors du passage des peptides aux protéines. La disponibilité d'une valeur quantitative au niveau des protéines a un double avantage : premièrement elle permet de trier les protéines suivant une valeur reflétant leur abondance absolue, et deuxièmement il devient possible d'utiliser au niveau protéique des tests statistiques conventionnels comme le test de Student. Le calcul du PAI réalisé par MFPaQ est une tentative de réponse à cette problématique mais nous n'avons jamais pu évaluer si elle constituait la méthode la plus efficace. Depuis d'autres méthodes ont été décrites dans la littérature comme le calcul de l'indice iBAQ (Schwanhausser, Busse et al. 2011). Ce dernier correspond pour une protéine donnée à la somme intensités peptidiques observées divisée par le nombre de peptides observables. J'envisage donc dans la future version de Prosper d'implémenter différents algorithmes d'estimation d'indices d'abondances protéiques. Des standards protéiques tels que les « Universal Protein Standards » vendus par le fournisseur Sigma-Adrich pourront alors être utilisés pour déterminer la méthode la plus précise.

Nous avons évoqué deux problèmes majeurs liés à l'analyse des données quantitatives (les valeurs manquantes, le lien entre abondances peptidiques et protéiques) mais on peut également citer d'autres objectifs qui ne sont pas complètement atteints par les solutions existantes : gestion de grandes séries de fichiers, gestion du fractionnement de l'échantillon, automatisation de la chaîne de traitement. Enfin le temps d'analyse est également un facteur important à prendre en compte. En effet, suivant le nombre d'échantillons à comparer, ce temps peut passer de plusieurs heures à plusieurs jours.

Ainsi au-delà de l'amélioration du résultat de quantification il est encore nécessaire d'optimiser la vitesse d'exécution des algorithmes mis en œuvre. Dans le but d'accélérer la phase d'extraction du signal j'ai commencé à mettre au point un nouveau format de fichier (.mzDB) disposant d'un système d'indexation optimisé pour l'analyse de données LC-MS. Différents standards de stockage de données brutes basés sur le format XML, tel que mzXML (Pedrioli, Eng et al. 2004) et mzML (Martens, Chambers et al. 2011), ont été développés par la communauté protéomique. Ces formats ont simplifié l'échange des données entre les différents laboratoires, mais certains auteurs ont démontré (Lin, Zhu et al. 2005; Askenazi, Parikh et al. 2009; Shah, Davidson et al. 2010) qu'ils ne constituaient pas la solution la plus efficace pour un traitement à haut-débit des données tel que l'extraction du signal MS dans des approches quantitatives sans marquage. Afin de diminuer le temps d'accès aux données brutes, j'ai donc développé ce nouveau format mzDB qui repose sur l'utilisation du format générique SQLite avec des index optimisés pour la lecture de données LC-MS. Ce projet est effectué en collaboration avec Sara Nasso (IMLS, Université de Zurich) qui a démontré dans une étude préliminaire (Nasso, Silvestri et al. 2010) l'intérêt d'un index de type R*Tree pour diminuer les temps d'accès aux données MS. Très récemment une initiative équivalente a été entreprise par Wilhelm *et al*. (Wilhelm, Kirchner et al. 2012), prouvant ainsi qu'il y a un réel besoin associé à la mise à disposition d'un format de fichier efficace pour l'accès aux données brutes. Contrairement à notre solution, leur implémentation est basée sur le format HDF5 (Poinot 2010) qui utilise un modèle de représentation hiérarchique des données au lieu d'un modèle relationnel. En pratique, l'utilisation du format SQLite est réellement plus simple que celle du format HDF5 car il est possible d'accéder aux données en lecture et en écriture via des requêtes SQL. De plus, l'utilisation d'un modèle relationnel permet de mettre en œuvre différentes stratégies d'indexation des données. Ainsi dans notre implémentation mzDB nous avons intégré trois indexations différentes des données MS :

- une dans la dimension temps (i.e. dans le sens d'acquisition de données) afin de lire efficacement les spectres MS,
- une autre dans la dimension m/z, de manière à charger rapidement des plages de masse pour toute la durée du gradient chromatographique,
- une dernière dans les deux dimensions à la fois, via l'utilisation du module R*Tree supporté nativement par SQLite, afin de récupérer des régions entières de pics en un temps très court.

Nous envisageons de comparer les performances relatives des systèmes mz5 et mzDB afin de déterminer la viabilité de notre projet sur le long terme. Sur le plan des performances et de la qualité d'extraction des intensités nous avons déjà obtenu des résultats préliminaires très encourageants : environ 30000 « features » extraites en moins de 3 minutes sur un ensemble de 3 fichiers LTQ-Orbitrap Velos. Ces travaux ont été présentés lors du dixième congrès international HUPO organisé à Genève en septembre 2011 (MZDB: AN OPTIMIZED FILE FORMAT FOR THE EFFICIENT ANALYSIS OF LC-MS DATASETS). Avec ce nouveau format l'analyse des données brutes d'un échantillon passera de plusieurs dizaines de minutes à quelques secondes. Il sera donc possible de comparer un grand nombre de processus analytiques et d'outils différents en un temps réduit. La partie suivante de ce projet va consister à finaliser les spécifications du format de façon à intégrer autant d'informations que celles représentées dans le format mzML.

Enfin, dans le cadre du projet ProFI, de nombreux développements vont être réalisés au sein du laboratoire dans le but de fournir un environnement dédié au stockage et au traitement des données de spectrométrie de masse. Ce nouveau logiciel qui a été nommé « Proline » va mobiliser l'ensemble des développeurs informatiques et des bioinformaticiens présents sur les différents sites partenaires. Il devrait permettre de combler les imperfections des solutions existantes ainsi que de répondre aux besoins actuels et futurs du traitement des données protéomiques en tirant parti de nos expériences passées, qui ont consistées pour ma part à la réalisation de ce travail de thèse.

Pour conclure, je suis persuadé que l'amélioration continue de la pertinence et de la qualité des informations extraites à partir des données de spectrométrie de masse sera une aide indispensable dans la recherche de réponses à des problématiques biologiques étudiées par des stratégies d'analyse protéomique.

Bibliographie

Alves, P., R. J. Arnold, M. V. Novotny, P. Radivojac, J. P. Reilly and H. Tang (2007). "Advancement in protein inference from shotgun proteomics using peptide detectability." Pac Symp Biocomput: 409-420.

Angel, T. E., U. K. Aryal, S. M. Hengel, E. S. Baker, R. T. Kelly, E. W. Robinson and R. D. Smith (2012). "Mass spectrometry-based proteomics: existing capabilities and future directions." Chem Soc Rev **41**(10): 3912-3928.

Annesley, T. M. (2003). "Ion suppression in mass spectrometry." Clin Chem **49**(7): 1041-1044.

Askenazi, M., J. R. Parikh and J. A. Marto (2009). "mzAPI: a new strategy for efficiently sharing mass spectrometry data." Nat Methods **6**(4): 240-241.

Bellew, M., M. Coram, M. Fitzgibbon, M. Igra, T. Randolph, P. Wang, D. May, J. Eng, R. Fang, C. Lin, J. Chen, D. Goodlett, J. Whiteaker, A. Paulovich and M. McIntosh (2006). "A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS." Bioinformatics **22**(15): 1902-1909.

Benjamini, Y. and Y. Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing." J. Roy. Statist. Soc. Ser. B **57**(1): 289--300.

Bildl, W., A. Haupt, C. S. Muller, M. L. Biniossek, J. O. Thumfart, B. Huber, B. Fakler and U. Schulte (2012). "Extending the dynamic range of label-free mass spectrometric quantification of affinity purifications." Mol Cell Proteomics **11**(2): M111 007955.

Blagoev, B., I. Kratchmarova, S. E. Ong, M. Nielsen, L. J. Foster and M. Mann (2003). "A proteomics strategy to elucidate functional protein-protein interactions applied to EGF signaling." Nat Biotechnol **21**(3): 315-318.

Bleasby, A. J., D. Akrigg and T. K. Attwood (1994). "OWL--a non-redundant composite protein sequence database." Nucleic Acids Res **22**(17): 3574-3577.

Bodenmiller, B., S. Wanka, C. Kraft, J. Urban, D. Campbell, P. G. Pedrioli, B. Gerrits, P. Picotti, H. Lam, O. Vitek, M. Y. Brusniak, B. Roschitzki, C. Zhang, K. M. Shokat, R. Schlapbach, A. Colman-Lerner, G. P. Nolan, A. I. Nesvizhskii, M. Peter, R. Loewith, C. von Mering and R. Aebersold (2010). "Phosphoproteomic analysis reveals interconnected system-wide responses to perturbations of kinases and phosphatases in yeast." Sci Signal **3**(153): rs4.

Bousquet-Dubouch, M. P., S. Nguen, D. Bouyssie, O. Burlet-Schiltz, S. W. French, B. Monsarrat and F. Bardag-Gorce (2009). "Chronic ethanol feeding affects proteasome-interacting proteins." Proteomics **9**(13): 3609-3622.

Bouyssie, D., A. Gonzalez de Peredo, E. Mouton, R. Albigot, L. Roussel, N. Ortega, C. Cayrol, O. Burlet-Schiltz, J. P. Girard and B. Monsarrat (2007). "Mascot file parsing and quantification (MFPaQ), a new software to parse, validate, and quantify proteomics data generated by ICAT and SILAC mass spectrometric analyses: application to the proteomics study of membrane proteins from primary human endothelial cells." Mol Cell Proteomics **6**(9): 1621-1637.

Burande, C. F., M. L. Heuze, I. Lamsoul, B. Monsarrat, S. Uttenweiler-Joseph and P. G. Lutz (2009). "A label-free quantitative proteomics strategy to identify E3 ubiquitin ligase substrates targeted to proteasome degradation." Mol Cell Proteomics **8**(7): 1719-1727.

Callister, S. J., R. C. Barry, J. N. Adkins, E. T. Johnson, W. J. Qian, B. J. Webb-Robertson, R. D. Smith and M. S. Lipton (2006). "Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics." J Proteome Res **5**(2): 277-286.

Carr, S., R. Aebersold, M. Baldwin, A. Burlingame, K. Clauser and A. Nesvizhskii (2004). "The need for guidelines in publication of peptide and protein identification data: Working Group on Publication Guidelines for Peptide and Protein Identification Data." Mol Cell Proteomics **3**(6): 531-533.

Choi, H. and A. I. Nesvizhskii (2008). "Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics." J Proteome Res **7**(1): 254-265.

Christin, C., R. Bischoff and P. Horvatovich (2011). "Data processing pipelines for comprehensive profiling of proteomics samples by label-free LC-MS for biomarker discovery." Talanta **83**(4): 1209-1224.

Claassen, M., L. Reiter, M. O. Hengartner, J. M. Buhmann and R. Aebersold (2011). "Generic comparison of protein inference engines." Mol Cell Proteomics.

Clauser, K. R., P. Baker and A. L. Burlingame (1999). "Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching." Anal Chem **71**(14): 2871-2882.

Clough, T., S. Braun, V. Fokin, I. Ott, S. Ragg, G. Schadow and O. Vitek (2011). "Statistical design and analysis of label-free LC-MS proteomic experiments: a case study of coronary artery disease." Methods Mol Biol **728**: 293-319.

Colinge, J., A. Masselot, M. Giron, T. Dessingy and J. Magnin (2003). "OLAV: towards high-throughput tandem mass spectrometry data identification." Proteomics **3**(8): 1454-1463.

Cox, J. and M. Mann (2008). "MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification." Nat Biotechnol **26**(12): 1367-1372.

Craig, R. and R. C. Beavis (2004). "TANDEM: matching proteins with tandem mass spectra." Bioinformatics **20**(9): 1466-1467.

Csete, M. E. and J. C. Doyle (2002). "Reverse engineering of biological complexity." Science **295**(5560): 1664-1669.

de Godoy, L. M., J. V. Olsen, G. A. de Souza, G. Li, P. Mortensen and M. Mann (2006). "Status of complete proteome analysis by mass spectrometry: SILAC labeled yeast as a model system." Genome Biol **7**(6): R50.

Dengjel, J., A. R. Kristensen and J. S. Andersen (2008). "Ordered bulk degradation via autophagy." Autophagy **4**(8).

Dieterle, F., A. Ross, G. Schlotterbeck and H. Senn (2006). "Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics." Anal Chem **78**(13): 4281-4290.

Domon, B. and R. Aebersold (2006). "Mass spectrometry and protein analysis." Science **312**(5771): 212-217.

Dupierris, V., C. Masselon, M. Court, S. Kieffer-Jaquinod and C. Bruley (2009). "A toolbox for validation of mass spectrometry peptides identification and generation of database: IRMa." Bioinformatics **25**(15): 1980-1981.

Elias, J. E. and S. P. Gygi (2007). "Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry." Nat Methods **4**(3): 207-214.

Elias, J. E., W. Haas, B. K. Faherty and S. P. Gygi (2005). "Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations." Nat Methods **2**(9): 667-675.

Eng, J., A. McCormack and J. Yates (1994). "An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database." J Am Soc Mass Spectrom **5**(11): 976-989.

Fenn, J. B., M. Mann, C. K. Meng, S. F. Wong and C. M. Whitehouse (1989). "Electrospray ionization for mass spectrometry of large biomolecules." Science **246**(4926): 64-71.

Florens, L., M. J. Carozza, S. K. Swanson, M. Fournier, M. K. Coleman, J. L. Workman and M. P. Washburn (2006). "Analyzing chromatin remodeling complexes using shotgun proteomics and normalized spectral abundance factors." Methods **40**(4): 303-311.

Geer, L. Y., S. P. Markey, J. A. Kowalak, L. Wagner, M. Xu, D. M. Maynard, X. Yang, W. Shi and S. H. Bryant (2004). "Open mass spectrometry search algorithm." J Proteome Res **3**(5): 958-964.

Gerster, S., E. Qeli, C. H. Ahrens and P. Buhlmann (2010). "Protein and gene model inference based on statistical modeling in k-partite graphs." Proc Natl Acad Sci U S A **107**(27): 12101-12106.

Graumann, J., N. C. Hubner, J. B. Kim, K. Ko, M. Moser, C. Kumar, J. Cox, H. Scholer and M. Mann (2008). "Stable isotope labeling by amino acids in cell culture (SILAC) and proteome quantitation of mouse embryonic stem cells to a depth of 5,111 proteins." Mol Cell Proteomics **7**(4): 672-683.

