THESE

présentée pour l'obtention du grade de

DOCTEUR DE L'UNIVERSITE LOUIS PASTEUR

DE STRASBOURG

Par

Christine CARAPITO

Vers une meilleure utilisation des données de spectrométrie de masse en analyse protéomique

Soutenue le 17 novembre 2006 devant la commission d'examen :

Dr. Alain VAN DORSSELAER	Directeur de thèse
Dr. Emmanuelle LEIZE	Co-directeur de thèse
Prof. Claude KEDINGER	Président et rapporteur interne
Dr. Jérôme GARIN	Rapporteur externe
Prof. Arnaud DUCRUIX	Rapporteur externe

A mes parents,

« L'éducation est une chose admirable, mais il est bon de se souvenir de temps en temps que rien de ce qui est digne d'être connu ne peut s'enseigner. »

Oscar Wilde

Merci !

Ce travail de thèse a été réalisé au Laboratoire de Spectrométrie de Masse Bio-Organique de l'Institut Pluridisciplinaire Hubert Curien de Strasbourg (UMR 7178, CNRS-ULP).

Je souhaite en premier lieu adresser ma profonde reconnaissance à Alain Van Dorsselaer pour m'avoir accueillie dans son laboratoire, pour m'avoir offert un cadre de travail « de rêve » et surtout pour sa présence, son soutien et la confiance qu'il m'a accordés. Mes plus vifs remerciements vont également à Emmanuelle Leize pour son encadrement, ses encouragements et sa confiance.

J'adresse mes vifs remerciements à Monsieur Claude Kédinger qui a accepté de présider mon jury de thèse ainsi qu'à Messieurs Jérôme Garin et Arnaud Ducruix, rapporteurs externes, qui ont consacré de leur temps à l'évaluation de ce travail.

Je remercie également l'ensemble des personnes avec lesquelles j'ai été amenée à collaborer et qui m'ont entraînée dans des thématiques aussi diverses que passionnantes, dans le désordre : J.P. Brizard et C. Brugidou, A. Castro et C. Clément, D. Muller et P. Bertin, V. Phalip et J.M. Jeltsch, R. Bernhardt, Susanne, Hoon et Britta, M. Schöller, N. Bühr et S. Viville, J.F. Launay, H. Ludwig et L. Bode, P. Cailliéret et M. Gobert, …

J'adresse mes chaleureux remerciements au CNRS et à la société Bruker Daltonics pour le financement de cette thèse.

Un grand merci à tous ceux avec qui j'ai partagé ces trois années au LSMBO pour les échanges, l'ambiance et la bonne humeur : Agnès, Andy, Audrey, Cédric, Christelle, Christine, Danièle, Dimitri, Elsa, Fabrice B., Fabrice V., Flavie, Grand Fred, Guillaume B., Guillaume C., Haiko, Hélène, Jean-Marc, Jonathan, Laetitia, Laurent, Marie-Laure, Nathalie, Noëlle, Ptit Fred, Raymond, Rossella, Sarah, Sébastien, Sophie, Véronique T., Véronique D., Virginie, …

Un grand merci à mon entourage et mes proches amis.

Un immense merci à ma « garde rapprochée », à mes plus fidèles supporters pour leur présence et leur amour, à savoir mes grands frères et mes parents.

Et enfin, mes plus affectueux remerciements à mon cher binôme et soutien quotidien.

PLAN GENERAL

PLAN DETAILLE

RESULTATS

PARTIE I : Amélioration de la prise de données MS/MS

Chapitre I : Le couplage nanoHPLC-trappe ionique 67

Chapitre II : Les nouveautés instrumentales du couplage : le système nanoHPLC-Chip Cube et la fragmentation ETD ... 79

PARTIE III : Applications portant sur des organismes non séquencés : le séquençage *de novo*

Chapitre I : L'identification des protéines par séquençage *de novo* 122

Chapitre II : Identification des protéines impliquées dans la résistance à l'arsenic chez la bactérie *Caenibacter arsenoxydans* 134

Publication : Identification of genes and proteins involved in the pleiotropic response to arsenic stress of Caenibacter arsenoxydans, a metalloresistant beta-proteobacterium with an unsequenced genome.
Carapito C, Muller D, Turlin E, Koechler S, Danchin A, Van Dorsselaer A, Leize-Wagner E, Bertin PN, Lett MC, *Biochimie*, 88, 595-606.

Chapitre III : Analyse protéomique de tissus de vigne (*Vitis vinifera L.*) ayant subi un stress herbicide 138

LISTE DES PRINCIPALES ABBREVIATIONS

ADH :	Alcohol Dehydrogenase
ADN :	Acide désoxyribonucléique
ARN :	Acide ribonucléique
BLAST :	Basic Local alignment Search Tool
BSA :	Bovine Serum Albumin
CAD :	Collision Activated Dissociation
CDS :	CoDing Sequence
CID :	Collision-Induced Dissociation
CRM :	Charged Residue Mechanism
Da :	Dalton
ECD :	Electron Capture Dissociation
ESI :	Electrospray Ionisation
ETD :	Electron Transfer Dissociation
FT-ICR :	Fourier Transform Ion Cyclotron Resonance
FWHM :	Full Width at Half Maximum
HPLC :	High Performance Liquid Chromatography
ICAT :	Isotope-Coded Affinity Tag
IEF :	Iso-Electrofocalisation
IEM :	Ion Evaporation Mechanism
IT :	Ion Trap
LC :	Liquid Chromatography
MALDI :	Matrix-Assisted Laser Desorption/Ionisation
MS :	Spectrométrie de Masse
MS/MS :	Spectrométrie de masse en tandem
MudPIT :	Multi-Dimensional Protein Identification Technology
Nano-ESI :	Nano-Electrospray
NanoLC-MS/MS :	Nano liquid chromatography tandem mass spectrometry
PCR :	Polymerase Chain Reaction
PMF :	Peptide Mass Fingerprinting
PTGS :	Post-transcriptional Gene Silencing
PTM :	Post-translational Modification
Q :	Quadripôle
RP :	Reverse Phase
SCX :	Strong Cation Exchange
SILAC :	Stable Isotope Labelling with Amino acids in Cell culture
SDS-PAGE :	Sodium DodecylSulfate PolyAcrylamide Gel Electrophoresis
TIGR :	Institute for Genomic Research
TOF :	Time of Flight

PRESENTATION DE CE TRAVAIL DE THESE

PRESENTATION DE CE TRAVAIL DE THESE

Dès le début des années 90, la spectrométrie de masse a commencé à jouer un rôle de plus en plus important en biologie. Elle compte actuellement parmi les méthodologies les plus prometteuses permettant des progrès majeurs dans la compréhension de nombreux processus biologiques. L'attribution du prix Nobel de chimie en 2002 à John B. Fenn et Koichi Tanaka pour le développement des méthodes d'ionisation douces ayant révolutionné l'analyse de macromolécules biologiques, respectivement, l'ESI (électrospray) et MALDI (ionisation/désorption laser assistée par matrice) illustre l'importance croissante de la spectrométrie de masse dans ce domaine.

L'année 1994 est marquée par les premières utilisations des termes « protéome » et « analyse protéomique par spectrométrie de masse ». Et jusqu'à ce jour, des progrès conjoints dans plusieurs domaines ont permis l'émergence de l'analyse protéomique par spectrométrie de masse : les développements instrumentaux de la spectrométrie de masse, les développements de nouvelles techniques de séparation des protéines en amont de l'analyse par spectrométrie de masse ainsi que la croissance rapide des banques de séquences accessibles, alimentées grâce aux projets de séquençage des génomes.

L'analyse protéomique par spectrométrie de masse ne se limite plus actuellement à la simple identification des protéines présentes dans un spot de gel 2D. De nouveaux défis ont été relevés dans les domaines de la caractérisation complète des protéines notamment par la détermination des modifications post-traductionnelles ou encore de la quantification.

La **partie bibliographique** de ce manuscrit est consacrée à la description des techniques et évolutions à l'origine de l'avènement et des progrès en analyse protéomique par spectrométrie de masse ces 20 dernières années.

Les trois facteurs clés sont présentés dans un premier chapitre :

- l'instrumentation de la spectrométrie de masse ESI et MALDI
- les techniques séparatives des protéines et peptides
- les banques de séquences, la ressource de base pour l'identification des protéines.

Le second chapitre est consacré à la présentation des stratégies d'identification des protéines avec des données de spectrométrie de masse ainsi que des directives récemment établies pour l'homogénéisation de la publication des données de protéomique. Un bref aperçu des approches quantitatives et de détermination des modifications post-traductionnelles est également présenté.

Dans ce contexte actuel où l'analyse protéomique par spectrométrie de masse sur des organismes séquencés est largement répandue, mon travail de thèse a consisté en des optimisations et des développements méthodologiques à deux niveaux :

- **L'amélioration de la prise des données de spectrométrie de masse en tandem** notamment par la mise en place et l'optimisation du couplage nanoLC-MS/MS sur des instruments de type trappe ionique.
- **Des développements méthodologiques liés à l'interprétation des données** en aval de l'analyse par spectrométrie de masse. Ces développements ont été réalisés dans le but d'élargir le champ d'application de l'analyse protéomique au-delà de la poignée d'organismes modèles pour lesquels des informations de qualité concernant les séquences de leur génome et de leur protéome sont disponibles.

Ces travaux ont été mis en application au travers de plusieurs collaborations pour lesquelles des réponses à des problématiques biologiques ont pu être apportées.

Ainsi, **les résultats** de ce travail sont organisés en 4 parties :

- **La première partie** est consacrée au volet instrumental de ce travail. Un premier chapitre présente les optimisations réalisées pour la mise en oeuvre expérimentale du couplage nanoLC-MS/MS sur des instruments de type trappe ionique. Un second chapitre présente la mise en place et l'évaluation de nouvelles techniques instrumentales : la technologie nanoHPLC-Chip cube et la fragmentation ETD (Electron Transfer Dissociation).

- **La seconde partie** des résultats illustre trois applications biologiques pour lesquelles les optimisations instrumentales préalablement décrites ont été mises en application pour répondre à des questions biologiques précises :
  - **Un premier chapitre** est consacré à l'établissement d'une carte protéomique de référence pour la levure *Schizosaccharomyces pombe*, un organisme modèle en plein essor. Une étude différentielle par gel d'électrophorèse 2D de

l'effet d'un minéralocorticoïde, l'aldostérone, sur cette même levure a également été réalisée.

- **Un second chapitre** décrit l'étude de l'exoprotéome d'un champignon pathogène *Fusarium graminearum* poussant sur des parois de houblon.
- **Un dernier chapitre** est consacré à l'étude de l'interactome virus-protéines hôtes recrutées lors de l'infection du riz par le virus de la Panachure Jaune du Riz, le *RYMV* (*Rice Yellow Mottle Virus*).

Les applications présentées dans cette partie portent sur des organismes dont les **génomes** ont été **séquencés** et dont les protéomes sont bien représentés dans les banques de séquences protéiques. Dans ces cas, des stratégies d'identification classiques ont permis l'identification des protéines d'intérêt.

- **La troisième partie** des résultats contient des applications portant sur des organismes dont le **génome** n'a **pas** encore été **séquencé** et qui ne sont pas représentés dans les banques protéiques accessibles. L'identification des protéines a donc nécessité le recours à des stratégies utilisant le **séquençage *de novo***.
  - **Le premier chapitre** est un chapitre méthodologique : il contient un état des lieux des outils d'aide au séquençage *de novo*, l'évaluation d'un certain nombre de ces outils et les moyens développés durant ce travail de thèse pour l'amélioration des recherches par séquençage *de novo*.
  - **Le second chapitre** porte sur l'étude du mécanisme de résistance à l'arsenic de la bactérie *Caenibacter arsenoxydans*, une β-protéobactérie dont le génome n'est pas séquencé.
  - **Le troisième chapitre** de cette partie décrit l'étude de l'effet d'un herbicide, la flumioxazin, sur différents tissus de plants de vigne utilisant une approche différentielle par gel d'électrophorèse 2D.

- **La quatrième partie** des résultats est consacrée à la mise en place d'une stratégie de **recherche** avec des données de spectrométrie de masse en tandem **dans des banques génomiques non interprétées**.
  - **Le premier chapitre** est consacré à la description de la stratégie mise en place durant ce travail de thèse pour les recherches avec des données nanoLC-MS/MS dans des séquences génomiques complètes, non interprétées.
  - **Le second chapitre** décrit l'application de cette stratégie à des études protéomiques portant sur des bactéries dont les génomes sont séquencés mais encore non annotés ou dont l'annotation pose des difficultés. L'identification

des protéines directement dans le génome, indépendamment des protéomes prédits *in silico* apporte une validation expérimentale de l'expression de la région identifiée sur le génome. L'apport des recherches dans le génome comme outil complémentaire de validation et d'aide à l'annotation des génomes a pu être mis en évidence par ce travail.

- **Le troisième chapitre** est consacré à la démonstration de la possibilité de discrimination de gènes paralogues spécifiquement exprimés au sein de familles multigéniques grâce à des recherches dans le génome. Cela a été démontré dans le cas de l'infection du riz par le *RYMV* : l'identification des protéines dans le génome complet du riz a permis de discriminer les gènes paralogues spécifiquement exprimés et codant pour les protéines recrutées par le virus lors de son entrée dans les cellules.

Ce travail de thèse a été co-financé par le CNRS et la société Bruker Daltonics qui m'a également apporté un support technique et scientifique.

PARTIE BIBLIOGRAPHIQUE

Des débuts de la protéomique aux défis actuels

Chapitre I : Les outils au service de l'analyse protéomique

Chapitre II : Les stratégies d'identification en analyse protéomique et les nouveaux défis

INTRODUCTION

En biologie, l'étude des protéines passe par la détermination de leur séquence primaire en acides aminés. Des années 50 jusqu'à la fin des années 80, la dégradation d'Edman était la méthode de choix pour la détermination des séquences d'acides aminés, notamment avec l'apparition, en 1967, du séquenceur automatique [Edman, 1950 ; Edman et coll., 1967].

Cependant, dès le début des années 90, la spectrométrie de masse a commencé à jouer un rôle de plus en plus important en biologie pour devenir aujourd'hui la méthode la plus sensible de caractérisation des biomolécules [Lottspeich, 1999 ; Griffiths et coll., 2001 ; Aebersold et coll., 2003 ; Domon et coll., 2006]. Plusieurs faits majeurs ont participé conjointement à l'émergence de ces nouvelles techniques :

- L'apparition des nouvelles méthodes d'ionisations douces, l'Electrospray (ESI) [Yamashita et coll., 1984a ; Yamashita et coll., 1984b ; Fenn et coll., 1989] et le MALDI (Matrix Assisted Laser Desorption Ionisation) [Karas et coll., 1988 ; Karas et coll., 1989 ; Tanaka et coll., 1988].
- Le développement de méthodes séparatives des protéines et des peptides en amont de l'analyse par spectrométrie de masse [Washburn et coll., 2001 ; Görg et coll., 2004].
- La croissance rapide du nombre de séquences protéiques accessibles dans les banques de données publiques, alimentées grâce aux projets de séquençage des génomes à haut débit et le développement de la bioinformatique pour la gestion de ces données [Aebersold et coll., 2003 ; Nesvizhskii et coll., 2005 ; Domon et coll., 2006].

Ces avancées marquantes sont présentées dans le premier chapitre de cette partie bibliographique. Le second chapitre est consacré à la présentation des grandes stratégies d'identification des protéines et des nouveaux défis que doit relever l'analyse protéomique.

Définitions

Le protéome et la protéomique :

Au milieu de l'année 1994, Mark Wilkins, un étudiant à l'Université Macquarie en Australie, était à la recherche du mot « juste » qui lui permettrait de remplacer de longues phrases durant la rédaction d'un article résumant ses travaux de thèse portant sur l'identification rapide de protéines. L'utilisation d'expressions comme « l'ensemble des protéines exprimées par un génome, ou dans un tissus ou dans une cellule » lui semblait inélégante et fastidieuse. C'est ainsi qu'il commença à jouer avec les mots pour exprimer l'équivalent protéique du génome et après avoir écarté « proteinome » et « protome », il opta pour le terme **protéome**, selon lui le plus simple à prononcer. En septembre 1994, Wilkins a utilisé le terme protéome lors d'une conférence scientifique en Italie et depuis ce jour le terme est resté [Wilkins et coll., 1996]. Cet historique du terme protéome est retracé dans « The proteomics payoff » par J. Cohen [Cohen, 2001].

Le terme **protéomique** était au départ utilisé pour désigner l'identification des protéines contenues dans des spots de gel 2D mais est actuellement associé à une grande variété de techniques d'analyses visant 3 objectifs principaux : l'identification, la caractérisation et la quantification de protéines [Hernandez et coll., 2006].

Dans la littérature, de nombreuses définitions ont été données au terme protéomique, de la plus restrictive à la plus vaste. En février 2002 s'est tenu un symposium à la « National Academy of Sciences » à Washington intitulé « Defining the mandate of proteomics in the post-genomics era » durant lequel de nombreux orateurs se sont penchés sur le problème de la définition de la protéomique [Kenyon et coll., 2002]. Deux définitions admises sont rapportées :

- La protéomique représente l'effort pour définir l'identité, la quantité, la structure ainsi que les fonctions biochimiques et cellulaires de l'ensemble des protéines d'un organisme, d'un organe ou d'un organelle et pour établir comment ces propriétés varient dans l'espace, le temps et les états physiologiques. D'après G. Kenyon.
- La protéomique désigne l'analyse de l'ensemble des protéines présentes dans une cellule ou dans un environnement définis ainsi que de leurs variations dans l'espace et le temps. D'après M. Cassman.

En 2003, dans une revue publiée dans *Nature*, Tyers et Mann annoncent que l'étude du protéome évoque non seulement l'ensemble des protéines dans une cellule donnée mais aussi l'ensemble des isoformes et modifications des protéines, les interactions entre protéines, la description structurale des protéines et des complexes et englobe finalement tout ce qui est « post-génomique » [Tyers et coll., 2003].

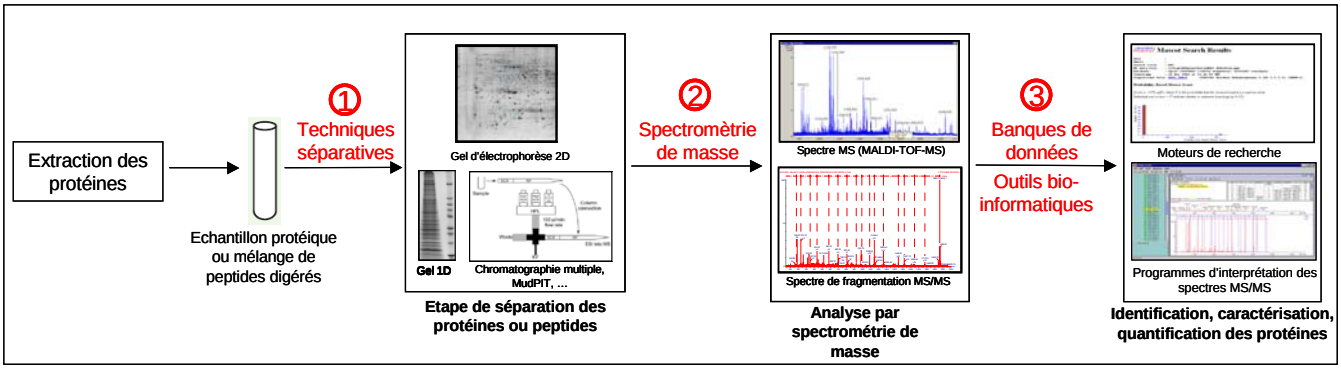
CHAPITRE I

Les outils au service de l'analyse protéomique

Des développements conjoints dans trois domaines sont à l'origine de l'ascension de l'analyse protéomique par spectrométrie de masse : (i) les développements instrumentaux de la spectrométrie de masse (ii) les développements des méthodes séparatives des protéines et peptides, et (iii) les projets de séquençage de génomes à haut débit nourrissant les banques de séquence qui constituent la ressource de base pour l'identification des protéines avec les données de spectrométrie de masse.

Ces trois domaines sont les clés des différentes étapes de l'analyse protéomique qui sont schématisées en figure 1. Après l'étape de préparation de l'échantillon protéique ou peptidique (si la digestion est réalisée dès le départ), ce dernier est séparé grâce à des techniques séparatives telles que les gels d'électrophorèse 1D ou 2D ou des chromatographies liquides. Les mélanges peptidiques sont ensuite analysés par spectrométrie de masse. Enfin, les identifications des protéines et les interprétations des données de spectrométrie de masse sont réalisées grâce à des outils bioinformatiques spécialisés.

Figure 1 : Schématisation des différentes étapes de l'analyse protéomique.



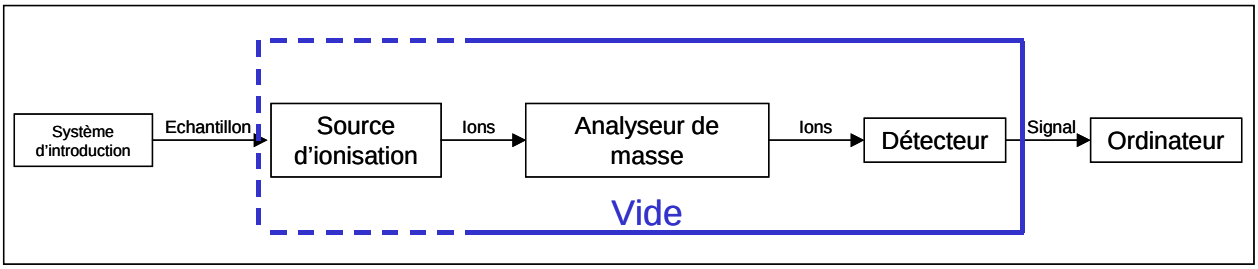
Les développements dans ces trois domaines sont présentés dans ce chapitre en commençant par les développements instrumentaux de la spectrométrie de masse qui sont au cœur de l'analyse protéomique par spectrométrie de masse et qui ont conditionné des développements spécifiques et adaptés dans les deux autres domaines.

1 Les outils de la spectrométrie de masse

Les débuts de la spectrométrie de masse datent du début du siècle dernier avec les travaux de Thompson et Aston [Thompson et coll., 1913]. Cependant, ce n'est que depuis ces vingt dernières années que la spectrométrie de masse joue un rôle croissant en biologie pour la caractérisation des biomolécules [Lottspeich, 1999 ; Griffiths et coll., 2001 ; Aebersold et coll., 2003]. Le succès de la spectrométrie de masse en biologie repose largement sur l'apparition de nouvelles techniques d'ionisation douces à la fin des années 1980 : l'**ionisation Electrospray** (ESI) [Yamashita et coll., 1984a ; Yamashita et coll., 1984b ; Fenn et coll., 1989] et le **MALDI** (Matrix Assisted Laser Desorption Ionisation) [Karas et coll., 1988 ; Karas et coll., 1989 ; Tanaka et coll., 1988]. Ces techniques ont rendu possible le transfert de biomolécules intactes en phase gazeuse pour l'analyse par spectrométrie de masse. John Fenn et Koichi Tanaka ont été récompensés par un prix Nobel de chimie en 2002 pour leurs travaux sur ces nouvelles méthodes d'ionisation.

Les différents éléments constituant un spectromètre de masse sont schématisés sur la figure 2 et sont décrits dans la suite de ce premier chapitre (cf 1.1 à 1.4).

Figure 2 : Schéma des différents constituants d'un spectromètre de masse.



1.1 Les sources d'ionisation ESI et MALDI

1.1.1 La source Electrospray

Les expériences de Dole à la fin des années 60 ont suggéré que le mécanisme ESI pouvait être utilisé pour mesurer des masses élevées de polymères (tels que des polystyrènes) en solution [Dole et coll., 1968]. Ce n'est que dans les années 80 que l'intérêt pour l'ESI renaît lorsque Fenn reprend ces travaux en couplant cette source à un analyseur quadripolaire [Yamashita et coll., 1984]. Il démontre alors que des ions peuvent être obtenus ainsi sans fragmentation à partir de molécules non volatiles. En 1988, le premier spectre de masse d'une protéine multi-chargée intacte de 40 kDa, l'Alcohol Dehydrogenase, est présenté [Meng et coll., 1988]. Les exemples de spectres de biomolécules obtenus par ESI se multiplient ensuite rapidement dans la littérature [Fenn et coll., 1989 ; Fenn et coll., 1990 ; Loo et coll., 1992].

Le phénomène ESI se produit à pression atmosphérique, sous l'effet d'un fort champ électrique. Le mécanisme macroscopique de l'ESI est aujourd'hui bien décrit: il assure le passage des ions présents en solution vers la phase gazeuse. Cependant quelques divergences persistent sur la désorption des ions de la solution vers la phase gazeuse.

Le processus d'ionisation-désorption ESI se divise en 3 étapes majeures :

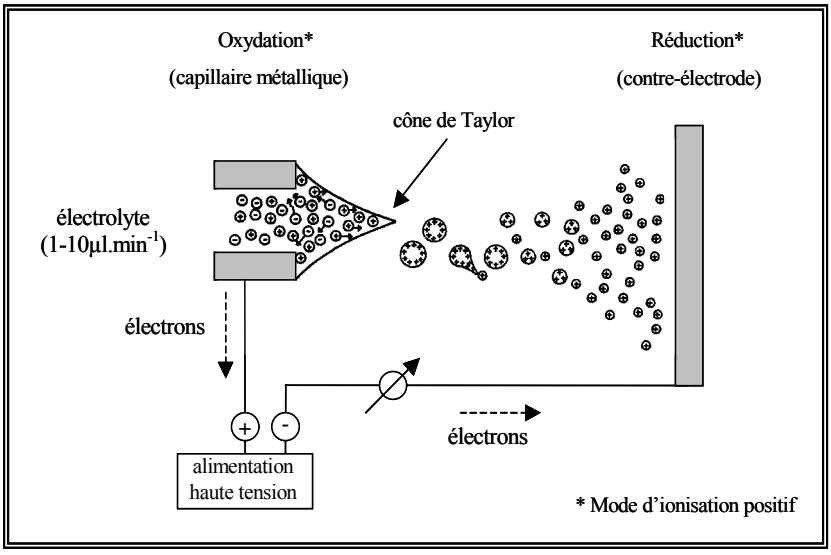
- **La production des gouttelettes chargées à partir de l'electrolyte en solution**
- **La fission des gouttelettes chargées par explosions coulombiennes**
- **Le transfert des ions désolvatés en phase gazeuse**

➢ La production des gouttelettes chargées à partir de l'electrolyte en solution

L'application d'un champ électrique intense (10^6 V/m) à la pointe du capillaire provoque une polarisation du liquide et une séparation des charges positives et négatives. L'accumulation des charges situées à la pointe du capillaire déstabilise la surface du liquide qui prend, pour un champ électrique critique, la forme d'un cône appelé « cône de Taylor ». De ce cône sont émises des gouttelettes chargées (Figure 3).

La source ESI peut donc être considérée comme une cellule électrolytique [Blades et coll., 1991]. Une assistance pneumatique, gaz de nébulisation, est requise pour améliorer une production stable et régulière de gouttelettes chargées [Ikonomou et coll, 1991].

Figure 3 : Schéma de la source ESI [Kebarle, 1993].



➢ La fission des gouttelettes chargées en « gouttelettes filles »

Dans la source chauffée, l'évaporation progressive du solvant va entraîner le rétrécissement de la taille des gouttelettes. La diminution du rayon de la gouttelette a lieu jusqu'à ce que la limite de stabilité de Rayleigh soit atteinte : à ce moment, les forces de répulsions électrostatiques sont égales aux forces de tension de surfaces [Kebarle, 2000]. L'équation ci-dessous relie la charge de la gouttelette à son rayon. Le rayon limite de Rayleigh dépend de la charge de la gouttelette (Q), de la tension superficielle du liquide (γ) et de la permittivité du vide (ε_0) :

$$Q^2 = 64\,\pi^2\,\varepsilon_0\,\gamma\,R_R^3$$

La gouttelette devient ensuite instable et produit des gouttelettes plus petites, appelées gouttelettes « filles » par fission coulombienne. Les gouttelettes filles subiront le même processus et cela sur plusieurs générations jusqu'à donner un ion complètement désolvaté.

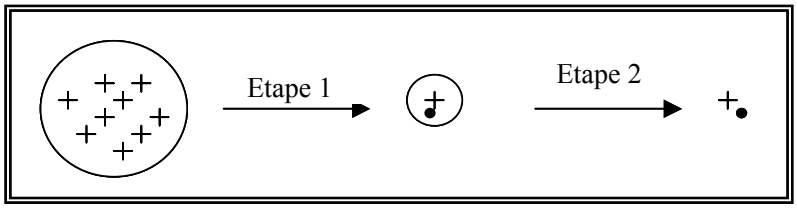
- Le transfert des ions désolvatés en phase gazeuse

Deux mécanismes décrivent la production d'ions désolvatés en phase gazeuse à partir de gouttelettes chargées : le **mécanisme de la charge résiduelle** (« Charged Residue Mechanism », CRM) proposé par Dole [Dole et coll., 1968] et le **mécanisme de l'évaporation ionique** (« Ion Evaporation Mechanism », IEM) d'Iribarne et de Thomson [Iribarne et Thomson, 1976].

Le modèle de Dole, le CRM

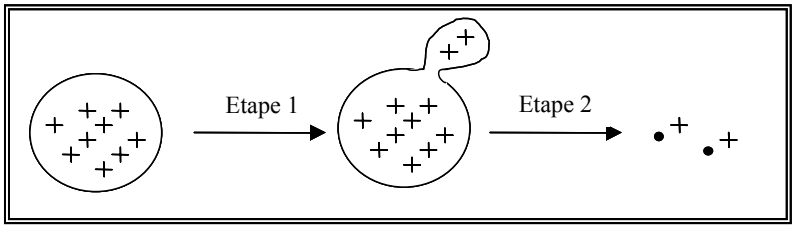
Selon ce modèle, la gouttelette mère va tout au long des différentes explosions coulombiennes générer des gouttelettes filles qui à leur tour vont subir ce même mécanisme jusqu'à ce qu' il n'y ait plus qu'une seule charge par gouttelette (Figure 4).

Figure 4 : Illustration du mécanisme de Dole, d'après [Leize, 1994].



Le modèle d'Iribarne et Thomson, l'IEM, encore appelé modèle d'évaporation des ions, propose qu'à un stade intermédiaire de sa vie (avant le rayon critique de Rayleigh), le champ électrique à la surface de la gouttelette soit suffisant pour désorber des ions directement de la gouttelette (« évaporation d'ions »). Généralement ce processus est très important pour une gouttelette de rayon R<10nm (Figure 5).

Figure 5 : Illustration du mécanisme d'Iribarne et Thomson, d'après [Leize, 1994].



Actuellement, il est admis que **les petits ions** sont produits majoritairement par **IEM** alors que les **protéines de tailles importantes** sont produites par **CRM** [Kebarle, 2000].

Le pré-requis pour la production d'ions en phase gazeuse est que les molécules d'analyte puissent être ionisées en solution. Lorsque les molécules d'analyte présentent plusieurs sites ionisables, des ions multichargés sont produits. En général, les protéines dénaturées présentent environ une charge par 1000Da [Hoffmann et coll., 2003]. Ainsi, comme ce sont toujours des ratios m/z qui sont mesurés, en observant des espèces multichargées, la gamme de masse effective du spectromètre de masse peut être étendue de centaines

de milliers de daltons : une hémocyanine de 2,2 MDa a pu être caractérisée au laboratoire [Sanglier et coll., 2003].

1.1.2 La source nano-electrospray ou nanospray (nano-ESI)

En mode d'ionisation ESI, le courant d'ion détecté par le spectromètre de masse dépend de la concentration de l'analyte et non du débit avec lequel l'échantillon est infusé [Ikonomou et coll., 1990]. Pour des faibles quantités d'échantillon, il devient donc intéressant de travailler à des débits plus faibles mais sur des solutions plus concentrées afin de gagner en sensibilité. De ces observations sont nées les **sources micro-électrospray** atteignant des débits inférieurs à 1µl/min [Emmet et coll., 1994], et en 1996 les **sources nanospray** (20nl/min) [Wilm et coll., 1996].

Dans ces montages, la taille des gouttelettes générées est beaucoup plus petite et l'assistance gazeuse du spray n'est plus nécessaire. La miniaturisation des sources ESI présente de nombreux avantages pour l'analyse : tolérance aux sels, augmentation de la sensibilité et augmentation de la précision de la mesure de masse des protéines en moyennant le signal sur une plus longue durée [Wilm et coll., 1996]. Les diverses applications du nanospray, notamment pour l'analyse protéomique montrent aujourd'hui l'importance d'un tel dispositif [Shevchenko et coll. 1996].

1.1.3 La source MALDI

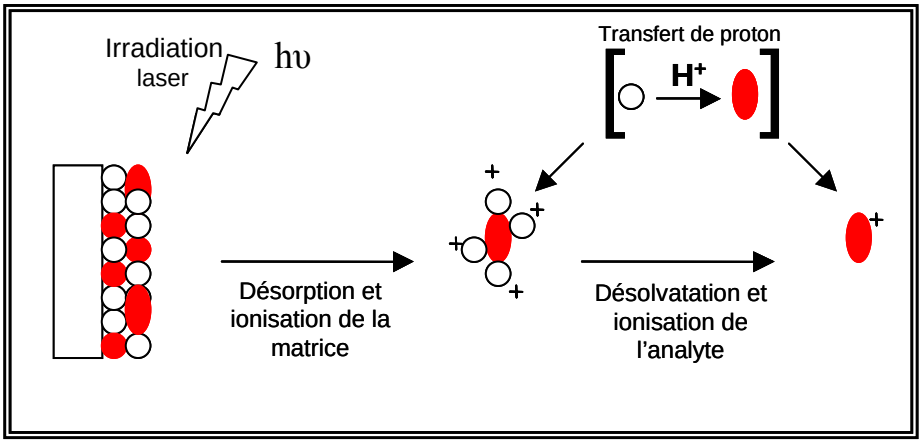
La technique MALDI a été introduite en 1988 par Karas et Hillencamp qui décrivaient l'analyse de protéines de poids moléculaires dépassant 10kDa [Karas et coll., 1987 ; Karas et coll., 1988]. Une description détaillée de la source MALDI est présentée dans une thèse précédente du laboratoire (Wagner, 2004), seul le principe général en sera décrit ici.

Cette technique consiste à irradier avec un faisceau laser pulsé de longueur d'onde donnée (généralement dans l'UV) un dépôt cristallin contenant une matrice organique et l'échantillon à analyser. Les matrices les plus couramment utilisées sont l'acide α-cyano-4-hydroxycinnamique [Beavis et coll., 1989] (peptides et protéines), l'acide sinapinique (protéines), l'acide 2,5-dihydrobenzoïque (oligosaccharides) ou encore l'acide 3-hydroxypicolinique (oligonucléotides). La qualité des spectres dépend de plusieurs facteurs liés à la préparation de l'échantillon comme le choix de la matrice, la méthode de dépôt pour la co-cristallisation matrice-analyte, le pH des solutions, …

Le processus d'ionisation par MALDI s'effectue sous vide ($\approx 10^{-7}$ mbar). Les phénomènes ayant lieu ne sont, à l'heure actuelle, ni entièrement ni unanimement établis. Plusieurs modèles ont été proposés mais de nombreuses ambiguïtés subsistent concernant certains aspects du processus. Cependant, selon les hypothèses les plus communément admises, le processus est constitué de trois événements majeurs (Figure 6) [Hoffmann et coll., 2003] :

- L'excitation des molécules de matrice par les photons du laser
- L'émission des molécules cibles dans la phase gazeuse
- L'ionisation des molécules en phase gazeuse [Knochenmuss et coll., 2003]

Figure 6 : L'ionisation MALDI, d'après Hoffmann et Stroobant [Hoffmann et coll., 2003].



La tolérance aux sels, souvent citée comme un des points forts de l'ionisation MALDI comparée aux autres modes d'ionisation, vient essentiellement de la nature solide du dépôt à analyser.

Enfin, le processus MALDI est indépendant des propriétés d'absorption et de la taille des composés à analyser et il permet donc de ce fait la désorption et l'ionisation d'analytes de très haut poids moléculaire (>100000Da).

1.2 Les analyseurs

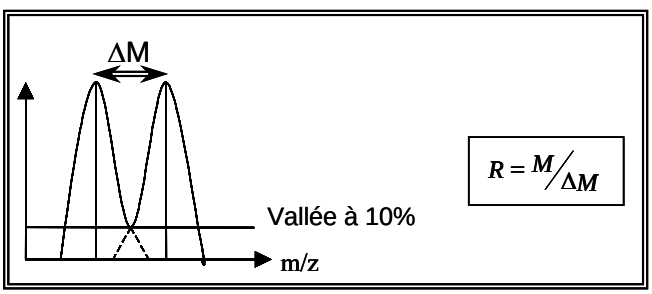
Les résultats obtenus lors d'une analyse protéomique sont principalement conditionnés par les performances de l'analyseur de masse utilisé. Les performances d'un analyseur sont définies par quelques paramètres clés : sa **résolution** (Figure 7), sa **précision de mesure de masse**, sa **gamme de masse**, sa **sensibilité** et sa capacité à faire de la **spectrométrie de masse en tandem**, MS/MS voire de la MS multiple.

Quatre types d'analyseurs sont couramment utilisés en analyse protéomique :

- les trappes ioniques (IT)
- les tubes de vol (TOF)
- les quadripôles (Q)
- les analyseurs de résonance cyclotronique des ions à transformée de Fourier (FT-ICR)

Ces analyseurs peuvent être utilisés soit seuls, soit assemblés pour créer des analyseurs hybrides bénéficiant des points forts de chacun des analyseurs assemblés.

Figure 7 : Définition de la résolution : La résolution correspond à la capacité de pouvoir distinguer une masse M d'une masse M+ΔM. On admet que deux pics sont résolus quand l'intensité de la vallée entre ces pics égale 10% de l'intensité du pic le plus faible (on parle dans ce cas de résolution à 10% de vallée). La résolution se calcule donc par le rapport entre la masse et la largeur du pic ΔM à 10% de son maximum. On parle maintenant couramment de résolution FWHM (Full Width at Half Maximum), dans ce cas, la largeur du pic est prise à 50% de son maximum.



Durant ce travail de thèse, différents instruments ont été utilisés : un MALDI-TOF-TOF Ultraflex (Bruker Daltonics), un Q-TOF II (Waters) et des trappes ioniques Esquire 3000+, HCTPlus, HCTUltra (Bruker Daltonics). Seuls les principes des analyseurs constituant ces différents instruments sont présentés.

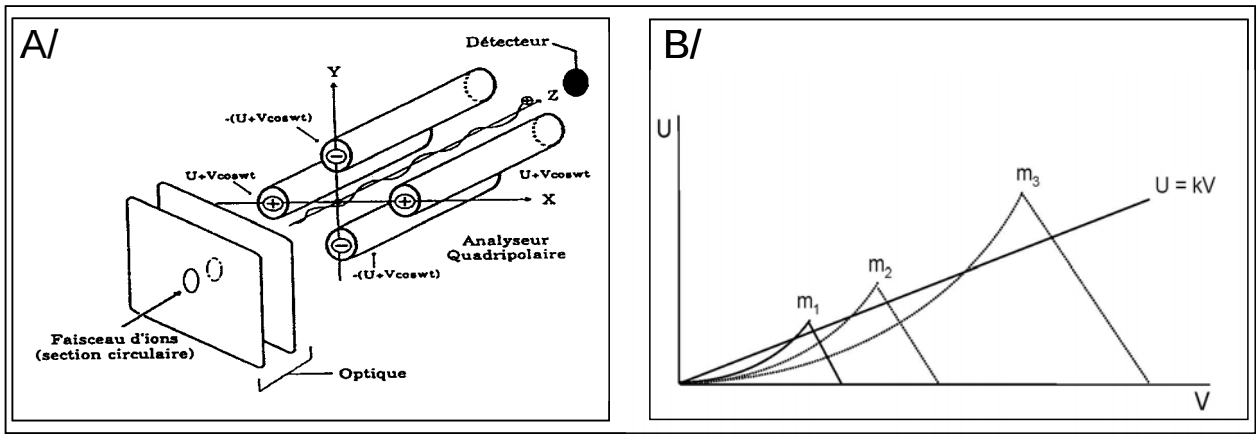
1.2.1 L'analyseur quadripôlaire (Q)

Le quadripôle est l'analyseur le plus couramment couplé à la source ESI car il est capable d'analyser un courant d'ion continu. Son principe de fonctionnement a été décrit en détail par Campana [Campana, 1980]. Il est constitué de 4 barres métalliques parallèles : les barres opposées, connectées électriquement, sont portées au même potentiel Φ_0 (ou $-\Phi_0$) donc deux barres adjacentes sont de potentiels opposés Φ_0 et $-\Phi_0$ (Figure 8A). Le potentiel Φ_0 comporte une composante sinusoïdale $V\cos\omega t$, où V est l'amplitude de la radio-fréquence, de fréquence ω et une composante continue U.

$$\begin{cases} \Phi_0 = U - V\cos\omega t \\ -\Phi_0 = -U + V\cos\omega t \end{cases}$$

Le quadripôle est un analyseur à balayage qui utilise la stabilité de la trajectoire des ions entre les barres pour les séparer en fonction de leur rapport m/z. Les ions accélérés suivant l'axe z pénètrent entre les barres du quadripôle et conservent leur vitesse suivant cet axe. Cependant, ils sont soumis aux accélérations résultantes des forces dues aux champs électriques suivant x et y. Les équations régissant la trajectoire des ions dans le quadripôle ont été établies en 1866 par le physicien Mathieu. La résolution de ces équations permet de déterminer les zones de stabilité des ions d'une masse donnée en fonction de U et de V (Figure 8B). D'après ce diagramme, un balayage maintenant un rapport U/V constant permettra de transmettre successivement des masses différentes.

Figure 8 : A/ Schéma d'un analyseur quadripolaire avec la représentation de la trajectoire des ions selon l'axe z, d'après Campana [Campana, 1980]. B/ Diagramme de stabilité des ions (de masses m_1, m_2 et m_3) en fonction de U et V dans un quadripôle.



1.2.2 L'analyseur à temps de vol (TOF)

Le principe de séparation des ions selon leur rapport m/z en fonction de leur temps de vol date des années 50 [Wiley et coll., 1955] mais ce n'est qu'avec l'apparition de la source MALDI en 1987 [Karas et coll., 1987] et grâce à des avancées technologiques que l'utilisation de cet analyseur fut enfin reconnue [Standing et coll., 2000]. Classiquement couplé à des sources d'ionisation MALDI (pulsé donc compatible avec une mesure de temps), l'apparition de l'injection orthogonale a permis son couplage avec des sources ESI.

L'analyseur à temps de vol est constitué de deux zones : une **zone d'accélération** des ions par application de potentiels élevés et une **zone libre de champ** dans lesquelles règne un vide poussé de l'ordre de 2.10^{-7}mbar (Figure 9).

La mise en équation ci-dessous permet d'établir, dans le cas idéal où tous les ions ont la même énergie cinétique initiale, que le temps de vol des ions est proportionnel à la racine carrée de leur rapport m/z :
$Ec = zeE$ avec Ec l'énergie cinétique d'un ion de masse m et de charge z accéléré dans un champ E avec e =charge de l'électron.

$v = \sqrt{\frac{2zeE}{m}}$ avec v la vitesse de l'ion de masse m et de charge z.

Alors, le temps t nécessaire à l'ion pour traverser le tube de vol de longueur L, sachant que $v = \frac{L}{t}$, sera donné par :

$$t = L\sqrt{\frac{m}{2zeE}}$$

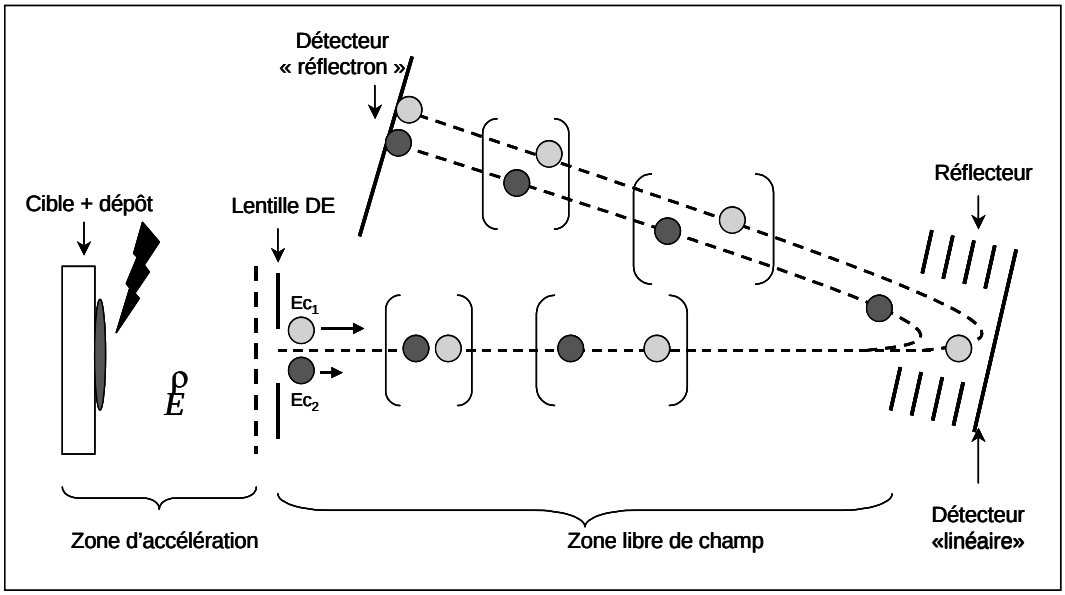
Les premiers analyseurs TOF souffraient d'une mauvaise résolution jusqu'à ce que deux dispositifs aient été mis en place permettant de compenser les distributions temporelles (temps de formation des ions), spatiales (lieu de la formation des ions) et cinétiques (énergie de formation des ions) des ions :

- **L'extraction retardée** (« Delay extraction ») [Cotter et coll., 1989 ; Vestal et coll., 1995]. L'utilisation d'une lentille supplémentaire permet d'appliquer pendant un bref délai suivant le tir laser un potentiel légèrement supérieur à celui de la cible s'opposant ainsi à la propagation des ions dans l'analyseur. Ceci permet au nuage gazeux de se dissiper et les ions seront alors tous sur la « même ligne de départ » pour pénétrer dans le tube de vol une fois le délai passé. La focalisation temporelle et spatiale est donc améliorée.
- **Le réflecteur électrostatique** [Mamyrin et coll., 1973]. Le réflecteur est composé d'une série d'anneaux ou de grilles sur lesquels sont appliqués des potentiels croissants, agissant comme un miroir électrostatique. Le champ crée par ce miroir électrostatique va permettre de renvoyer les ions vers le détecteur. Cependant, en fonction de l'énergie cinétique que possèdent les ions, ils évoluent plus ou moins profondément dans le réflecteur. Au final, les ions de même m/z sont refocalisés et arriveront au même moment au détecteur.

Ces améliorations combinées ont permis la construction de spectromètres TOF atteignant des résolutions de 20000, voire 40000 (Technologie MultiPass^{TM}, Bruker Daltonics) et en font un instrument de choix pour les mesures de masses précises des peptides trypsiques en analyse protéomique.

La figure 9 rassemble l'ensemble de ces dispositifs pour illustrer le principe de fonctionnement d'un spectromètre de masse de type MALDI-TOF.

Figure 9 : Schéma représentant le principe de fonctionnement d'un MALDI-TOF avec un réflecteur électrostatique et un système d'extraction retardée. ○ et ● sont deux ions de même m/z.



1.2.3 L'analyseur trappe ionique (IT)

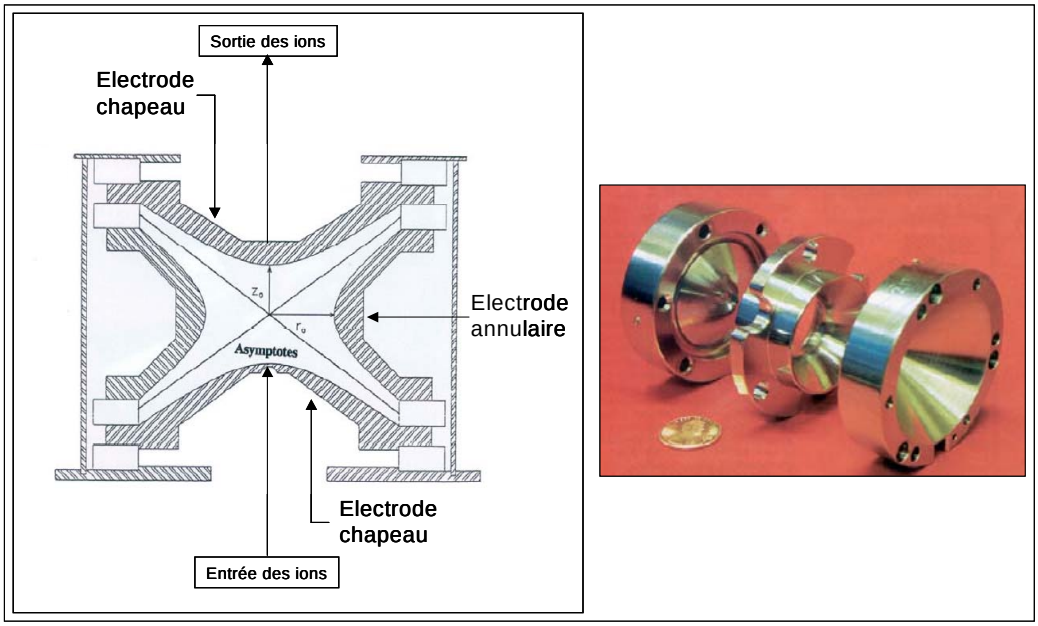
Dans les années 50, Paul et Steinwedel ont été les premiers à décrire des analyseurs de type trappe ionique [Paul et coll., 1960] et ce travail a été reconnu par un prix Nobel décerné en 1989 à Wolfgang Paul [Paul, 1990]. Suite aux nombreux développements technologiques apportés par G. Stafford et Finnigan MAT, la trappe ionique est devenue un analyseur de spectrométrie de masse, commercialisé en 1983 sous le nom de Ion Trap Detector (ITD) comme détecteur de chromatographie gazeuse [Stafford et coll., 1984 ; Kelley et coll., 1985]. En 1995, une nouvelle génération d'instruments de type trappe ionique (les LCQ et GCQ de Finnigan et l'Esquire de Franzen chez Bruker) a été introduite, employant des sources externes d'ions avec un système d'injection des ions générés à l'extérieur de la trappe. Enfin, d'autres groupes ont participé à l'amélioration des performances de cet analyseur et son principe, son fonctionnement et ses utilisations ont été largement décrits dans quelques ouvrages et revues [March et coll., 1989 ; Cooks et coll., 1991 ; March and Todd, 1995 ; March, 1997 ; March et coll., 1998]. Les points fondamentaux de son principe de fonctionnement sont rappelés ici.

- Description de l'analyseur

La trappe ionique est constituée de 3 électrodes, une électrode annulaire en forme de diabolo et deux électrodes chapeaux quasi-hyperboliques. Les ions entrent et sortent de la trappe par des orifices situés dans les électrodes chapeaux (Figure 10). L'électrode chapeau d'entrée des ions contient un orifice unique tandis

que l'électrode chapeau de sortie des ions présente plusieurs petits trous par lesquels les ions vont pouvoir sortir pour atteindre le détecteur.

Figure 10 : Schéma et photo de l'analyseur trappe ionique, d'après March [March, 1997]. r_0 est la distance entre le centre de la trappe et l'électrode annulaire et z_0 est la distance entre le centre de la trappe et les électrodes chapeaux.

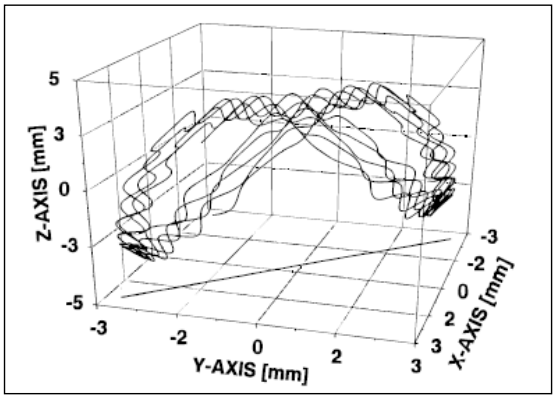


➢ <u>Le piégeage des ions dans la trappe</u>

Lors de l'injection des ions dans la trappe, une radiofréquence Vcosωt est appliquée sur l'électrode annulaire de la trappe. Le champ quadripôlaire crée par cette radiofréquence assure le piégeage des ions. Un gaz tampon, de l'hélium, est maintenu dans la trappe à une pression d'environ 5.10^{-3} mbar et permet d'améliorer l'efficacité de piégeage des ions : l'énergie cinétique des ions est réduite par des collisions avec les molécules de gaz ce qui permet de les refocaliser au centre de la trappe et donc d'améliorer leur piégeage.

Comme dans un quadripôle, la trajectoire tridimensionnelle des ions à l'intérieur du piège est imposée par le champ quadripôlaire et peut être définie par les équations différentielles de Mathieu : les ions suivent le champ quadripôlaire selon le modèle d'une courbe de Lissajous (Figure 11).

Figure 11 : Trajectoire des ions piégés dans la trappe par le champ quadripolaire. La trajectoire des ions suit une courbe de Lissajous.



Le champ quadripolaire appliqué peut être assimilé à un puits de potentiel de largeur $2z_0$ dirigé suivant l'axe z de la trappe. L'équation définissant la profondeur du puits de potentiel permet de décrire les effets de l'amplitude de la radiofréquence sur l'efficacité de piégeage des ions [Dehlmet et coll., 1967] :

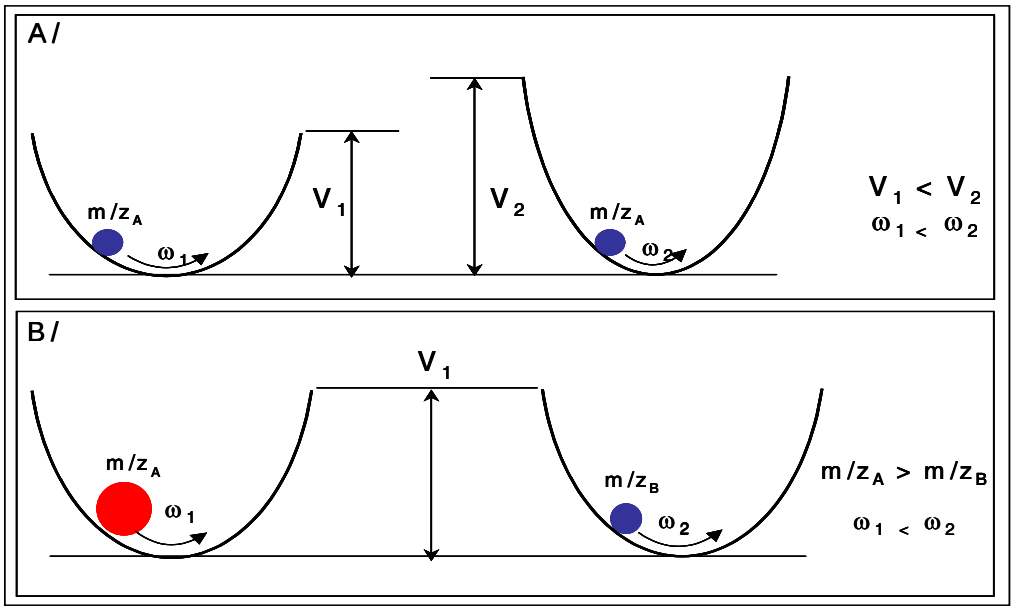
$$D_z = \frac{V^2}{m/z\left(r_0^{\,2} + 2z_0^{\,2}\right)\omega_{RF}^{\,2}}$$

avec D_z la profondeur du puits de potentiel, V l'amplitude de la radiofréquence, ω_{RF} la fréquence angulaire de la radiofréquence et r_0 et z_0, les plus petites distances qui séparent le centre de la trappe des électrodes annulaire et chapeaux, respectivement.

Ainsi, comme la profondeur du puits de potentiel est proportionnelle au carré de l'amplitude de la radiofréquence, il en résulte que plus l'amplitude de la radiofréquence est élevée, plus la capacité de piégeage des ions sera forte (Figure 12A). D'autre part, la profondeur du puits de potentiel est également inversement proportionnelle au rapport m/z, les ions de différents m/z ne subissent donc pas le même potentiel de piégeage. Plus l'amplitude de la radiofréquence est élevée, plus le piégeage sera favorable aux ions de haut m/z mais défavorable aux ions de petit m/z (Figure 12B).

L'amplitude et la fréquence de la radiofréquence appliquée sur l'électrode annulaire va donc directement influer sur la gamme des m/z pouvant être piégés. En effet, pour une radiofréquence donnée, certains ions de petit m/z auront des fréquences d'oscillations trop importantes pour pouvoir être piégés. Cette limite inférieure de piégeage des ions est appelée « cut-off » (voir aussi Partie I, Chapitre I des résultats pour une illustration du cut-off). L'utilisateur fixe lui-même (indirectement par le Trap Drive sur les instruments Bruker) l'amplitude de la radiofréquence sur les instruments commerciaux, donc la limite inférieure de détection des ions selon la gamme de masse qu'il souhaite analyser.

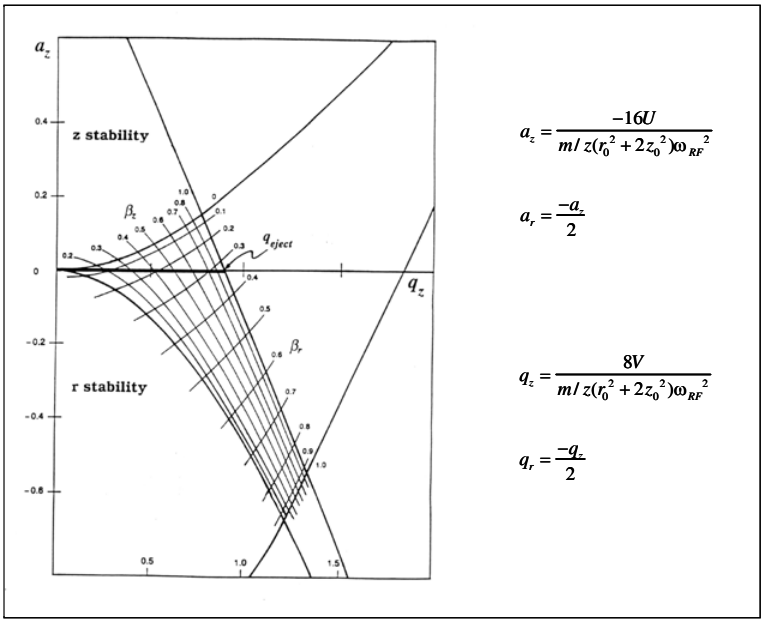
Figure 12 : A/ Illustration de l'effet de l'augmentation de l'amplitude de la radiofréquence sur la profondeur du puits pour un ion de même m/z_A. B/ Illustration de la profondeur du puits de potentiel pour une amplitude de la radiofréquence constante en fonction du rapport m/z de l'ion piégé (deux ions de m/z différents m/z_A et m/z_B sont présentés). D'après [Carte, 2001].



➢ Le mouvement et la stabilité des ions dans la trappe

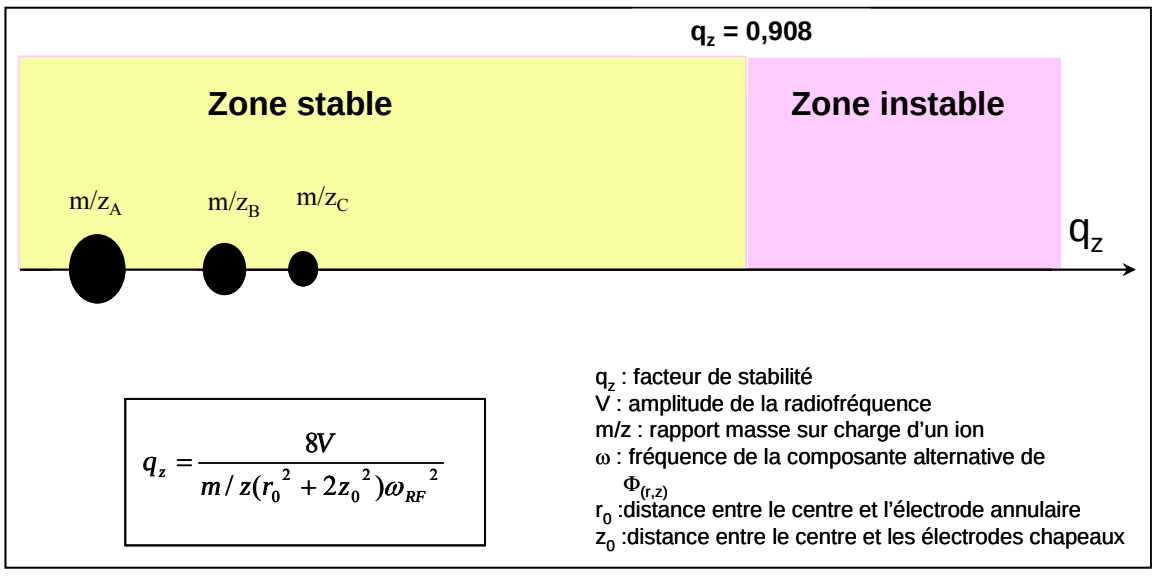
La résolution des équations de Mathieu conduit à la détermination dans l'espace de zones de stabilité des ions de masses différentes. En recouvrant les régions de stabilité correspondant à des solutions stables de l'équation de Mathieu dans les directions de r et de z, la zone illustrée en figure 13 apparaît. Les ions ayant des trajectoires stables sont piégés alors que les ions ayant des trajectoires instables sont défléctés sur les électrodes.

Figure 13 : Diagramme de stabilité dans l'espace (a_z, q_z) pour les régions de stabilité simultanément dans les deux directions r- et z- pour la trappe ionique.



Le mode de fonctionnement des analyseurs à trappe ionique ne fait intervenir qu'une composante alternative Vcosωt (en mode RF-only) pour établir le champ quadripolaire de piégeage dans la trappe. Le facteur a_z devient donc nul et le diagramme de stabilité peut être réduit à deux zones paramétrées par le facteur q_z et représentées en figure 14.

Figure 14 : Diagramme de stabilité des ions A, B, C ($m/z_A > m/z_B > m/z_C$) soumis à un champ quadripolaire.



Les deux zones sont délimitées par une frontière où $q_z=q_{eject}=0,908$:
Pour des valeurs de $\mathbf{q_z < 0,908}$, les ions ont une trajectoire stable et sont piégés.
Pour des valeurs de $\mathbf{q_z > 0,908}$, les ions sont défléctés sur les électrodes.
Cette représentation illustre également l'existence du cut-off de basses masses, car pour une radiofréquence donnée, il existe des ions de petits m/z pour lesquels q_z sera supérieur à 0,908 et qui seront donc éjectés.

- L'éjection des ions de la trappe

L'éjection simple

D'après le diagramme de stabilité (Figure 14) et de la définition du facteur q_z, un balayage en amplitude croissante de la radiofréquence permet d'éjecter de la trappe les ions suivant un ordre croissant de m/z. Or, cette technique d'éjection est limitée par l'amplitude de la tension maximale applicable (20kV d'après [March et coll., 1955]). Il existe alors des ions de m/z élevés pour lesquels l'amplitude maximale de la tension alternative ne sera pas suffisante pour les éjecter de la trappe. Dans ces conditions, l'éjection simple par balayage croissant de l'amplitude de la tension alternative limite donc la gamme de masse à une valeur de m/z=650.

La méthode la plus couramment appliquée pour augmenter la gamme de masse des analyseurs trappe ionique est la méthode d'éjection résonante.

L'éjection résonante

L'éjection résonante consiste à mettre les ions de m/z croissants en résonance de façon à les éjecter prématurément de la trappe pour pouvoir augmenter la gamme de masse des ions analysables. Les ions confinés dans la trappe ont un mouvement périodique oscillant, il est donc possible de conférer de l'énergie à leur mouvement ce qui, par analogie, revient à pousser un pendule à sa fréquence fondamentale.

Physiquement, cette éjection prématurée est réalisée en transférant aux ions de l'énergie cinétique suivant l'axe z de la trappe (Figure 10) par l'intermédiaire d'une tension alternative supplémentaire appliquée sur les électrodes chapeaux. La frontière sur le diagramme de stabilité est alors décalée vers des valeurs de q_z plus petites. De plus, les ions oscillent avec une fréquence qui leur est propre, il est donc possible de mettre en résonance précisément chaque ion individuellement.

D'après March et coll., les ions sont confinés au centre de la trappe dans un nuage ayant la forme d'un oignon. L'éjection résonante a comme effet de soustraire les ions de bas m/z résidant dans les couches internes de l'oignon des perturbations dues aux effets d'espace-charge. Les ions ainsi libérés de ces effets d'espace-charge sont éjectés ce qui conduit à une amélioration de la résolution et de la précision de la mesure de masse [March, 1997]. Brièvement, les phénomènes d'espace-charge sont observés dans des analyseurs de type trappe ionique et résultent d'une accumulation trop importante d'ions dans l'analyseur. Lorsque les ions sont accumulés en surnombre dans la trappe, le champ quadripôlaire est perturbé et les performances de l'analyseur en résolution et en précision de mesure de masse sont réduites (cf Partie I, chapitre I des résultats pour une description des effets d'espace-charge).

- L'isolation des ions dans la trappe

L'analyseur trappe ionique ouvre la possibilité de réaliser des expériences de spectrométrie de masse multiple, MS/MS, MS^3, MS^4, …MS^n. La maîtrise de la trajectoire des ions dans la trappe permet d'isoler

spécifiquement un ion de m/z choisi dans la trappe. Le processus d'isolation des ions de m/z choisi dans la trappe peut être décomposé en trois étapes :

- Les ions de m/z inférieurs au m/z choisi sont éjectés par un balayage croissant de l'amplitude de la radiofréquence appliquée sur l'électrode annulaire en atteignant une amplitude proche de l'amplitude nécessaire à l'éjection de l'ion précurseur choisi.
- Les ions de m/z supérieurs sont mis en résonance et éjectés par un balayage en fréquence partant d'une fréquence proche de la fréquence de résonance des ions précurseurs à isoler.

L'ion précurseur est isolé plus finement par des balayages en fréquence proches de la fréquence de résonance de l'ion précurseur.

➢ La fragmentation des ions précurseurs

La fragmentation est provoquée par l'ajustement de la fréquence de la tension alternative sur l'électrode chapeau de manière à ce qu'elle coïncide avec la fréquence propre de résonance des ions précurseurs : cette amplitude est réglée afin d'éviter l'éjection des ions de la trappe (l'amplitude de fragmentation est plus faible que l'amplitude d'éjection). Ainsi, dopés en énergie cinétique, les ions précurseurs entrent en collision avec les molécules d'hélium. Au fur et à mesure des collisions, leur énergie cinétique est convertie en énergie interne vibrationnelle et l'ion parent finit par se fragmenter.

Dans l'analyseur trappe ionique l'énergie de fragmentation est exclusivement transmise aux ions précurseurs. Ainsi, les ions fils issus d'une fragmentation ne peuvent donc, théoriquement, se fragmenter à nouveau pour donner des ions petits-fils, l'énergie qui est transmise aux ions précurseurs n'étant délivrée qu'à leur seule fréquence de résonance.

Par ailleurs, un avantage supplémentaire de l'analyseur à trappe ionique est la possibilité de faire de la MS^n. En effet, le processus à plusieurs étapes, piégeage-isolation-fragmentation, peut être répété plusieurs fois et ceci s'avère très utile pour l'élucidation structurale des composés analysés.

➢ Applications, limites et nouveaux développements

Cet analyseur à la fois robuste, sensible et de coût modéré est un instrument de choix pour l'analyse protéomique. Cependant, la principale limite de cet analyseur est la faible précision des mesures de masse due en partie à la faible capacité de piégeage d'ions dans la trappe avant de subir les effets d'espace-charge (cf Partie I, chapitre I des résultats). Pour les trappes Bruker Daltonics, la capacité d'ions dans la trappe a été considérablement augmentée entre les trappes de type Esquire et les nouvelles trappes HCT grâce à des changements dans la géométrie de l'analyseur ce qui a permis d'améliorer les performances instrumentales (cf Partie I, chapitre I des résultats).

Par ailleurs, une nouvelle génération d'instruments de type trappe ionique linéaire ou 2D, présentant des capacités de piégeage beaucoup plus élevées que les trappes 3D, ont été développées apportant des améliorations tant au niveau de la sensibilité que de la résolution et de la précision de masse [Hager, 2002 ; Schwartz et coll., 2002 ; Ouyang et coll., 2004].

1.2.4 Des analyseurs en tandem : l'exemple du Q-TOF

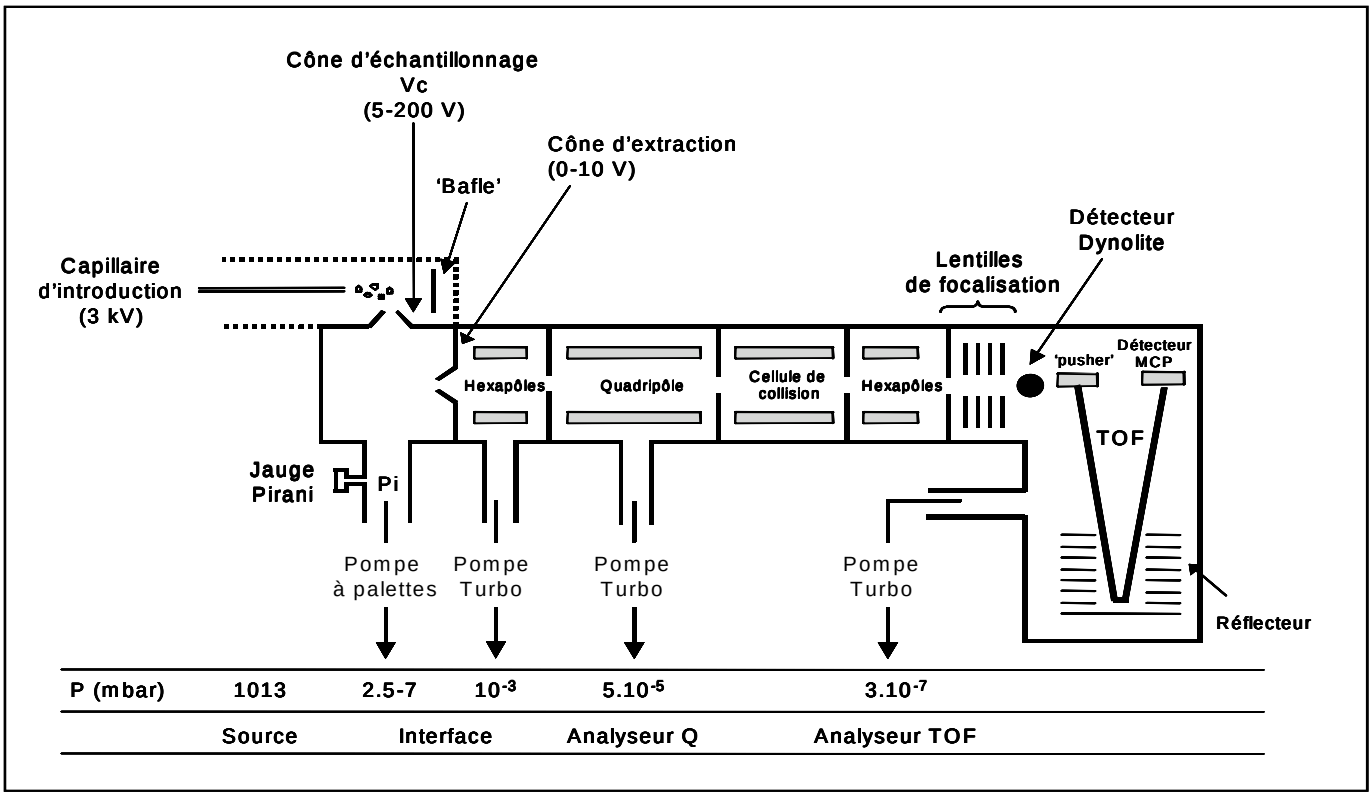
Cet appareil est constitué de l'assemblage de deux analyseurs, un premier analyseur quadripôlaire suivi d'un tube de vol, les deux étant séparés par une cellule de collision. Dans cette configuration le quadripôle peut fonctionner soit comme analyseur, soit comme guide d'ions (U=0).

Lorsque le quadripôle fonctionne comme guide d'ion (U=0), les ions sont analysés par le TOF, le quadripôle fonctionne alors en mode « RF-only ». Cependant, la transmission des ions dans le quadripôle sera dépendante de l'amplitude de la radiofréquence (cf 1.2.1) : la fonction « MS profile » permet de modifier la valeur de l'amplitude de la radiofréquence au cours d'une analyse de façon à laisser passer successivement différentes gammes de rapports m/z.

Lorsque le quadripôle fonctionne comme analyseur, un ion peut être sélectionné dans le quadripôle, transmis vers la cellule de collision dans laquelle il sera fragmenté puis les fragments générés seront analysés dans le second analyseur (le TOF). On parle alors de spectromètre de masse en tandem permettant de faire de la MS/MS sur des ions sélectionnés.

Dans cette configuration, le Q-TOF est particulièrement adapté pour la réalisation d'analyses protéomiques en couplage nanoLC-MS/MS : durant ces analyses, des cycles MS suivis de une ou plusieurs MS/MS sur les peptides sélectionnés permettront d'obtenir des informations sur les peptides trypsiques élués de la colonne chromatographique. C'est dans cette configuration que le Q-TOF II (Waters) a été utilisé durant ce travail (Figure 15).

Figure 15 : Représentation schématique du Q-TOF II (Waters) [Sanglier, 2002].



1.3 La fragmentation peptidique

La spectrométrie de masse en tandem donnant la possibilité de déterminer une partie, voire la totalité de la séquence en acides aminés des peptides trypsiques confère un pouvoir discriminant et une spécificité d'identification considérables à l'analyse protéomique par spectrométrie de masse. Or, l'identification des séquences peptidiques à partir de l'interprétation des spectres MS/MS souvent très complexes nécessite une connaissance des mécanismes de fragmentation des peptides en phase gazeuse.

- Les fragmentations induites par collision (CID) à basse et haute énergie

Le transfert des ions en phase gazeuse par ESI n'est pas un processus hautement énergétique : les ions entrant dans le spectromètre de masse ont donc des énergies internes faibles [Gaskell, 1997]. L'énergie acquise par les ions pour être dissociés est, la plupart du temps, apportée par activation par collisions avec des molécules de gaz non chargées (Collision Induced Dissociation, CID ou Collision Activated Dissociation, CAD). On peut distinguer deux types de fragmentations par CID, les **fragmentations basse énergie** et les **fragmentations haute énergie**. Les spectres MS/MS obtenus sur des analyseurs de type Q-TOF ou trappe ionique mettent en jeu des énergies de fragmentation de quelques eV, les collisions sont donc de basse énergie. A coté de cela, la spectrométrie de masse en tandem obtenue sur des analyseurs de type TOF/TOF ou magnétique fait intervenir des énergies de fragmentation de quelques keV et est appelée collision à haute énergie. Les fragmentations basses et hautes énergies génèrent des profils de fragmentation différents : durant ce travail de thèse, l'ensemble des expériences ayant été réalisées sur des instruments faisant intervenir des fragmentations basse énergie (Q-TOF et trappe ionique), seuls les mécanismes de fragmentation basse énergie sont décrits ici.

- La nomenclature des fragmentations peptidiques

L'observation de la fragmentation d'un grand nombre de peptides a permis à Roepstorff et Fohlmann d'établir une première nomenclature de fragmentation peptidique [Roepstorff et coll., 1984] qui a été complétée et simplifiée par Biemann quelques années plus tard [Biemann et coll., 1990]: cette nomenclature fait toujours référence dans le domaine aujourd'hui (Figure 16).

Deux types d'ions sont observés :

- Les ions pour lesquels la charge positive est portée par la partie contenant l'acide aminé N-terminal : **les ions des séries a, b et c**.
- Les ions pour lesquels la charge positive est portée par la partie contenant l'acide aminé C-terminal : **les ions des séries x, y et z.**

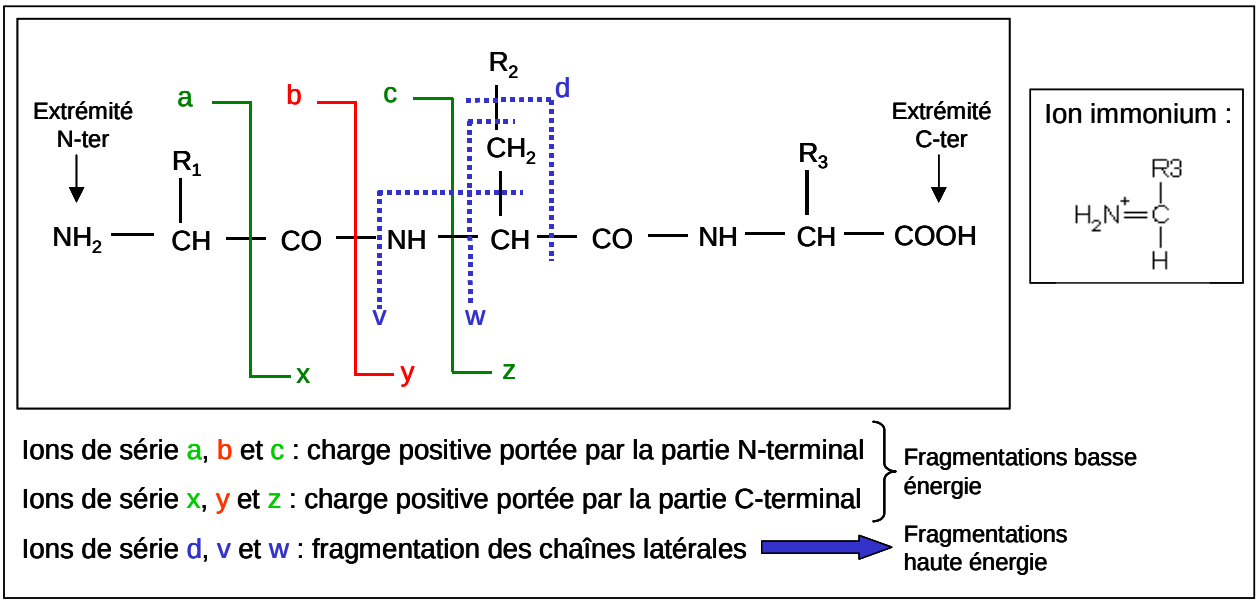
Les séries b- et y- sont observées à la fois lors des fragmentations basses et hautes énergies. Les différences de masse entre des ions consécutifs d'une même série permettent de déterminer l'identité des acides aminés consécutifs et donc de déduire la séquence du peptide à l'exception du couple d'acides aminés isobares Leucine/Isoleucine et du couple Glutamine/Lysine ayant une différence de masse de 0,036 Da, trop faible pour pouvoir être différenciés.

Des doubles fragmentations peuvent générer ce que l'on appelle des **fragments internes** : les fragments internes ne contenant qu'une seule chaîne latérale et formés par la combinaison d'une coupure de type a- et d'une coupure de type y- sont appelés **ions immoniums**. La présence de ces ions immoniums dans

les basses masses sur les spectres MS/MS est très informative pour l'interprétation des spectres et confirme la présence de certains acides aminés dans la séquence.

Les fragmentations des chaînes latérales générant des ions des séries d-, v- et w- n'apparaissent que lors de fragmentations à haute énergie.

Figure 16 : Illustration de la nomenclature des fragmentations d'après Biemann sur un tripeptide, les chaînes latérales sont notées R1, R2 et R3.



De la même manière, d'autres nomenclatures ont été développées afin d'élucider et d'attribuer les fragments issus de la fragmentation par spectrométrie de masse des oligonucléotides [McLuckey et coll., 1992] et des oligosaccharides [Domon et coll., 1988].

➢ Les fragmentations basse énergie et le modèle du « proton mobile »

Pour la plupart des peptides, en fragmentation basse énergie, les fragmentations au niveau des liaisons amides sont dominantes, ce qui résulte en des ions des **séries b- et y- majoritaires** et facilite la déduction de la séquence en acides aminés des peptides.

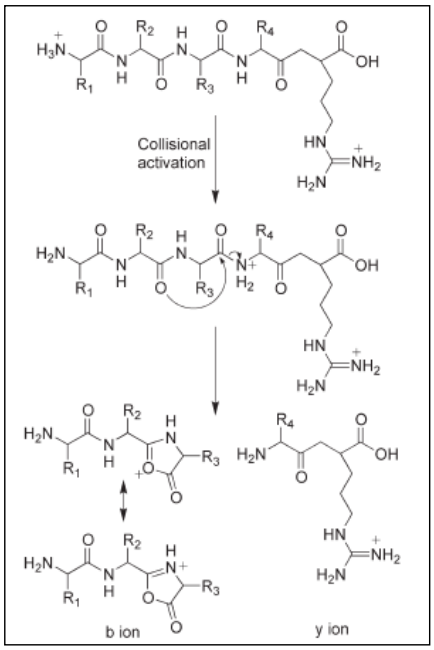
Plus précisément, l'énergie requise pour la fragmentation des peptides est apportée par activation par les collisions avec les molécules de gaz. L'addition d'énergie par activation collisionnelle altère la population initiale des formes protonnées du peptide et entraîne une augmentation de la population des formes ayant des énergies plus élevées que les formes les plus stables : dans ces molécules excitées, le proton est délocalisé à différents endroits sur le squelette peptidique. C'est la présence de ce proton qui va initier la fragmentation au niveau des liaisons peptidiques pour former des ions des séries b- et y- (Figure 17) : on dit que ces fragmentations basses énergies sont **dirigées par la charge** (« **charge-directed fragmentation** ») et suivent le **modèle du « proton mobile »** décrit en 1996 par Dongré et coll. [Dongré et coll., 1996].

L'énergie requise pour initier la fragmentation d'un peptide dépend de sa composition en acides aminés (présence ou absence d'acides aminés basiques), de sa séquence et de sa taille [Dongré et coll., 1996]. Pour des **peptides monochargés**, la présence d'acides aminés basiques (Arginine>Lysine) augmente

considérablement l'énergie nécessaire à leur fragmentation : le proton est « séquestré » sur la chaîne latérale de l'acide aminé basique et l'énergie requise pour le délocaliser sur le squelette peptidique est plus élevée.

Pour des **peptides trypsiques doublement chargés** (espèces majoritaires à 65-80% dans le cas de peptides trypsiques [Savitski et coll., 2005]), les deux charges sont généralement portées par l'extrémité N-terminale et par la chaîne latérale de la lysine ou de l'arginine C-terminale : la barrière énergétique pour transférer le proton N-terminal sur le squelette peptidique est faible (par rapport au proton séquestré par l'acide aminé basique en C-terminal) ce qui résulte en la génération des ions b- et y- (Figure 17). Les ions y- sont souvent plus intenses ce qui s'explique par leur plus grande stabilité (le proton étant séquestré par le résidu C-terminal) par rapport aux ions b-. Les ions b- n'ayant pas de site fortement favorable à la charge, ils sont sujets à de nouvelles fragmentations [Tang et coll., 1993].

Figure 17 : Schéma de fragmentation pour l'obtention d'ions b- et y- à partir de peptides multichargés par CID, le « proton mobile » est délocalisé sur la chaîne peptidique et entraîne la fragmentation « dirigée par la charge » du peptide [Lane, 2005].



Ces règles de fragmentation sont relativement simples pour le cas particulier des peptides trypsiques doublement chargés qui sont, par chance, les plus abondants dans nos analyses protéomiques. Par contre, la présence de certains acides aminés dans la séquence peptidique résulte en des sites de coupure préférentielle.

- Quelques exemples de coupures préférentielles lors de fragmentations à basse énergie :

Les fragmentations en **N-terminal des résidus proline** sont favorisées car l'azote de la proline est plus basique que les azotes des autres acides aminés de la chaîne peptidique et va donc préférentiellement séquestrer le proton mobile. En revanche, les fragments en C-terminal de la proline sont rares à cause de l'encombrement stérique du cycle de la proline [Breci et coll., 2003].

Pour les peptides contenant des arginines et un résidu acide, acide aspartique ou glutamique, un clivage préférentiel a lieu **en C-terminal du résidu acide** lorsque le nombre de charges du peptide est inférieur ou égal au nombre d'arginines : le proton est alors séquestré par l'arginine et c'est le proton du groupement acide de l'acide aminé acide qui va initier la fragmentation. Ce phénomène ne se produit pas lorsque

l'arginine est remplacée par une lysine, résidu moins basique, le proton étant alors moins séquestré [Wysocki et coll., 2000].

Un clivage préférentiel a lieu en **C-terminal des résidus histidine** pour les peptides doublement chargés contenant 1) une histidine dont la chaîne latérale est protonnée et 2) une arginine protonée. Ce phénomène n'a pas lieu pour les peptides monochargés contenant une arginine et une histidine car dans ces cas, l'unique proton est séquestré par l'arginine [Wysocki et coll., 2000].

De nombreuses équipes travaillent à l'établissement de ce type de règles de fragmentations très précises qui pourront ensuite être implémentées dans les algorithmes de recherche en banques de données avec des données MS/MS. L'établissement de telles règles est un outil précieux pour l'amélioration de l'efficacité et de la spécificité de ces algorithmes [Paizs et coll., 2005].

1.4 Conclusions

- Instrumentation

Les analyseurs et instruments décrits dans ce chapitre sont les plus répandus et les plus largement utilisés pour l'analyse protéomique par spectrométrie de masse. Chacun présentant ses forces et ses faiblesses par rapport aux performances attendues de la spectrométrie de masse en analyse protéomique.

Actuellement, à coté de ces instruments « traditionnels » de l'analyse protéomique, les spectromètres de type FT-ICR présentent un attrait particulier pour les protéomistes malgré leur coût onéreux, de part leur haute résolution et donc aussi haute précision des mesures de masse. Aussi, très récemment, un nouveau type d'analyseur appelé Orbitrap a vu le jour : c'est le premier analyseur introduit depuis les 30 dernières années se basant sur un nouveau principe physique (la séparation des ions dans un champ électrique oscillant) et il permet d'atteindre des performances en résolution et précision de masse proches de celles des instruments de type FT-ICR [Hardman et coll., 2003 ; Hu et coll., 2005] (Résolution=150000, large capacité de charge, haute précision de masse 2-5ppm, gamme de masse d'au moins 6000, gamme dynamique de 10).

- De nouvelles techniques de fragmentation

Dans ce chapitre, seules les fragmentations par CID ont été décrites. Or, en 2004, une nouvelle méthode de fragmentation des peptides basée sur la dissociation par transfert d'électrons (Electron Transfer Dissociation, ETD) a été développée par l'équipe de Hunt [Syka et coll., 2004] et adaptée sur des instruments de type trappe ionique. Cette méthode de fragmentation met en jeu des mécanismes comparables aux mécanismes de l'Electron Capture Dissociation (ECD) sur les instruments de type FT-ICR [Zubarev et coll., 1998]. L'acquisition d'une trappe ionique HCT Ultra PTM Discovery System (Bruker Daltonics) au laboratoire en début d'année 2006 nous a permis d'évaluer les performances et les possibilités de cette nouvelle méthode de fragmentation : une description instrumentale, le principe de l'ETD et les premiers résultats obtenus sont décrits dans la partie I, chapitre II des résultats.

2 Les techniques de séparation des protéines et peptides

En parallèle des développements instrumentaux de la spectrométrie de masse, les développements de techniques séparatives des protéines en amont de l'analyse par spectrométrie de masse ont également contribués à l'émergence de l'analyse protéomique par spectrométrie de masse [Yates, 2000 ; Görg et coll., 2004 ; Paoletti et coll., 2004].

L'approche de séparation des protéines par gel d'électrophorèse bidimensionnel (gel 2D), introduite dès les années 70 [Kenrick et coll., 1970] n'en reste pas moins aujourd'hui un outil incontournable et largement exploité en analyse protéomique (cf 2.1) [Görg et coll., 2004]. Néanmoins, des méthodes alternatives au gel 2D ont vu le jour plus récemment parmi lesquelles les techniques de séparation des protéines par chromatographie liquide multidimensionnelle (MudPIT) utilisée pour les approches de « shotgun » protéomique (cf 2.2.2) [Wolters et coll., 2001 ; Washburn et coll., 2001 ; Paoletti et coll., 2004].

2.1 Le gel d'électrophorèse bidimensionnel (gel 2D)

2.1.1 Principe et colorations

Dès 1975, O'Farrel reporte la détection de 1000 spots distincts sur un gel 2D [O'Farrel, 1975]. En 2004, Görg et coll. annoncent qu'un gel 2D peut résoudre jusqu'à 5000 protéines simultanément (2000 protéines en routine) et détecter et quantifier >1ng de protéine par spot [Görg et coll., 2004].

Lors de la séparation d'un mélange complexe de protéines par gel 2D, l'échantillon est séparé suivant une première dimension en fonction du point isoélectrique (pI) des protéines (iso-électrofocalisation, IEF) puis suivant une deuxième dimension en fonction de leurs masses moléculaires (électrophorèse SDS-PAGE) [Kenrick et coll., 1970]. Le gel 2D permet de séparer plusieurs centaines de formes protéiques et en particulier un grand nombre de variants post-traductionnels (formes phosphorylées, glycosylées, …), avec une dimension quantitative et une capacité unique de séparer des protéines entières.

Le remplacement des ampholytes par des gels à gradients de pH immobilisés (IPG) a constitué un pas décisif dans l'évolution pour la première dimension des gels 2D, améliorant considérablement la reproductibilité, la flexibilité et la capacité des séparations par gel 2D [Bjellqvist et coll., 1982 ; Görg et coll., 1988].

Par ailleurs, la sensibilité de détection des gels passe par la sensibilité des colorants utilisés. Les colorations les plus couramment utilisées en analyse protéomique sont :

- le bleu de Coomassie : malgré la faible sensibilité de cette coloration, elle présente le grand avantage d'être linéaire et donc très adaptée pour les études comparatives.

- la coloration à l'argent de part sa limite de détection 100 fois plus faible que la coloration au bleu de Coomassie [Shevchenko et coll., 1996]. Cependant, elle ne témoigne pas de l'exacte stœchiométrie des protéines en présence. Au-delà de certaines concentrations, la coloration devient non linéaire et des incompatibilités avec l'analyse successive par spectrométrie de masse MALDI et ESI ont été mises en évidence [Gevaert et coll., 2000]. Pour ce dernier point, un protocole améliorant la compatibilité de cette coloration avec la MS a été développé [Richert et coll., 2004].

- les **colorations fluorescentes**, le Sypro Ruby commercial ou le Rhutenium II [Rabilloud et coll., 2001] qui présentent une sensibilité intermédiaire entre les colorations au nitrate d'argent et au bleu de Coomassie [Rabilloud et coll., 2001].

2.1.2 Les limitations du gel 2D

Les limites des approches de séparation par gel 2D apparaissent à plusieurs niveaux et des améliorations tentent de minimiser ces limites [Rabilloud, 2002] :

- **L'automatisation** de la réalisation des gels reste difficile même si des progrès ont été réalisés dans ce sens : (i) la commercialisation de gels prêts à l'utilisation, (ii) la commercialisation de l'IPGphor ou d'autres instruments permettant l'automatisation de l'IEF, (iii) la semi-automatisation de la coloration à l'argent permettant la coloration d'une dizaine de gels simultanément, (iv) des nouveaux protocoles de coloration rapides, (v) des algorithmes performants d'analyse et de comparaison des images, (vi) des robots de prélèvement automatique des spots [Görg et coll., 2004], …
- **La résolution :** Face à l'énorme diversité des protéines cellulaires, la résolution des gels 2D n'est pas toujours suffisante pour séparer l'ensemble des protéines et la co-migration de plusieurs protéines par spot n'est pas rare [Gygi et coll., 2000 ; Campostrini et coll., 2005]. Pour améliorer encore la résolution des gels, des « zoom gels » et des gels à gamme de pH réduite ont été développés [Hoving et coll., 2000].
- **La gamme dynamique visible sur gel :** La gamme dynamique maximale visible sur gel 2D est de 10^4 au maximum. Or, pour donner un exemple révélateur, l'actine, la protéine généralement la plus abondante dans les cellules humaines, est présente à 10^8 molécules par cellule. A coté de cela, des récepteurs cellulaires ou des facteurs de transcription sont présents à 100-1000 molécules par cellule : cela représente donc une gamme dynamique de 10^5 à 10^6. A plus forte raison encore, dans des sérums, la gamme dynamique peut atteindre 10^9 [Rabiloud, 2002]. Des méthodes de pré-fractionnement sont donc nécessaires afin de visualiser les protéines minoritaires.
- **La visualisation des protéines hydrophobes et membranaires :** Les protéines membranaires sont largement sous-représentées sur les gels 2D et ce principalement pour 3 raisons : elles sont en général peu abondantes, leurs pI sont alcalins et elles sont faiblement solubles dans les milieux aqueux utilisés pour l'IEF [Santoni et coll., 2000]. D'importantes améliorations ont été apportées en incorporant des agents de solubilisation, surfactants et détergents, tels que la thiourée ou le CHAPS [Rabilloud et coll., 1997]. Cependant, malgré les progrès réalisés dans ce domaine, les problèmes d'extraction et de solubilisation des protéines membranaires resteront vraisemblablement les plus limitants pour les approches par gel 2D [Rabilloud, 2002].

Au vu de toutes ces limitations, même si le gel 2D est encore largement utilisé, des techniques alternatives ont été développées éliminant l'étape d'IEF. En effet, même si l'IEF est à l'origine de la haute résolution du gel 2D notamment pour la séparation des variants post-traductionnels, elle est aussi la cause de la plupart des limitations du gel d'électrophorèse 2D.

2.2 La chromatographie liquide (LC) et les alternatives au gel 2D

Les techniques de séparation par chromatographie liquide ont fait l'objet de nombreuses adaptations pour rendre le couplage LC-spectrométrie de masse ESI (compatible avec l'introduction des échantillons en solution) robuste et sensible [Delalande, 2003].

En ESI-MS l'intensité du signal est concentration-dépendante et non débit-dépendante [Ikonomou et coll. ; 1990]. La recherche d'une sensibilité toujours meilleure a donc naturellement conduit au développement de systèmes nanoLC permettant de travailler à des débits de l'ordre de 100 à 300 nl/min. Les avancées techniques réalisées dans le domaine des nano-colonnes de chromatographie et de l'instrumentation de nanoHPLC ont ainsi amené le couplage nanoLC au cœur de l'analyse protéomique actuelle : avec les systèmes actuels, le couplage nanoLC-MS ou nanoLC-MS/MS permet de détecter jusqu'à quelques femtomoles de matériel.

La nanoLC de phase inverse (RP : Reverse phase) est classiquement utilisée pour séparer et concentrer un mélange plus ou moins complexe de peptides. Cependant, la complexité des échantillons en analyse protéomique est souvent telle que la nanoLC ne peut pas être utilisée seule, sans autre technique de séparation en amont. En effet, des extraits protéiques cellulaires peuvent contenir plusieurs milliers de protéines qui génèreront des centaines de milliers de peptides trypsiques qui ne pourront pas être séparés efficacement sur nano-colonnes.

Cette technique est donc souvent couplée avec une autre technique séparative en amont. Elle peut être couplée au gel 2D [Lee et coll., 2005] mais d'autres alternatives ont été développées comme le couplage avec une séparation SDS-PAGE ou la chromatographie multidimensionnelle afin de s'affranchir des limites liées au gel 2D.

2.2.1 Le gel SDS PAGE (gel 1D) couplé à la nanoLC

Le gel SDS (sodium dodecylsulfate) PAGE (PolyAcrylamide Gel Electrophoresis) permet de séparer les protéines en fonction de leurs masses apparentes quelles que soient leurs caractéristiques physico-chimiques. Facile et rapide à mettre en œuvre, la séparation par SDS-PAGE peut être utilisée seule (sans séparation par IEF) pour pré-fractionner des échantillons complexes de protéines avant leur analyse par spectrométrie de masse. Néanmoins, son pouvoir de séparation reste relativement faible et la présence de nombreuses protéines par bande de gel 1D explique son couplage avec un système de nanoLC-MS/MS. Dans cette stratégie, deux étapes de séparations sont donc réalisées, une au niveau protéique et une au niveau peptidique [Rabilloud, 2002]:

- Une étape de **séparation des protéines** par gel SDS-PAGE. Les bandes de gel 1D contenant une ou plusieurs protéines sont découpées puis les protéines sont digérées in-gel.
- Une étape de **séparation des peptides** par nanoLC. Les peptides sont séparés sur des nano-colonnes de phase inverse, élués par un gradient eau/acetonitrile avant d'entrer dans le spectromètre de masse.

Le principal avantage de cette approche par rapport au gel 2D est qu'elle permet d'analyser les protéines basiques et hydrophobes ; son principal champ d'application est donc l'analyse des protéomes membranaires [Bell et coll., 2001 ; Xiong et coll., 2005].

2.2.2 Les techniques de chromatographie liquide multidimensionnelle

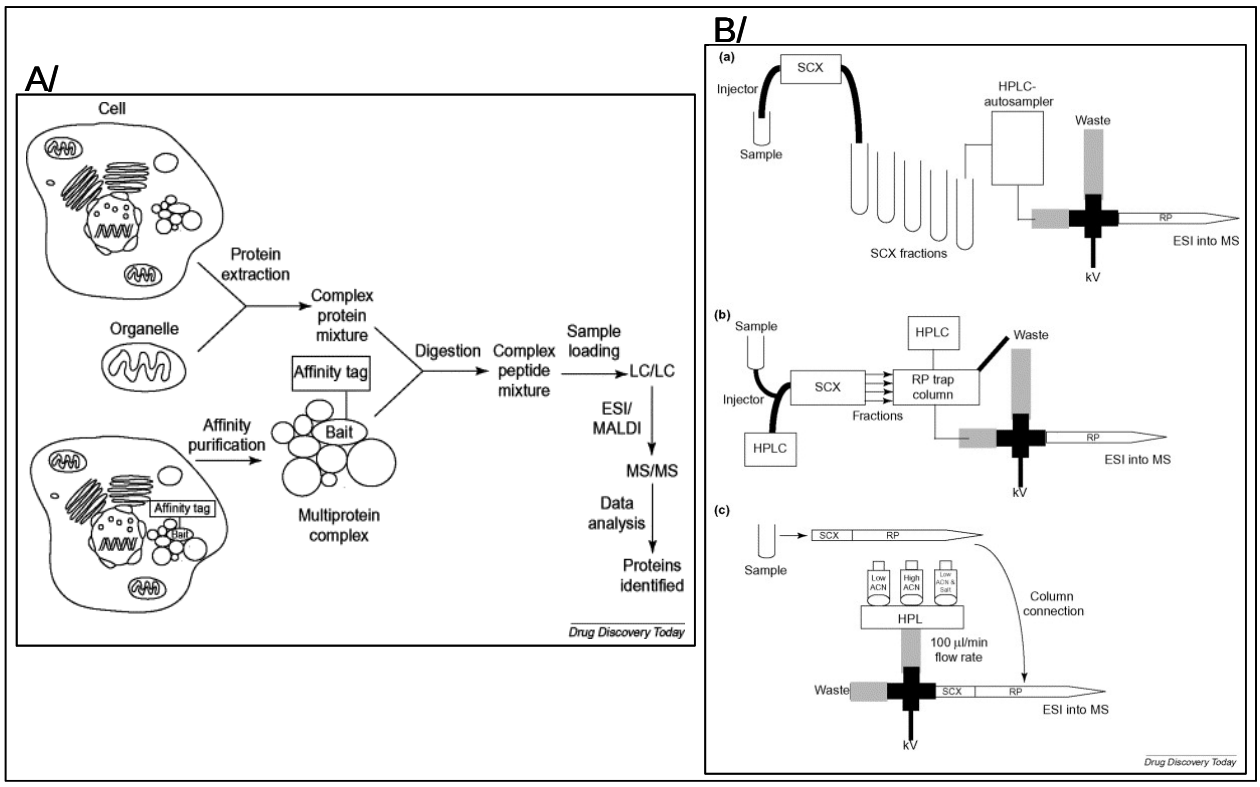
Les approches décrites précédemment impliquant le passage par une étape de gel 2D ou 1D font appel à des digestions in-gel des protéines, processus ne présentant pas de très bons rendements dû à la perte de certains peptides, notamment des plus grands lors de l'étape d'extraction des peptides du gel [Rabilloud, 2002]. Cette limitation a encouragé le développement de méthodologies sans gel. C'est ainsi qu'une approche par chromatographie multidimensionnelle a été introduite par le laboratoire de Yates en 1999 [Link et coll., 1999]. Initialement appelée **DALP (Direct Analysis of Large Protein Complexes)** car elle a été appliquée dans un premier temps à l'étude du complexe ribosomique 80S de *S. cerevisiae*, cette approche a ensuite été appelée **MudPIT (Multi-Dimensional Protein Identification Technology)** [Washburn et coll., 2001].

Ces nouvelles stratégies ont ouvert la voie à ce que l'on appelle la protéomique « **shotgun** ». Par analogie avec les approches « shotgun » de séquençage des génomes, les approches de « shotgun » protéomique désignent l'analyse directe de mélanges complexes de protéines pour générer rapidement un profil global des protéines présentes dans le mélange [Wolters et coll., 2001 ; Wu et coll., 2002]. Le mélange complexe de protéines est directement digéré en peptides qui sont séparés par des étapes de chromatographies multiples avant d'être analysés en MS et fragmentés en MS/MS comme illustré en figure 18A/ [Swanson et coll., 2005].

L'approche la plus utilisée consiste à coupler une première étape de séparation des peptides par une chromatographie d'échange de cations (SCX : Strong Cation Exchange) suivie d'une étape de chromatographie de phase inverse. Trois approches ont été décrites dans la littérature pour la réalisation des analyses SCX/RP/MS/MS (Figure 18B/) :

- **La stratégie off-line** (Figure 18 B/(a)) : L'étape de SCX est réalisée de façon indépendante, les fractions sont collectées avant d'être analysées individuellement en nanoLC-MS/MS avec séparation sur colonne de phase inverse classique. N'étant pas directement couplée à la colonne RP, cette approche est très flexible quant aux solvants utilisés pour la SCX (chlorure de sodium et chlorure de potassium peuvent être utilisés malgré leur incompatibilité avec l'analyse par MS).
- **La stratégie on-line** (Figure 18 B/(b)) : Dans cette configuration les peptides sont élués de la colonne SCX sur une colonne RP d'enrichissement. Cette étape permet d'éliminer les éluants utilisés incompatibles avec la MS avant l'élution des peptides de la colonne d'enrichissement sur la colonne RP couplée au spectromètre de masse [Nagele et coll., 2004]. Néanmoins, l'architecture de la valve dans cette configuration est très complexe, ce qui rend cette configuration délicate à mettre en œuvre. Par contre, une variante de cette configuration utilisant des solvants compatibles avec la MS comme l'acétate d'ammonium pour la SCX et ne nécessitant donc pas de colonne d'enrichissement a également été décrite [Xiang et coll., 2004].
- **La stratégie MudPIT** (Figure 18 B/(c)) : Les peptides sont chargés sur une colonne microcapillaire bi-phasique remplie en série par une résine SCX et par une résine de phase inverse. La colonne chargée est ensuite intégrée dans le système et les peptides sont élués de la résine SCX par paliers avec un gradient en sels pour se déposer en tête de la phase RP. Ils sont ensuite élués une seconde fois en fonction de leur hydrophobicité croissante vers le spectromètre de masse [Link et coll., 1999 ; Washburn et coll., 2001 ; Wolters et coll., 2001].

Figure 18 : A/ Représentation d'une expérience classique de « shotgun » protéomique B/ Différentes stratégies décrites pour la réalisation d'expériences de SCX/RP/MS/MS, la configuration off-line (a), la configuration on-line (b) et la configuration MudPIT (c). D'après Swanson [Swanson et coll., 2005].



Le nombre de protéines pouvant être identifiées par ces approches MudPIT peut largement dépasser le nombre de protéines identifiées par des approches par gel 2D : 1484 protéines uniques ont pu être identifiées par rapport à 300 protéines identifiées après gel 2D dans un lysat cellulaire total de *Saccharomyces cerevisiae* [Wolters et coll., 2001]. Cependant, avec ces approches, le nombre de peptides servant à l'identification des protéines est souvent faible (les identifications peuvent contenir des taux de faux positifs non négligeables) et l'interprétation des données « volumineuses » (classiquement 10000 à 100000 spectres MS/MS) générées durant ces analyses soulève donc quelques difficultés [Nesvizhskii et coll., 2005]. Dans les approches « shotgun », la digestion des protéines constitue la première étape et les séparations ne se font donc qu'au niveau peptidique. La relation entre les peptides et la protéine de départ est donc perdue. Or un peptide d'une séquence donnée peut être présent dans plusieurs protéines ou plusieurs isoformes de protéines ce qui rend les identifications souvent difficiles et ambiguës : la nécessité d'établissement d'une nomenclature et de critères précis pour la validation et la présentation des identifications réalisées est donc évidente. L'établissement des directives pour la publication de données de protéomique (cf chapitre II) s'applique d'ailleurs tout particulièrement aux données générées en « shotgun » protéomique [Bradshaw, 2005].

2.3 Conclusion

L'analyse de mélanges protéiques de plus en plus complexes étant envisagée, l'ensemble des techniques séparatives (au niveau protéique et peptidique) est utilisé en « associations » de plus en plus originales afin d'obtenir un maximum d'informations sur les protéines présentes. Néanmoins, même si des résultats obtenus avec différentes approches sont publiés, la publication de travaux comparatifs entre ces différentes approches est encore rare [Swanson et coll., 2005]. A titre d'exemples, quelques combinaisons et innovations décrites dans la littérature peuvent être citées :

- l'utilisation d'une **chromatographie échangeuse d'anions** en première dimension remplaçant la SCX [Mawuenyega et coll., 2003].
- l'utilisation d'une première étape d**'IEF des peptides** avant l'analyse nanoLC-MS/MS a été comparée à l'analyse avec une première étape de SCX et 13% de protéines supplémentaires ont ainsi été identifiées [Essader et coll., 2005].
- l'utilisation de **séparations à 3 dimensions** a été décrite : une étape de séparation des protéines par gel filtration, les fractions collectées sont digérées puis les peptides sont séparés par SCX off-line avant leur analyse en nanoLC-MS/MS [Jacobs et coll., 2004]. Beausoleil et coll. ont utilisé une première séparation des protéines par gel 1D SDS-PAGE avant digestion, séparation des peptides par SCX off-line puis nanoLC-MS/MS et ont ainsi pu identifier 2000 sites de phosphorylation dans la fraction nucléaire de cellules HeLa [Beausoleil et coll., 2004].

Enfin, une nouvelle méthode de séparation des protéines ou des peptides par isoélectrofocalisation, l'**IEF Off-Gel**^{TM} a été développée permettant de recouvrer les protéines fractionnées en solution, tout comme d'utiliser une importante charge de départ (>1mg de protéines) [Ros et coll., 2002 ; Michel et coll., 2003 ; Michel et coll., 2006].

3 Le séquençage des génomes et les banques de données de séquence

L'évolution de la protéomique n'aurait pas été possible sans les résultats antérieurs de la génomique. Les séquences de génomes disponibles sont la ressource de base pour l'identification rapide des protéines par corrélation entre des données de spectrométrie de masse de peptides et des banques de séquences.

Le terme **génomique** a été inventé par Thomas Roderick en 1986 lors d'une discussion sur le nom d'un journal, *Genomics*. Au départ, la génomique s'est principalement consacrée à l'analyse des données de séquences et à la détermination des séquences codantes sur le génome. La nouvelle tendance, après la détermination des séquences, s'oriente vers l'étude de la fonction des gènes, appelée **génomique fonctionnelle** [Hieter et coll., 1997]. L'**analyse protéomique** tient une place de choix dans ce contexte post-génomique dans lequel il est rapidement apparu que le séquençage des génomes n'était qu'une étape dans la compréhension des phénomènes biologiques.

Les premiers génomes complets à avoir été séquencés sont ceux d'*Haemophilus influenzae* (1,8 Mbps) en 1995 puis de la levure *Saccharomyces cerevisiae* (14 Mbps) en 1996 [Goffeau et coll., 1996]. Dès lors la course au séquençage a été lancée : en **août 2006**, un total de **2127 projets de séquençage de génomes** sont en cours et **409 génomes complets** ont déjà été publiés. L'ensemble des projets de séquençage de génomes est compilé dans la « Genomes OnLine Database (GOLD) » accessible sur le site http://www.genomesonline.org/ [Liolios et coll., 2006]. Ce sont les énormes progrès réalisés pour l'automatisation des stratégies de séquençage des génomes qui ont permis d'accroître ainsi la productivité des biologistes dans la détermination de nouvelles séquences.

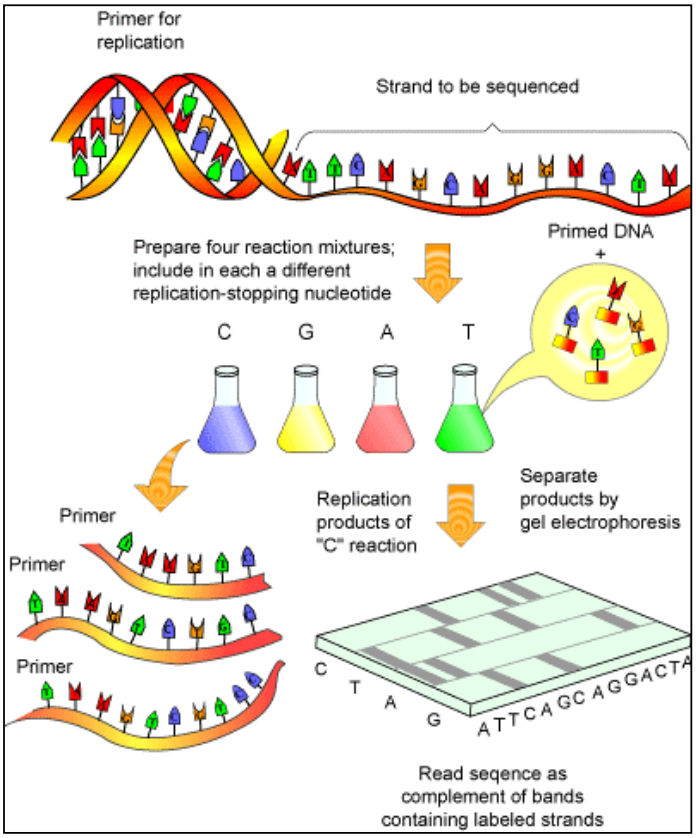
3.1 Le séquençage des génomes

3.1.1 Le séquençage de l'ADN

La même année, en 1977, deux méthodes de séquençage de l'ADN ont été publiées :

- **une méthode chimique** [Maxam et coll., 1977], la technique de **Maxam-Gilbert**, qui utilise des produits chimiques pour cliver l'ADN en fragments au niveau de bases spécifiques. Son principal inconvénient est la toxicité des composés chimiques utilisés.
- **une méthode biochimique** [Sanger et coll., 1977], la plus utilisée, **la technique de Sanger** qui lui a valu un prix Nobel en 1980 (Figure 19). Le principe de cette méthode consiste à initier la polymérisation de l'ADN à l'aide d'un petit oligonucléotide (amorce) complémentaire à une partie du fragment d'ADN à séquencer. L'élongation de l'amorce est réalisée par le fragment de Klenow (une ADN polymérase I dépourvue d'activité exonucléase 5'->3'). Les quatre désoxynucléotides (dATP, dCTP, dGTP, dTTP) sont ajoutés, ainsi qu'en faible concentration les quatre didésoxynucléotides (ddATP, ddCTP, ddGTP, ddTTP). Lorsqu'un didésoxynucléotide est incorporé à la nouvelle chaîne synthétisée, il empêche la poursuite de l'élongation. Ceci conduit donc à la génération de fragments d'ADN de taille variable qui sont ensuite séparés par électrophorèse sur un gel de polyacrylamide.

Figure 19 : Illustration de la méthode biochimique de séquençage de l'ADN, la méthode de Sanger.



3.1.2 Les stratégies de séquençage des génomes

Selon les projets, deux stratégies distinctes sont utilisées pour le séquençage des génomes (Figure 20) :

- La stratégie de **séquençage aléatoire global** ou « **en vrac** » (**"whole genome shotgun sequencing "**). Cette technique repose sur un principe simple : découper un génome en un **grand nombre de fragments de petite taille**. Les extrémités d'un certain nombre de ces fragments sont ensuite séquencées, puis ces séquences sont assemblées en contigs sur la base de leurs chevauchements grâce à de puissants programmes informatiques. Enfin, les contigs sont assemblés pour essayer de produire une séquence complète. Les difficultés d'une telle technique sont à deux niveaux : obtenir suffisamment de fragments pour couvrir le génome dans sa totalité et réussir l'assemblage des fragments.

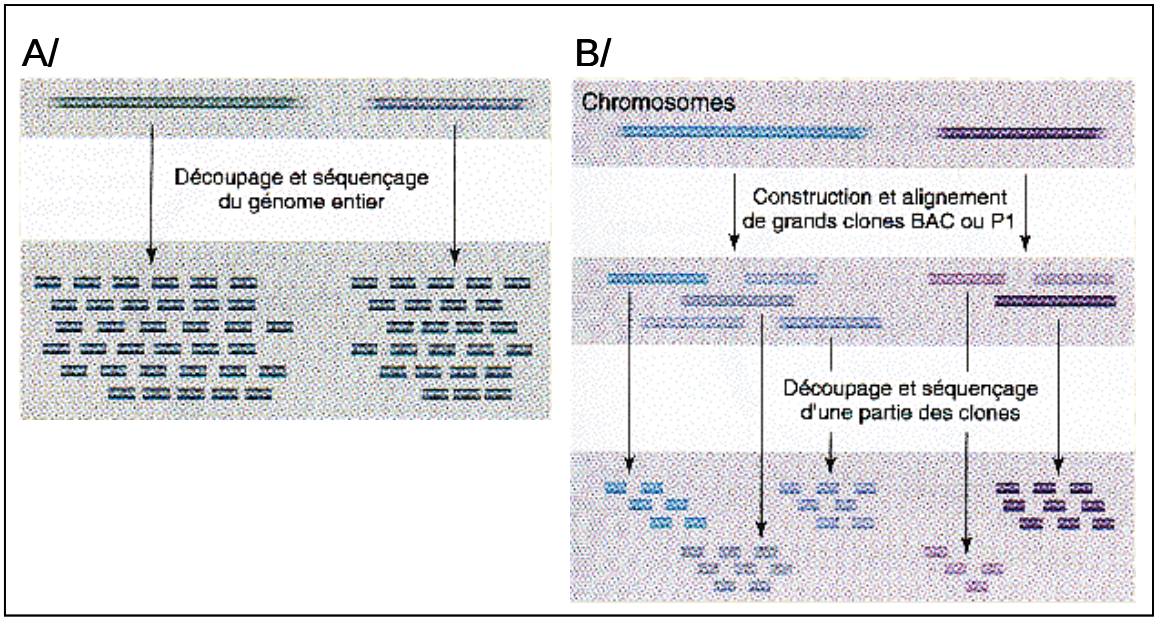
  La génération des fragments d'ADN de même que l'échantillonnage des fragments séquencés sont des processus aléatoires : de ce fait, plus le nombre d'extrémités de fragments séquencés sera important, plus la longueur des blocs de séquences que l'on est en mesure d'assembler augmentera et plus la fraction du génome couverte augmentera de même que la précision de la séquence (car la redondance des séquences permettra de corriger les éventuelles erreurs de séquençage). On parle de **profondeur** pour signifier combien de fois la taille du génome a été recouverte en taille globale de fragments séquencés (par exemple séquençage avec une profondeur 10X).

- La stratégie hiérarchique **« clone par clone »** (dite encore du "**shotgun hiérarchique**" ou "**top-down sequencing**") est celle qui a été adoptée par le consortium international pour le

séquençage du génome humain (HGP : Human Genome Project). Il s'agit d'une démarche en deux temps : l'**établissement d'une carte physique** dans un premier temps permettant d'ordonner un nombre restreint de fragments de grande taille, puis **le séquençage** (de type "shotgun") **de ces fragments**. La carte peut aussi être construite en même temps que le séquençage progresse mais quoi qu'il en soit, la réalisation de cette carte physique préalable est une étape longue et fastidieuse. Les grands fragments, aussi appelés clones, permettant d'établir la carte physique sont clonés dans des vecteurs spéciaux : les YAC (« *Yeast Artificial Chromosome* »), les BAC (« *Bacterial Artificial Chromosome* ») ou des vecteurs dérivés du phage P1 (les PAC).

Le séquençage aléatoire global s'est aujourd'hui imposé dans le cas des génomes bactériens. On l'associe aussi plus volontiers aux sociétés privées parce qu'il est rapide et économique. Toutefois, il faut remarquer que, pour de grands génomes comme celui de l'Homme, ces sociétés se sont souvent appuyées sur des données de cartographie établies par les chercheurs académiques. Au final, pour les grands génomes, on tend aujourd'hui à utiliser des stratégies mixtes, mêlant séquençage aléatoire global et "clone par clone".

Figure 20 : Illustration des deux grandes stratégies de séquençage des génomes. A/ Approche par séquençage aléatoire global (« whole genome shotgun sequencing »). B/ Approche hiérarchique « clone par clone » (« Top-down sequencing approach »).



3.2 Les banques de données de séquence

L'ère du séquençage des génomes a donné naissance à une nouvelle discipline scientifique permettant la gestion des grandes quantités de données numérisées : en 1997, Andrade et Sander définissent la **bioinformatique** comme un nouveau domaine de recherche qui, à partir de données biologiques et par l'utilisation de méthodes informatiques, permet de créer des connaissances nouvelles dans le domaine de la biologie elle-même [Andrade et coll., 1997].

L'aspect sans doute le plus apparent de l'importance prise par la bioinformatique est le développement de nombreuses banques de données biologiques : elles se comptent par centaines et le numéro du 1 Janvier

de chaque année de la revue *Nucleic Acids Research* est consacré aux banques de données en biologie [Database Issue, 2006]. Les informations y sont de nature variée (images, voies métaboliques, structures chimiques, cartographies physiques ou génétiques, taux d'expression de gènes, ...) ; les plus nombreuses et volumineuses concernent les données de séquences [Boguski, 1999]. Le volume de janvier 2006 est le plus volumineux, il contient 94 nouvelles banques et 68 déjà existantes.

Les principaux critères pris en compte pour qualifier et classer les banques de séquences sont le **caractère primaire** ou non des données, une vocation **généraliste** ou au contraire leur **spécialisation**, et le fait qu'elles considèrent principalement un type de données (ici les séquences) ou bien qu'elles en intègrent plusieurs, de diverses sources.

Les **données primaires** sont par définition les **séquences obtenues expérimentalement**, en excluant celles issues de calculs effectués par des programmes informatiques à partir des données primaires. Une stricte acceptation de cette définition en exclurait toutes les séquences protéiques déduites par traduction conceptuelle de séquences nucléotidiques (c'est à dire l'immense majorité de celles qui peuplent les banques), ainsi que les séquences d'ADN génomique obtenues par assemblage de fragments et qui sont des consensus calculés à partir des alignements des fragments chevauchants. Néanmoins, dans les faits, **les banques généralistes nucléotidiques et protéiques** sont considérées comme des banques de données primaires, à côté de celles contenant des données dérivées (alignements, motifs, ...) et des banques spécialisées focalisées sur un organisme ou une problématique biologique, et intégrant souvent d'autres informations que les séquences elles-mêmes.

3.2.1 Les banques de séquences nucléotidiques

Au niveau mondial, trois banques travaillent en étroite collaboration : leurs missions sont de collecter, maintenir et distribuer publiquement l'information primaire représentée par toutes les séquences nucléotidiques connues :

- **EMBL** [Cochrane et coll., 2006] créée en 1980 par l'EMBL (European Molecular Biology Laboratory), est maintenue à l'EBI (European Bioinformatics Institute), http://www.ebi.ac.uk/embl/.
- **GenBank** créée en 1982 au Los Alamos National Laboratory, est maintenue au NCBI (National Center for Biotechnology information), qui dépend du NIH (*National Institute of Health*) américain [Benson et coll., 2006], http://www.ncbi.nlm.nih.gov/Genbank/.
- **DDBJ** (DNA Data Bank of Japan) maintenue par le Centre d'Information Biologique de l'Institut National de Génétique, à Mishima s'est jointe à l'effort des deux précédentes en 1986 [Okubo et coll., 2006], http://www.ddbj.nig.ac.jp/.

Une **collaboration internationale**, officiellement formalisée depuis 1988, existe entre ces trois organismes. Cette collaboration implique un échange quotidien des données collectées, l'application de règles communes relatives à la nomenclature et aux formats utilisés et le fait qu'une entrée ne peut être modifiée que par celui des trois centres auquel elle avait été initialement soumise (pour éviter les « conflits » dans les mises à jour).

Ces trois banques intègrent l'ensemble des séquences nucléotidiques primaires, quelle que soit leur origine, l'organisme, l'espèce moléculaire (ADN, ARN sous toutes leurs formes, y compris les gènes synthétiques) ainsi que la méthodologie mise en jeu pour obtenir la séquence.

➢ Les conséquences de l'exhaustivité des banques

La mission des banques GenBank/EMBL/DDBJ est de constituer une archive complète de l'ensemble des séquences nucléotidiques primaires connues à ce jour. Néanmoins, l'exhaustivité de ces banques entraîne quelques corollaires inévitables :

- une **forte redondance** car la mission de ces banques implique d'intégrer toutes les nouvelles séquences produites et ce même si elles sont déjà présentes dans la banque.
- une **hétérogénéité des données** due aux différentes natures de molécules séquencées, à la qualité variable des séquences selon les projets et à la nature des annotations (les informations accompagnant les séquences sont sous la seule responsabilité des auteurs des séquences).

➢ L'organisation des données

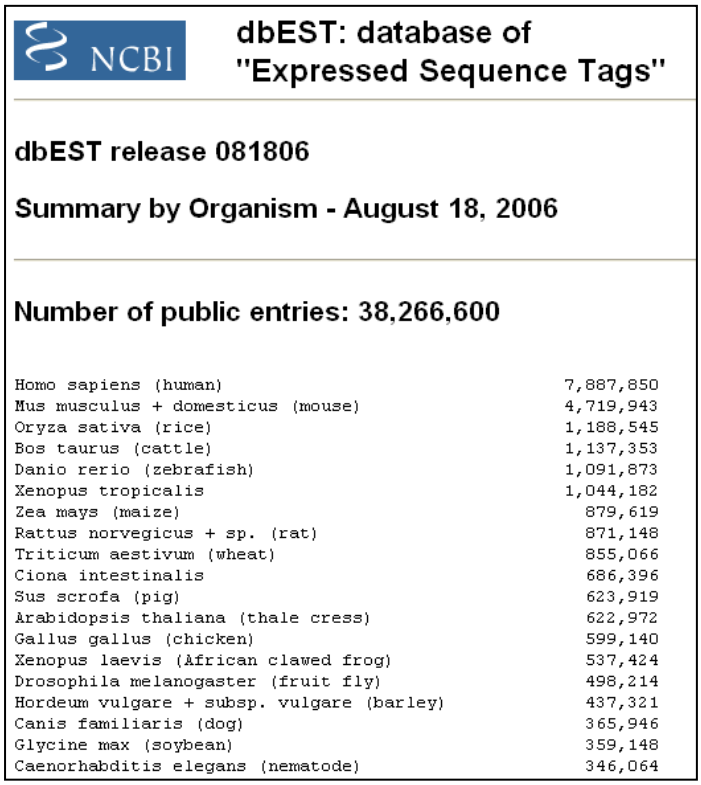
La première génération de séquences, celles des années 80-90, correspond essentiellement à des gènes isolés et clonés. Le début des années 90 est marqué par les premières productions massives de séquences d'EST (« Expressed Sequence Tags ») suivies ensuite par le séquençage des premiers chromosomes et génomes. Ces derniers projets de séquençage sont responsables de la croissance exponentielle du volume des banques de ces 15 dernières années (cf conclusion). Le format des données disponibles dépend de la stratégie de séquençage utilisée, de la taille du génome et de l'état d'avancement du séquençage : les données génomiques peuvent être accessibles sous formes de BACs, de YACs, de contigs, ou parfois d'une longue et unique séquence consensus lorsque l'assemblage du génome est terminée.

Chaque banque propose une organisation de ses entrées sous forme de regroupements en sections : les sections sont essentiellement définies par des critères taxonomiques alors que d'autres correspondent au regroupement de données selon leur nature, comme les sections réservées aux ESTs, aux STS (« Sequence Tagged Sites »), aux GSS (« Genome Survey Sequences ») ou aux HTG (« High Throughput Genomic Sequences »), …

➢ Les banques d'ESTs (« Expressed Sequence Tag »)

Les séquences d'ESTs sont des lectures uniques des extrémités de clones de banques d'ADN complémentaires. Ces séquences courtes ne contiennent donc que les parties codantes des gènes : les exons. Par conséquent, elles sont un outil important pour la détermination des exons dans les séquences génomiques d'eucaryotes supérieurs. Le NCBI et l'EBI proposent chacun une banque ne contenant que des ESTs. Un aperçu de la banque dbEST du NCBI est présenté en figure 21, indiquant la distribution des ESTs selon l'organisme en août 2006. Il apparaît que la grande majorité des ESTs disponibles proviennent *d'Homo spaiens* et de *Mus musculus*.

Figure 21 : Distribution des entrées de la banque dbEST du NCBI au 18 août 2006, http://www.ncbi.nlm.nih.gov/dbEST/.



NCBI

dbEST: database of "Expressed Sequence Tags"

dbEST release 081806

Summary by Organism - August 18, 2006

Number of public entries: 38,266,600

Homo sapiens (human)	7,887,850
Mus musculus + domesticus (mouse)	4,719,943
Oryza sativa (rice)	1,188,545
Bos taurus (cattle)	1,137,353
Danio rerio (zebrafish)	1,091,873
Xenopus tropicalis	1,044,182
Zea mays (maize)	879,619
Rattus norvegicus + sp. (rat)	871,148
Triticum aestivum (wheat)	855,066
Ciona intestinalis	686,396
Sus scrofa (pig)	623,919
Arabidopsis thaliana (thale cress)	622,972
Gallus gallus (chicken)	599,140
Xenopus laevis (African clawed frog)	537,424
Drosophila melanogaster (fruit fly)	498,214
Hordeum vulgare + subsp. vulgare (barley)	437,321
Canis familiaris (dog)	365,946
Glycine max (soybean)	359,148
Caenorhabditis elegans (nematode)	346,064

3.2.2 Les banques protéiques

Il existe une grande variété de banques protéiques disponibles qui diffèrent selon leur **exhaustivité**, leur **degré de redondance** et la **qualité de leurs annotations** [Apweiler et coll., 2004 ; Nesvizhskii et coll., 2005]. Le degré d'intervention de l'homme pour la vérification des séquences, les corrections et les ajouts d'annotations est un critère important de distinction des différentes banques protéiques accessibles.

On peut ainsi distinguer deux catégories de banques protéiques :

- Les grandes **banques généralistes de dépôt** (« sequence repositories ») qui contiennent essentiellement des séquences protéiques issues de la traduction de gènes prédits *in silico* par des algorithmes de prédiction sans vérification et avec des annotations fonctionnelles limitées voire inexistantes. Parmi ces banques, on peut citer la banque **GenPept** (GenBank Gene Products Data Bank) produite par le NCBI regroupant l'ensemble des protéines traduites à partir des banques nucléotidiques GenBank/EMBL/DDBJ, sans annotation et avec une grande redondance. Ces banques sont très peu utilisées en analyse protéomique.
- Les **banques de séquences « corrigées »** pour lesquelles un travail de nettoyage, de tri, de documentation et d'annotation a été réalisé par des experts. Ces banques peuvent être différenciées selon les moyens humains investis pour augmenter ce que l'on appelle la « valeur ajoutée » de la banque protéique. Les **banques protéiques spécialisées** focalisées sur une famille de protéines, un groupe de protéines ou consacrées à un organisme font partie de ces banques. La liste de ces banques est longue et seules les banques les plus couramment utilisées en analyse protéomique sont brièvement détaillées ici.

➢ La banque Entrez Protein du National Center for Biotechnology information, NCBInr

Cette banque généraliste se trouve à la jonction des deux types de banques car elle contient à la fois les séquences issues de la traduction automatique des banques GenBank/EMBL/DDBJ mais aussi des séquences annotées de Swiss-Prot, de la PIR (Protein Information Resource) [Cathy et coll., 2003], de RefSeq [Pruitt et coll., 2005] et de la PDB (Protein Data Bank) [Westbrook et coll., 2003]. Cette banque est donc volumineuse et redondante mais par contre elle présente l'avantage d'être relativement complète. Elle est accessible sur le site du NCBI : http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi.

➢ Le consortium UniProt

En 2003, le Swiss Institute of Bioinformatics (SIB), l'European Bioinformatics Institue (EBI) et le groupe de la Protein Information Resource (PIR) ont décidé d'unir leurs forces et de former le **consortium UniProt** centralisant les banques SwissProt, TrEMBL et PIR [Apweiler et coll., 2004 ; Bairoch et coll., 2005 ; Wu et coll., 2006].

La banque UniProt contient trois composantes : la banque « UniProt Knowledgebase » (**UniProtKB**), la banque « UniProt Archive » (**UniParc**) et la banque « UniProt non-redundant reference database » (**UniProt NREF**). Les banques Swiss-Prot, TrEMBL et PIR-PSD sont rassemblées dans UniProtKB. Ces banques sont accessibles sur le site http://www.expasy.uniprot.org/.

- **La banque PIR-PSD (Protein Information Resource Protein Sequence Database)**

Il s'agit de la plus ancienne des banques « corrigées » établie dès 1984 succédant à l'«*Atlas of Protein Sequence and Structure* » qui avait été maintenue pendant 20 ans auparavant par M. O. Dayhoff. La caractéristique de cette banque est qu'elle est organisée selon une classification par familles et super-familles de protéines et les annotations sont également basées sur cette organisation. Cette banque contient des annotations fonctionnelles, structurales, bibliographiques (liens vers PubMed, …) et génétiques [Cathy et coll., 2003] (http://pir.georgetown.edu/).

- **La banque Swiss-Prot, aussi appelée UniProtKB/Swiss-Prot depuis la création d'UniProt**

Créée en 1986 par Amos Bairoch, la banque Swiss-Prot est la banque corrigée de « haute valeur ajoutée » de référence (http://www.ebi.ac.uk/swissprot/). C'est la banque de séquences protéiques la moins redondante, la mieux annotée et la plus intégrée par des références croisées vers d'autres banques. Le processus d'intégration d'une nouvelle séquence dans Swiss-Prot comprend plusieurs étapes visant à contrôler la pertinence de son entrée dans la banque. Ces étapes sont réalisées par un groupe d'experts en annotations. Un soin particulier est apporté pour indiquer la nature, expérimentale ou bioinformatique, des informations fonctionnelles, ainsi que le niveau de confiance qui est accordé à ces informations [Schneider et coll., 2004]. Au 4 septembre 2006, la banque SwissProt comptait 231434 entrées.

Jusqu'au milieu des années 90, le groupe d'Amos Bairoch, en collaboration avec l'équipe produisant la banque EMBL, parvenait à suivre le rythme d'accroissement d'EMBL, appliquant l'expertise humaine à toutes les nouvelles séquences entrées. Cependant, avec l'accroissement rapide des tailles des banques nucléotidiques ces 10 dernières années, les moyens humains qui auraient permis de suivre cette croissance rapide, tout en remplissant les critères initiaux de qualité de Swiss-Prot, n'ont pu être obtenus. Plutôt que

d'opter pour une baisse en qualité et une hétérogénéité des données, une banque complémentaire à Swiss-Prot a été créée : la banque TrEMBL (Translation from EMBL).

- **La banque TrEMBL (Translation from EMBL) ou UniProtKB/TrEMBL**

La banque TrEMBL, construite à l'EBI, contient les séquences annotées *in silico* d'après la traduction de l'ensemble des séquences codantes de GenBank/EMBL/DDBJ hormis celles déjà présentes dans Swiss-Prot [Boeckmann et coll., 2003] (http://www.ebi.ac.uk/trembl/). Des programmes informatiques de pré-annotation des entrées de TrEMBL avant leur soumission à Swiss-Prot permettent d'accélérer la procédure d'annotation : (i)Les séquences redondantes sont repérées informatiquement et sont assemblées. (ii)Les séquences non annotées de TrEMBL sont comparées aux séquences déjà présentes dans Swiss-Prot grâce à des programmes de recherches de domaines conservés (InterPro [Mulder et coll., 2003]) et sont pré-classées par familles. Une fois pré-annotées, les entrées TrEMBL deviennent des entrées Sp-TrEMBL, prêtes à être expertisées par des humains qui jugeront de la pertinence biologique des pré-annotations et qui, selon le cas, les modifieront, les complèteront ou les invalideront. Au 4 septembre 2006, la banque TrEMBL contenait 3182016 entrées.

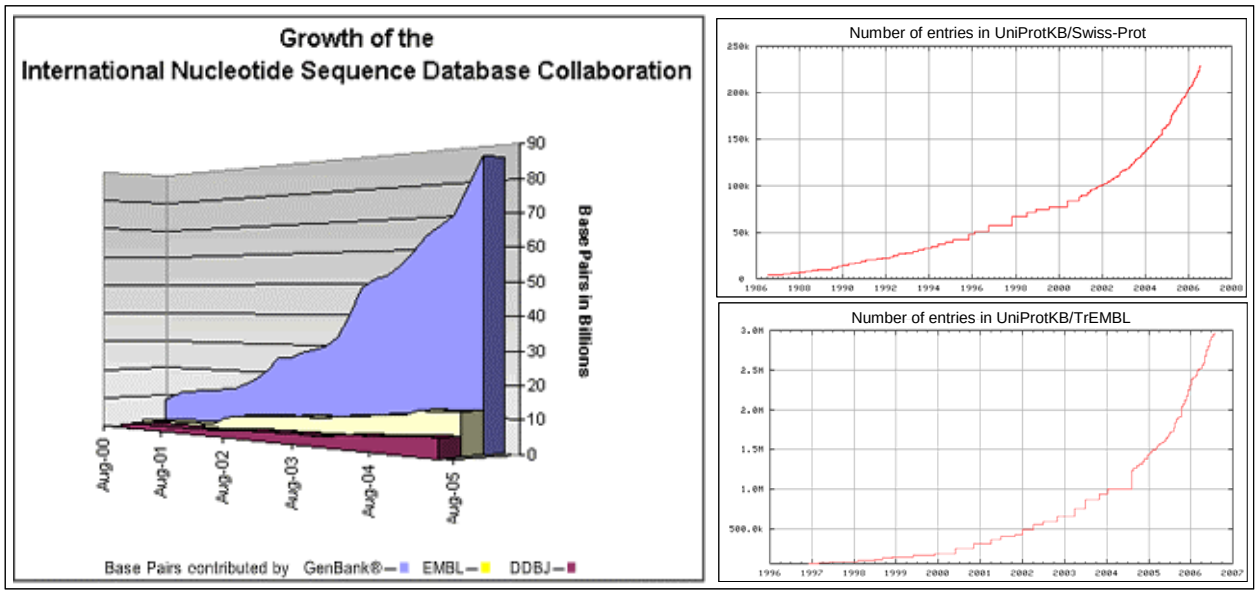
3.3 Conclusions

En 2006, les banques de séquences nucléotidiques GenBank/EMBL/DDBL contiennent des séquences d'ADN de plus de 205000 organismes différents et ne comptent pas moins de 51 milliards de nucléotides. La part des génomes complets est croissante : plus de 250 génomes bactériens complets sont représentés et près de 90 génomes eucaryotes pour lesquels une couverture et des assemblages significatifs sont disponibles [Benson et coll., 2006]. Aucun ralentissement de la croissance exponentielle des banques de données nucléiques et protéiques n'est préconisé pour ces prochaines années [Figure 22]. Outre les ressources informatiques nécessaires à la gestion efficace de tels volumes de données, de nombreuses questions sont soulevées quant aux développements futurs des banques de données. Homogénéité et non-redondance sont impossibles à concilier avec exhaustivité, le maintien des banques généralistes à côté desquelles se développent de nombreuses banques spécialisées semble être la solution adoptée. Le nombre croissant de banques spécialisées décrites dans la littérature souligne bien cette tendance [Database Issue, 2006].
Cependant, dans ce contexte plusieurs points méritent réflexion :

- Le **niveau de qualité** des banques accessibles : les données sont sous la seule responsabilité des auteurs des séquences et des annotations ce qui résulte en des données de qualité non homogène et non contrôlée.
- Le **maintien des banques** nécessite des **ressources financières** à long terme ce qui pose souvent des difficultés. En août 2006 dans une interview publiée dans *Protéines à la Une* (une publication électronique de vulgarisation scientifique du groupe Swiss-Prot de l'institut Suisse de Bioinformatique hébergé sur ExPASy, http://www.expasy.org/), Amos Bairoch décrit les moyens qu'il a du mettre en œuvre pour sauver SwissProt qui avait failli disparaître en 1996 faute de financement.

- Enfin, la **communication et l'échange d'informations** entre les différentes banques créées nécessitent l'élaboration et la conservation de langages communs qui permettront de maintenir l'interopérabilité entre l'ensemble des banques.

Figure 22 : Croissance exponentielle des tailles des banques nucléotidiques (avec la contribution des trois centres GenBank, EMBL et DDBK) et des banques protéiques UniProtKB/Swiss-Prot et UniProtKB/TrEMBL.
Sources : http://www.ncbi.nlm.nih.gov/Genbank/, http://us.expasy.org/sprot/relnotes/relstat.html et http://www.ebi.ac.uk/swissprot/sptr_stats/index.html.



CHAPITRE II

Les stratégies d'identification en analyse protéomique et les nouveaux défis

Les principales stratégies d'identification des protéines utilisant les outils décrits dans le chapitre précédent sont présentées dans cette partie. Dans un second temps les nouveaux défis confiés à l'analyse protéomique par spectrométrie de masse sont brièvement présentés.

1 Les stratégies d'identification des protéines en analyse protéomique

1.1 Historique

Les banques de données de séquences accessibles et les développements instrumentaux de la spectrométrie de masse des 20 dernières années ont donné naissance à toute une série de stratégies d'identification des protéines.

C'est ainsi qu'en 1993 un concept a été décrit simultanément par 5 groupes de chercheurs [Henzel et coll., 1993 ; James et coll., 1993 ; Mann et coll., 1993 ; Pappin et coll., 1993 ; Yates et coll., 1993] : la méthode de l'**empreinte peptidique massique** (« peptide mass fingerprinting », PMF [Pappin et coll., 1993]). Cette méthode consiste en l'identification des protéines par la mesure des masses des peptides générés suite à une digestion enzymatique, masses qui sont ensuite comparées à celles des peptides théoriques calculées à partir de la digestion *in silico* des protéines stockées dans les banques de séquences (cf 1.2.1).

Un an après, Eng, McCormack et Yates ont appliqué une technique similaire pour corréler des **spectres MS/MS** avec des peptides théoriques [Eng et coll., 1994]. Tandis que Mann et Wilm ont proposé, la

même année, de déduire des séquences partielles en acides aminés à partir des différences de masses observées entre les fragments sur les spectres MS/MS puis d'utiliser ces informations pour lancer des requêtes « tolérantes aux erreurs » dans les banques de données [Mann et coll., 1994]. A partir là, de nombreux algorithmes de recherche à partir de données de spectrométrie de masse en tandem ont été développés, chacun présentant ses particularités et participant à l'amélioration des performances de l'identification des protéines par des données de MS/MS (cf 1.3).

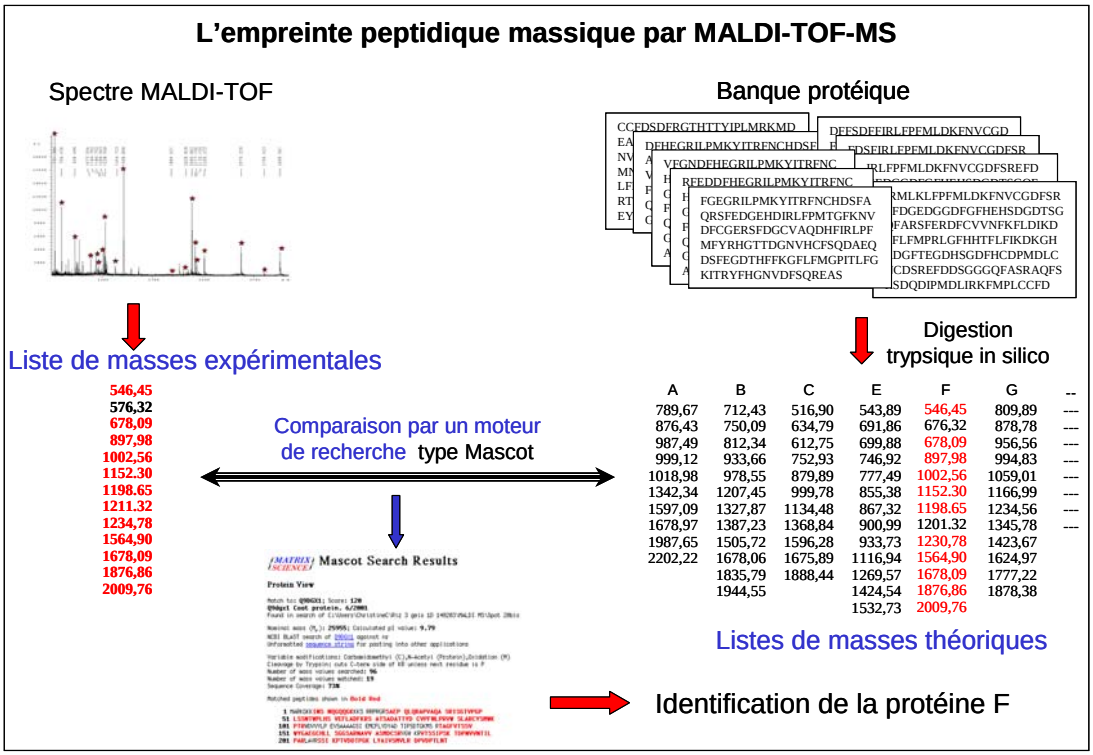
1.2 L'empreinte peptidique massique (PMF)

1.2.1 Les identifications par PMF

Cette stratégie consiste en la comparaison de la liste de masses expérimentales mesurées à l'ensemble des listes de masses théoriques calculées *in silico* pour les protéines présentes dans les banques données. Le principe des recherches par PMF est illustré en figure 1.

De nombreux programmes de recherche par PMF ont été développés ces 10 dernières années tels que : MOWSE [Pappin et coll., 1993], Mascot [Perkins et coll., 1999], MS-Fit [Clauser et coll., 1999], PeptIdent [Wilkins et coll., 1999], ProFound [Zhang et coll., 2000] ou encore plus récemment Aldente [Tuloup et coll., 2003].

Figure 1 : Illustration de la stratégie d'identification des protéines par empreinte peptidique massique.



Classiquement, les identifications de protéines par PMF sont réalisées sur des extraits de peptides de spots de gels 2D analysés par MALDI-TOF-MS. Les analyses par MALDI-TOF-MS sont adaptées pour le PMF car elles sont rapides et facilement automatisables, tolérantes aux sels et aux détergents, très sensibles

et présentent une bonne résolution (plus de 10000 en général) et une bonne précision de masse (10-20 ppm) [Horn et coll., 2004]. Néanmoins, des identifications par PMF peuvent aussi être réalisées à partir de données de LC-MS obtenues sur des instruments permettant de mesurer des masses avec une bonne précision (dernières générations d'ESI-TOF ou FT-ICR).

Les critères permettant de valider une identification par PMF sont :

- le **pourcentage de recouvrement de séquence** obtenu ou plus précisément le **nombre de peptides identifiés** car le pourcentage de recouvrement dépend de la taille de la protéine et peut donc être élevé même si le nombre de peptides identifiés est faible dans le cas de petites protéines.
- les **erreurs** observées sur les mesures de masse
- mais aussi le nombre de peptides présents sur le spectre n'appartenant pas à la protéine identifiée
- la **comparaison entre le pI et le poids moléculaire théorique** de la protéine identifiée à sa position sur le gel lorsqu'il s'agit d'analyses de spots de gel 2D.

Les critères d'acceptation des identifications par PMF sont inclus dans les directives de publication de bonnes données de protéomique décrites en 1.5 [Carr et coll., 2004 ; Bradshaw, 2005 ; Wilkins et coll., 2006].

1.2.2 Les limites du PMF

Malgré ses avantages qui en font une technique de choix pour l'analyse protéomique à haut débit des protéines contenues dans des spots de gels 2D, cette stratégie d'identification présente de nombreuses limites :

- Les échantillons analysés doivent être relativement « simples » car **les mélanges complexes** vont donner lieu à des superpositions d'empreintes peptidiques difficiles à interpréter et qui vont rendre les identifications rapidement impossibles [Clauser et coll., 1999].

- La digestion de **protéines de faible poids moléculaire** ou faiblement abondantes va produire un petit nombre de peptides pouvant être identifiés ce qui influera sur la qualité et la confiance accordée à l'identification de ces protéines.

- La **précision des mesures de masse** doit être très élevée pour limiter les identifications de faux positifs [Clauser et coll., 1999].

- Avec la croissance exponentielle de la **taille des banques de données**, la **spécificité du PMF** est de plus en plus **limitée** et l'identification de faux positifs de plus en plus probable [Gattiker et coll., 2002]. En effet, plusieurs compositions en acides aminés peuvent avoir des masses égales et si on ne dispose que de la masse du peptide, rien ne permettra de valider l'une ou l'autre des séquences. En augmentant le nombre de séquences dans les banques, le nombre de masses théoriques augmente et donc également le nombre de masses redondantes.

Une solution à ces limites est l'utilisation d'instruments à très haute résolution du type MALDI-FT-ICR MS générant des mesures de masse très précises (≈1-2ppm) et permettant ainsi d'améliorer considérablement la qualité et la spécificité des identifications par PMF [Horn et coll., 2004].

Néanmoins, les instruments de type FT-ICR n'étant pas très répandus dans les laboratoires de protéomique, dans bien des cas, le recours à la spécificité apportée par les informations de séquence obtenues par des analyses MS/MS est nécessaire [Steen et coll., 2004].

1.3 Les approches par nanoLC-MS/MS

Hunt et coll. ont été les premiers, en 1992, à utiliser des approches LC-MS/MS pour l'analyse de mélanges complexes de peptides [Hunt et coll., 1992]. L'identification des protéines par des analyses en couplage nanoLC-MS/MS est aujourd'hui au cœur de l'analyse protéomique par spectrométrie de masse [Lane, 2005].

Les approches en couplage nanoLC-MS/MS présentent deux avantages majeurs par rapport aux stratégies par PMF :

- L'analyse de mélanges beaucoup plus complexes est possible : l'étape de chromatographie de phase inverse permet la séparation des peptides avant leur entrée dans le spectromètre de masse. Les peptides élués par le gradient d'acetonitrile entrent tour à tour dans le spectromètre de masse par ordre croissant d'hydrophobicité et même si la séparation n'est pas totale, la compétition au moment de l'ionisation sera considérablement réduite par rapport à une injection simultanée de l'ensemble des peptides.

- Les informations de séquence apportées par la MS/MS permettent d'augmenter considérablement la spécificité et la qualité des identifications.

Lors d'analyses nanoLC-MS/MS, la sélection des ions parents pour la fragmentation peut se faire manuellement mais se fait dans la plupart des cas automatiquement : la tâche de sélection des ions est alors dynamiquement contrôlée par le logiciel pilotant l'instrumental [Aebersold et Goodlett, 2001]. Néanmoins, des paramètres de sélection des ions précurseurs sont définis par l'utilisateur (cf Partie I, Chapitre I, 2.3.3 et 2.3.4 des résultats). Les règles de la fragmentation peptidique sont décrites dans le chapitre I, 1.3.

Les fichiers générés durant ces analyses sont très volumineux par rapport à des analyses en PMF : de 100 à 300 ions sont généralement sélectionnés et fragmentés en MS/MS et les fichiers sauvegardés font de plusieurs dizaines à quelques centaines de Mo en taille. La conservation de ces données nécessite donc une infrastructure informatique de stockage des données adaptée.

De la même manière que pour le PMF, des algorithmes de recherche automatique en banques de données avec les données de MS/MS ont été développés. L'ensemble des stratégies de recherche disponibles actuellement pour l'identification automatique des protéines avec des données de spectrométrie de masse en tandem a été compilé par récemment par Hernandez et coll. [Hernandez et coll., 2006].

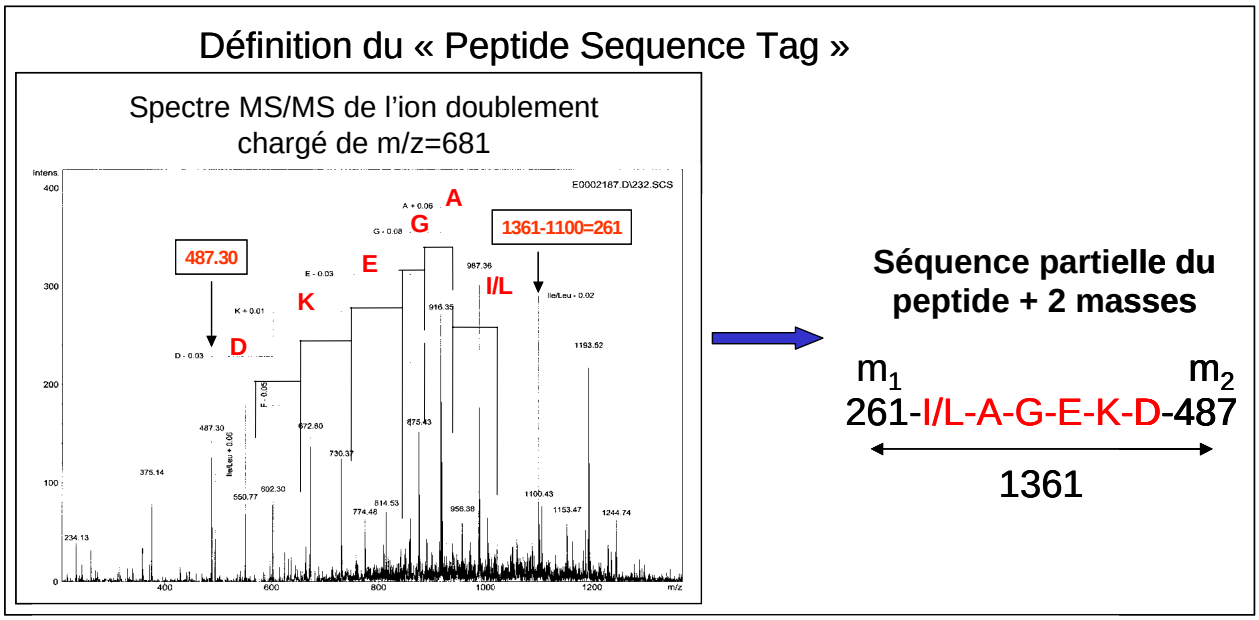
1.3.1 Les moteurs de recherche en banques de données

- <u>L'approche par « Peptide Sequence Tag»</u>

Cette approche décrite par Mann et Wilm en 1994 [Mann et coll., 1994] repose sur le principe suivant :

Les spectres MS/MS contiennent toujours au moins une courte série de fragments permettant la lecture non ambiguë d'un fragment de séquence du peptide (« tag »). En plus de ce tag, cette approche utilise également la masse de l'ion parent ainsi que les masses m_1 et m_2 pouvant être déterminées sur le spectre. m_1 et m_2 définissent les distances entre le tag de séquence déterminé et les extrémités du peptide. Le « peptide sequence tag » est donc constitué de 3 parties, le fragment de séquence et les masses m_1 et m_2 et sera soumis au moteur de recherche PeptideSearch pour l'identification de la protéine (Figure 2).

Figure 2 : Détermination du « Peptide Sequence Tag » [Mann et coll., 1994].



En 1994, Mann et Wilm en étaient arrivés à la conclusion que 2 ou 3 acides aminés constituant le Tag sont nécessaires pour identifier un peptide trypsique unique dans une banque de données telle que Swiss-Prot [Mann et coll., 1994]. Or, depuis 1994, le nombre d'entrées dans Swiss-Prot est passé de moins de 40000 [http://us.expasy.org/sprot/relnotes/relstat.html] à 230133 au 24 juillet 2006 [http://srs6.ebi.ac.uk/]. Au vu de ces chiffres, les exigences pour pouvoir considérer une identification non ambiguë ont évolué ces 10 dernières années (cf 1.3.2).

Par ailleurs, cette approche requiert non seulement un temps d'interprétation non négligeable et la spécificité d'un seul tag n'est souvent plus suffisante.

D'autres programmes, plus rapides et reposant sur des algorithmes différents, ont donc été développés.

- Les approches par Peptide Fragment Fingerprinting (PFF)

La dénomination **Peptide Fragment Fingerprinting** a été introduite par Blueggel et coll. en 2004 par analogie avec la dénomination PMF [Blueggel et coll., 2004]. Toute une série d'algorithmes peuvent être regroupés sous ces approches par PFF [Hernandez et coll., 2006] :

-Mascot (http://www.matrixscience.com/),

- Sequest (http://fields.scripps.edu/sequest/),

- Phenyx (http://www.phenyx-ms.com/),

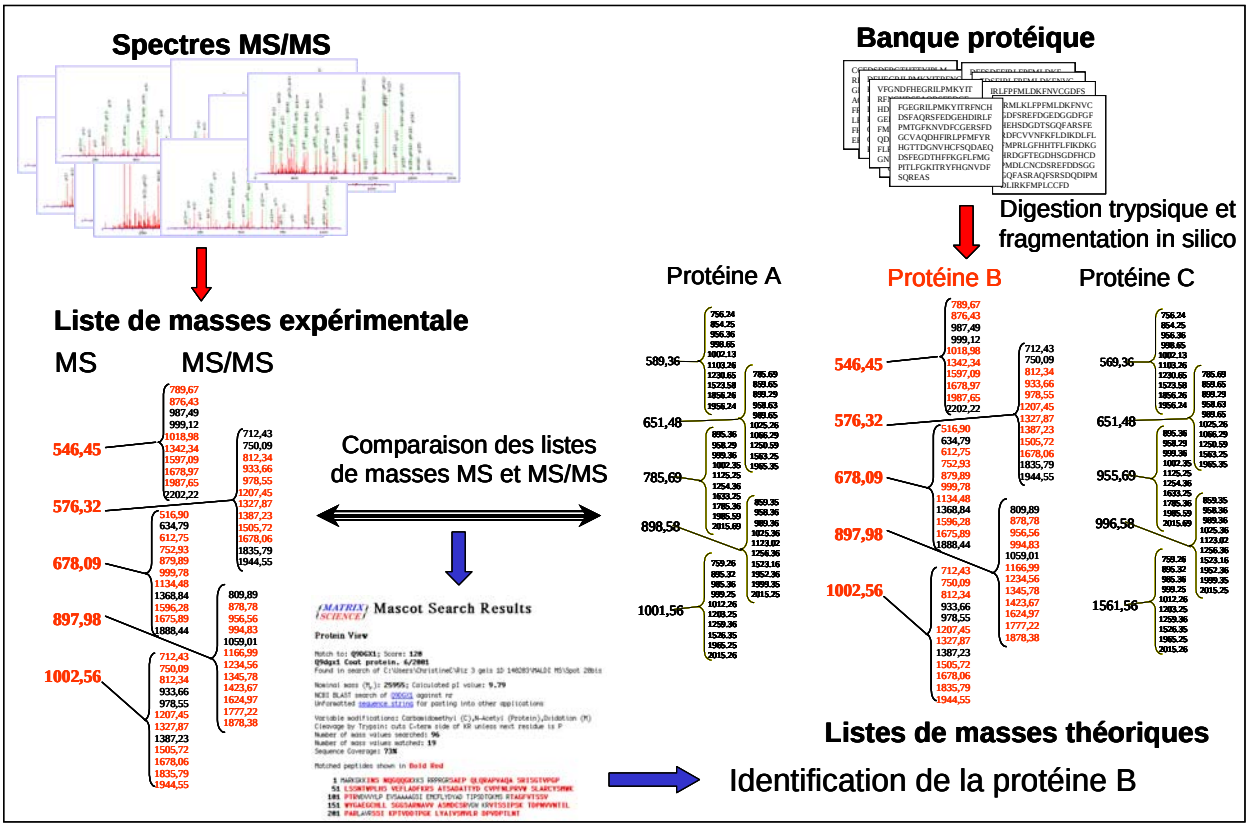
- Sonar (http://bioinformatics.genomicsolutions.com/service/prowl/sonar.html),
- PepFrag (http://prowl.rockefeller.edu/prowl/pepfragch.html),
- MS-Tag (http://prospector.ucsf.edu/ucsfhtml4.0/mstagfd.htm),
- probID (http://projects.systemsbiology.net/probid/)

Ces approches sont basées sur des comparaisons entre les spectres MS/MS expérimentaux et les spectres MS/MS générés à partir de l'ensemble des peptides trypsiques théoriques des protéines contenues dans les banques de données.