Gupta, N. and P. A. Pevzner (2009). "False discovery rates of protein identifications: a strike against the two-peptide rule." J Proteome Res **8**(9): 4173-4181.

Gygi, S. P., B. Rist, S. A. Gerber, F. Turecek, M. H. Gelb and R. Aebersold (1999). "Quantitative analysis of complex protein mixtures using isotope-coded affinity tags." Nat Biotechnol **17**(10): 994-999.

Han, D. K., J. Eng, H. Zhou and R. Aebersold (2001). "Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry." Nat Biotechnol **19**(10): 946-951.

Hanke, S., H. Besir, D. Oesterhelt and M. Mann (2008). "Absolute SILAC for accurate quantitation of proteins in complex mixtures down to the attomole level." J Proteome Res **7**(3): 1118-1130.

Hansen, K. C., G. Schmitt-Ulms, R. J. Chalkley, J. Hirsch, M. A. Baldwin and A. L. Burlingame (2003). "Mass spectrometric analysis of protein mixtures at low levels using cleavable 13C-isotope-coded affinity tag and multidimensional chromatography." Mol Cell Proteomics **2**(5): 299-314.

Henzel, W. J., T. M. Billeci, J. T. Stults, S. C. Wong, C. Grimley and C. Watanabe (1993). "Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases." Proc Natl Acad Sci U S A **90**(11): 5011-5015.

Henzel, W. J., C. Watanabe and J. T. Stults (2003). "Protein identification: the origins of peptide mass fingerprinting." J Am Soc Mass Spectrom **14**(9): 931-942.

Hoffert, J. D., G. Wang, T. Pisitkun, R. F. Shen and M. A. Knepper (2007). "An automated platform for analysis of phosphoproteomic datasets: application to kidney collecting duct phosphoproteins." J Proteome Res **6**(9): 3501-3508.

Hunt, D. F., J. R. Yates, 3rd, J. Shabanowitz, S. Winston and C. R. Hauer (1986). "Protein sequencing by tandem mass spectrometry." Proc Natl Acad Sci U S A **83**(17): 6233-6237.

Ideker, T., T. Galitski and L. Hood (2001). "A new approach to decoding life: systems biology." Annu Rev Genomics Hum Genet **2**: 343-372.

Ishihama, Y., Y. Oda, T. Tabata, T. Sato, T. Nagasu, J. Rappsilber and M. Mann (2005). "Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein." Mol Cell Proteomics **4**(9): 1265-1272.

Jaitly, N., A. Mayampurath, K. Littlefield, J. N. Adkins, G. A. Anderson and R. D. Smith (2009). "Decon2LS: An open-source software package for automated processing and visualization of high resolution mass spectrometry data." BMC Bioinformatics **10**: 87.

James, P., M. Quadroni, E. Carafoli and G. Gonnet (1993). "Protein identification by mass profile fingerprinting." Biochem Biophys Res Commun **195**(1): 58-64.

Johnson, R. S., S. A. Martin, K. Biemann, J. T. Stults and J. T. Watson (1987). "Novel fragmentation process of peptides by collision-induced decomposition in a tandem mass spectrometer: differentiation of leucine and isoleucine." Anal Chem **59**(21): 2621-2625.

Kall, L., J. D. Storey, M. J. MacCoss and W. S. Noble (2008). "Posterior error probabilities and false discovery rates: two sides of the same coin." J Proteome Res **7**(1): 40-44.

Kall, L., J. D. Storey and W. S. Noble (2008). "Non-parametric estimation of posterior error probabilities associated with peptides identified by tandem mass spectrometry." Bioinformatics **24**(16): i42-48.

Keller, A., A. I. Nesvizhskii, E. Kolker and R. Aebersold (2002). "Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search." Anal Chem **74**(20): 5383-5392.

Kruger, M., M. Moser, S. Ussar, I. Thievessen, C. A. Luber, F. Forner, S. Schmidt, S. Zanivan, R. Fassler and M. Mann (2008). "SILAC mouse for quantitative proteomics uncovers kindlin-3 as an essential factor for red blood cell function." Cell **134**(2): 353-364.

Lam, H., E. W. Deutsch, J. S. Eddes, J. K. Eng, N. King, S. E. Stein and R. Aebersold (2007). "Development and validation of a spectral library searching method for peptide identification from MS/MS." Proteomics **7**(5): 655-667.

Li, L., C. M. Barshick, J. T. Millay, A. V. Welty and F. L. King (2003). "Determination of bromine in flame-retardant plastics using pulsed glow discharge mass spectrometry." Anal Chem **75**(16): 3953-3961.

Li, Y. F., R. J. Arnold, Y. Li, P. Radivojac, Q. Sheng and H. Tang (2009). "A bayesian approach to protein inference problem in shotgun proteomics." J Comput Biol **16**(8): 1183-1193.

Lin, S. M., L. Zhu, A. Q. Winter, M. Sasinowski and W. A. Kibbe (2005). "What is mzXML good for?" Expert Rev Proteomics **2**(6): 839-845.

Liu, H., R. G. Sadygov and J. R. Yates, 3rd (2004). "A model for random sampling and estimation of relative protein abundance in shotgun proteomics." Anal Chem **76**(14): 4193-4201.

Lopez-Ferrer, D., S. Martinez-Bartolome, M. Villar, M. Campillos, F. Martin-Maroto and J. Vazquez (2004). "Statistical model for large-scale peptide identification in databases from tandem mass spectra using SEQUEST." Anal Chem **76**(23): 6853-6860.

Lottspeich, F. (1999). "Proteome Analysis: A Pathway to the Functional Analysis of Proteins." Angew Chem Int Ed Engl **38**(17): 2476-2492.

Luber, C. A., J. Cox, H. Lauterbach, B. Fancke, M. Selbach, J. Tschopp, S. Akira, M. Wiegand, H. Hochrein, M. O'Keeffe and M. Mann (2010). "Quantitative proteomics reveals subset-specific viral recognition in dendritic cells." Immunity **32**(2): 279-289.

Ma, Z. Q., S. Dasari, M. C. Chambers, M. D. Litton, S. M. Sobecki, L. J. Zimmerman, P. J. Halvey, B. Schilling, P. M. Drake, B. W. Gibson and D. L. Tabb (2009). "IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering." J Proteome Res **8**(8): 3872-3881.

Mann, M., P. Hojrup and P. Roepstorff (1993). "Use of mass spectrometric molecular weight information to identify proteins in sequence databases." Biol Mass Spectrom **22**(6): 338-345.

Mann, M. and M. Wilm (1994). "Error-tolerant identification of peptides in sequence databases by peptide sequence tags." Anal Chem **66**(24): 4390-4399.

Martens, L., M. Chambers, M. Sturm, D. Kessner, F. Levander, J. Shofstahl, W. H. Tang, A. Rompp, S. Neumann, A. D. Pizarro, L. Montecchi-Palazzi, N. Tasman, M. Coleman, F. Reisinger, P. Souda, H. Hermjakob, P. A. Binz and E. W. Deutsch (2011). "mzML--a community standard for mass spectrometry data." Mol Cell Proteomics **10**(1): R110 000133.

Meyer-Arendt, K., W. M. Old, S. Houel, K. Renganathan, B. Eichelberger, K. A. Resing and N. G. Ahn (2011). "IsoformResolver: A peptide-centric algorithm for protein inference." J Proteome Res **10**(7): 3060-3075.

Mueller, L. N., M. Y. Brusniak, D. R. Mani and R. Aebersold (2008). "An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data." J Proteome Res **7**(1): 51-61.

Mueller, L. N., O. Rinner, A. Schmidt, S. Letarte, B. Bodenmiller, M. Y. Brusniak, O. Vitek, R. Aebersold and M. Muller (2007). "SuperHirn - a novel tool for high resolution LC-MS-based peptide/protein profiling." Proteomics **7**(19): 3470-3480.

Nasso, S., F. Silvestri, F. Tisiot, B. Di Camillo, A. Pietracaprina and G. M. Toffolo (2010). "An optimized data structure for high-throughput 3D proteomics data: mzRTree." J Proteomics **73**(6): 1176-1182.

Navarro, P. and J. Vazquez (2009). "A refined method to calculate false discovery rates for peptide identification using decoy databases." J Proteome Res **8**(4): 1792-1796.

Nesvizhskii, A. I. (2010). "A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics." J Proteomics **73**(11): 2092-2123.

Nesvizhskii, A. I. and R. Aebersold (2004). "Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem MS." Drug Discov Today **9**(4): 173-181.

Nesvizhskii, A. I. and R. Aebersold (2005). "Interpretation of shotgun proteomic data: the protein inference problem." Mol Cell Proteomics **4**(10): 1419-1440.

Nesvizhskii, A. I., A. Keller, E. Kolker and R. Aebersold (2003). "A statistical model for identifying proteins by tandem mass spectrometry." Anal Chem **75**(17): 4646-4658.

Oda, Y., K. Huang, F. R. Cross, D. Cowburn and B. T. Chait (1999). "Accurate quantitation of protein expression and site-specific phosphorylation." Proc Natl Acad Sci U S A **96**(12): 6591-6596.

Old, W. M., K. Meyer-Arendt, L. Aveline-Wolf, K. G. Pierce, A. Mendoza, J. R. Sevinsky, K. A. Resing and N. G. Ahn (2005). "Comparison of label-free methods for quantifying human proteins by shotgun proteomics." Mol Cell Proteomics **4**(10): 1487-1502.

Olsen, J. V., P. A. Nielsen, J. R. Andersen, M. Mann and J. R. Wisniewski (2007). "Quantitative proteomic profiling of membrane proteins from the mouse brain cortex, hippocampus, and cerebellum using the HysTag reagent: mapping of neurotransmitter receptors and ion channels." Brain Res **1134**(1): 95-106.

Olsen, J. V., S. E. Ong and M. Mann (2004). "Trypsin cleaves exclusively C-terminal to arginine and lysine residues." Mol Cell Proteomics **3**(6): 608-614.

Ong, S. E., B. Blagoev, I. Kratchmarova, D. B. Kristensen, H. Steen, A. Pandey and M. Mann (2002). "Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics." Mol Cell Proteomics **1**(5): 376-386.

Ong, S. E. and M. Mann (2005). "Mass spectrometry-based proteomics turns quantitative." Nat Chem Biol **1**(5): 252-262.

Ong, S. E. and M. Mann (2006). "A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC)." Nat Protoc **1**(6): 2650-2660.

Pappin, D. J., P. Hojrup and A. J. Bleasby (1993). "Rapid identification of proteins by peptide-mass fingerprinting." Curr Biol **3**(6): 327-332.

Patterson, S. D. and R. H. Aebersold (2003). "Proteomics: the first decade and beyond." Nat Genet **33 Suppl**: 311-323.

Pedrioli, P. G., J. K. Eng, R. Hubley, M. Vogelzang, E. W. Deutsch, B. Raught, B. Pratt, E. Nilsson, R. H. Angeletti, R. Apweiler, K. Cheung, C. E. Costello, H. Hermjakob, S. Huang, R. K. Julian, E. Kapp, M. E. McComb, S. G. Oliver, G. Omenn, N. W. Paton, R. Simpson, R. Smith, C. F. Taylor, W. Zhu and R. Aebersold (2004). "A common open representation of mass spectrometry data and its application to proteomics research." Nat Biotechnol **22**(11): 1459-1466.

Peng, J., J. E. Elias, C. C. Thoreen, L. J. Licklider and S. P. Gygi (2003). "Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome." J Proteome Res **2**(1): 43-50.

Perkins, D. N., D. J. Pappin, D. M. Creasy and J. S. Cottrell (1999). "Probability-based protein identification by searching sequence databases using mass spectrometry data." Electrophoresis **20**(18): 3551-3567.

Petricoin, E. E., C. P. Paweletz and L. A. Liotta (2002). "Clinical applications of proteomics: proteomic pattern diagnostics." J Mammary Gland Biol Neoplasia **7**(4): 433-440.

Poinot, M. (2010). "Five Good Reasons to Use the Hierarchical Data Format." Computing in Science & Engineering **12**(5): 84-90.

Price, T. S., M. B. Lucitt, W. Wu, D. J. Austin, A. Pizarro, A. K. Yocum, I. A. Blair, G. A. FitzGerald and T. Grosser (2007). "EBP, a program for protein identification using multiple tandem mass spectrometry datasets." Mol Cell Proteomics **6**(3): 527-536.

Raymond, A. A., A. Gonzalez de Peredo, A. Stella, A. Ishida-Yamamoto, D. Bouyssie, G. Serre, B. Monsarrat and M. Simon (2008). "Lamellar bodies of human epidermis: proteomics characterization by high throughput mass spectrometry and possible involvement of CLIP-170 in their trafficking/secretion." Mol Cell Proteomics **7**(11): 2151-2175.

Roepstorff, P. and J. Fohlman (1984). "Proposal for a common nomenclature for sequence ions in mass spectra of peptides." Biomed Mass Spectrom **11**(11): 601.

Roux-Dalvai, F., A. Gonzalez de Peredo, C. Simo, L. Guerrier, D. Bouyssie, A. Zanella, A. Citterio, O. Burlet-Schiltz, E. Boschetti, P. G. Righetti and B. Monsarrat (2008). "Extensive analysis of the cytoplasmic proteome of human erythrocytes using the peptide ligand library technology and advanced mass spectrometry." Mol Cell Proteomics **7**(11): 2254-2269.

Sadygov, R. G. and J. R. Yates, 3rd (2003). "A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases." Anal Chem **75**(15): 3792-3798.

Schulze, W. X. and M. Mann (2004). "A novel proteomic screen for peptide-protein interactions." J Biol Chem **279**(11): 10756-10764.

Schwanhausser, B., D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen and M. Selbach (2011). "Global quantification of mammalian gene expression control." Nature **473**(7347): 337-342.

Seidler, J., N. Zinn, M. E. Boehm and W. D. Lehmann (2010). "De novo sequencing of peptides by MS/MS." Proteomics **10**(4): 634-649.

Selbach, M. and M. Mann (2006). "Protein interaction screening by quantitative immunoprecipitation combined with knockdown (QUICK)." Nat Methods **3**(12): 981-983.

Shah, A. R., J. Davidson, M. E. Monroe, A. M. Mayampurath, W. F. Danielson, Y. Shi, A. C. Robinson, B. H. Clowers, M. E. Belov, G. A. Anderson and R. D. Smith (2010). "An efficient data format for mass spectrometry-based proteomics." J Am Soc Mass Spectrom **21**(10): 1784-1788.