L'algorithme de **Mascot** est basé sur une **approche probabilistique** (« Probability-based matching) [Perkins et coll., 1999]. Les masses des fragments MS/MS calculées à partir des peptides présents dans la banque de données sont comparées aux masses expérimentales des fragments mesurées. Un score est calculé qui reflète la significativité de la correspondance entre les spectres théoriques et expérimentaux (Figure 3).

L'algorithme de **Sequest** est basé sur un **processus d'auto-correlation** [Eng et coll., 1994] : les peptides présents dans les banques de données sont utilisés pour construire des spectres MS/MS théoriques et le degré d'auto-correlation entre le spectre expérimental et un spectre théorique permet de déterminer le meilleur « match ».

Figure 3 : L'algorithme de Mascot est basé sur la comparaison de la liste de masses expérimentales obtenue avec les listes de masses des fragments théoriques calculées à partir des protéines présentes dans la banque de données.



1.3.2 Les limites des algorithmes de recherche dans les banques de données et les outils de validation

Les algorithmes sont de plus en plus rapides et performants et permettent de lancer des requêtes dans des banques très volumineuses : aussi bien des larges banques de séquences protéiques que des banques d'ESTs ou des banques génomiques (dont ils assurent la traduction automatique dans les 6 cadres de lecture) [Kuster et coll., 2001 ; Choudhary et coll., 2001]. Cependant, ces algorithmes ne sont pas infaillibles et les développements de méthodes de validation des identifications ont été intensifs ces 5 dernières années [Keller et coll., 2002 ; Nesvizhskii et coll., 2003 ; Nesvizhskii et coll., 2004]. **Des banques de données aléatoires** peuvent ainsi être utilisées pour évaluer les taux de faux positifs [Peng et coll., 2003 ; Elias et coll., 2005].

Peng et coll. ont évalué le taux de faux positifs en fonction du nombre de peptides ayant permis l'identification d'une protéine et de la spécificité de l'enzyme utilisée lors de la recherche dans les banques : ils ont ainsi estimé que le **taux de faux positifs** passait de 0.8% à près de 10% lorsque l'on tolère l'identification de peptides semi-trypsiques (ne présentant qu'une extrémité trypsique). De la même manière, ils ont estimé qu'un taux de faux positifs de 26% pour des protéines identifiées avec un unique peptide est réduit à 1,4% pour des identifications avec 2 peptides et devient nul pour des identifications avec 3 ou plus de peptides. Les identifications de protéines avec un seul peptide ne doivent être limitées qu'à des cas exceptionnels [Steen et coll. 2004 ; Carr et coll., 2004].

En 1994, un Peptide Sequence Tag de 2-3 acides aminés était suffisant pour identifier un peptide unique dans SwissProt mais en **2004** (cf 1.3.1), on estime qu'une séquence de moins de **7 acides aminés** n'est plus suffisante pour identifier un peptide unique [Steen et coll., 2004]. Les exigences vont devenir de plus en plus importantes avec la croissance exponentielle des banques.

Dans cette optique de validation des résultats et d'estimation des taux de faux positifs, un programme intégré dans Mascot permet de générer, à partir des banques classiquement utilisées, des banques de séquences inversées ou « aléatoires » dites **banques Decoy** (dans les banques inversées, les séquences des entrées sont lues du C-ter vers le N-ter et dans les banques « aléatoires », les séquences sont mélangées mais le nombre d'entrées est conservé). Le taux de faux positifs peut alors être facilement estimé grâce au nombre d'identifications significatives mises en évidence lors des recherches dans ces banques « dégénérées » [Elias et coll., 2005].

Enfin, toutes ces stratégies sont **« non tolérantes aux erreurs »** (les algorithmes se basent sur des identités strictes entre données expérimentales et données théoriques) et ne sont donc adaptées que pour l'identification de **protéines présentes dans les banques de données**. Une stratégie alternative utilisant le séquençage *de novo* des peptides fragmentés en MS/MS est utilisée pour l'interprétation des données provenant d'organismes non séquencés.

1.3.3 Les approches par séquençage *de novo*

Plusieurs projets réalisés durant ce travail de thèse ont porté sur des organismes non séquencés et ont nécessité le recours à des stratégies utilisant le séquençage *de novo*. La partie III des résultats est consacrée à

la mise en place de cette stratégie et aux études protéomiques qui l'ont requise : une description détaillée du principe et des outils de cette approche y est présentée.

1.4 L'importance du choix de la banque de données utilisée pour les recherches

Le bon choix de la banque de données dans laquelle sont lancées les requêtes est primordial car dans la plupart des cas, seules les protéines dont la séquence est présente dans la banque peuvent être identifiées (cf Partie IV, Chapitre I des résultats). Malgré des données de MS de qualité, un mauvais choix de banque de données peut donc conduire à une absence d'identification.

Toutes les approches décrites sont basées sur des **mesures de masse de peptides**. Les recherches impliquent donc en premier lieu des **identifications de peptides.** Ce n'est ensuite que dans un second temps que les algorithmes assemblent les peptides pour l'identification de protéines. De ce fait, tous les types de banques de séquences peuvent théoriquement être utilisés, des banques de séquences protéiques mais aussi des banques d'ESTs et des banques génomiques [Kuster et coll., 2001 ; Choudhary et coll., 2001]. Néanmoins, les analyses protéomiques utilisent majoritairement **les banques protéiques**.

Pour certains organismes modèles, les banques protéiques sont le fruit d'un énorme travail de vérification et d'annotation par des experts et elles sont dans ces cas fiables et donnent une bonne image du protéome de l'organisme : c'est le cas pour la banque SwissProt qui contient les protéomes bien annotés pour une petite série d'organismes modèles.

Par contre, lorsqu'il s'agit d'identifier des protéines d'organismes autres que la petite poignée d'organismes modèles, la qualité des protéomes présents dans la banque est très variable selon les organismes. Il est donc important de consulter les informations relatives aux différentes banques de données avant de les utiliser pour en connaître notamment le nombre d'entrées concernant l'organisme étudié, la façon dont ces entrées ont été générées (par simple prédiction automatique ou si une annotation manuelle a été réalisée) et quel est l'état d'avancement du séquençage du génome, …

Le choix de la banque est aussi fonction de l'objectif final de l'expérience [Nesvizhskii et coll., 2005]. Par exemple :

- Dans les cas où l'identification de polymorphismes de séquence est cruciale pour l'interprétation biologique des résultats, des banques de séquences comme NCBInr ou UniProtKB contenant un maximum de variants de séquences sont adaptées.
- Dans les cas où la qualité de l'annotation et la facilité de l'interprétation sont primordiales, il est plus avantageux d'utiliser des banques plus restreintes mais bien annotées telles que SwissProt.

Potentiellement l'utilisation de **banques d'ESTs** et de **banques génomiques** permet d'identifier de nouvelles protéines absentes des banques protéiques ou de nouveaux variants d'épissage de protéines connues [Nesvizhskii et coll., 2004 ; Choudhary et coll., 2001 ; Kuster et coll., 2001]. Lors de l'importation de séquences en acides nucléiques dans les moteurs de recherche, aucune information quant au cadre de lecture n'est contenue dans la séquence : les moteurs de recherche assurent donc la traduction des séquences nucléiques dans les six cadres de lecture possibles.

Néanmoins, les banques d'ESTs ne sont que rarement utilisées pour des recherches en protéomique car peu nombreux sont les organismes pour lesquelles des banques d'ESTs de qualité ont été générées. Pour que la banque d'ESTs soit significative pour un organisme donné, il faut qu'un très grand nombre d'ESTs ait été séquencé (cf Chapitre I, 3.2.1). Par ailleurs, lorsque le nombre d'ESTs est significatif, les recherches dans les banques correspondantes prennent un temps assez long (la banque dbEST du NCBI contient plus de 7,8 millions d'ESTs humains). Enfin, les collections d'ESTs sont souvent « contaminées » par des séquences courtes et de faible qualité.

Une partie de ce travail de thèse a été consacrée au développement d'une stratégie permettant les recherches avec des données de spectrométrie de masse en tandem dans des **données génomiques non interprétées**. La mise au point de cette approche, ses avantages et les applications pour lesquelles elle a permis d'apporter des réponses sont présentées dans la quatrième partie des résultats de ce manuscrit.

1.5 Les directives de publication des données de protéomique

Ces dernières années, un nombre important de travaux utilisant l'analyse protéomique par spectrométrie de masse pour l'identification de protéines a été publié. Cependant, dans ces publications, les critères de confiance utilisés pour la validation des identifications est très variable. Ceci est particulièrement remarquable avec l'avènement des stratégies de « shotgun » protéomique générant des données volumineuses dans lesquelles la relation entre les peptides et la protéine précurseur est perdue et impliquant souvent des identifications avec un petit nombre de peptides.

C'est ainsi qu'en 2004, à l'initiative du journal ***Molecular & Cellular Proteomics***, un groupe de travail s'est réuni afin d'établir des directives pour la publication de « données de protéomique de qualité » [Carr et coll., 2004]. Les outils de validation des données de spectrométrie de masse en tandem décrits en 1.3.2 s'inscrivent dans ce même désir de garantir la qualité des données protéomiques publiées [Nesvizhskii et coll., 2003 ; Nesvizhskii et coll., 2004 ; Elias et coll., 2005].

Une nouvelle version des directives établies par un groupe de travail réuni à la Maison de la Chimie à Paris en mai 2005 est accessible sur le site web du journal *Molecular & Cellular Proteomics* [Bradshaw, 2005]. L'objectif de ces directives est de formuler des critères standards pour que les éditeurs, les reviewers et les lecteurs puissent évaluer plus facilement les informations fournies assurant l'intégrité et l'exactitude des données protéomiques publiées. De la même manière, des directives ont été publiées dans le journal ***Proteomics*** en janvier 2006 [Wilkins et coll., 2006].

Ces directives tentent d'harmoniser la qualité et les formats des données de protéomique et donnent des instructions précises sur les informations requises pour la publication de ces données, aussi bien les informations relatives à l'instrumentation utilisée, aux critères de traitement des données, de recherche dans les banques et de validation des identifications. Les identifications réalisées durant ce travail de thèse ont suivi ces directives et la publication des résultats a été formatée suivant les instructions qu'elles définissent.

1.6 Conclusions

L'utilisation combinée de différents algorithmes de recherche et de programmes de validation suivie de la confrontation de leurs résultats permet d'accroître la qualité des identifications et la confiance qui leur est accordée. Néanmoins la réalisation systématique de recherches multiples et la comparaison des résultats de ces recherches (surtout si la comparaison est réalisée manuellement) restent longues et fastidieuses. Dans cette optique, des outils sont développés actuellement, comme la plate-forme *Proteinscape* (Bruker Daltonics) intégrant différents algorithmes de recherche et combinant leurs résultats en calculant des méta-scores [Chamrad et coll., 2003]. Ce sont très probablement ces types d'outils qui seront préconisés dans l'avenir : des approches cycliques faisant travailler successivement plusieurs algorithmes (PFF, séquençage *de novo*, …) en tirant profit des avantages de chacun (rapidité de recherche, tolérance aux erreurs, …) permettront sans aucun doute une amélioration de la qualité des interprétations.

2 Les nouveaux défis en protéomique

D'après Hernandez et coll., la définition du terme protéomique s'est vue élargie ces dernières années à une grande variété de technique d'analyses couvrant 3 axes principaux : **l'identification** mais aussi **la caractérisation** et **la quantification** de protéines [Hernandez et coll., 2006]. A côté des techniques d'identification des protéines précédemment décrites, de **nouveaux défis** ont donc été lancés à la protéomique. Ainsi, ces dernières années, des stratégies plus particulièrement destinées à la caractérisation fine, notamment la détermination de modifications post-traductionnelles, et à la quantification précise des protéines ont été développées. Ce sont ces dernières stratégies qui sont brièvement présentées dans ce dernier paragraphe.

2.1 Les stratégies de quantification

Dans une revue publiée dans le journal Nature en 2003, Aebersold et Mann soulignent que pour que la détermination des protéines exprimées dans une cellule ou dans un tissus donnés soit significative, les données doivent au minimum être semi-quantitatives : une simple liste de l'ensemble des protéines présentes dans un environnement donné n'est pas suffisante [Aebersold et coll., 2003]. Or, les spectromètres de masse sont en fait de « mauvais » instruments de quantification en l'absence de standards internes [Lane, 2005]. En effet, la relation entre la concentration en analyte dans l'échantillon de départ et l'intensité du signal mesurée avec le spectromètre de masse dépend d'une multitude de facteurs qui sont difficilement contrôlables : l'accessibilité de l'enzyme lors de la digestion, la solubilité des peptides lors de l'extraction et l'efficacité d'ionisation aussi bien en MALDI qu'en ESI [Steen et coll., 2004].

Des méthodes permettant une quantification des protéines plus précise qu'à partir de l'analyse d'images de gels 2D, non adaptés à la séparation des protéines hydrophobes et membranaires [Santoni et coll., 2000 ; cf Chapitre I, 2.1.2], ont été développées ces dernières années. Ces techniques sont principalement basées sur l'utilisation d'isotopes stables.

L'idée clé de ces stratégies est que deux formes d'une même molécule qui ne diffèrent que par la substitution d'isotopes stables vont se comporter de façon identique lors de l'analyse MS et ne seront

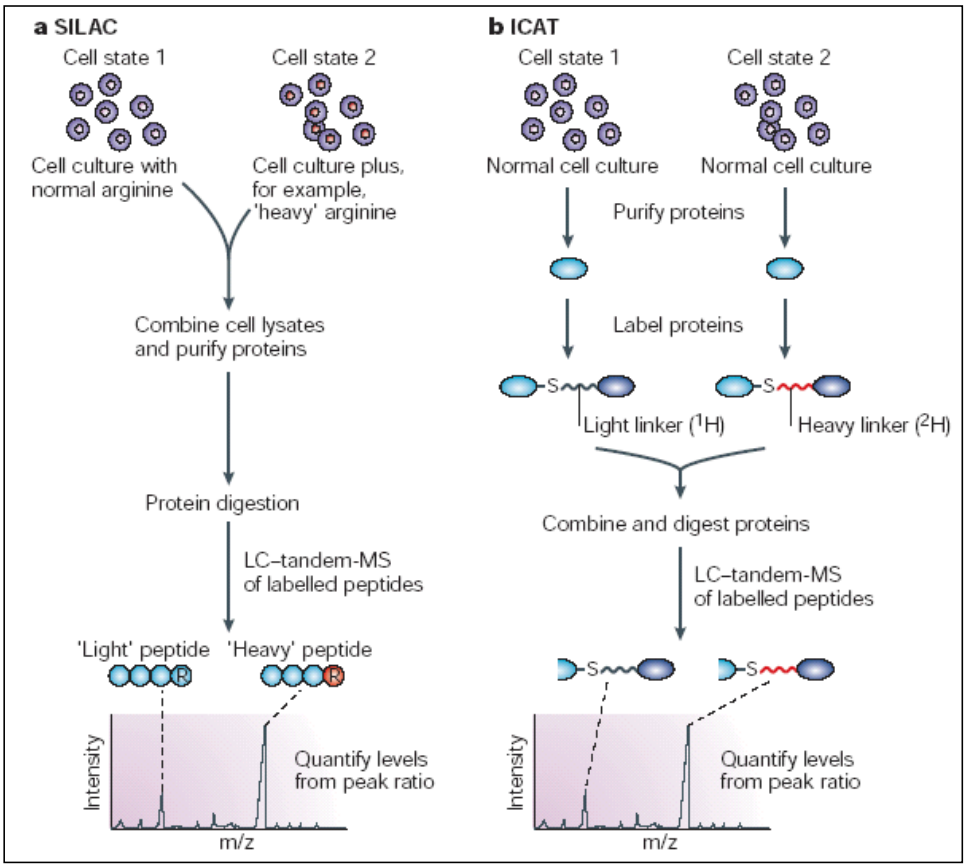
différenciées que par un écart de masse. Ainsi, le ratio entre ces deux pics indiquera précisément les quantités relatives des deux espèces. L'incorporation d'isotopes stables (principalement ^{2}H, ^{18}O, ^{13}C et ^{15}N) peut être réalisée de plusieurs manières [Lane, 2005]:

- par **réactions chimiques** : stratégie ICAT (Isotope-coded affinity tag) [Gygi et coll., 1999]
- par **marquage métabolique utilisant des acides aminés marqués** : stratégie SILAC (Stable isotope labelling with amino acids in cell culture) [Ong et coll., 2002]
- par **incorporation enzymatique d'O^{18}** durant la protéolyse [Mirgorodskaya et coll., 2000 ; Heller et coll., 2003].

En figure 4 sont illustrées les deux stratégies de quantification les plus couramment utilisées à l'heure actuelle. Brièvement, **l'approche SILAC** consiste en l'incorporation d'un acide aminé « lourd » dans le protéome d'une population cellulaire. La quantification relative est alors réalisée en comparaison avec des cellules ayant grandies dans un milieu de culture « normal ». Les lysats cellulaires sont mélangés, les protéines purifiées et digérées. Les peptides sont séparés par LC-MS et la coélution des peptides « lourds » et « légers » permet de réaliser la quantification relative de ces peptides.

Dans **l'approche ICAT**, les deux extraits cellulaires à comparer sont d'abord marqués avec un réactif, l'un avec un réactif « lourd » (^{2}H) et l'autre avec un réactif « léger » (^{1}H). Les extraits sont ensuite mélangés et les protéines digérées. Le marqueur possède à son extrémité un groupement biotine qui permet la capture sélective des peptides marqués suite à un passage sur colonne d'affinité. Seuls les peptides marqués seront donc analysés et quantifiés relativement.

Figure 4 : Illustration des stratégies de quantification par les méthodes SILAC et ICAT. D'après Steen et Mann [Steen et coll., 2004].



Récemment, la stratégie ICAT a été optimisée pour quantifier des protéines membranaires. Une étape de séparation des protéines par gel 1D après le marquage, suivie d'une digestion des protéines in-gel a été introduite [Ramus et coll., 2006].

En parallèle de ces techniques permettant une quantification relative des protéines, une technique permettant une quantification absolue a été décrite par Gerber et coll. [Gerber et coll., 2003]. Cette technique utilise des peptides synthétiques (ces peptides peuvent présenter des modifications post-traductionnelles ciblées), spécifiques de la protéine d'intérêt à quantifier, ayant incorporé des isotopes stables. L'échantillon protéique à quantifier est séparé par gel 1D puis les protéines sont digérées in-gel en présence des peptides synthétiques avant l'analyse des extraits peptidiques et la quantification absolue par LC-MS/MS.

2.2 La caractérisation des modifications post-traductionnelles (PTM)

Le second défi lancé à la protéomique réside dans la caractérisation complète des protéines. En effet, pour certaines applications, l'identification des protéines n'est plus suffisante et une connaissance de leurs modifications post-traductionnelles est nécessaire.

Les modifications post-traductionnelles sont des évènements chimiques qui altèrent les propriétés des protéines après leur traduction, soit par clivage protéolytique soit par addition de groupements sur un ou plusieurs acides aminés. Leur détermination est complexe mais génère des informations indispensables pour la compréhension des fonctions biologiques des protéines. Les PTMs incluent entre autres la phosphorylation, l'acétylation, la méthylation, la création de ponts disulfures, la déamidation ou encore l'ubiquitinylation [Mann et coll., 2003].

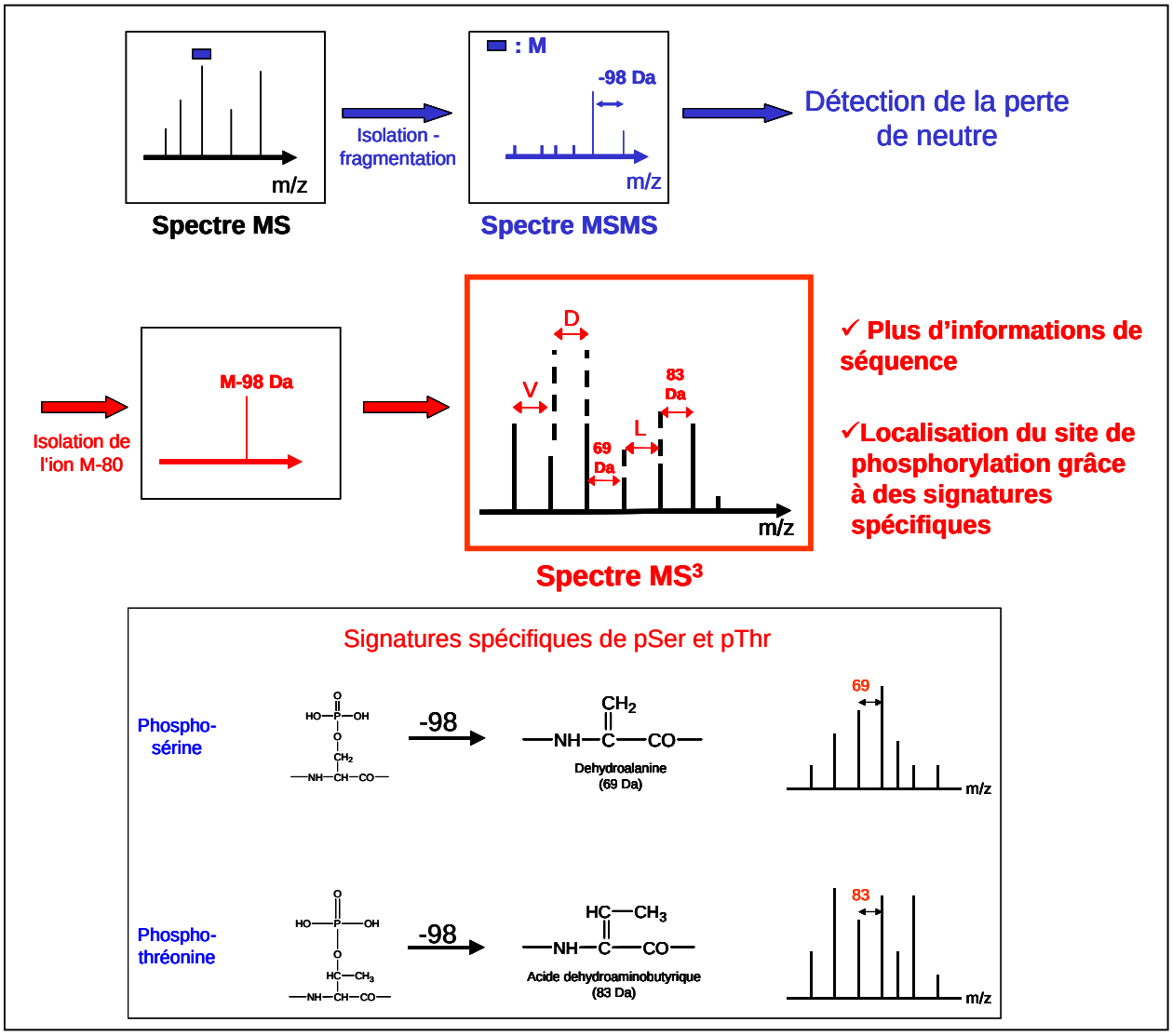
La **spectrométrie de masse** est l'**outil idéal pour la détermination de PTMs** car ces modifications entraînent des différences de masse « mesurables » au niveau des acides aminés, peptides ou protéines concernées.

De nombreuses stratégies permettant la détection, plus ou moins automatisée des PTMs par des techniques de spectrométrie de masse ont été développées ces dernières années. Jensen décrit l'ensemble des avancées et méthodologies disponibles à ce jour [Jensen, 2006]. Ces méthodologies nécessitent aussi bien des techniques d'enrichissement en protéines modifiées en amont de la spectrométrie de masse que des stratégies d'acquisition des données de MS ciblées sur les modifications ou encore des algorithmes de prédiction et de détection automatique des modifications.

En figure 5 est illustrée une stratégie permettant la détermination automatique de sites de phosphorylation en utilisant la MS^3 en couplage nanoLC-MS/MS sur la trappe HCT Ultra (Bruker Daltonics). L'acquisition en mode MS-MS/MS automatique est paramétrée de façon à ce qu'une perte de neutre (= perte de la PTM, c'est à dire une différence de masse paramétrée, comme par exemple –98 (H_3PO_4) ou –80 (HPO_3) pour les phosphorylations) soit détectée automatiquement sur le spectre MS/MS. La fragmentation de ces ions parents génère en général majoritairement l'ion ayant perdu la phosphorylation (coupures préférentielles en CID, cf chapitre I, 1.3). Dans un second temps, ces ions parents sur lesquels une perte de neutre est détectée sont sélectionnés pour une MS^3 (l'ion parent moins la perte de neutre est

sélectionné pour la MS^3) : c'est durant l'étape de MS^3 que l'ion va fragmenter et que les sites de phosphorylations pourront être détectés et spécifiquement localisés.

Figure 5 : Cette figure illustre la stratégie utilisée pour la détection de sites de phosphorylations en utilisant un mode d'acquisition MS^3 automatique avec un paramétrage de perte de neutre sur des instruments de type trappe ionique.



3 Conclusions

La **caractérisation** des protéines (c'est à dire obtenir des pourcentages de recouvrement de séquence de 100%) est devenue ces dernières années un maître-mot car dans bien des cas, la simple identification des protéines avec des pourcentages de recouvrement moyens n'est plus suffisante pour la compréhension totale de leur rôle biologique. Une seule PTM peut constituer la seule différence entre deux états. Les nouvelles techniques instrumentales et les algorithmes d'interprétation des données spécialisés sont nombreux. La tendance générale de l'analyse protéomique focalisée, il y a quelques années encore, sur l'amélioration des techniques d'analyses à haut débit pour l'identification du plus grand nombre de protéines s'oriente plutôt aujourd'hui vers le développement de techniques privilégiant la qualité plutôt que le nombre d'identifications.

CONCLUSION

En guise de conclusion, j'ai tenté de rassembler l'ensemble des développements et stratégies décrites dans cette partie bibliographique en un schéma. Les avantages et les inconvénients de chacune des approches mis en évidence lors de ce travail bibliographique y sont repris de façon synthétique.

Dans un premier volet, les points forts et points faibles des 3 principales stratégies de séparation des protéines et peptides, le gel d'électrophorèse 1D, le gel d'électrophorèse 2D et les chromatographies multiples sont décrits. Le second volet présente les deux approches instrumentales les plus couramment utilisées en analyse protéomique par MS : l'approche par MALDI-TOF-MS (ou LC-MS) pour générer des empreintes peptidiques massiques et les approches par nanoLC-MS/MS dont les deux principaux avantages sont l'étape de séparation des peptides en amont de l'analyse MS et les informations de fragmentation des peptides obtenues grâce aux instruments de spectrométrie de masse en tandem. Enfin, le troisième volet illustre les deux principales stratégies d'identification de l'analyse protéomique par MS, les identifications par empreinte peptidique massique et les identifications par empreinte de fragmentation peptidique.

Ce schéma sera repris dans la conclusion générale de ce manuscrit complété avec les apports et réflexions ce travail de thèse.

La stratégie de séparation des protéines

❑ Le gel d'électrophorèse 2D

-Image globale des protéines avec une relativement bonne résolution (2000 protéines séparées en routine)
-Quantification relative possible par comparaison d'images de gel correspondant à deux états
-Différentes colorations possibles avec des degrés de linéarité et de sensibilité variables.

-Problème des protéines hydrophobes
-Faible gamme dynamique
-Automatisation difficile
-Mauvais rendement de l'extraction des peptides après digestion *in-gel*

❑ Le gel d'électrophorèse 1D

-Pas de problème de sous-représentation des protéines membranaires
-Facile à mettre en œuvre
-Couplage avec la nanoLC

-Faible pouvoir résolutif, utilisée en pré-fractionnement uniquement
-Mauvais rendement de l'extraction des peptides après digestion *in-gel*
-Ne permet pas de quantification

❑ Les stratégies par chromatographies multiples

-Combinaisons de plusieurs types de chromatographie possible (phase inverse, SCX, échangeuse d'anions, gel filtration, …)
-Automatisation possible
-Analyse de mélanges très complexes, pouvoir résolutif élevé
-Compatible avec les approches quantitatives ICAT, SILAC, …

-Montage des chromatographies multiples en série délicats
-Problèmes de compatibilité entre éluants et analyse ESI
-Problèmes de reproductibilités et de validation des identifications pour les mélanges peptidiques très complexes

L'instrumentation utilisée

❑ Approche par MALDI-TOF-MS (ou LC-MS)

-Analyse rapide (1min pour l'acquisition d'un spectre)
-Bonne précision de mesure (<20ppm)
-Bonne sensibilité (femtomolaire)
-Bonne tolérance aux sels

-Problème de suppression de signal donc inadaptée pour les mélanges complexes
-Inadaptée pour les petites protéines car peu de peptides trypsiques
-Pas d'informations de séquence
-Pas de quantification

❑ Approches nanoLC-MS/MS

-Etape de séparation des peptides, permet donc l'analyse de mélanges plus complexes
-Détermination de séquence grâce à la MS/MS
-Spécificité et pouvoir discriminant apportés par la MS/MS
-Identification de modifications post-traductionnelles
-Quantification possible

-Analyse plus longue (environ 1h pour un gradient nanoLC-MS/MS)
-Gestion informatique de fichiers d'analyses très volumineux
-Couplage nanoLC délicat : problèmes fréquents de bouchages, surpressions, fuites, …
-Reproductibilité difficile dans le cas de mélanges très complexes

La stratégie d'identification

❑PMF (Empreinte peptidique massique)

-Identification rapide pour des protéines uniques ou des mélanges de faible complexité
-Adaptée pour l'identification de protéines présentes dans les banques de données

-Identifications difficiles lorsque plusieurs PMF sont superposées sur un même spectre
-Approche non tolérante aux erreurs
-Faible pouvoir discriminant

❑PFF (Empreinte de fragmentation peptidique)

-Spécificité et pouvoir discriminant élevés grâce à l'information de MS/MS
-Information de séquence
-Identification des modifications post-traductionnelles
-Adaptée pour l'identification de protéines présentes dans les banques de données

-Approche non tolérante aux erreurs
-Recherches en banque de données longues, en fonction de la banque de données choisie, car fichiers volumineux

Bibliographie

Aebersold R. and Goodlett D. R.

Mass spectrometry in proteomics. *Chem. Rev.*, **2001**, 101, 269-295.

Aebersold R. and Mann M.

Mass-spectrometry-based proteomics. *Nature*, **2003**, 422, 198-207.

Andrade M. A. and Sander C.

Bioinformatics: from genome data to biological knowledge. *Curr Opin Biotechnol*, **1997**, 8, 675-683.

Apweiler R., Bairoch A. and Wu C. H.

Protein sequence databases. *Curr. Opin. Chem. Biol.*, **2004**, 8, 76-80.

Bairoch A., Apweiler R., Wu C. H., Barker W. C., Boeckmann B., Ferro S., Gasteiger E., Huang H., Lopez R., Magrane M., Martin M. J., Natale D. A., O'Donovan C., Redaschi N. and Yeh L. S.

The Universal Protein Resource (Uniprot). *Nucleic Acids Res.*, **2005**, 33, 154-159.

Beausoleil S. A., Jedrychowski M., Schwartz D., Elias J. E., Villen J., Li J., Cohn M. A., Cantley L. C. and Gygi S. P.

Large-scale characterization of HeLa cell nuclear phosphoproteins. *Proc Natl Acad Sci U S A*, **2004**, 101, 12130-12135.

Beavis R. C. and Chait B. T.

Cinnamic acid derivatives as matrices for ultraviolet laser desorption mass spectrometry of proteins. *Rapid Commun Mass Spectrom*, **1989**, 3, 432-435.

Bell A. W., Ward M. A., Blackstock W. P., Freeman H. N., Choudhary J. S., Lewis A. P., Chotai D., Fazel A., Gushue J. N., Paiement J., Palcy S., Chevet E., Lafreniere-Roula M., Solari R., Thomas D. Y., Rowley A. and Bergeron J. J.

Proteomics characterization of abundant Golgi membrane proteins. *J Biol Chem*, **2001**, 276, 5152-5165.

Benson D. A., Karsch-Mizrachi I., Lipman D. J., Ostell J. and Wheeler D. L.

GenBank. *Nucleic Acids Res.*, **2006**, 34, 16-20.

Bianchetti L., Thompson J.D., Lecompte O., Plewniak F. and Poch O.

vALID: validation of protein sequence quality based on multiple alignment data. *J. Bioinform. Comput. Biol.*, **2005**, 3, 1-19.

Biemann K.

Appendix 5. Nomenclature for peptide fragment ions (positive ions). *Methods Enzymol*, **1990**, 193, 886-887.

Bjellqvist B., Ek K., Righetti P. G., Gianazza E., Gorg A., Westermeier R. and Postel W.

Isoelectric focusing in immobilized pH gradients: principle, methodology and some applications. *J Biochem Biophys Methods*, **1982**, 6, 317-339.

Blades A. T., Ikonomou M. G. and Kebarle P.

Mechanism of electrospray mass spectrometry. Electrospray as an electrolysis cell. *Anal Chem*, **1991**, 63, 2109-2114.

Blueggel M., Chamrad D. and Meyer H. E.

Bioinformatics in proteomics. *Curr Pharm Biotechnol*, **2004**, 5, 79-88.

Boguski, M. S.

Biosequence exegesis. *Science*, **1999**, 286, 453-455.

Bradshaw R. A.

Revised draft guidelines for proteomic data publication. *Mol. Cell. Proteomics*, **2005**, 4, 1223-1225.

Breci L. A., Tabb D. L., Yates J. R. I. and Wysocki V. H.

Cleavage N-terminal to proline : Analysis of a database of peptide tandem mass spectra. *Anal Chem*, **2003**, 75, 1963-1971.

Chamrad D. C., Koerting G., Gobom J., Thiele H., Klose J., Meyer H. E. and Blueggel M.

Interpretation of mass spectrometry data for high-throughput proteomics. *Anal Bioanal Chem*, **2003**, 376, 1014-1022.

Campana J. E.

Elementary theory of the quadrupole mass filter. *Int. J. Mass Spectrom. Ion Processes*, **1980**, 33, 101-117.

Campostrini N., Areces L. B., Rappsilber J., Pietrogrande M. C., Dondi F., Pastorino F., Ponzoni M. and Righetti P. G.

Spot overlapping in two-dimensional maps: a serious problem ignored for much too long. *Proteomics*, **2005**, 5, 2385-2395.

Carr S., Aebersold R., Baldwin M., Burlingame A., Clauser K. and Nesvizhskii A.

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*, **2004**, 3, 531-533.

Carte N.

La trappe ionique et l'ionisation electrospray : un nouveau potentiel pour la caractérisation de biomolécules. *Thèse de l'Université Louis Pasteur de Strasbourg*, **2001**.

Wu C. H., Yeh L. S. L., Huang H., Arminski L., Castro-Alvear J., Chen Y., Hu Z., Kourtesis P., Ledley R. S., Suzek B. E., Vinayaka C. R., Zhang J. and Barker W. C.

The Protein Information Resource. *Nucleic Acids Res.*, **2003**, 31, 345-347.

Choudhary J. S., Blackstock W. P., Creasy D. M. and Cottrell J. S.

Interrogating the human genome using uninterpreted mass spectrometry data. *Proteomics*, **2001**, 1, 651-667.

Clauser K. R., Baker P. R. and Burlingame A. L.

Role of accurate mass measurement (+/- 10ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal. Chem.*, **1999**, 71, 2871-2882.

Cochrane, G, Aldebert P., Althorpe N., Andersson M., Baker W., Baldwin A., Bates K., Bhattacharyya S., Browne P., van den Broek A., Castro M., Duggan K., Eberhardt R., Faruque N., Gamble J., Kanz C., Kulikova T., Lee C., Leinonen R., Lin Q., Lombard V., Lopez R., McHale M., McWilliam H., Mukherjee G., Nardone F., Pilar Garcia Pastor M., Sobhany S., Stoehr P., Tzouvara K., Vaughan R., Wu D., Zhu W and Apweiler R.

EMBL Nucleotide Sequence Database: developments in 2005. *Nucleic Acids Research*, **2006**, 34, 10-15.

Cohen, J.

The proteomics payoff. *Technol. Rev.*, **2001**, October, 55-60.

Cooks R. G., Glish G. L., McLuckey S. A. and Kaiser R. E.

Ion trap mass spectrometry. *Chemical & Engineering News*, **1991**, 25, 26-41.

Database Issue

Nucleic Acids Research, **2006**, 34.

Dehmelt H. G.

Radiofrequency spectroscopy of stored ions. I:Storage. *Adv. At. Mol. Phys.*, **1967**, 3, 53-72.

De Hoffmann E. and Stoobant V.

Mass spectrometry : Principles and applications. *Wiley, Chichester*, **2003**.

Delalande F.

Application du couplage chromatographie liquide-spectrométrie de masse à l'étude de la biodisponibilité de peptides issus de produits laitiers et à la protéomique. *Thèse de l'Université Louis Pasteur de Strasbourg*, **2003**.

Dole M., Mack L. L., Himes R. L., Mobley R. C., Ferguson L. D. and Alice M.B.

Molecular beams of macroions. *J. Chem. Phys.*, **1968**, 49, 2240-2249.

Domon B. and Aebersold R.

Mass spectrometry and protein analysis. *Science*, **2006**, 312, 212-217.

Domon B. and Costello C. E.

Structure elucidation of glycosphingolipids and gangliosides using high-performance tandem mass spectrometry. *Biochemistry*, **1988**, 27, 1534-1543.

Dongré A. R., Jones J. L., Somogyi A. and Wysocki V. H.

Influence of peptide composition, gas-phase basicity, and chemical modification on fragmentation efficiency : Evidence for the mobile proton model. *J. Am. Chem. Soc.*, **1996**, 118, 8365-8374.

Edman P.

Method for determination of the amino acid sequence in peptides. *Acta Chem. Scand.*, **1950**, 4, 283-293.

Edman P. and Begg G.

A protein sequenator. *Eur. J. Biochem.*, **1967**, 1, 80-91.

Elias J. E., Haas W., Faherty B. K. and Gygi S. P.

Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nat Methods*, **2005**, 2, 667-675.

Emmet M. R. and Caprioli R.M.

Micro-electrospray : ultra-high-sensitivity- analysis of peptides and proteins. *J. Am. Soc. Mass Spectrom.*, **1994**, 5, 605-613.

Eng J. K., McCormack A. L. and Yates J. R. III.

An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.*, **1994**, 5, 976-989.

Essader A. S., Cargile B. J., Bundy J. L. and Stephenson J. L. Jr.

A comparison of immobilized pH gradient isoelectric focusing and strong-cation-exchange chromatography as a first dimension in shotgun proteomics. *Proteomics*, **2005**, 5, 24-34.

Fenn J. B., Mann M., Meng C. K., Wong S. F. and Whitehouse C. M.

Electrospray ionization for mass spectrometry of large molecules. *Science*, **1989**, 246, 64-71.

Fenn J. B., Mann M., Meng C. K., Wong S. F. and Whitehouse C. M.

Electrospray ionization-principles and practice. *Mass Spectrom. Rev.*, **1990**, 9, 37-70.

Gaskell S. J.

Electrospray : principles and practice. *J. Mass Spectrom.*, **1997**, 32, 677-688.

Gattiker A., Bienvenut W. V., Bairoch A. and Gasteiger E.

FindPept, a tool to identify unmatched masses in peptide mass fingerprinting protein identification. *Proteomics*, **2002**, 2, 1435-1444.

Gerber S. A., Rush J., Stemman O., Kirschner M. W. and Gygi S. P.

Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc Natl Acad Sci U S A*, **2003**, 100, 6940-6945.

Gevaert K. and Vandekerckhove J.

Protein identification methods in proteomics. *Electrophoresis*, **2000**, 21, 1145-1154.

Goffeau A., Barrell B. G., Bussey H., Davis R. W., Dujon B., Feldmann H., Galibert F., Doheisel J. D., Jacq C., Johnston M., Louis E. J., Mewes H. W., Murakami Y., Philippsen P., Tettelin H. and Oliver S. G.

Life with 6000 genes. *Science*, **1996**, 274 (5287), 546-567.

Gorg A., Postel W., Domscheit A. and Gunther S.

Two-dimensional electrophoresis with immobilized pH gradients of leaf proteins from barley (Hordeum vulgare): method, reproducibility and genetic aspects. *Electrophoresis*, **1988**, 9, 681-692.

Gorg A., Weiss W. and Dunn M. J.

Current two-dimensional electrophoresis technology for proteomics. *Proteomics*, **2004**, 4, 3665-3685.

Griffiths W. J., Jonsson A. P., Liu S., Rai D. K. and Wang Y.

Electrospray and tandem mass spectrometry in biochemistry. *Biochem J*, **2001**, 355, 545-561.

Gygi S. P., Corthals G. L., Zhang Y., Rochon Y. and Aebersold R.

Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc Natl Acad Sci U S A*, **2000**, 97, 9390-9395.

Gygi S. P., Rist B., Gerber S. A., Turecek F., Gelb M. H. and Aebersold R.

Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol*, **1999**, 17, 994-999.

Hager J. W.

A new linear ion trap mass spectrometer. *Rapid Commun Mass Spectrom*, **2002**, 16, 512-526.

Hardman M. and Makarov A. A.

Interfacing the orbitrap mass analyzer to an electrospray ion source. *Anal Chem*, **2003**, 75, 1699-1705.

Heller M., Mattou H., Menzel C. and Yao X.

Trypsin catalyzed 16O-to-18O exchange for comparative proteomics: tandem mass spectrometry comparison using MALDI-TOF, ESI-QTOF, and ESI-ion trap mass spectrometers. *J Am Soc Mass Spectrom*, **2003**, 14, 704-718.

Henzel W. J., Billeci T. M., Stults J. T., Wong S. C., Grimley C. and Watanabe C.

Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proc. Natl. Acad. Sci. USA*, **1993**, 90, 5011-5015.

Hernandez P., Müller M. and Appel R. D.

Automated protein identification by tandem mass spectrometry : issues and strategies. *Mass Spectrom. Rev.*, **2006**, 25, 235-254.

Hieter P. and Boguski M.

Functional genomics: It's all how you read it. *Science*, **1997**, 278, 601-602.

Horn D. M., Peters E. C., Klock H., Meyers A. and Brock A.

Improved protein identification using automated high mass measurement accuracy MALDI FT-ICR MS peptide mass fingerprinting. *Int. J. Mass Spectrom.*, **2004**, 238, 189-196.

Hoving S., Voshol H. and van Oostrum J.

Towards high performance two-dimensional gel electrophoresis using ultrazoom gels. *Electrophoresis*, **2000**, 21, 2617-2621.

Hu Q., Noll R. J., Li H., Makarov A. A., Hardman M. and Cooks R. G.

The Orbitrap: a new mass spectrometer. *J. Mass Spectrom.*, **2005**, 40, 430-443.

Hunt D. F., Henderson R. A., Shabanowitz J., Sakaguchi K., Michel H., Sevilir N., Cox A. L., Appella E. and Engelhard V. H.

Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry. *Science*, **1992**, 255, 1261-1263.

Ikonomou M. G., Blades A. T. and Kebarle P.

Electrospray-Ion spray: A comparaison of mechanisms and performance. *Anal Chem*, **1991**, 63, 1989-1998.

Ikonomou M., Blades A. T. and Kebarle P.

Investigations of the Electrospray Interface for Liquid ChromatographyIMass Spectrometry. *Anal Chem*, **1990**, 62, 957-967.

Iribarne J. V. and Thomson B. A.

On the evaporation of small ions from charged droplets. *J. Chem. Phys.*, **1976**, 64, 2287-2294.

Jacobs J. M., Mottaz H. M., Yu L. R., Anderson D. J., Moore R. J., Chen W. N., Auberry K. J., Strittmatter E. F., Monroe M. E., Thrall B. D., Camp D. G. 2nd. and Smith R. D.

Multidimensional proteome analysis of human mammary epithelial cells. *J Proteome Res*, **2004**, 3, 68-75.

James P., Quadroni M., Carafoli E. and Gonnet G.

Protein identification by mass profile fingerprinting. *Biochem. Biophys. Res. Commun.*, **1993**, 195, 58-64.

Jensen O. N.

Interpreting the protein language using proteomics. *Nat Rev Mol Cell Biol*, **2006**, 7, 391-403.

Karas M., Bachmann D., Bahr U. and Hillenkamp F.

Matrix-assisted ultraviolet laser desorption of non-volatile compounds. *Int. J. Mass Spectrom. Ion Processes*, **1987**, 78, 53-68.

Karas M., Bahr U., Ingendoh A. and Hillenkamp F.

Laser desorption ionization mass spectrometry of proteins of mass 100000 to 250000 Daltons. *Angew Chem Int Ed Engl*, **1989**, 28, 760-761.

Karas M. and Hillenkamp F.

Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal Chem*, **1988**, 60, 2299-2301.

Kebarle P.

A brief overview of the present status of the mechanism involved in electrospray mass spectrometry. *J. Mass Spectrom.*, **2000**, 35, 804-817.

Keller A., Nesvizhskii A. I., Kolker E. and Aebersold R.

Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem*, **2002**, 74, 5383-5392.

Kelley P. E., Stafford G. C. S. and Stephens D. R.

US patent 4 540 884, **1985**.

Kenrick K. G. and Margolis J.

Isoelectric focusing and gradient gel electrophoresis: a two-dimensional technique. *Anal Biochem*, **1970**, 33, 204-207.

Kenyon G. L., DeMarini D. M., Fuchs E., Galas D. J., Kirsch J. F., Leyh T. S., Moos W. H., Petsko G. A., Ringe D., Rubin G. M. and Sheahan L. C.

Defining the mandate of proteomics in the post-genomics era: workshop report. *Mol. Cell. Proteomics*, **2002**, 1, 763-780.

Knochenmuss R. and Zenobi R.

MALDI ionization: the role of in-plume processes. *Chem Rev*, **2003**, 103, 441-452.

Kuster B., Mortensen P., Andersen J. S. and Mann M.

Mass spectrometry allows direct identification of proteins in large genomes. *Proteomics*, **2001**, 1, 641-650.

Lane C. S.

Mass spectrometry-based proteomics in the life sciences. *Cell. Mol. Life Sci.*, **2005**, 62, 848-869.

Lee I. N., Chen C. H., Sheu J. C., Lee H. S., Huang G. T., Yu C. Y., Lu F. J. and Chow L. P.

Identification of human hepatocellular carcinoma-related biomarkers by two-dimensional difference gel electrophoresis and mass spectrometry. *J Proteome Res*, **2005**, 4, 2062-2069.

Leize E.

Caractèrisation d'édifices supramoléculaires par spectrométrie de masse avec ionisaton électropsray. *Thèse de l'Université Louis Pasteur de Strasbourg*, **1994**.

Link A. J., Eng J., Schieltz D. M., Carmack E., Mize G. J., Morris D. R., Garvik B. M. and Yates J. R. III.

Direct analysis of protein complexes using mass spectrometry. *Nat Biotechnol*, **1999**, 17, 676-682.

Liolios K., Tavernarakis N., Hugenholtz P. and Kyrpides N. C.

The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. *Nucleic Acids Res.*, **2006**, 34, 332-334.

Loo J. A., Loo R. R., Light K. J., Edmonds C. G. and Smith R. D.

Multiply charged negative ions by electrospray ionization of polypeptides and proteins. *Anal Chem*, **1992**, 64, 81-88.

Lottspeich F.

Proteome Analysis: A Pathway to the Functional Analysis of Proteins. *Angew. Chem. Int. Ed. Engl.*, **1999**, 38, 2476-2492.

Mamyrin B. A., Karataev V. I., Shmikk D.V. and Zagulin V. A.

Mass reflectron. New nonmagnetic time-of-flight high-resolution mass spectrometer. *Zh. Eksp. Theor. Fiz.*, **1973**, 64, 82-89.

Mann M. and Jensen O. N.

Proteomic analysis of post-translational modifications. *Nat Biotechnol*, **2003**, 21, 255-261.

Mann M., Hojrup P. and Roepstorff P.

Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biol. Mass Spectrom.*, **1993**, 22, 338-345.

Mann M. and Wilm M.

Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal Chem*, **1994**, 66, 4390-4399.

March R. E.

An introduction to quadrupole ion trap mass spectrometry. *J. Mass Spectrom.*, **1997**, 32, 351-369.

March R. E.

Quadrupole ion trap mass spectrometry : Theory, simulation, recent developments and applications. *Rapid Commun Mass Spectrom*, **1998**, 12, 1543-1554.

March R. E., Hugues R. J. and Todd J. F. J.

Quadrupole storage mass spectrometry. *Wiley Interscience, New York*, **1989**.

March R. E. and Todd J. F. J.

Practical aspects of ion trap mass spectrometry. Ion trap instrumentation, *CRC Series Modern Mass Spectrometry*, **1995**, Volume II.

Mawuenyega K. G., Kaji H., Yamuchi Y., Shinkawa T., Saito H., Taoka M., Takahashi N. and Isobe T.

Large-scale identification of Caenorhabditis elegans proteins by multidimensional liquid chromatography-tandem mass spectrometry. *J Proteome Res*, **2003**, 2, 23-35.

McLuckey S. A., Van Berkel G. J. and Glish G. L.

Tandem Mass Spectrometry of Small, Multiply charged Oligonucleotides. *J. Am. Soc. Mass Spectrom.*, **1992**, 3, 60-70.

Meng C. K., Mann M. and Fenn J. B.

Of protons or proteins "A beam's a beam for a' that." *Z. Phys. D.*, **1988**, 10, 361-368.

Michel P. E., Crettaz D., Morier P., Heller M., Gallot D., Tissot J. D., Reymond F. and Rossier J. S.

Proteome analysis of human plasma and amniotic fluid by Off-Gel isoelectric focusing followed by nano-LC-MS/MS. *Electrophoresis*, **2006**, 27, 1169-1181.

Michel P. E., Reymond F., Arnaud I. L., Josserand J., Girault H. H. and Rossier J. S.

Protein fractionation in a multicompartment device using Off-Gel isoelectric focusing. *Electrophoresis*, **2003**, 24, 3-11.

Mirgorodskaya O. A., Kozmin Y. P., Titov M. I., Korner R., Sonksen C. P. and Roepstorff P.

Quantitation of peptides and proteins by matrix-assisted laser desorption/ionization mass spectrometry using (18)O-labeled internal standards. *Rapid Commun Mass Spectrom*, **2000**, 14, 1226-1232.

Mulder N. J., Apweiler R., Attwood T. K., Bairoch A., Barrell D., Bateman A., Binns D., Biswas M., Bradley P., Bork P., Bucher P., Copley R. R., Courcelle E., Das U., Durbin R., Falquet L., Fleischmann W., Griffiths-Jones S., Haft D., Harte N., Hulo N., Kahn D., Kanapin A., Krestyaninova M., Lopez R., Letunic I., Lonsdale D., Silventoinen V., Orchard S. E., Pagni M., Peyruc D., Ponting C. P., Selengut J. D., Servant F., Sigrist C. J. A., Vaughan R. and Zdobnov E. M.

The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, **2001**, 29, 37-40.

Nagele E., Vollmer M. and Horth P.

Improved 2D nano-LC/MS for proteomics applications: a comparative analysis using yeast proteome. *J Biomol Tech*, **2004**, 15, 134-143.

Nesvizhskii A. I. and Aebersold R.

Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem MS. *Drug Discov Today*, **2004**, 9, 173-181.

Nesvizhskii A. I. and Aebersold R.

Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics*, **2005**, 4, 1419-1440.

Nesvizhskii A. I., Keller A., Kolker E. and Aebersold R.

A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem*, **2003**, 75, 4646-4658.

Nielsen P. and Krogh A.

Large-scale prokaryotic gene prediction and comparison to genome annotation. *Bioinformatics*, **2005**, 21, 4322-4329.

O'Farrrell

High resolution two-dimentional electrophoresis of proteins. *J. Biol. Chem.*, **1975**, 250, 4007-4021.

Okubo K., Sugawara H., Gojobori T. and Tateno Y.

DDBJ in preparation for overview of research activities behind data submissions. *Nucleic Acids Res.*, **2006**, 34, 6-9.

Ong S. E., Blagoev B., Kratchmarova I., Kristensen D. B., Steen H., Pandey A. and Mann M.

Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics*, **2002**, 1, 376-386.

Ouyang Z., Wu G., Song Y., Li H., Plass W. R. and Cooks R. G.

Rectilinear ion trap : concepts, calculations, and analytical performance of a new mass analyzer. *Anal Chem*, **2004**, 76, 4595-4605.

Paizs B. and Suhai S.

Fragmentation pathways of protonated peptides. *Mass Spectrom. Rev.*, **2005**, 24, 508-548.

Paoletti A. C., Zybailov B. and Washburn M. P.

Principles and applications of multidimensional protein identification technology. *Expert Rev Proteomics*, **2004**, 1, 275-282.

Pappin D. D. J., Hojrup P. and Bleasby A. J.

Rapid identification of proteins by peptide-mass finger printing. *Curr. Biol.*, **1993**, 3, 327-332.

Paul W.

Elektromagnetische Käfige für geladene und neutrale Teilchen (Nobel-Vortrag). *Angew. Chem.*, **1990**, 102, 780-789.

Paul W. and Steinwedel H.

US Patent 2 939 952, **1960**.

Peng J., Elias J. E., Thoreen C. C., Licklider L. J. and Gygi S. P.

Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J Proteome Res*, **2003**, 2, 43-50.

Perkins D. N., Pappin D. D. J., Creasy D. M. and Cottrell J. S.

Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **1999**, 20, 3551-3567.

Pruitt K. D., Tatusova T. and Maglott D. R.

NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **2005**, 33, 501-504.

Rabilloud T.

Two-dimensional gel electrophoresis in proteomics: Old, old fashioned, but still climbs up the mountains. *Proteomics*, **2002**, 2, 3-10.

Rabilloud T., Adessi C., Giraudel A. and Lunardi J.

Improvement of the solubilization of proteins in two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis*, **1997**, 18, 307-316.

Rabilloud T., Strub J. M., Luche S., Van Dorsselaer A. and Lunardi J.

A comparison between Sypro Ruby and ruthenium II tris (bathophenanthroline disulfonate) as fluorescent stains for protein detection in gels. *Proteomics*, **2001**, 1, 699-704.

Ramus C., Gonzalez de Peredo A., Dahout C., Gallagher M. and Garin J.

An optimized strategy for ICAT quantification of membrane proteins. *Mol. Cell. Proteomics*, **2006**, 5, 68-78.

Richert S., Luche S., Chevallet M., Van Dorsselaer A., Leize-Wagner E. and Rabilloud T.

About the mechanism of interference of silver staining with peptide mass spectrometry. *Proteomics*, **2004**, 4, 909-916.

Roepstorff P. and Fohlman J.

Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed Mass Spectrom*, **1984**, 11, 601.

Ros A., Faupel M., Mees H., Oostrum J., Ferrigno R., Reymond F., Michel P., Rossier J. S. and Girault H. H.

Protein purification by Off-Gel electrophoresis. *Proteomics*, **2002**, 2, 151-156.

Sanglier S.

La spectrométrie de masse : un nouvel outil pour l'étude des interactions faibles en biologie. *Thèse de l'Université Louis Pasteur de Strasbourg*, **2002**.

Sanglier S., Leize E., Van Dorsselaer A. and Zal F.

Comparative ESI-MS study of approximately 2.2 MDa native hemocyanins from deep-sea and shore crabs: from protein oligomeric state to biotope. *J Am Soc Mass Spectrom*, **2003**, 14, 419-429.

Santoni V., Molloy M. and Rabilloud T.

Membrane proteins and proteomics : un amour impossible? *Electrophoresis*, **2000**, 21, 1054-1070.

Savitski M. M., Nielsen M. L., Kjeldsen F. and Zubarev R. A.

Proteomics-grade de novo sequencing approach. *J. Proteome Res*., **2005**, 4, 2348-2354.

Schneider M., Tognolli M. and Bairoch A.

The Swiss-Prot protein knowledgebase and ExPASy : providing the plant community with high quality proteomic data and tools., *Plant Physiology and Biochemistry*, **2004**, 42, 1013-1021.

Schwartz J. C., Senko M. W. and Syka J. E. P.

A two-dimensional quadrupole ion trap mass spectrometer. *J Am Soc Mass Spectrom*, **2002**, 13, 659-669.

Shevchenko A., Wilm M., Vorm O. and Mann M.

Mass spectrometric sequencing of proteins from silver-stained polyacrylamide gels. *Anal Chem*, **1996**, 68, 850-858.

Stafford G. C. S., Kelley P. E., Syka J. E. P., Reynolds W. E. and Todd J. F.

Recent improvements in, and analytical applications of advanced ion trap technology. *Int. J. Mass Spectrom. Ion Processes*, **1984**, 60, 85-98.

Standing K. G.

Timing the flight of biomolecules : a personal perspective. *Int. J. Mass Spectrom.*, **2000**, 200, 597-610.

Steen H. and Mann M.

The ABC's (and XYZ's) of peptide sequencing. *Nat Rev Mol Cell Biol*, **2004**, 5, 699-711.

Syka J. E. P., Coon J. J., Schroeder M. J., Shabanowitz J. and Hunt D. F.

Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *PNAS*, **2004**, 101(26), 9528-9533.

Swanson S. K. and Washburn M. P.

The continuing evolution of shotgun proteomics. *Drug Discov Today*, **2005**, 10, 719-725.

Tanaka K., Waki H., Ido Y., Akita S., Yoshida Y. and Yoshida T.

Protein and polymer analyses up to m/z 100,000 by laser ionization time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.*, **1988**, 2, 151-153.

Tang X. J., Thibault P. and Boyd R. K.

Frangmentation reactions of multiply-protonated peptides and implications for sequencing by tandem mass spectrometry with low-energy collision-induced dissociation. *Anal Chem*, **1993**, 65, 2824-2834.

Thompson J. J.

Rays of positive electricity and their applications to chemical analysis. *Longmans Green, London*, **1913**.

Tuloup M., Hernandez C., Coro I., Hoogland C., Binz P. A. and Appel R. D.

Aldente and BioGraph : An improved peptide mass fingerprinting protein identification environment. *Swiss Proteomics Society 2003 Congress [Fontis Media]*, **2003**, 174-176.

Tyers M. and Mann M.

From genomics to proteomics. *Nature*, **2003**, 422, 193-197.

Wagner E.

Nouvelles approches de la spectrométrie de masse en tendem pour la caractérisatoin de biomolécules. Applications à la protéomique. *Thèse de l'Université Louis Pasteur de Strasbourg*, **2004**.

Washburn M. P., Wolters D. and Yates J. R. III.

Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.*, **2001**, 19, 242-247.

Westbrook J., Feng Z., Chen L., Yang H. and Berman H. M.

The Protein Data Bank and structural genomics. *Nucleic Acids Res.*, **2003**, 31, 489-491.

Wiley W. C. and McLaren I. H.

Time-of-flight mass spectrometry with improved resolution. *Rev. Sci. Instrum.*, **1955**, 26, 1150-1157.

Wilkins M. R., Appel R. D., Van Eyk J. E., Chung M. C., Gorg A., Hecker M., Huber L. A., Langen H., Link A. J., Paik Y. K., Patterson S. D., Pennington S. R., Rabilloud T., Simpson R. J., Weiss W. and Dunn M. J.

Guidelines for the next 10 years of proteomics. *Proteomics*, **2006**, 6, 4-8.

Wilkins M. R., Gasteiger E., Wheeler C. H., Lindskog I., Sanchez J. C., Bairoch A., Appel R. D., Dunn M. J. and Hochstrasser D. F.

Multiple parameter cross-species protein identification using MultiIdent--a world-wide web accessible tool. *Electrophoresis*, **1998**, 19, 3199-3206.

Wilkins M. R., Pasquali C., Appel R. D., Ou K., Golaz O., Sanchez J. C., Yan J. X., Gooley A. A., Hughes G., Humphery-Smith I., Williams K. L. and Hochstrasser D. F.

From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Biotechnology (N Y)*, **1996**, 14, 61-65.

Wilm M. and Mann M.

Analytical properties of the nanoelectrospray ion source. *Anal Chem*, **1996**, 68, 1-8.

Wolters D. A., Washburn M. P. and Yates J. R., 3rd

An automated multidimensional protein identification technology for shotgun proteomics. *Anal Chem*, **2001**, 73, 5683-5690.

Wu C. C. and MacCoss M. J.

Shotgun proteomics: tools for the analysis of complex biological systems. *Curr Opin Mol Ther*, **2002**, 4, 242-250.

Wu C. H., Apweiler R., Bairoch A., Natale D. A., Barker W. C., Boeckmann B., Ferro S., Gasteiger E., Huang H., Lopez R., Magrane M., Martin M. J., Mazumder R., O'Donovan C., Redaschi N. and Suzek B.

The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.*, **2006**, 34, 187-191.

Wysocki V. H., Tsaprailis G., Smith L. L. and Breci L. A.

Mobile and localized protons : a framework for understanding peptide dissociation. *J. Mass Spectrom.*, **2000**, 35, 1399-1406.

Xiong Y., Chalmers M. J., Gao F. P., Cross T. A. and Marshall A. G.

Identification of Mycobacterium tuberculosis H37Rv integral membrane proteins by one-dimensional gel electrophoresis and liquid chromatography electrospray ionization tandem mass spectrometry. *J Proteome Res*, **2005**, 4, 855-861.

Xiang R., Shi Y., Dillon D. A., Negin B., Horvath C. and Wilkins J. A.

2D LC/MS analysis of membrane proteins from breast cancer cell lines MCF7 and BT474. *J Proteome Res*, **2004**, 3, 1278-1283.

Yamashita M. and Fenn J.

Electrospray ion source. Another variation on the Free-jet theme. *J. Phys. Chem.*, **1984a**, 88, 4451-4460.

Yamashita M. and Fenn J.

Negative Ion Production with the electrospray Ion Source. *J. Phys. Chem.*, **1984b**, 88, 4671-4675.

Yates J. R. III.

Mass spectrometry. From genomics to proteomics. *Trends Genet,* **2000**, 16, 5-8.

Yates J. R. III, Speicher S., Griffin P. R. and T. H.

Peptide mass maps : A highly informative approach to protein identification. *Anal. Biochem.*, **1993**, 214, 397-408.

Zhang W. and Chait B. T.

ProFound: an expert system for protein identification using mass spectrometric peptide mapping information. *Anal Chem*, **2000**, 72, 2482-2489.

Zubarev R. A., Kelleher N. L. and McLafferty F. W.

Electron capture dissociation of multiply charged protein cations. A nonergodic process. *J. Am. Chem. Soc.*, **1998**, 120, 3265-3266.

RESULTATS

PREMIERE PARTIE :
Amélioration de la prise de données MS/MS

Chapitre I : Le couplage nanoHPLC-Trappe ionique

Chapitre II : Les nouveautés instrumentales du couplage : le système nanoHPLC-Chip Cube et la fragmentation ETD

CHAPITRE I

Le couplage nanoHPLC-trappe ionique

Une trappe ionique Esquire 3000+ (Bruker Daltonics) était installée au laboratoire à mon arrivée et était utilisée pour des analyses nanoLC-MS/MS en couplage avec une nanoHPLC Agilent Series 1100 (Agilent Technologies). J'ai eu l'occasion de me familiariser au principe de fonctionnement de la trappe et au couplage nanoLC-MS/MS sur ce système durant mes premiers mois de thèse. Puis le laboratoire a fait l'acquisition d'une trappe ionique de nouvelle génération, une trappe de type HCT Plus (High Capacity Trap, Bruker Daltonics). Une partie de mon travail de thèse a consisté en l'installation, l'évaluation des performances et l'optimisation de cette nouvelle trappe ionique pour l'analyse des protéines et des peptides et en la mise en place de cette nouvelle trappe en couplage nanoLC-MS/MS avec la nanoHPLC Agilent Series 1100.

Dans ce chapitre seront décrites quelques spécificités et optimisations instrumentales qui auront permis d'améliorer les performances de la trappe ionique mais aussi du couplage nanoLC-MS/MS.

1 La trappe HCTPlus

1.1 Optimisations des paramètres d'acquisition de la trappe

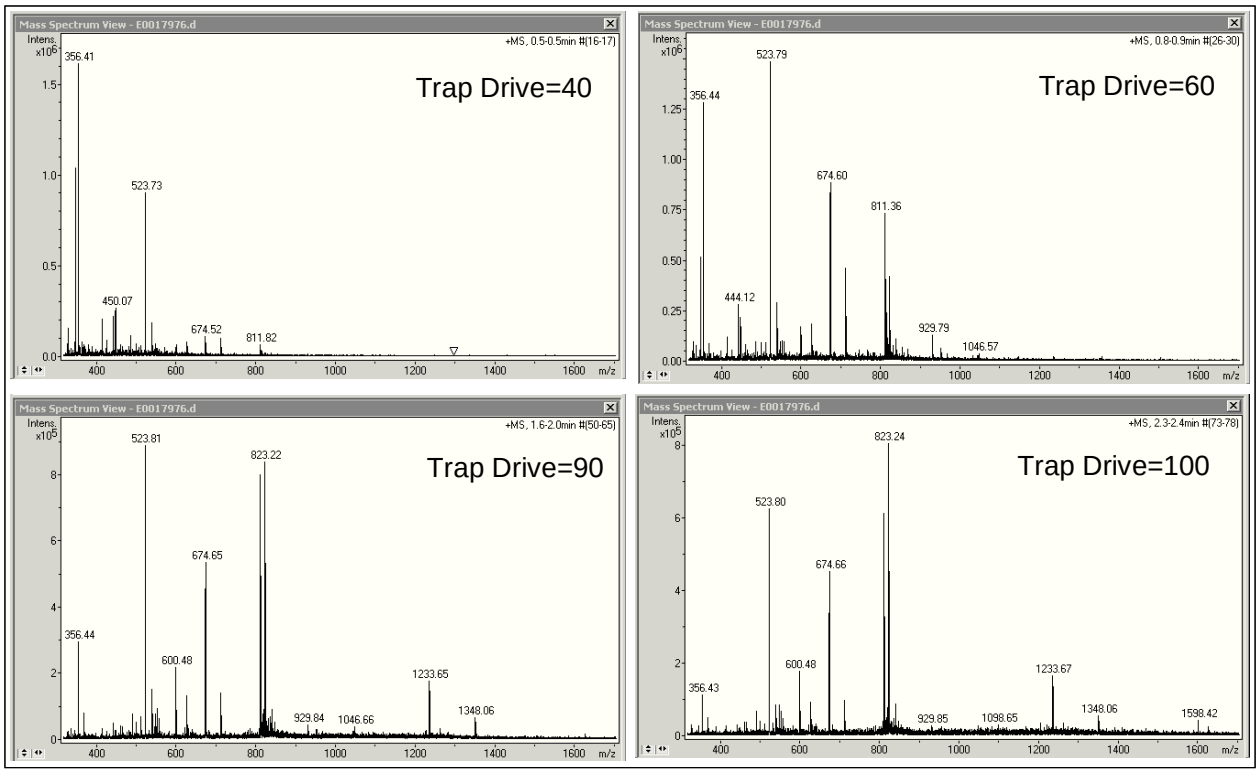
1.1.1 Le cut-off ou la limite inférieure de piégeage des ions

Les ions sont piégés dans la trappe sous l'effet d'un champ quadripolaire généré par la radiofréquence appliquée sur l'électrode annulaire de la trappe. Comme décrit dans la partie bibliographique (Chapitre I.1.2.3), le choix de l'amplitude de la radiofréquence détermine la gamme de masse m/z qu'il est possible de piéger dans la trappe. Pour une amplitude de cette radiofréquence donnée, un « cut-off » est défini, correspondant à la valeur du plus petit m/z pouvant être piégé dans l'analyseur. Sur les instruments commerciaux, l'amplitude de la radiofréquence est réglée indirectement par l'utilisateur. Selon le type d'instrument utilisé, ce paramètre de réglage porte une appellation différente : sur les trappes ioniques de type HCT, ce réglage se fait par l'optimisation du paramètre appelé Trap Drive. Plus le Trap Drive sera fixé

à une valeur haute, plus le piégeage des ions lourds (m/z élevés) sera efficace et inversement un Trap Drive bas favorisera le piégeage des ions légers (m/z bas). En figure 1 est illustré l'effet de la modification du paramètre Trap Drive lors de l'injection, dans des conditions d'ESI standard, d'un mélange de peptides (1picomol/µl) en infusion directe à 1µl/min. Le bon paramétrage de cette valeur est primordial et en travaillant en infusion, une optimisation préalable de cette valeur s'impose en fonction de la masse du composé analysé. En effet, la figure 1 montre bien que ce paramètre n'est pas anodin car un mauvais réglage de Trap Drive peut entraîner jusqu'à l'absence totale de visualisation de l'ion d'intérêt. Le logiciel d'acquisition de la trappe propose d'ailleurs un volet pour « utilisateurs non experts » qui permet une optimisation automatique de cette valeur en fonction de la masse d'intérêt que l'utilisateur précisera.

Cette optimisation devient plus délicate lorsqu'il s'agit d'étudier des mélanges complexes ou lors du travail en couplage (donc en acquisition automatique) car le piégeage dans la trappe ne peut être simultanément optimal pour les petits et les hauts m/z. Un compromis doit donc être trouvé dans ces conditions afin d'obtenir les spectres les plus informatifs et de meilleure qualité possible.

Figure 1 : Illustration de l'importance du réglage du paramètre Trap Drive définissant l'amplitude de la radiofréquence appliquée sur l'électrode annulaire de la trappe et donc le cut-off de piégeage des ions. Plus la valeur du trap drive est basse, plus les ions de m/z petits sont favorisés, plus la valeur du trap drive est élevée, plus les ions de m/z hauts sont favorisés.
Les peptides présents dans le mélange sont :
La Leu-enkephaline, m/z=712,378 (1+) et 356,689 (2+) ; l'Angiotensine, m/z=1046,542 (1+) et 523,771 (2+) ; la Substance P, m/z=1347,736 (1+) et 674,368 (2+) ; la Bombésine, m/z=1620,807 (1+) et 810,903 (2+) ; l'ACTH, m/z=1233,099 (2+) et 822,399 (3+).



1.1.2 Les phénomènes d'espace-charge et le réglage de l'ICC dans la trappe (Ion Charge Control)

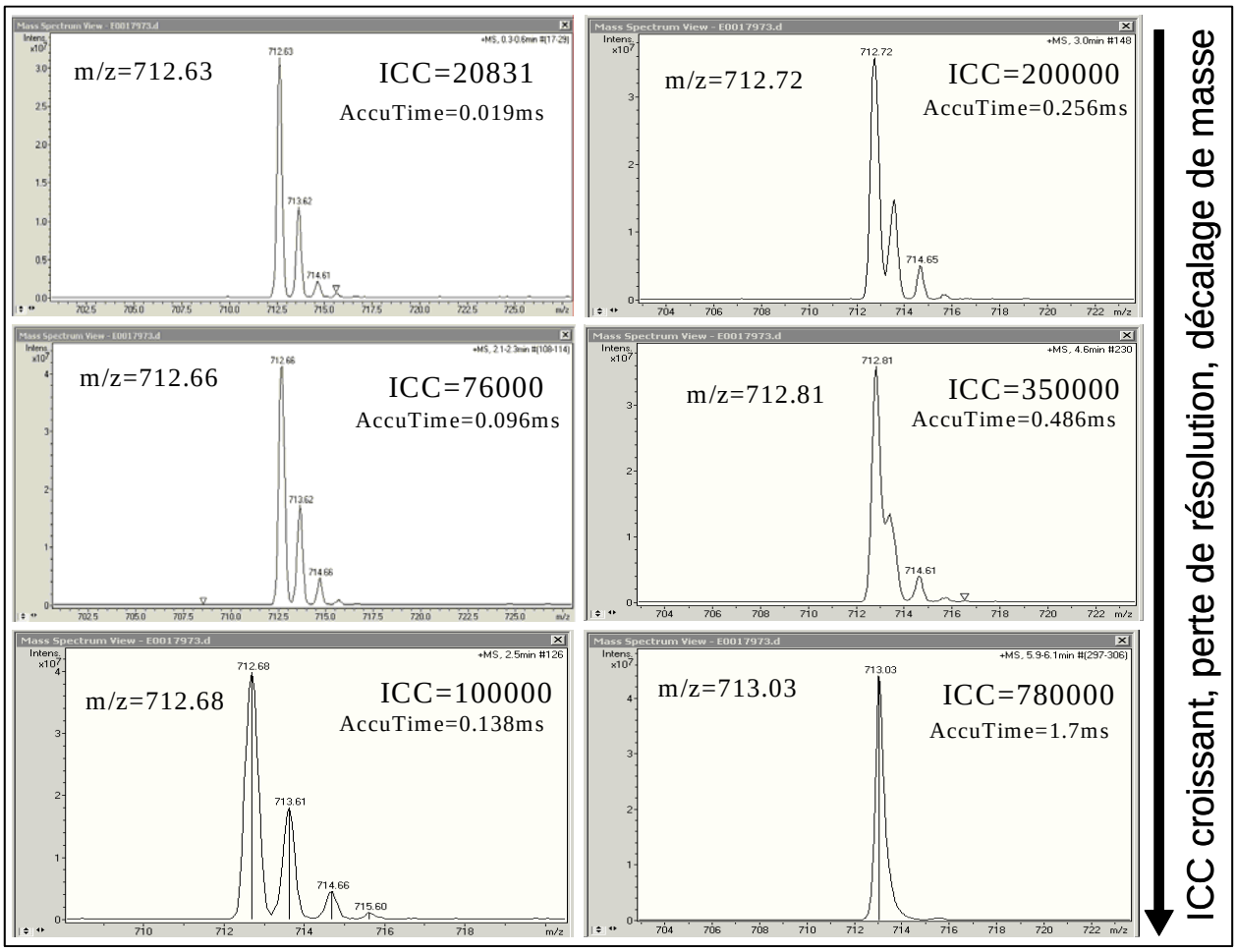
Une fois l'amplitude de la radiofréquence réglée pour un piégeage optimal des ions, il s'agit ensuite de contrôler un phénomène important pouvant affecter la précision de la mesure de masse et la résolution dans les instruments de type trappe ionique : le phénomène d'espace-charge [Cox et coll., 1995]. Ce phénomène résulte d'une accumulation trop importante des ions dans l'analyseur : les ions accumulés en surnombre dans la trappe induisent des répulsions coulombiennes, créant un champ additionnel qui modifie ponctuellement le champ quadripolaire dans l'espace. Le champ quadripolaire est perturbé, devient alors beaucoup moins efficace pour le piégeage des ions et les performances de l'analyseur en résolution et en précision de masse sont réduites [Guan et coll., 1994 ; Li et coll., 1997]. Lorsque trop d'ions de même polarité sont piégés dans la trappe, il en résulte une éjection retardée des ions due à un décalage du diagramme de stabilité vers des plus hauts q_z ce qui entraîne une augmentation dans les valeurs de m/z mesurées [Dobson et coll., 2004].

Les options d'Ion Charge Control (ICC) sur les trappes HCT permettent de contrôler et de réguler le flux d'ions entrants dans la trappe. L'utilisateur entre une valeur cible "smart target" à atteindre (valeur sans unité) qu'il a au préalable optimisée afin de limiter au maximum les espaces-charges. Le temps d'accumulation des ions est automatiquement ajusté afin d'atteindre la valeur cible. Lorsque le flux d'ions augmente, le temps d'accumulation est raccourci, lorsque le flux d'ions baisse, le temps d'accumulation est augmenté jusqu'à une limite de temps également définie par l'utilisateur. Cette limite de temps (Max Accu Time) permet de ne pas passer trop de temps à l'acquisition d'un spectre et de passer au spectre suivant lorsque le signal est vraiment faible. La figure 2 illustre les phénomènes d'espace-charge observés sur la leu-enkephaline en augmentant le nombre d'ions accumulés dans la trappe. On constate à la fois le décalage dans la mesure de masse et la perte de résolution.

De la même manière que le Trap Drive, l'optimisation de la quantité d'ions dans la trappe se fait en fonction de la gamme de masse étudiée et de la nature de l'échantillon.

C'est principalement la possibilité d'accumuler beaucoup plus d'ions avant l'apparition des effets d'espace-charge dans les trappes ioniques de type HCT (High Capacity Trap) par rapport aux trappes de type Esquire qui a permis les gains de sensibilité considérables. Pour des analyses en routine, l'ICC a été optimisé à 100000 en mode MS sur les trappes HCT alors qu'avec la trappe Esquire 3000+, au-delà de 40000, les effets d'espace-charges étaient déjà très importants.

Figure 2 : Observation des phénomènes d'espace-charge sur l'état de charge 1+ de la leu-enkephaline. Plus l'ICC est élevé, plus le temps d'accumulation augmente, plus la quantité d'ions dans la trappe augmente et plus les effets d'espace-charge sont observés (décalage vers des masses plus élevées et perte de résolution).



1.1.3 Le choix du mode de balayage

Plusieurs modes de vitesses de balayage sont disponibles sur les nouvelles générations d'instruments de type trappe ionique. En particulier un mode de balayage très rapide, le mode UltraScan à 26000m/z/s. La vitesse de balayage utilisée influe sur la résolution et la sensibilité de l'analyseur. Yang et coll. ont étudié en détail l'influence de l'augmentation de la vitesse de balayage sur la sensibilité et la résolution et ont décrit jusqu'à un facteur 200 de gain en sensibilité en augmentant la vitesse de balayage 12 fois [Yang et coll., 2005].

Comme illustré en figure 3, en passant d'une vitesse de balayage à 8100m/z/s à une vitesse de balayage de 26000m/z/s, un gain en sensibilité est observé. Par contre, la résolution et donc la précision sur la mesure de masse sont considérablement réduites. Le gain en sensibilité est principalement du à des effets d'espace-charge réduits lorsque la vitesse de balayage est augmentée. La perte en résolution s'explique par l'éjection plus rapide, donc moins spécifiques des ions.

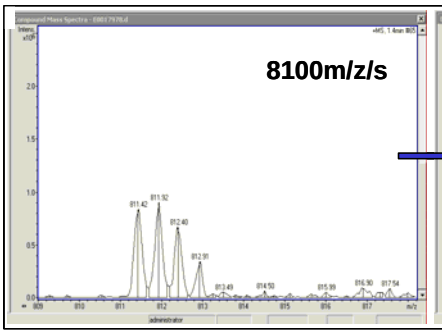
En fonction de l'analyse effectuée, le choix de la vitesse de balayage a son importance. Le gain en sensibilité du mode Ultrascan est très utile lorsqu'il s'agit de détecter des composés très peu abondants. Par contre, la perte de la résolution isotopique et donc de la possibilité de détermination de l'état de charge est

un désavantage considérable lorsque l'on s'intéresse à des composés présentant jusqu'à 3 voire 4 charges, observables avec le mode balayage standard (« mode standard enhanced » à 8100m/z/s). Il s'agira donc de bien choisir, en fonction de la masse et de l'état de charge des composés analysés de privilégier l'intensité du signal ou la résolution.

Par contre, le mode de balayage rapide présente des avantages considérables pour la MS/MS car le gain en sensibilité permet d'isoler plus d'ions et ouvre donc la possibilité de réaliser des MSn de degré n plus élevé.

Figure 3 : A/ Illustration du gain en sensibilité mais de la perte de résolution en passant du mode de balayage à 8100m/z/s au mode de balayage à 26000m/z/s, l'échelle d'intensité est respectée. B/ Les différents modes de balayage disponibles sur les trappes HCTPlus, HCTUltra et HCTUltra PTM Discovery System.

A/



B/

Mode	Vitesse de balayage (m/z/s)	Gamme de masse	Largeur des pics (FWHM)
UltraScan	26000	50-3000	0,6
Standard Enhanced	8100	50-3000	0,4
Standard Maximum	800	50-3000	0,25
Extended	27000	50-6000	3

Perte en résolution mais gain en sensibilité

1.2 La MS/MS sur les trappes HCTPlus

Des nouvelles fonctionnalités et possibilités de paramétrage sur les trappes ioniques HCT ont permis de révolutionner la qualité des spectres MS/MS :

- **<u>La nouvelle géométrie de la trappe permet une plus grande capacité de charge</u>**

  La plus grande quantité d'ions pouvant être accumulée dans la trappe permet d'augmenter significativement la sensibilité en MS mais aussi et surtout en MS/MS. En effet, les spectres MS/MS sur les trappes de type Esquire étaient souvent de faible qualité car trop faibles en intensité. En figure 4 est présenté le spectre MS/MS de l'ion de m/z=653,4 (2+), peptide trypsique de la BSA, obtenu d'une part sur la trappe Esquire 3000+ et d'autre part sur la trappe HCTPlus pour 300fmol de digest trypsique de BSA injectées. Les principaux ions sont présents sur les deux spectres mais sur le spectre MS/MS généré avec l'Esquire 3000+, l'intensité globale du spectre est beaucoup plus faible, les ions fragments minoritaires n'étaient donc pas détectés.

- **<u>La possibilité de paramétrer une valeur d'ICC indépendante en MS/MS</u>**

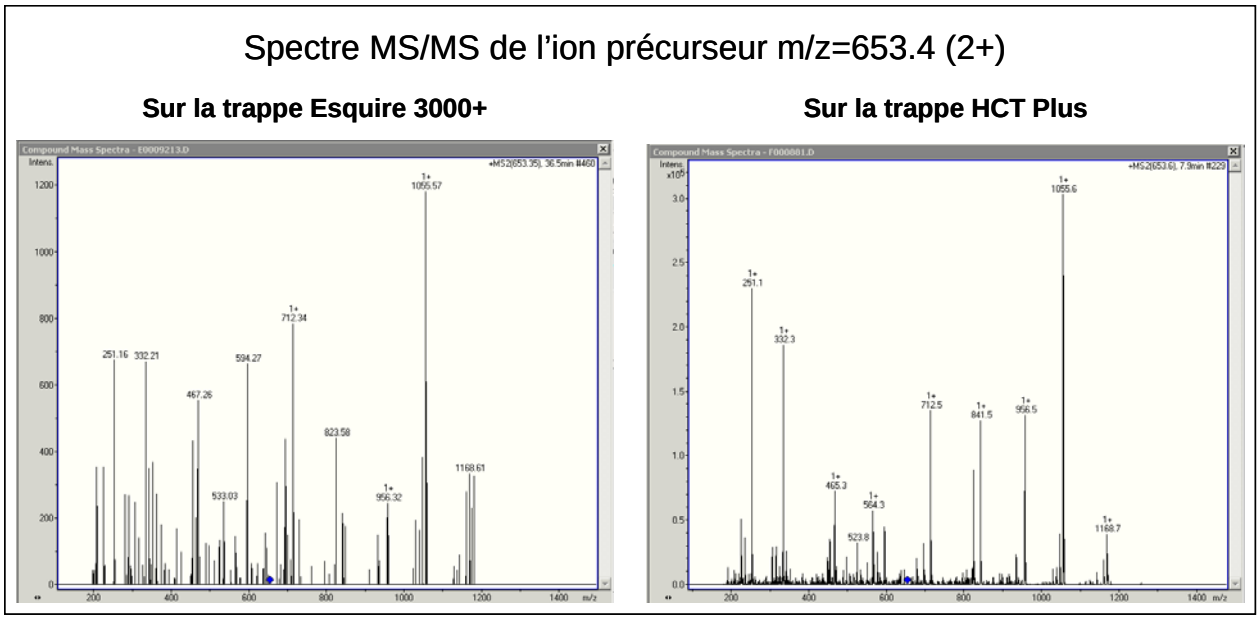
  Sur la trappe HCTPlus, l'ICC de la MS peut être paramétré à une valeur relativement basse pour limiter les phénomènes d'espace-charge et avoir une bonne précision et une bonne résolution sur la mesure de la masse de l'ion parent. Indépendamment de cette valeur d'ICC, la valeur d'ICC pour la MS/MS peut être fixée à une valeur beaucoup plus élevée de sorte à accumuler une quantité d'ions beaucoup plus importante avant l'isolation et la fragmentation. En effet, durant l'étape d'accumulation des ions avant l'isolation de l'ion parent, plus la quantité d'ions parents accumulés est importante, plus le nombre d'ions à fragmenter sera important et plus le spectre MS/MS sera informatif. Durant cette étape d'accumulation, les éventuelles pertes de résolution et décalages de la masse de l'ion parent dus aux effets d'espace-charge ne sont pas critiques à condition que la fenêtre d'isolation soit assez

large pour pouvoir englober le massif isotopique de l'ion même s'il est décalé par les effets d'espace-charge.

- **Le mode de balayage UltraScan**

  De la même manière, le spectre MS peut être réalisé en mode de balayage standard (mode « standard enhanced » à 8100 m/z/s) afin d'obtenir une résolution et une précision suffisantes pour pouvoir sans problème déterminer les états de charge 2+ ou 3+ des ions parents. Par contre, sur les spectres MS/MS des ions 2+ (voire 3+), les fragments sont principalement des ions monochargés, la résolution n'est donc plus capitale et le gain en sensibilité est privilégié : les spectres MS/MS sont donc généralement acquis en mode Ultrascan à 26000m/z/s.

Figure 4 : Comparaison du spectre MS/MS de l'ion parent de m/z=653.4 (2+), peptide trypsique de la BSA, obtenu sur la trappe Esquire 3000+ et la trappe HCTPlus suite à l'injection de 300fmol de digest trypsique de BSA.



Globalement la qualité des spectres MS/MS obtenus sur la trappe HCT est significativement plus élevée que celle des spectres précédemment obtenus sur l'Esquire 3000+. Les interprétations par séquençage *de novo* sont beaucoup plus aisées et la mauvaise réputation des instruments de type trappe ionique pour ce type d'applications par rapport aux instruments de type Q-TOF n'est plus justifiée au vu de la qualité des spectres MS/MS générés.

2 Le couplage nanoLC-MS/MS

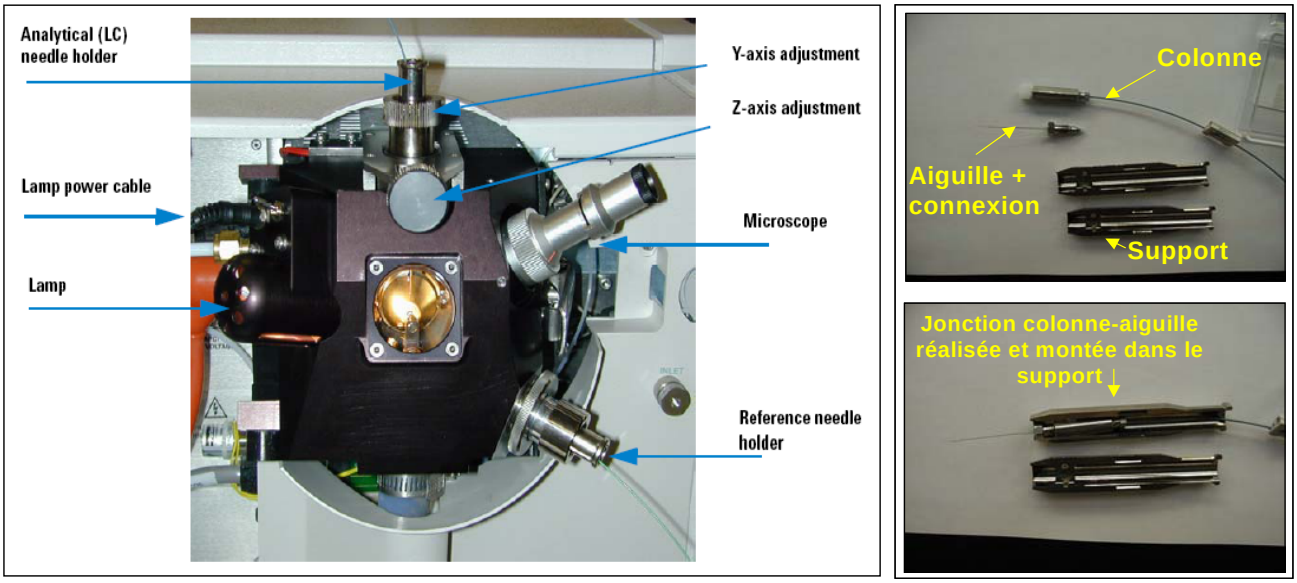
Dans cette partie, les résultats des optimisations effectuées pour la mise en place du couplage nanoLC-MS/MS pour l'analyse protéomique en routine seront brièvement décris.

2.1 Configuration du système

Le système nanoHPLC Agilent Technologies Series 1100 (Agilent Technologies) a été couplé à la trappe ionique HCTPlus (Bruker Daltonics). La source nanospray, Dual Nanospray ion source (G3253A Agilent Technologies) a été mise en place et le montage de cette source est décrit en figure 5. Les colonnes

chromatographiques de phase inverse de 75µm de diamètre interne (colonnes Zorbax 300SB-C18, 15cm x 75µm, particules de 5µm) sont connectées à l'aiguille de nanospray de 20µm de diamètre interne et 360µm de diamètre externe (PicoTip Emitter, New Objective, FS360-20-10-CE-20). L'ensemble est fixé dans un support avant d'être introduit dans la source nanospray de la trappe (Figure 5).
En amont de la colonne analytique, une pré-colonne d'enrichissement est utilisée sur ce système : Zorbax 300SB-C18, 5mm x 0.3 mm, particules de 5µm (Agilent Technologies).

Figure 5 : Photos de la source nanospray [Agilent G3253A Dual Nanospray Ion source, 2005], détails sur le système d'assemblage de l'aiguille nanospray et de la colonne chromatographique dans le support.



2.2 Optimisation des gradients de chromatographie

L'optimisation des gradients de chromatographie est aussi primordiale que l'optimisation des paramètres d'acquisition de la trappe. L'ionisation electrospay étant concentration-dépendante, l'obtention de pics chromatographiques les plus fins possibles est indispensable pour pouvoir avoir une sensibilité optimale. La qualité des spectres MS/MS sera directement dépendante de la quantité d'ions parents accumulés dans la trappe donc même si les paramétrages de la trappe sont optimaux, si les pics chromatographiques sont dispersés et la concentration des ions faible dans le spray, les signaux ne seront pas très bons. L'étape d'optimisation du gradient est donc importante et dépendra de la complexité du mélange. L'objectif ultime étant de séparer un maximum de pics et de parvenir à les sélectionner pour la MS/MS à leur intensité maximale.

La configuration du système nanoHPLC est telle que les volumes-morts et les volumes des capillaires reliant les différentes composantes du système sont relativement importants et causent quelques difficultés non négligeables. En effet, même avec des connexions garanties « zéro volume mort » reliant la sortie de la colonne chromatographique à l'aiguille de nébulisation, à un débit de 230nl/min, même un volume de quelques nl est critique et peut être à l'origine d'une dispersion et d'un élargissement des pics chromatographiques. Par ailleurs, les volumes des capillaires utilisés pour relier la sortie des pompes à la pré-colonne puis à la colonne sont relativement importants. Ces volumes importants induisent des délais

dans l'application du gradient sur la colonne et le temps d'équilibration du système doit être ajusté en conséquence. Ainsi, dans ces conditions, il s'agit de trouver le meilleur compromis entre une bonne séparation chromatographique et des temps d'acquisition pas trop longs.

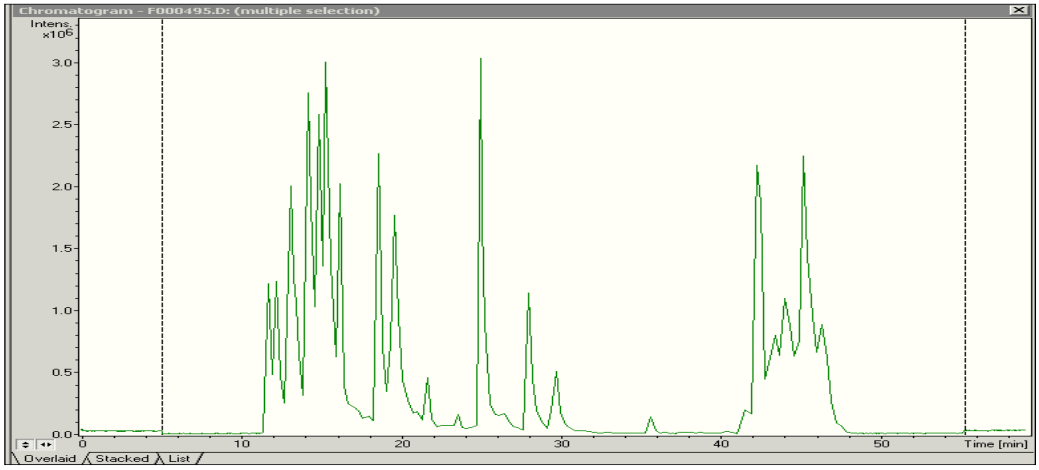
Le chargement des peptides sur la pré-colonne d'enrichissement est réalisé par une pompe capillaire à un débit de 30µl/min. Après le chargement de l'échantillon en tête de pré-colonne, celle-ci passe dans le circuit de la colonne et est balayée par le gradient de chromatographie délivré par la pompe nano à un débit de 230nl/min, en sens inverse de son chargement (backflush mode). Les peptides trypsiques sont pour la plupart élués entre 10 et 40% d'acetonitrile. La pente du gradient doit être optimisée afin d'obtenir la meilleure résolution chromatographique possible (c'est à dire, la meilleure finesse des pics et la meilleure séparation des pics). Le délai à prendre en considération entre l'application du gradient et son arrivée sur la colonne est d'environ 12min ce qui implique aussi un temps de rééquilibration du système non négligeable à considérer. Tous ces paramètres étant pris en considération, le tableau 1 présente le gradient de 60min qui a été optimisé pour des analyses d'échantillons protéomiques en routine. Cependant, il est bien évident qu'en fonction de la complexité et de la nature des échantillons analysés, ce gradient est ajusté individuellement.

Tableau 1 : Gradient chromatographique optimisé pour des analyses protéomiques en routine sur le couplage nanoHPLC Agilent / Trappe ionique HCTPlus

Temps (min)	%ACN	
0	15	Chargement de l'échantillon sur la pré-colonne
5	15	
33	35	Gradient d'élution des peptides
34	90	
39	90	
40	15	Reéquilibration du système
60	15	

Un exemple de chromatogramme obtenu suite à l'injection de 300fmol de digest de BSA dans ces conditions d'analyse est présenté en figure 6. Les pics chromatographiques font environ 30 à 45s de large et les paramètres d'acquisition de la trappe décrits dans les paragraphes suivants seront optimisés pour des pics chromatographiques de cette largeur.

Figure 6 : Exemple de chromatogramme obtenu pour l'injection de 300fmol de digest de BSA en couplage nanoLC-MS/MS avec le gradient du tableau 1.

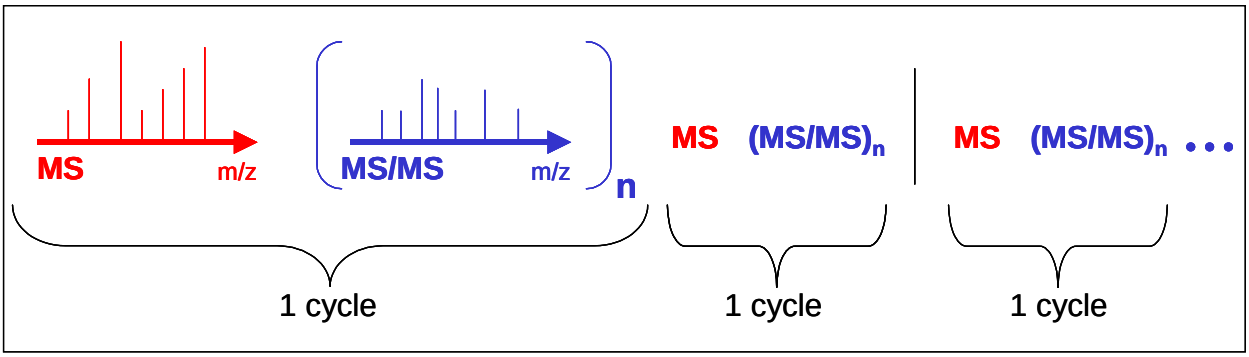


2.3 Optimisation des paramètres d'acquisition en mode MS-MS/MS automatique

2.3.1 Les cycles d'acquisition en mode MS-MS/MS automatique

En mode d'acquisition MS/MS automatique, la sélection des ions parents pour la fragmentation est réalisée de manière automatique selon des critères optimisés par l'utilisateur. L'acquisition fonctionne par cycles MS suivie d'une ou de plusieurs MS/MS sur des ions parents sélectionnés sur le spectre MS (Figure 7).

Figure 7 : Fonctionnement par cycle du mode d'acquisition MS-MS/MS automatique : acquisition d'un spectre MS sur lequel sont sélectionnés le ou les ions parents avant d'être isolé(s) puis fragmenté(s) pour la MS/MS.



Dans nos conditions d'analyse, les pics chromatographiques font entre 30 et 45s. Or le nombre d'ions précurseurs sélectionnés conditionne la durée d'un cycle et la durée d'un cycle doit être adaptée à la largeur des pics d'élution des peptides séparés sur la colonne afin d'obtenir un maximum d'informations de fragmentation.

Les paramétrages décrits dans les paragraphes suivants influencent directement les durées d'acquisition des cycles et doivent donc être optimisés.

En général, trois ions précurseurs par spectre MS/MS sont sélectionnés pour la méthode d'analyse protéomique de routine.

2.3.2 Le mode SPS (Smart Parameter Setting)

Les paramètres d'acquisition de la trappe doivent être optimisés en fonction de la masse de l'ion parent sélectionné : le réglage du « trap drive » est par exemple critique pour le piégeage optimal d'un ion de m/z donné comme expliqué en 1.1.1. Or en mode d'acquisition automatique, cette optimisation ne peut être réalisée en direct par l'expérimentateur. Dans ce cas, la fonction SPS, Smart Parameter Setting, permet une optimisation automatique des paramètres d'acquisition de la trappe, notamment du trap drive, systématiquement en fonction de l'ion parent sélectionné. L'optimisation automatique n'est pas aussi fine qu'une optimisation réalisée manuellement mais elle permet néanmoins une amélioration considérable de la qualité des résultats obtenus.

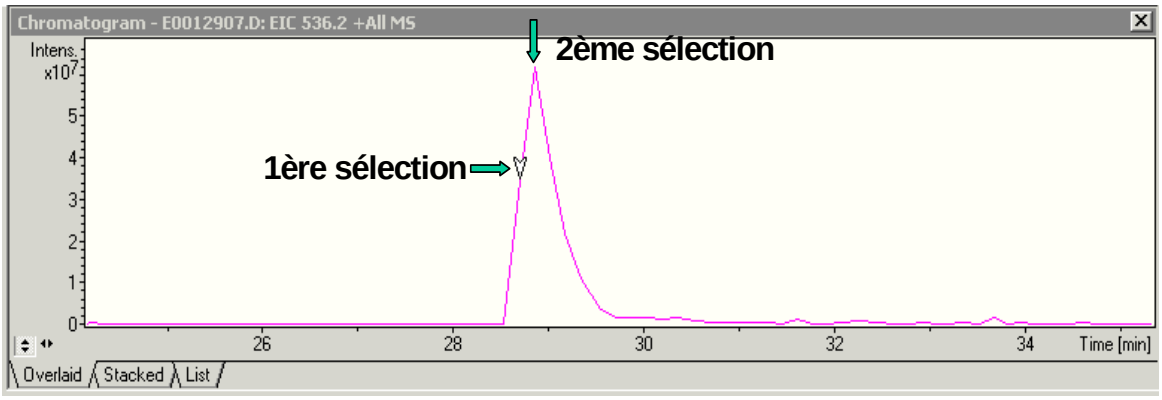
2.3.3 Le seuil de sélection des ions parents et l'exclusion temporaire

L'intensité des ions détectés doit être supérieure à une valeur seuil afin que ceux-ci puissent être sélectionnés comme ions parents pour la MS/MS. Ce seuil doit être finement optimisé car il conditionnera le moment de la sélection des ions parents. Lorsque ce seuil est placé trop bas, les ions parents sont sélectionnés dès qu'ils dépassent le seuil ce qui ne correspond pas forcément à leur maximum d'intensité. Ce n'est donc pas à ce moment que leur fragmentation sera la plus informative. En revanche, lorsque le seuil de sélection est placé trop haut, le risque de perdre des informations sur des ions minoritaires est important.

Par ailleurs, il est possible d'exclure temporairement ou définitivement des ions parents de la sélection pour la MS/MS une fois qu'ils ont été fragmentés. Cette option permet d'éviter la sélection multiple d'un ion qui serait largement majoritaire dans une analyse et empêcherait ainsi la sélection d'ions minoritaires. Cette exclusion doit cependant être utilisée avec prudence car elle implique toujours l'exclusion d'une fenêtre de 2 à 3 Da par ion exclu. Or plusieurs peptides trypsiques de séquence différente peuvent avoir la même masse et il faut éviter de perdre des informations sur ces peptides parce qu'ils seraient exclus de la sélection. Il est donc souvent plus avantageux d'utiliser cette exclusion de façon temporaire, c'est à dire d'exclure les ions pendant 2min puis de permettre ensuite leur re-sélection. En effet, même s'il arrive que des peptides aient la même masse, il est plus rare qu'ils aient la même masse et aient en plus le même temps de rétention chromatographique.

Ainsi, prenant en compte le seuil de sélection et la possibilité d'exclusion temporaire, un compromis doit être trouvé entre perdre le moins d'informations sur des ions minoritaires et obtenir le plus d'informations sur les ions sélectionnés. Pour cela, la solution qui est apparue comme fournissant les meilleurs résultats est la suivante : placer le seuil de sélection à une valeur basse et autoriser la sélection du même ion parent deux fois de suite. Cette solution permet de ne pas exclure de la sélection des ions minoritaires. Et d'autre part, le fait d'autoriser la sélection du parent deux fois permet de le fragmenter en général une première fois au début de son pic chromatographique et le temps d'un cycle passant, lors de la deuxième sélection, il est sélectionné près du sommet du pic chromatographique (Figure 8).

Figure 8 : Chromatogramme d'ion extrait (EIC) pour l'ion m/z=536,2. L'optimisation de la sélection des ions parents a permis de le sélectionner une première fois, puis une seconde fois à son intensité maximale.



2.3.4 Le choix des ions parents sélectionnés pour la MS/MS

Les peptides issus de la digestion trypsique sont généralement doublement (espèces majoritaires à 65-80% [Savitski, et coll., 2005]) ou triplement chargés lorsqu'ils sont analysés en mode d'ionisation ESI positif : ils présentent au moins deux sites de protonation, un sur l'amine N-terminale et l'autre sur la chaîne latérale de la lysine ou de l'arginine du côté C-terminal. Les ions doublement et triplement chargés fragmentant mieux que les ions monochargés [Dongré et coll., 1996] (voir Partie bibliographique, chapitre I.1.3), une option permettant de « préférer » les ions doublement chargés peut être sélectionnée. Les ions doublement chargés seront alors sélectionnés de façon privilégiée, même s'ils sont moins intenses sur le spectre MS, par rapport aux ions monochargés. Des ions monochargés ne seront sélectionnés que lorsqu'aucun ion doublement chargé présent sur le spectre ne dépassera le seuil de sélection. La possibilité d'exclure totalement les ions monochargés est également ouverte mais n'est pas utilisée puisque la fragmentation des ions monochargés peut malgré tout apporter des informations s'ils ne sont pas sélectionnés au détriment des ions doublement chargés.

2.3.5 Le mode de balayage

Comme discuté précédemment, le gain en sensibilité apporté par le mode balayage Ultrascan le place en mode de balayage privilégié pour l'acquisition des spectres MS/MS. Dans les spectres MS/MS, la plupart des ions étant monochargés, une résolution maximale n'est pas prioritaire. Par contre, la sensibilité étant un facteur critique pour la qualité de la MS/MS, on favorisera donc le gain en sensibilité par rapport à la perte de résolution.

2.3.6 Le nombre de spectres moyennés

Sur les instruments de type trappe ionique, chaque spectre apparaissant à l'écran est un spectre moyen de plusieurs spectres dont le nombre est défini par l'utilisateur. Plus le nombre de spectres moyennés sera grand et plus le spectre résultant sera représentatif du signal étudié (le rapport signal/bruit de chacun des pics est accru). Cependant, la multiplication du nombre de spectres moyennés augmente le temps d'acquisition. Il faut donc trouver un compromis satisfaisant à la fois nos conditions d'analyse chromatographique et la qualité de la représentation du signal.

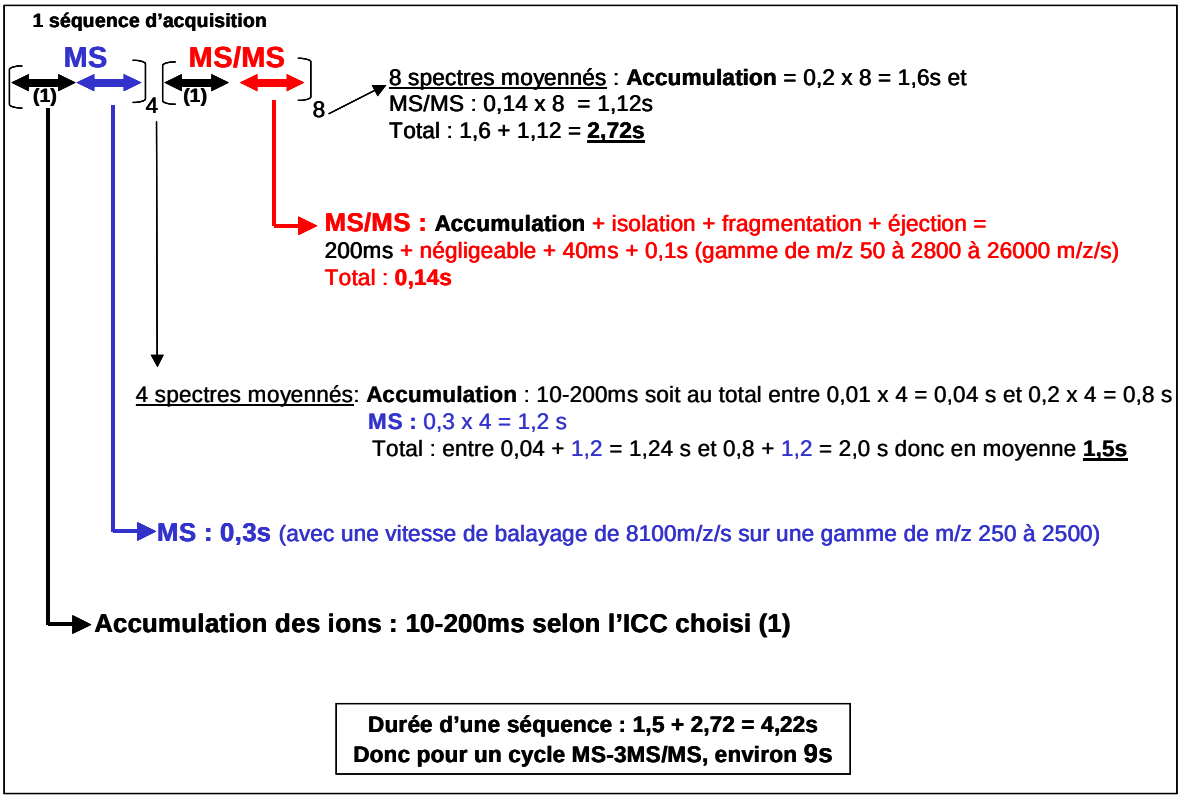
Des tests d'ajustement des nombres de spectres moyennés en MS et en MS/MS ont été réalisés et le

compromis qui a été choisi en mode MS est de 4 et en mode MS/MS de 8.

2.3.7 Détail sur la durée d'acquisition d'un cycle MS-MS/MS

En tenant compte des conditions et des paramètres précédemment décrits (vitesse de balayage, nombre de spectres moyennés), la figure 9 résume en détail les contributions en temps des différentes étapes dans un cycle d'acquisition. Globalement, l'acquisition d'un cycle complet, 1 MS suivie de 3 MS/MS prend environ 9s ce qui permet d'acquérir entre 250 et 300 spectres MS/MS durant le gradient de l'analyse nanoLC-MS/MS (environ 40min, Tableau 1).

Figure 9 : Détail sur la durée d'un cycle d'acquisition MS-3MS/MS dans les conditions d'analyse et avec les paramétrages de vitesse de balayage et de nombre de spectres moyennés décrits précédemment.



3 Conclusions

L'objectif de ce chapitre était d'illustrer l'importance de l'optimisation à la fois des paramètres d'acquisition de la trappe mais aussi des conditions chromatographiques car les deux sont étroitement liés. Si les conditions chromatographiques ne sont pas bonnes, les résultats des analyses ne pourront pas être bons malgré un très bon réglage de l'analyseur trappe. Et inversement, même avec une excellente séparation chromatographie, lorsque l'analyseur est mal paramétré, les résultats en seront affectés.

En réalité, il s'agit toujours de trouver le meilleur compromis entre l'ensemble des facteurs influant sur la qualité des résultats, les optimisations ne sont pas figées, il faut les adapter constamment en fonction de la nature et de la complexité des échantillons à analyser. Ainsi, l'objectif de ce chapitre n'était pas d'établir une méthode de base mais plutôt de soulever et d'expliquer les facteurs critiques auxquels il faut porter une particulière attention pour l'obtention de bons résultats de nanoLC-MS/MS avec ce type de couplage.

CHAPITRE II

Les nouveautés instrumentales du couplage : le système nanoHPLC-Chip Cube et la fragmentation ETD

De nouvelles composantes ont été installées sur ce couplage nanoHPLC Series 1100 (Agilent Technologies)/HCTPlus (Bruker Daltonics) en début d'année 2006: un système nanoHPLC-Chip Cube (Agilent technologies) et la fragmentation ETD (Electron Transfer Dissociation) grâce à l'acquisition d'une trappe HCTUltra PTM Discovery System (Bruker Daltonics).

1 Le système nanoHPLC-Chip Cube (Agilent Technologies)

La technologie nanoHPLC-Chip Cube étant encore à ses débuts, des tests de performances, de robustesse du système et une optimisation systématique des gradients et des paramétrages d'acquisition des données sur la trappe ont été réalisées afin d'établir une méthode d'analyse en routine pour des échantillons de protéomique. Un résumé des résultats obtenus et les conclusions quant aux avantages mais aussi aux limites de ce système sont brièvement décrits dans ce chapitre.

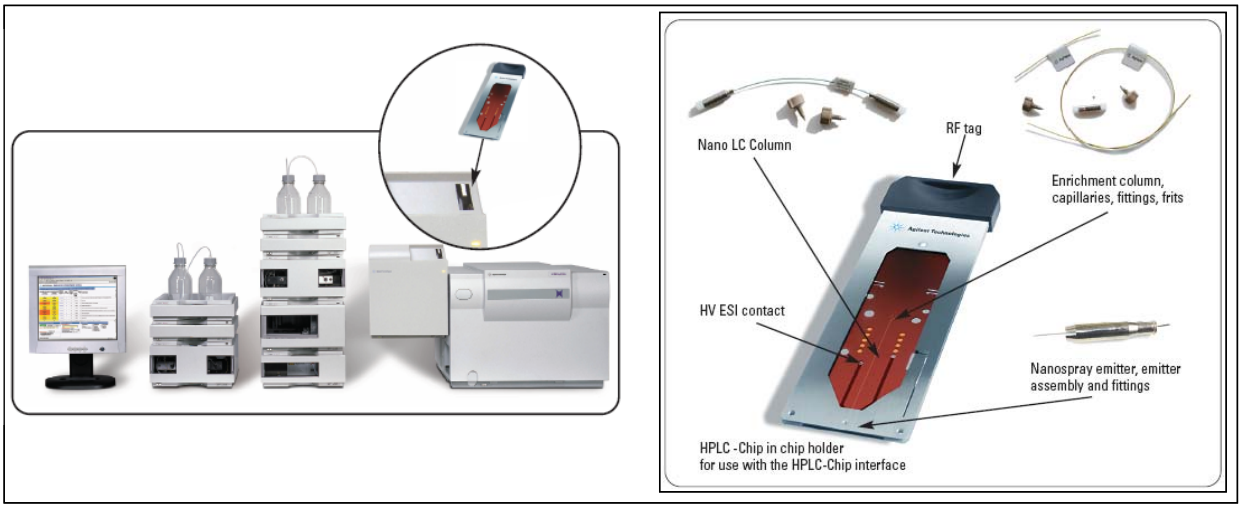
1.1 Configuration du système

1.1.1 L'interface nanoHPLC-Chip cube

Dans la suite de ce chapitre, le terme chip sera utilisé pour décrire la partie réutilisable et amovible de l'interface HPLC-Chip-Cube-MS intègrant une pré-colonne d'enrichissement, une colonne de séparation, le nébulisateur nanospray et l'ensemble des connections nécessaires (Figure 1). Ces chips sont constituées de films de polyimide laminés à l'intérieur desquels des canaux sont creusés par ablation-laser. Le volume de capillaires est limité dans ce système, ce qui limite les possibilités de fuites et élimine les volumes morts post-colonne [Yin et coll., 2005 ; Vollmer et coll., 2005a]. Ce sont ces volumes morts post-colonnes qui sont en grande partie responsable de la dispersion des pics chromatographiques dans les montages nanoHPLC classiques.

Par ailleurs, le fait de ne plus avoir à remplacer fréquemment les aiguilles de nébulisation, à optimiser le positionnement du spray, … mais de simplement avoir à introduire la chip dans le cube apporte à l'expérimentateur un confort d'utilisation notable.

Figure 1 : Figures extraites de la note d'application Agilent [Vollmer et coll., 2005b] illustrant la configuration de l'interface nanoHPLC-Chip cube-MS . L'ensemble des éléments nécessaires pour le système nanoHPLC classique est regroupé dans la Chip : la pré-colonne d'enrichissement, la colonne de séparation et le montage de nanospray (holder + aiguilles).



1.1.2 Les différentes chips disponibles

Des chips avec différents types de colonnes sont disponibles à l'heure actuelle : des colonnes de phase inverse courtes (43mm, 75µm) et longues (150mm, 75µm), des colonnes carbone-graphite pour la séparation des glycopeptides, des chips à infusion (pour l'injection directe au pousse-seringue sans avoir à remettre en place la source ESI classique), des chips de calibration et de diagnostic et des chips de séparation de petites molécules.

1.1.3 La traçabilité des chips

L'ensemble des informations relatives à l'utilisation d'une chip est enregistré grâce à un système de puce et code barre : des informations telles que le temps d'utilisation des chips, le nombre d'injections réalisées, … peuvent être consultées à n'importe quel moment ce qui facilite le suivi de l'utilisation des chips.

1.1.4 La configuration du balayage de la pré-colonne

La configuration du système permet de balayer la pré-colonne d'enrichissement soit en mode « forward-flush » c'est à dire dans le même sens que le chargement soit en mode « backflush » c'est à dire dans le sens inverse du chargement [Vollmer et coll., 2005b]. Les performances chromatographiques n'étant pas dégradées lorsque la pré-colonne est balayée en mode « forward-flush » vu le faible volume de la pré-colonne (40nl), l'utilisation de ce mode de balayage est préconisée. Le fait de travailler en mode « forward-flush » est plus robuste car cela permet d'éviter que d'éventuels contaminants qui auraient été accumulés en tête de pré-colonne ne soient directement élués sur la colonne au moment du basculement. Néanmoins le

passage en mode « backflush » après un certain nombre d'analyses effectuées sur la chip permet d'allonger la durée de vie de la chip.

1.2 Optimisation des gradients

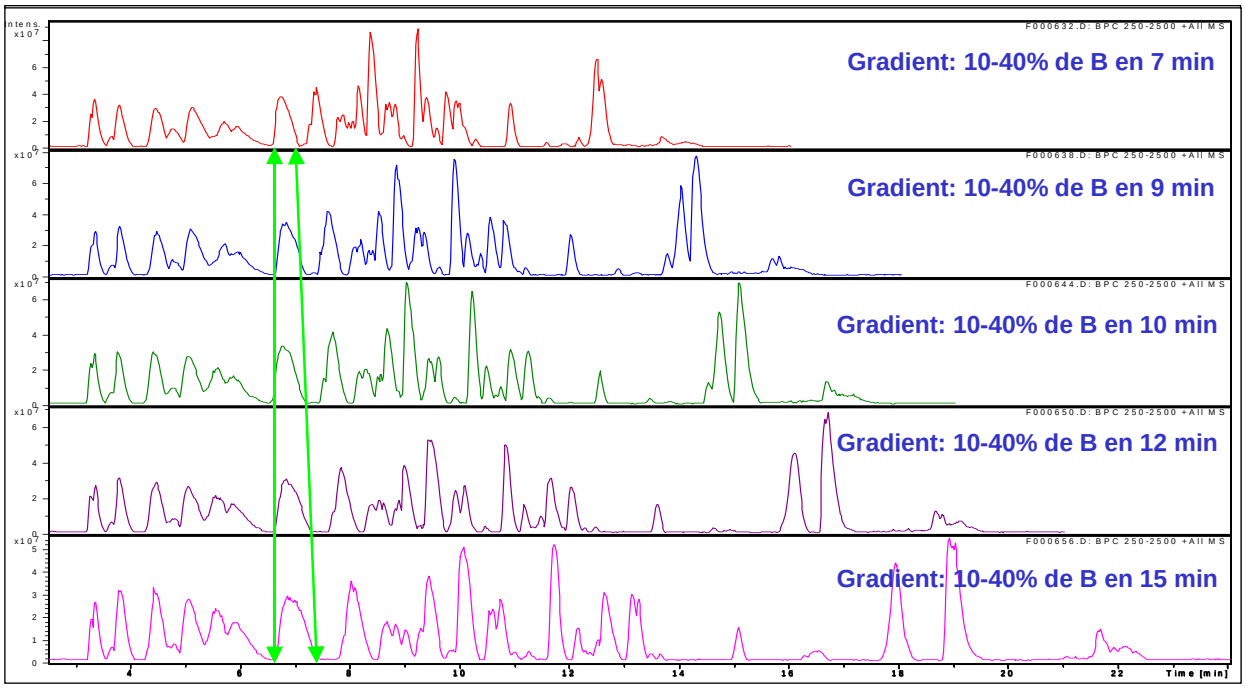
Les principaux avantages de la configuration de ce système proviennent de la réduction considérable des volumes morts post-colonne et des volumes de capillaire. Cette configuration permet en effet :

- **Une réduction considérable de la dispersion des pics chromatographiques donc l'obtention de pics beaucoup plus fins**. Les jonctions liquides, même garanties « zéro-volume mort », permettant la connexion de la colonne chromatographique à l'aiguille de nébulisation dans le système nanospray classique induisent des volumes morts post-colonnes critiques, à l'origine de la dispersion des pics chromatographiques. A des débits de 200-300nl/min et pour des pics d'élution équivalent à des volumes de moins de 100nl, le moindre volume mort est critique. Ainsi, avec les chips, la largeur des pics chromatographiques a pu être réduite à 10s par rapport aux pics de 30 à 45s obtenus avec les systèmes nanoHPLC classiques.
- **Une application du gradient plus rapide.** Le chemin de capillaires à parcourir étant plus court, l'application du gradient sur la colonne est plus rapide qu'avec la configuration nanoHPLC-MS classique : un retard d'environ 15min entre l'application du gradient et l'élution correspondante du pic a pu être réduit à 2min avec le système nanoHPLC-chip-cube.
- **Une réduction du temps d'équilibration.** Le temps d'équilibration nécessaire au reconditionnement de la colonne est également réduit : avec le système classique, un temps d'équilibration de 20min était nécessaire alors qu'avec le système HPLC-chip, ce temps a pu être réduit à 5min. Cette durée d'équilibration réduite contribue significativement à la réduction globale du temps d'une analyse.

Dans la configuration actuelle, le système est conçu pour délivrer un débit régulier et un spray stable entre 100nl/min et 400nl/min sans nécessité d'une assistance de gaz de nébulisation [Yin et coll., 2005]. Plusieurs débits ont été testés et une comparaison des résolutions chromatographiques, impliquant à la fois finesse des pics et séparation entre deux pics consécutifs a permis de fixer un débit optimal à 300nl/min. Avec des débits plus élevés (400-600nl/min), certains pics chromatographiques sont plus fins ce qui s'explique par une dispersion post-colonne des pics réduite et des effets de diffusion longitudinale moins importants à débit plus élevé (équation de Van Deemter). Cependant la résistance au transfert de masse, également décrite par l'équation de Van Deemter, conduit à un élargissement des pics chromatographiques avec des débits plus élevés. Le débit optimal de 300nl/min a été fixé pour l'ensemble des analyses suivantes.

Les peptides trypsiques étant élués pour la grande majorité entre 10 et 40% d'acetonitrile, des gradients linéaires entre 10%-40% d'acetonitrile ont été testés avec des pentes variables (Figure 2). Une pente ralentie permet théoriquement de séparer plus de peptides et d'avoir une meilleure résolution chromatographique mais lorsque la pente du gradient est trop ralentie, les pics chromatographiques finissent par s'élargir entraînant une perte en sensibilité (Figure 2).

Figure 2 : Visualisation d'une étape d'optimisation des gradients illustrant l'effet du ralentissement de la pente du gradient sur la résolution chromatographique.



Toutes les observations et tests réalisés ont permis d'aboutir à un gradient optimisé pour les analyses protéomiques en routine (Tableau 1). L'échantillon est chargé sur la pré-colonne d'enrichissement par une pompe auxiliaire à 4µl/min. Une fois le chargement de l'échantillon terminé, la pré-colonne passe dans le circuit de la colonne et à un débit de 300nl/min, un gradient linéaire de 8 à 40% d'acetonitrile en 7min est appliqué suivi d'un passage à 70% d'acetonitrile en 2min et d'un pallier à 70% d'acetonitrile durant 1min. Puis 5 min d'équilibration du système à 8% d'acetonitrile sont réalisées.

Tableau 1 : Gradient optimisé pour des analyses protéomiques sur le système nanoHPLC-Chip cube.

Temps (min)	%ACN
3	8%
10	40%
12	70%
13	8%
18	8%

1.3 Des temps d'analyse plus courts donc adaptation de l'acquisition MS-MS/MS

Les pics chromatographiques obtenus avec le système nanoChip étant beaucoup plus fins, un réajustement des paramètres d'acquisition sur la trappe a du être réalisé. Comme illustré dans le chapitre précédent, l'acquisition durant des analyses nanoLC-MS/MS se fait par des cycles MS suivie d'une ou de plusieurs MS/MS. L'objectif à atteindre est toujours le même : régler la sélection des ions parents de sorte à pouvoir sélectionner un maximum d'ions parents et à les sélectionner lorsqu'ils sont à leur intensité maximale donc au sommet du pic chromatographique.

En mode MS, le mode de balayage « standard enhanced » à 8100m/z/s est maintenu car le passage en mode ultrascan entraînerait une trop grande perte de résolution et de précision dans la mesure de la masse des ions parents. En mode MS/MS le mode de balayage ultrascan est maintenu. Par contre, le gain en sensibilité permet de réduire le nombre de spectres moyennés par spectre affiché et donc par la même le temps d'acquisition d'un cycle MS suivie de trois MS/MS. En mode MS, le nombre de spectres moyennés est maintenu à 4 par contre, en mode MS/MS, il a également pu être réduit à 4 (au lieu de 8) sans pertes d'informations.

1.4 Evaluation de la sensibilité du système

La sensibilité du système nanoHPLC-Chip cube a été évaluée grâce à l'injection d'un mélange de peptides trypsiques de digestion de BSA (Bovine Serum Albumin) en quantité croissante de 10fmol à 100fmol (Figures 3). Six peptides trypsiques appartenant à la BSA ont pu être identifiés sans difficulté suite à l'injection de 10fmol de digest. Ces tests de sensibilité ont été réalisés en nanoLC-MS mais aussi en nanoLC-MS/MS de sorte à pouvoir lancer ensuite des recherches Mascot avec les fichiers .mgf générés et de pouvoir juger du recouvrement de séquence et de la qualité des spectres MS/MS obtenus sur le digest de BSA (Tableau 2).

Figure 3 : A/ Evaluation de la sensibilité des chips contenant une pré-colonne d'enrichissement (4mm, 40 nl, Zorbax 300SB-C18, 5µm) et une colonne analytique de 75 µm de diamètre interne (43 mm, Zorbax 300SB-C18, 5µm). Des quantités croissantes, de 10 fmol à 90 fmol, de peptides trypsiques de BSA ont été injectées. B/ Zoom sur deux chromatogrammes d'ions extraits pour les peptides de m/z=927,4 et de m/z=722,7.

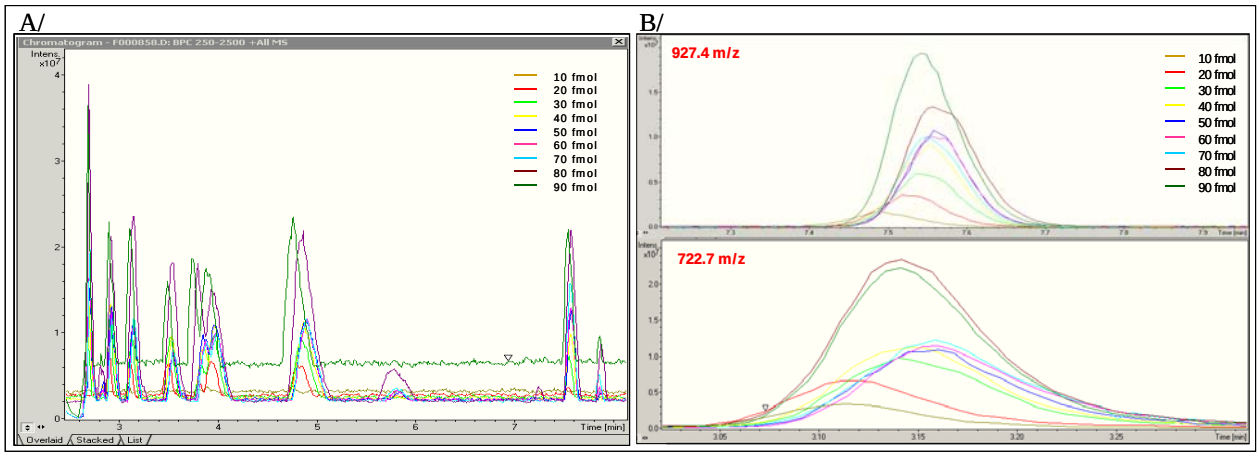


Tableau 2 : Pourcentage de recouvrement MS/MS obtenu et nombre de peptides identifiés pour des quantités croissantes de digest de BSA injectées.

Quantité de BSA injectée en fmol	Pourcentage de recouvrement moyen	Nombre de peptides moyen identifiés
10	5	6
20	14	12
30	19	15
40	19	15
50	20	16
60	21	17
70	24	19
80	24	20

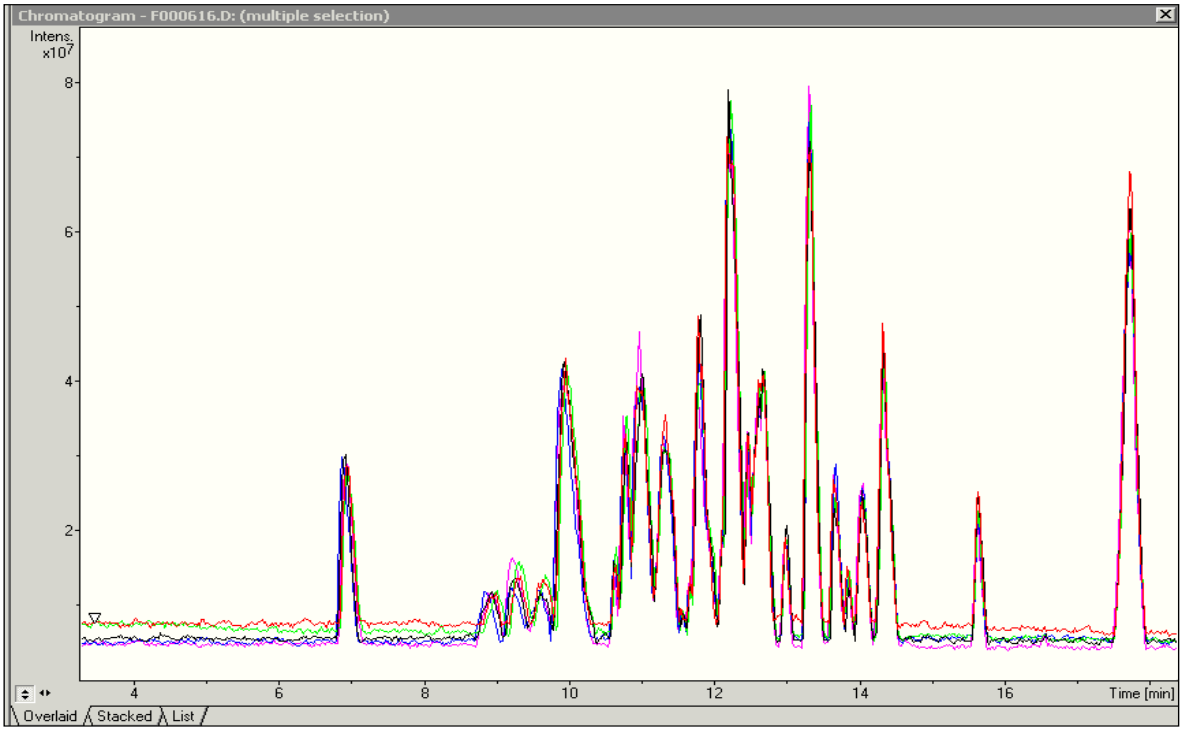
Un gain en sensibilité d'un facteur 4-5 a pu être observé avec le système nanoHPLC-chip par rapport au système nanoHPLC classique. Ce gain en sensibilité s'explique principalement par les pics chromatographiques qui sont beaucoup plus fins : les pics étant plus fins, la concentration des peptides est plus élevée dans le spray et le signal de masse sera donc plus intense.

En mode MS/MS, le gain en sensibilité est également significatif et une nette amélioration des scores Mascot obtenus en MS/MS pour les peptides trypsiques de la BSA a pu être observée.

1.5 La reproductibilité du système

La reproductibilité du système a été évaluée par des injections successives d'un mélange de peptides trypsiques issus de la digestion de BSA. La figure 4 présente la superposition des chromatogrammes obtenus pour 5 injections successives de 100 fmol de digest.

Figure 4 : Superposition de 5 chromatogrammes de 5 injections successives de 100 fmol d'un digest de BSA.



1.6 Avantages et limites du système

1.6.1 Les avantages

- **Des temps d'analyse raccourcis**

Avec le système nanoHPLC classique avec des colonnes de 75 µm de diamètre interne, les temps d'analyse étaient d'environ 60min par échantillon. Le fait de pouvoir réduire ce temps d'une analyse à moins de 20min permet de tripler le nombre d'échantillons analysés. Ceci présente un avantage considérable pour les projets pour lesquels un nombre important d'analyses est à réaliser.

- **Un gain en sensibilité**

Le gain en sensibilité qu'apporte l'utilisation des chips ouvre de nouvelles possibilités. En effet, une partie des extraits peptidiques des spots de gel (qui étaient injectés dans leur intégralité avec le système classique), peut à présent être conservée pour des analyses complémentaires. Il est par exemple souvent utile de réinjecter l'échantillon en modifiant le gradient pour permettre une meilleure séparation des peptides dans une zone précise du gradient ou d'utiliser un programme d'exclusion des peptides préalablement identifiés et ainsi pouvoir sélectionner et fragmenter des espèces plus minoritaires. Combiné avec le gain de temps d'analyse, le gain en sensibilité permet donc, dans un temps donné, de réaliser des analyses complémentaires et plus poussées des échantillons.

- **Une meilleure qualité de MS/MS**

Le gain en sensibilité observé se répercute également sur la qualité des spectres MS/MS. Les pics chromatographiques étant plus fins, les ions sélectionnés sont, au moment de leur sélection, beaucoup plus intenses. Or plus le nombre d'ions accumulés dans la trappe sera élevé, plus la qualité du spectre MS/MS sera grande. Des spectres MS/MS plus informatifs seront également plus facilement interprétables par séquençage *de novo*.

- **Un confort pour l'expérimentateur**

Le fait de ne plus avoir à utiliser les systèmes d'aiguilles et les montages délicats du couplage nanoLC-MS/MS classique est un réel confort pour l'utilisateur. Auparavant, la durée de vie des aiguilles de nébulisation était très variable et leur robustesse un peu aléatoire. Un temps non négligeable était donc consacré quotidiennement aux réglages du spray, au remplacement des aiguilles bouchées, délicates à casser et à mettre en place, … La mise en place de la chip ne nécessite aucun réglage et le spray est stable et robuste sur une longue période lorsque le système est propre.

1.6.2 Les limites

Malgré tous ces avantages qu'apportent ce système, il présente toutefois aussi ses limites dans l'état actuel.

- **La présence de contaminants**

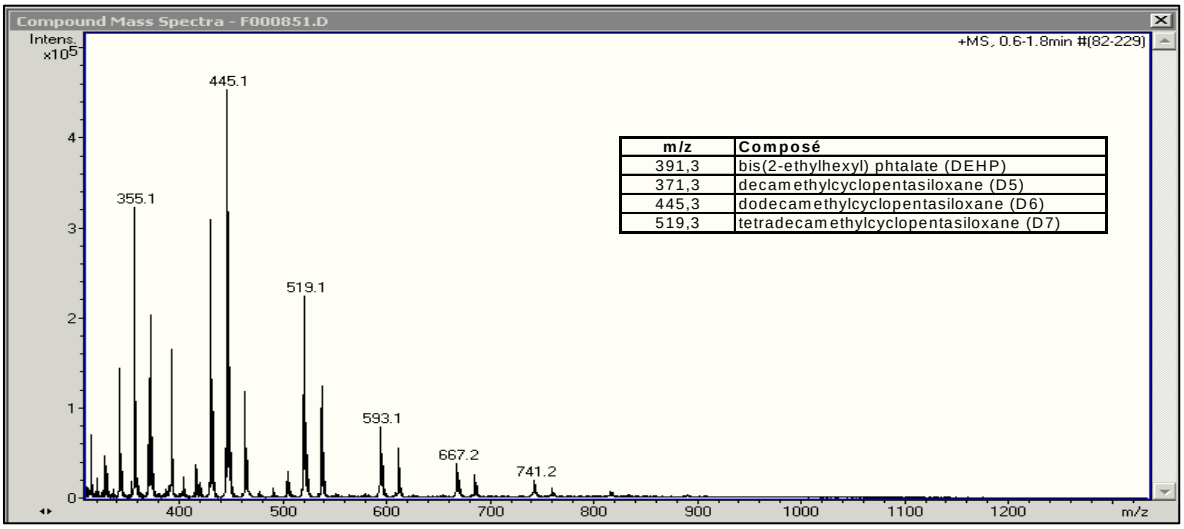
Une limite du système est la présence permanente de contaminants sur les spectres. Ces contaminants sont des composés du type phtalate ou siloxane et ont des m/z compris entre 300 et 600, ce qui ne les exclus pas de la gamme de masse couverte par les peptides trypsiques. Un exemple de spectre illustrant ces contaminants est présenté en figure 5.

La présence de ce bruit de fond était annoncée mais son intensité ne devait pas dépasser une intensité équivalente à un signal d'un peptide à 1fmol [Yin et coll., 2005]. Or ces contaminants se révèlent être bien plus intenses qu'annoncés (une intensité de 10^5 est plutôt équivalente à un peptide à 80 fmol) et leur présence réduit inévitablement la sensibilité des mesures car ces composés participent, au même titre que les peptides élués à la compétition d'ionisation dans le spray.

Ces contaminants, et particulièrement les polysiloxanes proviennent des revêtements anti-feu des circuits électroniques de la trappe. Et comme la géométrie du cube est telle que la source est ouverte, ces composés volatiles sont attirés dans la chambre d'ionisation, ionisés et analysés en continu dans la trappe.

La solution préconisée par la société Agilent est la fermeture de la source par un système de balai-brosse afin de limiter l'entrée de ces contaminants mais se pose alors le problème de stabilité du spray. En effet, la présence d'oxygène dans la chambre d'ESI stabilise le spray en évitant des phénomènes de charge à la pointe de l'aiguille de nébulisation dus à l'accumulation des ions positifs. Ainsi, le moyen de limiter à la fois le bruit de fond et de maintenir la stabilité du spray est d'introduire un pourcentage d'oxygène dans le gaz de désolvatation. Un système de mélange air-azote est en cours d'installation au laboratoire.

Figure 5 : Spectre illustrant la présence d'un bruit de fond permanent sur le système nanoHPLC-Chip Cube ainsi que la liste des composés caractérisant les principaux contaminants.



m/z	Composé
391,3	bis(2-ethylhexyl) phtalate (DEHP)
371,3	decamethylcyclopentasiloxane (D5)
445,3	dodecamethylcyclopentasiloxane (D6)
519,3	tetradecamethylcyclopentasiloxane (D7)

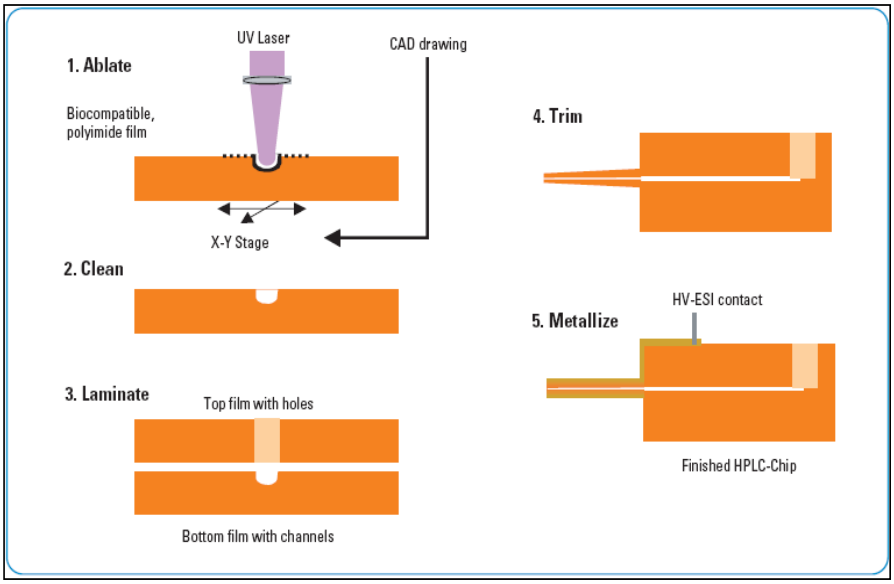
- **La maintenance des pompes nanoHPLC**

Le fonctionnement du système, tel qu'il est en place, ne permet pas de maintenir les pompes allumées lorsque la chip est retirée. Or il est bien connu que les arrêts du débit sur des pompes nano peuvent perturber l'état général de fonctionnement des pompes, les systèmes utilisant des nano-débits étant très sensibles. Il faut donc malgré tout limiter au maximum les temps d'arrêt des pompes nano.

- **La fabrication et la robustesse des chips**

La fabrication des chips est encore en cours d'optimisation notamment en ce qui concerne le revêtement des pointes de nébulisation. Comme illustré en figure 6, la dernière étape de fabrication des chips consiste en la métallisation de la pointe de nébulisation. C'est cette métallisation qui permettra l'application de la haute tension sur la pointe de nébulisation et donc la formation du spray. Cette étape est critique pour la robustesse et la stabilité du spray. Le platine utilisé pour la métallisation des chips de première génération a été remplacé par l'or ce qui a donné plus de robustesse à ces dernières chips.

Figure 6 : Les différentes étapes de fabrication des chips [Application note, Agilent Technologies, 2006, 5989-5492EN].



- **Les chips multi-pré-colonnes**

Même en travaillant en mode « forward-flush », les chips se voient souvent dégradées et inutilisables en raison d'une pré-colonne défectueuse quand bien même les performances de la colonne ne sont pas altérées. Ainsi, la conception de chips comportant plusieurs pré-colonnes serait très utile et permettrait d'augmenter significativement la durée de vie des chips. Ces chips multi-précolonnes sont en cours de développement.

1.7 Conclusions

Le laboratoire a fait l'acquisition de l'interface HPLC-Chip Cube au mois de février 2006 et durant ces quelques mois de travail et d'optimisation sur le système, les résultats obtenus sont très prometteurs. Malgré les limites soulignées dans ce dernier paragraphe, globalement le système nous a permis de gagner en sensibilité, d'obtenir des spectres MS/MS de meilleure qualité et d'augmenter le débit des analyses. Par ailleurs, le confort de travail de l'utilisateur est considérablement amélioré.

2 La dissociation par transfert d'électron (ETD)

Une trappe HCT Ultra PMT Discovery System (Bruker Daltonics) a été mise en place au laboratoire en février-mars 2006. Cette trappe contient une source à ionisation chimique permettant de réaliser des fragmentations de peptides ou de protéines par dissociation par transfert d'électron (ETD=Electron Transfer Dissociation). La technique ETD, les mécanismes de fragmentation et les premiers résultats que nous avons pu obtenir sur cet instrument seront brièvement décrits dans ce chapitre. Puis je ferai un point sur les avantages et les limites que nous avons pu constater sur le système dans l'état actuel.

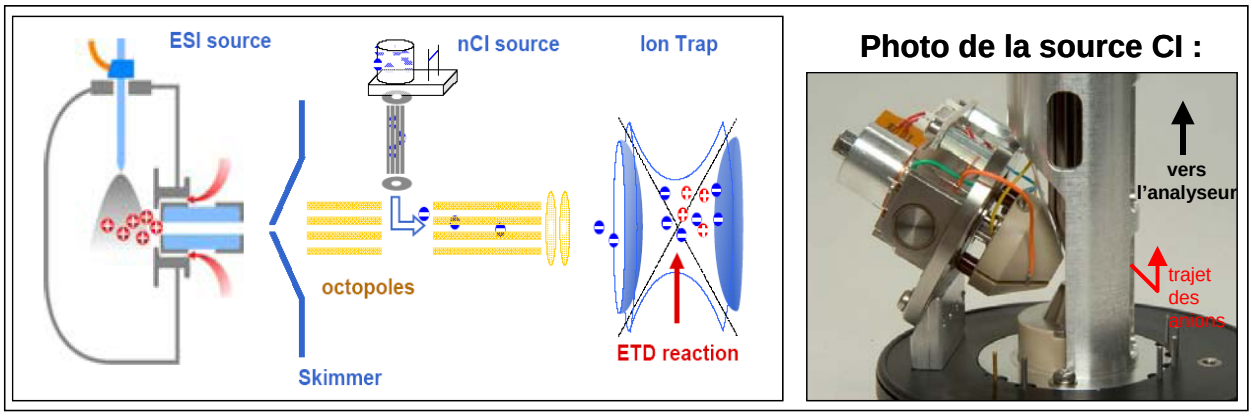
2.1 Le principe de l'ETD et configuration de la trappe HCTUltra PTM Discovery System

En 1998, McLafferty et coll. ont introduit une nouvelle méthode de fragmentation des peptides et protéines adaptée pour les spectromètres de masse de type FT-ICR-MS : la dissociation par capture d'électron (ECD=Electron Capture Dissociation) [Zubarev et coll., 1998]. Cette méthode présente des avantages par rapport à la fragmentation basée sur les dissociations induites par collisions (CID) mais présente le désavantage de n'être compatible qu'avec des instruments de type FT-ICR-MS, les spectromètres de masse les plus coûteux. C'est ainsi qu'une méthode impliquant des mécanismes de fragmentation voisins a été développée pour des instruments de type trappe ionique, moins coûteux et plus répandus dans les laboratoires de spectrométrie de masse : la dissociation par transfert d'électron (ETD=Electron Transfer Dissociation). L'équipe de Donald Hunt a apporté une contribution majeure au développement de cette méthode [Syka et coll., 2004 ; Coon et coll., 2005].

Le principe de l'ETD consiste à transférer un électron sur le peptide protoné (cation multichargé) à partir d'un réactif (un anion). Ce réactif transporteur d'électron est formé à l'extérieur de l'analyseur dans une source à ionisation chimique avant d'être confiné dans l'analyseur avec les peptides protonnés où auront lieu les transferts d'électron du réactif aux peptides protonnés.

Dans la trappe HCTUltra PTM Discovery, la source à ionisation chimique a été introduite entre les deux octapoles (Figure 7).

Figure 7 : Schéma de la trappe HCTUltra PTM Discovery avec l'introduction de la source à ionisation chimique (source CI) entre les deux octapôles [Hartmer et coll., 2005].

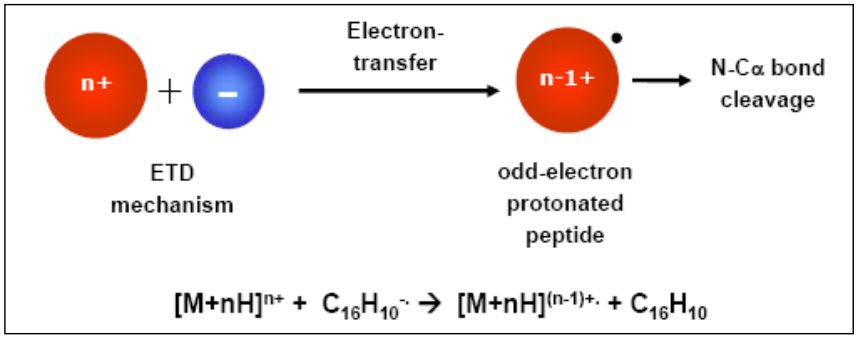


Dans cette source, un hydrocarbure aromatique, le fluoranthène est utilisé pour générer les anions radicaux en utilisant du méthane comme gaz réactant :

$$C_{16}H_{10} + e^- \rightarrow C_{16}H_{10}^{\bullet -}$$

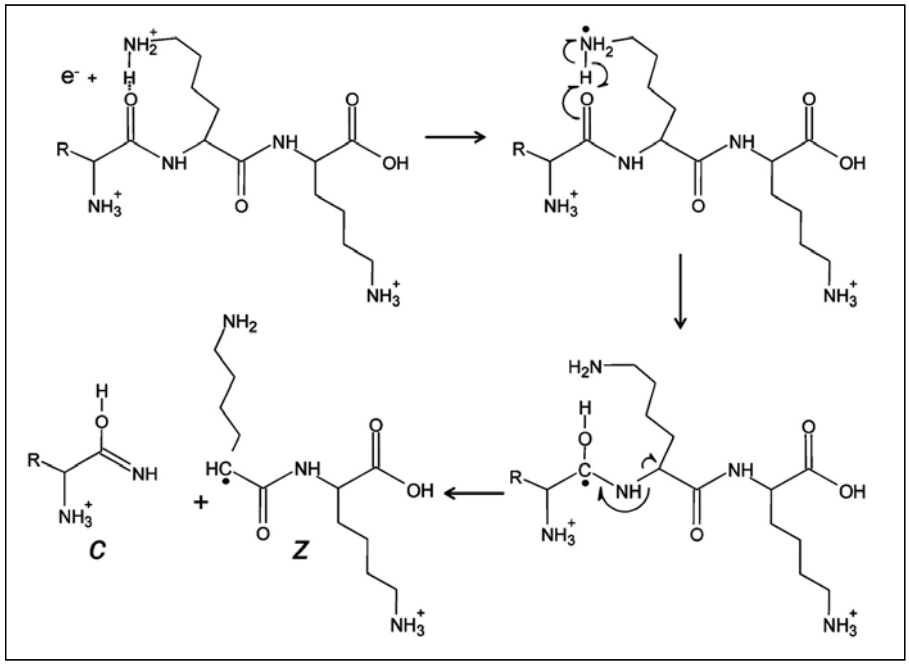
Des réactions de transfert d'électron auront lieu en phase gazeuse entre les anions fluoranthène et les peptides protonnés (Figure 8).

Figure 8 : Schéma du transfert d'électron de l'anion vers le peptide protonné [Lubeck et coll., 2006].



Les mécanismes de fragmentation des peptides sont similaires à ceux décrits pour les dissociations par ECD (Figure 9).

Figure 9 : Schéma de fragmentation pour la génération d'ions c- et z- après la réaction de transfert d'électron sur le peptide protonné [Syka et coll., 2004].



Les principaux avantages de la fragmentation ETD sont :

- Les fragmentations se font sur la chaîne peptide indépendamment de la séquence peptidique, il n'y a **pas de coupures préférentielles** comme c'est le cas en CID. Les spectres ETD génèreront donc **plus d'informations de séquence** que les spectres CID.

- Les **modifications post-traductionnelles sont maintenues** pendant les fragmentations ETD car les coupures ne se font pas préférentiellement au niveau des liaisons fragiles comme c'est le cas en CID. Les modifications pourront donc plus facilement être déterminées et localisées.
- La possibilité de fragmenter **des peptides plus gros** voire **des protéines**.

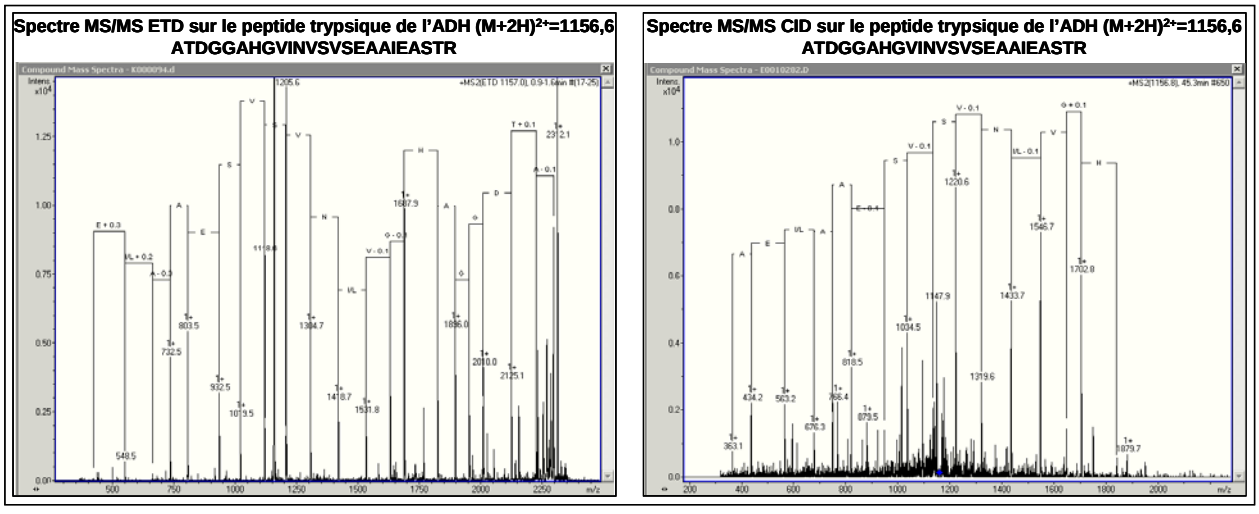
2.2 Les premiers résultats obtenus

2.2.1 La fragmentation de peptides trypsiques : comparaison CID-ETD

En figure 10 sont illustrés deux spectres MS/MS réalisés sur un peptide trypsique de l'Alcohol Dehydrogenase (ADH) de m/z=1556,6 et de séquence ATDGGAHGVINVSVSEAAIEASTR. La fragmentation de ce peptide a été réalisée d'une part en mode CID classique et d'autre part en mode ETD. Les deux spectres sont informatifs au niveau de la séquence en acides aminés du peptide qu'ils permettent de déterminer. Cependant les spectres présentent quelques différences notables. Les spectres générés en mode ETD permettent dans la plupart des cas de conserver l'information de la masse monoisotopique de l'ion parent. En effet, les phénomènes de réduction d'état de charge étant nombreux, lorsqu'un ion triplement chargé est fragmenté, les espèces doublement et monochargées seront abondantes sur le spectre ETD. De la même manière, le processus ETD conduit à la fragmentation aléatoire de la chaîne peptidique au niveau des liaisons N-C_α ce qui résulte en général en des fragments d'intensité assez homogènes. En revanche, en mode de fragmentation CID, des fragments préférentiels sont toujours prédominants, les profils des spectres MS/MS sont fortement dépendants à la fois de la composition en acides aminés et de l'ordre d'enchaînement de ces acides aminés. Un spectre CID peut donc tout aussi bien être très informatif et présenter des fragments sur toute la chaîne peptidique ou alors ne présenter que quelques fragments préférentiels très intenses.

Néanmoins, le principe de fonctionnement de la trappe et particulièrement l'application du cut-off fait que les ions de basses masses sont absents à la fois des spectres CID et ETD.

Figure 10 : Exemple d'un spectre MS/MS d'un peptide trypsique de m/z=1156,6 et de séquence ATDGGAHGVINVSVSEAAIESTR, fragmenté en mode ETD d'un coté et en mode CID de l'autre.



2.2.2 Les analyses nanoLC-MS/MS avec la fragmentation ETD

Des tests de couplage nanoLC-MS/MS en mode de fragmentation ETD sont en cours. Les acquisitions en mode nanoLC-MS/MS automatique peuvent être réalisées de différentes manières :

- Soit en mode CID exclusivement soit en mode ETD exclusivement.
- Soit en mode CID/ETD alterné. Ce mode permet de réaliser à la fois un spectre CID et un spectre ETD sur tous les ions sélectionnés.

Cette deuxième possibilité est particulièrement informative. Cependant, pour l'instant les algorithmes de recherche classiquement utilisés ne permettent pas encore de soumettre des fichiers « peak list » provenant d'analyses combinant à la fois des spectres de fragmentation CID et ETD. En effet, avec Mascot, la seule possibilité actuellement disponible pour lancer des requêtes avec des données ETD est de choisir comme instrument FTMS-ECD, l'algorithme considère alors des fragments des séries c-, y-, z- et z+1 qui correspondent également aux fragments générés en ETD.