Smith, R. D., G. A. Anderson, M. S. Lipton, C. Masselon, L. Pasa-Tolic, Y. Shen and H. R. Udseth (2002). "The use of accurate mass tags for high-throughput microbial proteomics." OMICS **6**(1): 61-90.

Stanislas, T., D. Bouyssie, M. Rossignol, S. Vesa, J. Fromentin, J. Morel, C. Pichereaux, B. Monsarrat and F. Simon-Plas (2009). "Quantitative proteomics reveals a dynamic association of proteins to detergent-resistant membranes upon elicitor signaling in tobacco." Mol Cell Proteomics **8**(9): 2186-2198.

Steen, H. and M. Mann (2004). "The ABC's (and XYZ's) of peptide sequencing." Nat Rev Mol Cell Biol **5**(9): 699-711.

Sturm, M., A. Bertsch, C. Gropl, A. Hildebrandt, R. Hussong, E. Lange, N. Pfeifer, O. Schulz-Trieglaff, A. Zerck, K. Reinert and O. Kohlbacher (2008). "OpenMS - an open-source software framework for mass spectrometry." BMC Bioinformatics **9**: 163.

Taylor, C. F., N. W. Paton, K. S. Lilley, P. A. Binz, R. K. Julian, Jr., A. R. Jones, W. Zhu, R. Apweiler, R. Aebersold, E. W. Deutsch, M. J. Dunn, A. J. Heck, A. Leitner, M. Macht, M. Mann, L. Martens, T. A. Neubert, S. D. Patterson, P. Ping, S. L. Seymour, P. Souda, A. Tsugita, J. Vandekerckhove, T. M. Vondriska, J. P. Whitelegge, M. R. Wilkins, I. Xenarios, J. R. Yates, 3rd and H. Hermjakob (2007). "The minimum information about a proteomics experiment (MIAPE)." Nat Biotechnol **25**(8): 887-893.

Trinkle-Mulcahy, L., S. Boulon, Y. W. Lam, R. Urcia, F. M. Boisvert, F. Vandermoere, N. A. Morrice, S. Swift, U. Rothbauer, H. Leonhardt and A. Lamond (2008). "Identifying specific protein interaction partners using quantitative mass spectrometry and bead proteomes." J Cell Biol **183**(2): 223-239.

Tsou, C. C., C. F. Tsai, Y. H. Tsui, P. R. Sudhir, Y. T. Wang, Y. J. Chen, J. Y. Chen, T. Y. Sung and W. L. Hsu (2010). "IDEAL-Q, an automated tool for label-free quantitation analysis using an efficient peptide alignment approach and spectral data validation." Mol Cell Proteomics **9**(1): 131-144.

Vinther, J., M. M. Hedegaard, P. P. Gardner, J. S. Andersen and P. Arctander (2006). "Identification of miRNA targets with stable isotope labeling by amino acids in cell culture." Nucleic Acids Res **34**(16): e107.

Washburn, M. P., D. Wolters and J. R. Yates, 3rd (2001). "Large-scale analysis of the yeast proteome by multidimensional protein identification technology." Nat Biotechnol **19**(3): 242-247.

Weatherly, D. B., J. A. Atwood, 3rd, T. A. Minning, C. Cavola, R. L. Tarleton and R. Orlando (2005). "A Heuristic method for assigning a false-discovery rate for protein identifications from Mascot database search results." Mol Cell Proteomics **4**(6): 762-772.

Wilhelm, M., M. Kirchner, J. A. Steen and H. Steen (2012). "mz5: space- and time-efficient storage of mass spectrometry data sets." Mol Cell Proteomics **11**(1): O111 011379.

Xie, F., T. Liu, W. J. Qian, V. A. Petyuk and R. D. Smith (2011). "Liquid chromatography-mass spectrometry-based quantitative proteomics." J Biol Chem **286**(29): 25443-25449.

Yates, J. R., 3rd, S. Speicher, P. R. Griffin and T. Hunkapiller (1993). "Peptide mass maps: a highly informative approach to protein identification." Anal Biochem **214**(2): 397-408.

Zhang, B., M. C. Chambers and D. L. Tabb (2007). "Proteomic parsimony through bipartite graph analysis improves accuracy and transparency." J Proteome Res **6**(9): 3549-3557.

Zhu, W., J. W. Smith and C. M. Huang (2010). "Mass spectrometry-based label-free quantitative proteomics." J Biomed Biotechnol **2010**: 840518.

Liste des publications

1: Gautier V, Mouton-Barbosa E, Bouyssié D, Delcourt N, Beau M, Girard JP, Cayrol C, Burlet-Schiltz O, Monsarrat B, Gonzalez de Peredo A (2012). “Label-free Quantification and Shotgun Analysis of Complex Proteomes by One-dimensional SDS-PAGE/NanoLC-MS: EVALUATION FOR THE LARGE SCALE ANALYSIS OF INFLAMMATORY HUMAN ENDOTHELIAL CELLS.” Mol Cell Proteomics. 11(8):527-39.

2: Mischak H, Kolch W, Aivaliotis M, Bouyssié D, Court M, Dihazi H, Dihazi GH, Franke J, Garin J, Gonzalez de Peredo A, Iphöfer A, Jänsch L, Lacroix C, Makridakis M, Masselon C, Metzger J, Monsarrat B, Mrug M, Norling M, Novak J, Pich A, Pitt A, Bongcam-Rudloff E, Siwy J, Suzuki H, Thongboonkerd V, Wang LS, Zoidakis J, Zürbig P, Schanstra JP, Vlahou A (2010). “Comprehensive human urine standards for comparability and standardization in clinical proteome analysis.” Proteomics Clin Appl 4(4):464-78.

3: Mouton-Barbosa E, Roux-Dalvai F, Bouyssié D, Berger F, Schmidt E, Righetti PG, Guerrier L, Boschetti E, Burlet-Schiltz O, Monsarrat B, Gonzalez de Peredo A (2010). “In-depth exploration of cerebrospinal fluid by combining peptide ligand library treatment and label-free protein quantification.” Mol Cell Proteomics 9(5):1006-21.

4: Bousquet-Dubouch MP, Nguen S, Bouyssié D, Burlet-Schiltz O, French SW, Monsarrat B, Bardag-Gorce F (2009). “Chronic ethanol feeding affects proteasome-interacting proteins.” Proteomics (13):3609-22.

5: Stanislas T, Bouyssie D, Rossignol M, Vesa S, Fromentin J, Morel J, Pichereaux C, Monsarrat B, Simon-Plas F (2009). "Quantitative proteomics reveals a dynamic association of proteins to detergent-resistant membranes upon elicitor signaling in tobacco.” Mol Cell Proteomics 8(9):2186-98.

6: Raymond AA, Gonzalez de Peredo A, Stella A, Ishida-Yamamoto A, Bouyssie D, Serre G, Monsarrat B, Simon M (2008). “Lamellar bodies of human epidermis: proteomics characterization by high throughput mass spectrometry and possible involvement of CLIP-170 in their trafficking/secretion.” Mol Cell Proteomics 7(11):2151-75.

7: Roux-Dalvai F, Gonzalez de Peredo A, Simó C, Guerrier L, Bouyssié D, Zanella A, Citterio A, Burlet-Schiltz O, Boschetti E, Righetti PG, Monsarrat B (2008). “Extensive analysis of the cytoplasmic proteome of human erythrocytes using the peptide ligand library technology and advanced mass spectrometry.” Mol Cell Proteomics 7(11):2254-69.

8: Bouyssié D, Gonzalez de Peredo A, Mouton E, Albigot R, Roussel L, Ortega N, Cayrol C, Burlet-Schiltz O, Girard JP, Monsarrat B (2007). “Mascot file parsing and quantification (MFPaQ), a new software to parse, validate, and quantify proteomics data generated by ICAT and SILAC mass spectrometric analyses: application to the proteomics study of membrane proteins from primary human endothelial cells.” Mol Cell Proteomics 6(9):1621-37.

Annexe

Label-free Quantification and Shotgun Analysis of Complex Proteomes by One-dimensional SDS-PAGE/NanoLC-MS

EVALUATION FOR THE LARGE SCALE ANALYSIS OF INFLAMMATORY HUMAN ENDOTHELIAL CELLS*[S]

Violette Gautier‡§¶, Emmanuelle Mouton-Barbosa§¶, David Bouyssié‡§¶, Nicolas Delcourt§, Mathilde Beau‡§, Jean-Philippe Girard‡§‖, Corinne Cayrol‡§‖, Odile Burlet-Schiltz‡§, Bernard Monsarrat‡§, and Anne Gonzalez de Peredo‡§‡‡**

To perform differential studies of complex protein mixtures, strategies for reproducible and accurate quantification are needed. Here, we evaluated a quantitative proteomic workflow based on nanoLC-MS/MS analysis on an LTQ-Orbitrap-VELOS mass spectrometer and label-free quantification using the MFPaQ software. In such label-free quantitative studies, a compromise has to be found between two requirements: repeatability of sample processing and MS measurements, allowing an accurate quantification, and high proteomic coverage of the sample, allowing quantification of minor species. The latter is generally achieved through sample fractionation, which may induce experimental bias during the label-free comparison of samples processed, and analyzed independently. In this work, we wanted to evaluate the performances of MS intensity-based label-free quantification when a complex protein sample is fractionated by one-dimensional SDS-PAGE. We first tested the efficiency of the analysis without protein fractionation and could achieve quite good quantitative repeatability in single-run analysis (median coefficient of variation of 5%, 99% proteins with coefficient of variation <48%). We show that sample fractionation by one-dimensional SDS-PAGE is associated with a moderate decrease of quantitative measurement repeatability while largely improving the depth of proteomic coverage. We then applied the method for a large scale proteomic study of the human endothelial cell response to inflammatory cytokines, such as TNFα, interferon γ, and IL1β, which allowed us to finely decipher at the proteomic level the biological pathways involved in endothelial cell response to proinflammatory cytokines. ***Molecular & Cellular Proteomics 11: 10.1074/mcp.M111.015230, 527–539, 2012.***

From ‡Centre National de la Recherche Scientifique, Institut de Pharmacologie et de Biologie Structurale, F-31077 Toulouse, France, and §Université de Toulouse, Université Paul Sabatier, Institut de Pharmacologie et de Biologie Structurale, F-31077 Toulouse, France

Received October 21, 2011, and in revised form, April 10, 2012

Published, MCP Papers in Press, April 19, 2012, DOI 10.1074/mcp.M111.015230

With recent advances in mass spectrometry, label-free quantitative proteomic approaches have progressed and are now considered as reliable and efficient methods to study protein expression level changes in complex mixtures. These approaches, which have been reviewed recently (1, 2), are based on the measurement either of the MS/MS sampling rate of a particular peptide or of its MS chromatographic peak area, these values being directly related to peptide abundance. The increase of instrument sequencing speed has benefited MS/MS spectral counting approaches by improving MS/MS sampling of peptide mixtures, whereas the introduction of high resolution analyzers such as FT-Orbitrap has boosted the use of methods based on peptide intensity measurements by greatly facilitating the matching of peptide peaks in different complex maps acquired independently. The most obvious advantage of these methods over isotopic labeling techniques is their ease of use at the sample preparation step, because they do not require any preliminary treatment to introduce a label into peptides or proteins. Being more straightforward, they also do not present the classical drawbacks of labeling methods, *i.e.*, cost, applicability to limited types of samples (mostly cultured cells in the case of metabolic labeling) and the limited number of conditions that can be compared. On the other hand, the use of label-free strategies is hampered by two main difficulties: 1) the variability of all sample processing steps before MS analysis and of the analytical measurement itself, because the samples to be compared are processed and analyzed individually, and 2) the complexity of the MS data analysis step, which requires proper realignment, normalization, and peptide peaks matching across different nanoLC-MS runs.

Many bioinformatic tools have been developed in recent years for the quantification of MS data generated in label-free experiments, either by spectral counting (3–5) or by peptide MS signal intensity measurement (6–9). In the later field, a lot of emphasis has been put on peptide pattern-based methods, in which the software performs feature detection in LC-MS maps through analysis of the characteristic isotopic pattern of a peptide ion in the *m/z* dimen-

sion, and on its chromatographic elution peak in the retention time (RT)[1] dimension. The total ion current integrated under this MS feature can then be used as a quantitative measurement of the peptide concentration. The primary advantage of this approach is that any signal detected by the mass spectrometer in the MS survey scan can be in principle analyzed and quantified, whether or not the peak has been selected for MS/MS sequencing. Bioinformatic programs based on peptide feature detection as the starting step for label-free analysis include among others SuperHirn (8), MSInspect (6), OpenMS (9), Decon2LS (7), or the commercial software Progenesis LC-MS. Although they offer an attractive and powerful analysis of the data, algorithms based on recognition of peptide features and LC-MS maps alignment require intensive computer calculation, making the quantification time-consuming and difficult to perform on a large number of LC-MS files. In addition, integration of MS features quantitative data with MS/MS identification results from search engines occurs as a second step, and depending on the bioinformatic tool used, retrieving quantitative values for the list of identified and validated peptides, and then for the associated list of proteins, can be difficult to implement. Finally, because the LC-MS maps are usually analyzed individually, low intensity features near the cut-off value set for the recognition process are detected in an irreproducible way, and most of the available software generates quantitative data sets containing many missing values, which complicates further statistical analysis of the results.

On the other hand, another approach to extract quantitative data from MS survey scans is based on the reverse process, *i.e.*, making use initially of peptide identification results to go back in the MS scans to obtain peptide intensity values. For each peptide ion identified from MS/MS sequencing, experimentally measured RT and monoisotopic *m/z* values can be used as a starting point to retrieve the associated extracted ion chromatograms (XIC) of this ion. In that case, confident extraction of a peptide signal (*versus* chemical noise) is supported by the identification result, and because the charge state of the ion is known, definition of isotopic patterns and extraction of intensity values for the different isotopes of a same peptide ion is facilitated. Such a method, which is in principle more simple and rapid, has been used in a few software packages such as Serac (10), Quoil (11), Ideal-Q (12), and MFPaQ (13, 14). A drawback of this method, however, is that only identified peptides can be quantified. For analysis of highly complex peptide mixtures, MS/MS undersampling thus limits the number of identified and quantified proteins. Depending on the software, this problem can be alleviated by a cross-assignment of peptide signals across different replicate LC-MS/MS runs: if a peptide ion is identified in only one or a few runs, its signal can be extracted in the other analytical runs by using a predicted RT value, even if identification results are missing for this particular peptide in these runs because of MS/MS undersampling. Thus, acquisition of multiple replicate runs allows to increase the number of identified and thus quantified peptides and proteins. Nevertheless, the performances of identity-based methods are still strictly linked to the number of identifications and to the depth of the proteomic analysis on highly complex samples.