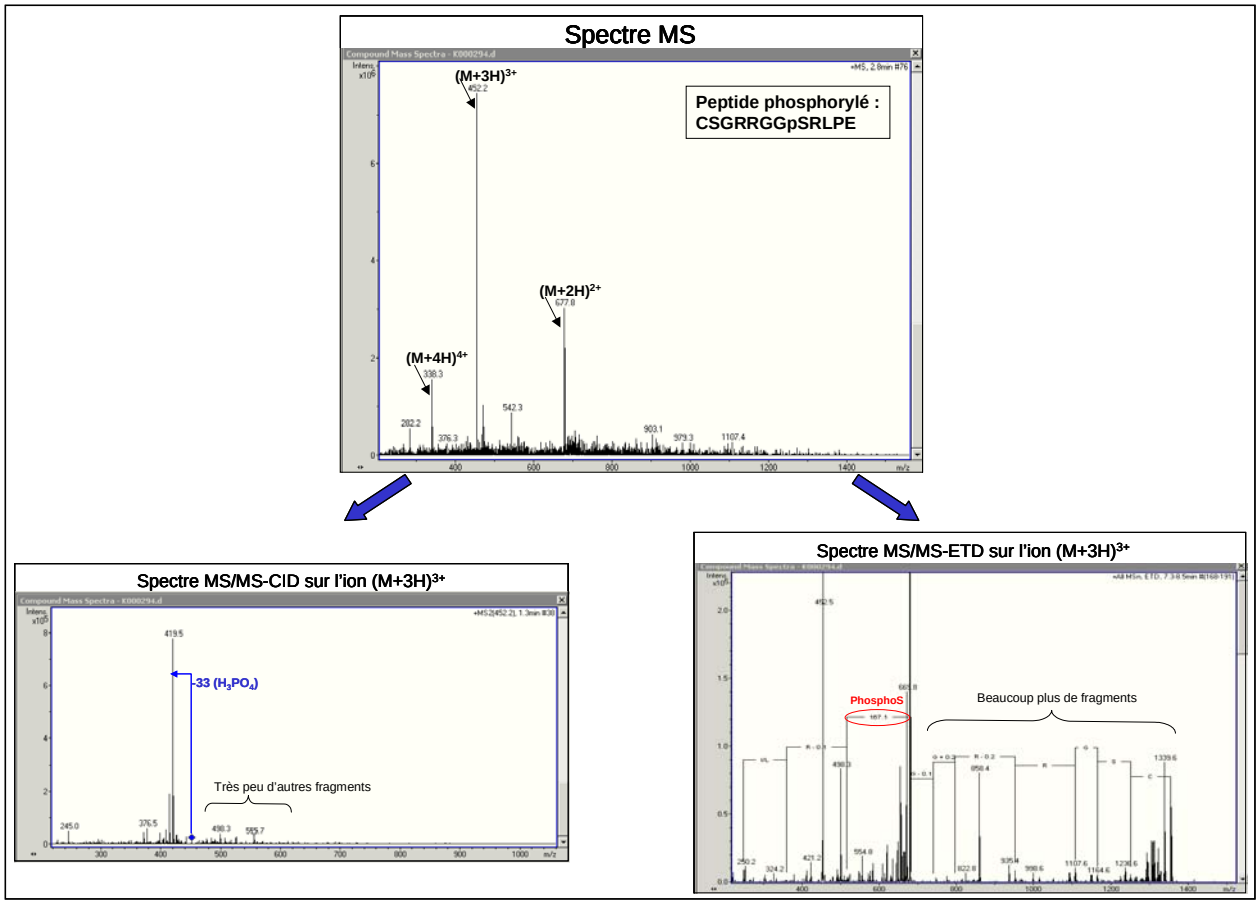
Bien qu'il soit possible de créer dans Mascot une méthode où à la fois les fragments c-, z- mais aussi b-, y- pourraient être détectés, ceci augmente considérablement le nombre de possibilités d'identification des fragments et réduit donc considérablement la spécificité de la recherche. Cette solution a donc été rejetée.

Par contre, dans la prochaine version de Mascot il sera possible de lancer des requêtes avec des « peak list » combinées CID-ETD, une entête sera ajoutée devant chaque ion précurseur dans la « peak list » précisant si les masses de fragments le concernant ont été obtenues par CID ou par ETD. Mascot reconnaîtra ces entêtes et ferra les recherches en considérant les fragments en conséquence.

2.2.3 La détermination des sites de phosphorylation

Comme précédemment décrit dans l'introduction, un des avantages principaux de la fragmentation par ETD est la conservation des modifications post-traductionnelles. En mode CID classique, les fragmentations vont avoir lieu préférentiellement au niveau des liaisons les plus fragiles. Les fragments correspondant à la perte du phosphate dans le cas des peptides phosphorylés seront donc largement majoritaires ce qui conduira dans la plupart des cas à des spectres MS/MS peu informatifs (Figure 11). En mode ETD par contre, ces liaisons fragiles ne seront pas préférentiellement brisées ce qui permettra de localiser précisément le site de phosphorylation. En figure 11 par exemple, la sérine phosphorylée a pu être identifiée de même qu'une bonne partie de la séquence en acides aminés du peptide.

Figure 11 : Exemple d'un peptide phosphorylé pour lequel la fragmentation CID n'est pas informative mais pour lequel la fragmentation ETD permet à la fois de déterminer la séquence en acides aminés et de localiser précisément la phosphorylation.



Des optimisations de méthodes en couplage nanoLC-MS/MS pour l'identification des phosphorylations sont également en cours. En effet, des acquisitions en mode perte de neutre sont possibles : des valeurs de pertes de masses peuvent être définies par l'utilisateur (par exemple, pour la détection de phosphorylation, des pertes de –80 pour HPO_3 ou –98 pour H_3PO_4 seront prédéfinies) et ces pertes seront automatiquement détectées. Les ions parents pour lesquels ces pertes de neutre sont détectées sur le spectre CID peuvent alors être automatiquement refragmentés en mode ETD, ou une MS3 automatique pourra être réalisée sur l'ion parent moins la perte de masse et un spectre ETD sur l'ion parent sera alors réalisé dans un second temps.

Pour le retraitement de ce type d'analyses, la version de Mascot permettant les interrogations combinées avec des données de CID et d'ETD est très attendue.

2.2.4 Essais sur des grands peptides et protéines

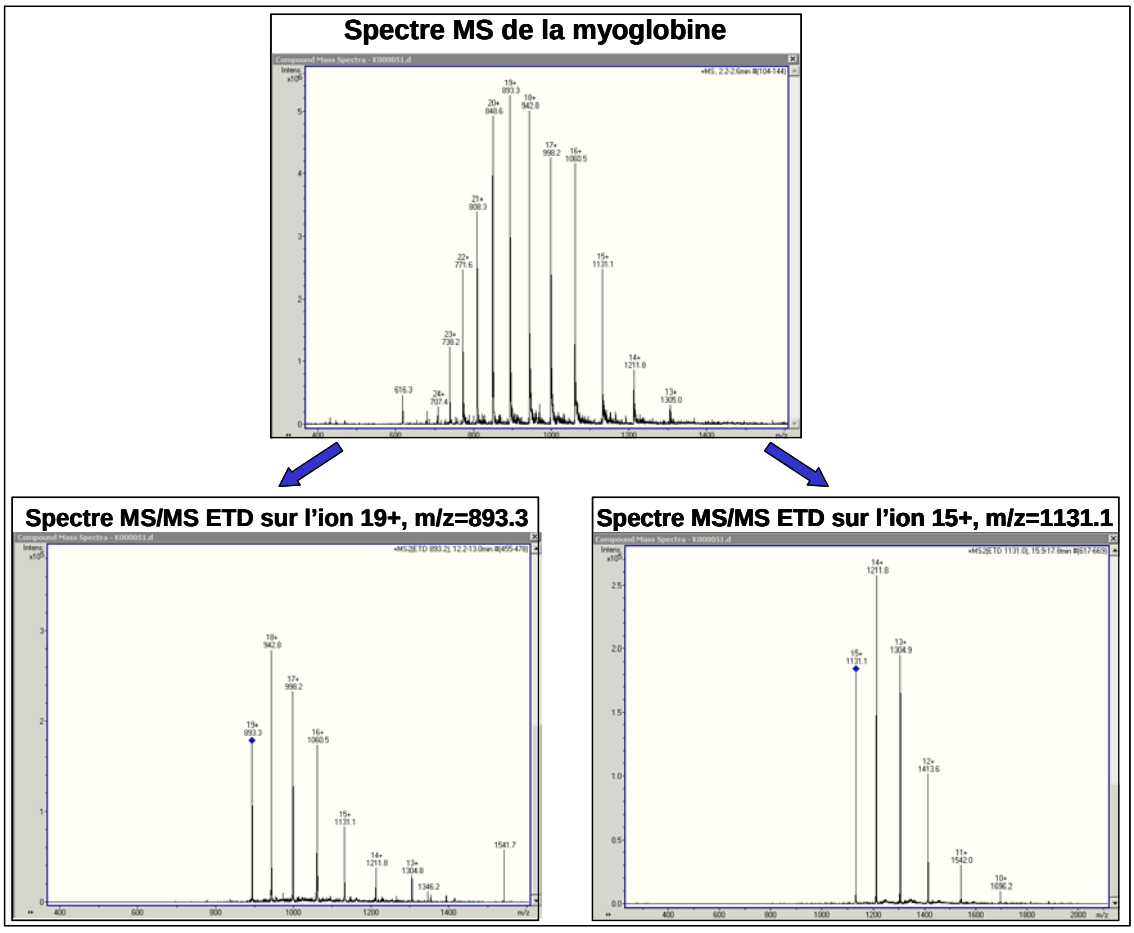
Un autre intérêt du développement de la fragmentation ETD a été de pouvoir l'appliquer à des grands peptides voire des protéines afin de déterminer des fragments de séquence plus longs qu'en CID [Syka et coll., 2004].

Des premiers tests sur des grands peptides et protéines ont été réalisées, les résultats obtenus jusque là sont encourageants pour de grands peptides (peptides allant de 2500 à 5000Da). Pour se placer dans des conditions idéales de fragmentation de ces peptides, il est utile d'ajuster la composition du solvant d'infusion

(entre 0,5 et 2% d'acide) de sorte à favoriser les protonations afin de pouvoir fragmenter les états de charge 4+ voire 3+ des espèces présentes.

Par contre, des essais sur des protéines entières ont été réalisés comme l'illustre la figure 12 avec la myoglobine de masse 16951,5Da. Dans ces cas, on observe principalement des phénomènes de réduction des états de charge sur les spectres : les anions agissent à la fois comme donneurs d'électrons mais aussi comme bases. A partir d'un ion $(M+3H)^{3+}$ on obtient à la fois des ions $(M+3H)^{+2\cdot}$ et $(M+3H)^{+\cdot}$ mais aussi des ions $(M+2H)^{2+}$ et $(M+H)^{+}$. D'après Syka et coll., 30 à 50 % des produits de transferts d'électrons sont des ions dont les fragments non dissociés sont assemblés par des liaisons non covalentes [Syka et coll., 2004]. Une étape de CID successive permet de fragmenter ces états de charge réduit. Nos optimisations vont donc dans cette direction : utiliser dans un premier temps la fragmentation ETD pour réduire les états de charge et dans un deuxième temps activer par CID ces espèces à charges réduites pour les fragmenter.

Figure 12 : Spectre MS de la myoglobine (16951,5 Da) ainsi que les spectres de fragmentation ETD des états de charge 19+ et 15+. Ces figures illustrent bien les phénomènes de réduction des états de charge.



Coon et coll. ont publié des résultats intéressants sur la fragmentation de grands peptides et de protéines utilisant l'ETD [Coon et coll., 2005] suivie par des réactions de déprotonation des ions multichargés générés. Les réactions de transfert de protons pour réduire les états de charge élevés sont réalisées en utilisant des anions d'acide benzoïque déprotoné [Coon et coll., 2005]. Les spectres ETD étant souvent difficilement interprétables car complexes et présentant de nombreux ions multichargés (pour lesquels la résolution de la trappe n'est pas suffisante pour déterminer sans ambiguïté les états de charge par

déconvolution), la réduction des états de charge pour n'obtenir qu'une majorité de fragments monochargés facilite considérablement l'interprétation.

2.3 Quelques astuces d'optimisation du signal et de la fragmentation

Pour l'instant, aucun protocole d'optimisation des meilleures conditions pour la fragmentation ETD n'est encore disponible. Au fil de ces quelques mois de découverte de l'instrument et des possibilités de la fragmentation ETD, quelques mises au point et réglages permettant d'améliorer les résultats de fragmentation ont pu être mis en évidence. Dans cette partie seront brièvement présentés les résultats de ces observations et quelques procédures importantes à suivre pour améliorer la qualité des fragmentations.

2.3.1 La quantité d'anions fluorenthène accumulés dans la trappe

L'optimisation du nombre d'anions accumulés dans la trappe doit se faire systématiquement avant chaque série d'analyses. L'objectif à atteindre est une valeur d'ICC d'environ 600000 pour être dans de bonnes conditions pour la fragmentation ETD.

La meilleure façon de procéder pour réaliser l'optimisation de la génération des anions et de leur transmission jusqu'à l'analyseur est la suivante:

- Dans un premier temps, fixer un temps d'accumulation court, 1ms, afin de ne pas accumuler trop d'anions et de conserver une bonne résolution isotopique du fluoranthène de m/z=202. Un ICC entre 30000 et 50000 doit être atteint en optimisant bien les paramètres de transmission.
- Il suffit alors dans un deuxième temps de multiplier le temps d'accumulation par le facteur nécessaire pour obtenir une valeur d'ICC de 600000. Dans de bonnes conditions, des temps d'accumulation de l'ordre de 10ms sont nécessaires.

2.3.2 Le temps de réaction dans l'analyseur

Le temps de réaction dans l'analyseur doit également être attentivement optimisé. Dans un premier temps, il est utile de fixer un temps de réaction relativement court puis de l'augmenter progressivement afin d'observer l'apparition de plus en plus de fragments. Il s'agit ensuite de trouver le temps de réaction optimal : en augmentant le temps de réaction, on finit par observer une perte en sensibilité car les transferts multiples d'électrons aboutissent à la génération d'espèces neutres qui ne vont plus être détectables. En partant d'un temps de réaction de 80ms, on arrive en général à un temps de réaction optimal entre 130 et 180ms.

2.3.3 Le cut-off de l'ETD

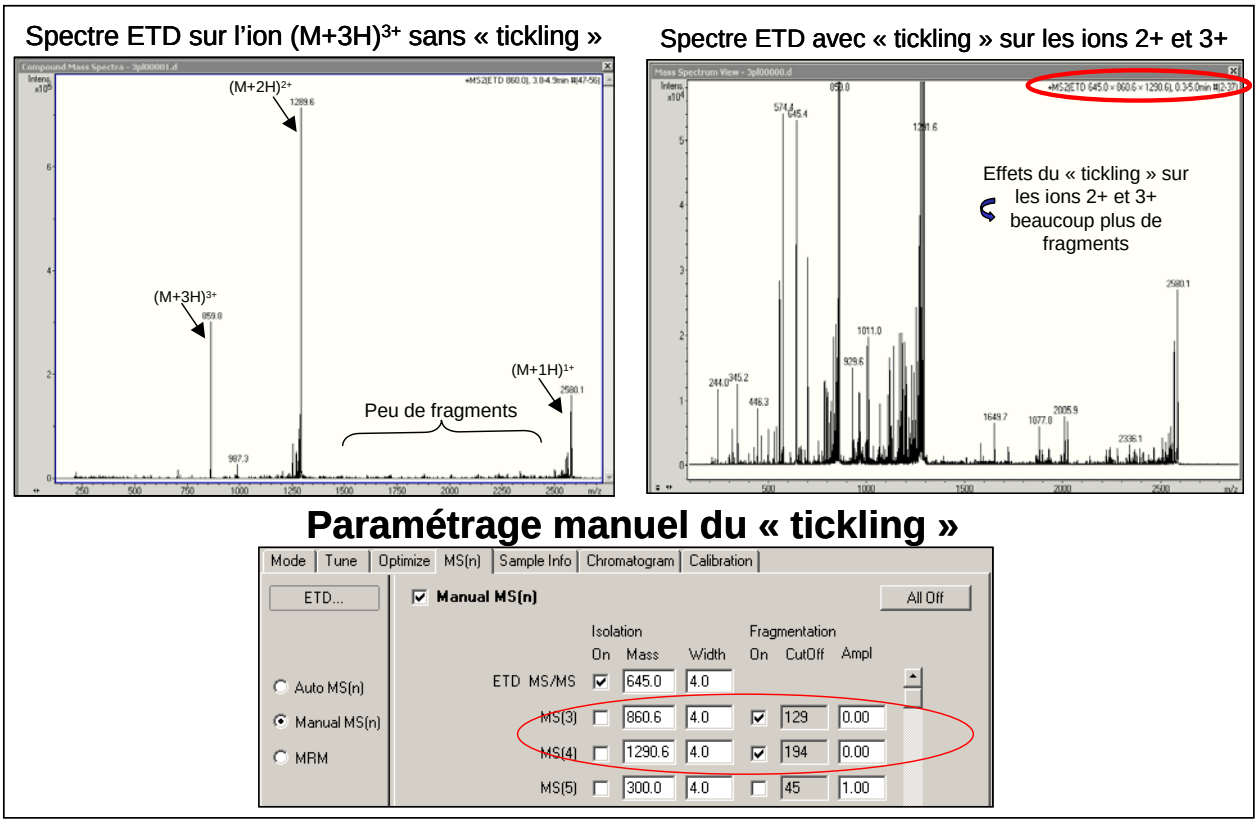
En mode de fragmentation ETD, un second cut-off (indépendant du cut-off décrit au chapitre I, § 1.1.1 de cette partie des résultats) doit être optimisé. Cette valeur de cut-off règle l'augmentation de l'amplitude de la radiofréquence sur l'électrode annulaire permettant de confiner à la fois les anions et les peptides protonnés au centre de la trappe. Une valeur de 135 pour ce cut-off est en général préconisée mais cette valeur peut être légèrement augmentée et on observe alors parfois une plus grande efficacité des fragmentations. Cependant, en dépassant une valeur de 180, les anions sont éjectés et l'efficacité de la fragmentation chute.

Pour les grands peptides et les protéines, présentant plus de charges, la valeur de ce cut-off peut être diminuée (100) ainsi que le temps de réaction (80ms). Les ions à fragmenter étant plus chargés, les attractions électrostatiques entre les ions positifs et les anions sont plus fortes et les collisions se font plus facilement (sans avoir ni à trop les confiner ni à les faire réagir longtemps).

2.3.4 Les modes "smart decomposition" et le "tickling" des ions

La fragmentation ETD est plus efficace sur des hauts états de charge, les ions doublement chargés fragmentent en général difficilement. Ceci est un désavantage lorsqu'il s'agit d'étudier des peptides trypsiques qui sont majoritairement doublement chargés. Dans ces cas, une opération de « tickling » des ions permet d'augmenter l'efficacité de fragmentation des peptides. Cette opération se règle par l'option « smart decomposition » : lorsque l'on fragmente un peptide doublement chargé, il faut paramétrer l'option smart decomposition à z=2, lorsque l'on fragmente un peptide triplement chargé à z=3, … Cette opération de « tickling » consiste, dans le cas d'un peptide doublement chargé, en une excitation supplémentaire, ciblée sur la masse calculée de l'ion monochargé correspondant, équivalente à une excitation CID afin d'améliorer sa fragmentation. Cette opération peut être paramétrée par l'option smart decomposition (dans ce cas, les masses des peptides monochargés sont calculées automatiquement et l'excitation se fait de façon transparente pour l'utilisateur) ou elle peut être paramétrée par l'utilisateur directement dans la fenêtre de paramétrage des MS multiples (Figure 13). Cette deuxième possibilité permet dans bien des cas d'améliorer l'efficacité de la fragmentation : lorsque la fragmentation ETD génère un (ou plusieurs) ion(s) majoritaire(s), l'utilisateur peut appliquer une excitation supplémentaire à ces différents ions majoritaires (Figure 13).

Figure 13 : Illustration de l'effet du « tickling » des ions sur l'efficacité de fragmentation des peptides.



2.4 Les limites du système

2.4.1 La perte de sensibilité

- Les fragmentations générant plus de fragments qu'en mode CID, la perte en sensibilité globale en mode ETD est relativement importante.
- D'autre part, des temps de réaction trop longs conduisent à la génération de neutres non mesurables en spectrométrie de masse qui contribuent également à la perte d'intensité de signal.
- Le fait d'avoir des ions multichargés sur les spectres MS/MS ETD nécessite une bonne résolution afin de pouvoir déterminer, dans la mesure du possible, les états de charge des fragments. Avec ces exigences de résolution, l'utilisation du mode de balayage Ultrascan n'est donc pas appropriée ce qui est également désavantageux pour la sensibilité (voir chapitre I, 1.1.3 de cette partie de résultats).

2.4.2 Les difficultés d'interprétation des spectres

Les analyseurs de type trappe ionique étant limités en résolution, la déconvolution des ions multichargés devient rapidement une tâche très compliquée voire impossible de même que l'interprétation des spectres ETD. Ce problème est un peu amélioré en utilisant le mode de balayage « standard enhanced » pour privilégier une résolution maximale au détriment de la sensibilité mais reste malgré tout réel.

2.5 Conclusion

L'objectif de ce chapitre était de présenter les premiers résultats obtenus et de mettre en forme les réflexions relevées au fil de ces quelques mois d'utilisation de la trappe HCT Ultra PMT Discovrey System. Ces résultats ne sont pour l'instant que préliminaires mais peuvent constituer une base pour les optimisations futures de la fragmentation ETD.

Bibliographie

Agilent 1200 Series HPLC-Chip/MS system. The easy-to-use HPLC-Chip for reliable high sensitivity nanospray LC/MS. *Agilent Technologies Application Note*, **2006**, publication number 5989-5492EN.

Agilent G3253A Dual Nanospray Ion source. *User's Guide Agilent Technologies*, Manual Part Number G3253-90015, **2005**.

Dobson G., Murrell J., Despeyroux D., Wind F. and Tabet J. C.

Investigation into factors affecting precision in ion trap mass spectrometry using different scan directions and axial modulation potential amplitudes. *Journal of Mass Spectrom.*, **2004**, 39, 1295-1304.

Dongré A. R., Jones J. L., Somogyi A. and Wysocki V. H.

Influence of peptide composition, gas-phase basicity, and chemical modification on fragmentation efficiency : Evidence for the mobile proton model. *J. Am. Chem. Soc.*, **1996**, 118, 8365-8374.

Coon J. J., Ueberheide B., Syka J. E. P., Dryhurst D. D., Ausio J., Shabanowitz J. and Hunt D. F.

Protein identification using sequential ion/ion reactions and tandem mass spectrometry. *PNAS*, **2005**, 102(27), 9463-9468.

Cox K. A., Cleven C. D. and Cooks R. G.

Mass shifts and local space charge effects observed in the quadrupole ion trap at higher resolution. *Int. J. of Mass Spectrom.*, **1995**, 144, 47-65.

Guan S. and Marshall A. G.

Equilibrium space charge distribution in a quadrupole ion trap. *J. Am. Soc. Mass Spectrom.*, **1994**, 5, 64-71.

Hartmer R. and Lubeck M.

New approach for characterization of post translational modified peptides using ion trap MS with combined ETD/CID fragmentation. *The applications book Medical/Biological*, Bruker Daltonics, **2005**.

Li G. Z., Guan S. and Marshall A. G.

Comparison of equilibrium ion density distribution and trapping force in Penning, Paul and combined Ion traps. *J. Am. Soc. Mass Spectrom.*, **1997**, 9, 473-481.

Lubeck M., Shen M., Hartmer R., Brekenfeld A. and Baessmann C.

Improved detection and characterization of phosphopeptides using electron transfer dissociation ion trap MS. *Bruker Daltonics ABRF Poster*, **2006**, abrf06-6.

O'Connor P. B., Cournoyer J. J., Pitteri S. J., Chrisman P. A. and McLuckey S. A.

Differentiation of aspartic and isoaspartic acids using electron transfer dissociation. *J. Am. Soc. Mass Spectrom.*, **2006**, 17, 15-19.

Savitski M. M., Nielsen M. L., Kjeldsen F. and Zubarev R. A.

Proteomics-grade de novo sequencing approach. *J. Proteome Res.*, **2005**, 4, 2348-2354.

Syka J. E. P., Coon J. J., Schroeder M. J., Shabanowitz J. and Hunt D. F.

Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *PNAS*, **2004**, 101, 9528-9533.

Vollmer M. and Miller C. A.

Comparison of proteome analysis by HPLC-Chip/MS vs. Nanospray LC/MS. *ABRF05 Poster 1*, **2005a.**

Vollmer M., Miller C. and Glatz B.

Performance of the Agilent 1100 HPLC-Chip/MS System. *Technical Overview*, **2005b**, Publication number 5989-4148EN.

Yang C. G. and Bier M. E.

Investigation of the Rapid Scan on an Electrospray Ion Trap Mass Spectrometer. *Anal. Chem.*, **2005**, 77, 1663-1671.

Yin H., Killeen K., Brennen R., Sobek D., Werlich M. and van de Goor T.

Microfluidic chip for peptide analysis with an integrated HPLC column, sample enrichment column and nanoelectrospray tip. *Anal. Chem.*, **2005**, 77, 527-533.

Zhang Z.

Prediction of low-energy collision-induces dissociation spectra of peptides. *Anal. Chem.*, **2004**, 76, 3908-3922.

Zubarev R. A., Kelleher N. L. and McLafferty F. W.

Electron capture dissociation of multiply charged protein cations. A nonergodic process. *J. Am. Chem. Soc.*, **1998**, 120, 3265-3266.

RESULTATS

DEUXIEME PARTIE :
Applications portant sur des organismes dont le génome est séquencé et bien annoté

Chapitre I : Une carte protéomique 2D de référence et étude différentielle de l'effet d'un minéralocorticoïde, l'aldostérone, chez *Schizosaccharomyces pombe*

Chapitre II : Etude de l'exoprotéome du champignon filamenteux *Fusarium graminearum* se développant sur des parois de houblon

Chapitre III : Etude de l'interactome virus-protéines hôtes recrutées lors de l'infection du riz par le virus de la Panachure Jaune du Riz

CHAPITRE I

Une carte protéomique de référence et étude différentielle de l'effet d'un minéralocorticoïde, l'aldostérone chez *Schizosaccharomyces pombe*

Les deux études rapportées dans ce chapitre ont été réalisées en collaboration avec le laboratoire *FR 8.3 Biochemie de l'Universität des Saarlandes* à Saarbrücken en Allemagne sous la direction du Professeur Rita Bernhardt. Deux étudiants en thèse étaient directement impliqués dans ces collaborations : Kyung-Hoon Hwang pour l'établissement de la carte protéomique de référence et Susanne Böhmer pour l'étude différentielle de l'effet de l'aldostérone.

1 Etablissement d'une carte protéomique de référence pour *Schizosaccharomyces pombe*

1.1 Contexte de cette étude

1.1.1 *Schizosaccharomyces pombe*, un système modèle

La levure *Schizosaccharomyces pombe (S. pombe)* a été isolée et décrite pour la première fois en 1890 à partir de bière de millet provenant de Zanzibar en Afrique orientale (*pombe* signifiant bière en Swahili) par le biologiste allemant P. Lindner. Cette levure eucaryote unicellulaire est appelée aussi levure fissipare, fission yeast, car elle ne se multiplie pas, comme *Saccharomyces cerevisiae*, par bourgeonnement mais par fission transversale [Kurtzman et coll., 2000]. Des découvertes essentielles dans le domaine du contrôle du cycle cellulaire chez *S. pombe*, l'identification des molécules clés contrôlant le cycle, la caractérisation de leurs fonctions et la mise en évidence de leur caractère universel ont valu le prix Nobel de médecine et de physiologie en 2001 à Leland Hartwell, Paul Nurse et Tim Hunt. Un bon nombre de raisons confortent le fait

que cette levure soit, en bien des aspects, très proche des systèmes vertébrés (encore bien plus que la levure *S. cerevisiae*) et en font un organisme modèle de choix [Forsburg et coll., 2006]:

- Des protéines humaines ont été exprimées avec succès dans *S. pombe* [Bureik et coll., 2002, Idiris et coll., 2006]
- De nombreux travaux sur le contrôle du cycle cellulaire, la mitose et la méiose, la réparation et la recombinaison de l'ADN ont été conduits chez *S. pombe* [Fantes et coll., 2000; Davis et coll., 2001]
- Le système membranaire intracellulaire est plus développé chez *S. pombe* que chez *S. cerevisiae* [Chappell et coll., 1989]
- Les domaines de transactivation des mammifères ont un fonctionnement identique chez *S. pombe* [Remacle et coll., 1997]

1.1.2 Le génome et le protéome de S. pombe

Huit ans après la publication du génome de *Saccharomyces cerevisiae*, le premier génome eucaryote entièrement séquencé [Goffeau et coll., 1996], le génome de *S. pombe* a été publié en 2002 [Wood et coll., 2002]. Voici quelques caractéristiques et statistiques de ce génome mises à jour au 26 janvier 2006 par le Sanger Institute (http://www.sanger.ac.uk) :

- Son génome fait environ 14,1 Mbps
- Il présente 3 chromosomes de 3,5 à 5,7 Mbps
- Il comporte 4979 gènes codants dont 1538 donc 30,9% ont été caractérisés et publiés (45% de ses gènes contiennent des introns)
- Son pourcentage de GC est de 36% de GC

Le génome de *S. pombe* a été largement annoté et son protéome est bien représenté dans les banques de données protéiques accessibles avec 3098 entrées annotées pour *S. pombe* dans la banque UniProtKB/SwissProt et 10939 liens dans la banque protéique généraliste NCBInr.

1.1.3 Objectif de l'étude

Au vu du grand intérêt porté par bon nombre de biologistes pour *S. pombe* comme système modèle et du peu d'informations disponibles quant à l'étude de son protéome, l'objectif de cette étude a été la réalisation d'une carte 2D de référence de *S. pombe* la plus complète possible. Pour cela, tous les moyens ont été mis en oeuvre pour la réalisation de gels d'électrophorèse 2D les plus résolutifs et reproductibles possibles ainsi que pour l'analyse des spots par spectrométrie de masse MALDI-TOF-MS complétée par des analyses nanoLC-MS/MS.

1.2 Choix de la stratégie expérimentale mise en oeuvre

1.2.1 Préparation des échantillons

Toutes les informations relatives à la préparation des échantillons, à la culture cellulaire, à l'extraction des protéines puis à la préparation et à l'analyse des gels 2D de référence sont données dans la publication détaillée des résultats en 1. 3.

1.2.2 Analyse par spectrométrie de masse

Dans un premier temps, une stratégie d'identification par empreinte peptidique massique (PMF) a été mise en oeuvre pour l'ensemble des spots découpés sur les gels 2D. Cette approche par MALDI-TOF-MS a donné lieu à de nombreuses identifications. Cependant, dans un nombre significatif de spots, soit un trop grand nombre de protéines a rendu les échantillons trop complexes pour l'analyse par PMF, soit les identifications présentaient des ambiguïtés. Dans ces cas, des analyses par nanoLC-MS/MS sur des couplages de type CapLC/Q-TOF II (Waters) ou nanoHPLC Agilent Serie 1100/trappe ionique HCTPlus (Bruker Daltonics) ont permis de compléter les résultats obtenus.

1.2.3 Stratégie d'identification des protéines

Les recherches en banque de données ont été réalisées avec le moteur de recherche Mascot™ dans la banque protéique accessible NCBInr. Une tolérance de 70ppm est autorisée, aucune restriction de taxonomie ni de pI ni de poids moléculaire n'est spécifiée, un manquement de coupure trypsique est toléré et enfin la carbamidomethylation des cystéines ainsi que l'oxidation des méthionines sont autorisées comme modifications variables. Tous les spots ont été analysés par MALDI-TOF-MS et une analyse de ces premiers résultats a permis d'écarter les identifications sans ambiguïtés et de prévoir l'analyse de tous les cas ambigus par nanoLC-MS/MS. Les requêtes avec des données de nanoLC-MS/MS ont été réalisées avec des paramétrages identiques à ceux des recherches PMF hormis pour la tolérance de masse qui a été fixée à 0,25Da en MS et en MS/MS. Pour les protéines identifiées de fonction encore inconnue, des recherches d'homologies de séquence ont été lancées par BLAST afin d'attribuer des fonctions ou de détecter au-moins des domaines conservés.

1.3 Résultats publiés

La carte protéomique établie dans le cadre de cette étude a fait l'objet d'une publication acceptée dans le journal *Proteomics* en mars 2006.

En résumé, un total de 1500 spots a été visualisé sur les gels 2D colorés à l'argent et un total de 800 spots sur les gels colorés au bleu dans une gamme de pH 3-10. 80% des spots se situant dans une gamme de pH 4-7, une seconde série de gels restreints à la gamme de pH 4-7 a été réalisée afin d'obtenir une meilleure résolution. Ainsi, près de 1000 et 500 spots ont pu être visualisés sur ces gels, respectivement en coloration à l'argent et en coloration au bleu. Au total, 386 spots ont été découpés et 365 protéines ont pu être identifiées (représentant 158 protéines différentes). Par rapport à la première carte protéomique de référence pour *S. pombe* publiée en mai 2005 dans le journal *Proteomics*, 118 nouvelles protéines ont pu être identifiées ici [Sun et coll., 2005].

Afin de suivre les directives de publication de données de protéomique [Carr et coll., 2004 ; Bradshaw, 2005 ; Wilkins et coll., 2006], les critères stringents de validation des identifications ont été appliqués :

- Pour les analyses MALDI-TOF-MS un minimum de 6 peptides identifiés avec une erreur de moins de 30 ppm est requis.

- Pour les analyses nanoLC-MS/MS un minimum de 2 peptides présentant des spectres MS/MS de haute qualité est requis et les identifications avec 2 peptides seulement ne sont validées qu'une fois les spectres MS/MS individuellement vérifiés à la main par séquençage *de novo*.

Dans bien des cas, la comparaison des résultats obtenus en MALDI-TOF-MS et nanoLC-MS/MS ont permis de mettre en évidence la complémentarité de ces deux techniques d'analyse, des peptides différents ont pu être identifiés et les pourcentages de recouvrement globaux augmentés (voir fig 2.C et fig 3.B de la publication attachée).

1.4 Conclusions et perspectives

La carte protéomique de référence qui a été réalisée ici, plus complète que celle qui a été précédemment réalisée, pourra constituer un outil de base en parallèle du génome pour les études futures utilisant *S. pombe* comme organisme modèle. Les perspectives de ce travail sont de continuer à faire évoluer cette carte de référence, de la compléter, d'y apporter de nouvelles annotations, d'utiliser ces données de protéomique pour les confronter aux annotations du génome afin de construire une base de connaissances la plus complète possible du génome et du protéome de *S. pombe*.

Proteome analysis of *Schizosaccharomyces pombe* by two-dimensional gel electrophoresis and mass spectrometry

Kyung-Hoon Hwang, **Christine Carapito**, Susanne Böhmer, Emmanuelle Leize, Alain Van Dorsselaer et Rita Bernhardt

Proteomics, 2006, Vol. 6, Pages 4115–4129

Pages 4115 à 4129 :

2 Etude différentielle de l'effet d'un minéralocorticoïde : l'aldostérone

2.1 Contexte de cette étude

2.1.1 L'aldostérone

Les hormones stéroïdes sont impliquées dans toute une série de processus importants dans l'organisme humain et elles se divisent en 3 classes : les hormones sexuelles, les glucocorticoïdes et les minéralocorticoïdes. L'aldostérone, le principal minéralocorticoïde, contrôle l'homéostasie de l'eau et des sels dans le corps : une surproduction de cette hormone conduit donc à des augmentations de la pression sanguine et peut causer la fibrose de tissus cardiaques [Brilla, 2000]. Deux modes d'action des hormones stéroïdes sont connus. Le mode d'action génomique par fixation sur des récepteurs nucléaires spécifiques induisant des effets positifs ou négatifs sur des gènes cibles est bien connu [Beato et coll., 1996; Beato et coll., 2000]. Le second mode d'action des hormones stéroïdes récepteur-indépendant n'influence pas directement l'expression des gènes mais opère plutôt par l'activation de cascades de signalisation en affectant le pH intracellulaire ou par d'autres biais [Sylvia et coll., 1993; Wehling et coll., 1996; Wehling, 1997; Losel et coll., 2003]. Plus particulièrement, il a été montré que l'aldostérone a des effets rapides et non-génomiques dans différents tissus cibles incluant le système vasculaire et le coeur [Wehling et coll., 1998; Losel et coll., 2002]. Bien que des études de ces réponses rapides, récepteurs-indépendantes, aient été décrites à tous les niveaux biologiques, de nombreux aspects de ces réponses nécessitent encore des recherches intensives et des points clés dans ces réponses doivent encore être élucidés.

2.1.2 Choix du modèle d'étude *S. pombe*

Les principaux avantages de *S. pombe* comme organisme modèle ont déjà été décrits en 1.1.1. Cependant, dans le cadre de cette étude, une raison supplémentaire vient conforter le choix de *S. pombe* comme système modèle : cette levure ne contenant pas de récepteurs nucléaires aux stéroïdes, l'étude de l'action récepteur-indépendante des hormones stéroïdes au niveau des profils protéiques est possible.

2.1.3 Objectif de l'étude

L'objectif de cette étude est de déterminer, au niveau protéique, les acteurs qui sont impliqués dans la voie d'action récepteur-indépendante de l'aldostérone. Pour cela, une approche protéomique différentielle par gel d'électrophorèse 2D suivie d'analyses par spectrométrie de masse pour l'identification des protéines différentiellement exprimées est mise en oeuvre. De plus, afin de valider l'approche mais aussi le choix de *S. pombe* comme système modèle, l'effet de l'aldostérone sur une des protéines identifiée par l'approche protéomique a été vérifié et validé sur des cellules humaines en culture grâce à un contrôle avec des anticorps spécifiques.

2.2 Choix de la stratégie expérimentale mise en oeuvre

2.2.1 Préparation des échantillons

En se basant à la fois sur des études précédemment réalisées avec l'aldostérone et en tenant compte de la structure de *S. pombe*, des temps d'incubation et une gamme de concentration d'aldostérone appropriés ont été déterminés pour cette étude : une gamme de concentration de 1 à 10µM d'aldostérone et un temps d'incubation de 3,5h ont été choisis. Le corticostérone, un précurseur de l'aldostérone est utilisé comme contrôle additionnel pour l'analyse différentielle par gel 2D. Dans un premier temps, après incubation avec de la ^{35}S-methionine, les extraits protéiques contrôles et ceux avec aldostérone ont été déposés sur gel d'électrophorèse 1D afin de déterminer globalement si l'aldostérone avait un effet sur les profils protéiques. Ceci ayant été montré, des gels d'électrophorèse 2D et des analyses approfondies de ces images de gel ont été réalisées avec le logiciel PDQuest. L'ensemble des gels 2D, contrôles, avec corticostérone et avec aldostérone ont été réalisés en triplicata. Grâce à des analyses statistiques, des spots différentiellement exprimés spécifiquement et consécutivement à la présence d'aldostérone ont pu être identifiés. De plus, grâce à l'utilisation de corticostérone en second contrôle, une série de différences répondant plus généralement à des effets des hormones stéroïdes ont pu être mises en évidence.

2.2.2 Analyse par spectrométrie de masse

Les spots d'intérêts ont été analysés à la fois par MALDI-TOF-MS et nanoLC-MS/MS afin de bénéficier de la complémentarité de ces deux techniques d'analyse et d'obtenir un maximum de recouvrement pour les protéines identifiées. Les analyses par MALDI-TOF-MS ont été réalisées sur un Ultraflex^{TM} TOF-TOF (Bruker Daltonics) et les analyses nanoLC-MS/MS sur un couplage CapLC/Q-TOF II (Waters). Lors de l'établissement d'une carte protéomique de référence un maximum de protéines doit être identifié et la multiplicité des protéines par spot ne pose donc pas de réel problème. En revanche, lors d'une analyse différentielle la présence de plusieurs protéines par spot est problématique et pour l'interprétation du différentiel cet aspect ne doit pas être négligé [Campostrini et coll., 2005]. En effet, l'identification de plusieurs protéines par spot rend difficile l'attribution de la sous- ou sur-expression d'un spot à l'une ou l'autre des protéines présentes. Ce problème a été rencontré dans plusieurs cas dans cette étude et une analyse détaillée des résultats de spectrométrie de masse a permis de trancher en faveur de l'une ou l'autre des protéines.

2.2.3 Stratégie d'identification des protéines

Les données de spectrométrie de masse ont été soumises à des requêtes Mascot^{TM} dans la banque protéique généraliste NCBInr avec les mêmes paramètres de recherche que ceux utilisés pour l'établissement de la carte protéomique de référence. Cependant, le calcul systématique des pourcentages de recouvrement totaux pour les protéines identifiées combinant les résultats de PMF et de nanoLC-MS/MS a été réalisé et le bénéfice de cette complémentarité a été mis en valeur dans la publication des résultats.

2.3 Résultats publiés

L'ensemble des résultats a fait l'objet d'une publication acceptée dans le journal *Biological Chemistry* en février 2006.

En résumé, 38 spots différentiellement exprimés en présence d'aldostérone ont été identifiés suite à l'analyse d'image réalisée avec le logiciel PDQuest. 9 spots présentent une intensité au-moins double et 29 spots une intensité diminuée d'au-moins 2 fois dans les échantillons traités par l'aldostérone. Grâce à la comparaison des résultats aldostérone/corticostérone, 6 spots sont considérés comme spécifiquement modulés par l'aldostérone et 32 spots semblent répondre plus généralement à une action des hormones stéroïdes.

Dans certains spots, plusieurs protéines ont été identifiées sans ambiguïté ce qui rend impossible l'attribution de la responsabilité de la variation d'intensité du spot à l'une ou l'autre de ces protéines sans avoir recours à des expériences supplémentaires de quantification. L'interprétation de l'analyse différentielle devient dans ces cas très difficiles. Cependant dans 13 spots, une unique protéine a été identifiée et les calculs des pourcentages de recouvrement combinés obtenus par MALDI-TOF-MS et nanoLC-MS/MS illustrent bien la complémentarité et l'intérêt de combiner les deux approches dès que cela est possible (résultats illustrés dans le Tableau 2 de la publication).

Dans le spot 7, une protéine d'un intérêt tout particulier a été identifiée, la protéine vip1 : dans une première analyse de l'extrait peptidique de ce spot, quelques peptides appartenant à deux autres protéines ont été identifiés (une transaldolase et une actine). Dans ce cas, une analyse détaillée du spectre de masse MALDI-TOF ainsi que plusieurs observations ont permis l'attribution de la sous-expression de ce spot à la protéine vip1 (voir discussion des résultats dans la publication).

Enfin, dans le but de valider l'utilisation de *S. pombe* comme organisme modèle pour cette étude mais aussi pour valider l'approche protéomique mise en oeuvre ici, la sous-expression d'une des protéines identifiées a été validée dans des cellules humaines HCT116 en culture par immunoblotting.

2.4 Conclusions et perspectives

En conclusion, un effet recepteur-indépendant de l'aldostérone a pu être clairement identifié grâce à l'approche protéomique mise en oeuvre dans cette étude. En plus de l'intérêt biologique des protéines identifiées largement discuté dans la publication, deux aspects importants du point de vue analytique peuvent être retenus de ce travail :

- **La présence de plusieurs protéines dans un spot de gel 2D** complique souvent l'interprétation de l'analyse différentielle et fait apparaître une des limites des approches protéomiques différentielles par gel 2D [Lill, 2003]. En effet, malgré l'amélioration de la sensibilité, de la reproductibilité et des pouvoirs résolutifs des gels 2D, ces approches se voient souvent remplacées par des approches quantitatives basées sur des marquages isotopiques (du type ICAT ou SILAC) [Gygi et coll., 1999 ; Ong et coll., 2002]. Même si le gel 2D reste un outil largement utilisé en analyse protéomique, ses limites, notamment pour les approches quantitatives, sont démontrées (sous-représentation des protéines membranaires, une gamme dynamique trop faible pour la détection des protéines minoritaires, la présence de plusieurs protéines par spot ou de plusieurs spots par protéine) [Lill, 2003]. Par ailleurs, les améliorations de la sensibilité, des

limites de détection et de la gestion des mélanges complexes en MS ont accentué ce problème : si les techniques de MS permettaient, il y a 10 ans, l'identification d'une protéine dans un spot, plusieurs protéines pourraient aujourd'hui être identifiées dans ce même spot ce qui rend l'interprétation du différentielle difficile.

- **La complémentarité des approches par MALDI-TOF-MS et nanoLC-MS/MS** a pu être mise en évidence. L'intérêt d'utiliser la complémentarité de ces deux approches pour augmenter les pourcentages de recouvrement et donc caractériser plus finement les protéines identifiées est grand.

Maintenant que le modèle et l'approche sont validés, de nouvelles analyses sont en cours afin de compléter les résultats déjà obtenus passant par l'utilisation de gels encore plus résolutifs.

Analysis of aldosterone-induced differential receptorindependent protein patterns using 2D-electrophoresis and mass spectrometry

Susanne Böhmer, **Christine Carapito**, Britta Wilzewski1, Emmanuelle Leize, Alain Van Dorsselaer and Rita Bernhardt

Biol. Chem., 2006, Vol. 387, pages 917–929

Pages 917 à 929 :

CHAPITRE II

Etude de l'exoprotéome du champignon filamenteux *Fusarium graminearum* se développant sur des parois de houblon

Cette étude a été réalisée en collaboration avec le Dr. Vincent Phalip et le Dr. Jean-Marc Jeltsch de l'équipe « Phytopathologie » du département « Biotechnologie des intéractions macromoléculaires » de l'UMR 7100 de l'Ecole Supérieure de Biotechnologie de Strasbourg.

1 Contexte de cette étude

1.1 L'interaction hôte-pathogène : le modèle *Humulus lupulus / Fusarium graminearum*

Les agents pathogènes responsables des maladies, certaines léthales, qui affectent les plantes, se caractérisent par leur nature (virus, bactérie, champignon), leur mode d'action et par les effets qu'ils provoquent. Mis à part leur rôle essentiel dans le recyclage des nutriments par dégradation des tissus végétaux [Evans et coll., 2001], les champignons filamenteux sont aussi les pathogènes les plus redoutés des plantes devant les virus et bactéries causant des pertes considérables dans l'agriculture mondiale [Lepoivre, 2003; Agrios, 1997]. Le laboratoire de phytopathologie de l'ESBS a choisi le couple hote-pathogène *Humulus lupulus / Fusarium graminearum* comme modèle d'étude des interactions hôtes-pathogènes [Hatsch, 2004].

Fusarium graminearum (téléomorphe *Gibberella zeae*) est un ascomycète de l'ordre des Hypocréales. Ce champignon filamenteux est réputé pour provoquer de nombreuses pathologies notamment dans les grandes cultures céréalières (blé, orge). En plus des pertes occasionnées par une diminution de rendement, la production de mycotoxines rend certaines récoltes impropres à la consommation humaine ou animale. Le genre *Fusarium* est régulièrement isolé dans des prélèvements de lianes de houblon en dépérissement.

Le houblon, *Humulus lupulus L.*, est une plante dioïque de la famille des Cannabinacées comportant deux genres : *Cannabis* et *Humulus*. Le plant femelle est cultivé pour ses inflorescences (cônes), qui sont l'épice par excellence des brasseurs. 80% de la production française est localisée en Alsace, où une variété locale, le Strisselspalt, est majoritairement cultivée.

Les interactions entre hôte et pathogène peuvent être nombreuses. *F. graminearum* ne disposant pas de structure lui permettant de franchir la paroi végétale, il doit trouver d'autres moyens pour pénétrer à l'intérieur de la plante. Dans le cas des céréales, il pénètre au niveau des plumeaux des jeunes épis. Cette zone de la plante n'est pas totalement développée et la paroi végétale en ce point est peu épaisse : il a été montré que l'épaisseur de la paroi végétale au niveau du site d'infection est diminuée [Wanjiru et coll., 2002]. Ainsi les enzymes de dégradation de la paroi cellulaire (les CWDE : Cell Wall Degradation Enzyme) libérées par le champignon jouent un rôle important dans le processus infectieux et le champignon dispose d'une grande collection de ces enzymes lytiques capables de lyser chaque couche de la paroi végétale : les cellulases, les xylanases et pectinases.

L'interaction entre les deux organismes sera donc examinée au niveau de la dégradation de la paroi cellulaire de la plante par le champignon.

1.2 Le génome et le protéome de *Fusarium graminearum*

La taille du génome de *Fusarium graminearum* est évaluée à 40 Mbps repartis sur 4 chromosomes et le nombre de gènes présents dans ce génome est estimé à 12000. Le projet de séquençage initié par le Whitehead Institute (*F. graminearum* Sequencing Project, Whitehead Institute/MIT Center for Genome Research) et soutenu financièrement par le département d'agriculture des Etats-Unis, a conduit au séquençage de la souche PH-1 (NRRL 31084), souche largement répartie au niveau mondial et impliquée dans les cas de pathologie. Un séquençage par la technique du "whole-genome shotgun sequencing" avec un taux de couverture de dix fois a été réalisé et la séquence du génome a été mise à la disposition de la communauté scientifique en mars 2003. Les premières annotations du génome ont été mises en ligne au Whitehead Institute ainsi que sur le serveur bioinformatique du laboratoire de phytopathologie (http://biotec.u-strasbg.fr).

Le protéome de *F. graminearum* est actuellement bien représenté dans les banques de données protéiques généralistes accessibles. En effet, en juillet 2006, la banque protéique UniProtKB contient 11823 entrées de *Gibberella zeae*. Par ailleurs, des banques d'ORFs prédits et traduits ont été générées au laboratoire de phytopathologie.

1.3 Objectif de l'étude

L'objectif de cette étude est l'analyse de l'exoprotéome de *F. graminearum* lorsque celui-ci se développe sur des parois de houblon afin de découvrir la nature de la machinerie de dégradation libérée par le champignon durant le processus d'infection de la plante. Cette étude constituera une première étape importante dans l'élucidation des mécanismes d'interactions *Fusarium* / plantes dicotylédones.

2 Choix de la stratégie expérimentale mise en oeuvre

2.1 Préparation des échantillons

Les parois cellulaires du houblon sont préparées comme précédemment décrit par Sposato et coll. pour la préparation de parois de maïs [Sposato et coll., 1995]. Le champignon *F. graminearum* est isolé à partir de plants de houblon en dépérissement avant d'être inoculé et mis en culture dans du milieu minimal M3 [Mitchell et coll., 1997] soit en présence de 1% de glucose soit en présence de 0,8% de préparation de parois cellulaires de houblons comme seule source de carbone. Après 6 et 9 jours de croissance de *F. graminearum* sur glucose et sur parois cellulaires, les surnageants de culture sont récupérés et les protéines déposées sur gels d'électrophorèse SDS-PAGE. Dans un premier temps, une approche par gel 1D a été choisie pour comparer la complexité des surnageants protéiques obtenus sur glucose par rapport à ceux obtenus sur parois cellulaires. Mais au vu de la complexité des surnageants protéiques des cultures sur parois, une stratégie par gel d'électrophorèse 2D a très rapidement été mise en place afin d'augmenter le pouvoir de résolution et d'identifier plus de protéines.

2.2 Analyse par spectrométrie de masse

Après coloration au bleu de Coomassie R-250, les gels 1D ont été découpés systématiquement en bandes de 2mm sur toute la hauteur du gel et 192 spots ont été carottés sur le gel 2D. Après lavage, réduction-alkylation et digestion in-gel des spots, les extraits peptidiques ont été analysés sur des couplages nanoLC-MS/MS : un couplage nanoLC Agilent Serie 1100 / trappe ionique HCT Plus (Bruker Daltonics) et un couplage CapLC / Q-TOF II (Waters). Plusieurs facteurs ont orienté le choix de la stratégie vers l'utilisation de la nanoLC-MS/MS plutôt que vers une stratégie par empreinte peptidique massique. D'une part, les spots de gel mono-dimensionnel étaient très riches en protéines, bien au-delà des limites de complexité d'échantillons analysables par MALDI-TOF-MS. D'autre part, la spécificité apportée par l'information de MS/MS était primordiale afin de pouvoir discriminer des protéines présentant des hauts degrés d'homologie de séquence, les enzymes de dégradation de la paroi cellulaire constituant de grandes familles de protéines.

2.3 Stratégie d'identification des protéines

Les données de nanoLC-MS/MS ont été soumises à des requêtes Mascot™ avec les paramétrages suivant : sans restriction taxonomique, avec un manquement de coupure trypsique toléré, 0,25Da d'erreur en MS et en MS/MS, la carbamidomethylation des cystéines et l'oxidation des méthionines spécifiées en modifications variables. Plusieurs banques de données ont été utilisées. Dans un premier temps, la banque protéique NCBInr a été interrogée. Puis dans un deuxième temps, les recherches ont également été réalisées dans la banque protéique FGBD, accessible sur le site http://mips.gsf.de/genre/proj/fusarium et dans une banque d'ORFs traduits générée au laboratoire de phytopathologie (http://biotec.u-strasbg.fr). Ceci pour minimiser les possibilités de manquer des identifications du simple fait de l'absence de la protéine dans la banque de données publique du NCBInr. Les résultats de ces recherches ont été confrontés et comparés : seules les protéines identifiées avec un minimum de 5 peptides présentant des spectres MS/MS de haute qualité (score Mascot supérieur à 40) ont été retenues sans ambiguïté. Dans les autres cas où moins de

peptides ont été identifiés, une analyse individuelle et manuelle des spectres MS/MS suivie d'une requête BLAST a été réalisée. De nombreuses protéines identifiées correspondaient à des ORFs traduits, prédits *in silico* mais pour lesquels aucune attribution de fonction n'avait encore était réalisée. Pour tous ces cas, des analyses BlastP ont permis l'identification des homologues les plus proches chez *Neurospora crassa*, *Ustilago maydis* ou encore *Magnaporthe grisea* et le programme FunCat (http://mips.gsf.de/projects/funcat) a été utilisé pour l'assignation des fonctions.

3 Résultats publiés

Les résultats obtenus dans ce travail ont fait l'objet d'une publication acceptée en octobre 2005 dans le journal *Current Genetics*.

Par opposition aux conditions de culture sur glucose, un riche arsenal de glycosyl hydrolases, de proteases et de lipases sont produites par le champignon lorsqu'il est mis en culture sur les parois de houblon. En effet, 84 protéines ont pu être identifiées dans les cultures sur parois de houblon parmi lesquelles plus de 40 sont classées par la classification FunCat dans les enzymes de dégradation des polysaccharides ou de catabolisme des carbohydrates. Seulement trois sur l'ensemble des protéines identifiées ne présentent pas de peptide signal de sécrétion ce qui confirme bien que la grande majorité des protéines identifiées ici sont réellement sécrétées.

En parallèle, la composition de la paroi cellulaire du houblon a été déterminée et la comparaison entre les enzymes identifiées dans cette étude comme étant spécifiquement sécrétées pour la dégrader et les constituants de cette paroi montre de très fortes corrélations.

Des enzymes appartenant à des grandes familles ont été identifiées ici et ce sont les hauts pourcentages de recouvrement d'identification des protéines mais aussi la spécificité apportée par la qualité des spectres MS/MS qui rendent la discrimination entre les différents membres, souvent proches en séquence, de ces grandes familles d'enzyme possible. Dans de telles applications, les exigences de qualité des données de spectrométrie de masse sont élevées et le choix de la stratégie à mettre en oeuvre doit tenir compte de ces exigences.

4 Conclusions et perspectives du travail

Au vu de ces résultats, l'analyse de l'exoprotéome apparaît comme une approche puissante pour l'étude des interactions entre champignons et hôtes et la connaissance des mécanismes d'attaque des champignons en fonction de l'hôte touché est un atout majeur pour la compréhension de ces pathologies de plantes.

Pour ce qui est des perspectives de ces travaux, la prochaine étape consistera en la caractérisation des exoprotéomes en additionnant un paramètre cinétique par la réalisation du même type d'expériences à différents stades du processus d'infection. Ainsi, une chronologie dans la sécrétion des enzymes attaquant les parois bactériennes pourra être établie ce qui permettra de déterminer les acteurs les plus précoces de la machinerie de dégradation.

Diversity of the exoproteome of *Fusarium graminearum* grown on plant cell wall

Vincent Phalip, François Delalande, **Christine Carapito**, Florence Goubet, Didier Hatsch, Emmanuelle Leize-Wagner, Paul Dupree, Alain Van Dorsselaer and Jean-Marc Jeltsch

Current Genetics, 2005, Vol. 48, Pages 366-379

Pages 366 à 379 :

CHAPITRE III

Etude de l'interactome virus-protéines hôtes recrutées lors de l'infection du riz par le virus de la Panachure Jaune du Riz

Cette étude a été réalisée en collaboration avec l'équipe du Dr Christophe Brugidou et Jean-Paul Brizard appartenant à l'UR génomique du riz de l'UMR CNRS/IRD/UP « Génome et développement des plantes » de Montpellier.

1 Contexte de cette étude

1.1 L'interaction virus-protéines hôtes : RYMV/*Oryza sativa*

1.1.1 Oryza sativa et ses agents pathogènes

Après le blé, le riz est la deuxième céréale cultivée dans le monde. Dans de nombreux pays, il est la principale source de calories dans l'alimentation humaine. Cependant, plusieurs agents pathogènes (champignon, bactérie, virus,...) sont des fléaux considérables pour la riziculture. Plus de quinze virus pathogènes du riz ont déjà été répertoriés [Hibino, 1996; Siré et coll., 2002]. Les dégâts les plus importants sont occasionnés par la maladie du Tungro en Asie (« Rice Tungro Bacilliform Virus » (RTBV)) associée au « Rice Tungro Spherical Virus » (RTSV), le « Rice Hoja Blanca Virus » (RHBV) en Amérique du sud et **la panachure jaune ou « Rice Yellow Mottle Virus » (RYMV)** en Afrique. De manière générale, les dégâts occasionnés par les virus se sont sévèrement accrus depuis les années 60, début de l'intensification de la riziculture, de la double culture et de l'introduction de variétés à haut potentiel agronomique [Hibino, 1996].

Parmi les différentes espèces de riz, deux sont principalement cultivées : *Oryza sativa L.* et *Oriza glaberrima Steud*. Elles appartiennent au genre *Oryza*, groupe *sativa* et sont diploïdes (2n=24). Au sein de l'espèce *Oryza sativa* on distingue deux variétés analogues à des sous-espèces, les groupes *indica* et *japonica* qui ne présentent pas la même sensibilité vis à vis des virus et notamment du RYMV. Les variétés du groupe *indica* cultivées en conditions irriguées sont très sensibles au RYMV. En revanche, les variétés du riz appartenant au groupe *japonica* pluvial sont partiellement résistantes. Bien que ce virus, au niveau mondial, ne soit pas économiquement important, il est dans certaines parties du continent africain un fléau grandissant qui justifie son étude. Les pertes de rendement varient entre 20% et 100% selon les années et les régions africaines [Awoderu, 1991; Abo et coll., 1998]. Ainsi, au cours des 20 dernières années, des programmes de sélection ont été développés dans le but de transférer la résistance du groupe *japonica* à des variétés productives, adaptées aux conditions irriguées.

1.1.2 Le virus RYMV (Rice Yellow Mottle Virus)

Le virus de la panachure jaune du riz (rice yellow mottle virus, RYMV) avec la pyriculariose et les nématodes sont les principales maladies du riz rencontrées en Afrique. La maladie se manifeste par des panachures jaunes sur la surface foliaire de la plante. L'infection commence souvent en bordure de champ, puis progresse vers l'intérieur pour envahir toute la parcelle. Ce virus est endémique à l'Afrique et a été décrit pour la première fois au Kenya [Bakker, 1970]. Il présente des caractéristiques communes avec les sobémovirus : un seul ARN simple brin de polarité positive non polyadénylé et de petite taille (Figure 1), constitué de particules de symétrie icosaèdrique T=3 (Figure 2) produites en très grande quantité dans la plante [Opalka et coll., 2000].

Figure 1 : Organisation génétique du RYMV.

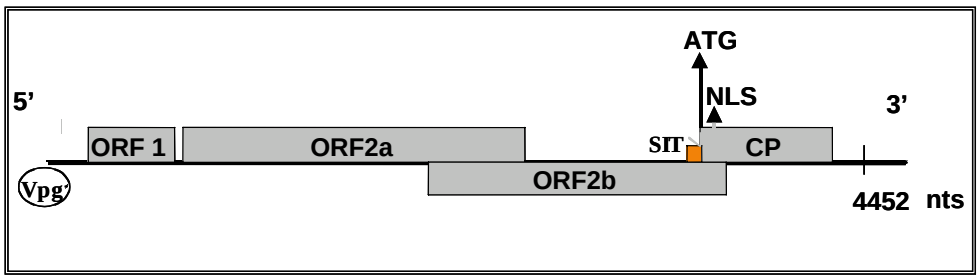
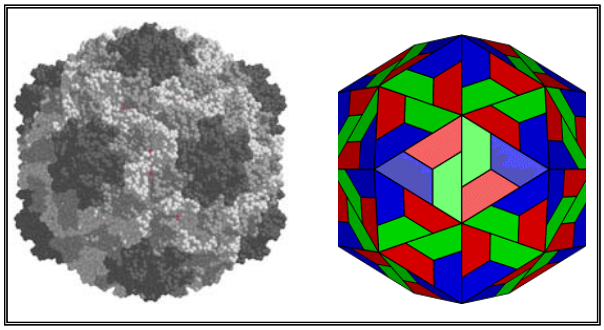


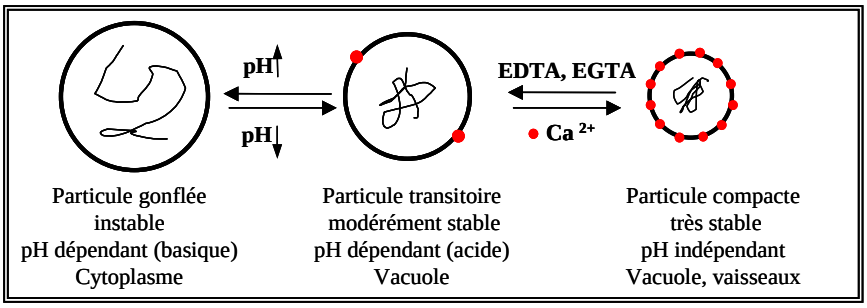
Figure 2 : Modélisation du RYMV, virus icosahédrique de symétrie T=3.



1.1.3 L'interaction virus-protéines hôtes

Au cours du cycle biologique du virus (réplication) et du cycle infectieux (migration dans la plante) dans la plante, les protéines de capside (CP) de la particule virale jouent un rôle déterminant notamment dans les étapes clés (infection, réplication, …) en interagissant avec des protéines hôtes. Le virus recruterait des protéines de la cellule hôte, par l'intermédiaire de ses protéines de capside, pour faciliter sa multiplication. Le RYMV existe sous trois isoformes différentes. L'équipe du Dr Brugidou a déjà montré que ces trois isoformes avaient probablement une localisation sub-cellulaire et une fonction biologique différente et a utilisé les caractéristiques du RYMV et notamment les propriétés électrostatiques des ses différentes isoformes (Figure 3) qui permettent dans certaines conditions d'accrocher des protéines de plantes [Brugidou et coll., 2002].

Figure 3 : Représentation schématique des différentes isoformes du RYMV et de leurs caractéristiques (stabilité et localisation dans la plante), d'après [Brugidou et coll., 2002].



1.2 Le génome et le protéome d'*Oryza sativa*

Le riz possède le plus petit génome connu chez les monocotylédones, environ 430 Mbps, répartis en douze chromosomes [Harumuganathan et coll., 1991]. Il est diploïde et dispose de peu de séquences répétées (moins de 40% contre 80% chez le blé).

En 2000, la firme Monsanto a annoncé la première ébauche de séquence du génome d'*Oryza sativa japonica*. Deux autres ébauches de séquences ont été publiées dans *Science* en avril 2002, l'une sur *O. sativa indica* réalisée par le BGI (Beijing Genomics Institute) [Yu et coll., 2002] et l'autre sur *O. sativa japonica* réalisée par la firme Syngenta [Goff et coll., 2002]. Un quatrième projet de séquençage du génome du riz a débuté en septembre 1997. Ce dernier projet est un projet international (IRGSP, *International Rice Genome Sequencing Project*) regroupant 10 nations (le Japon, le Brésil, la Chine, la Grande-Bretagne, l'Inde, la Corée, Taiwan, la Thaïlande, les Etats-Unis et la France). L'ensemble des données compilées (les trois premières ébauches et la séquence complète du génome du riz du projet IRGSP) met le riz au premier plan des génomes des végétaux séquencés. La carte génétique du riz est l'une des plus complète, après celle d'*Arabidopsis thaliana* pour les dicotylédones. Ses caractéristiques génétiques en font un excellent modèle d'étude et une plante de référence pour l'étude du génome des monocotylédones, de même qu'*Arabidopsis thaliana* est la plante modèle pour l'étude du génome des dicotylédones [Yuan et coll., 2001].

En juillet 2006, 1703 séquences protéiques sont annotées *Oryza sativa indica* et 121737 annotées *Oryza sativa japonica* dans la banque protéique du NCBI et UniProtKB compte 62588 entrées annotées *Oryza sativa* dont 842 appartenant à UniProtKB/SwissProt. La protéome du riz est donc largement représenté dans les banques protéiques accessibles. En parallèle, un vaste programme d'annotation du génome du riz est en cours au TIGR (The Institute for Genomic Research). Dans ce travail la banque de protéines prédites (TIGR Rice Pseudomolecules release 4) générée au TIGR et contenant 62827 entrées a été utilisée [Yuan et coll., 2005].

1.3 L'objectif de cette étude

Tandis que certaines interactions virus-protéines hôtes particulières sont bien décrites [Park et coll., 2001 ; Léonard et coll., 2000 ; Gurer et coll., 2002], aucune étude montrant le recrutement en parallèle et simultané de plusieurs protéines hôtes n'a été publiée à ce jour. Or pour faire face à la complexité du cycle de vie des virus (décapsidation-encapsidation, translation, replication et transport) et aux mécanismes de défense de la plante infectée, les virus recruteraient des protéines hôtes pour combler les fonctions protéiques nécessaires non codées dans leur génome. Dans ce contexte, les objectifs de cette étude sont les suivants :

- La démonstration du recrutement et de la spécificité du recrutement de protéines hôtes par le virus dès son entrée dans les cellules
- l'établissement d'une carte de référence de ces complexes virus-protéines hôtes

2 Choix de la stratégie expérimentale mise en oeuvre

2.1 Préparation des échantillons

Les plants de 2 cultivars de riz, IR64 (*Oryza sativa indica*) très sensible à l'infection du RYMV et Azucena (*Oryza sativa japonica*) partiellement résistante à l'infection par le RYMV ont été cultivés pendant 15 jours dans des conditions contrôlées sous serre. Après cette période, les plants ont été mécaniquement inoculés avec du RYMV purifié et les feuilles récupérées après 1, 2 et 3 semaines d'inoculation. Un protocole original utilisant la chromatographie d'exclusion stérique a permis la purification des complexes virus-protéines hôtes formés *in vivo* et *in vitro* et ces complexes ont ensuite été séparés sur gel d'électrophorèse 1D. Le détail des conditions de culture et des protocoles est décrit dans la publication des résultats en 3.

2.2 Analyse par spectrométrie de masse

Dans un premier temps, les bandes de gels 1D découpés systématiquement sur toute la hauteur des gels ont été analysées par MALDI-TOF-MS mais la complexité des échantillons a rapidement conduit à l'analyse de l'ensemble des spots par nanoLC-MS/MS sur le couplage CapLC-Q-TOF (Waters). En effet, à titre d'exemple, le tableau 1 illustre bien la limite de complexité des échantillons atteinte dans ces spots de gel pour l'analyse par MALDI-TOF-MS. Lorsque les mélanges de protéines sont trop complexes, l'identification par empreinte peptidique massique devient difficile et en général seule l'identification des 2-3

protéines les plus intenses est possible contrairement à l'analyse par nanoLC-MS/MS qui a permis l'identification de 9 protéines différentes dans ce spot.

Tableau 1 : Comparaison des résultats obtenus par les analyses MALDI-TOF-MS et nanoLC-MS-MS réalisées sur la même bande (spot 160) issue d'un gel 1D.

Spot 160	**MALDI-TOF-MS** (Analyse U11142CC)	**nanoLC-MS/MS** (Analyse Q010868CC)
	Pourcentage de recouvrement	Pourcentage de recouvrement
PHENYLALANINE AMMONIA-LYASE	15%	14%
PUTATIVE EXOGLUCANASE	17%	19%
HEAT SHOCK PROTEIN 70	~~	20%
PUTATIVE TRANSKETOLASE	~~	9%
PUTATIVE SUBTILISIN-LIKE PROTEASE	~~	8%
PEROXISOMAL MULTIFUNCTIONAL PROTEIN	~~	12%
BETA-D-GLUCAN EXOHYDROLASE	~~	10%
RUBISCO SMALL SUBUNIT	~~	15%
ARABINOXYLAN ARABINOFURANOHYDROLASE ISOENZYME AXAH-I	~~	2%

2.3 Stratégie d'identification des protéines

Les données de nanoLC-MS/MS ont été soumises à des requêtes Mascot™ dans la banque de données de protéines prédites pour *Oryza sativa* générées au TIGR (http://www.tigr.org/tdb/e2k1/osa1/). Des paramètres de recherche classiques ont été utilisés : 0,25Da d'erreur sont tolérés en MS et en MS/MS, un manquement de coupure trypsique est autorisé et l'oxidation des methionines et la carbamidomethylation des cystéines sont définies comme modifications variables.

3 Résultats publiés

Les résultats obtenus dans ce travail ont fait l'objet d'une publication acceptée en septembre 2006 dans le journal *Molecular & Cellular Proteomics*.

Les résultats relatifs à la spécificité de recrutement des protéines hôtes, les comparaisons entre variétés sensibles/résistances, les comparaisons de recrutement *in vivo/in vitro*, dans les systèmes non-hôtes spécifiques ainsi que la distribution fonctionnelle des protéines recrutées sont discutés en détail dans la publication.

Cependant, un aspect analytique particulièrement intéressant, seulement brièvement abordé dans la publication, est à retenir de cette étude. En effet, la protéine de capside (CP) qui sert d'appât pour le recrutement des protéines hôtes est présente en grande quantité dans l'ensemble des bandes de gel 1D sur toute la hauteur du gel. Cette présence ubiquitaire de la CP soulève un problème lors de l'analyse par nanoLC-MS/MS. Le fonctionnement par cycles 1 spectre MS suivi de 3 spectres MS/MS sur le couplage nanoLC-MS/MS ne permet pas la sélection de l'ensemble les ions présents pour la MS/MS, étant limité par les temps d'acquisition des spectres. Classiquement, l'instrument est paramétré de sorte à sélectionner les

ions les plus intenses en préférant les ions doublement chargés, espèces majoritaires dans le cas de peptides trypsiques et fragmentant le mieux (voir Chapitre II. 1.3 de l'introduction bibliographique). Ceci a conduit dans le cas présent à l'identification majoritaire des peptides trypsiques de la CP et a donc gêné, voire masqué, l'identification des autres protéines, les plus intéressantes mais plus faiblement présentes dans ces spots.

Pour contourner ce problème, un programme d'exclusion des peptides trypsiques de la CP a été généré. N'ayant aucun moyen d'éliminer « physiquement » les peptides de cette CP, la possibilité de créer dans les logiciels d'acquisition de données, des programmes d'exclusion permet de les éliminer « virtuellement » de la sélection pour la fragmentation. Pour cela :

- Les masses des peptides trypsiques di- et tri-chargés de la CP (numéro d'accession Swiss-Prot : Q9DGX1) ont été calculées *in silico* grâce au programme PeptideMass disponible sur le site ExPASY (http://ca.expasy.org/tools/peptide-mass.html).
- Une liste de 56 masses des peptides trypsiques di- et tri-chargés de la CP sur une gamme de m/z 400 à 1400 a pu être générée.
- Cette liste de masse est entrée dans un programme d'exclusion du logiciel MassLynx contrôlant l'acquisition automatique des données MS/MS sur le Q-TOF (Waters).

Afin d'évaluer l'impact de ce programme d'exclusion, l'analyse d'une série de spots a été réalisée en doublon avec et sans exclusion : le tableau 2 illustre l'exemple d'un spot dans lequel 2 nouvelles protéines ont pu être identifiées et les pourcentages de recouvrement des protéines précédemment identifiées sont significativement augmentés grâce à l'exclusion de la CP.

Tableau 2 : Comparaison des résultats de deux analyses faites sur la même bande de gel 1D avec et sans programme d'exclusion des ions di- et tri-chargés de la CP. L'utilisation du programme d'exclusion permet d'identifier deux nouvelles protéines. De plus, celles déjà identifiées sans programme d'exclusion le sont avec un meilleur recouvrement de séquence en peptides.

Spot 21	**Sans exclusion** (Analyse Q009595CC)		**Avec exclusion** (Analyse Q009558CC)	
	Nombre de peptides	% de recouvrement	Nombre de peptides	% de recouvrement
Protéine de capside Q9DGX1	**12**	**60%**	~~	~~
Protéine A	4	18%	10	47%
Protéine B	2	9%	8	35%
Protéine C	8	48%	10	57%
Protéine D	4	15%	7	33%
Protéine E	~~	~~	5	30%
Protéine F	~~	~~	5	23%

L’utilisation d’un tel programme présente toutefois un inconvénient car pour chaque masse exclue, une fenêtre d’exclusion doit être spécifiée afin d’éviter que les différents isotopes d’un même peptide ne soient sélectionnés. Ainsi en fixant une largeur de 2 Da, l’exclusion des 56 masses des peptides trypsiques de la CP revient à exclure 112 m/z. Il faut donc être vigilant et n’utiliser ce genre de programme que pour l’exclusion d’un nombre limité de masses afin d’éviter de se retrouver avec une trop large gamme de masse exclue.

4 Conclusion et perspectives

Cette étude a permis la mise en évidence du recrutement de protéines hôtes spécifiques par le virus RYMV lorsqu’il infecte le riz. Les protéines des complexes virus-protéines hôtes purifiés ont été identifiées par spectrométrie de masse et des expériences de validation fonctionnelle (par mutagénèse ou silencing) de certaines des protéines identifiées sont en cours. Dans tous les cas, le développement d’une telle stratégie peut être envisagé pour d’autres couples hôtes-virus et pourrait conduire à l’identification de nouvelles cibles thérapeutiques.

Dans la perspective de réalisation d’expériences de gene silencing, une analyse particulière de certaines des analyses de cette étude a été menée et sera décrite dans le chapitre III de la partie IV des résultats.

Proteome Analysis of Plant-Virus Interactome: Comprehensive Data for Virus Multiplication Inside their Hosts.

Jean Paul Brizard [1], Christine Carapito [2], François Delalande[2], Alain Van Dorsselaer [2], Christophe Brugidou [1].

1 Institut de Recherche pour le Développement, UMR 5096 (CNRS-IRD-UP), Montpellier, France. **2** Laboratoire de Spectrométrie de Masse Bio-Organique - UMR CNRS/ULP 7512, Strasbourg, France.

Running title: Proteomics of Virus-Plant Interactions

Corresponding author:
Jean-Paul Brizard
Centre IRD UMR 5096
911, Avenue Agropolis
BP64501
34394 Montpellier Cedex5, France
e-mail: brizard@mpl.ird.fr
Tel (0)4-67-41-62-39, Fax (0)4-67-41-61-81

Abbreviations:

RYMV	rice yellow mottle virus
CP	coat protein
wpi	weeks post infection
SCPMV	southern cow pea mosaic virus
FHV	flock house virus
PEG	polyethylene glycol
TAPS	N-[tris(hydroxylmethyl)methyl]-3-aminopropane-sulfonic acid

Summary

Known host-parasite molecular interactions are widespread among parasite families, but these interactions have to be particularly large considering the viruses which generally encode for few proteins. Since some particular virus-host interactions are well described, no global study yet showed multiple and simultaneous interactions in a host-parasite biological system. To prove that these multiple interactions occur in biological conditions, the complexes formed by a plant virus (Rice Yellow Mottle Virus) and the proteins of its natural host (rice) were extracted and purified from infected tissue sample. Remarkably, mass spectrometry permitted to identify a large number of proteins from the complexes which are involved in different functions not encoded by the virus but probably essential for its biological life cycle. This proteinic recruiting was strongly confirmed by the repetition of experiments using different pairs of virus-host, and the use of high salt concentration to extract the complexes. We mainly identified proteins involved in plant defense, in metabolism, translation and protein synthesis and some proteins involved in transport. This study demonstrates that viruses are able to recruit many proteins from their hosts to ensure their development. Among different pairs of virus-host, similar proteins functions were identified suggesting a particular importance of these proteins for viruses. The identification of particular paralog proteins among multigenic families suggests the high specificity of the recruiting for some protein functions.

Introduction

Although a new virus was discovered recently, the Mimivirus that encodes about 1,200 genes [1], viruses such as some phages, poxviruses, phycodnaviruses, iridoviruses and asfarviruses encode only about 500 genes and plant viruses belonging to the *Geminiviridae* and *Reoviridae* encode only 4 to 12 genes [2].

To cope with the virus life cycle complexity (decapsidation-encapsidation, translation, replication, and transport) and host defenses, viruses probably recruit host proteins to carry out needed functions that are not encoded by the viral genome. To confirm this virus recruiting, we have used the rice yellow mottle sobemovirus (RYMV) as virus model, and the rice (*Oryza sativa*) as plant model which is fully sequenced and widely used for genomic studies [3]. RYMV causes substantial economic losses in rice production in Africa [4]. Its distribution and biology are well known [5]. This virus is a monopartite positive RNA strand virus about 4450 nucleotides, and its genome organization is quite simple as it is composed of four ORFs that encode at least five proteins [5]. Three isoforms (compact, transitional and swollen) were identified according to the presence of divalent ions and pH, and were localized in different cell compartments such as cytoplasm, nucleolus, vacuole, vesicle and cell wall [6, 7].

Viral particles were also suspected to be localized in chloroplast [6].
In this paper, for the first time a range of analyses were reported, from the isolation of virus–host protein complexes to the identification of host proteins by mass spectrometry, allowing us to demonstrate the specific recruiting of numerous host proteins by the virus. A purification method of virus–host proteins complexes was developed from infected material, using gel exclusion chromatography, separation on SDS-PAGE, analysis by nanoLC-MS/MS after tryptic digestion and protein identification [8].
This new molecular methods could also be applied to human and animal viruses for the purpose of finding new therapeutic targets.