In a study focusing on label-free quantitative analysis of clinical samples (14), we previously described an approach based on the use of the MFPaQ software to circumvent the undersampling problem. Following extensive proteomic analysis of cerebrospinal fluid after treatment with combinatorial libraries of peptide ligands and one-dimensional SDS-PAGE fractionation, we generated an identification database containing sequences of identified peptides, along with their *m/z* and retention time-associated values that were then used to extract the XIC of these peptides in the one-shot analytical runs of unfractionated samples. This method was well suited for the analysis of clinical series in which very limited or no fractionation at all is performed on the samples, because of the large number of analyses (number of patients and technical replicates), and we showed that it indeed allowed significant increase of the number of proteins correctly quantified in replicate runs of individual samples. However, not all of the peptides from the database could be retrieved in the individual runs, because of the limited dynamic range of the instrument during the one-shot analysis of complex peptide mixtures. To overcome also the dynamic range limitation, a commonly used and efficient approach is to prefractionate individually the samples to be compared and perform nanoLC-MS/MS analysis of each fraction separately. Although it requires longer analytical time, this shotgun type of analysis clearly offers an improved coverage of the sample and allows the detection of low abundance proteins that remain undetected when the whole sample is analyzed in one run. To that aim, one-dimensional SDS-PAGE is often selected as a robust and simple method to fractionate most kinds of protein samples, even membrane ones, and is particularly used on SILAC or ICAT labeled proteomes, because the two samples to be compared are gathered and can be processed simultaneously. However, when label-free quantification is to be performed, parallel processing steps such as electrophoretic migration, gel cutting, and in-gel digestion represent different sources of variability that may alter the final quantitative comparison of the samples.

In the present study, our objective was to perform an in-depth quantitative analysis of the endothelial cell (EC) proteome using a label-free approach. First, we thus checked whether SDS-PAGE fractionation of the individual samples, which gives the best dynamic range on a global analysis, is

[1] The abbreviations used are: RT, retention time; XIC, extracted ion chromatogram; EC, endothelial cells; IFNγ, interferon γ; HUVEC, human umbilical vein endothelial cells; FDR, false discovery rate; PAI, protein abundance index; CV, coefficient of variation; MHC, major histocompatibility complex.

compatible with accurate label-free quantitation based on peptide signal intensity measurement. We evaluated the performances of a label-free quantitative workflow in terms of repeatability and number of quantified proteins, with or without protein fractionation by one-dimensional SDS-PAGE, for the analysis of a complex cellular proteome. We applied the MFPaQ software, which uses an identity-based extraction approach, to quantify the data obtained from the nanoLC-MS/MS analysis of a total lysate of primary cultured human vascular ECs. New data normalization and integration procedure dedicated to shotgun experiments were introduced in the software, allowing integration at the protein level the quantitative data from different fractions and correction of errors related to nonreproducible electrophoretic migration of proteins. We showed that the approach based on peptide XIC extraction provides good quality quantitative data on the identified proteome and that high repeatability is obtained on proteins quantified in single run analysis (median CV of 5%, 99% proteins with CV values of $<$48%). When the protein sample is fractionated by one-dimensional SDS-PAGE, the repeatability of the quantitative measurement decreases, although in a moderate way (median CV of 7%, 99% proteins with CV values of $<$62%), and concomitantly the depth of proteomic coverage is largely increased. We then applied the method for a large scale proteomic study of the response of ECs to proinflammatory treatments with TNFα/IFNγ or IL1β. It allowed us to identify and quantify more than 5400 unique proteins, providing an in-depth analysis of the endothelial cell proteome and a detailed characterization of the proteomic variations associated with the inflammatory response.

MATERIALS AND METHODS

EC Culture and Cytokine Stimulation—Primary human umbilical vein ECs (HUVECs) were purchased from Clonetics, grown in ECGM medium (Promocell, Heidelberg, Germany), and used after four passages for proteomic analyses. Cytokine treatment was performed by incubating the ECs for 12 h in OptiMEM medium (Invitrogen) with a combination of TNFα (25 ng/ml; R & D Systems) and IFNγ (50 ng/ml; R & D Systems) or with IL1β (5 ng/ml; R & D Systems).

Protein Sample Processing—The cells were lysed in a buffer containing 2% of SDS and sonicated, and protein concentration was determined by detergent-compatible assay (DC assay; Bio-Rad). Protein samples were reduced in Laemmli buffer (final composition: 25 mM DTT, 2% SDS, 10% glycerol, 40 mM Tris, pH 6.8) for 5 min at 95 °C. Cysteine residues were alkylated by addition of iodoacetamide at a final concentration of 90 mM and incubation for 30 min at room temperature in the dark. During the alkylation reaction, the pH of the samples was adjusted using small amounts of 1 M Tris, pH 8. Protein samples were loaded on a homemade one-dimensional SDS-PAGE gel (separating gel 1.5 mm $\times$ 5 cm, 12% acrylamide polymerized in SDS 0.1%, 375 mM Tris, pH 8.8, and stacking gel 1.5 mm $\times$ 1.5 cm, 4% acrylamide polymerized in 0.1% SDS, 125 mM Tris. For one-shot analysis of the entire mixture, no fractionation was performed, and the electrophoretic migration was stopped as soon as the protein sample (15 μg) entered the separating gel. The gel was briefly stained with Coomassie Blue, and a single band, containing the whole sample, was cut. For shotgun analysis, electrophoretic migration was performed to fractionate the protein sample (100 μg) into 12 gel bands. For replicate and comparative analyses, the samples were processed on adjacent migration lanes that were cut simultaneously with a long razor blade. To evaluate gel to gel repeatability, different gels were prepared and migrated in parallel, and the same number of homogeneous gel slices were cut successively on the separate gels, following the same cutting pattern. Gel slices were washed by two cycles of incubation in 100 mM ammonium bicarbonate for 15 min at 37 °C, followed by 100 mM ammonium bicarbonate/acetonitrile (1:1) for 15 min at 37 °C. The proteins were digested by 0.6 μg of modified sequencing grade trypsin (Promega) in 50 mM ammonium bicarbonate, overnight at 37 °C. The resulting peptides were extracted from the gel by incubation in 50 mM ammonium bicarbonate for 15 min at 37 °C and twice in 10% formic acid/acetonitrile (1:1) for 15 min at 37 °C. The three collected extractions were pooled with the initial digestion supernatant, dried in a SpeedVac, and resuspended with 17 μl of 5% acetonitrile, 0.05% TFA.

NanoLC-MS/MS Analysis—The Resulting peptides were analyzed by nanoLC-MS/MS using an Ultimate3000 system (Dionex, Amsterdam, The Netherlands) coupled to an LTQ-Orbitrap Velos mass spectrometer (Thermo Fisher Scientific, Bremen, Germany). Five μl of each sample were loaded on a C-18 precolumn (300-μm inner diameter $\times$ 5 mm; Dionex) at 20 μl/min in 5% acetonitrile, 0.05% TFA. After 5 min of desalting, the precolumn was switched online with the analytical C-18 column (75 μm inner diameter $\times$ 15 cm; PepMap C18, Dionex) equilibrated in 95% solvent A (5% acetonitrile, 0.2% formic acid) and 5% solvent B (80% acetonitrile, 0.2% formic acid). The peptides were eluted using a 5 to 50% gradient of solvent B during 80 min at 300 nl/min flow rate. The LTQ-Orbitrap Velos was operated in data-dependent acquisition mode with the XCalibur software. Survey scan MS were acquired in the Orbitrap on the 300–2000 *m/z* range with the resolution set to a value of 60,000. The 10 most intense ions per survey scan were selected for CID fragmentation, and the resulting fragments were analyzed in the linear trap (LTQ). Dynamic exclusion was employed within 60 s to prevent repetitive selection of the same peptide.

Database Search and Data Validation—The Mascot Daemon software (version 2.3.2; Matrix Science, London, UK) was used to perform database searches, using the Extract_msn.exe macro provided with Xcalibur (version 2.0 SR2; Thermo Fisher Scientific) to generate peaklists. The following parameters were set for creation of the peaklists: parent ions in the mass range 400–4500, no grouping of MS/MS scans, and threshold at 1000. A peaklist was created for each analyzed fraction (*i.e.*, gel slice), and individual Mascot (version 2.3.01) searches were performed for each fraction. The data were searched against *Homo sapiens* entries in Uniprot protein database (release 2010_09, September 21, 2010; 1,215,533 sequences). Carbamidomethylation of cysteines was set as a fixed modification, and oxidation of methionine and protein N-terminal acetylation were set as a variable modifications. Specificity of trypsin digestion was set for cleavage after Lys or Arg, and two missed trypsin cleavage sites were allowed. The mass tolerances in MS and MS/MS were set to 5 ppm and 0.6 Da, respectively, and the instrument setting was specified as "ESI-Trap." To calculate the false discovery rate (FDR), the search was performed using the "decoy" option in Mascot. Peptide identifications extracted from Mascot result files were validated at a final peptide FDR of 5%. Peptide matches were validated if their score was greater than the Mascot homology threshold (when available, otherwise the Mascot identity threshold was used) for a given Mascot *p* value. The FDR at the peptide level was calculated as described in Navarro and Vázquez (15). Using this method, the *p* value was automatically adjusted to obtain a FDR of 5% at the peptide level. Validated peptides were assembled into proteins groups following the principle of parsimony (Ocam's razor), which involves the creation of the minimal list of protein groups explaining the list of peptide spec-

trum matches. Protein groups were then rescored for the protein validation process. For each peptide match belonging to a protein group, the difference between its Mascot score and its homology threshold (or identity threshold) was computed for a given *p* value (automatically adjusted to increase the discrimination between target and decoy matches), and these "score offsets" were then summed to obtain the protein group score. Protein groups were validated based on this score to obtain a FDR of 1% at the protein level (FDR = number of validated decoy hits/(number of validated target hits + number of validated decoy hits) × 100). In the case of sample fractionation on one-dimensional SDS-PAGE, the MFPaQ software was used to create a unique nonredundant protein list from the identification results of each fraction by clustering protein groups containing sequences matching the same set of peptides. If a final group was composed of several TrEMBL and SwissProt entry names, a SwissProt entry was singled out, and the associated protein description was reported in the final lists (supplemental Tables I and II).

Data Quantification—Quantification of proteins was performed using the label-free module implemented in the MFPaQ v4.0.0 software (http://mfpaq.sourceforge.net/). For each sample, the software uses the validated identification results and XICs of the identified peptide ions in the corresponding raw nanoLC-MS files, based on their experimentally measured RT and monoisotopic *m/z* values The time value used for this process is retrieved from Mascot result files, based on an MS2 event matching to the peptide ion. If several MS2 events were matched to a given peptide ion, the software checks the intensity of each corresponding precursor peak in the previous MS survey scan. The time of the MS scan that exhibits the highest precursor ion intensity is attributed to the peptide ion and then used for XIC extraction as well as for the alignment process. Peptide ions identified in all the samples to be compared were used to build a retention time matrix to align LC-MS runs. If some peptide ions were sequenced by MS/MS and validated only in some of the samples to be compared, their XIC signal was extracted in the nanoLC-MS raw file of the other samples using a predicted RT value calculated from this alignment matrix by a linear interpolation method. Quantification of peptide ions was performed based on calculated XIC areas values. To perform normalization of a group of comparable runs, the software computed XIC area ratios for all the extracted signals between a reference run and all the other runs of the group and used the median of the ratios as a normalization factor. To perform protein relative quantification in different samples, a protein abundance index was calculated, defined as the average of XIC area values for at most three intense reference tryptic peptides identified for this protein (the three peptides exhibiting the highest intensities across the different samples were selected as reference peptides, and these same three peptides were used to compute the PAI of the protein in each sample; if only one or two peptides were identified and quantified in the case of low abundant proteins, the PAI was calculated based on their XIC area values). In the case of SDS-PAGE fractionation, integration of quantitative data across the fractions was performed as indicated in the text, by summing the PAI values for fractions adjacent to the fraction with the best PAI (the same three consecutive fractions for all the samples to be compared). For differential studies, a Student's *t* test on the PAI values was used for statistical evaluation of the significance of expression level variations. For proteins specifically detected in one condition and not in the other, the *t* test *p* value was calculated by assigning a noise background value to the missing PAI values. A 2-fold change and *p* value of 0.05 were used as combined thresholds to define biologically regulated proteins.

Quantitative PCR Experiments—Total RNA from HUVEC cells (mock treated, TNFα + IFNγ-treated, or IL1β-treated) was isolated using the Absolute RNA kit from Stratagene (Agilent Technologies, Santa Clara, CA), and cDNAs were synthesized using SuperSript III First strand cDNA synthesis system for RT-PCR (Invitrogen) according to the manufacturer's instructions. Quantitative PCR was performed using the ABI7300 Prism SDS real time PCR detection system (Applied Biosystems, Foster City, CA) with a SYBR Green PCR Master Mix kit (Applied Biosystems) and a standard temperature protocol. The results are expressed as relative quantities and calculated by the 2-ΔΔCT method. *Actin* was used as a control gene for normalization. Three separate experiments were performed. Primers used were purchased from Qiagen (QuantiTect primer assay), except Actin, GAPDH, NFKB2, ICAM1, and VCAM1 (from Sigma Genosys).

RESULTS

Analytical Workflow—A total lysate of cultured primary human vascular ECs was used for all experiments and processed in all cases through one-dimensional SDS-PAGE, as shown in Fig. 1. When the samples were to be analyzed in one analytical nanoLC-MS/MS run (no fractionation), the electrophoretic migration was stopped immediately after the protein samples entered the separating part of the gel, so that the whole sample was isolated into a unique gel band and subsequently in-gel digested. In our hands, processing in this way, the total cell lysate for tryptic digestion gave slightly better proteomic coverage than digestion in solution. For sample fractionation and shotgun analysis, migration was performed so that 12 gel bands could be cut afterward along the migration lanes. Gel cutting was performed systematically with a long razor blade to simultaneously cut all the corresponding gel bands for the different samples to be compared, perpendicularly to the migration direction. All in gel digestion steps were manually performed in parallel. The resulting tryptic digests were analyzed by nanoLC-MS/MS on an Orbitrap-Velos instrument with high sequencing speed to improve the MS/MS sampling and analytical coverage of the samples. MS scans were recorded in the Orbitrap, and MS/MS CID spectra were recorded in the ion trap using a classical parallel acquisition mode to obtain high resolution MS^1 data for peptide quantification while optimizing the number of MS^2 sequencing events to increase peptide identifications. Database searches using MS/MS sequencing data were performed through Mascot, and the results files were parsed and validated based on target decoy calculated FDRs, set at 5% for peptides and 1% for proteins. After realignment in time of nanoLC-MS runs, the software uses the *m/z* and time values associated to validated peptides ions of validated proteins, to extract the XIC of each of them. If some peptide ions were sequenced by MS/MS and validated only in some of the samples to be compared, their XIC signal was extracted in the nanoLC-MS raw file of the other samples using a predicted RT value and a time tolerance window. For protein quantification, a PAI was calculated, defined as the average of XIC area values of at most three intense reference tryptic peptides identified for this protein.

Repeatability of the Label-free Quantification without Sample Fractionation—We first evaluated the repeatability of the label-free analytical workflow by comparing replicate LC-MS

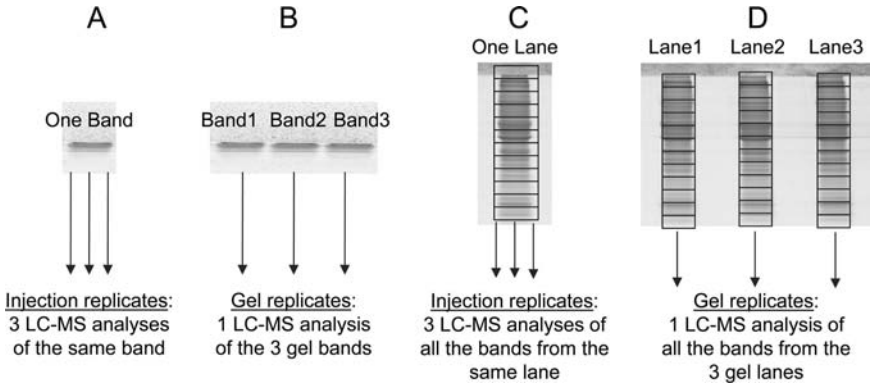
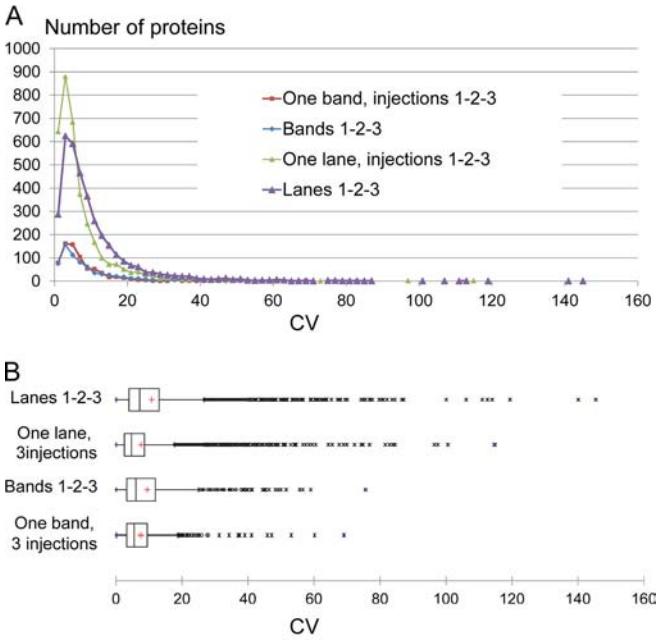


FIG. 1. **Experimental design to estimate the accuracy of label-free quantification with or without sample fractionation.** The same endothelial cell lysate was loaded on a one-dimensional SDS gel and either collected in a single band or fractionated into 12 gel bands cut along the migration lane, and to assess how repeatability is affected by each step of the analytical process, we compared either nanoLC-MS/MS injection replicates or gel replicates. Four experiments were performed. *A*, the protein sample (15 μg) was collected in a single band and digested, and the corresponding peptide digest was analyzed three times by nanoLC-MS/MS. *B*, three identical protein samples (15 μg each) were loaded on the gel and collected in three bands, and after digestion, one-third of each corresponding peptide digest was analyzed once by nanoLC-MS/MS. *C*, the protein sample (100 μg) was fractionated by electrophoresis into 12 gel fractions, the 12 bands were digested, and each of the corresponding peptide digests was consecutively analyzed three times by nanoLC-MS/MS. *D*, three identical protein samples (150 μg each) were loaded on the gel and fractionated into 12 gel bands, and after digestion, the corresponding molecular weight bands of each gel lane were consecutively analyzed once by nanoLC-MS/MS (one-third of resulting peptide digests for each band).

analyses of the same sample, without any fractionation. The first experiment consisted in triplicate nanoLC-MS/MS injections of the tryptic digest prepared from one gel band containing the whole protein mixture (Fig. 1*A*). In that case, sources of errors in the final quantitative results include only the variability of the nanoLC separation, of the mass spectrometry measurement, as well as of potential inconsistencies related to bioinformatic extraction of peptide XICs by the software. To evaluate the additional variability related to upstream sample processing steps (gel loading, gel migration, manual band cutting, in-gel trypsin digestion and peptide extraction), three nanoLC-MS/MS analyses were then performed on the tryptic digests obtained from triplicate gel bands containing each the same sample loaded on the gel (Fig. 1*B*). In both cases, the number of proteins identified from the three analytical runs was very similar (respectively 718 and 715 proteins for injection replicates or gel replicates; supplemental data 1). Although some of these proteins were identified by MS/MS in only one or two of the triplicates, the cross-assignment procedure used in MFPaQ allowed extraction of their MS signal in the runs, which did not contain any identification data for these particular proteins. As shown in supplemental data 2, this method generated a very modest number of missing values for quantification, at both the peptide and protein levels, leading to quantification of 715 and 686 proteins in these two experiments. To evaluate repeatability, the CVs of the PAI values obtained for these proteins were calculated. As shown in Fig. 2, the distribution of CVs for proteins quantified in the three gel replicates is very similar to that of CVs obtained with three injection replicates. The median CV is 5 and 6%, respectively, for the two experiments, and the interquartile range of the CV distribution is slightly increased in the case of gel replicates compared with injection replicates. Experimental steps such as gel migration or gel band processing may account for this little decrease of quantification accuracy observed for gel replicates. However, when the sample is isolated in only one band, such processes are supposed to be quite reproducible. Indeed, they seem to bring only a little additional variability, because the results show that a high percentage of the protein population is still correctly quantified (99% of proteins have CVs under 50%), with a relatively small absolute number of outlier proteins with extreme CV values. These results also confirm that label-free quantification using the identity-based signal extraction procedure in MFPaQ allows an accurate quantification of more than 600 proteins on a complex sample analyzed in a single run. This can also be seen from the correlation plots and the distribution of protein PAI ratios calculated between replicate nanoLC-MS/MS analyses (supplemental data 3).

Label-free Quantification after One-dimensional SDS-PAGE Shotgun Analysis—The sample was then submitted to one-dimensional SDS-PAGE and fractionated into 12 gel bands. Again, to assess how repeatability is affected by each step of the analytical process, we performed either three LC-MS/MS analyses of the 12 gel bands from the same migration lane or LC-MS/MS analyses of the gel bands from three replicate migration lanes of the same sample loaded on the gel. In this latter case, the bands within a particular molecular weight from the three lanes were analyzed successively, and peptide identifications from each of them were used to extract XICs in the corresponding bands, by cross-assignment of peptide signals in replicate LC-MS/MS runs. As expected, after fractionation the analytical coverage of the protein mixture was greatly improved, because more than 3500 unique proteins



C

	One band, no fractionation		Fractionation, 12 gel bands	
	injection replicates (1band)	gel replicates (3x1band)	injection replicates (12 bands)	gel replicates (3x12 bands)
Total number of identified proteins	725	696	3553	3590
Number of proteins with CV>50%	3	5	36	66
Number of proteins with CV>100%	0	0	3	8
Maximum CV	69,1	75,6	114,8	145,4
Median CV	5,5	6,0	4,7	7,2
99% CV distribution	37,4	48,4	49,9	61,2
95% CV distribution	19,6	30,8	24,3	32,9
75% CV distribution	9,4	11,9	8,6	13,1

FIG. 2. **Coefficients of variation for protein PAI values between triplicate LC-MS measurements.** The results are shown for experiments without fractionation (one gel band analyzed three times or three replicate gel bands analyzed once by LC-MS) or with one-dimensional SDS-PAGE fractionation (each of the 12 molecular weight fractions from one gel lane analyzed three times consecutively or analysis of three gel lanes). *A*, histogram of CVs distribution in the different experiments. *B*, box and whisker plots showing the dispersion of protein CVs near the median value. The *bottoms* and *tops* of the *boxes* correspond to the 25th and 75th percentiles of the CV distribution, and *whiskers* correspond to the lowest and highest values within 1.5× interquartile range of these limits. Extreme values falling out of the box plots correspond to outliers. *C*, number of proteins and quantitative repeatability in each experiment.

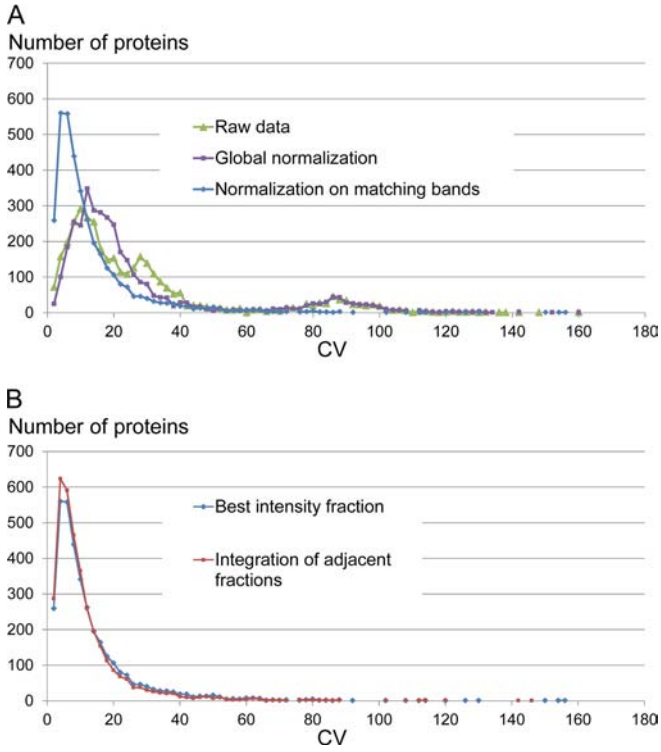


FIG. 3. **Effect of normalization and integration procedures on the quantitative results after SDS-PAGE fractionation.** *A*, histograms of CVs for 1) nontransformed protein PAI values calculated on raw data, 2) normalized protein PAI values transformed with a global correction factor (ratio of summed PAI values for all proteins along each migration lane in replicate experiments), or 3) protein PAI values calculated from normalized XIC signals for each group of matching gel bands. *B*, histograms of CVs for protein PAI values calculated after normalization of XIC signals in matching gel bands, taking into account only the fraction with the best PAI for proteins identified in several gel bands, or after integration by summing PAI values on three consecutive gel bands near the best intensity fraction.

were identified in both experiments (supplemental data 1). Even on this larger population, the signal extraction performed by the software allowed retrieval of quantitative data for almost 99% of the proteins after triplicate sample fractionation through one-dimensional gel (supplemental data 2). Preprocessing of the raw quantitative data was performed to remove the systematic effects and variations because of the measurement process. As for one-shot analysis, a normalization step was used to take into account variability of the nanoLC ESI-MS signal, and in the case of gel replicates, unequal amounts of protein were loaded on the gel and gel processing variability. When this normalization procedure was performed at the scale of the whole experiment (*i.e.*, by comparing signal intensity of all the peptides detected all along the migration lane in replicate experiments), a global correction factor was calculated and used to correct the protein PAI values of replicates experiments against a reference. As shown in Fig. 3*A*, this process improves to some extent the CVs of PAI values for proteins detected in replicate gel lanes but only in a limited way. A significantly better correction was achieved by comparing intensities of peptides detected in matching gel bands, allowing derivation of 12 different normalization factors applied separately to correct quantitative values in each group of molecular weight gel fractions replicates. Obviously, this approach is best suited to correct LC-MS variability, because it compares samples that were measured within a shorter lapse of time and that contain similar protein subpopulations. An automatic normalization

TABLE I
Statistics of peptide migration across the SDS-PAGE fractions

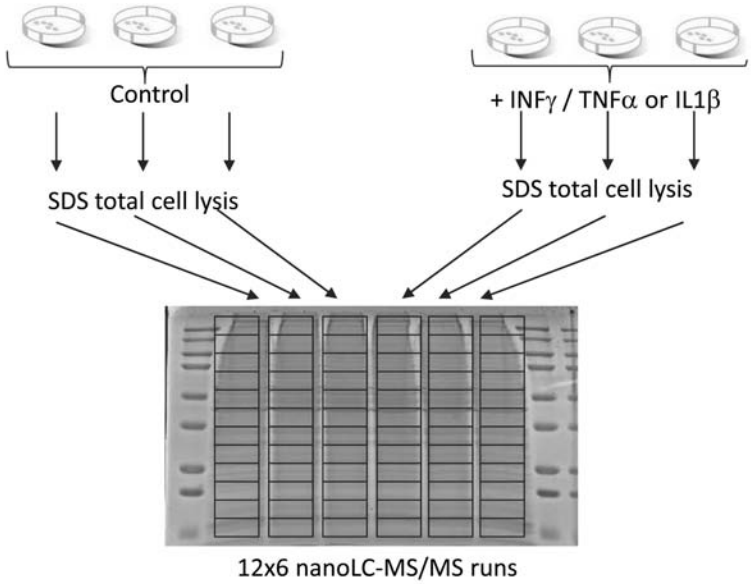
Peptide apex count distribution indicates the number of peptide precursor ions that were detected at their maximal intensity in the same matching fractions across all three replicates (one fraction), in the same fraction in only two replicates (two fractions), or in different fractions in the three replicates (three fractions). Peptide apex distance distribution illustrates the maximal gap between apex fractions when peptides were detected in nonmatching fractions across the replicates.