Experimental Procedures

Plant and virus materials

We used two rice cultivars showing a contrasted response to RYMV infection: a highly susceptible one IR64 (*Oryza sativa indica*), and a partially resistant and tolerant one Azucena (*Oryza sativa japonica*). IR64 is a high-yielding cultivar developed at the International Rice Research Institute (IRRI) and Azucena is a traditional upland cultivar from the Philippines. For both cultivars, seeds were sown separately and plants were grown in a confined greenhouse in controlled conditions: 12 h light at 28°C and 12 h dark at 24°C. Two weeks after seedling, plants were mechanically inoculated with purified RYMV particles from a virulent isolate of Burkina Faso (BF1) at a concentration of 100µg/ml in inoculation buffer (20 mM phosphate, pH 7) as described previously [9]. This experiment was repeated twice. Leaves of non-inoculated and RYMV inoculated plants were harvested at 1, 2 and 3 weeks post inoculation (wpi), frozen in nitrogen and conserved at - 80°C.
To compare the specificity of the recruited host proteins we used *Phaseolus vulgaris* (non host for RYMV and host for Southern cowpea mosaic virus, SCPMV) and *Nicotiana tabacum* (non host for RYMV). Infected and/or non- infected plants were cultivated and harvested in the same conditions as rice cultivars.
Different viruses were used in our experiments: RYMV isolate BF1 (access. Q9DGX1 Swiss Prot Database) was used to infect rice cultivars, for *in vitro* binding experiments with rice host plant and with *N. tabacum* as non host plant. SCPMV belongs to the same genus *Sobemovirus*, but has a different genomic organisation and a limited dicotyledonous host-range [10]. Flock house virus (FHV) is a member of the insect and animal virus family *Nodaviridae*.

RYMV extraction. Rice leaves were ground in liquid nitrogen and homogenized in 0.1M phosphate buffer, pH 5.0. Virus particles were precipitated with 6% PEG 8000 and resuspended in phosphate buffer. Further purification was performed by centrifugation through a 10 to 40% sucrose gradient. Virus concentration was estimated by spectrophotometry using an extinction coefficient of 6.5.

Plant Protein Extraction. About 10 g of frozen leaves were crushed in a liquid nitrogen-cooled mortar, and 50 ml of buffer extraction pH7.5 (Tris-HCl 50 mM, NaCl 100 mM, EDTA 10 mM, Glucose 25 mM, EGTA 5 mM, DTT 5 mM, Glycerol 5%, 0.1% Triton X100, Protease Inhibitor Cocktail Tablets (Roche Applied Science) added to the resulting powder. The suspension was centrifuged at 22000 g for 30 min at 3°C and the supernatant was filtered through a 0.45µm syringe filter.

Extraction of Virus-Protein Complexes. Frozen infected leaves (1, 2, 3 wpi) were treated with the protocol used for plant protein extraction (see below), then concentrated to 10 ml by ultra filtration with Centricon Plus 20 (Millipore) at 3°C.

***In vitro* Virus-Proteins Binding Assay.** 3.5 mg of purified virus were added to 100 mg proteins extracted with buffer as described above, then concentrated to 10 ml by ultra filtration with Centricon Plus 20 at 3 °C. Contact between virus and proteins were approximately 90 min until injection on chromatographic column. All purification steps were done in a cold room at 4°C.

Purification of Virus-Protein Complexes. The concentrated mixtures were injected on a 1000 mm long/26 mm diameter column filled with Sephacryl S500 (Amersham Biosciences), and TAPS buffer pH 7.6, 100 mM NaCl as eluent. The ÄKTAprime system (Amersham) was used for injection, detection and collection of the proteins. Fractions corresponding to the first 2/3 peak of virus-proteins complexes were collected and lyophilized (c_1, Figure 1).

SDS-PAGE of Virus-Proteins Complexes. The lyophilized pellets were resuspended with deionized water (Milli-Q) and the resulting solution was desalted using PD10 columns (Amersham Biosciences). The amount of protein was estimated using the BCA Protein Assay Kit (Pierce) and about 40 µg of proteins were denatured for 5 mn at 100°C after adding of 2X SDS-loading buffer (100mM Tris-HCl pH6.8, 200mM DTT, 4% SDS, 0.2% Bromophenol Blue, 20% Glycerol) and loaded on pre-cast gel 4-12 % acrylamide (Invitrogen).
Gels were stained with colloidal Coommassie blue, and sample bands were obtained from gels under

sterile conditions to avoid keratin contamination (Figure 2).

Mass Spectrometry and Protein Identification

In-gel digestion was performed with an automated protein digestion system, MassPREP Station (Waters, Milford Massachusetts, USA). The gel plugs were washed three times with a mixture of 50%/50% NH_4HCO_3 (25 mM)/ACN. The cysteine residues were reduced with dithiothreitol at 57°C and alkylated with iodoacetamide. After dehydration with acetonitrile, the proteins were digested in-gel with 30 µL of 12.5 ng/µL of modified porcine trypsin (Promega, Madison, WI, USA) in 25 mM NH_4HCO_3 overnight and at room temperature. Then a double extraction was performed**,** first with 60% acetonitrile in 5% formic acid and a second one with 100% acetonitrile. The resulting peptide extracts were directly injected for nanoscale capillary liquid chromatography-tandem mass spectrometry (nanoLC-MS/MS) analysis.

NanoLC-MS/MS analysis of the digested proteins was performed using a CapLC capillary LC system (Waters, USA) coupled to a hybrid quadrupole orthogonal acceleration time-of-flight tandem mass spectrometer (Q-TOF II, Waters). Chromatographic separations were conducted on a reverse-phase (RP) capillary column (Pepmap C18, 75µm i.d., 15 cm length, LC Packings) with a 200 nL/min flow rate.

Mass data acquisitions were piloted by MassLynx software (Waters, USA) using automatic switching between MS and MS-MS modes as described [11,12]. To improve the quality of MS/MS spectra during nanoLC-MS/MS analysis, we empirically derived energy curves depending on the m/z value of the selected precursor ion: for each m/z value, 3 different collision energies were applied. Fragmentation was performed using argon as the collision gas.

The mass data recorded during nanoLC-MS/MS analysis were processed and converted into MassLynx .pkl peak lists format prior to searching with the search engine Mascot (Matrix Science, London, UK). The searches were performed on a local Mascot server running on a 3 GHz Pentium IV processor with a tolerance on mass measurements of 70 ppm in MS mode and 0.25 Da in MS/MS mode. One missed cleavage per peptide was allowed and some variable modifications were taken into account such as carbamidomethylation for cysteine, oxidation for methionine and N-acetylation of the protein. Searches were performed without constraining protein molecular weight or isoelectric point. The Rice (Nippon Bare) pseudomolecule database (release 3) accessible from the TIGR (The Institute for Genomic Research http://www.tigr.org/tdb/e2k1/osa1/) website was used for the identification of rice proteins and a compilation database gathering SwissProt, TrEMBL and TrEMBLnew protein databases for the identifications of the other host plant proteins (*Phaseolus vulgaris* and *Nicotiana tabacum*). A total of 531 proteins were annotated for *Phaseolus vulgaris* and 1988 proteins for *Nicotiana tabacum* in this compilation database (July 2005).

Our rice protein identification method using the complete pseudomolecule database, allowed us, in some cases, to discriminate paralog genes among multigenic families [8].

Due to the strategy developed here, the virus coat protein was present in all gel bands and in very high concentration independently of the spot position on the gel. To circumvent the superabundance of the viral coat protein, we used an informatics exclusion program to prevent the selection of the coat protein peptides during the nanoLC-MS/MS analysis. This procedure allowed the identification of lower abundant proteins.

A python language script was used to extract data from files generated by Mascot software, and the Access Data Base software (Microsoft) was used to gather and compare datasets according to the different conditions of the experiments. In this study, we validated an identification when the protein was identified by at least two peptides both having an MS/MS ion score higher than 39.

Results

Methodology used to isolate RYMV host protein complexes *in vivo* and *in vitro*

RYMV-host protein complexes were isolated from infected plants (Figure 1A) or from *in vitro* binding experiments with soluble proteins extracted from non infected leaves supplemented with purified virus (Figure 1B).

According to our protocol (see Experimental Procedures), after injection of purified virus (6.5 mg) we detected an elution peak at 216 minutes. The first elution peak of soluble proteins extracted from non infected *Oryza sativa* (IR64) leaves, which was the control without virus (negative control), was eluted at 255 minutes (Figure 1A, green) and overlapped slightly with the peak at 216 minutes corresponding to the experiment with the virus purified alone (Figure 1A, c2, red). The eluted profile of soluble proteins extracted from infected leaves showed an additional peak at 206 minutes corresponding to virus-protein complexes (Figure 1A, c1 blue). The same additional peak at 206 minutes was found in experiments using the buffer extraction added with 0.5M or 1M NaCl (Figure 2). In experiments with *in vitro* virus-host protein complexes, the same additional peak at 206 minutes was detected at a lower concentration in accordance with the amount of added purified virus (3.5 mg) (Figure 1B, c1). After the purification/elution cycle, electron microscopy revealed the presence of intact virus particles in the

collected fraction corresponding to peak c2 (Figure 1C). In collected fraction corresponding to peak c1, virus particles were associated with electron dense materials (Figure 1D). This method allowed us to isolate *in vivo* and *in vitro* virus-host protein complexes and was further used to identify virus-host protein partners by nanoLC-MS/MS.

SDS-PAGE separation

At different times of plant development (3, 4 and 5 weeks post seedling, wps) and during the time course of infection (1, 2 and 3 weeks post infection, wpi) virus–host protein complexes were isolated from *in vivo* and *in vitro* experiments, denatured and separated by SDS-PAGE gel (Figure 3A). Gel protein profiles were reproducible when virus, plant varieties and stage of development were identical (data not shown). On the contrary the protein profiles for *O. sativa* IR64 variety (Figure 3A) differed, according to the stage of development and the time course of infection for *in vivo* experiments or stage of development for *in vitro* experiments. When comparing the same stage of development, we observed specific profiles for *in vivo* and *in vitro* experiments. This observation supports the idea that complexes isolated from infected plants were already present in the *in vivo* tissues, and were not formed during the protein extraction process. Thus, we observed for visible bands as a, b, c, d and e, different accumulations between *in vivo* and *in vitro* experiments at different stages of development (Figure 3A). To identify these proteins, systematic cutting and nanoLC-MS/MS analysis of these gel lanes were performed.

Identification of virus host protein partners by mass spectrometry

As expected, we have identified the RYMV coat protein (CP, accession number: Q9DGX1) around 26 kDa in accordance with the results obtained by western blot using monoclonal antibody against the virus (data not shown). For the same total amount (20 µg) of proteins loaded on each lane, the amount of CP increased during the time course of infection, whereas, equal amounts were observed for complexes purified from *in vitro* experiments (Figure 3A).

The virus coat protein was present at such a high concentration that it was identified in all gel bands independently from location in the gel. As the nanoLC-MS/MS automatic acquisition program selected the most prevalent peptides for fragmentation, the high abundance of the coat protein prevented the identification of less abundant proteins. To circumvent this problem, we have established an exclusion program to informatically exclude the masses of the coat protein tryptic peptides, thereby preventing their selection and allowing the selection of peptides belonging to other proteins [8]. As shown in Figure 3B, 137 gel bands were analyzed by nanoLC-MS/MS for the experiments with the IR64 cultivar with RYMV and 2017 proteins were identified in the rice pseudomolecules release 3 database (data not shown). Nevertheless, most of the proteins were identified several times and we presented here only the 223 non-redundant identified proteins (Table 1).

Functional distribution of recruited host proteins

For the 223 identified proteins using the IR64 cultivar, the following distribution was found: 19 % were identified only *in vivo*, 41 % were identified both in *vivo* and *vitro*, and 40 % were present only *in vitro*. The major functional category corresponded to proteins involved in metabolism functions (Me, Figure 4), mainly glycolysis, photosynthesis, amino acid, lipid and cell-wall metabolism. The second category corresponded to functions involved in translation and protein synthesis (T) including translation factors, elongation factors, tRNA synthetases, protein disulfide isomerase, chaperone proteins, and proteasome. The third category was related to defense (D), with protein chaperones (i.e. 70, 82 and 90kDa), proteins involved in defense pathways such as superoxide dismutase (SOD), phenylalanine-ammonia lyase (PAL), homocystein S-methyltransferase, lipoxygenase, proteins related to oxidative stress with thioredoxin, peroxiredoxin, oxidoreductase NAD binding, glutathion-S transferase as well as pathogenesis related proteins including peroxidase and chitinases. Other categories, less represented, were related to unknown functions (U), transport (Tp) and transcription (Tr).

Proteins identified only in *in vivo* experimentations (19 %) might be explained by the specificity of a recruiting made *in planta*, and by the inability for these already coated virus particles to recruit more proteins during the extraction process. Proteins identified only *in vitro* experimentations (40 %) might be explained by a recruiting due to the solubilization of proteins that were not accessible for the virus *in planta*. Some of these proteins were identified in *in vivo* experimentations with Azucena cultivar (i.e. putative aldehyde oxidase 11669.m05835) and might be involved in a specific recruiting *in planta*, but for the other we could not exclude an artifact recruiting. For the common proteins, it was highly probable that they could be recruited *in planta*, and they were recruited *in vitro* because they are extracted in *in vitro* experimentations.

Paralogs Identification

In some cases, we were able to discriminate which paralog gene among a multigenic family was a specifically recruited and present in virus-host

protein complex (see Material and Methods). This was the case for the paralog 11668.m03971 for phenylalanine ammonia-lyase (Defense category) specifically identified among 10 paralogs (TIGR database) from this multigenic family. In addition, some specific paralogs were recruited only *in vivo* such as peroxidase, putative 11682.m00373, others *in vitro* such as reversibly glycosylated polypeptide 11670.m05573. Additionally, we observed the same paralogs identified *in vivo* and *in vitro* (peroxidase, putative 11667.m02169 for example). Interesting was the example of G3PDH (Glyceraldehyde-3-phosphate dehydrogenase in Metabolism category) where the number of recruited paralogs varied with the experimental conditions: 4 paralogs (Table 1) were recruited *in vivo* and *in vitro* by the virus at one week post infection, 3 paralogs were still recruited *in vivo* at 2 wpi instead of 4 *in vitro*, and finally no paralog from G3PDH was recruited *in vivo* at 3 wpi whereas 3 paralogs were identified *in vitro* at 3 wpi.

These results suggested that some specific paralogs among multigenic families were recruited at specific points during the time course of virus infection.

Specificity of host proteins recruiting

In order to investigate the specificity of host proteins recruiting by RYMV, we have used the following strategy: (i) We used different NaCl concentrations to extract and isolate complexes from *in vivo* or *in vitro* experiments in order to confirm that some of the protein interactions occurred at low ionic strength, and may have been unspecific (ii) We then investigated whether the recruited proteins were host and virus dependent by testing different virus-host pairs.

Varying NaCl concentration: Protein profiles were quite different according to the salt concentration used in *in vivo* and *in vitro* extractions (Figure 5A). As expected, the amount of CP from *in vitro* experiments at different salt concentrations was similar. On the contrary, in *in vivo* experiments we observed a higher amount of CP at 0.5 M and 1M NaCl, suggesting that a higher amount of complexes was extracted with a high ionic strength buffer, probably due to the better extraction of some subcellular compartments. The higher number of proteins identified at 0.5 M and 1M of NaCl confirmed this result, specifically the ribosomal proteins which were localized in nucleolus, mitochondria and chloroplasts and also histon proteins which were localized in nucleus (*in vivo* and *in vitro*, Figure 5B), and furthermore, proteins from proteasome which localized to cytoplasm are well identified at low as well as high ionic strength (as cytoplasmic compartment is easier to extract) (Table 2). We confirmed that the amount of proteins extracted from Azucena cultivar infected leaves showed an 18% increasing when using buffer added with 0.5M and 1M NaCl compared to 0.1M NaCl (data not shown). We noticed a decrease in the number of proteins that were common with proteins identified at 0.1 M NaCl due to the possible abolition of low ionic strength interactions. We observed an expected decrease of total proteins at 1 M of NaCl (151 vs 163 at 0.5 M for *in vivo* and 62 vs 93 at 0,5M for *in vitro*). The common proteins that were still identified at 1M NaCl suggest that the binding was specific and that they strongly interacted with virus particles (Table2).

Our results suggested two effects of salt concentration. First, an increase of salt concentration to 0.5 M increased the number of proteins identified in complexes, resulting from a better extraction of complexes due to the ability of high salt concentration to extract proteins highly bounded to the membranes [13, 14], and to break membrane compartments (osmotic pressure) but not solubilise the membranes as a detergent could do it. On the contrary, the increase of salt concentration to 1M reduced the number of common proteins that were also identified at 0.1M NaCl, suggesting that proteins still identified at high salt concentration were strongly specific of the complexes.

Different pairs of virus-host plant: To see further whether this affinity was host and virus dependent, we studied complexes extracted and purified from various virus-hosts combinations (Figure 6).

First, we compared at 1, 2 and 3 wpi the *in vivo* interaction of RYMV with two sub-species of *O.sativa*: the susceptible *O. sativa* ssp *indica* (IR64) and the tolerant *O sativa* ssp *japonica* (Azucena) (Figure 6A). More proteins were identified for Azucena than for IR64 (171 vs 135 proteins identified when we cumulated 1, 2 and 3 wpi identifications). Among these proteins, a large number (100) were common for both sub-species (Table 1), confirming that the recruiting among sub-species was quite similar. It was likely that the proteins identified differentially in Azucena and IR64 take part to the tolerance or susceptibility of these subspecies (Table 1).

We further investigated the recruiting between different viruses and different host-plants: (i) a compatible interaction between another sobemovirus (SCPMV) and his host plant *Phaseolus vulgaris* (ii) three incompatible interactions, one between an insect virus FHV and IR64, and two other pairs: RYMV-*N. tabacum* and RYMV-*P. vulgaris.* We identified some identical protein functions among all the different complexes (Figure 5B), but we could not precisely conclude about the percentage of similarity of these recruiting activities due to the lack of the protein databases concerning these two others plants (Table 1).

For the interaction between FHV and IR64, we identified 41 common proteins also found with the interaction RYMV-IR64 (Table 1), what could explain the ability of FHV to replicate in rice [9], and suggest that a set of proteins from a host plant could interact with different viruses.
The results obtained with RYMV, FHV and SCPMV suggested that host protein recruiting occurred for different viruses and that two unrelated viruses (RYMV and FHV) could recruit the same proteins from the same plant.

Discussion

We know that the viral genome encodes a small number of genes, and the virus is thus necessarily dependent on host machinery to achieve its life cycle. The recruiting of the translational machinery from host plants is now well established for different viruses, for instance in the case of Potyviruses, the interaction of eIF4E, eIFiso4E, eiF4G and the viral protein VPg [15,16,17,18]. Nevertheless, the question of how the virion acts in the cellular context remains.
Indeed, for each stage of the virus life cycle, it is necessary to have the required host partners in the right conditions of time, concentration, localization and conformation. It is also necessary for the virus to have a high affinity for some host proteins. The external part of virus particles could play this crucial function in recruiting host proteins. To demonstrate this recruiting, we developed a method using RYMV having very stable particles [7] and replicating to a high level [4]. Thus, using size exclusion chromatography, virus-protein complexes were purified from infected plants, and from *in vitro* binding experiments using purified virus and soluble proteins extracted from non-infected plants. This method is reproducible and allows us to purify enough material to analyze complexes by SDS–PAGE and nanoLC-MS/MS. Host proteins from the complexes were separated and gave reproducible protein profiles for the same experimental conditions. To demonstrate the specific recruiting by the virus, we studied three critical stages for RYMV: 1 wpi (beginning of replication), 2 wpi (replication in systemically infected leaves, first symptoms) and 3 wpi (end of viral replication in susceptible variety IR64 and development of symptoms) [4]. We showed that the recruiting of host proteins were different according to the infection stages. Comparing the same infection stage *in vitro* and *in vivo*, we demonstrated that the complexes isolated from infected plants were not formed during the extraction, but presumably, preexisted *in vivo* during the infection process. The number of identified proteins for each stage (1, 2 and 3 wpi) corresponds to a population of different complexes representative of the global situation within infected plants, as the virus has the ability to infect different cell compartments. Looking at the stained SDS-PAGE gels, we clearly observed some recruited proteins showing different quantitative profiles according to the different stages of infection (band a, b, c, d and e, Figure 2). Because of the reproducibility of this recruiting, these results reinforce the idea that virus recruits specific host proteins during the infection process.
Among the recruited proteins, some of them have a higher affinity for RYMV. This was supported by the identification of 32 proteins that were still binding to the virus at 1 M of NaCl *in vivo* (out of 72 at 0.1M NaCl for IR64 2 wpi), and among them, 25 proteins were identified *in vitro* at 1M of NaCl (Figure 5B). We suppose that these proteins are most likely bound directly at the surface of the virus particle. The other proteins that were identified in lower salt concentration could have a lower affinity with the surface of the virus or could bind host proteins previously recruited by the virus.
Interestingly, we found a recruiting coherence through some functional categories. In metabolism category we identified a high number of enzymes involved in glycolysis, malate and citrate cycles, presumably recruited by the virus to produce energy for virus replication. In the defense category, we identified proteins involved in ROS (reactive oxygen species) [19] and detoxification (superoxide radical and hydrogen peroxide) which are presumably recruited by the virus to maintain an oxido-reduction environment compatible with viral replication. In addition, in the protein synthesis category, we identified a set of proteins involved in translation processes with ribosome, elongation factors, protein chaperones, protein disulfide isomerase and proteins involved in protein turn-over with the proteasome 20S. All these proteins would be recruited by the virus to optimize virus translation efficiency, as soon as virus starts decapsidation.
Nevertheless we can rule out that some proteins identified belong to plant defense system, and were neutralized by sequestration.
Virus recruiting is likely to occur for other viruses, as we saw a recruiting for the couple SCPMV-*Phaseolus vulgaris* (Figure 6B). RYMV is able to recruit proteins from *Nicotiana Tabacum* and from *Phaseolus vulgaris* in *in vitro* experimentations (Figure 6B). The results obtained *in vitro* with the pair FHV-IR64 are interesting since 41 proteins were identified that where also found with the pair RYMV-IR64 (*in vivo* and/or *in vitro*), suggesting a common set of proteins recruited from rice by viruses able to replicate in this plant (it was demonstrated that the FHV is able to replicate and is systemically spread in rice plants expressing movement protein genes [20]). Inside the rice host sub-species studied, we have shown that most recruited proteins are identical (Figure 6A), the proteins recruited that are different between the two sub-species could be related to the susceptibility or

tolerance effect observed for the IR64 and Azucena varieties.

Some proteins are identified whatever the experimental conditions (i.e. mitochondrial chaperonin60 11676.m02851), suggesting that they have a very strong affinity for viruses and that they play an important role in the virus biological process.

We discriminated some paralog genes belonging to a multigenic family that were specifically recruited. Some of the proteins identified in the complexes from the IR64-RYMV and Azucena-RYMV experiments were shown as deregulated in IR64 and Azucena suspension cells undergoing RYMV infection [11]. About fifteen similar protein functions were found in other study using cDNA-AFLP to discover genes induced or repressed during virus infection (unpublished data).

Some of the proteins have been identified in different virus–host interactions such as Hsp 60 with Hepatitis B virus [21] or HIV [22], Hsp70 with plant closteroviruses [23] or HIV [24] what suggests that viruses may recruit the same protein functions in different hosts. Other proteins, non identified in our experiments, were identified interacting with different viruses, as pectin methylesterases [25,26], homeodomain proteins [27], rab acceptor-related proteins [28], β-1,3-glucanase-interacting proteins [29], Fas-mediated apoptosis enhancer Daxx [30], SUMO-1 protein [31,32].

All these results suggest that the recruiting of proteins is probably a common process for different viruses.

In this paper we described for the first time an efficient method to extract virus-host proteins complexes and to identify by mass spectrometry the proteins involved in these complexes. The analysis with contrasted pathogenic isolates, in host and non-host interaction contexts, should help to identify molecular interaction mechanisms involved in viral infection. The functional relevance of these proteins remains to be evaluated using mutagenesis or silencing strategies. This method of analysis may help to identify new target proteins that may be useful to find new markers for plant selection, or to develop new strategies to abort virus infection processes. It is also conceivable to use this experimental approach to isolate virus/host proteins complexes from different organisms in an attempt to find new therapeutic targets in human and animal virus diseases.

Acknowledgments

We thank David Biron, Marc Van Regenmortel, Claude Fauquet and Eugénie Hébrard for helpful comments on this manuscript. We also wish to thank Gaël Lagrange for help with bioinformatics tools, Denis Fargette for providing the RYMV isolate BF1, David Hacker for providing the SCPMV and Anne Schneemann for providing the FHV. Financial support for this research was provided by the Institut de Recherche pour le Développement (IRD).C.C. was sponsored by Bruker Daltonics and CNRS, F.D. by Aventis.

REFERENCES:

1. Raoult, D., Audic, S., Robert, C., Abergel, C., Renesto, P. et al. (2004) The 1.2-megabase genome sequence of mimivirus. *Science* **306**(5700), 1344-1350
2. Fauquet, C. M., Mayo, M. A., Maniloff, J., Desselberger, U., Ball, L. A. (2005) *Virus taxonomy: eighth report of the international committee on taxonomy of viruses*. San Diego, CA, Elsevier Academic Press (2005).
3. Delseny, M., Salses, J., Cooke, R., Sallaud, C., Regad, F. et al. (2001) Rice genomics: present and future. *Plant Physiol Biochem* **39**, 323-334 (2001).
4. Kouassi, N.K., N'Guessan, P., Albar, L., Fauquet, C., Brugidou, C. (2005) Distribution and characterization of rice yellow mottle virus: a threat to african farmers. *Plant Disease* **89**, 124-133
5. Fargette, D., Pinel, A., Abubakar, Z., Traore, O., Brugidou, C. et al. (2004) Inferring the evolutionary history of rice yellow mottle virus from genomic, phylogenetic, and phylogeographic studies. *J Virol* **78**(7), 3252-61
6. Opalka, N., Brugidou, C., Bonneau, C., Nicole, M., Yeager, M. et al. (1998) Movement of rice yellow mottle virus between xylem cells through pit membranes. *Proc Na. Acad Sci USA* **95**, 3323-3328
7. Brugidou, C., Opalka, N., Yeager, M., Beachy, R.N., Fauquet, C. (2002) Stability of rice yellow mottle virus and cellular compartmentalization during the infection process in *Oryza sativa* (L.). *Virology* **297**, 98-108
8. Delalande, F., Carapito, C., Brizard, J. P., Brugidou, C., Van Dorsselaer, A. (2005) Multigenic families and proteomics: extended protein characterization as a tool for paralog gene identification. *Proteomics* **5**(2), 450-60
9. N'Guessan, P., Pinel, A., Caruana, M. L., Frutos, R., Sy, A. A. et al. (2000) Evidence of the presence of two serotypes of rice yellow mottle sobemovirus in Côte d'Ivoire. *Eur J Plant Pathol* **106**, 167-178
10. Tamm, T., Truve, E. (2000) Sobemoviruses. *J Virol* **74**(14), 6231-6241
11. Ventelon-Debout, M., Delalande, F., Brizard, J. P., Diemer, H., Van Dorsselaer, A. et al. (2004) Proteome analysis of cultivar-specific deregulations of Oryza sativa indica and O. sativa japonica cellular suspensions undergoing rice yellow mottle virus infection. *Proteomics* **4**(1), 216-25
12. Richert, S., Luche, S., Chevallet, M., Van Dorsselaer, A., Leize-Wagner, E. (2004) About the mechanism of interference of silver staining with peptide mass spectrometry. *Proteomics* **4**(4), 909-16
13. Robertson, D., Mitchell, G. P., Gilroy, J. S., Gerrish, C., Bolwell, G. P. et al. (1997) Differential extraction and protein sequencing reveals major differences in patterns in primary cell wall proteins from plants. *The Journal of Biological Chemistry* **272**(25), 15841-8
14. Usman, L. A., Ameen, O. M., Ibiyemi, S. A., Muhammad, N. O. (2005) The extraction of proteins from the neem seed (*Indica azadirachta A. Juss*). *African Journal of Biotechnology* **4**(10), 1142-4
15. Wittmann, S., Chatel, H., Fortin, M.G., Laliberte, J.F. (1997) Interaction of the viral protein genome linked of turnip mosaic potyvirus with the translational eukaryotic initiation factor (iso) 4E of arabidopsis

thaliana using the yeast two-hybrid system. *Virology* **21**,234(1), 84-92
16. Léonard, S., Plante, D., Wittmann, S., Daigneault, N., Fortin, M. G. et al. (2000) Complex formation between potyvirus VPg and translation eukaryotic initiation factor 4E correlates with virus infectivity. *J Virol* **74**(17), 7730-7737
17. Ruffel, S., Dussault, M. H., Palloix, A., Moury, B., Bendahmane, A. et al. (2002) A natural recessive resistance gene against potato virus Y in pepper corresponds to the eukaryotic initiation factor 4E (eIF4E). *Plant J* **32**(6), 1067-75
18. Yoshii, M., Nishikiori, M., Tomita, K., Yoshioka, N., Kozuka, R. et al. (2004) The arabidopsis cucumovirus multiplication 1 and 2 loci encode translation initiation factors 4E and 4G. *J Virol* **78**(12), 6102-6111
19. Mittler, R. (2002) Oxidative stress, antioxidants and stress tolerance. *Trends Plant Sci* **7**(9), 405-10
20. Dasgupta, R., Garcia, B.H.II, Goodman, R.M. (2001) Systemic spread of an RNA insect virus in plants expressing plant viral movement protein genes. *Proc Natl Acad Sci USA* **98**(9), 4910-4915
21. Park, S.G., Jung, G. (2001) Human hepatitis B virus polymerase interacts with the molecular chaperonin Hsp60. *J Virol* **75**(15), 6962-8
22. Bartz, S. R., Pauza, C.D., Ivanyi, J., Jindal, S., Welch, W. J. et al. (1994) An Hsp60 related protein is associated with purified HIV and SIV. *J Med Primatol* **23**(2-3), 151-154
23. Alzhanova, D.V., Napuli, A.J., Creamer, R., Dolja, V. V. (2001) Cell-to-cell movement and assembly of a plant closterovirus: roles for the capsid proteins and Hsp70 homolog. *EMBO J* **20**(24), 6997-7007
24. Gurer, C., Cimarelli, A., Luban, J. (2002) Specific incorporation of heat shock protein 70 family members into primate lentiviral virions. *J Virol* **76**(9), 4666-70
25. Dorokhov, Y. L., Makinen, K., Frolova, O. Y., Merits, A., Saarinen, J. et al. (1999) A novel function for a ubiquitous plant enzyme pectin methylesterase: the host-cell receptor for the tobacco mosaic virus movement protein. *FEBS Lett* **461**(3), 223-8
26. Chen, M. H., Citovsky, V. (2003) Systemic movement of a tobamovirus requires host cell pectin methylesterase. *Plant J* **35**(3), 386-92
27. Desvoyes, B., Faure-Rabasse, S., Chen, M. H., Park, J. W., Scholthof, H. B. (2002) A novel plant homeodomain protein interacts in a functionally relevant manner with a virus movement protein. *Plant Physiol* **129**(4), 1521-32
28. Huang, Z., Andrianov, V. M., Han, Y., Howell, S. H. (2001) Identification of arabidopsis proteins that interact with the cauliflower mosaic virus (CaMV) movement protein. *Plant Mol Biol* **47**(5), 663-75
29. Fridborg, I., Grainger, J., Page, A., Coleman, M., Findlay, K. et al. (2003) TIP, a novel host factor linking callose degradation with the cell-to-cell movement of potato virus X. *Mol Plant Microbe Interact* **16**(2), 132-40
30. Li, X. D., Makela, T. P., Guo, D., Soliymani, R., Koistinen, V. et al. (2002) Hantavirus nucleocapsid protein interacts with the Fas-mediated apoptosis enhancer Daxx. *J Gen Virol* **83**(Pt 4), 759-66
31. Kaukinen, P., Vaheri, A., Plyusnin, A. (2003) Non-covalent interaction between nucleocapsid protein of tula hantavirus and small ubiquitin-related modifier-1, SUMO-1. *Virus Res* **92**(1), 37-45
32. Maeda, A., Lee, B. H., Yoshimatsu, K., Saijo, M., Kurane, I. et al. (2003) The intracellular association of the nucleocapsid protein (NP) of hantaan virus (HTNV) with small ubiquitin-like modifier-1 (SUMO-1) conjugating enzyme 9 (Ubc9). *Virology* **305**(2), 288-97

Figure and Table Legends:

Table 1: Proteins identified from in vivo and in vitro complexes.

Non redundant proteins identified in the 137 gel bands (Figure 2B) and from the different *in vivo* and *in vitro* experiments. In the first part, proteins were classified according their similarity and number of identifications among the Azucena and IR64 cultivar *in vivo*, in second part proteins were classified according their differences and number of identifications between Azucena and IR64 cultivar *in vivo*. Columns annotated with 1, 2, 3 wpi corresponds to experiments realized with rice plants 1, 2 and 3 weeks post inoculation with the virus. Others columns refer to the experiments realized with different combinations of virus and host. Number inside dark rectangles indicates the number of peptides used to identify the protein and crosses indicate that a similar function was identified in SwissProt, TrEMBL or TrEMBLnew protein databases. Some of the multigenic families are represented by 2 or more paralogs which were identified by a discriminating peptide.

Table 2: Proteins identified from in vivo and in vitro complexes at different salt concentration.

Non redundant proteins identified with the IR64 cultivar from the different *in vivo* and *in vitro* experiments using a buffer extraction added with 0.1M, 0.5M and 1M NaCl. Proteins were classified according their identification in all experiments to proteins identified only with 1M NaCl *in vivo*.

Figure 1: Elution profiles of IR64 protein extraction solutions from in vivo and in vitro binding experiments.

10 ml of concentrated protein extract solution are injected with an automated injection syringe (ATKTA Prime system), and elution was carried out at 1.5ml/mn with buffer elution (TAPS 20 mM pH 7.6, 100 mM NaCl). Dead volume represents the fraction of membrane fragments and/or protein aggregates that are still present in solution. Collected fractions containing RYMV-host proteins complexes correspond to the first 2/3 part of peaks labeled c1.

(A) Size exclusion chromatography elution profiles of purified RYMV (red), of soluble proteins extracted from IR64 non infected leaves (Control experiment in green) and IR64 infected leaves (blue).

(B) Size exclusion chromatography of *in vitro* binding experiment with soluble proteins extracted from IR64 healthy plants added with purified RYMV.

(C) Negative contrast electron micrograph of collected fraction stained with uranyl acetate containing purified RYMV particle from fraction corresponding to the peak c1 eluted at 216 mn.

(D) RYMV particle associated with host materials in fraction corresponding to the peak c2 eluted at 206 mn. The bar represents 100 nm.

Figure 3: Gels of denatured RYMV-IR64 protein complexes.

(A) SDS-page gel electrophoresis of RYMV-IR64 protein complexes collected after size exclusion chromatography from *in vitro* IR64 infected leaves, and from *in vitro* binding experiments. Plants at 1, 2 and 3 weeks post infection (wpi) correspond to plants at 3, 4 and 5 weeks post seedling (wps).

(B) Same gel showing the numbering and localization of samples analyzed by LC-MS-MS. Each gel lane was cut and analyzed after tryptic digestion.

Figure 4: Distribution of proteins identified from RYMV-IR64 experiments in Table 1.

Central circle is divided in proteins detected only in *in vivo* experimentations, only in *in vitro* experimentations, and detected *in vivo* as well *in vitro* experimentations (common). For each group, functional distribution into Me (metabolism), D (defense-stress), Tr (transcription), T (translation/protein synthesis), Tp (transport), Si (signal transduction), U (unknown), To (transposon).

Figure 5: Effects of increased NaCl concentration during extraction-purification of complexes.

(A) SDS-page gel electrophoresis of RYMV-host protein complexes extracted and purified by size exclusion chromatography from 2 wpi *in vivo* IR64 infected leaves or from *in vitro* binding experiments at different salt concentrations. Extractions and exclusion size chromatography were realized with buffers added with 0.1 M NaCl, 0.5M NaCl then 1M NaCl.

(B) Histogram of non-redundant proteins identified by LC-MS-MS (see Table 2) in comparison with a control C (IR64 2 wpi 0.1M NaCl *in vivo*, and IR64 4 wps 0.1M NaCl *in vitro*). Ribosomal proteins and histon proteins (except ribosomal protein S15, putative 11686.m01022) were identified only at 0.5 M and 1M NaCl.

Figure 6: Common protein from different complexes.

(A) Histogram of cumulated number of non-redundant proteins identified by LC-MS-MS (see Table 1) for IR64 and Azucena cultivars infected by RYMV at 1, 2 and 3 wpi.

(B) Histogram of common proteins identified in FHV-IR64 *in vitro* interaction (see Table 1) and common protein functions identified in different interactions (*in vitro* with RYMV-N. tabacum and RYMV-P. vulgaris; *in vivo* with SCPMV and P. vulgaris)* (see Table 1).

* Identification was made using the TIGR rice pseudomolecule for experimentations with IR64 and Azucena cultivar, and with SwissProt, TrEMBL and TrEMBLnew for Phaseolus vulgaris and Nicotiana tabacum.

Table 1

Name of the protein	accession number	MW	IR64 *in vitro*			IR64 *in vivo*			Azucena *in vivo*			IR64 FHV
			1	2	3	1	2	3	1	2	3	
Metabolism (Me)												
4-alpha-glucanotransferase, putative	11673.m04620	105744	10	4	11	3			7	4		9
5-methyltetrahydropteroyltriglutamate--homocysteine S-methyltransferase	11686.m04282	84584	16	14	19	25	20	13	10	17	28	
5-methyltetrahydropteroyltriglutamate--homocysteine S-methyltransferase	11686.m04283	84666	16	14	20	26	19	14	10	15	33	
6,7-dimethyl-8-ribityllumazine synthase, putative	11670.m04040	22471			5							
6-phosphogluconate dehydrogenase, decarboxylating	11680.m00122	52721	2					3				
adenosine kinase	11668.m03967	36993	2					3				
adenosylhomocysteinase	11687.m02417	49299	6		7			3		4		
alpha-mannosidase	11687.m02967	114157	3							7	8	
aminotransferase, class V	11674.m03932	44397	9				4		6	2	2	
nucellin-like aspartic protease	11687.m00750	44732						3			2	
ATP synthase F1, alpha subunit	11681.m00716	55325	2									
ATP synthase F1, alpha subunit	11670.m01497	55665			8							13
ATP synthase F1, beta subunit	11667.m04784	59463	4		9							5
ATP synthase F1, beta subunit	11676.m01781	60260	10		4	5			3			9
beta-glucosidase	11669.m04987	56836	10			15	18	22		9	7	
Carbonic anhydrase, putative	11667.m04353	29117	14			10			12	12	4	
citrate synthase I	11668.m01280	56640			4							4
citrate (Si)-synthase, eukaryotic	11668.m00931	52229				3				5		
CoA-ligase, putative	11667.m01918	67855	2	13	15	19	11	14	5	13	21	2
3-hydroxyacyl-CoA dehydrogenase, C-terminal domain, putative	11668.m01643	78460	4		8	7	9	14		11	7	
3-hydroxyacyl-CoA dehydrogenase, NAD binding domain, putative	11667.m02406	79392			2	4						
2-oxoglutarate dehydrogenase, E2 component, dihydrolipoamide succinyltransferase	11670.m03088	48283			6	3			3	4		5
dihydrolipoamide S-acetyltransferase	11674.m03308	52541		2		2				4		2
D-isomer specific 2-hydroxyacid dehydrogenase, NAD binding domain, putative	11668.m00017	44786					18	2				
ferredoxin-dependent glutamate synthase (fragments)	11673.m04580	176883	7	18	42	49	47	53	33	54	74	
ferredoxin--nadp reductase, leaf isozyme, chloroplast precursor(ec 1.18.1.2) (fnr)	11680.m00092	39982	3		4		7	3		2	5	
ferredoxin--nitrite reductase	11668.m05166	72403	6								3	
Similar to ferredoxin-nitrite reductase (EC 1.7.7.1) precursor - rice	11667.m02487	55745	8									
flavodoxin, putative	11667.m05707	21705	3									
FMN-dependent dehydrogenase	11673.m00499	40219	10						2			
fructose-1,6-bisphosphatase	11667.m06481	37012	3						2			
fructose-1,6-bisphosphatase, putative	11670.m01491	42245	8					6			9	
Fructose-bisphosphate aldolase class-I	11667.m06837	38799	10	13	13	17	15	19	14	15	18	
Fructose-bisphosphate aldolase class-I	11674.m00179	38370			4	4	5	5	5	5	6	
Fructose-bisphosphate aldolase class-I	11680.m03968	37731	4	11	10	17	14	14	15	13	15	
Fructose-bisphosphate aldolase class-I	11682.m03129	38839	8	12	10	16	16	19	15	10	16	
Fructose-bisphosphate aldolase class-I	11687.m00638	41980	15	6	12	5	9	15	4	3	20	7
Fructose-bisphosphate aldolase class-I, putative	11667.m00198	178342						2			2	
GDSL-like Lipase/Acylhydrolase	11682.m01126	38959					9	7			10	
glucose-6-phosphate isomerase, putative	11674.m03720	67229	2									
similar to glucosidase II alpha subunit	11669.m01126	72311			4							
glutamate dehydrogenase	11669.m05875	44341	2								2	
glutamate dehydrogenase 2 (ec 1.4.1.3) (gdh 2). [mouse-ear cress	11670.m04472	45594	2									
glutamate-1-semialdehyde-2,1-aminomutase	11674.m04228	50237	6									
Glutamine synthetase, catalytic domain	11668.m04892	39176			3							
Glutamine synthetase, catalytic domain, putative	11670.m05561	49428	6	4			2				2	4
glyceraldehyde-3-phosphate dehydrogenase, C terminal domain, putative	11670.m03659	42689	11	9	14	9	6		8	9	4	9
glyceraldehyde-3-phosphate dehydrogenase, type I	11670.m03922	36673	7	6	7	5	5		9	6	4	
glyceraldehyde-3-phosphate dehydrogenase, type I	11674.m00238	36390	11	7	7	9	9		10	8	11	
glyceraldehyde-3-phosphate dehydrogenase, type I, putative	11669.m00306	47081	6	4		5				3		7
glycine cleavage system T protein	11670.m05219	43968	8						3			
glycine dehydrogenase	11667.m05023	111356	16	4	6				8	11	8	
glycine dehydrogenase	11680.m04002	111401	20	4	6				6	10	8	
putative glycine hydroxymethyltransferase	11669.m05342	56380			3							
glycolate oxidase	11669.m05782	40384	9						2			
Glycosyl hydrolases family 17	11667.m07188	35682	3						9			
isocitrate dehydrogenase, NADP-dependent	11667.m04493	46013	10									
lactoylglutathione lyase, putative	11674.m00866	32553						4				
legumin-like protein	11682.m00174	38195						4				
malate dehydrogenase, NAD-dependent	11682.m04796	35414	2					3			4	
cytoplasmic malate dehydrogenase	11676.m02982	35546	2					6			9	
putative glyoxysomal malate dehydrogenase	11669.m05674	36699	6						4	4		
Malic enzyme, N-terminal domain, putative	11667.m05146	71456	4		9			9		10	12	
Manganese-stabilizing protein / photosystem II polypeptide	11667.m02976	34861			2							
NAD binding domain of 6-phosphogluconate dehydrogenase, putative	11668.m03371	30496	3									
nucleoside diphosphate kinase i (ec 2.7.4.6) (ndk i) (ndp kinase i)(ndpk i)	11673.m02949	16851	2						3			

Figure 1

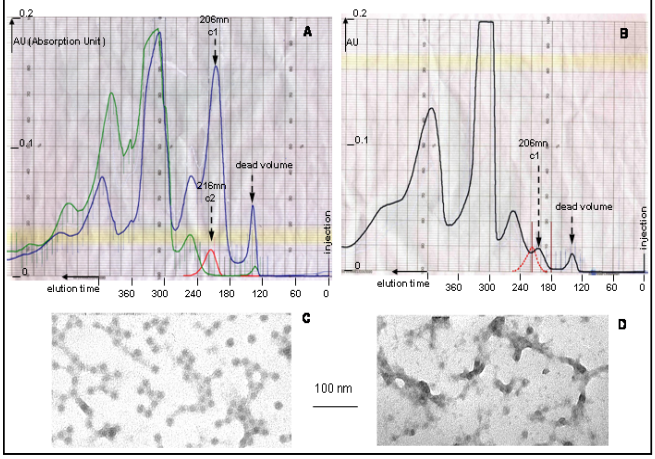


Figure 2

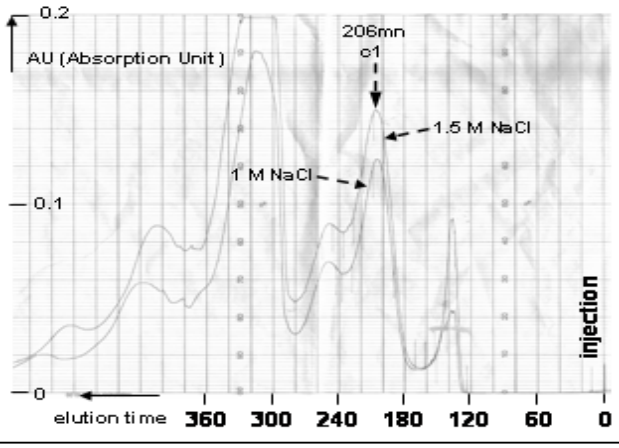


Figure 3

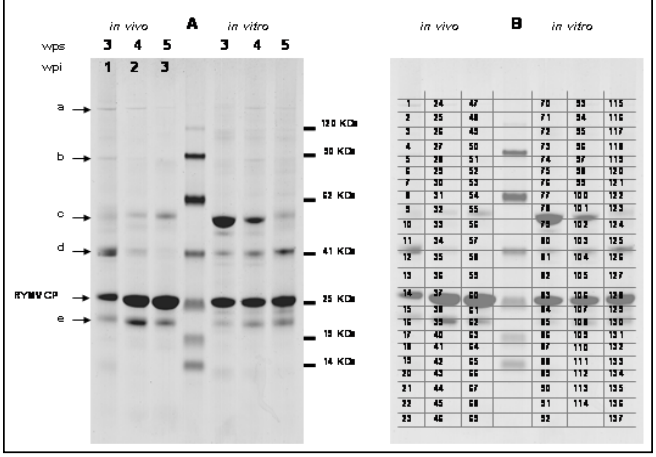


Figure 4

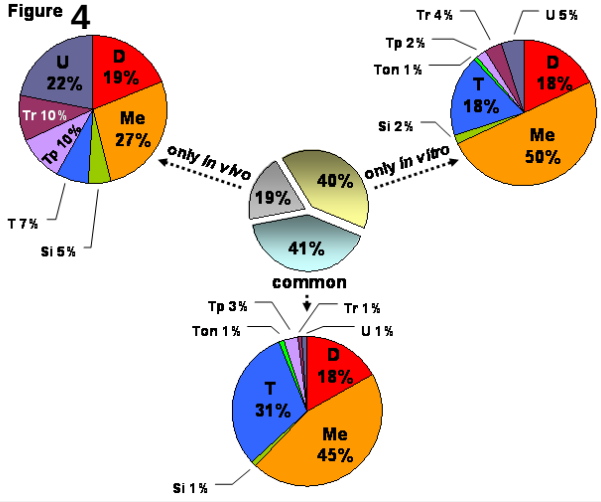


Figure 5

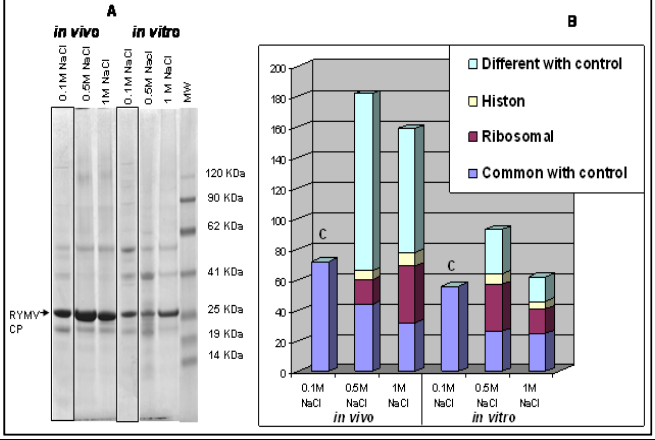
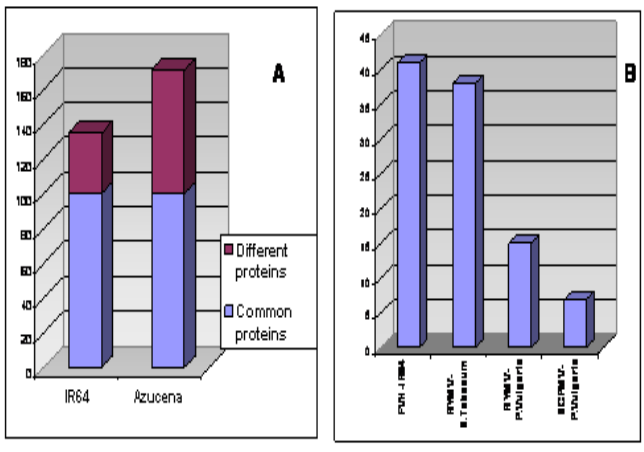


Figure 6



Bibliographie

Abo M. E., Sy A. A. and Alegbejo M. D.

Rice yellow mottle virus (RYMV) in Africa: Evolution, distribution, economic significance on sustainable rice production and management strategies. *Journal of sustainable Agriculture,* **1998**, 11, 85-114.

Agrios G. N.

Plant pathology, *4th edn. Academic Press*, London, **1997**.

Arumuganathan K and Earle E. D.

Nuclear DNA content of some important plant species. *Plant Mol Biol Rep,* **1991**, 9, 208-219.

Awoderu V. A.

Rice yellow mottle virus in West Africa. *Tropical Pest Management,* **1991**, 37, 356-362.

Bakker W.

Rice yellow mottle, a mechanically transmissible virus disease of rice in Kenya. *Netherland J. Plant Pathology,* **1970**, 77, 201-206.

Beato M., Chavez S. and Truss M.

Transcriptional regulation by steroid hormones. *Steroids*, **1996**, 61, 240-251.

Beato M. and Klug J.

Steroid hormone receptors: an update. *Hum Reprod Update 6*, **2000**, 225-236.

Bradshaw, R. A.

Revised draft guidelines for proteomic data publication. *Mol. Cell. Proteomics*, **2005**, 4, 1223-1225.

Brilla C.

Aldosterone and myocardial fibrosis in heart failure. *Herz*, **2000**, 25, 299-306.

Brugidou C., Opalka N., Yeager M., Beachy R. N. and Fauquet C.

Stability of rice yellow mottle virus and cellular compartmentalization during the infection process in Oryza sativa (L.). *Virology*, **2002**, 297, 98-108.

Bureik M., Schiffler B., Hiraoka Y., Vogel F. and Bernhardt R.

Functional expression of human mitochondrial CYP11B2 in fission yeast and identification of a new internal electron transfer protein, etp1. *Biochemistry*, **2002**, 41, 2311-2321.

Campostrini N., Areces L. B., Rappsilber J., Pietrogrande M.C., Dondi F., Pastorino F., Ponzoni M. and Righetti P. G.

Spot overlapping in two-dimensional maps : A serious problem ignored for much too long. *Proteomics*, **2005**, 5, 2385-2395.

Carr S., Aebersold R., Baldwin M., Burlingame A., Clauser K. and Nesvizhskii A.

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*, **2004**, 3, 531-533.

Chappell T. and Warren G.

A galactosyltransferase from the fission yeast *Schizosaccharomyces pombe*. *J. Cell Biol*., **1989**, 109, 2693-2702.

Davis L. and Smith G. R.

Meiotic recombination and chromosome segregation in Schizosaccharomyces pombe. *Proc. Natl. Acad. Sci.,* **2001**, 98, 8395-8402.

Evans T. and Hedger J. N.

Degradation of plant cell wall polymers, *in Gadd JM (ed) Fungi in bioremediation. Cambridge University Press, Cambridge*, **2001**, 1-26.

Fantes P. and Beggs J.

The Yeast Nucleus. *Oxford Univ. Press, Oxford*, **2000**.

Forsburg S. L. and Rhind N.

Basic methods for fission yeast. *Yeast*, **2006**, 23, 173-183.

Goff S. A., Ricke D., Lan T. H., Presting G., Wang R., Dunn M., Glazebrook J., Sessions A., Oeller P., Varma H., Hadley D., Hutchison D., Martin C., Katagiri F., Lange B. M., Moughamer T., Xia Y., Budworth P., Zhong J., Miguel T., Paszkowski U., Zhang S., Colbert M., Sun W. L., Chen L., Cooper B., Park S., Wood T. C., Mao L., Quail P., Wing R., Dean R., Yu Y., Zharkikh A., Shen R., Sahasrabudhe S., Thomas A., Cannings R., Gutin A., Pruss D., Reid J., Tavtigian S., Mitchell J., Eldredge G., Scholl T., Miller R. M., Bhatnagar S., Adey N., Rubano T., Tusneem N., Robinson R., Feldhaus J., Macalma T., Oliphant A. and Briggs S.

A draft sequence of the rice genome (*Oryza sativa L. ssp. Japonica*). *Science*, **2002**, 296, 92-100.

Goffeau A., Barrell B. G., Bussey H., Davis R. W., Dujon B., Feldmann H., Galibert F., Doheisel J. D., Jacq C., Johnston M., Louis E. J., Mewes H. W., Murakami Y., Philippsen P., Tettelin H. and Oliver S. G.

Life with 6000 genes. *Science*, **1996**, 274 (5287), 546-567.

Gurer C., Ciarelli A. and Luban J.

Specific incorporation of heat shock protein 70 family members into primate lentiviral virions. *J Virol*, **2002**, 76(9), 4666-4670.

Gygi S. P., Rist B., Gerber S. A., Turecek F., Gelb M. H. and Aebersold R.

Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol*, **1999**, 17, 994-999.

Hatsch D.

Interaction hôte/pathogène : étude du modèle *Humulus lupulus / Fusarium graminearum*. identification, génomique et transcriptomique du pathogène. *Thèse de l'Université Louis Pasteur*, **2004**.

Hibino H.

Biology and epidemiology of rice viruses. *Annu. Rev. Phytopathol.*, **1996**, 34, 249-274.

Idiris A., Bi K., Tohda H., Kumagai H. and Giga-Hama Y.

Construction of a protease-deficient strain set for the fission yeast Schizosaccharomyces pombe, useful for effective production of protease-sensitive heterologous proteins. *Yeast*, **2006**, 23, 83-99.

Kurtzman C. P. and Blanz P. A.

The yeasts: a taxonomic study, *In: Kurtzman, C. P., Fell, J. W. (Eds.), 4th edn. Elsevier, Amsterdam*, **2000**, 69-74.

Leonard S., Plante D., Wittmann S., Daigneault N., Fortin M. G. and Laliberte J. F.

Complex formation between potyvirus VPg and translation eukaryotic initiation factor 4E correlates with virus infectivity. *J Virol*, **2000**, 74(17), 7730-7737.

Lepoivre P.

Phytopathologie : Bases moléculaires et biologiques des pathosystèmes et fondements des stratégies de lutte. *De boeck Université, Les Presses Agronomiques de Gembloux*, **2003**.

Lill J.

Proteomic tools for quantitation by mass spectrometry. *Mass Spectrom Rev*, **2003**, 22, 182-194.

Losel R. M., Feuring M., Falkenstein E. and Wehling M.

Nongenomic effects of aldosterone: cellular aspects and clinical implications. *Steroids*, **2002**, 67, 493-498.

Losel R. and Wehling M.

Nongenomic actions of steroid hormones. *Nat Rev Mol Cell Biol*, **2003**, 4, 46-56.

Mitchell D. B., Weimann K., Vogel B. J., Pasamontes L. and van Loon A. P. G. M.

The phytase subfamily of histidine acid phosphatases: isolation of genes for two novel phytases from the fungi Aspergillus terreus and Myceliophthora thermophila. *Microbiology*, **1997**, 143, 245-252.

Ong S. E., Blagoev B., Kratchmarova I., Kristensen D. B., Steen H., Pandey A. and Mann M.

Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics*, **2002**, 1, 376-386.

Opalka N., Tihova M., Brugidou C., Kumar A., Beachy R. N., Fauquet C. M. and Yeager M.

Structure of native and expanded sobemoviruses by electron cryo-microscopy and image reconstruction. *J Mol Biol*, **2000**, 303, 197-211.

Park S. G. and Jung G.

Human hepatitis B virus polymerase interacts with the molecular chaperonin Hsp60. *J Virol*, **2001**, 75, 6962-6968.

Remacle J. E., Albrecht G., Brys R., Braus G. H. and Huylebroeck D.

Three classes of mammalian transcription activation domain stimulate transcription in *Schizosaccharomyces pombe*. *EMBO J*, **1997**, 16, 5722-5729.

Siré C. and Brugidou C.

Les virus du riz : état des lieux, impact et méthodes de lutte. *Virologie*, **2002**, 6, 431-443.

Sposato P., Ahn J.H. and Walton J.D.

Characterization and disruption of a gene in the maize pathogen *Cochliobolus carbonum* encoding a cellulase lacking a cellulose binding domain and hinge region. *Mol Plant Microbe*, **1995**, 8, 602-609.

Sylvia V. L., Schwartz Z., Schuman L., Morgan R. T., Mackey S., Gomez R. and Boyan B. D.

Maturation-dependent regulation of protein kinase C activity by vitamin D3 metabolites in chondrocyte cultures. *J Cell Physiol*, **1993**, 157, 271-278.

Sun N., Jang J., Lee S., Kim S., Lee S., Hoe K. L., Chung K. S., Kim D. U., Yoo H. S., Won M. and Song K. B.

The first two-dimensional reference map of the fission yeast, *Schizosaccharomyces pombe* proteins. *Proteomics*, **2005**, 5, 1574-1579.

Wanjiru W.M., Zhensheng K. and Buchenauer H.

Importance of cell wall degrading enzymes produced by Fusarium graminearum during infection of wheat heads. *Eur J Plant Pathol*, **2002**, 108, 803-810.

Wehling, M.

Specific, nongenomic actions of steroid hormones. *Annual Review of Physiology*, **1997**, 59, 365-393.

Wehling M., Bauer M. M., Ulsenheimer A., Schneider M., Neylon C. B. and Christ M.

Nongenomic effects of aldosterone on intracellular pH in vascular smooth muscle cells. *Biochem Biophys Res Commun*, **1996**, 223, 181-186.

Wehling M., Spes C. H., Win N., Janson C. P., Schmidt B. M. W., Theisen K. and Christ M.

Rapid cardiovascular action of aldosterone in man. *J Clin Endocrinol Metab*, **1998**, 83, 3517-3522.

Wilkins M. R., Appel R. D., Van Eyk J. E., Chung M. C., Gorg A., Hecker M., Huber L. A., Langen H., Link A. J., Paik Y. K., Patterson S. D., Pennington S. R., Rabilloud T., Simpson R. J., Weiss W. and Dunn M. J.

Guidelines for the next 10 years of proteomics. *Proteomics*, **2006**, 6, 4-8.

Wood V., Gwilliam R., Rajandream M. A., Lyne M., Lyne R., Stewart A., Sgouros J., Peat N., Hayles J., Baker S., Basham D., Bowman S., Brooks K., Brown D., Brown S., Chillingworth T., Churcher C., Collins M., Connor

R., Cronin A., Davis P., Feltwell T., Fraser A., Gentles S., Goble A., Hamlin N., Harris D., Hidalgo J., Hodgson G., Holroyd S., Hornsby T., Howarth S., Huckle E. J., Hunt S., Jagels K., James K., Jones L., Jones M., Leather S., Mcdonald S., Mclean J., Mooney P., Moule S., Mungall K., Murphy L., Niblett D., Odell C., Oliver K., O'neil S., Pearson D., Quail M. A., Rabbinowitsch E., Rutherford K., Rutter S., Saunders D., Seeger K., Sharp S., Skelton J., Simmonds M., Squares R., Squares S., Stevens K., Taylor K., Taylor R. G., Tivey A., Walsh S., Warren T., Whitehead S., Woodward J., Volckaert G., Aert R., Robben J., Grymonprez B., Weltjens I., Vanstreels E., Rieger M., Schafer M., Muller-Auer S., Gabel C., Fuchs M., Düsterhöft A., Fritzc C., Holzer E., Moestl D., Hilbert H., Borzym K., Langer I., Beck A., Lehrach H., Reinhardt R., Pohl T. M., Eger P., Zimmermann W., Wedler H., Wambutt R., Purnelle B., Goffeau A., Cadieu E., Dreano S., Gloux S., Lelaure V., Mottier S., Galibert F., Aves S. J., Xiang Z., Hunt C., Moore K., Hurst S. M., Lucas M., Rochet M., Gaillardin C., Tallada V. A., Garzon A., Thode G., Daga R. R., Cruzado L., Jimenez J., Sanchez M., Del Rey F., Benito J., Dominguez A., Revuelta J. L., Moreno S., Armstrong J., Forsburg S. L., Cerutti L., Lowe T., Mccombie W. R., Paulsen I., Potashkin J., Shpakovski G. V., Ussery D., Barrell B. G. and Nurse P.

The genome sequence of *Schizosaccharomyces pombe*. *Nature*, **2002**, 415, 871-880.

Yu J., Hu S., Wang J., Wong G. K., Li S., Liu B., Deng Y., Dai L., Zhou Y., Zhang X., Cao M., Liu J., Sun J., Tang J., Chen Y., Huang X., Lin W., Ye C., Tong W., Cong L., Geng J., Han Y., Li L., Li W., Hu G., Huang X., Li W., Li J., Liu Z., Li L., Liu J., Qi Q., Liu J., Li L., Li T., Wang X., Lu H., Wu T., Zhu M., Ni P., Han H., Dong W., Ren X., Feng X., Cui P., Li X., Wang H., Xu X., Zhai W., Xu Z., Zhang J., He S., Zhang J., Xu J., Zhang K., Zheng X., Dong J., Zeng W., Tao L., Ye J., Tan J., Ren X., Chen X., He J., Liu D., Tian W., Tian C., Xia H., Bao Q., Li G., Gao H., Cao T., Wang J., Zhao W., Li P., Chen W., Wang X., Zhang Y., Hu J., Wang J., Liu S., Yang J., Zhang G., Xiong Y., Li Z., Mao L., Zhou C., Zhu Z., Chen R., Hao B., Zheng W., Chen S., Guo W., Li G., Liu S., Tao M., Wang J., Zhu L., Yuan L. and Yang H.

A draft sequence of the rice genome (*Oryza sativa L. ssp. indica*). *Science*, **2002**, 296, 79-92.

Yuan Q., Ouyang S., Wang A., Zhu W., Maiti R., Lin H., Hamilton J., Haas B., Sultana R., Cheung F., Wortman J. and Buell C. R.

The Institute for Genomic Research Osa1 Rice Genome Annotation Database. *Plant Physiology*, **2005**, 138, 18-26.

Yuan Q., Quackenbush J., Sultana R., Pertea M., Salzberg S. L. and Buell A. R.

Rice Bioinformatics. Analysis of Rice Sequence Data and Leveraging the Data to Other Plant Species. *Plant Physiology*, **2001**, 125, 1166-1174.

RESULTATS

TROISIEME PARTIE :
Applications portant sur des organismes non séquencés : le séquençage *de novo*

Chapitre I : Identification des protéines par séquençage *de novo*

Chapitre II : Identification des protéines impliquées dans la résistance à l'arsenic chez la bactérie *Caenibacter arsenoxydans*

Chapitre III : Analyse protéomique de tissus de vigne (*Vitis vinifera L.*) ayant subi un stress herbicide

CHAPITRE I

L'identification des protéines par séquençage *de novo*

1 Pourquoi le recours au séquençage *de novo* ?

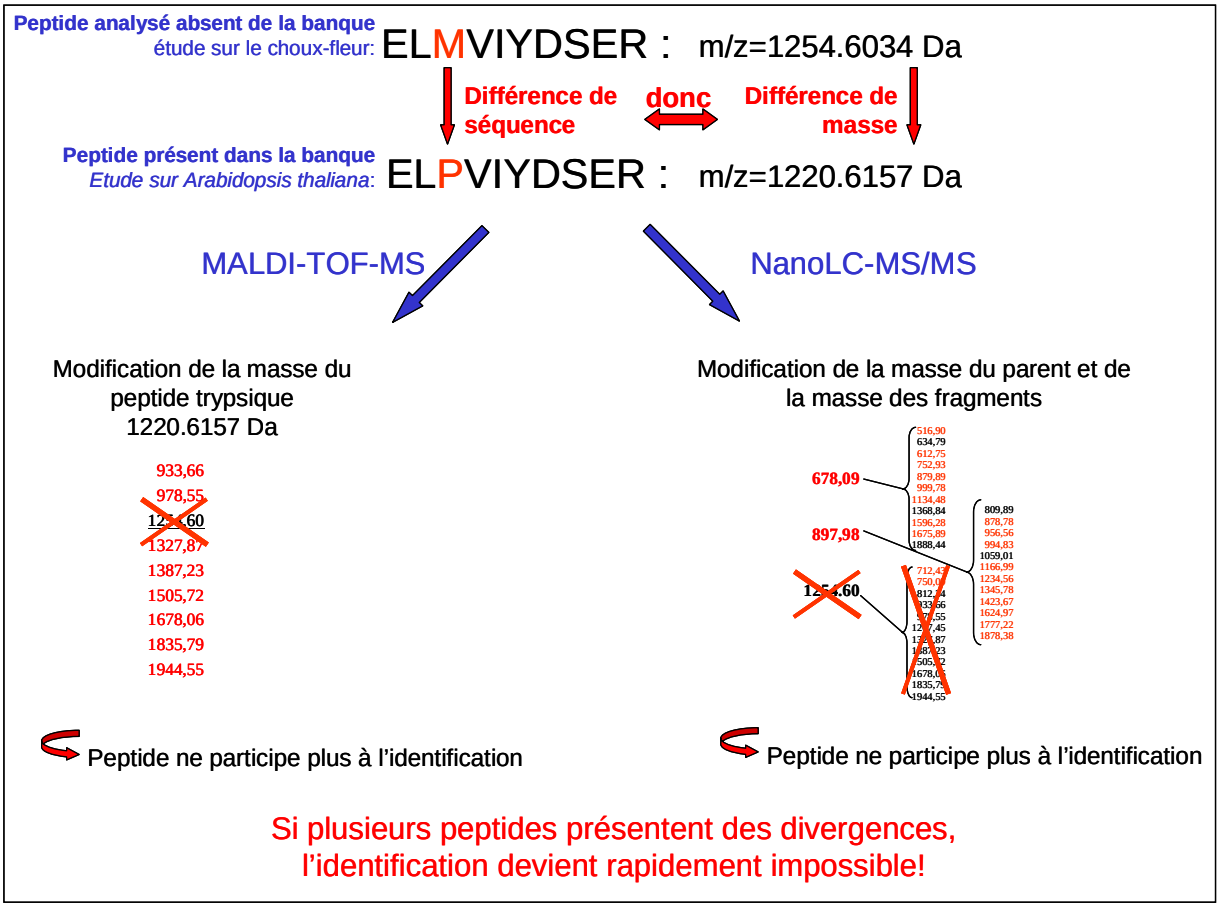
1.1 Limites des moteurs de recherche classique pour l'étude des organismes non séquencés

Même si leurs algorithmes sont différents, la plupart des programmes conventionnels d'identification de protéines comme Mascot [Perkins et coll., 1999] ou SEQUEST [Eng et coll., 1994] se basent sur des identités strictes de masses entre les peptides analysés et les peptides théoriques issus de la digestion *in silico* de l'ensemble des protéines présentes dans les banques de données. Dans ces approches « non tolérantes aux erreurs », l'exigence d'identité parfaite entre les peptides analysés et les peptides théoriques présents dans les banques présente plusieurs avantages :

- Elle augmente significativement la spécificité et la rapidité des recherches en banque [Clauser et coll., 1999].
- Elle permet d'être moins exigeant par rapport à la qualité des spectres MS/MS car les identifications sont possibles même quand les spectres MS/MS ne sont pas de qualité exceptionnelle.
- Elle permet d'identifier des protéines avec confiance avec un petit nombre de peptides fragmentés seulement.

Néanmoins, comme illustré en figure 1, la moindre différence de séquence et donc de masse entre les peptides analysés et ceux présents dans les banques est critique et conduit rapidement à l'impossibilité d'identifier la protéine. Cet aspect constitue un véritable goulot d'étranglement pour les études protéomiques portant sur des organismes non séquencés et/ou non représentés dans les banques de données protéiques accessibles et tendrait à restreindre le champ d'application des études protéomiques à une poignée d'organismes modèles entièrement séquencés et annotés.

Figure 1 : Illustration de la limite des moteurs de recherche conventionnels basés sur des identités de masses strictes pour l'identification de protéines absentes des banques de données.

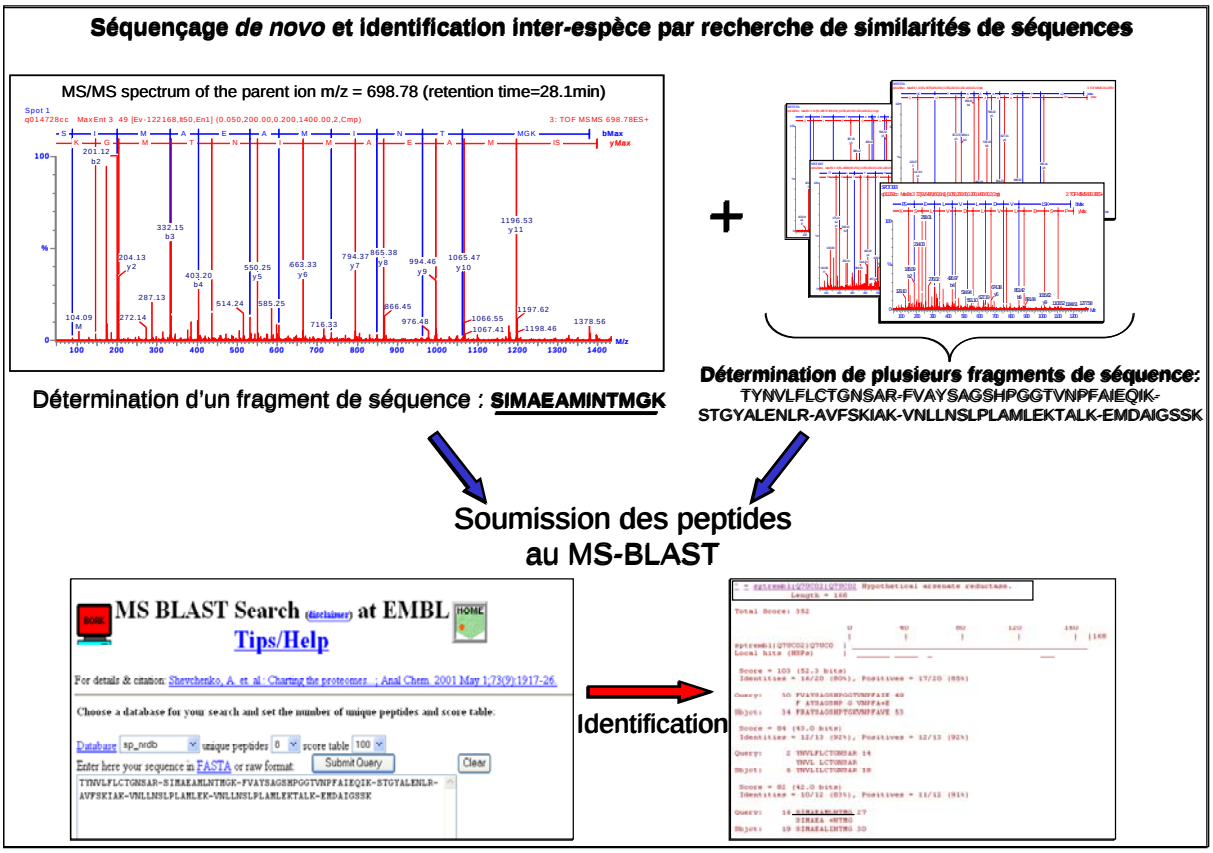


Afin d'élargir le panel d'organismes pour lesquels les protéines peuvent être identifiées par des approches protéomiques, des méthodes alternatives avec des critères de recherche et une interprétation différente des données de spectrométrie de masse en tandem ont vu le jour [Shevchenko et coll., 2002 ; Liska et coll., 2003] : les approches basées sur le séquençage *de novo* des peptides.

1.2 Le séquençage *de novo* et les identifications inter-espèces par recherche de similarités de séquences

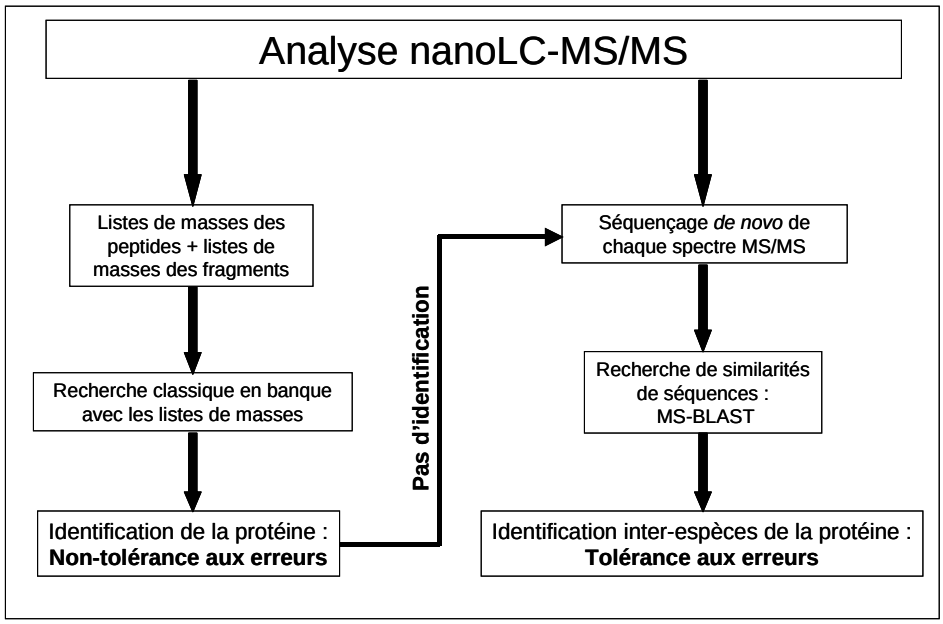
L'approche par séquençage *de novo* consiste en l'interprétation des spectres de masse en tandem afin de déduire directement de ces spectres des fragments de séquence en acides aminés, voire la séquence complète lorsque le spectre MS/MS le permet : l'étude des règles de fragmentation des peptides ont rendu cette interprétation possible (voir fragmentation des peptides Chapitre II.1.3 de l'introduction bibliographique). Les fragments ou « tags » de séquence d'acides aminés ainsi déterminés sont ensuite soumis à des programmes de recherche de similarités de séquences du type BLAST (Basic Local Alignement Search Tool) [Altschul et coll., 1997] ou FASTA [Pearson et coll., 1988] (Figure 2).

Figure 2 : Représentation de la stratégie d'identification des protéines par séquençage *de novo* suivie par des requêtes MS-BLAST.



Cette approche, contrairement aux moteurs de recherches protéomiques conventionnels, est tolérante aux erreurs de séquence : elle permet d'identifier des protéines non présentes dans les banques de données par homologie avec des protéines d'organismes phylogénétiquement proches qui sont présentes dans les banques (Figure 3). Les possibilités de travailler et d'identifier des protéines d'organismes « exotiques » dont le génome n'a pas encore été séquencé sont donc beaucoup plus vastes qu'avec les programmes classiquement utilisés.

Figure 3 : Illustration des deux voies d'interprétation des données de nanoLC-MS/MS : la première voie, classique, du type Mascot, SEQUEST, … qui nécessite des identités strictes de séquence et la seconde voie, par séquençage *de novo* et MS-BLAST, tolérante aux erreurs.



1.3 Les difficultés et contraintes de cette approche

Même si cette approche ouvre de nouvelles possibilités, elle présente aussi une série d'exigences et de contraintes :

- Les **spectres MS/MS doivent être de très bonne qualité** pour pouvoir en déduire un tag de séquence suffisamment long (4-6 acides aminés consécutifs). Cette exigence de qualité est beaucoup plus importante que pour des recherches classiques par comparaison de listes de masses. En effet, si les séries des ions fragments permettant de lire les différences de masses correspondant à des acides aminés sont interrompues, l'interprétation devient rapidement impossible par séquençage *de novo*.
- L'**interprétation** des spectres est **longue** et **fastidieuse** et la **vérification manuelle** de l'ensemble des spectres, voire l'interprétation manuelle complète des spectres est souvent nécessaire malgré le développement de nouveaux algorithmes d'aide au séquençage *de novo* de plus en plus performants.
- Les **algorithmes classiques** de recherches de similarités de séquence **du type BLAST**, optimisés pour des séquences longues et exactes, **ne sont pas adaptés pour les séquences courtes générées par spectrométrie de masse en tandem.** Des outils plus adaptés ont donc été développés (voir en 2.3).