Peptide apex	No. of precursor ions	
	Three migration lanes	Three injections of one lane
Count distribution		
One fraction	33,932	34,769
Two fractions	1,402	364
Three fractions	91	0
Distance distribution		
0	33,932	34,769
1	1,269	312
2	53	22
3	21	10
4	12	4
5	11	1
6	10	3
7	9	6
8	54	1
9	38	4
10	15	1
11	1	0

procedure to correct peptide intensities of matching fractions in the case of fractionation experiments was thus included in MFPaQ and used in this study.

In addition, to correct for gel migration variability from lane to lane, an integration procedure was also included to sum up the signal of proteins detected in several fractions over the SDS-PAGE lane. However, bands along the migration lane will be corrected with different normalization factors, and integration of signal from very distant gel bands can generate quantitative errors. To evaluate gel migration variability, MFPaQ was used to retrieve the apex of the electrophoretic gel migration pattern for each peptide identified in each replicate experiment. Table I shows the apex count distribution, reflecting the number of peptides that were detected at their maximal intensity in nonmatching fractions in the three different replicates. As expected, the apex of the vast majority of peptide ions was detected in the same gel band, but in the case of replicate gel lanes, for 1402 of the peptide ions (~4% of the total number of precursors) the "best" gel band is identical in only two replicates of three, and for a small minority of them (91 peptide ions, 0.25%), it is different in all three replicates. To some extent, these figures include cases that may be explained by LC-MS variability: indeed, in the case of LC-MS replicates, there is also a small degree of disparity between apex fractions (364 peptide ions, 1% of total, for which the maximal intensity is measured in a nonreproducible way in one injection replicate). However, variability of the electrophoretic migration of proteins along the gel lanes and manual cutting of the fractions account for the majority of the discrepancies in the case of replicate gel lanes. Of the 1493 peptide ions for which conflicts were detected, as shown in Table I, the maximal distance between apex fractions is 1 for 1269 precursors, *i.e.*, the maximal intensity is measured in matching gel bands or in an adjacent band for all three replicates. In many cases, this is probably due to gel cutting inside protein migration patterns and unequal partitioning of these proteins into adjacent gel bands depending on the migration lane. In a small number of other cases, the apex fractions are more distant, probably because of migration problems, irreproducible degradation or precipitation of some proteins, or wrong signal extraction by the software. To correct the most frequent artifacts associated with the SDS-PAGE fractionation process, without introducing additional errors, we thus decided to integrate quantitative data by summing the PAI values for fractions adjacent to the fraction with the best PAI (the same three consecutive fractions for all the replicates to be compared). Fig. 3*B* shows the result of this integration procedure on the CVs of PAI values for proteins detected in replicate gel lanes, compared with CVs calculated by retrieving only the best PAI in one fraction (identified across all the replicates, and the same matching fraction for all of them). Although the distributions are globally very similar, integration brings a small improvement on the CVs, in particular by reducing the number of extreme values (89 proteins of 3585 are measured with a CV higher than 50% when the PAI is retrieved from the best intensity fraction, *versus* 66 proteins when integration is performed). Thus, PAI values were summed up from three consecutive fractions in the case of fractionation experiments.

Finally, as shown in Fig. 2, in the case of sample fractionation, the number of quantified proteins clearly increases, but this is associated with a higher number of extreme values falling of the normal distribution of CVs for both experiments, a significant number of proteins quantified with CVs above 50%, and a higher interquartile range for gel replicates. In the case of replicate injections, quantitative errors occur again from the same causes than in the first one-shot experiment (variations in the nanoLC peptide separation, MS analysis, and bioinformatic processing), and globally, the accuracy of the quantification is thus similar (median CV of ~5% and comparable interquartile ranges). However, the presence of extreme values can be explained by the higher number of low abundance species that are quantified compared with one-shot measurements. Indeed, by increasing the depth of proteome analysis, the fractionation strategy generates quantitative data on low intensity signals that may be subject to larger fluctuations from one run to another or that may be incorrectly extracted by the software. This is illustrated by CV to PAI plots, which reflect a significant decrease of quantitative repeatability for lower PAI values (supplemental data 4). On the other hand, when the 12 gel bands from the three different migration lanes are analyzed independently, additional errors



	INFγ / TNFα experiment						IL1β experiment					
	control1	control2	control3	assay1	assay2	assay3	control1	control2	control3	assay1	assay2	assay3
Number of identified proteins /lane	4044	3881	4095	4029	3981	3907	4471	4506	4542	4691	4533	4387
Total number of quantified proteins	4842						5477					
Variant proteins (fold>2, p<0.05)	207						153					

FIG. 4. **Experimental design and identification results of the large scale quantitative proteomic study of endothelial cells.** Three independent biological experiments were performed by stimulating HUVECs with inflammatory cytokines in culture (either a combination of TNFα and IFNγ or IL1β). Total cell lysates from control and stimulated samples were loaded and fractionated on six parallel gel lanes and cut into 12 gel [GRAPHIC]bands. The table indicates the number of proteins identified for each gel lane, and the total number of proteins identified and quantified in each experiment, as well as the number of proteins detected as differentially expressed.

related to the one-dimensional SDS-PAGE fractionation process (migration and gel band cutting), which was expected to be the most important source of variability, are introduced. The distribution of CVs is shifted compared with what was obtained for injection replicates and now has a median of 7%. Thus, the gel fractionation process contributes to the variability of the measurement. However, even in that case, still 99% of the protein population has CVs for PAI values under 70%. These values illustrate the variability of the gel fractionation process when samples are loaded on adjacent lanes, on the same gel. To further evaluate gel to gel repeatability, which may be an important parameter when numerous samples have to be processed, we also performed triplicate fractionation experiments on different gels. As shown in supplemental data 5, the median CV of proteins PAI shifts from 7% when they are fractionated and quantified on one gel, to 9% when they are quantified from samples fractionated on different gels, and the distribution of CVs is slightly broader. In conclusion, sample fractionation largely improves the depth of proteome coverage, although this is obtained at the expense of quantification accuracy. However, the repeatability of the method is still acceptable for a differential quantitative study, performed with statistical analysis of replicate gel migration lanes.

Large Scale Label-free Quantitative Proteomic Analysis of Human Primary ECs under Inflammatory Conditions—The workflow was then used in the context of a real differential biological analysis, in which we stimulated primary HUVECs with TNFα/IFNγ or IL1β, which represent potent proinflammatory cytokines that trigger inflammatory and immunological responses. The cells were lysed directly in SDS buffer and sonicated, and the resulting protein extract was loaded on a one-dimensional gel. Three biological experiments were performed, with three control samples and three stimulated samples fractionated independently on six migration lanes (Fig. 4). Using the fractionation workflow, we could identify and quantify 4842 and 5477 proteins, respectively, from ECs in the TNFα/IFNγ and IL1β experiments (supplemental data 6 and 7). Statistical analysis was performed on protein PAI values calculated after normalization and integration, as described above. For defining expression changes, two criteria were applied to derive confident data sets of modulated proteins: Student's *t* test *p* value <0.05 and expression fold change >2, as described in previous studies (16). Based on these cut-off values, 207 proteins were found to exhibit a significant variation following TNFα/IFNγ stimulation (175 up-regulated and 32 down-regulated) (supplemental data 6). Endothelial cell response to IL1β stimulation was slightly more restricted, because we measured 153 modulated proteins (119 over-regulated and 34 down-regulated) (supplemental data 7). Functional analysis of modulated proteins using the Protein-

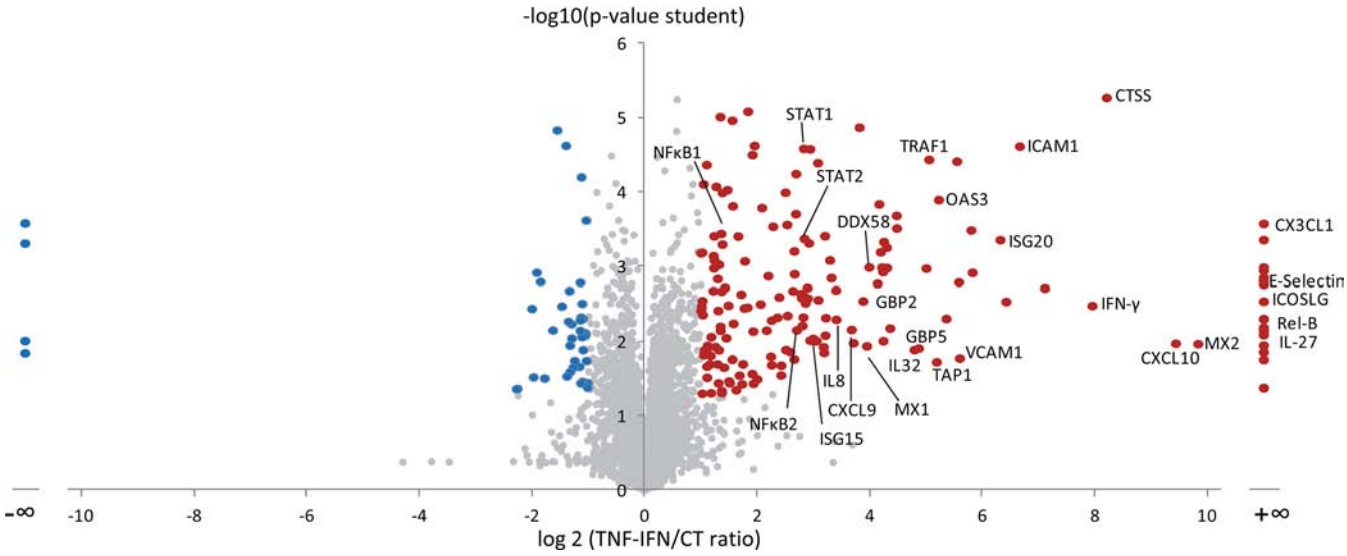


FIG. 5. **Quantitative analysis of endothelial cells proteome following TNFα-IFNγ stimulation.** The volcano plot of the statistical significance of expression level changes (*t* test *p* value) as a function of protein expression ratio between control and inflamed endothelial cells. The *red* and *blue dots* indicate up-regulated and down-regulated genes, respectively.

Center software shows an important enrichment of functional categories related to inflammation and immune response (supplemental data 8). Fig. 5 shows the volcano plot representing statistical significance in function of protein variation between treated and control ECs in the case of TNFα/IFNγ stimulation. Among the most induced proteins, we found many well known cell surface membrane proteins involved in leukocyte recognition and recruitment (E-selectin, ICAM-1, V-CAM1, and ICOSLG), proteins involved in antigen processing and presentation through the class I major histocompatibility complex, but also inflammatory mediators, such as signaling molecules and transcription factors downstream the TNFα pathway (TRAF1, NF-κB, and RELB) or IFNγ pathway (JAK1 and STAT transcription factors), as well as many characteristic interferon-induced proteins involved in antiviral response, as illustrated in Fig. 6. Interestingly, 42 proteins were found to be up-regulated both by IL1β and TNFα/IFNγ (supplemental data 9). Most of them have been described as NF-κB target genes, confirming the role of NF-κB as a key mediator of IL1 and TNFα pathways. To corroborate the results obtained using our quantitative workflow, we tested by quantitative PCR the up-regulation of a series of genes corresponding to modulated proteins identified by the proteomic approach. All of the genes tested confirmed the results of the proteomic study, including strongly induced genes coding for proteins well known to be involved in the inflammatory process and also other genes moderately up-regulated, corresponding to proteins less described in the literature to be part of endothelial cell response to cytokines, such as ROBO1 (supplemental data 10). Altogether, this study shows that the quantitative label-free workflow used here can successfully identify the pathways activated under inflammatory conditions, and it provided a detailed proteomic characterization of the response triggered by inflammatory cytokines in ECs.

DISCUSSION

Global analysis and quantitative comparison of large proteomes is a fruitful approach to get insights into molecular mechanisms of complex biological systems. To obtain a comprehensive picture of such systems, proteomic analysis must be as deep as possible, to map and quantify a large range of protein species, even low abundant ones. Although they have been greatly improved in recent years, the dynamic range and the sequencing speed of mass spectrometers still represent limiting factors for discovery-based proteomics, and in classical experimental LC-MS designs, they restrict the list of proteins that can be detected and quantified in a single-run analysis. To extend the list of identified proteins and obtain quantitative data on minor species, sample prefractionation is thus generally combined to nanoLC-MS analysis, either at the protein level (mainly by SDS-PAGE) or at the peptide level (often by SCX or isoelectric focusing). In recent studies, several thousand proteins could be identified from eukaryotic cells following sample fractionation (1, 17–19). This upstream separation step is often performed on isotopically labeled and mixed samples, ensuring accurate quantification. Here, we evaluated the repeatability of an analytical workflow combining SDS-PAGE fractionation and label-free quantification based on MS signal analysis. Some features of the label-free quantification performed through the MFPaQ software in this study were 1) extraction in raw MS files of XICs from identified and carefully validated peptides, 2) use of a global index for relative quantification at the protein level, derived from the intensity values of at most three intense peptides, and 3)

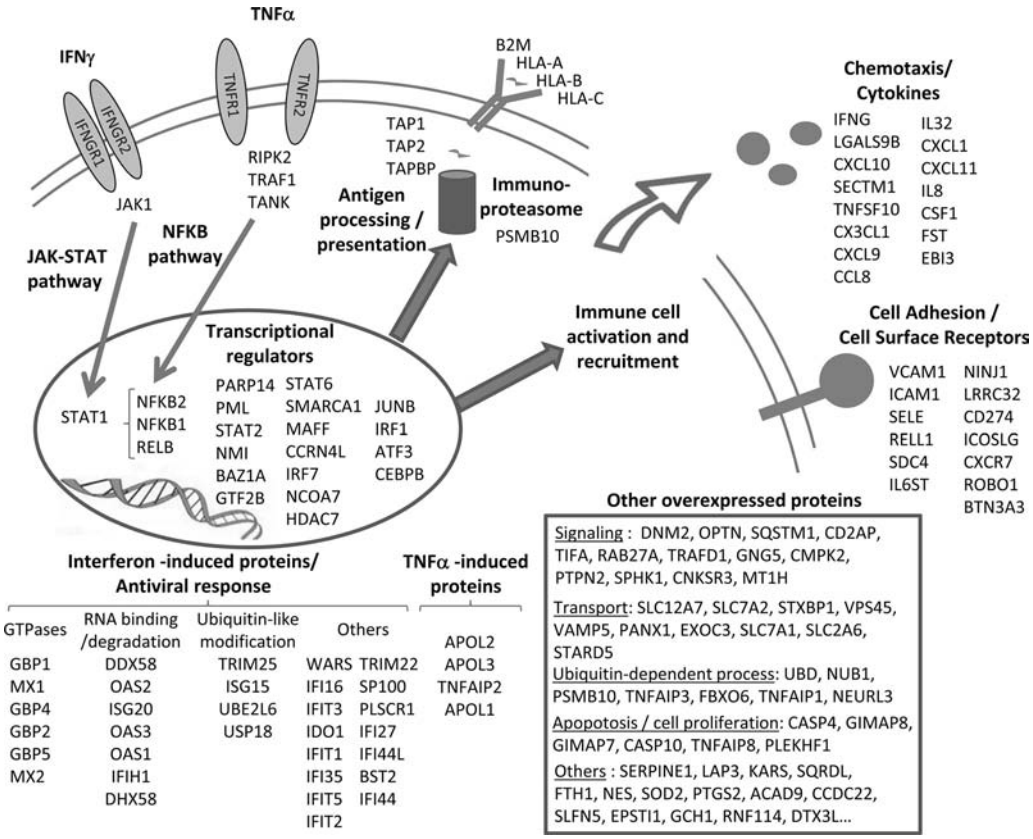


FIG. 6. **Biological pathways activated upon inflammatory response of endothelial cells after TNFα/IFNγ stimulation.** Proteins found to be up-regulated from the proteomic analysis are illustrated and classified in function of the biological processes in which they are involved or their subcellular location, according to literature data.

integration of the quantitative data from different fractions and overview of the shotgun experiment through the MFPaQ interface. The identity-based approach allowed extraction of signal for confidently identified peptides in an automated batch mode, directly on the 72 raw files of the comparative experiment (two conditions, three replicates, and twelve fractions). Quantitative data can be viewed at the peptide level through the MFPaQ interface, which displays the XICs of the peptide ions in all of the raw files corresponding to matched fractions but was also directly integrated by the software at the protein level, by computing the mean value of at most three intense peptides per protein. In the case of relatively abundant proteins identified with more than three peptides, this allowed calculation of PAI values on the highest quality signals, for a more accurate quantification. However, minor species identified with less than three peptides were also quantified based on the available peptide signals. Finally, data normalization and integration procedures were used in MFPaQ to correct LC-MS variability and errors related to nonreproducible electrophoretic migration of proteins in the case of sample fractionation.

Overall, our approach proved to behave in a robust way for the quantification of a complex proteome. Our results show that for one-shot analysis, label-free quantification can be achieved with good accuracy (median CV of 5%, 99% proteins with CV values $<$ 48%). Sample fractionation largely improved the depth of proteomic coverage, and this was associated with a moderate decrease of quantitative measurement repeatability (median CV of 7%, 99% proteins with CV values of $<$62%). Thus, prefractionation by SDS-PAGE appears to be compatible with label-free quantitation for the extensive analysis of complex proteomes. In the present study, it provided a detailed characterization of the proteomic variations associated with the inflammatory response in human primary ECs. Although they can be maintained for some time in culture, these primary cells are not easily amenable to SILAC labeling, and label-free methods were particularly convenient for their quantitative analysis. For each condition (control or stimulated cells), triplicate samples were fractionated by SDS-PAGE, and analysis of each gel lane led to the identification of up to 4600 unique proteins, based on a protein FDR of 1%. Globally, analysis of the six different gel lanes by nanoLC-MS/MS and cross-assignment of peptide signals between samples led to the identification and quantitation of more than 5400 unique proteins in the IL1β experiment. In a recent study, the use of very long LC-MS gradients on 50-cm-long columns was described for in-depth analysis of complex proteomes without prefractionation (20). Such experi-

mental strategies are probably not yet routinely applicable because they generally require high pressure chromatography devices and generate very large raw files that may be difficult to process with most current bioinformatic tools. However, they appear to be a promising approach that in principle could combine the advantages of extensive proteomic coverage and quantitative accuracy that can be obtained in single-run analysis. However, the quantitative repeatability of these several-hours-long LC-MS gradients is still to be assessed on replicate analytical runs. Regarding analytical time, analysis of 12 fractions of a one-dimensional gel lane in 2-h-long LC-MS gradients on conventional LC systems is three times longer, but technically easier to implement, than analysis of the whole sample on a long column with an 8-h gradient. On the other hand, sample prefractionation still probably represents up to now the most efficient way to get the deepest analytical coverage of a complex proteome, and the present study shows that the additional variability associated with this upstream process does not preclude quantitative analysis. Thus, although there is a trade-off between analytical time, quantitative accuracy, and proteomic coverage, putting the emphasis on this last parameter would probably require both sample fractionation and extensive peptide separation with long gradients for very extensive characterization of complex proteomes like the human one. Here, by using only a shotgun approach based on one-dimensional protein fractionation, sufficient depth was obtained here to detect changes on very low abundance proteins such as some transcription factors and signaling molecules. Although this label-free approach requires more analytical time than a SILAC-based experiment, because the samples are injected separately, it also avoids possible quantitation errors caused by superposition of one peptide of the SILAC pair with other different isotopic peptide patterns. In our hands, it also yielded a higher number of identified proteins, because the same MS/MS sequencing time is spent on less complex mixtures, containing half the number of peaks compared with isotopically labeled and mixed samples. Thus, MS intensity-based label-free quantification associated with SDS-PAGE fractionation appears as a valuable strategy for the differential analysis of complex proteomes.

The shotgun approach used in our study provided an in-depth characterization of the EC proteome and the label-free quantitative proteomic workflow allowed deciphering the inflammatory response of these cells. TNFα and IFNγ are potent pleiotropic cytokines that exert a number of biological effects and trigger a set of complex molecular programs in response to microbial or viral infection. IFNγ is produced mainly by NK cells and T helper type I cells and, through binding to its specific type II IFN receptor, activates the JAK-STAT signaling pathway, to induce the expression of a large number of genes (21, 22). In this large scale proteomic experiment, we measured an up-regulation of the JAK1 kinase and STAT1 transcription factor, which are known to mediate IFNγ response and regulate genes downstream of γ-activated sequence elements. We could also detect an increase of proteins involved in the TNFα signal transduction pathway, such as TRAF1, TANK, and RIPK2, converging to the activation of the NF-κB transcription factor (23). Accordingly, we measured overexpression of NF-κB subunits (NFKB2, NFKB1, and RELB) and a decrease of the inhibitor of NF-κB (IKBB), which controls nuclear translocation of NF-κB and undergoes proteasomal degradation upon TNFα signaling (24). In addition, several other proteins involved in transcriptional regulation were shown to be up-regulated after stimulation by the two cytokines (Fig. 6), such as members of the STAT family (STAT2 and STAT6); the PARP-14 protein, which enhances STAT6-dependent transcription (25); or the IRF1 secondary transcription factor, which is induced by STAT1 and plays a key role in orchestrating the IFN-induced inflammatory response (26).

One major biological process that makes part of this response in ECs is recruitment and activation of leukocytes to the inflammatory site. ECs line the blood vessel walls, and upon stimulation by cytokines, they secrete chemokines, which are chemoattractants for lymphocytes and monocytes, and express at their surface adhesion molecules that capture circulating leukocytes. We measured in our analysis the strong up-regulation of a panel of chemokines, such as Fractalkine, IL8, CXCL10, CXCL11, CXCL9, CXCL1, or CCL8, and of other secreted signaling molecules such as IL27b, or IL32. Indeed, IL32 was recently shown to be a critical regulator of EC function, which is strongly increased upon IL1β or TNFα stimulation and mediates in particular the expression of cell surface adhesion molecules involved in lymphocytes binding such as VCAM1 (27). Although our analysis was performed on a whole cell lysate and not focused on membrane proteins, we could clearly measure the overexpression of several cell surface proteins involved in cell-cell interactions. Leukocyte adhesion molecules such as E-selectin, ICAM1, and VCAM1 were the among the most strongly induced gene products and represent major players in the initial rolling and arrest step of leukocytes-EC interaction along the blood vessel walls (28). Simultaneously, molecules known to promote procoagulant activity at the EC surface such as plasminogen activator inhibitor 1 (Serpine 1) were also induced (29). Other cell surface proteins were shown to be overexpressed in response to TNFα/IFNγ treatment, such as the ICOS-ligand protein, which is an important costimulator in EC-mediated T cell activation (30); the ROBO1 receptor that may play a role in leukocyte migration (31); or the programmed cell death 1 receptor ligand PDL1, involved in immune regulation (32). Additionally, the expression of cell surface class I MHC molecules was also increased upon stimulation by inflammatory cytokines. ECs constitutively express class I MHC molecules *in vivo*, which are significantly decreased during cell culture but can be restored upon IFNγ or TNFα treatment (33). Following stimulation, we observed concomitantly the induction of all the machinery for antigen processing and presentation, including

the immunoproteasome responsible for degradation of cytoplasmic endogenous or viral proteins; TAP proteins involved in antigenic peptide transport to the endoplasmic reticulum; and Tapasin, which binds to the TAP complex and allows antigen loading to assembled MHC molecules (34). Finally, a wide range of interferon-induced proteins were detected as strongly up-regulated, such as small GTPases (guanylate-binding proteins, Mx1, and Mx2) and the 2′-5′-oligoadenylate synthase family, which play an essential role on viral RNA degradation and the innate immune response to viral infection (21).

The EC response to IL1β stimulation, as characterized from the second large scale proteomic experiment, shared many features with the response induced by the TNFα/IFNγ treatment. Major biological processes of EC inflammatory activation were again highlighted, *i.e.*, secretion of chemoattractant molecules and other cytokines (CXCL6, interleukin 8, CXCL1, CXCL2, CCL2, granulocyte colony-stimulating factor, macrophage colony-stimulating factor, interleukin 27, and interleukin 32), expression of cell surface leukocyte ligands (ICAM1, VCAM1, selectin, ICOS ligand, Syndecan-4), as well as antigen processing and presentation through MHC class I molecules (immunoproteasome subunits, TAP1, Tapasin, and HLA molecules). As IL1β signals through the NF-κB pathway, many proteins induced by IL1β were also induced by TNFα (see supplemental data 9). Both cytokine are, for example, endogenous pyrogens that cause fever, and in both experiments, we found an up-regulation of the prostaglandin G/H synthase 2 of the prostaglandin G/H synthase 2 (cyclooxygenase-2, COX2), which is responsible for synthesis of prostaglandin E_2 prostaglandin, the key molecule for activation of thermosensitive neurons in the hypothalamus (35, 36). Additionally, in the IL1β experiment, we detected the induction of phospholipase A_2, which hydrolyzes glycerophospholipids to produce arachidonic acid, the rate-limiting step in the synthesis of prostaglandin E_2 by COX2. In this experiment, we could also specifically detect the induction of the cysteine protease caspase 1, which is directly involved in cleavage of proactive IL1β into its mature form, as well new regulatory molecules such as TC1, which has been described as a novel endothelial inflammatory regulator that is up-regulated by IL1β and amplifies NF-κB signaling via a positive feedback (37).

Many proteins could be identified that were not previously described as activated in ECs, deserving further studies to determine their exact function in the inflammatory process. For example, the ROBO1 receptor protein has been described to be involved in axon guidance and neuronal precursor cell migration (38), but its potential role in mediating cell-cell interactions at the endothelial surface under inflammatory conditions has been poorly described (39). Here, we show that this protein is overexpressed in HUVECs after TNFα/IFNγ stimulation, and the induction of the corresponding gene was confirmed by quantitative PCR for both TNFα/IFNγ and IL1β treatment. Another example is the circadian deadenylase Nocturnin, which was found to be significantly induced by both TNFα/IFNγ and IL1β stimulations. This protein, that is under circadian regulation, can also mediate immediate early gene responses, and it has been hypothesized that it could be involved in the post-transcriptional regulation of both rhythmic and acutely inducible mRNAs, by controlling mRNA decay through poly(A) tail removal (40). Indeed, very recently it was shown that Nocturnin can be induced by endotoxin lipopolysaccharide and that it stabilizes the proinflammatory transcript inducible nitric-oxide synthase (40), suggesting that Nocturnin could play a role in the circadian response to inflammatory signals. The proteomic data obtained here indicate that it is also induced in endothelial cells upon stimulation with TNFα/IFNγ and IL1β and thus support the idea that this protein could play a general role in the regulation of cytokine-induced inflammatory response.

In conclusion, this is the most extensive proteomic study of EC to date, performed on the widely used *in vitro* primary endothelial cell model HUVEC. It allowed identification in these endothelial cells of more than 5400 proteins, adding some more depth to a large scale data set previously published (41), in which ~3800 proteins were identified and 1300 proteins could be quantified by ^{18}O labeling, following treatment with the proangiogenic factor vascular endothelial growth factor. The present study provides the first complete characterization at the proteomic level of the EC response to inflammatory cytokines such as TNFα, IFNγ, and particularly IL1β. The list of proteins modulated by these factors, as characterized here in a global way, can thus represent a reference to study the function of other newly discovered interleukins of the IL1 family that may trigger similar responses but also some specific pathways.

* This work was supported by grants from the Agence Nationale de la Recherche (Programme Plates-formes Technologiques du Vivant); Fondation pour la Recherche Médicale (Programme Grands Equipements); Ibisa (Infrastructures en Biologie, Santé, et Agronomie); and FEDER (Fonds Européen de Développement Régional); and a fellowship from Région Midi-Pyrénées (to N. D.).

[S] This article contains supplemental material.

¶ These authors contributed equally to this study.

‖ Supported by the "Ligue Nationale contre le Cancer" (LIGUE 2009).

** To whom correspondence may be addressed: Institut de Pharmacologie et de Biologie Structurale, 205 route de Narbonne, 31077 Toulouse cedex 4, France. E-mail: bernard.monsarrat@ipbs.fr.