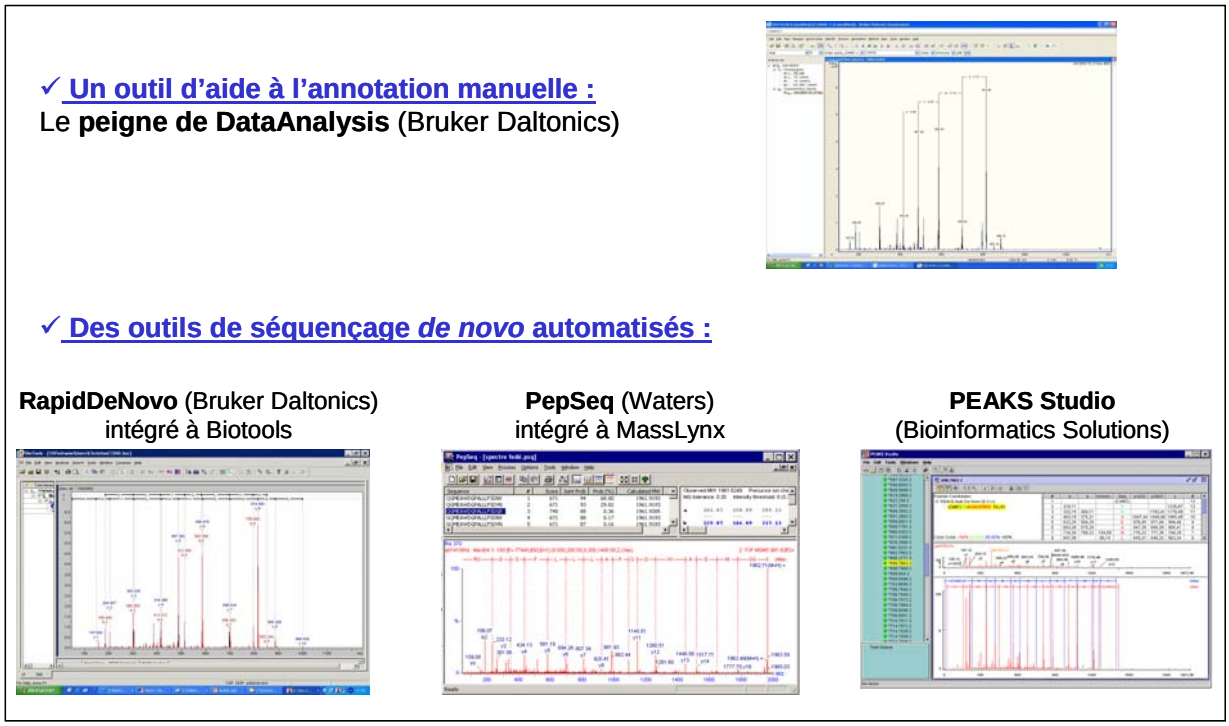
2 Les outils informatiques

2.1 Les logiciels d'aide à l'interprétation des spectres par séquençage *de novo*

L'interprétation manuelle des spectres, tableau des masses, règle et calculatrice en main est rapidement devenue un facteur limitant car très longue et fastidieuse. La course à la mise au point d'algorithmes d'aide au séquençage *de novo* a donc démarré et un grand nombre de logiciels et d'algorithmes sont actuellement disponibles. Ainsi, des algorithmes variés sont proposés présentant des performances en qualité et en temps d'interprétation variables.

Plusieurs de ces logiciels sont disponibles au laboratoire, des outils d'aide à l'annotation manuelle à des logiciels semi-automatisés voire entièrement automatisés. Les outils disponibles au laboratoire sont présentés en figure 4.

Figure 4 : Les logiciels d'aide à l'interprétation des spectres par séquençage *de novo* disponibles au laboratoire.



2.2 Evaluation de ces outils

Les outils disponibles reposent sur des algorithmes différents, chacun présentant ses forces et ses faiblesses. Il est important d'être conscient des limites de chacun des algorithmes et de ne pas négliger la vérification manuelle des séquences.

2.2.1 L'outil peigne de DataAnalysis

L'outil peigne de DataAnalysis est, même s'il n'est pas automatisé, très utile puisqu'il permet d'annoter manuellement mais rapidement les spectres. Les différences de masse correspondant aux différents acides aminés sont pré-enregistrées dans le programme et il suffit de se déplacer sur le spectre pour annoter les écarts entre les ions. Cet outil est à la fois très utile quotidiennement pour la vérification rapide du bon état de fonctionnement de l'appareil mais aussi pour la vérification précise et ponctuelle de variations de séquences ou de modifications post-traductionnelles dont on soupçonne la présence. En effet, une vérification quotidienne de la séquence d'un peptide d'autolyse de la trypsine dont on connaît bien la séquence permet de faire un rapide contrôle du bon fonctionnement de la fragmentation.

2.2.2 Les outils de traitement spectres à spectres, traitement semi-automatisé : RapidDeNovo (Bruker Daltonics), PepSeq (Waters)

Pour chaque spectre MS/MS ces outils vont faire travailler un algorithme qui va proposer une ou plusieurs séquences candidates, avec des pourcentages de confiance.

Les algorithmes qu'utilisent ces programmes sont optimisés pour des spectres spécifiquement générés par un type de machine. Même si globalement les mécanismes de fragmentation sont les mêmes pour des fragmentations CID basse énergie, les fonctionnements des analyseurs conduisent à de légères différences dans les profils de fragmentation selon le type d'instrument utilisé. Par exemple, le principe de

fonctionnement de l'analyseur trappe empêchera la visualisation des ions immoniums dans les basses qui peuvent être un indice supplémentaire de la présence des acides aminés dans des spectres de Q-TOF.

Ainsi, le programme PepSeq est très performant pour des spectres de fragmentation générés par des instruments de type Q-TOF. De la même manière, l'algorithme RapidDeNovo a été optimisé pour des spectres générés par des trappes ioniques en mode CID ou ETD, pour des spectres PSD (Post Source Decay), ISD (In Source Decay) ou encore des spectres de fragmentation haute énergie générés par des instruments de type MALDI-TOF/TOF.

Pour ces différents algorithmes, le succès de la détermination de la séquence correcte dépend bien sûr de la qualité des données spectrales en elles-mêmes mais il dépend aussi fortement du paramétrage des différentes options. Ces algorithmes font appel à de nombreux paramétrages (tolérances d'erreurs de masse sur l'ion parent et les ions fragments, seuils d'intensité, stringence des calculs, …) et c'est à l'utilisateur d'acquérir l'expertise nécessaire pour affiner ces paramétrages selon la qualité du spectre à traiter.

En conclusion, ces algorithmes donnent souvent de bons résultats mais ils nécessitent un lourd investissement en temps de réglage des paramètres et de calcul lorsque l'on a des analyses nanoLC-MS/MS contenant 100 à 300 spectres MS/MS à traiter.

2.2.3 Le logiciel PEAKS Studio (Bioinformatics Solutions)

Ayant découvert ce nouveau logiciel sur Internet dès sa commercialisation en juillet 2003, la possibilité de téléchargement gratuit d'une version de démonstration pendant une période de 30 jours nous a permis de le tester puis d'en faire l'acquisition au vu de la qualité des résultats obtenus.

L'algorithme original de PEAKS est conceptuellement différent de la majorité des autres algorithmes de séquençage *de novo* existant. Globalement, l'ensemble des algorithmes repose sur un modèle graphique des spectres dans lequel les pics sur les spectres correspondraient à des nœuds dans un graphique. Et l'algorithme tente alors de trouver un chemin permettant de connecter tous les nœuds entre eux afin de reconstruire la succession complète des acides aminés du Nter vers le Cter. La principale limite de ces algorithmes est que lorsqu'il y a une interruption dans la chaîne, due à l'absence d'un ion y- (ou b-), la poursuite de la détermination de la séquence devient difficile.

L'algorithme de PEAKS repose sur un modèle mathématique qui est décrit par Ma et coll. [Ma et coll., 2003] et fonctionne en 4 étapes :

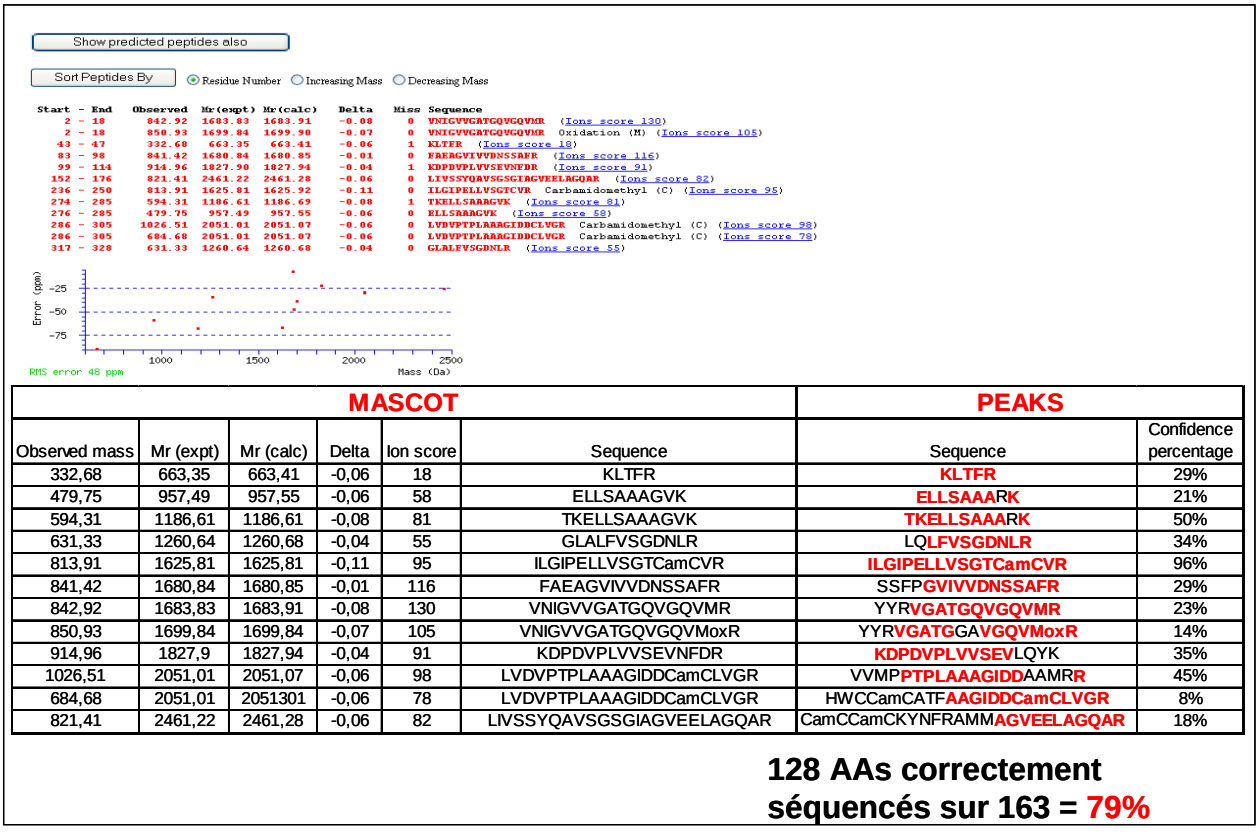
- Pré-traitement des données MS/MS brutes incluant le filtrage du bruit, le centrage des pics et la deconvolution des ions multi-chargés.
- Calcul d'un grand nombre de séquences candidates possibles
- Affinage du score et sélection des meilleurs candidats
- Calcul d'un score de confiance global pour la séquence du peptide et d'un score de confiance individuel pour chacun des acides aminés.

Le principal avantage de cet algorithme est que l'absence d'un pic dans la succession d'une série y- (ou b-) ne pose pas de problème majeur : PEAKS pourra trouver deux tags de séquence dans un même peptide avec un niveau très élevé de confiance entrecoupés par un tag de faible confiance. Son second avantage est que le traitement par PEAKS est beaucoup plus rapide qu'avec les autres logiciels : les données nanoLC-MS/MS brutes sont soumises au logiciel, sans pré-traitement nécessaire, et le temps d'interprétation d'un spectre MS/MS est d'environ 5 secondes.

Cependant, un point faible peut être relevé quant au fonctionnement de cet algorithme, point faible qui est aussi réel pour les algorithmes de PepSeq et de RapidDeNovo : l'ensemble de ces algorithmes se base toujours sur la masse globale du peptide et vont toujours proposer une séquence complète du peptide allant jusqu'à la masse du parent. Ce point est quelquefois limitant car souvent le spectre de fragmentation est très beau sur une partie seulement de la gamme de masse alors que l'algorithme va toujours jusqu'au bout de la séquence au détriment parfois de la bonne détermination de la belle région du spectre uniquement.

Nous avons évalué les performances du logiciel PEAKS en analysant ses résultats sur des séquences peptidiques connues et en comparant ses résultats avec ceux obtenus avec PepSeq ou encore RapidDeNovo. La figure 5 présente une évaluation des résultats obtenus par PEAKS sur des données de nanoLC-MS/MS d'un spot de gel 2D de souris qui ont été, dans un premier temps, soumises à une requête Mascot classique. Les données nanoLC-MS/MS brutes ont été en parallèle soumises à PEAKS pour le séquençage *de novo* des peptides. Le tableau de la figure 5 illustre la comparaison entre les séquences des peptides qui ont servi à l'identification de la protéine par Mascot et les séquences déterminées pour ces mêmes peptides par PEAKS. Un total de 79% des acides aminés ont été correctement déterminés par PEAKS.

Figure 5 : Evaluation du logiciel PEAKS par comparaison entre des séquences de peptides connus identifiés par Mascot et les séquences déterminées par séquençage *de novo* par PEAKS avec les pourcentages de confiance. Les acides aminés en rouge ont été correctement déterminés par PEAKS.



MASCOT						PEAKS	
Observed mass	Mr (expt)	Mr (calc)	Delta	Ion score	Sequence	Sequence	Confidence percentage
332,68	663,35	663,41	-0,06	18	KLTFR	KLTFR	29%
479,75	957,49	957,55	-0,06	58	ELLSAAAGVK	ELLSAAARK	21%
594,31	1186,61	1186,61	-0,08	81	TKELLSAAAGVK	TKELLSAAARK	50%
631,33	1260,64	1260,68	-0,04	55	GLALFVSGDNLR	LQLFVSGDNLR	34%
813,91	1625,81	1625,81	-0,11	95	ILGIPELLVSGTCamCVR	ILGIPELLVSGTCamCVR	96%
841,42	1680,84	1680,85	-0,01	116	FAEAGVIVVDNSSAFR	SSFPGVIVVDNSSAFR	29%
842,92	1683,83	1683,91	-0,08	130	VNIGVVGATGQVGQVMR	YYRVGATGQVGQVMR	23%
850,93	1699,84	1699,84	-0,07	105	VNIGVVGATGQVGQVMoxR	YYRVGATGGAVGQVMoxR	14%
914,96	1827,9	1827,94	-0,04	91	KDPDVPLVVSEVNFDR	KDPDVPLVVSEVLQYK	35%
1026,51	2051,01	2051,07	-0,06	98	LVDVPTPLAAAGIDDCamCLVGR	VVMPPTPLAAAGIDDAAMRR	45%
684,68	2051,01	2051301	-0,06	78	LVDVPTPLAAAGIDDCamCLVGR	HWCCamCATFAAGIDDCamCLVGR	8%
821,41	2461,22	2461,28	-0,06	82	LIVSSYQAVSGSGIAGVEELAGQAR	CamCCamCKYNFRAMMAGVEELAGQAR	18%

128 AAs correctement séquencés sur 163 = 79%

Présentant des performances élevées et permettant un gain en temps d'analyse considérable, le logiciel PEAKS est aujourd'hui principalement utilisé au laboratoire pour tous les projets nécessitant des interprétations par séquençage *de novo*.

Après l'étape de séquençage, des recherches de similarités de séquence doivent être effectuées avec les séquences déterminées afin d'identifier la fonction de la protéine par homologie de séquence.

2.3 Le MS-BLAST, un outil adapté pour les données de spectrométrie de masse en tandem

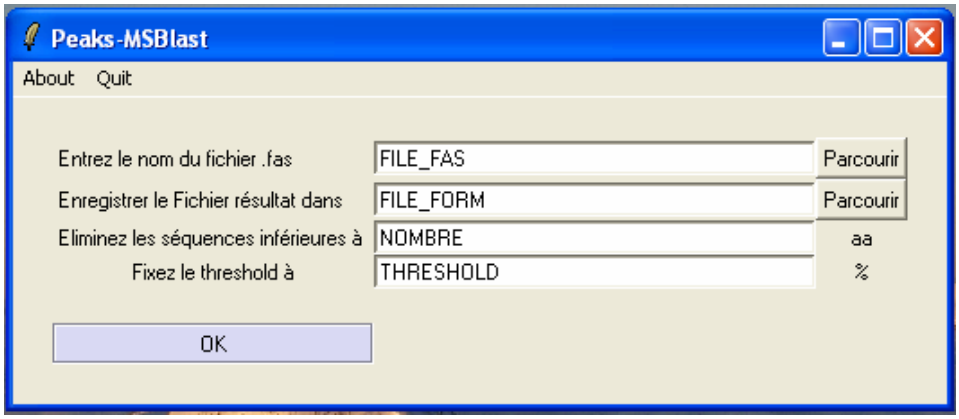
Les algorithmes de recherche de similarités de séquences conventionnels du type BLAST [Altschul et coll., 1997] ou FASTA [Pearson et coll., 1988] ont été optimisés pour la soumission de séquences exactes et de longueurs supérieures à 35 acides aminés [Altschul et coll., 1996 ; Pearson et coll., 1997]. Or les séquences déterminées à partir des spectres MS/MS ne sont en général que des fragments de séquences de peptides trypsiques et ne dépassent pas 10 à 15 acides aminés dans les meilleurs cas. En 2001, Shevchenko et coll. ont développé un protocole de recherche adapté pour ce type de données qu'ils ont appelé MS-BLAST [Shevchenko et coll., 2001 ; http://dove.embl-heidelberg.de/Blast2/msblast.html]. Le MS-BLAST utilise certaines options de l'algorithme WU-BLAST2 [Altschul et coll., 1996 ; Gish, 1996] mais déploie une matrice de scoring adaptée pour des données de MS. En effet, le MS-BLAST ne tolère pas d'interruptions (gaps) à l'intérieur d'un peptide mais en revanche les interruptions entre peptides ne sont aucunement pénalisées et peuvent être de n'importe quelle longueur. Cela permet de soumettre l'ensemble des tags de séquence obtenus pour l'ensemble des spectres MS/MS traités en une seule requête [Shevchenko et coll., 2003]. MS-BLAST identifie des HSPs (high-scoring segment pairs) entre des peptides soumis et des peptides présents dans les banques et leur attribue un score individuel. Lorsque plusieurs HSPs sont identifiées sur une même protéine, le score global tiendra compte de la somme des scores des HSPs individuelles. Ces scores seront comparés à des « scores seuils » calculés et fixés au préalable par l'algorithme pour attester du degré de confiance de l'identification. Le calcul du score ne tient pas compte du nombre de peptides soumis dans une requête ce qui présente le grand avantage de pouvoir soumettre l'ensemble des peptides déterminés par séquençage *de novo* et même d'y inclure des peptides redondants lorsqu'il y a doute et hésitation entre plusieurs séquences pour un même spectre. Une évaluation poussée des performances de l'algorithme MS-BLAST a été réalisée par Habermann et coll. [Habermann et coll., 2004] et les résultats de plus de 20000 requêtes MS-BLAST ont conduit à la détermination d'un taux d'identification de faux positifs de moins de 3%.

D'autres programmes basés sur l'algorithme FASTA ont également été adaptés pour les données de spectrométrie de masse [Mackey et coll., 2002 ; Huang et coll., 2001]. Cependant, même si ces programmes sont plus flexibles car ils considèrent les permutations isobariques dans les séquences peptidiques et d'autres paramètres, ces logiciels nécessitent des temps de recherche très longs, des ressources informatiques importantes et leurs scores sont dépendants et décroissent lorsque le nombre de séquences redondantes soumises augmente.

2.4 Automatisation des étapes de la stratégie par des scripts

Afin d'automatiser les étapes de cette approche et d'éviter aussi d'introduire des erreurs lors des nombreuses étapes de « copier-coller » des séquences, un script a été développé au laboratoire permettant de soumettre automatiquement les données de séquences générées par PEAKS au MS-BLAST (Figure 6).

Figure 6 : Ce programme permet de soumettre le fichier de sortie des séquences générées par PEAKS au format .fas, de le transformer au format nécessaire pour la soumission au MS-BLAST et de lancer ensuite automatiquement la requête MS-BLAST. Nous y avons implémenté un filtre permettant d'éliminer les séquences générées ayant un nombre d'acides aminés inférieur à un nombre X choisi ou celles ayant un pourcentage de confiance inférieur à un seuil X choisi.



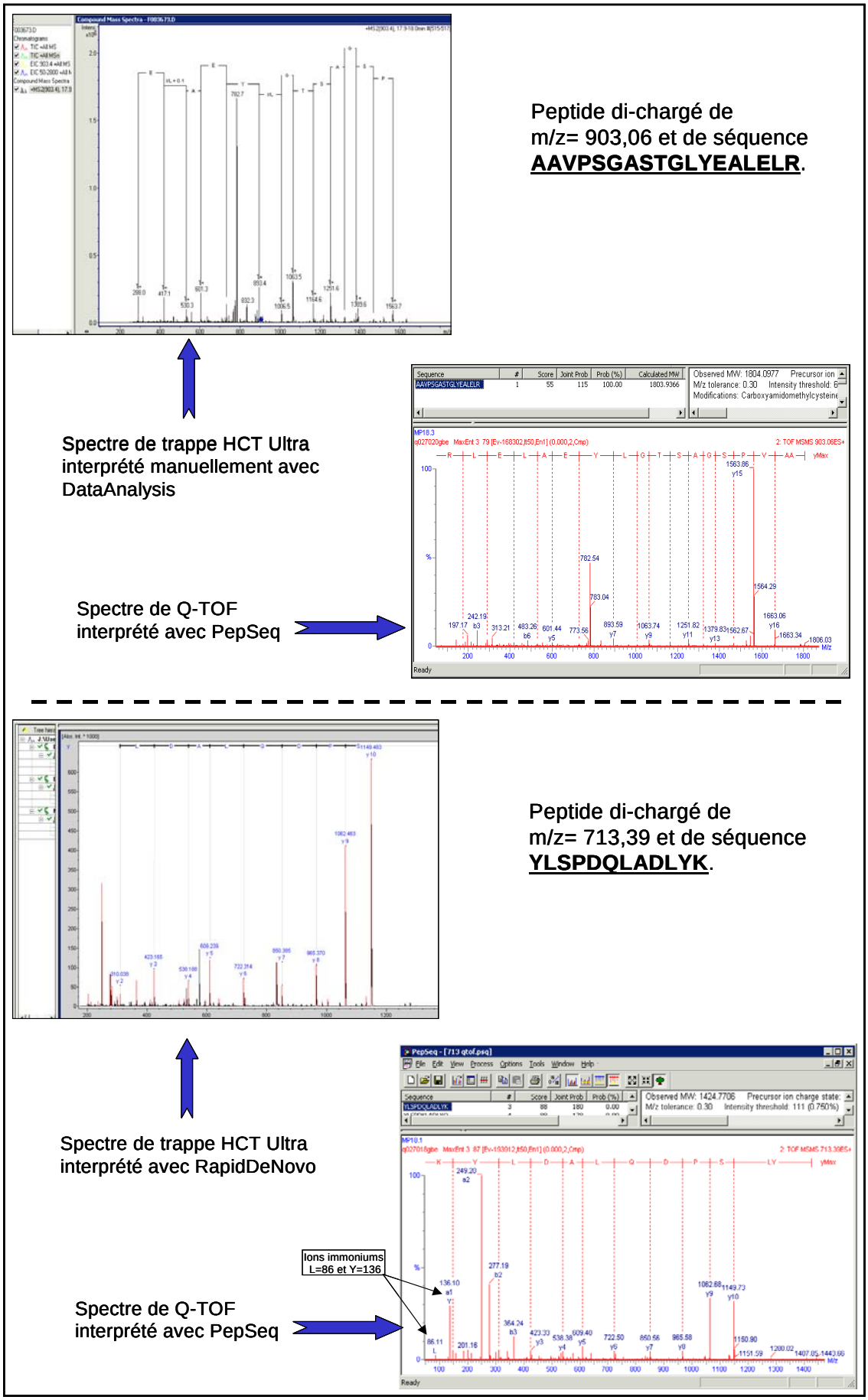
Par ailleurs, le programme MS-BLAST dont le code est disponible gratuitement, à la demande, pour les organismes publics a été installé en interne au laboratoire afin de pouvoir y intégrer nos propres banques de données. Le second avantage de cette installation en interne est aussi de ne plus être tributaires des longues files d'attente habituelles lorsque des requêtes sont soumises par l'interface web [http://dove.embl-heidelberg.de/Blast2/msblast.html].

3 Le séquençage *de novo* sur les instruments de type trappe ionique

Disposant au laboratoire à la fois d'un instrument de type Q-TOF (Q-TOF II, Waters) et de trappes ioniques (Esquire 3000+ puis trappe HCT Plus et HCT Ultra, Bruker Daltonics), les résultats obtenus sur ces différentes machines ont souvent été comparés pour la qualité des spectres MS/MS, les pourcentages de recouvrement des protéines identifiées, les précisions sur les mesures de masse, ... Les instruments de type trappe ionique d'ancienne génération (Esquire 3000+) jouissaient d'une « mauvaise réputation » pour le séquençage *de novo*, les spectres MS/MS étant trop peu intenses et trop « bruiteux » ce qui les rendaient difficilement interprétables par séquençage *de novo*. Cependant, les progrès instrumentaux significatifs qui ont été réalisés entre cette génération de trappe et les trappes de type High Capacity Trap (HCT) ont permis un véritable bon en avant en ce qui concerne la qualité des spectres MS/MS (voir Chapitre I de la Partie I des résultats).

Ainsi, à titre de comparaison, en figure 7, deux exemples de spectres MS/MS obtenus à partir d'un même échantillon dont la moitié a été injectée en couplage CapLC-Q-TOF (Waters) et la seconde moitié sur le système nanoLC (Agilent)/ HCT Ultra (Bruker Daltonics), traités et annotés par les différents outils disponibles au laboratoire.

Figure 7 : Comparaison entre des spectres MS/MS obtenus sur un Q-TOF II et sur une trappe ionique HCT Plus, annotés avec DataAnalysis, RapidDenovo (Bruker Daltonics) et PepSeq (Waters).



Globalement, les ions détectés sur ces spectres MS/MS sont comparables, ce qui est attendu vu la nature des fragmentations et des énergies mises en jeu lors des fragmentations basses énergies à la fois dans les appareils de type Q-TOF et trappe ionique. En effet, les spectres des deux instruments sont aussi informatifs les uns que les autres (avec cependant leurs spécificités dues aux principes de fonctionnements différents de ces deux types d'instruments comme par exemple l'absence des ions de basses masses due au cut-off dans les instruments de type trappe ionique). Les ions y- sont largement majoritaires et ont permis dans ces exemples, soit la lecture manuelle de la séquence (avec le peigne de DataAnalysis) soit aux algorithmes de déterminer automatiquement et correctement l'enchaînement des acides aminés.

Dans sa première version, l'algorithme du logiciel PEAKS avait été optimisé pour des données de type Q-TOF mais étant très intéressée pour l'utilisation de PEAKS avec les données de trappe ionique générées au laboratoire, des échanges de données et une collaboration avec les concepteurs du logiciel ont abouti à une version bien optimisée pour des spectres MS/MS de trappe ionique. En effet, dans la première version de PEAKS, la détection d'ions immoniums des différents acides aminés dans les basses masses était un critère important dans l'algorithme. Or, de part le principe même de fonctionnement de la trappe ionique, ces ions de basses masses ne peuvent pas être détectés et l'algorithme a donc du être adapté afin d'accorder moins d'importance à la présence de ces ions de basses masses.

Une évaluation du même type qu'en figure 5 a été réalisée avec des données de nanoLC-MS/MS générées avec la trappe HCT Ultra (Figure 8). Les pourcentages d'acides aminés correctement séquencés par *de novo* sont comparables à ceux précédemment obtenus avec les données de Q-TOF. De ce fait, les trappes sont utilisées, à ce jour, au même titre que le Q-TOF pour les applications nécessitant le recours au séquençage *de novo*.

Figure 8 : Evaluation des performances de PEAKS pour des données générées par des instruments de type trappe ionique. Les séquences identifiées lors de la recherche Mascot sont comparées aux séquences déterminées par PEAKS. Les acides aminés en rouge correspondent aux acides aminés correctement séquencés par PEAKS.

MASCOT						PEAKS	
Observed mass	Mr (expt)	Mr (calc)	Delta	Ion score	Sequence	Sequence	Confidence percentage
403,22	804,42	804,42	-0,01	45	VGVNGFGR	VGVNGFGR	51%
1247,12	2492,22	2492,31	-0,09	130	NTDIEIVAVNDLTDNATLAHLLK	WWAEIVAVNDLTDNVWAHPEK	47%
404,22	806,42	806,43	-0,01	49	FDSILGR	FDSILGR	52%
1008,00	2013,99	2014,01	-0,02	131	LPQDVSLEGDDTIVIGDTK	LPQDVSLEGDDTIVIGDTK	99%
828,75	2483,23	2483,14	0,09	50	YDGSQNIISNASCamCTTNCamCLGPLAK	LWKWEFFNASTTTNCamCLGPRP	2,80%
567,29	1132,57	1132,61	-0,04	58	VLNDEFGIVK	VLNDEFGIVK	57%
1192,59	2383,17	2367,13	0,04	54	GLMoxTTIHAYTQDQNLQDGPHK	WWEMHAYDNDQNLQDYQK	9%
526,77	1051,52	1051,66	-0,14	81	AIGLVLPELK	AIGLVLPELK	57%
496,76	991,5	991,55	-0,04	49	GKLDGYALR	GKLDGYALR	88%
856,43	1710,85	1710,94	-0,09	85	VPIPTGSVTDLTAELAK	VPIPTGSVTDLTAERR	51%
618,32	1234,62	1234,59	0,03	104	SASVEDINAAMK	SASVEDINAAMK	41%
626,32	1250,62	1250,58	0,04	72	SASVEDINAAMoxK	TGSVEDINAAMoxK	21%
378,71	755,40	755,42	-0,02	42	AAAEGPLK	AAAEGPLK	49%
394,22	786,42	786,42	0,00	30	VIDNQAK	VIDNQAK	21%
887,94	1773,87	1773,77	0,10	98	VVSWYDNEWGYSNR	VVSWYDNEWGYSNR	99%
499,76	997,51	997,62	-0,11	93	LADLVALVGK	LADLVALVGK	99%

172 AAs correctement séquencés sur 201 = 86%

4 Conclusions

Au vu du nombre de projets de séquençage de génomes qui se multiplient et des tailles des banques nucléotidiques et protéiques qui augmentent de façon exponentielle (voir chapitre I de l'introduction bibliographique), la question de la place promise aux approches utilisant le séquençage *de novo* dans l'avenir peut se poser. En effet, de plus en plus d'organismes voient leur génome séquencé et on peut donc se demander si les approches de recherches protéomiques classiques, non-tolérantes aux erreurs, ne vont pas permettre l'identification des protéines d'un très grand nombre d'espèces car chaque espèce pourra bientôt bénéficier d'un génome assez proche du sien séquencé. Car même si les algorithmes de séquençage *de novo* sont de plus en plus rapides et performants et que l'ensemble de la stratégie séquençage *de novo*/MS-BLAST est automatisable, cette approche reste toujours beaucoup moins rapide que les approches classiques (dans lesquelles c'est justement la spécificité exigée par l'identité parfaite des séquences qui a permis des gains en temps de recherche considérables).

Cette réflexion paraît toutefois un peu disproportionnée car même si le séquençage des génomes avance à une vitesse spectaculaire, leurs traductions en protéines et annotations ne sont pas prêtes d'être achevées [Pennisi, 2003]. D'ici à ce que l'ensemble des protéomes des génomes séquencés soient bien représentés dans les banques de données protéiques, une place de choix est encore promise à des approches basées sur les recherches de similarités de séquences dans des organismes modèles mieux connus.

Habermann et coll. [Habermann et coll., 2004] ont tenté d'évaluer la distance phylogénétique minimale permettant d'identifier les protéines non disponibles dans les banques grâce à des protéines orthogues proches présentes dans les banques par MS-BLAST. Ils ont par exemple pu conclure de leurs travaux qu'en recherchant des séquences de peptides de souris chez l'homme, 60% des protéines ont pu être identifiées par MS-BLAST avec trois peptides seulement et plus de 80% des protéines ont pu être identifiées dans des requêtes avec 15 peptides. Ils estiment d'ailleurs que 80% des protéines peuvent être identifiées par des approches de recherche de similarité de séquences à l'intérieur de la classe des mammifères. De la même manière, Liska et coll. [Liska et coll., 2003] ont illustré l'impact positif qu'a eu le séquençage complet du génome *d'Arabidopsis thaliana* pour la protéomique des plantes en général. Par ailleurs, le séquençage complet du génome du riz a ouvert de vastes possibilités pour les études protéomiques chez des céréales économiquement importantes comme le maïs, le blé ou l'orge car 98% des protéines connues de ces espèces ont des protéines homologues chez le riz [Goff et coll., 2002]. En revanche, le succès des recherches par similarité de séquence dans les lignées ayant divergées beaucoup plus tôt dans l'évolution, comme par exemple pour les champignons, sera moindre [Habermann et coll., 2004].

Ainsi, le succès des identifications de protéines par ces approches dans l'avenir dépendra fortement du choix des organismes qui auront leur génome séquencé, de sorte que chaque organisme non séquencé puisse au moins "bénéficier" du séquençage d'un organisme phylogénétiquement proche.

Je pense qu'un bel avenir est promis à des moteurs de recherche qui combineront efficacement à la fois plusieurs approches de recherche avec des données de MS/MS. Par exemple, on peut envisager dans un premier temps un algorithme de recherche classique de PFF (Peptide Fragmentation Fingerprinting), basée sur l'identité parfaite des séquences, qui permet de cribler rapidement les banques qui sont de plus en plus volumineuses. Puis dans un second temps, les informations de séquences obtenues par séquençage *de novo* sur les spectres MS/MS de qualité non attribués pourront être recherchées par des algorithmes de recherche de similarités de séquence pour compléter les identifications, détecter des mutations, …

CHAPITRE II

Identification des protéines impliquées dans la résistance à l'arsenic chez la bactérie *Caenibacter arsenoxydans*

Cette étude a été réalisée en collaboration avec l'équipe de Génétique Moléculaire, Génomique, Microbiologie (CNRS/ULP-UMR 7156) du Professeur Philippe Bertin à l'Institut Botanique de l'Université Louis Pasteur de Strasbourg. Chronologiquement, notre collaboration avec le laboratoire de P. Bertin a démarré par le travail présenté dans ce chapitre. A cette date, donc fin 2003, le génome de la bactérie qui s'appelait *Caenibacter arsenoxydans* n'était pas encore séquencé. Ce travail consiste en une étude différentielle par gel d'électrophorèse 2D pour l'identification des protéines impliquées dans la résistance à l'arsenic chez cette bactérie. Une stratégie par séquençage *de novo* s'est imposée, cette bactérie ne présentant, au vu des premiers résultats, aucune forte homologie avec des bactéries dont le génome était disponible à ce moment-là. Cette bactérie changera ensuite de nom en mars 2006 pour devenir *Herminiimonas arsenicoxydans* et la suite du projet est décrite dans le Chapitre II de la Partie IV des résultats.

1 Contexte de cette étude

1.1 *Caenibacter arsenoxydans* et sa résistance à l'arsenic

Provenant tant de sources naturelles qu'anthropogéniques, l'arsenic est largement répandu dans l'environnement, essentiellement sous 3 états d'oxydation: As(-III) (arsine), As(+III) (arsenite) et As(+V) (arseniate). Une contamination des eaux de distribution par les formes solubles et toxiques, arsenite et arseniate, est fréquemment rapportée et l'arsenic a été identifié comme un risque majeur pour la santé humaine en divers endroits du monde [Smith et coll., 2002]. L'arsenic est généralement associé au soufre (comme dans l'arsénopyrite (FeAsS)), au fer et à divers autres métaux dans les minéraux sulfurés et dans les

milieux hydrothermaux résultant d'une activité volcanique récente ou ancienne. L'écologie de ce métalloïde est fortement dépendante de transformations microbiennes qui affectent la mobilité, la biodisponibilité ainsi que la toxicité de l'arsenic dans l'environnement. Des bactéries impliquées dans les processus de biotransformations tels que réduction, oxydation et méthylation ont été décrites [Oremland et coll., 2003 ; Muller et coll., 2003]. A ce jour, le mécanisme de transformation le mieux étudié est la réduction de l'arseniate (AsV) en arsenite (AsIII). En revanche, la physiologie, l'enzymologie et la génétique de l'oxydation de l'arsenite ainsi que les processus de régulation sous-jacents restent à ce jour assez méconnus.

La bactérie *Caenibacter arsenoxydans* est une β-protéobactérie hétérotrophe isolée dans des boues d'une station d'épuration industrielle capable d'oxyder l'arsenite (As[III]) en arsenate (As[V]) [Weeger et coll., 1999]. L'observation des cellules en microscopie révèle une morphologie en bâtonnet légèrement incurvé présentant à son extrémité un flagelle polaire. Une meilleure connaissance des mécanismes impliqués dans la résistance de cette bactérie à l'arsenic pourrait être avantageusement mise à profit dans des procédés de bioremédiation.

1.2 Le génome et le protéome de la bactérie

Lorsque ce projet a démarré en septembre 2003, le génome de *C. arsenoxydans* n'était pas encore séquencé mais un projet de séquençage aléatoire global avec une profondeur de 10X du génome, en collaboration avec le Génoscope venait d'être lancé (voir Chapitre II de la partie IV des résultats). A la date de ces analyses, aucune donnée concernant le génome ni le protéome de *C. arsenoxydans* n'était disponible.

1.3 Objectif de l'étude

L'objectif de cette étude est de déterminer les gènes et protéines différentiellement exprimés et accumulés lorsque la bactérie *C. arsenoxydans* est cultivée en présence d'arsenic. Pour cela, des expériences de mutagenèse par transposition aléatoire, suivies par une analyse quantitative des ARNm ont été réalisées afin d'identifier les effets de l'arsenic au niveau de l'expression des gènes. En parallèle, une étude protéomique différentielle par gel d'électrophorèse 2D afin d'analyser ces effets au niveau de l'accumulation des protéines a été menée.

2 Choix de la stratégie expérimentale mise en œuvre

2.1 Préparation des échantillons

Les cultures bactériennes induites par l'arsenic ont été supplémentées d'une solution d'arsenite de sodium ($NaAsO_2$) pour obtenir une concentration finale d'As [III] de 2,66 mM.

Les extraits protéiques obtenus dans les différentes cultures avec/sans arsenic ont été déposés sur gel d'électrophorèse 2D. Les protocoles détaillés de cette préparation d'échantillons sont décrits dans la publication des résultats.

2.2 Analyse par spectrométrie de masse

Le choix de la technique d'analyse par spectrométrie de masse pour l'identification des protéines devait inévitablement tenir compte de l'absence de données de génome et de protéome de l'organisme. Une approche permettant d'obtenir des informations de séquence était indispensable, l'ensemble des analyses a été réalisé par nanoLC-MS/MS sur un système CapLC/Q-TOF II (Waters).

2.3 Stratégie d'identification des protéines

Une première soumission des données de nanoLC-MS/MS au moteur de recherche Mascot a été réalisée dans la banque protéique NCBInr. Ces requêtes ont été sans succès dans la grande majorité des spots, à l'une ou l'autre exception près pour lesquelles un peptide a pu être identifié par homologie chez *Ralstonia solanacearum* ou *Bordetella pertussis*, des espèces taxonomiquement assez proches de *C. arsenoxydans*. Dans un second temps, l'ensemble des spectres a été interprété systématiquement par séquençage *de novo* avec le logiciel PEAKS suivi par des requêtes MS-BLAST.

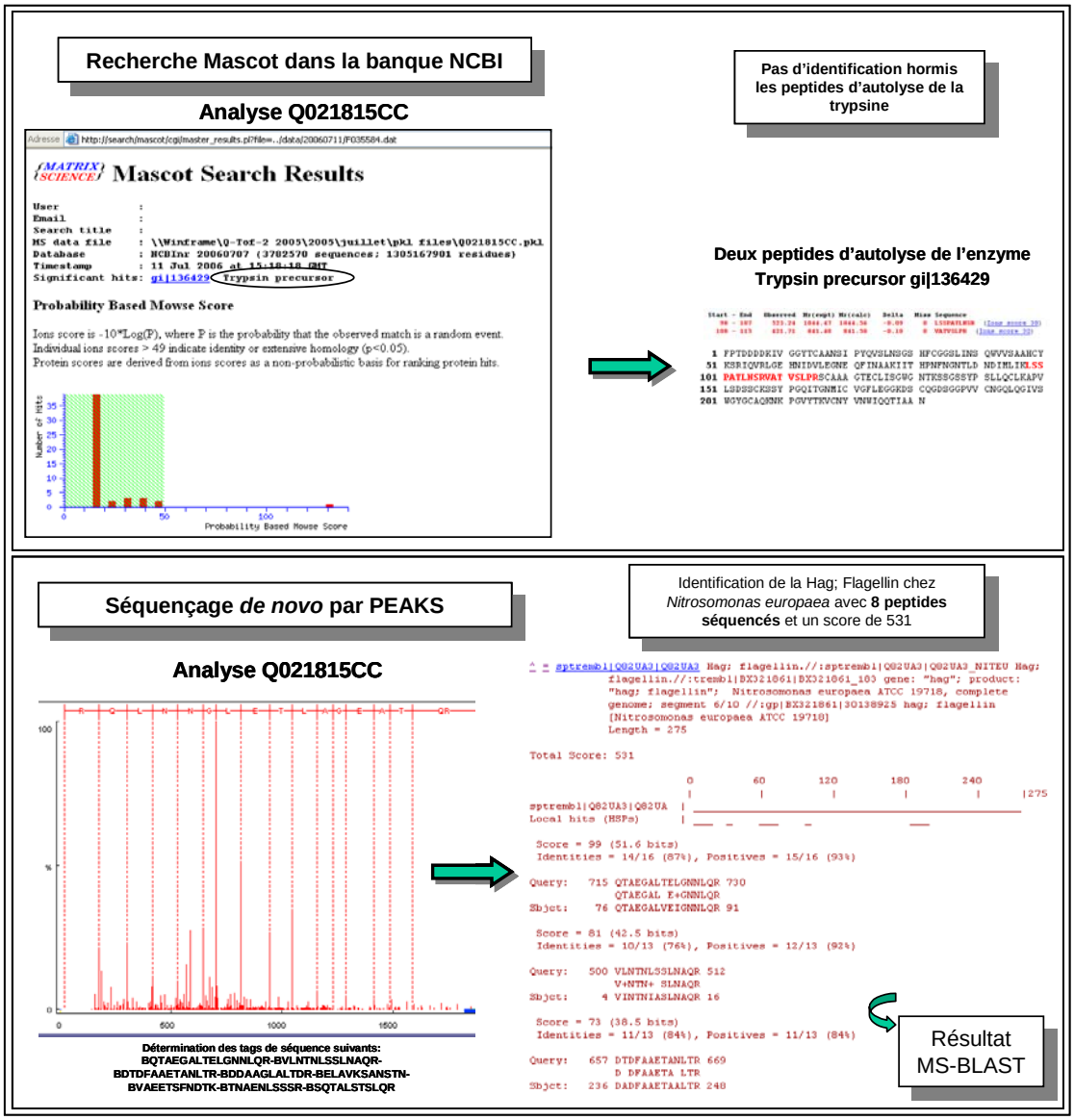
3 Résultats publiés

Les résultats obtenus dans ce travail ont fait l'objet d'une publication acceptée dans le journal *Biochimie* en novembre 2005.

Au total, 16 gènes et 22 protéines impliqués dans la réponse au stress arsénieux chez *C. arsenoxydans* ont pu être identifiés. Deux gènes impliqués dans l'oxydation de l'arsenic (*aoxA* et *aoxB*) avaient été précédemment décrits [Muller et coll., 2003] et les analyses faites ici ont permis de compléter ces résultats par l'identification d'un nouveau gène et deux nouvelles protéines directement impliqués dans le mécanisme de résistance à l'arsenic. En plus de cela, les autres protéines et gènes identifiés montrent que la réponse à l'arsenic touche également d'autres fonctions cellulaires comme celles du métabolisme énergétique ou de la structure de l'enveloppe cellulaire. Il est important de noter aussi l'intérêt d'étudier cette réponse à la fois au niveau des ARNm et des protéines car ces deux approches permettent d'apporter des indications complémentaires pour la compréhension de ce genre de mécanismes.

Du point de vue de la stratégie analytique mise en œuvre pour l'analyse protéomique, l'utilisation du séquençage *de novo* a été indispensable ici car sans la possibilité d'utiliser cette approche « tolérante aux erreurs », l'ensemble des protéines n'aurait pu être identifié. En effet, comme illustré en figure 1, cette bactérie ne présentant aucune forte homologie avec les organismes bactériens séquencés à ce jour, les requêtes Mascot ne donnaient aucun succès d'identification. Une explication complémentaire à la justification de l'«originalité » de ce génome réside dans le fait que cette bactérie évolue dans des milieux très complexes et très riches en organismes que sont les boues d'épuration et a donc été sujette à de nombreux transferts horizontaux de gènes.

Figure 1 : Comparaison des résultats obtenus lors d'une recherche Mascot classique dans la banque NCBInr avec les résultats obtenus par séquençage *de novo* suivi par MS-BLAST pour des données d'une analyse nanoLC-MS/MS.



4 Conclusions et perspectives du travail

Cette étude a permis de mettre en évidence des caractéristiques remarquables chez la bactérie *C. arsenoxydans* dans son adaptation à croître dans des milieux fortement contaminés par l'arsenic. La poursuite et les perspectives de ce travail seront décrits dans le chapitre III de la Partie IV des résultats car le séquençage du génome de la bactérie ayant été achevé entre temps, des analyses beaucoup plus vastes du génome et du protéome ont pu être réalisées. Cette application illustrera l'ampleur et les possibilités qu'ouvre la connaissance de la séquence du génome pour l'élucidation et la compréhension du fonctionnement de l'organisme.

Identification of genes and proteins involved in the pleiotropic response to arsenic stress in *Caenibacter arsenoxydans*, a metalloresistant beta-proteobacterium with an unsequenced genome

Christine Carapito, Daniel Muller, Evelyne Turlin, Sandrine Koechler, Antoine Danchin, Alain Van Dorsselaer, Emmanuelle Leize-Wagner, Philippe N. Bertin et Marie-Claire Lett

Biochimie, 2006, Vol. 88, Pages 595-606

Pages 595 à 606:

CHAPITRE III

Analyse protéomique de tissus de vigne (*Vitis vinifera L.*) ayant subi un stress herbicide

Cette étude a été réalisée en collaboration avec le Dr. Antonio Castro et le Dr. Christophe Clément du laboratoire de stress, défenses et reproduction des plantes de l'Université de Reims Champagne Ardenne.

1 Contexte de cette étude

1.1 *Vitis vinifera* et l'herbicide Flumioxazin (fmx)

Vitis vinifera (2n=38) appartient à la famille des *Vitaceae*. Cette famille, présente majoritairement dans la zone intertropicale, contient 17 genres, essentiellement des lianes pérennes ou herbacées. Seul le genre *Vitis* produit des fruits comestibles. Il contient environ 60 espèces dioïques distribuées de façon équitable entre les continents américain et asiatique. Cette diversité représente une ressource considérable de gènes de résistance aux agents pathogènes de la vigne [Mullins et coll., 1992 ; Jansen et coll., 2006]. *Vitis vinifera* est la seule espèce originaire d'Eurasie mais elle a été dispersée par l'homme à travers le monde pour la production de raisins de table et de vin. Elle est ainsi devenue une espèce majeure dans l'agriculture européenne et son importance économique est capitale pour certaines régions. Pour augmenter sa compétitivité économique et l'adapter aux nouveaux objectifs de production résultant des changements climatiques, de l'émergence de nouvelles maladies et des impératifs de protection de l'environnement, il est nécessaire d'accroître nos connaissances sur sa physiologie et ses pathologies.

La flumioxazin (2-[7-fluoro-3,4-dihydro-3-oxo-4-(2-propynyl)-2*H*-1,4-benzoxazin-6-yl]-4,5,6,7-tetrahydro-1H-isoindole-1,3(2H)-dione) fait partie des nouveaux herbicides synthétiques dont l'utilisation est autorisée et largement répandue depuis peu dans les vignobles [Tomlin, 2000]. Il s'agit d'un herbicide à large spectre de la famille des N-phenylphthalimide inhibant la protoporphyrinogen IX oxidase, une enzyme clé de la biosynthèse de la chlorophylle [Lee et coll., 1994]. De précédentes études ont déjà mis en évidence des effets physiologiques néfastes du fmx sur la vigne [Saladin et coll., 2003a ; Saladin et coll., 2003b].

Ces dernières années, les effets d'un certain nombre de facteurs abiotiques sur des plantes modèles ont été déterminés par des approches protéomiques : la sécheresse chez *Oryza sativa* [Salekdeh et coll., 2002], la réponse à des températures élevées chez le maïs [Lund et coll., 1998], la pollution par des métaux lourds chez *Oryza sativa* [Hajduch et coll., 2001], … Dans ce contexte, une approche de ce type a été développée afin d'étudier les effets de traitements à l'herbicide fmx au niveau du protéome de différents tissus de *Vitis vinifera*.

1.2 Le génome et le protéome de *Vitis vinifera*

En juillet 2006, la banque protéique du NCBI ne compte que 914 entrées annotées *Vitis vinifera*, la banque UniProtKB en compte 588 dont 26 dans UniProtKB/Swiss-Prot. Très peu d'informations sont donc disponibles en ce qui concerne le protéome de *Vitis vinifera*. Néanmoins, à l'automne 2004, le conseil scientifique du génoscope français a déposé un projet de séquençage du génome de *Vitis vinifera* par « whole genome shotgun sequencing » avec une profondeur 3X. En parallèle l'Institut National de la Recherche Agronomique (INRA) a fait la demande de séquençage complémentaire pour atteindre un niveau 3X. Et finalement en juillet 2005, la France et l'Italie ont signé un accord de collaboration afin d'atteindre un niveau 12X pour le génome de la vigne et pour le séquençage en parallèle d'une souche de levure utilisée en vinification ainsi que de l'agent pathogène *Botrytis cinerea*. Le génome est donc en cours de séquençage au génoscope, des séquences de gènes sont déjà disponibles mais beaucoup de séquences partielles uniquement.

1.3 Objectif de l'étude

Des études poussées des effets des herbicides sur les plantes cibles sont toujours largement menées de même que de leurs effets sur les rendements des récoltes. Par contre l'analyse des effets au niveau moléculaire de la toxicité de ces substances sur les organismes non directement ciblés est quelques fois négligée. C'est dans ce contexte qu'une approche protéomique a été développée ici afin de déterminer, *in vivo*, les effets du traitement au fmx au niveau du protéome des racines, tiges et feuilles de la vigne. Une approche différentielle par gel d'électrophorèse 2D a été réalisée et les spots différentiellement exprimés ont été analysés et identifiés par nanoLC-MSMS.

2 Choix de la stratégie expérimentale mise en œuvre

2.1 Préparation des échantillons

En résumé, des plants de vigne ont été cultivés sur milieu contenant 10µM de fmx pendant 1, 2, 4, 6, et 21 jours avant d'être récoltés. Des extraits protéiques de tiges, de racines et de feuilles ont été préparés et déposés sur gel d'électrophorèse 2D. Le détail des procédures de préparation des extraits protéiques ainsi que des gels d'électrophorèse est décrit dans la publication des résultats (3). Un total de 120 gels 2D a été réalisé, incluant des gels témoin, des gels correspondant aux différents temps d'incubation et des duplicats afin d'attester de la reproductibilité des expériences. Les images de gels ont été analysées et comparées grâce au logiciel Melanie II (Bio-Rad).

2.2 Analyse par spectrométrie de masse

Au vu de la faible représentation de *Vitis vinifera* dans les banques protéiques, une approche par nanoLC-MS/MS a été choisie car l'information de séquence apportée par la MS/MS est indispensable pour l'étude d'organismes non séquencés. Les analyses ont été réalisées avec un couplage CapLC-Q-TOF II (Waters).

2.3 Stratégie d'identification des protéines

Dans un premier temps, l'ensemble des données de nanoLC-MS/MS ont été soumises à des requêtes Mascot dans la banque NCBInr sans appliquer de restriction de taxonomie. Malgré le peu d'entrées annotées pour *Vitis vinifera* dans NCBI, quelques protéines ont pu être identifiées par cette approche non-tolérante aux erreurs par homologie chez *Arabidopsis thaliana*, *Oryza sativa* ou d'autres espèces phylogénétiquement proches (*Hordeum vulgare*, *Rosa hybrid*, …). Cependant, comme précédemment décrit dans le chapitre I de cette partie de résultats, les homologies doivent être très fortes pour pouvoir faire des identifications inter-espèces par des approches non-tolérantes aux erreurs comme le propose l'algorithme de Mascot. Dans un deuxième temps, l'ensemble des spectres MS/MS a donc été traité par séquençage *de novo* et les identifications des protéines ont été confirmées et complétées grâce aux séquences ainsi déterminées. Les logiciels PEAKS Studio (Bioinformatics Solutions) [Ma et coll., 2003] et PepSeq (Waters) ont été utilisés pour le traitement des spectres. Les identifications de fonction des protéines consécutives ont été réalisées par MS-BLAST.

3 Résultats publiés

Les résultats obtenus dans cette étude ont fait l'objet d'une publication acceptée en juillet 2005 dans *Journal of Experimental Botany*.

En résumé, 11 spots ont été détectés comme différentiellement accumulés dans les extraits de racines, 13 spots dans les extraits de tiges et 9 spots dans les extraits de feuilles. Un total de 29 spots ont pu être identifiés avec succès grâce aux données de nanoLC-MS/MS. Les protéines identifiées ici sont en bonne corrélation avec les effets du fmx précédemment identifiés au niveau physiologique [Saladin et coll., 2003a ; Saladin et coll., 2003b] et tous ces résultats combinés suggèrent une action systématique de l'herbicide dans tous les tissus de la vigne, probablement après absorption par les racines. L'intérêt biologique de cette approche protéomique et plus particulièrement des protéines identifiées est largement décrit dans la publication.

Mis à part ces intérêts, l'aspect le plus important à retenir du point de vue analytique est la nécessité et l'utilité du recours au séquençage *de novo* pour l'identification de protéines lorsque l'on travaille sur des organismes non séquencés et non représentés dans les banques de données protéiques.

En effet, dans les données de nanoLC-MS/MS, de nombreux spectres MS/MS n'étaient pas attribués lors des recherches Mascot classiques alors que leur qualité permet d'en déduire une séquence en acides aminés sans ambiguïté.

Comme illustré dans le tableau 1, le séquençage de tous ces spectres non attribués et les requêtes MS-BLAST consécutives ont permis de largement compléter la qualité des identifications.

Pour les spots GRP3 et GRP4, les protéines avaient été identifiées avec un seul peptide dans les requêtes Mascot, donc sans grande confiance : ces identifications ont pu être complétées avec 2 et 4 peptides, respectivement, grâce au séquençage *de novo*. Le spot GRP5 n'a conduit à aucune identification par Mascot alors que 3 belles séquences ont pu être déterminées par séquençage *de novo* et ont conduit à l'identification de la D-Protein chez *Hordeum vulgare* avec un score MS-BLAST de 180. Enfin, l'identification de la 2,3-bisphosphoglycerate-independent phosphoglycerate mutase a pu être complétée par une séquence.

Tableau 1 : Résultats d'identification de 4 spots de racine combinant une recherche Mascot à une approche par séquençage *de novo* : les peptides identifiés par la recherche Mascot sont présentés en caractères noirs et les peptides identifiés par séquençage *de novo* sont donnés en caractères gras, italiques et rouges.

Spot	Tissus	Masses théoriques/Masses expérimentales	Protéines identifiées (*espèce*)	Numéro d'accession	Peptides identifiés	Total score MS-BLAST
GRP2	Racine	61804/53394	2,3-bisphosphoglycerate-independent phosphoglycerate mutase (*Prunus dulcis*)	O24246	LDQLQLLLK-GVDAQIASGGGR-AVEIAEK-***LYEGEGFK***	235
GRP3	Racine	54911/51887	UTP-glucose-1P-uridyltransferase (*Arabidopsis thaliana*)	P57751	LVEADALK-***LVQLETAAGAAIR-SGFINLVSR***	189
GRP4	Racine	37635/41278	Orcinol O-methyltransferase *(Rosa hybrid)*	Q8L5K8	VIIIDMMMENQK-***CTVLDLPHVRAD-FNEAMASDAR-VDVGGGTGDAGFSGYK-EEGYVLTHAS***	347
GRP5	Racine	34801/34021	D-protein (*Hordeum vulgare*)	Q8VWY8	***SSHALALVGQK-LLSSGDAGPPHR-AYVFGGELTPR***	180

4 Conclusions et perspectives du travail

Jusqu'alors aucune étude relatant des mécanismes de stress induits *in vivo* par des pesticides au niveau du protéome de plantes non ciblées n'a été publiée. Les résultats obtenus ici sont en bonne corrélation avec ceux précédemment obtenus au niveau physiologique pour le même système et cette étude constitue donc une base pour des études futures. Certaines des protéines identifiées pourront être utilisées comme marqueurs biochimiques de la présence de fmx dans les tissus de la vigne de même que pour l'évaluation de l'impact du fmx sur d'autres plantes non-ciblées de l'écosystème du vignoble. Afin d'éviter de masquer les effets du fmx par des fluctuations extérieures liées à l'environnement pouvant induire des stress sur les plants de vigne dans le vignoble, cette étude a été réalisée sur des plants cultivés *in vitro* dans un premier temps. La première perspective de cette étude est donc de transposer et de vérifier ces données sur des plants de vigne cultivés en extérieur. Par ailleurs, des protéines appartenant à la famille multigénique des PR-10 (pathogenesis related proteins) ont été identifiées dans plusieurs spots, des analyses complémentaires vont permettre de caractériser plus précisément les différents membres de cette grande famille : des recherches avec les données de nanoLC-MS/MS directement dans le génome seront réalisées afin de discriminer parmi les gènes paralogues de cette famille ceux réellement exprimés. Des analyses MS complémentaires pour la caractérisation d'éventuelles modifications post-traductionnelles de ces PR-10 sont également prévues.

Enfin, ce travail constituant une preuve des effets néfastes du fmx sur la physiologie de la vigne, il pourrait contribuer à l'établissement de règles d'utilisation plus écologiques de certains herbicides chimiques.

Proteomic analysis of grapevine (Vitis vinifera L.) tissues subjected to herbicide stress

Antonio J. Castro, **Christine Carapito**, Nathalie Zorn, Christian Magné, Emmanuelle Leize, Alain Van Dorsselaer et Christophe Clément

Journal of Experimental Botany, 2005, Pages 1-13

Bibliographie

Altschul S. F. and Gish W.

Local alignment statistics. *Methods enzymol.*, **1996**, 266; 460-480.

Altschul S. F., Madden T. L., Shaffer A. A., Zhang J., Zhang Z., Miller W. and Lipman D. J.

Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.,* **1997**, 25, 3389-3402.

Clauser K. R., Baker P. and Burlingame A. L.

Role of accurate mass measurement (+/- 10ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal Chem*, **1999**, 71, 2871-2882.

Eng J., McCormack A. and Yates J.

An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.,* **1994**, 5, 976-989.

Gish W.

WU-BLAST2.0. blast.wustl.edu, **1996**.

Goff S. A., Ricke D., Lan T. H., Presting, G., Wang R., Dunn M., Glazebrook J., Sessions A. Oeller P., Varma H., Hadley D., Hutchison D., Martin C., Katagiri F., Lange B. M, Moughamer T., Xia Y., Budworth P., Zhong J., Miguel T., Paszkowski U., Zhang S., Colbert M., Sun W. L, Chen L., Cooper B., Park S., Wood T. C., Mao L., Quail P., Wing R., Dean R., Yu Y., Zharkikh A., Shen R., Sahasrabudhe S., Thomas A., Cannings R., Gutin A., Pruss D., Reid J., Tavtigian S., Mitchell J., Eldredge G., Scholl T., Miller R. M., Bhatnagar S., Adey N., Rubano T., Tusneem N., Robinson R., Feldhaus J., Macalma T., Oliphant A. and Briggs S.

A draft sequence of the rice genome (*Oryza sativa L. ssp. japonica*). *Science*, **2002**, 296, 92-100.

Habermann B., Oeagema J., Sunyaev S. and Shevchenko A.

The power and the limitations of cross-species protein identification by mass spectrometry-driven sequence similarity searches. *Mol. Cell. Proteomics*, **2004**, 3, 238-249.

Hajduch M., Rakwal R., Agrawal G. K., Yonekura M. and Pretova A.

High-resolution two-dimensional electrophoresis separation of proteins from metal-stressed rice (*Oryza sativa L.*) leaves: drastic reductions/fragmentation of ribulose-1,5-bisphosphate carboxylase/oxygenase and induction of stress-related proteins. *Electrophoresis*, **2001**, 22, 2824-2831.

Huang L., Jacob R. J., Pegg S.C., Baldwin M. A., Wang C. C., Burlingame A. L. and Babbitt P. C.

Functional assignment of the 20 S proteasome from *Trypanosoma brucei* using mass spectrometry and new bioinformatics approaches. *J Biol Chem*., **2001**, 276(30), 28327-28339.

Jansen R. K., Kaittanis C., Saski C., Lee S. B., Tomkins J., Alverson A. J. and Daniell H.

"Phylogenetic analyses of *Vitis* (Vitaceae) based on complete chloroplast genome sequences : effects of taxon sampling and phylogenetic methods on resolving relationships among rosids." *BMC Evolutionary Biology*, **2006**, 6, 32L.

Liska A. J. and Shevchenko A.

Expanding the organismal scope of proteomics: cross-species protein identification by mass spectrometry and its implications. *Proteomics*, **2003**, 3, 19-28.

Lund A. A., Blum P. H., Bhattramakki D. and Elthon T. E.

Heatstress response of maize mitochondria. *Plant Physiology*, **1998**, 116, 1097-1110.

Ma B., Zhang K., Hendrie C., Liang C., Li M., Doherty-Kirby A. and Lajoie G.

PEAKS: Powerful Software for Peptide De Novo Sequencing by MS/MS. *Rapid Commun. Mass Spectrom.*, **2003**, 17, 2337-2342.

Mackey A. J., Haystead T. A. and Pearson W. R.

Getting more from less: algorithms for rapid protein identification with multiple short peptide sequences. *Mol. Cell. Proteomics*., **2002**, 1(2), 139-147.

Muller D., Lièvremont D., Simeonova D., Hubert J. C. and Lett M. C.

Arsenite oxidase aox genes from a metal-resistant β-Proteobacterium. *J. Bacteriol.,* **2003**, 185, 135-141.

Mullins M. G., Bouquet A. and Williams L. E.

"Biology of grapevine". *Cambridge University Press*, **1992**, Cambridge.

Oremland R. S. and Stolz J. F.

The ecology of arsenic. *Science*, **2003**, 300, 939-944.

Pearson W. R. and Lipman D. J.

Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U.S.A.,* **1988**, 85, 2444-2448.

Pearson W. R., Wood T., Zhang Z. and Miller W.

Comparison of DNA sequences with protein sequences. *Genomics*, **1997**, 46, 24-36.

Pennisi E.

Bioinformatics. Gene counters struggle to get the right answer. *Science,* **2003**, 301, 1040-1041.

Perkins D. N., Pappin D. J., Creasy D. M. and Cottrell J. S.

Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **1999**, 20(18), 3551-3567.

Saladin G., Magné C. and Clément C.

Impact of flumioxazin herbicide on growth and carbohydrate physiology in *Vitis vinifera L. Plant Cell Reports*, **2003a**, 21, 821-827.

Saladin G., Clément C. and Magné C.

Stress effects of flumioxazin herbicide on grapevine (*Vitis vinifera L.*) grown *in vitro*. *Plant Cell Reports*, **2003b**, 21, 1221-1227.

Salekdeh G. H., Siopongco J., Wade L. J., Ghareyazie B. and Bennett J.

Proteomic analysis of rice leaves during drought stress and recovery. *Proteomics,* **2002**, 2, 1131-1145.

Shevchenko A., Chernushevic I., Wilm M. and Mann M.

"*De novo*" sequencing of peptides recovered from in-gel digested proteins by nano-electrospray tandem mass spectrometry. *Mol. Biotechnol.,* **2002**, 20, 107-118.

Shevchenko A., Sunyaev S., Liska A., Bork P. and Shevchenko A.

Nanoelectrospray tandem mass spectrometry and sequence similarity searching for identification of proteins from organisms with unknown genomes. *Methods Mol. Biol.*, **2003**, 211, 221-234.

Shevchenko A., Sunyaev S., Loboda A., Shevchenko A., Bork P., Ens W. and Standing K. G.

Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching. *Anal. Chem.*, **2001**, 73(9), 1917-1926.

Smith A. H., Lopipero P. A., Bates M. N. and Steinmaus C. M.

Arsenic epidemiology and drinking water standards. *Science*, **2002**, 296, 2145-2146.

Tomlin CDS.

Flumioxazin. *In: Tomlin CDS, ed. Pesticide manual. Surrey, British Crop Protection Council*, **2000**, 441-442.

Weeger W., Lièvremont D., Perret M., Lagarde F., Hubert J. C., Leroy M. and Lett M. C.

Oxidation of arsenite to arsenate by a bacterium isolated from an aquatic environment. *Biometals*, **1999**, 12, 141-149.

RESULTATS

QUATRIEME PARTIE :
Applications des recherches de données nanoLC-MS/MS dans des génomes complets non annotés

Chapitre I : Une stratégie de recherche avec des données nanoLC-MS/MS dans le génome complet non annoté

Chapitre II : Les recherches dans les génomes : un outil d'aide à l'annotation des génomes bactériens

Chapitre III : Recherche dans le génome complet pour la discrimination des gènes paralogues au sein de grandes familles multigéniques

CHAPITRE I

Une stratégie de recherche avec des données nanoLC-MS/MS dans le génome complet non annoté

Dans l'ère actuelle de séquençage des génomes à haut débit, un grand nombre de génomes sont entièrement séquencés mais leur annotation est loin d'être achevée, complète et vérifiée. Les protéomes d'un nombre significatif de génomes sont donc peu ou mal représentés dans les banques de données protéiques. Dans ce contexte, cette partie des résultats sera consacrée à la description d'une stratégie permettant les recherches avec des données de nanoLC-MS/MS directement dans des données génomiques brutes non interprétées. Deux applications particulières de cette stratégie permettront de mettre en évidence les avantages de cette approche.

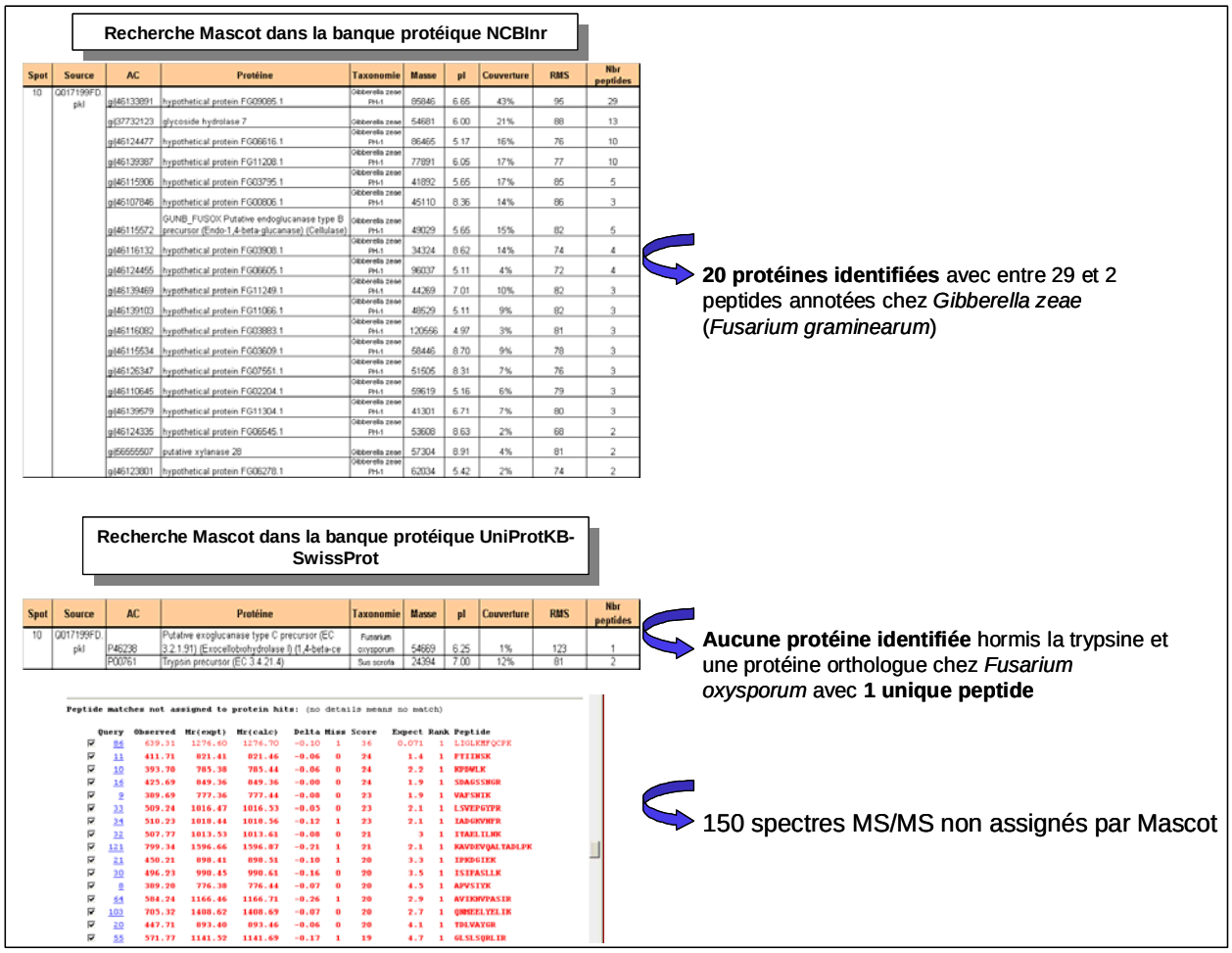
1 Pourquoi rechercher directement dans le génome ?

1.1 L'importance du choix de la banque de données dans laquelle sont effectuées les recherches protéomiques

Indépendamment de la qualité des données de spectrométrie de masse en elles-mêmes, le choix de la banque dans laquelle sont recherchées ces données est également crucial pour le succès de l'identification des protéines. Les moteurs de recherche classiquement utilisés comme Mascot étant basés sur des identités strictes et étant non tolérantes aux erreurs, l'absence de la protéine étudiée dans la banque conduira souvent à l'échec de l'identification de la protéine. La figure 1 présente un exemple d'analyse nanoLC-MS/MS réalisée sur une bande de gel 1D de protéines du champignon pathogène *Fusarium graminearum* (voir Chapitre II Partie II des résultats) : les données ont été recherchées avec Mascot à la fois dans la banque protéique NCBInr et dans la banque UniProtKB/SwissProt. Le protéome de *Fusarium graminearum* n'a pas encore été intégré dans la banque UniProtKB/SwissProt, les annotations n'étant pas de qualité suffisante et la recherche n'a donc conduit à aucune identification à l'exception d'un unique peptide identifié dans une protéine orthologue chez *Fusarium oxysporum*. Dans certains cas, lancer les requêtes dans SwissProt est avantageux

non seulement pour un gain de temps mais aussi pour éviter d'encombrer les résultats avec de nombreuses protéines redondantes (comme c'est souvent le cas avec des banques généralistes comme NCBInr ou UniProtKB). Cependant pour des organismes plus récemment séquencés et moins bien connus, il faut veiller à ce que les protéines étudiées y soient bien présentes pour éviter de manquer des identifications.

Figure 1 : Comparaison des résultats d'une requête Mascot lancée dans UniProtKB/SwissProt et dans la banque protéique NCBInr pour illustrer l'importance du choix de la banque de données dans laquelle les requêtes sont lancées.



1.2. Les banques protéiques accessibles contiennent des erreurs

Les 10 dernières années ont été extrêmement productives en nombre de génomes séquencés et portée par la compétition dans le domaine, l'automatisation et les nouvelles techniques, la communauté scientifique des « génomistes » a largement dépassé ses propres ambitions de séquençage [Liolios, 2006 ; Brent, 2005]. Cependant la séquence des génomes est la base qui devrait permettre d'atteindre le second objectif : établir la liste complète de l'ensemble des gènes et protéines exprimés à partir de cette séquence brute. Or cette tâche est beaucoup plus difficile qu'escomptée et les produits du génome sont bien plus complexes et variés que ce qui avait été anticipé. Même pour le génome humain, pour lequel de grands moyens instrumentaux, financiers et humains ont été investis, les avis divergent toujours sur le nombre de gènes codant pour des

protéines [Pennisi, 2003] et aucun outil actuellement à la disposition des annotateurs de génome ne permet de dresser un catalogue précis et sans failles de l'ensemble des gènes codants [Brent, 2005].

De plus, même si les données de transcriptomique se multiplient et aident à l'augmentation de la qualité des prédictions de gènes chez les eucaryotes, les protéomes prédits disponibles dans les banques de données protéiques accessibles (contenant à 80% des données uniquement prédites *in silico*, sans autre validation) ne sont pas exhaustives et contiennent des erreurs [Bianchetti, 2005].

En ce qui concerne les organismes procaryotes, même s'ils ne contiennent pas les complexes structures introns-exons, l'annotation de leur génome n'en est pas pour autant une tâche facile [Nielsen et coll., 2005]. Pour l'ensemble des 335 génomes bactériens entièrement séquencés et les plus de 900 projets de séquençage en cours [Liolios, 2006], la qualité des données et des prédictions de gènes est très variable ce qui rend les comparaisons très difficiles. D'autant plus que les erreurs introduites dans les banques se propagent car beaucoup de processus sont basés sur des recherches d'homologies avec ce qui est déjà connu et admis comme étant correct [Doerks et coll., 1999]. En effet, des publications révèlent que pour certains génomes procaryotes, jusqu'à 60% des entrées annotées dans GenBank ou RefSeq contiennent des mauvaises prédictions de codons start, et ceci particulièrement pour le cas des génomes riches en GC [Nielsen et coll., 2005].

1.3 Intérêts des recherches directement dans le génome non interprété

Dans ce contexte, une approche alternative aux recherches classiques dans les banques de données protéiques, utilisant les données génomiques brutes, sans annotation au préalable, pour rechercher les données de nanoLC-MS/MS a été développée pendant ce travail de thèse. Rechercher des données MS/MS directement dans le génome non interprété présente des avantages et ouvre de nouvelles possibilités :

- Les programmes de séquençage des génomes assurent un niveau d'erreur de moins d'une base sur 10000 [http://www.genome.gov/10000923].
- Le génome complet contient la totalité des informations et l'ensemble des séquences codantes des protéines sont donc présentes. Les recherches ne seront donc plus tributaires de la bonne prédiction des protéines ni de leur présence dans les banques de données protéiques. Les recherches dans le génome permettent ainsi l'identification de nouveaux gènes codants qui n'auraient pas été prédits par les algorithmes de prédiction (parce qu'elles ne présentent pas les caractéristiques de régions codantes traditionnelles).
- Les recherches dans le génome non interprété peuvent se faire sans avoir à attendre la fin du processus d'annotation du génome qui peut parfois prendre beaucoup de temps. Le démarrage des études protéomiques n'est donc plus conditionné par la mise en accessibilité du protéome prédit pour l'organisme étudié.
- Identifier les peptides exprimés directement dans le génome permet non seulement l'identification de la protéine mais aussi sa localisation sur le génome.
- Identifier le gène à l'origine de la protéine exprimée, indépendamment du protéome prédit, constitue une preuve expérimentale de l'expression du gène, cette approche est donc un outil de validation expérimentale de l'expression d'un gène.
- La comparaison entre les résultats d'identification de protéines obtenus par cette méthode et le protéome prédit permet de mettre en évidence d'éventuelles erreurs de prédiction de codon

start, de frameshift, … cette approche, étant indépendante des prédictions *in silico*, elle constitue donc un outil d'aide à l'annotation des génomes.

Une approche de recherche dans les génomes a été développée au laboratoire et testée pour un certain nombre de génomes d'organismes couvrant une bonne partie du règne vivant, de génomes procaryotes à des génomes d'eucaryotes supérieurs :

- des génomes bactériens comme *Herminiimonas arsenicoxydans* ou *Mycobacterium smegmatis*
- le génome de *Trypanosoma brucei*,
- le génome de *Toxoplasma gondii*,
- le génome d'*Oryza sativa*,
- et même le génome d'*Homo sapiens*.

2 La stratégie de recherche dans les génomes

2.1 La recherche des données de protéomique dans les génomes n'est que très peu décrite dans la littérature

En 1995, l'équipe de Yates a démontré la possibilité de rechercher des séquences nucléotidiques avec des données de spectrométrie de masse en tandem [Yates et coll., 1995]. Depuis cette date, seulement un petit nombre de travaux ont été publiés utilisant des recherches de données nanoLC-MS/MS dans des séquences génomiques :

- deux articles techniques apportant les preuves de faisabilité des recherches de données nanoLC-MS/MS dans des banques d'ESTs et le génome humain d'une part [Choudhary et coll., 2001] et dans les génomes humain et d'*Arabidopsis thaliana* d'autre part [Kuster et coll., 2001]. Les travaux de Choudhary et coll. ont été réalisés par les concepteurs du programme Mascot et cet article discute les possibilités techniques de Mascot pour la gestion de grands génomes, les performances en temps de recherche et compare les résultats obtenus dans la banque d'ESTs humains à ceux obtenus dans le génome humain et dans la banque protéique MSDB. Le second travail utilise le programme PepSea et l'approche par Peptide Sequence Tag pour effectuer des recherches dans le génome humain et dans celui d'*Arabidopsis thaliana* [Kuster et coll., 2001].
- Un projet d'établissement de génome et de protéome complet de la bactérie *Mycoplasma mobile* utilise également des recherches dans le génome avec le programme SEQUEST pour établir ce qu'ils appellent une carte protéogénomique de cette bactérie à petit génome (≈777 kbps) [Jaffe et coll., 2004 ; Jaffe et coll., 2004].
- Enfin, un travail sur le ciliome de *Tetrahymena thermophila*, organisme unicellulaire eucaryote modèle caractérisé par un dimorphisme nucléaire [Smith et coll., 2005].L'approche décrite dans ce papier utilise Mascot pour réaliser les recherches avec des données nanoLC-MS/MS dans une banque de données génomiques, les peptides identifiés sont localisés sur le génome, les séquences encadrant les peptides identifiés (150 acides aminés de part et d'autre) sont extraites, combinées et soumises à des requêtes BLAST.