‡‡ To whom correspondence may be addressed: Institut de Pharmacologie et de Biologie Structurale, 205 route de Narbonne, 31077 Toulouse cedex 4, France. E-mail: gonzalez@ipbs.fr.

REFERENCES

1. Wiśniewski, J. R., Zougman, A., Nagaraj, N., and Mann, M. (2009) Universal sample preparation method for proteome analysis. *Nat. Methods* **6,** 359–362
2. Vaudel, M., Sickmann, A., and Martens, L. (2010) Peptide and protein quantification: A map of the minefield. *Proteomics* **10,** 650–670
3. Park, S. K., Venable, J. D., Xu, T., and Yates, J. R., 3rd (2008) A quantitative analysis software tool for mass spectrometry-based proteomics. *Nat. Methods* **5,** 319–322

4. Searle, B. C. (2010) Scaffold: A bioinformatic tool for validating MS/MS-based proteomic studies. *Proteomics* **10,** 1265–1269
5. Heinecke, N. L., Pratt, B. S., Vaisar, T., and Becker, L. (2010) PepC: Proteomics software for identifying differentially expressed proteins based on spectral counting. *Bioinformatics* **26,** 1574–1575
6. Bellew, M., Coram, M., Fitzgibbon, M., Igra, M., Randolph, T., Wang, P., May, D., Eng, J., Fang, R., Lin, C., Chen, J., Goodlett, D., Whiteaker, J., Paulovich, A., and McIntosh, M. (2006) A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS. *Bioinformatics* **22,** 1902–1909
7. Jaitly, N., Mayampurath, A., Littlefield, K., Adkins, J. N., Anderson, G. A., and Smith, R. D. (2009) Decon2LS: An open-source software package for automated processing and visualization of high resolution mass spectrometry data. *BMC Bioinformatics* **10,** 87
8. Mueller, L. N., Rinner, O., Schmidt, A., Letarte, S., Bodenmiller, B., Brusniak, M. Y., Vitek, O., Aebersold, R., and Müller, M. (2007) SuperHirn: A novel tool for high resolution LC-MS-based peptide/protein profiling. *Proteomics* **7,** 3470–3480
9. Sturm, M., Bertsch, A., Gröpl, C., Hildebrandt, A., Hussong, R., Lange, E., Pfeifer, N., Schulz-Trieglaff, O., Zerck, A., Reinert, K., and Kohlbacher, O. (2008) OpenMS: An open-source software framework for mass spectrometry. *BMC Bioinformatics* **9,** 163
10. Old, W. M., Meyer-Arendt, K., Aveline-Wolf, L., Pierce, K. G., Mendoza, A., Sevinsky, J. R., Resing, K. A., and Ahn, N. G. (2005) Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol. Cell. Proteomics* **4,** 1487–1502
11. Hoffert, J. D., Wang, G., Pisitkun, T., Shen, R. F., and Knepper, M. A. (2007) An automated platform for analysis of phosphoproteomic datasets: Application to kidney collecting duct phosphoproteins. *J. Proteome Res.* **6,** 3501–3508
12. Tsou, C. C., Tsai, C. F., Tsui, Y. H., Sudhir, P. R., Wang, Y. T., Chen, Y. J., Chen, J. Y., Sung, T. Y., and Hsu, W. L. (2010) IDEAL-Q, an automated tool for label-free quantitation analysis using an efficient peptide alignment approach and spectral data validation. *Mol. Cell. Proteomics* **9,** 131–144
13. Bouyssié, D., Gonzalez de Peredo, A., Mouton, E., Albigot, R., Roussel, L., Ortega, N., Cayrol, C., Burlet-Schiltz, O., Girard, J. P., and Monsarrat, B. (2007) Mascot file parsing and quantification (MFPaQ), a new software to parse, validate, and quantify proteomics data generated by ICAT and SILAC mass spectrometric analyses: Application to the proteomics study of membrane proteins from primary human endothelial cells. *Mol. Cell. Proteomics* **6,** 1621–1637
14. Mouton-Barbosa, E., Roux-Dalvai, F., Bouyssié, D., Berger, F., Schmidt, E., Righetti, P. G., Guerrier, L., Boschetti, E., Burlet-Schiltz, O., Monsarrat, B., and Gonzalezde Peredo, A. (2010) In-depth exploration of cerebrospinal fluid by combining peptide ligand library treatment and label-free protein quantification. *Mol. Cell. Proteomics* **9,** 1006–1021
15. Navarro, P., and Vázquez, J. (2009) A refined method to calculate false discovery rates for peptide identification using decoy databases. *J. Proteome Res.* **8,** 1792–1796
16. Sun, N., Pan, C., Nickell, S., Mann, M., Baumeister, W., and Nagy, I. (2010) Quantitative proteome and transcriptome analysis of the archaeon *Thermoplasma acidophilum* cultured under aerobic and anaerobic conditions. *J. Proteome Res.* **9,** 4839–4850
17. Graumann, J., Hubner, N. C., Kim, J. B., Ko, K., Moser, M., Kumar, C., Cox, J., Schöler, H., and Mann, M. (2008) Stable isotope labeling by amino acids in cell culture (SILAC) and proteome quantitation of mouse embryonic stem cells to a depth of 5,111 proteins. *Mol. Cell. Proteomics* **7,** 672–683
18. Geiger, T., Cox, J., and Mann, M. (2010) Proteomic changes resulting from gene copy number variations in cancer cells. *PLoS Genet.* **6,** pii
19. Luber, C. A., Cox, J., Lauterbach, H., Fancke, B., Selbach, M., Tschopp, J., Akira, S., Wiegand, M., Hochrein, H., O'Keeffe, M., and Mann, M. (2010) Quantitative proteomics reveals subset-specific viral recognition in dendritic cells. *Immunity* **32,** 279–289
20. Thakur, S. S., Geiger, T., Chatterjee, B., Bandilla, P., Froehlich, F., Cox, J., and Mann, M. (2011) Deep and highly sensitive proteome coverage by LC-MS/MS without pre-fractionation. *Mol. Cell. Proteomics* 10.1074/mcp.M110.003699
21. Boehm, U., Klamp, T., Groot, M., and Howard, J. C. (1997) Cellular responses to interferon-γ. *Annu. Rev. Immunol.* **15,** 749–795
22. Stark, G. R., Kerr, I. M., Williams, B. R., Silverman, R. H., and Schreiber, R. D. (1998) How cells respond to interferons. *Annu. Rev. Biochem.* **67,** 227–264
23. Baud, V., and Karin, M. (2001) Signal transduction by tumor necrosis factor and its relatives. *Trends Cell Biol.* **11,** 372–377
24. Traenckner, E. B., Pahl, H. L., Henkel, T., Schmidt, K. N., Wilk, S., and Baeuerle, P. A. (1995) Phosphorylation of human IκB-α on serines 32 and 36 controls IκB-α proteolysis and NF-κB activation in response to diverse stimuli. *EMBO J.* **14,** 2876–2883
25. Mehrotra, P., Riley, J. P., Patel, R., Li, F., Voss, L., and Goenka, S. (2011) PARP-14 functions as a transcriptional switch for Stat6-dependent gene activation. *J. Biol. Chem.* **286,** 1767–1776
26. Taniguchi, T., Ogasawara, K., Takaoka, A., and Tanaka, N. (2001) IRF family of transcription factors as regulators of host defense. *Annu. Rev. Immunol.* **19,** 623–655
27. Nold-Petry, C. A., Nold, M. F., Zepp, J. A., Kim, S. H., Voelkel, N. F., and Dinarello, C. A. (2009) IL-32-dependent effects of IL-1β on endothelial cell functions. *Proc. Natl. Acad. Sci. U.S.A.* **106,** 3883–3888
28. Springer, T. A. (1994) Traffic signals for lymphocyte recirculation and leukocyte emigration: The multistep paradigm. *Cell* **76,** 301–314
29. van Hinsbergh, V. W., Kooistra, T., van den Berg, E. A., Princen, H. M., Fiers, W., and Emeis, J. J. (1988) Tumor necrosis factor increases the production of plasminogen activator inhibitor in human endothelial cells *in vitro* and in rats *in vivo*. *Blood* **72,** 1467–1473
30. Khayyamian, S., Hutloff, A., Büchner, K., Gräfe, M., Henn, V., Kroczek, R. A., and Mages, H. W. (2002) ICOS-ligand, expressed on human endothelial cells, costimulates Th1 and Th2 cytokine secretion by memory CD4+ T cells. *Proc. Natl. Acad. Sci. U.S.A.* **99,** 6198–6203
31. Prasad, A., Qamri, Z., Wu, J., and Ganju, R. K. (2007) Slit-2/Robo-1 modulates the CXCL12/CXCR4-induced chemotaxis of T cells. *J. Leukocyte Biol.* **82,** 465–476
32. Singh, A. K., Stock, P., and Akbari, O. (2011) Role of PD-L1 and PD-L2 in allergic diseases and asthma. *Allergy* **66,** 155–162
33. Lapierre, L. A., Fiers, W., and Pober, J. S. (1988) Three distinct classes of regulatory cytokines control endothelial cell MHC antigen expression. Interactions with immune γ interferon differentiate the effects of tumor necrosis factor and lymphotoxin from those of leukocyte α and fibroblast β interferons. *J. Exp. Med.* **167,** 794–804
34. Li, S., Paulsson, K. M., Chen, S., Sjögren, H. O., and Wang, P. (2000) Tapasin is required for efficient peptide binding to transporter associated with antigen processing. *J. Biol. Chem.* **275,** 1581–1586
35. Dinarello, C. A. (1999) Cytokines as endogenous pyrogens. *J. Infect. Dis.* **179,** (Suppl. 2) S294–S304
36. Dinarello, C. A., Gatti, S., and Bartfai, T. (1999) Fever: Links with an ancient receptor. *Current biology* **9,** R147–R150
37. Kim, J., Kim, Y., Kim, H. T., Kim, D. W., Ha, Y., Kim, J., Kim, C. H., Lee, I., and Song, K. (2009) TC1(C8orf4) is a novel endothelial inflammatory regulator enhancing NF-κB activity. *J. Immunol.* **183,** 3996–4002
38. Wong, K., Park, H. T., Wu, J. Y., and Rao, Y. (2002) Slit proteins: Molecular guidance cues for cells ranging from neurons to leukocytes. *Curr. Opin. Genet. Dev.* **12,** 583–591
39. Legg, J. A., Herbert, J. M., Clissold, P., and Bicknell, R. (2008) Slits and roundabouts in cancer, tumour angiogenesis and endothelial cell migration. *Angiogenesis* **11,** 13–21
40. Niu, S., Shingle, D. L., Garbarino-Pico, E., Kojima, S., Gilbert, M., and Green, C. B. (2011) The circadian deadenylase Nocturnin is necessary for stabilization of the iNOS mRNA in mice. *PLoS One* **6,** e26954
41. Jorge, I., Navarro, P., Martínez-Acedo, P., Núñez, E., Serrano, H., Alfranca, A., Redondo, J. M., and Vázquez, J. (2009) Statistical model to analyze quantitative proteomics data obtained by $^{18}O/^{16}O$ labeling and linear ion trap mass spectrometry: Application to the study of vascular endothelial growth factor-induced angiogenesis in endothelial cells. *Mol. Cell. Proteomics* **8,** 1130–1149

Développement de nouveaux outils bioinformatiques pour l'exploitation des données de spectrométrie de masse en protéomique haut-débit

David Bouyssié

Résumé

En biologie, la spectrométrie de masse est devenue l'outil incontournable pour l'identification des protéines. Associée à des techniques de séparation, elle est aussi utilisée pour mesurer la variation d'abondance des protéines entre plusieurs échantillons. Cependant, la très grande quantité et complexité des données liées à ce type d'analyse requièrent des programmes informatiques sophistiqués et adaptés.

Mon travail de doctorat a consisté à répondre aux différentes problématiques liées à l'exploitation des données nanoLC-MS/MS, à savoir la validation des résultats d'identification ainsi que la quantification relative des protéines pour des approches mettant en œuvre ou non un marquage isotopique. Le logiciel MFPaQ, dont deux versions sont présentées dans ce document, en est le principal résultat. La version 3 intègre des fonctionnalités telle que la validation des données Mascot, la génération de listes non-redondantes de protéines et la quantification d'analyses ICAT. La version 4, évolution majeure du logiciel, incorpore des algorithmes adaptés à l'analyse quantitative de données MS sans marquage, ainsi que la gestion des stratégies de marquage SILAC et $^{14}N/^{15}N$. Son utilisation a facilité la réalisation d'études protéomiques, dont certaines, auxquelles j'ai plus particulièrement participé, sont présentées. Afin de répondre aux futurs enjeux informatiques de la protéomique, j'ai entrepris dans un second temps le développement du logiciel Prosper, qui dispose d'une architecture d'organisation des données permettant de réaliser des requêtes croisées sur l'ensemble des échantillons analysés. Il constitue aussi un outil prototype pour l'élaboration de nouveaux algorithmes.

Mots-clés : Protéomique, Spectrométrie de masse, nanoLC-MS/MS, Mascot, Quantification de mélanges complexes de protéines, ICAT, SILAC, 14N/15N, label-free, carte LC-MS, logiciel.

Abstract

In biology, mass spectrometry has become an indispensable tool for protein identification. Associated with separation techniques, it can also be used to measure the variation of protein abundance between different samples. However, due to the huge quantity and complexity of the data produced by this kind of analysis, sophisticated and suitable computer programs are needed.

My PhD work was to address the different issues related to the processing of nanoLC-MS/MS data, namely the validation of the identification results, and the relative quantification of proteins using approaches based or not on isotopic labeling. The MFPaQ program, two versions of which are presented here, is the main result of this work. Version 3 includes features such as Mascot data validation, generation of non-redundant protein lists and quantification of ICAT analyses. Version 4, which represents a major upgrade of the software, incorporates additional algorithms for quantitative analysis of label-free MS data, as well as for the handling of the $^{14}N/^{15}N$ and SILAC labeling strategies. This bioinformatic tool has been used for various proteomic studies, some of which are discussed here. In order to meet future IT challenges in proteomics, I undertook later the development of the Prosper software, which is based on an optimized architecture for organizing data, and allows performing cross-queries on all analysed samples. It also constitutes a prototype tool for the development and evaluation of new algorithms.

Keywords: Proteomics, Mass spectrometry, nanoLC-MS/MS, Mascot, complex protein mixture quantification, ICAT, SILAC, 14N/15N, label-free, LC-MS map, software.