Chacun de ces travaux souligne des avantages des recherches de données nanoLC-MS/MS dans des génomes non interprétés et dans la continuité de ces travaux, nous avons mis une stratégie en place au laboratoire permettant d'effectuer des recherches dans les génomes, quelque soit l'organisme et la taille de son génome. Cette stratégie a été automatisée grâce à des programmes réalisés au laboratoire et a été appliquée à deux contextes particuliers :

- comme approche complémentaire d'aide à la prédiction de gènes et à l'annotation des génomes procaryotes.
- comme approche permettant la discrimination des gènes paralogues spécifiquement exprimés au sein de familles multigéniques chez *Oryza sativa*.

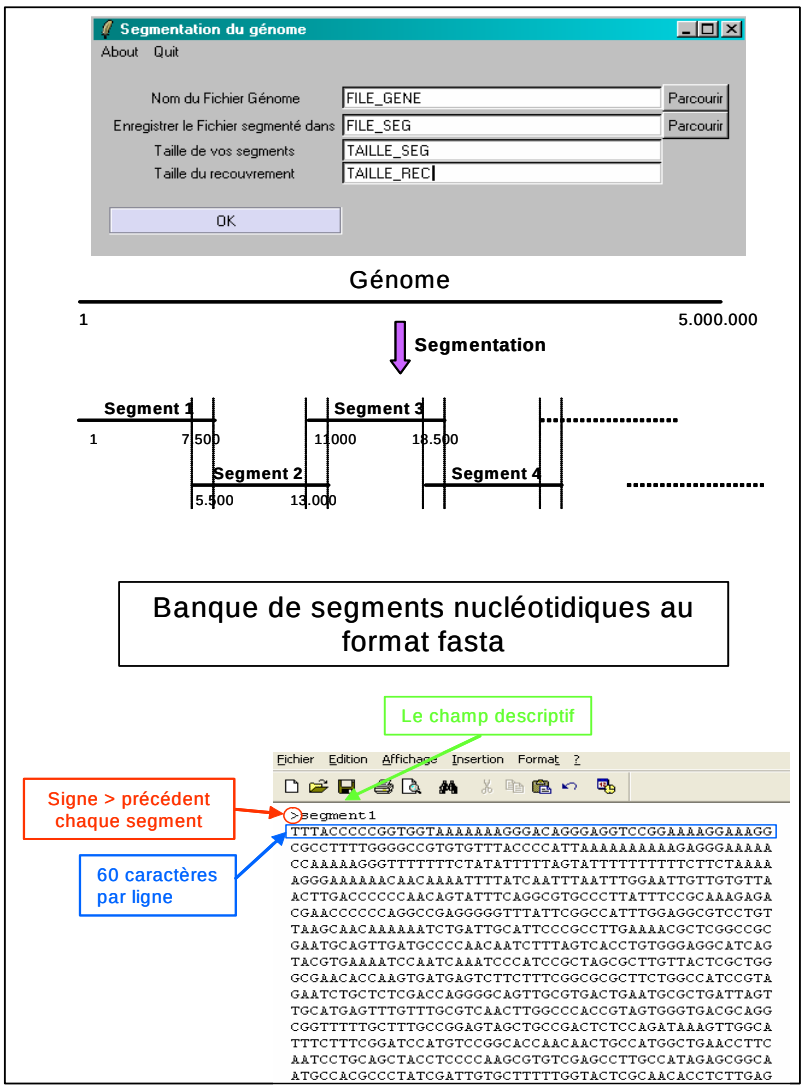
2.2 Description de la stratégie mise en place au laboratoire

2.1.1 Le découpage du génome

Un grand nombre de génomes sont séquencés mais selon la stratégie de séquençage utilisée et l'état d'avancement du projet de séquençage, les données de séquence du génome rendues disponibles peuvent l'être sous différentes formes. Dans certains cas, une séquence consensus a été établie et le génome peut donc être disponible sous la forme d'un unique fichier texte. Dans d'autres cas, les assemblages ne sont que partiels et la séquence du génome est disponible soit sous forme de BACs (Bacterial Artificial Chromosome), de contigs, soit sous la forme de chromosomes partiellement ou complètement assemblés en une séquence consensus, … Il faut donc trouver un moyen pour gérer les différents formats dans lesquels les données sont disponibles.

L'interface graphique de Mascot est limitée en nombre d'acides aminés affichables (limitation à 30000 acides nucléiques donc 10000 acides aminés), des séquences très longues comme par exemple des chromosomes entiers ne sont pas visualisables dans les fenêtres classiques de « Protein view ». Ainsi, pour contourner ce problème, un programme de découpage des génomes en language TcL (Tool Command Language) "segmentation du génome" a été réalisé au laboratoire afin de découper les données génomiques en des segments courts et pouvoir visualiser normalement les résultats. Ce programme génère en sortie un fichier au format Fasta (donc compatible avec l'importation dans Mascot) des séquences nucléotidiques découpées en segments de longueur choisie avec des recouvrements de séquence de longueur choisie également (Figure 2). Ce programme fonctionne à la fois pour des génomes qui se présentent sous la forme d'une entrée unique et pour les génomes qui sont disponibles sous formes de grands fragments ou de chromosomes. Dans ces cas l'intitulé de l'entrée fasta est conservé et les numérotations des segments sont ajoutées à la suite (Figure 2).

Figure 2 : Programme permettant de découper le génome en segments de longueur choisie avec des recouvrements de séquence de longueur choisie également. Ce programme génère ensuite une banque de segments au format .fasta qui pourra être importée dans Mascot.



2.1.2 *Importation dans Mascot et traduction dans les six cadres de lecture*

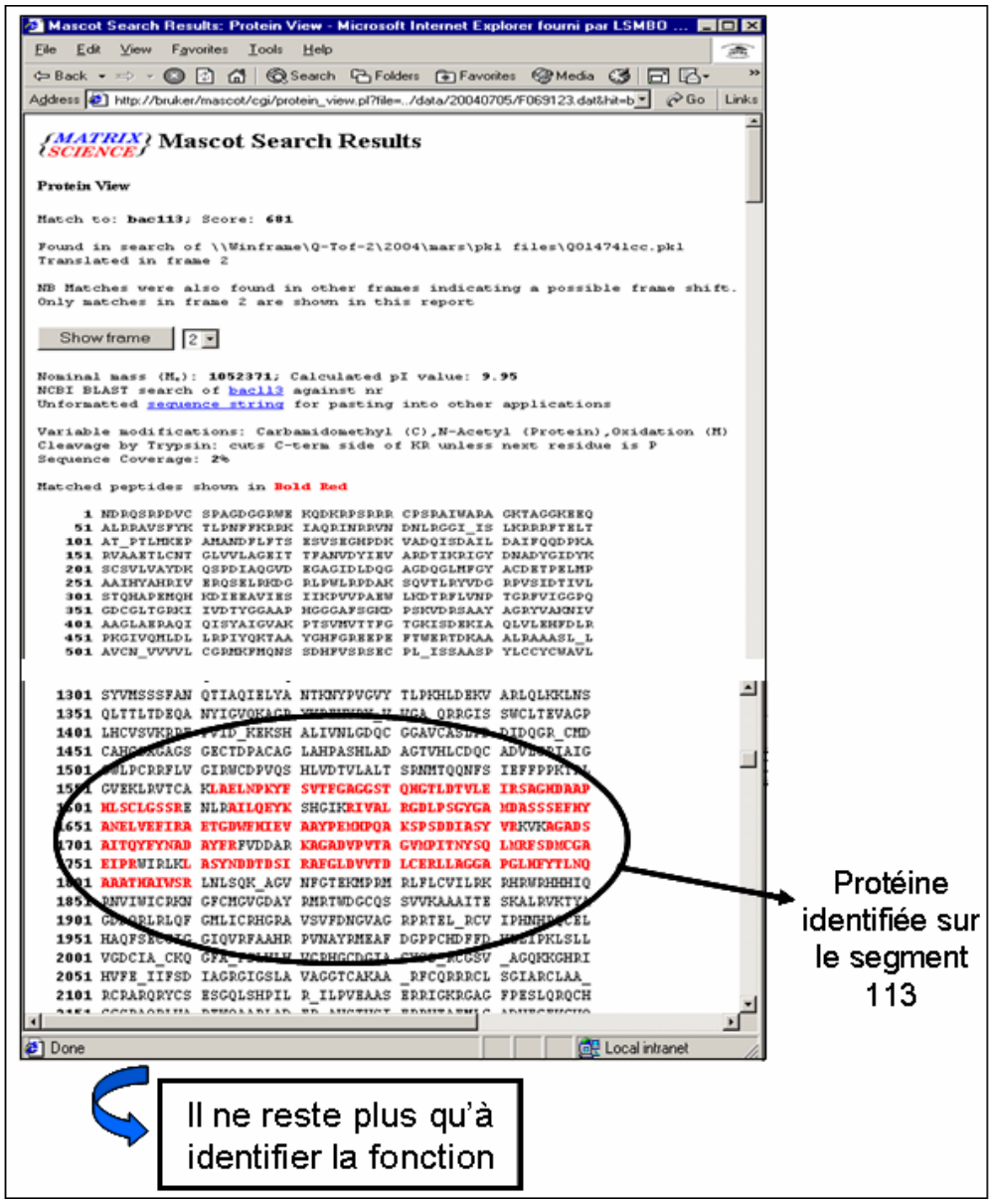
Une fois le génome formaté, la banque des segments nucléotidiques peut être facilement importée dans le programme Mascot, à condition de disposer d'une licence Mascot et que le programme soit installé en interne. Il suffit de préciser que les données importées sont des données nucléotidiques et lors du chargement en mémoire de la banque dans Mascot, la traduction dans les 6 cadres de lecture possible est alors réalisée automatiquement.

2.1.3 *Requête Mascot et identification de la région codante sur le génome*

Les requêtes Mascot sont lancées dans cette banque génomique avec les paramétrages classiques pour des données de nanoLC-MS/MS à savoir : la trypsine est spécifiée comme enzyme de coupure, un manquement de coupure est toléré, 0,25 Da d'erreur sont tolérés en mode MS et MS/MS et la carbamidomethylation des cystéine et l'oxidation des méthionines sont définies comme modifications variables.

En figure 4 est présenté un exemple de résultat Mascot de recherche dans le génome.

Figure 4 : Exemple de résultat Mascot obtenu par recherche dans un génome bactérien non interprété, fragmenté en segments de 7500 acides nucléiques (soit 2500 acides aminés). La région codante est identifiée et localisée sur le génome, il ne reste plus qu'à déterminer la fonction de la région identifiée.



L'ensemble des peptides identifiés est regroupé sur un segment et permet ainsi la localisation de la région codante sur le génome. Il ne reste ensuite plus qu'à déterminer la fonction de la protéine identifiée. Pour cela, deux possibilités sont ouvertes : soit le génome a été annoté et l'annotation de la fonction du gène identifiée et disponible, soit les peptides identifiés sont soumis au MS-BLAST afin d'identifier la fonction de la protéine par homologie de séquence avec des protéines connues. La deuxième solution a été optée dans notre approche.

2.1.4 Identification de la fonction de la protéine par MS-BLAST

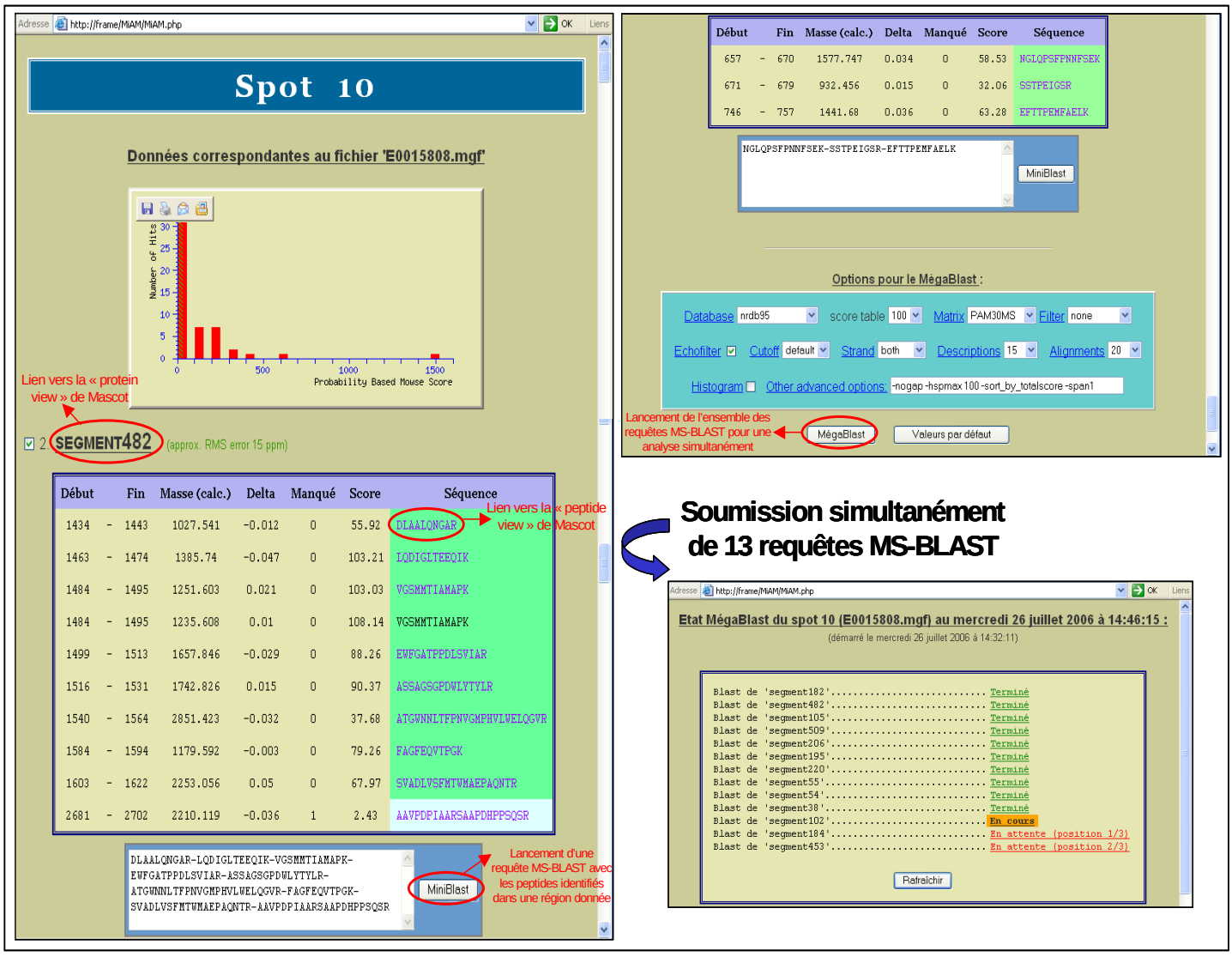
En effet, souhaitant travailler indépendamment des données de prédiction de gènes et d'annotation de génomes, le programme MS-BLAST est utilisé pour déterminer la fonction des régions codantes identifiées. Le fait de travailler indépendamment des prédictions et annotations du génome permet de disposer d'une stratégie entièrement alternative pour laquelle aucun travail d'annotation du génome au préalable n'est

nécessaire. Les identifications des protéines peuvent donc être réalisées sans avoir à attendre la fin du processus de prédiction et d'annotation des gènes. Par ailleurs, les résultats obtenus de façon indépendante par cette approche pourront par la suite être comparés aux résultats de prédiction et ainsi constituer une voie de confirmation et de validation des prédictions.

Seuls les peptides identifiés durant la recherche Mascot sont extraits pour être soumis au MS-BLAST, contrairement aux travaux précédemment publiés [Smith et coll., 2005]. Le choix de ne pas élargir la séquence soumise au BLAST avec les régions inter-peptides ou avec des régions avoisinantes des peptides identifiés permet de ne soumettre et de ne se baser pour l'identification que sur des séquences peptidiques dont la séquence a été déterminée et pourra être vérifiée grâce aux données MS/MS.

Un programme permettant d'extraire les peptides identifiés uniquement et d'automatiser le lancement des requêtes MS-BLAST a également été réalisé au laboratoire. Ce programme, appelé MiAM, est présenté en figure 5.

Figure 5 : Quelques aperçus du programme MiAM développé au laboratoire permettant de visualiser les résultats avec des liens vers les différentes fenêtres de Mascot par simple clic sur les segments identifiés (lien vers la « Protein View » de Mascot) ou sur les peptides identifiés (lien vers la « Peptide View » de Mascot). Ensuite les BLAST peuvent être lancés individuellement (MiniBLAST) ou simultanément (MégaBLAST) lorsque plusieurs protéines sont identifiées dans un spot. Les soumissions au MS-BLAST se font de façon transparente et les résultats sont consultables ultérieurement. Enfin, MiAM génère un rapport des résultats validés par l'expérimentateur qui peut être exporté vers le logiciel Excel.

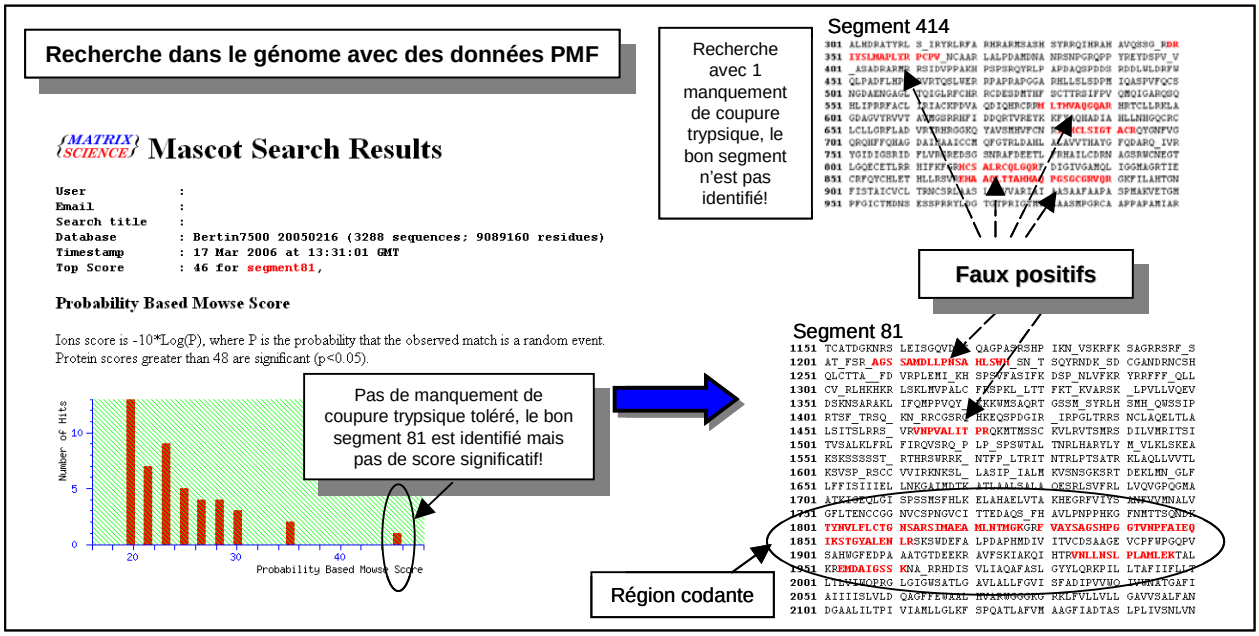


2.3 La spécificité apportée par les informations de séquence contenues dans les données MS/MS est indispensable

L'évaluation de la possibilité de lancer des requêtes dans le génome non interprété avec des données d'empreinte peptidique massique a été réalisée. Pour cela, une partie des extraits peptidiques d'une série de spots a été analysée dans un premier temps par MALDI-TOF-MS avant d'être injectés en nanoLC-MS/MS et les résultats d'identification ont été comparés. Comme illustré en figure 6, ces requêtes n'ont pas permis d'aboutir à des résultats positifs : les données génomiques traduites dans les 6 cadres de lecture donnent lieu à tellement de masses de peptides trypsiques théoriques possibles, qu'un nombre significatif de peptides vont être identifiés aléatoirement uniquement de part leur masse. Ce problème est accentué lorsque les requêtes Mascot sont lancées en tolérant 1 manquement de coupure trypsique, le nombre de faux positifs identifiés est très grand et, dans la plupart des cas, le bon segment n'est même pas identifié. Par contre, en n'autorisant pas de manquement de coupure trypsique, le bon segment est souvent identifié en première position mais, comme illustré en figure 6, des faux positifs entourent la séquence codante identifiée. Il est donc difficilement possible d'accorder une confiance à ces identifications par empreinte peptidique massique et la spécificité apportée par les informations MS/MS est indispensable pour le succès des recherches dans les génomes.

L'impossibilité d'identifier des protéines dans des données génomiques par PMF avait déjà été soulevée [Küster et coll., 2001].

Figure 6 : Illustration d'une recherche Mascot dans le génome avec des données d'empreinte peptidique massique.



3 La protéomique devient alors un outil d'aide à l'annotation des génomes

Rechercher ainsi directement dans le génome permet à l'analyse protéomique par spectrométrie de masse en tandem de constituer une véritable aide et un complément direct au travail d'annotation des génomes et à la compréhension globale de la structure des génomes. L'ensemble de la stratégie d'identification développée étant totalement indépendante des informations de prédiction de gènes et d'annotation du génome, ces données peuvent ensuite être comparées. Les régions codantes identifiées par spectrométrie de masse sur le génome sont des régions codantes « expérimentalement validées ».

L'utilité des données de protéomique et particulièrement des informations de séquence déterminées par MS/MS pour enrichir et compléter l'annotation des génomes est évidente. L'ampleur du projet PeptideAtlas dont le but final est d'annoter entièrement et de valider les génomes eucaryotes avec des données expérimentales obtenues par l'analyse des protéines en est un bon exemple [Desiere et coll., 2005 ; Desiere et coll., 2006]. Au départ, seul le génome humain était visé mais le projet s'est déjà élargi à *Drosophila melanogaster*, à la levure *Saccharomyces cerevisiae* et des projets d'élargissement à la souris et *Arabidopsis thaliana* sont en cours.

Enfin, les recherches dans les génomes permettront peut-être de mettre en évidence de nouvelles régions codantes dans des régions du génome qui étaient prédites comme non-codantes car elles ne présentaient pas les caractéristiques classiques de régions codantes. En effet, avec les approches de recherche protéomique classique en banques de données protéiques, l'exigence de la présence au préalable dans les banques des peptides pouvant être identifiés conduit à un dilemme puisqu'elle empêche la découverte de nouvelles informations sur les génomes [Baginsky et coll., 2006].

4 Conclusions

Cette stratégie de recherche a été appliquée à une série d'organismes étudiés au laboratoire, à la fois à des génomes procaryotes et eucaryotes. Deux applications particulières pour lesquelles cette stratégie a été utilisée sont développées dans les 2 chapitres suivants :

- Dans le second chapitre, l'utilisation de cette stratégie est décrite pour l'étude de protéomes bactériens dont les génomes ont été séquencés mais peu annotés et pour lesquels la prédiction des régions codantes pose des difficultés. Le gain en temps et en informations que la recherche dans les génomes a permis de réaliser par rapport à la stratégie par séquençage *de novo*/MS-BLAST est démontré ainsi que les précieuses informations que cette stratégie a permis d'apporter pour compléter l'annotation de ces génomes bactériens.
- Dans le troisième chapitre, l'application de cette stratégie à un génome eucaryote complexe, celui d'*Oryza sativa*, est illustrée. L'objectif de ce chapitre est de démontrer que les recherches dans le génome permettent la discrimination de gènes paralogues au sein de grandes familles multigéniques afin de connaître la séquence du gène spécifiquement exprimé.

CHAPITRE II

Les recherches dans les génomes : un outil d'aide à l'annotation des génomes bactériens

Cette étude a été réalisée en collaboration avec l'équipe de Génétique Moléculaire, Génomique, Microbiologie (CNRS/ULP-UMR 7156) du Professeur Philippe Bertin à l'Institut Botanique de l'Université Louis Pasteur de Strasbourg. Ce projet s'inscrit dans la continuité du projet d'étude de la réponse à l'arsenic par approche différentielle par gel 2D chez *Caenibacter arsenoxydans* (Chapitre II Partie III des résultats).

1 Contexte de cette étude

Le contexte biologique de cette étude a été décrit en détail dans le chapitre II (Partie III des résultats). Au moment du démarrage du projet, le génome de la bactérie *Caenibacter arsenoxydans* n'était pas séquencé et l'utilisation d'une approche par séquençage *de novo* / MS-BLAST a été indispensable pour identifier les protéines différentiellement exprimées en présence d'arsenic dans le milieu de culture bactérien [Carapito et coll., 2006]. Le séquençage du génome étant en parallèle en cours au Génoscope, la disponibilité de la séquence du génome a permis un élargissement du projet.

La bactérie présentant des caractéristiques tout à fait intéressantes et originales, un projet d'une étude protéomique plus globale a été envisagé (carte 2D de référence des protéines cytoplasmiques et protéome membranaire). Cependant, l'utilisation d'une stratégie séquençage *de novo*/MS-BLAST représente un investissement en temps d'interprétation considérable lorsqu'il s'agit de traiter un très grand nombre de spots. Dans ce projet, c'est la connaissance de la séquence du génome (même si celui-ci n'était pas encore annoté) qui a permis le développement d'une stratégie de recherche directement dans le génome non interprété et qui a permis d'envisager un projet de cette ampleur. L'utilisation du moteur de recherche Mascot pour effectuer les recherches directement dans le génome non interprété a permis un gain en temps mais aussi un gain en qualité considérables par rapport à l'approche par séquençage *de novo*. Pour ces recherches, des programmes développés au laboratoire ont permis d'automatiser l'ensemble des étapes des

recherches et une visualisation très informative des résultats de protéomique obtenus, alignés sur la séquence du génome.

Enfin, l'établissement de la carte protéomique ayant été réalisée en même temps que l'annotation du génome, les résultats obtenus par l'analyse protéomique ont pu être alignés avec ceux obtenus par prédictions et annotations *in silico*. Les apports de l'analyse protéomique comme outil complémentaire d'aide et de correction à l'annotation des génomes ont pu être largement mis en évidence.

A noter également qu'entre temps, la bactérie a changé de nom pour s'appeler maintenant *Herminiimonas arsenicoxydans*.

2 Les recherches dans le génome d'*Herminiimonas arsenicoxydans*

2.1 Les échantillons et les analyses nanoLC-MS/MS

Les gels 2D ont été préparés suivant le même protocole que les gels 2D du chapitre II de la partie III des résultats. Pour cette étude, les spots ont ensuite été découpés systématiquement. En ce qui concerne les protéines membranaires, les extraits membranaires ont été séparés par gel 1D et les bandes de gel 1D ont été découpées systématiquement en bandes de 2mm de large sur toute la longueur du gel.

Une partie des spots de gel 2D ont été analysés par MALDI-TOF-MS mais la spécificité de la MS/MS est nécessaire pour le succès de l'identification par des recherches dans le génome (Chapitre précédent). L'ensemble des spots a donc été analysé par nanoLC-MS/MS sur le couplage nanoHPLC (Agilent Series 1100)-trappe ionique HCT Plus (Bruker Daltonics).

2.2 Stratégie de recherche dans le génome

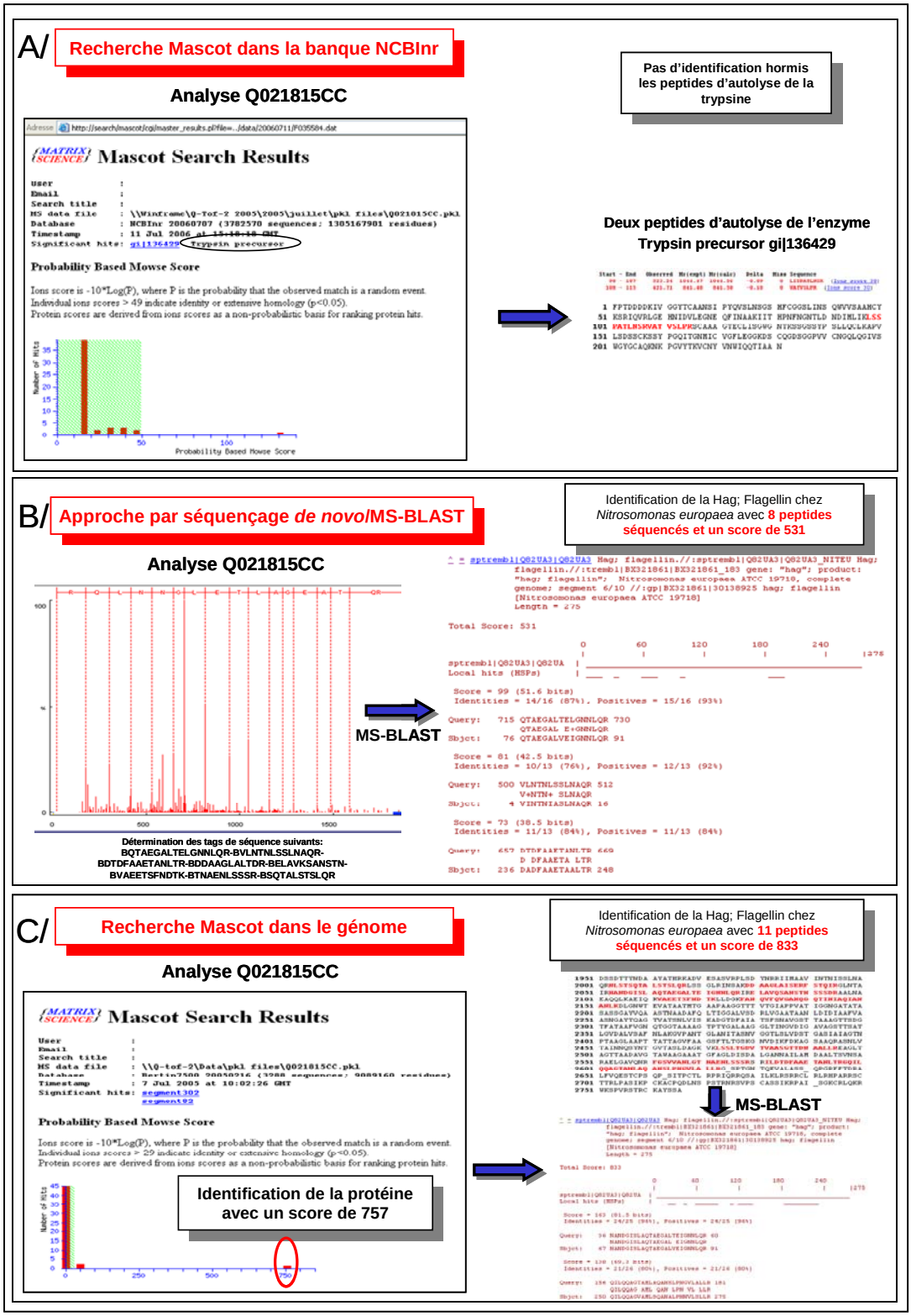
La séquence du génome d'*Herminiimonas arsenicoxydans* a été fournie par le Génoscope sous la forme d'un fichier unique contenant une séquence consensus de l'ensemble du génome. La taille du génome est de 3,4 Mbps et ce génome a été découpé en segments de longueur choisie grâce au programme "segmentation du génome" qui a été développé au laboratoire. La taille des segments a été optimisée à 7500 acides nucléiques soit 2500 acides aminés avec des recouvrements de séquence de 2000 acides nucléiques de part et d'autre. La taille des segments ainsi que des recouvrements a été optimisée de sorte à éviter les identifications de protéines à la jonction de deux segments et pour pouvoir visualiser les résultats par les « protein view » de Mascot. Cette banque de segments ainsi générée est importée dans Mascot où elle est traduite dans les 6 cadres de lecture. Les requêtes avec les données de nanoLC-MS/MS sont ensuite lancées dans cette banque comme décrit dans le chapitre précédent.

3 Comparaison des résultats obtenus par recherche dans le génome et par séquençage *de novo*/MS-BLAST

Dans le but de valider l'approche de recherche dans les génomes, les premières données ayant permis l'identification des protéines différentiellement exprimées chez *H. arsenicoxydans* en présence d'arsenic par séquençage *de novo* (voir Chapitre II de la partie III des résultats), ont été recherchées dans la banque génomique générée comme décrit ci-dessus.

Afin d'illustrer la comparaison des résultats obtenus, la figure (Figure 1, Partie III, Chapitre III.3.) présentant les résultats de l'approche par séquençage *de novo*/MS-BLAST pour le fichier de données nanoLC-MS/MS Q021815CC.pkl a été reprise et les résultats obtenus pour ces mêmes données par la recherche dans le génome y ont été ajoutés (Figure 1)

Figure 1: Comparaison des résultats obtenus pour les données de l'analyse Q021815CC.pkl par les trois approches d'interprétation: A/ Résultats obtenus par recherche Mascot classique dans la banque de données protéiques NCBI. B/ Résultats obtenus par l'approche séquençage *de novo* des spectres MS/MS suivi de requête MS-BLAST. C/ Résultats obtenus par recherche Mascot dans le génome non interprété suivi par requête MS-BLAST pour l'identification de la fonction de la région codante identifiée.



Par l'approche séquençage *de novo*/MS-BLAST, 8 séquences de peptides déterminées par séquençage *de novo* ont permis l'identification par homologie de séquence de la protéine Hag :Flagellin chez *Nitrosomonas europaea* avec un score MS-BLAST de 531. La recherche dans le génome a permis d'identifier la région codante de la protéine sur le segment 302, avec 11 peptides identifiés. La soumission de ces 11 peptides au MS-BLAST a permis d'identifier la Hag :Flagellin chez *Nitrosomonas europaea* avec score MS-BLAST de 833. La recherche dans le génome a donc non seulement permis d'identifier la bonne fonction protéique mais elle a aussi permis de localiser la région codante pour la protéine exprimée sur le génome et enfin, l'identification a pu être confirmée avec un plus grand nombre de peptides.

De la même manière, le tableau 1 présente la comparaison de quelques identifications précédemment obtenus lors de l'étude protéomique différentielle (Chapitre II Partie III des résultats) à ceux obtenus par des recherches avec les mêmes données dans le génome. Comme l'illustre ce tableau, les protéines précédemment identifiées ont toutes été identifiées dans le génome, avec l'ensemble des peptides séquencés par *de novo* et même un nombre significatif de peptides supplémentaires.

Tableau 1 : Comparaison des peptides identifiés par l'approche séquençage *de novo*/MS-BLAST et des peptides identifiés par les recherches dans le génome. Les peptides en gras sont ceux identifiés par les deux approches.

Identification de la fonction protéique grâce au MS-BLAST	Sequences déterminées par séquençage de novo ayant permis l'identification inter-espèces	Score MS-BLAST	Peptides identifiés lors de la recherche dans le génome	Score MS-BLAST
Adenosylhomocysteinase	***ESLVDGIKR*** -***ATDVMIAGK***- ***FDNLYGCGR*** -***IAETEMPGLMAIR*** - ***VAVIAGYGDVGK*** -***TEETTTGVHR***	**443**	***IAETEMPGLMAIR*** -EEFAAAQPLK-ITGSIHMTIQTAVLIQTLEALGAK-GETLDDYWEYTHR-IFEWPNDDK-GAPVYSNMILDDGGDATLLLHLGTR-DASVLNNPGSEEEICLFNSIK-HLAVDATWYSK-LPQILGV***TEETTTGVHR*** LAFPAINVNDSVTK-SK***FDNLYGCR*** -***ESLVDGIKR*** -ATDVMIAGK- ***VAVIAGYGDVGK*** -ALSAQVWVTEIDPICALQAAMEGYR-DQAIVCNIGHFDNEIEVAALK-IILLAEGR-LVNLGCGTGHPSYVMSSSFANQTIAQIELYANTK-NYPVGVYTLPK-LNSQLTTLTDEQANYIGVQK-AGPYKPEHYR	**2090**
Putative PMBA protein	***MGQGVNYVTGDYSR- IQYPVEEITIAGN-TVEAAYNIAR- GASGFWVE-NSEGASVYA- KSTFLLDVTGK-LAAPETIGR***	**472**	GGVETIEQNKDK-GMGVTVYLGEER-ALKD***TVEAAYNIAR*** -FTAEDDCAGLPDEDMLER-DLQLCYPWLISTEEAVELAK-GEAAAFAVDKR-IT***NSEGASVYA*** QQSHFVSANSR-GFIGGYPYSR-HTISVAPIAGK-***LAAPETIGR*** -KCPVLFEAPLAAGLLGAFVQAVSGGALYR- ***KSTFLLDSLGK*** -EVVKDGVVQGYFLSTYSAR-KLDTGLLVTEL***MGQGVNYVTGDYSR*** -***GASGFWVE*** K-GV***IQYPVEEITIAGN*** MK-DIFMQVVAIGADTLMR	**1392**
Imidazole glycerol phosphate synthase	***AVLAASLFHYGQHTVQEAK- GAGEILLTSMDR-IIPCLDVTAGR- FMSEQGIAMR-IVVAIDAK- GVNFLELR-TGLDAI***	**481**	***IIPCLDVTAGR*** -***GVNFLELR*** -DAGDPVEIAR-RYDDQGADELTFLDITATSDGR-YGSQC***IVVAIDAK*** -A***TGLDAI*** EWAK-KMESL***GAGEILLTSMDR*** -VGFDLDLTR-SVSDAISIPVIASGGVGGLQDLVDGVK-AD***AVLAASIFHYGQHTVQEAK*** -***FMSEQGIAMR***	**967**
Elongation factor Tu	***LLDQGQAGDNVGVLLR- FLLPVEDVFSISGR- TQVTTCTGVEMFR-ALDSYIPTPER- FPGDDLPIIK-LTAAIATVLS- TTLTAAIA-IIVFLNK-EHILLAR***	**645**	TT***LTAAIATVLS*** K-HYAHVDCPGHADYVK-***EHILLAR*** -QVGVPY***IIVFLNK*** -YE***FPGDDLPIIK*** -AVDGA***FLLPVEDVFSISGR*** -VGESLEIVGIR-D***TQVTTCTGVEMFR*** -***LLDQGQAGDNVGVLLR*** -HFTGEIYVLSKDEGGR-TTDVTGSIELPKDK-TVGAGVVAK	**1053**

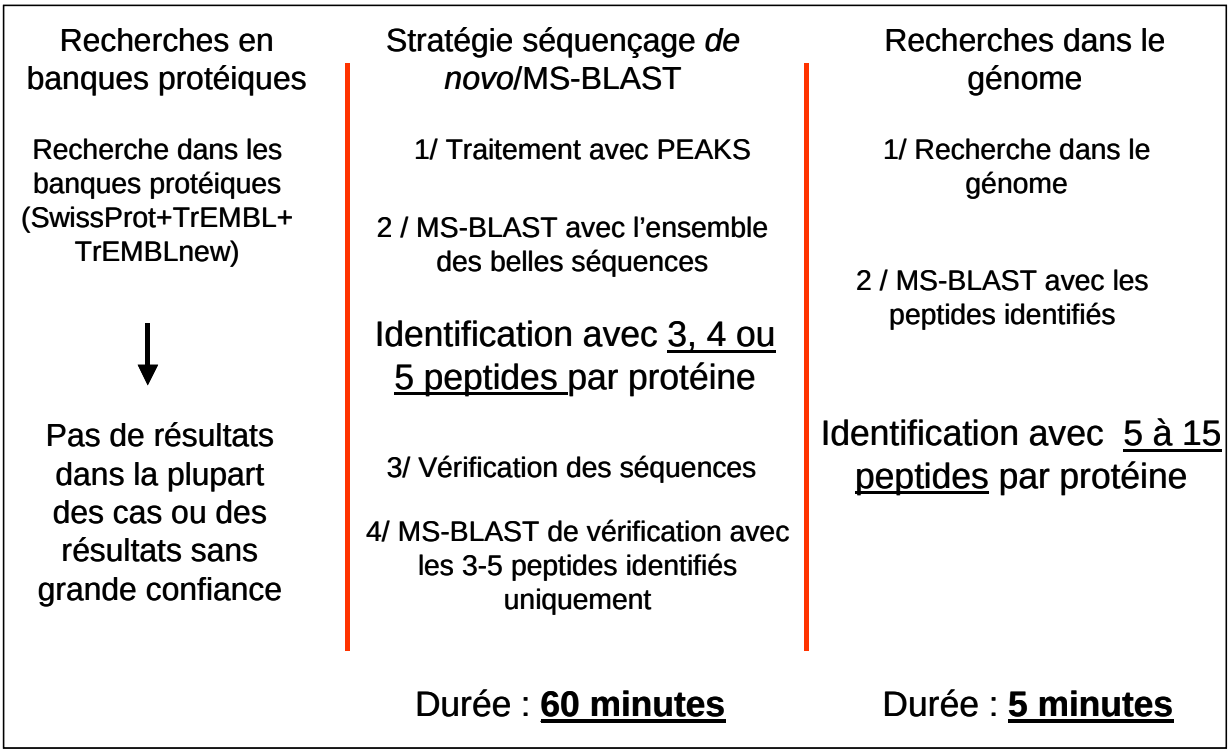
Les **peptides supplémentaires** identifiés dans les **recherches dans le génome** s'expliquent par plusieurs raisons :

- les peptides identifiés dans les recherches dans le génome ne doivent pas nécessairement présenter des spectres MS/MS de qualité suffisante pour pouvoir être interprétés par séquençage *de novo*. Les critères de qualité significatifs pour Mascot sont suffisants. Un certain nombre de peptides identifiés pourront donc être retenus alors que leur interprétation *de novo* aurait posé quelques difficultés.
- dans l'approche d'identification des protéines par séquençage *de novo*, un maximum de peptides sont séquencés mais seuls les peptides présentant des homologies de séquence avec

les peptides des protéines présentes dans la banque criblée par le MS-BLAST peuvent être retenus pour l'identification. En effet, cette approche est tolérante aux erreurs mais elle ne se base néanmoins que sur les protéines déjà connues pour identifier les peptides homologues. En revanche, dans les recherches dans le génome, même les peptides qui ne présentent aucune homologie avec des peptides présents dans les banques et qui sont spécifiques à l'organisme étudié peuvent être identifiés comme appartenant à la région codante vu que le génome contient toute l'information.

Enfin, en figure 2 sont schématisées les différentes étapes des trois stratégies de recherche et une estimation des temps d'interprétation de ces différentes approches. Cette figure illustre bien les avantages des recherches dans le génome à plusieurs niveaux, elles sont plus rapides et permettent d'obtenir un maximum d'informations.

Figure 2 : Comparaison des trois approches de recherche avec des données de nanoLC-MS/MS dans le cas d'un organisme non représenté dans les banques de données protéiques. La méthode classique de recherche dans les banques protéiques ne permet souvent d'aboutir à aucune identification. La stratégie par séquençage de novo/MS-BLAST nécessite un temps d'interprétation long (vérification du séquençage par l'expérimentateur) et une haute qualité des spectres MS/MS. Enfin, la recherche dans le génome permet d'obtenir un maximum d'informations en un temps d'interprétation plus court.



4 Etablissement des cartes protéomiques de référence pour *Herminiimonas arsenicoxydans*

4.1 La carte cytoplasmique et le logiciel InPACT

Une fois l'approche de recherche dans le génome validée, les analyses pour l'établissement de la carte protéomique d'*Herminiimonas arsenicoxydans* ont pu démarrer. Un total de 437 spots de gel 2D ont été découpés et analysés par nanoLC-MS/MS avec le couplage nanoHPLC Agilent Series 1100 / Trappe ionique HCT Plus (Bruker Daltonics). Ces analyses ont abouti à l'identification de 2398 peptides uniques ayant permis l'identification de 1870 protéines avec des redondances, représentant un total de 367 protéines distinctes.

Un logiciel de visualisation de ces résultats protéomiques a été développé au laboratoire, le logiciel InPACT pour « InterACtive Proteomic Tool » (Figure 3). Ce logiciel compile toutes les données d'identification des protéines et permet de visualiser ces données en se déplaçant sur l'image du gel. Par clic sur le spot d'intérêt, une fenêtre présentant les résultats d'identification de la (ou des) protéine(s) détectée(s) dans le spot s'ouvre (Figure 4).

Figure 3 : Le logiciel InPACT développé au laboratoire pour la compilation des résultats protéomiques obtenus pour la carte 2D de *H. arsenicoxydans*. La fonction de zoom permet de faire des agrandis sur les zones d'intérêt et de se déplacer sur le gel.

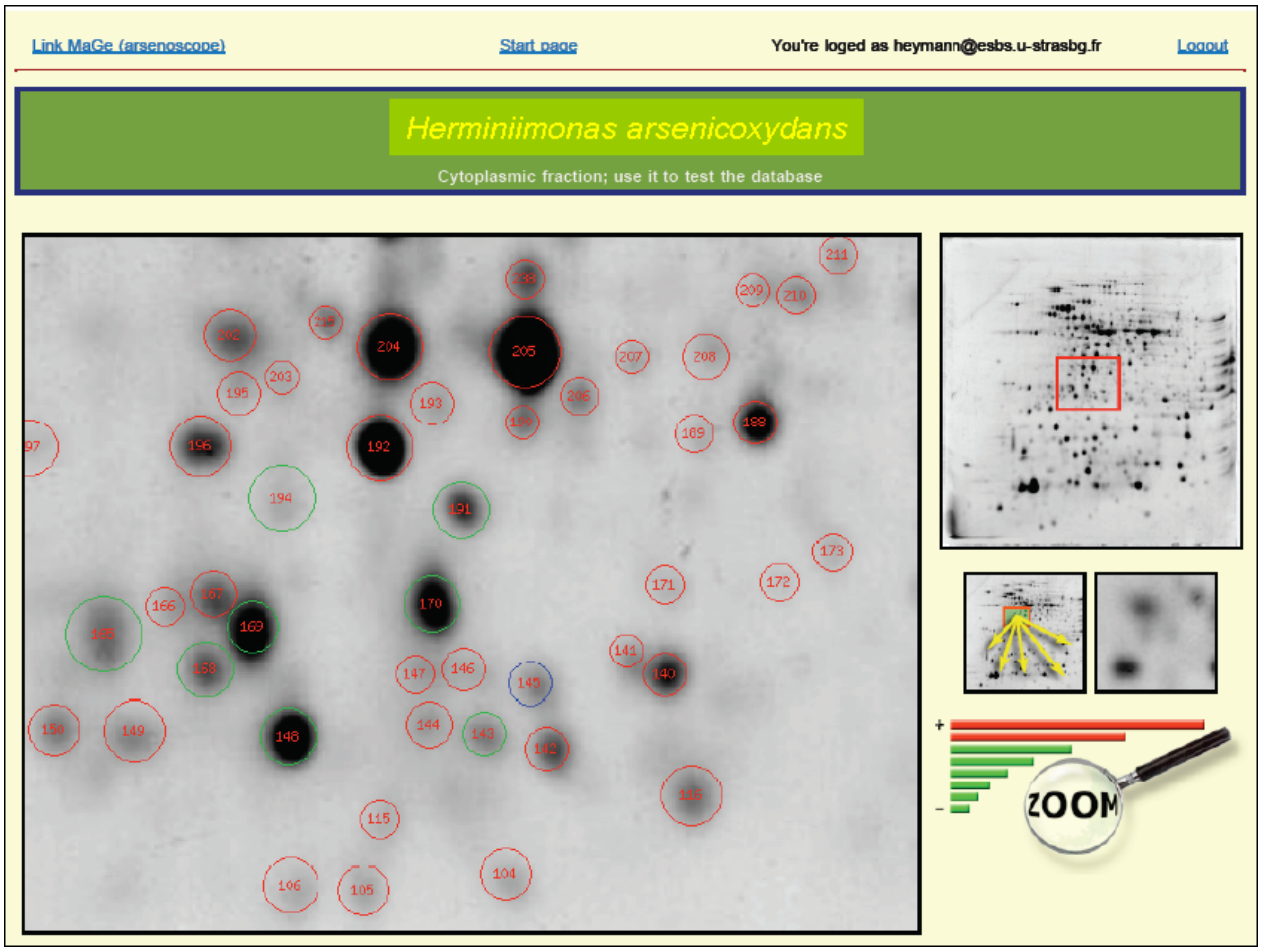
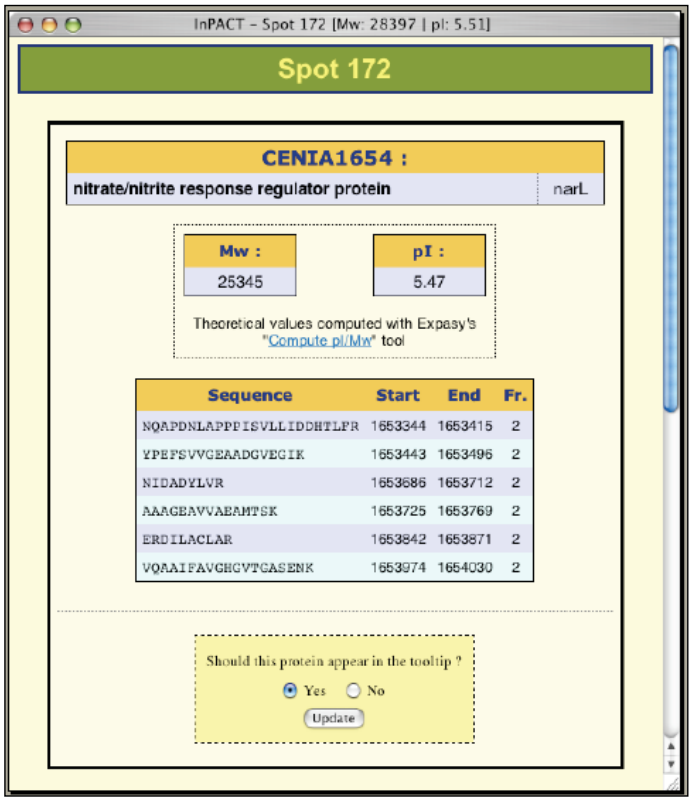


Figure 4 : Visualisation des résultats d'identification des protéines dans InPACT après avoir cliqué sue un spot du gel 2D.



4.2 Le protéome membranaire séparé par gel d'électrophorèse 1D : les recherches dans le génome sur des mélanges complexes.

En parallèle de la carte 2D de référence des protéines cytoplasmiques, des analyses du protéome membranaire ont également été réalisées. Pour cela, les extraits de protéines membranaires séparés par gel 1D découpé systématiquement en bandes d'environ 2mm de large, ont été analysés de la même manière que les spots de gel 2D. Ces bandes de gel 1D découpées sont très riches en protéines et les extraits peptidiques après digestion sont donc très complexes.

Les gradients de nanoLC-MS/MS ont été adaptés et rallongés pour permettre la sélection de plus d'ions pour la fragmentation mais malgré cela, les mélanges étant tellement complexes, moins de peptides trypsiques sont sélectionnés pour la MS/MS par protéine. Il faut donc s'attendre à obtenir des pourcentages de recouvrement plus faibles.

Néanmoins, les recherches dans le génome ont permis d'aboutir à l'identification sans ambiguïté d'une douzaine de protéines en moyenne avec un minimum de 2 peptides (avec un score MS/MS supérieur à 40) par bande de gel 1D. Les recherches dans les génomes peuvent donc être utilisées même pour des mélanges complexes de protéines (Figure 5).

Figure 5 : Exemple d'identifications dans le génome de protéines d'une bande de gel 1D. 11 protéines ont été identifiées sans ambiguïté.

Spot	Fichier	Identification	Organisme	Accession number	Sequences	Segment
spot 23	E0015812.mgf	Phasin family protein.	Burkolderia mallei	Q62IY3	NQFEAQLAILNSFASTAFDSVQK-IVDLNLNVVK-SSLQESSAAAQQLISAK-DPQEFFSLSAAQAQPTAEK-HLVGITTSTQAEFAK-AAETQIAETNRK-VLSLVEELSK-NAPAGSENAIALIK-SAIGNANAGYEQLSK-HAVETLEGNLNAAASQFASAAEK	segment132
		Cytochrome c family protein.	Burkolderia mallei	Q62FM1	TIGAGSEGLAPEAVSAR-IKPVADEGFTLK-FGDAGAWGPR-IAQGYETLVK-GGNPDLDDVEVER-VVVLLANSAGGK-VPAPAAAPAADAAAK	segment531
		HYPOTHETICAL TRANSMEMBRANE PROTEIN.	Ralstonia solanacearum	Q8Y028	QANQAAQLYEELQK-FPGTAYAQMTALAAAK-SAFDANDFK-LAGILLDEK-LLSASFPAQFAGVVEDR	segment206
		AGR_L_212p.	Agrobacterium tumefaciens	Q7CVU6	DIAGAMNGVLADVFALYLK-NFHWHMSGPHFR-DYHLLLDDQAVQIYAMTDLIAER-ILDNDADYVEPNDMLAELR-TWFLFEASR	segment242
		PROBABLE LIPOPROTEIN.	Ralstonia solanacearum	Q8Y1M6	FQQVQGELSSALSR-LMAVSENYPQLK-DLQSQLEGTENR-AISAPPAVNFGK	segment103
		Transport-associated	Methylobacillus flagellatus	gi:68212434	TPGTIIDDTVITTK-TALLADSDIK-IAGDTEGVASIINNLTIK	segment383
		Superoxide dismutase [Fe] (EC 1.15.1.1).	Bordetella pertussis	P37369	ETLEFHYGK-HHQTYVTNLNNLIK-GTEFENASLEDIIK-SCVTNFGSGWTWLVK	segment396
		Heat shock protease.	Bordetella bronchiseptica	Q7VWH2	VITAPSSSTELWLIDSVR-NSALVAVSTGLLQSMTK-VVGYFVDR	segment350
		NAD(P) transhydrogenase, alpha subunit (EC 1.6.1.1).	Methylococcus capsulatus	Q603N5	VATVPEVVEK-APSAEEIGLIR-AVIEAANAFGR	segment52
		Cytochrome c family protein			LEGATGSVALPASSLPASDGK-LPDQPATGTSDSGNAYGSTQADSK	segment546
		Ubiquinol-cytochrome c reductase (EC 1.10.2.2).	Chromobacterium violaceum	Q7NQX9	LQDIGLTEEQIK-EWFGATPPDLSVIAR	segment482

5 Les recherches dans le génome, un outil d'aide à l'annotation du génome

5.1 InPACT permet la confrontation des résultats de protéomique avec les prédictions de gènes *in silico*

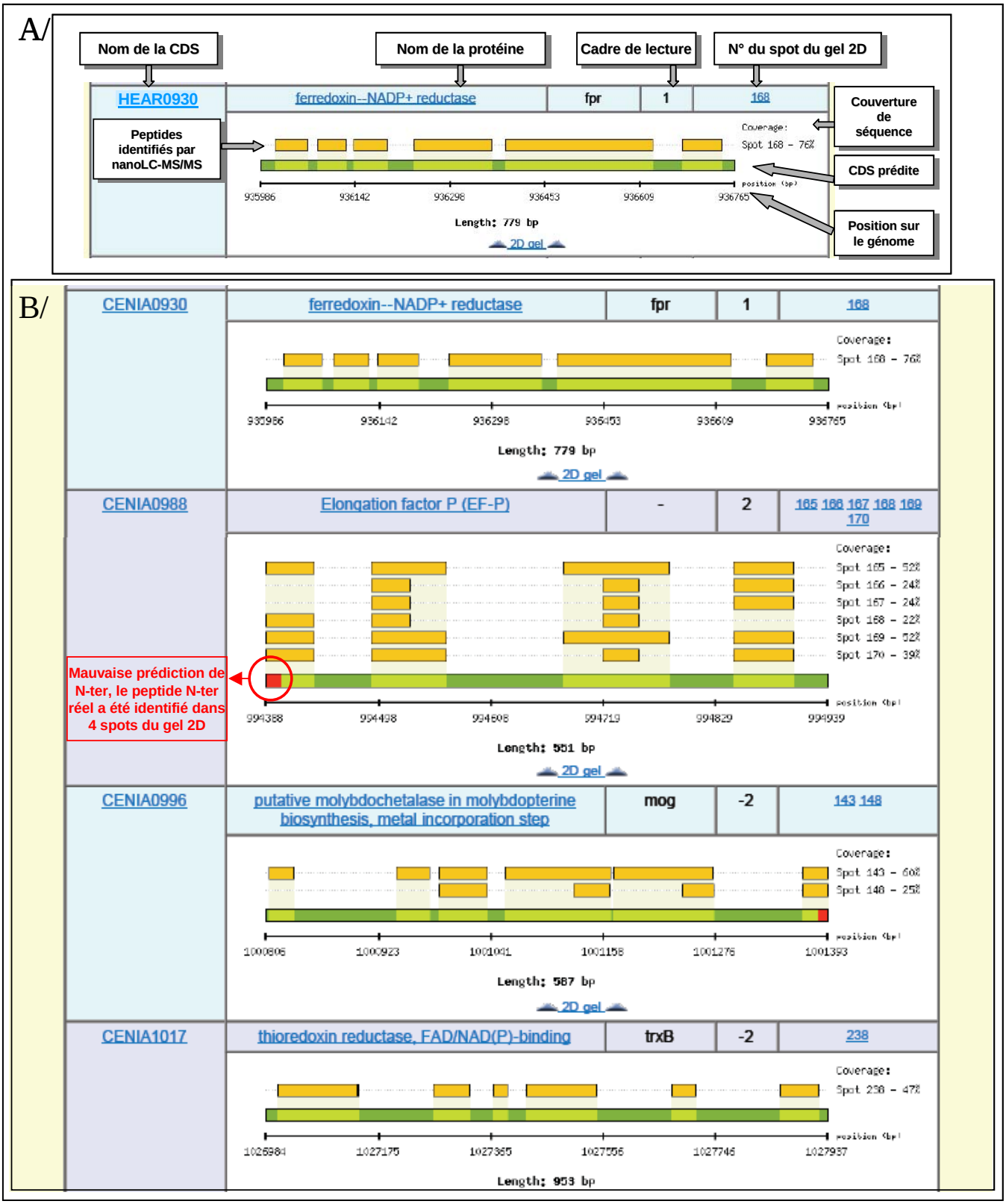
Les analyses protéomiques ont été réalisées en même temps que la prédiction *in silico* et l'annotation du génome d'*H. arsenicoxydans* par des annotateurs. L'annotation du génome a été réalisée grâce au logiciel MaGe (Magnifying Genomes, Microbial genome Annotation System) développé au Genoscope [Vallenet et coll., 2006] et le programme InPACT a été construit pour interagir avec le logiciel MaGe. En effet, InPACT permet d'extraire les régions codantes (CDS : CoDing Sequence) prédites par MaGe correspondant aux régions dans lesquelles des peptides ont été identifiés par MS et de visualiser l'alignement de ces peptides sur la CDS prédite (Figure 6). Cette représentation permet de détecter rapidement si les peptides identifiés sont en accord avec la CDS prédite ou si des peptides sont identifiés en dehors de la région codante prédite.

Pour cela un code couleur a été mis en place et comme illustré en figure 6, les peptides identifiés en dehors des CDS apparaissent en rouge. De la même manière, un code couleur est utilisé pour l'encerclement des spots sur l'image du gel : les spots dans lesquels des différences entre les résultats de protéomique et les prédictions de gènes ont été identifiées peuvent directement être repérées sur l'image globale du gel par leur couleur (Figure 3).

Cette confrontation des résultats de protéomique et de prédictions de gène permet donc :

- de visualiser rapidement des erreurs de prédiction dans les CDS, soit des mauvaises prédictions de codon d'initiation ou de terminaison.
- de détecter des régions codantes en dehors des CDS prédites. Les données de protéomique apporte donc ainsi la preuve expérimentale de l'expression d'une région codante qui aurait échappé aux algorithmes de prédiction de gènes.

Figure 6 : InPACT permet la visualisation des peptides identifiés par spectrométrie de masse alignés sur les CDS prédites sur le génome. A/ Exemple annoté de la visualisation des résultats dans InPACT. B/ Quelques exemples de protéines identifiées : certaines CDS ont été identifiées dans plusieurs spots avec des peptides communs, d'autres différents et les peptides identifiés en dehors de la CDS prédite sont présentés en rouge.



5.2 Correction des mauvaises prédictions de codons d'initiation

Les principales erreurs de prédiction qui ont pu être mises en évidence sont des erreurs de prédiction de codons d'initiation.

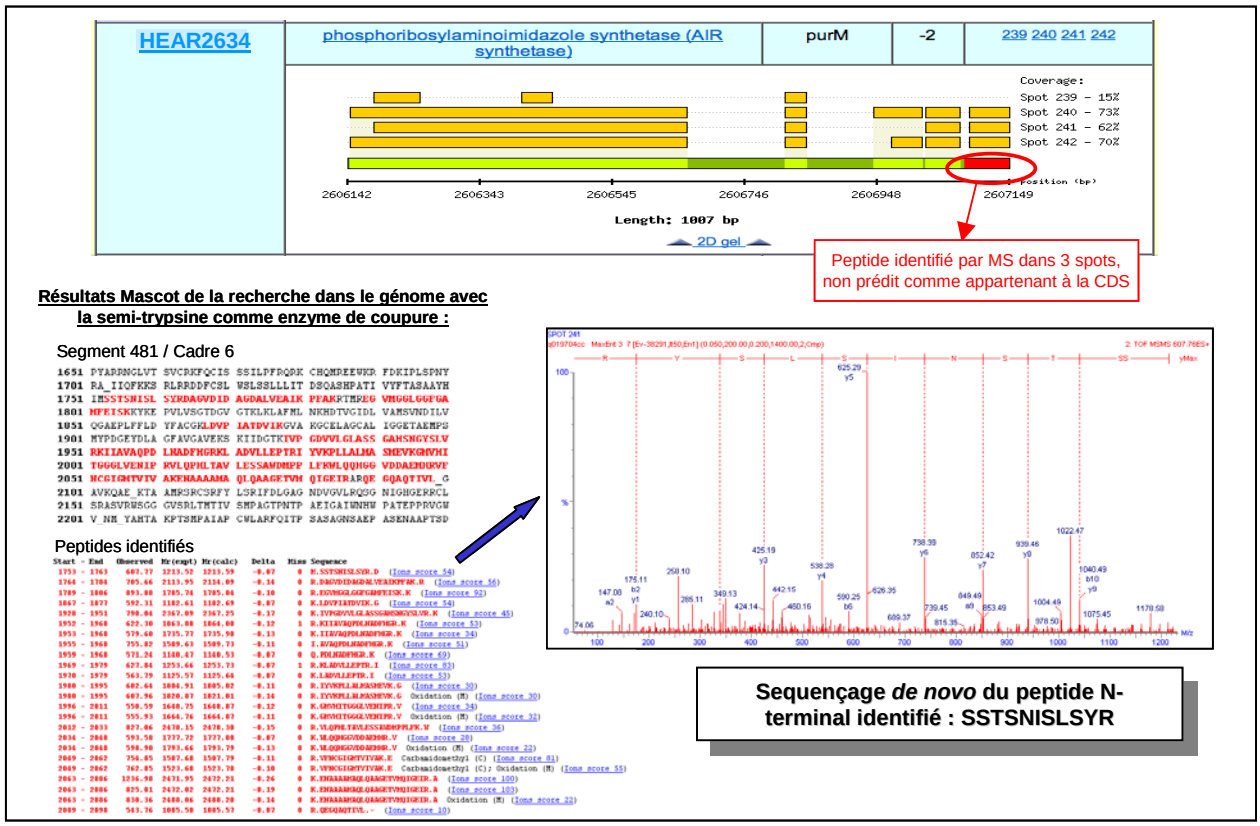
L'évolution de ces erreurs a été intéressante : dans un premier temps, une quinzaine de mauvaises prédictions de codon d'initiation a été mise en évidence pour l'ensemble des CDS prédites. Puis, les processus d'annotation *in silico* mais aussi de vérification et de correction par des experts se sont poursuivis et un certain nombre de ces erreurs ont été corrigées. Cependant, à chaque nouvel algorithme utilisé et nouvelle version de prédictions, certaines erreurs ont disparu mais de nouvelles sont apparues. Ceci démontrant bien que les algorithmes de prédiction ne sont pas parfaits et que chacun de ces algorithmes présente ses forces mais aussi ses faiblesses.

5.2.1 La visualisation des peptides N-terminaux

Pour pouvoir visualiser les peptides N-terminaux des protéines et donc les codons d'initiation, un problème se pose lors de la recherche Mascot dans le génome. En effet, lorsque l'on spécifie la trypsine comme enzyme de coupure spécifique, les peptides identifiés par Mascot ne pourront qu'exclusivement être des peptides trypsiques précédés d'une lysine ou d'une arginine (hormis les peptides N-terminaux et C-terminaux des protéines lorsque l'on recherche dans des banques protéiques). Or lors des recherches dans le génome, ce dernier a été fragmenté en segments réguliers de longueur définie sans se préoccuper des codons d'initiation des gènes. Les débuts des segments ne correspondent donc pas à des codons d'initiation donc aucunement à des peptides N-terminaux. Les peptides N-terminaux réels sont répartis à l'intérieur des segments. Les peptides N-terminaux n'apparaîtront donc pas comme des peptides trypsiques (pas précédés d'une lysine ou d'une arginine) et échapperont donc à l'identification par Mascot lors des recherches dans le génome. Or ce sont particulièrement ces peptides N-terminaux qui sont utiles pour valider le réel codon d'initiation des gènes prédits. Ainsi, afin de pouvoir voir ces peptides N-terminaux, les recherches Mascot ont été relancées une seconde fois en précisant la semi-trypsine comme enzyme de coupure : cela permet d'identifier des peptides semi-trypsiques, c'est-à-dire qui ne présentent les spécificités d'un peptide trypsique qu'à une seule de leurs extrémités.

C'est grâce à cette seconde requête Mascot que toute une série de peptides N-terminaux et C-terminaux a pu être identifiée, certains permettant de corriger des mauvaises prédictions comme illustré dans l'exemple de la figure 7.

Figure 7 : Exemple de détermination et de validation de la prédiction d'un mauvais codon d'initiation. Un peptide en amont du codon start prédit a été identifié dans 3 spots par des recherches Mascot dans le génome avec la semi-trypsine comme enzyme de coupure. La séquence du peptide N-terminal identifié a été vérifiée par séquençage *de novo* de son spectre de fragmentation.



5.2.2 *La prédiction des codons d'initiation chez les bactéries riches en GC, le cas de Mycobacterium smegmatis*

La prédiction des codons d'initiation pose problème chez un certain nombre de bactéries et principalement les bactéries appartenant à la famille des génomes riches en GC [Ou et coll., 2004 ; Zhu et coll., 2004 ; Nielsen et coll., 2005].

H. arsenicoxydans n'entre pas dans la catégorie des organismes très riches en GC car son génome présente un pourcentage en GC moyen. Cependant, il a été décrit que ce pourcentage en GC était irrégulièrement réparti sur son génome et que certaines régions pouvaient à ce titre être considérées comme très riches en GC [Muller et coll., 2006, en soumission].

Par ailleurs, dans ce contexte de difficultés de prédiction de codon d'initiation, cette stratégie de recherche dans le génome et de confrontation avec des données de prédiction a été appliquée à l'étude d'une bactérie très riche en GC, *Mycobacterium smegmatis*, étudiée au laboratoire. De sérieuses difficultés de prédiction de codons d'initiation et d'annotation se posaient aux bioinformaticiens s'intéressant à cette bactérie. Un programme d'identification d'un maximum de peptides N-terminaux par des approches protéomiques afin d'aider à l'amélioration des algorithmes de prédiction des codons d'initiation est en cours au laboratoire en collaboration avec le laboratoire de bioinformatique de l'Institut de Génétique et de Biologie Moléculaire et Cellulaire de Strasbourg. En parallèle des approches de marquage spécifique des peptides N-terminaux mises en place, la recherche dans le génome, sans traitement particulier, a déjà permis

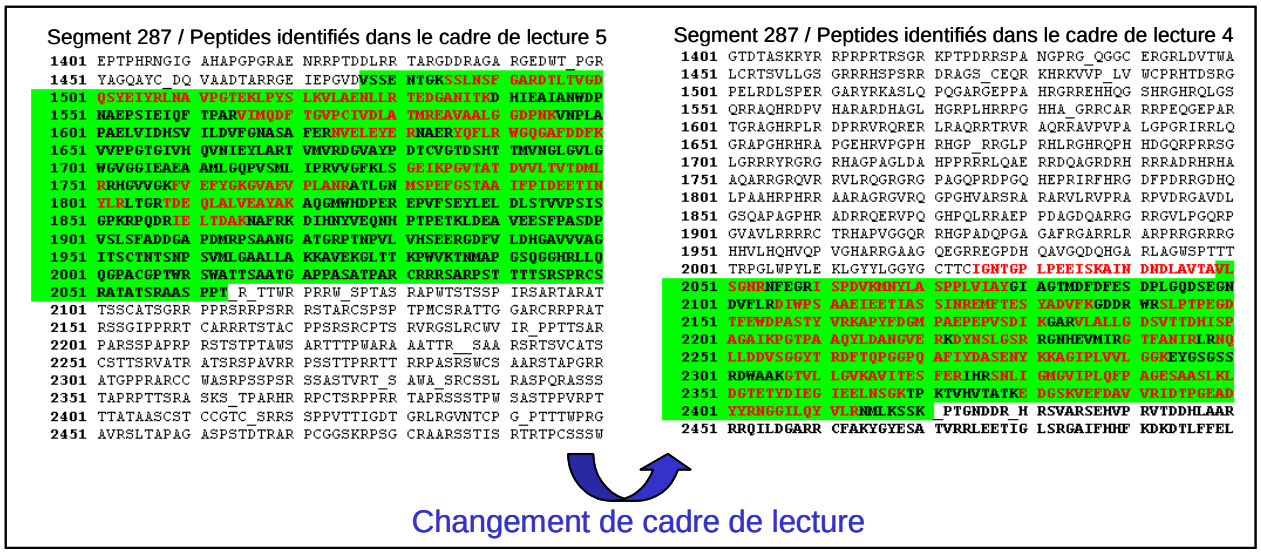
d'identifier un grand nombre de peptides N-terminaux et de mettre en évidence un nombre significatif d'erreurs de prédiction (jusqu'à 40 % de mauvaises prédictions par Glimmer).

5.3 Mise en évidence des décalages de cadre de lecture (frameshifts)

Un autre phénomène rend également la tâche d'annotation de certaines bactéries difficiles, les décalages dans le cadre de lecture. Ces décalages peuvent soient être dus à la présence d'erreurs dans la séquence soit à des phénomènes de « frameshifts » naturels qui sont cependant plus rares que les erreurs de séquençage [Farabaugh, 1996 ; Baranov et coll., 2002 ; Perrodou et coll., 2006].

Les recherches dans le génome permettent de détecter ces phénomènes puisque dans le cas d'un décalage dans le cadre de lecture, des peptides appartenant à la même protéine sont identifiés dans différents cadres de lecture. Et dans les meilleurs des cas, le peptide jonction peut même être identifié ce qui apportera la preuve expérimentale de la localisation du décalage dans le cadre de lecture. Un exemple de détection de « frameshift » est illustré en figure 8.

Figure 8 : Mise en évidence d'un phénomène de décalage dans le cadre de lecture. Des peptides ont été identifiés dans deux cadres de lectures différents dans la même région du génome et la requête MS-BLAST avec l'ensemble des peptides identifiés confirme bien qu'il s'agit d'une unique protéine.



6 Publication des résultats

Cette étude a fait l'objet de la rédaction de plusieurs publications :

- Une publication décrivant la stratégie d'identification des protéines dans le génome, le logiciel InPACT, la carte protéomique et l'apport de la stratégie pour l'aide à l'annotation des génomes a été soumise au *Journal of Proteome Research* en septembre 2006.

  Uninterpreted genome sequences used for protein identifications with tandem mass spectrometry data: An efficient complementary tool for genome structure elucidation

- Une publication décrivant particulièrement tous les aspects concernant la résistance à l'arsenic chez *Herminiimonas arsenycoxidans* tant au niveau du génome que du protéome a été soumise en juillet 2006 :

  A tale of two oxidation states: colonization of arsenic-rich environments revealed by the genome of *Herminiimonas arsenicoxydans*

  Muller D., Médigue C., Koechler S., Barbe V., Barakat M., Talla E., Bonnefoy V., Krin E., Arsène-Ploetze F., Carapito C., Chandler M., Cournoyer B., Cruveiller S., Dossat C., Duval S., Heymann M., Leize E., Lieutaud A., Lièvremont D., Makita Y., Mangenot S., Nitschke W., Ortet P., Perdrial N., Schoepp B., Siguier N., Simeonova D. D., Rouy Z., Segurens B., Turlin E., Vallenet D., Van Dorsselaer A., Weiss S., Weissenbach J., Lett M. C., Danchin A. and Bertin P.N.

- Enfin, une dernière publication décrivant en détail les résultats des cartes protéomiques est en cours de rédaction.

7 Conclusion et perspectives

Le développement de cette stratégie de recherche dans les génomes a permis :

- De démarrer le projet de carte protéomique de référence 2D pour *H. arsenicoxydans* sans avoir à utiliser ni une approche par séquençage *de novo* ni à avoir à attendre la fin du processus d'annotation du génome.
- De soulever l'utilité des recherches dans les génomes pour valider expérimentalement et corriger des prédictions de CDS *in silico*. Une telle approche protéomique constitue donc une aide complémentaire à l'annotation des génomes.

Les résultats obtenus par cette approche pour le projet portant sur *Mycobacterium smegmatis* dont le génome est très riche en GC, seront utilisés par les bioinformaticiens dans l'objectif de l'amélioration des algorithmes de prédiction des régions codantes dans les génomes riches en GC.

CHAPITRE III

Recherche dans le génome complet pour la discrimination des gènes paralogues au sein de grandes familles multigéniques

Cette étude a été réalisée en collaboration avec l'équipe du Dr Christophe Brugidou appartenant à l'UR génomique du riz de l'UMR CNRS/IRD/UP « Génome et développement des plantes » de Montpellier dans la continuité du projet précédemment décrit dans le chapitre III de la Partie II des résultats.

1 Contexte de cette étude

1.1 Gènes paralogues et familles multigéniques

Dans le génome de tous les organismes existent des gènes dont les similitudes sont telles qu'elles impliquent une parenté entre eux : on dit que ce sont des gènes homologues paralogues. Ils proviennent tous d'un gène ancestral par duplications successives suivies d'une divergence plus ou moins importante des copies. L'ensemble des gènes ainsi apparentés forme une famille multigénique (Figure 1). Plus généralement, la notion d'homologie implique la présence d'un ancêtre commun mais deux types d'homologies peuvent ensuite être distinguées (Figure 2) :

- **la paralogie** qui décrit la relation d'homologie existante entre deux gènes qui ont divergé suite à un phénomène de duplication de gène.
- **l'orthologie** qui décrit la relation d'homologie existante entre deux gènes qui ont divergé suite à un phénomène de spéciation.

Figure 1 : Illustration de la composition d'une famille multigénique.

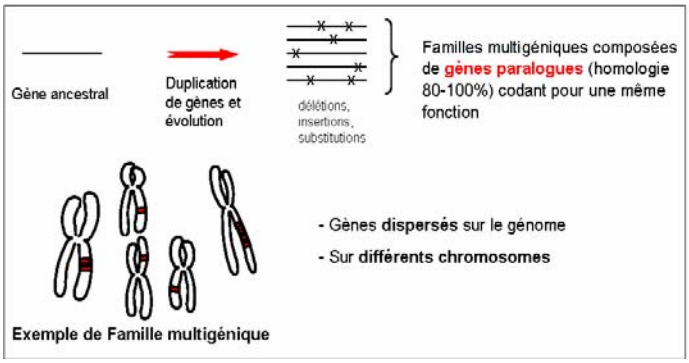
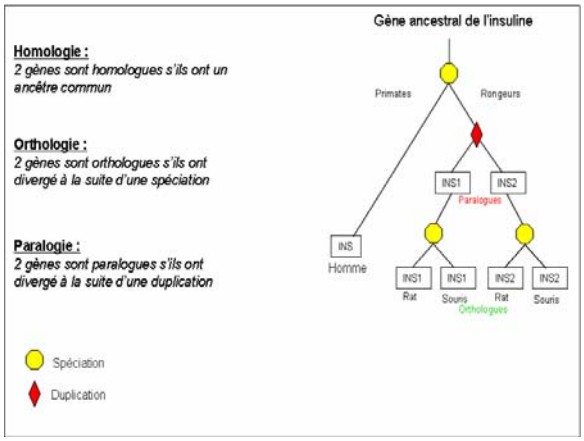
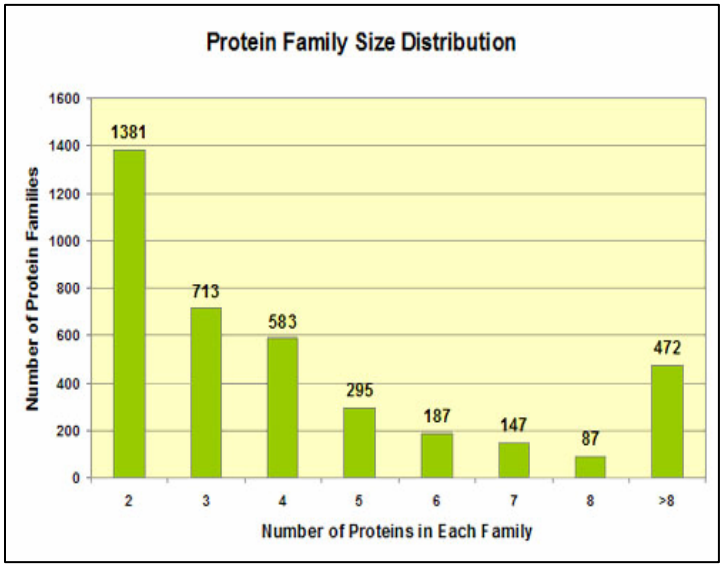


Figure 2 : Explication des notions d'homologie, de paralogie et d'orthologie.



Plus particulièrement pour le génome d'*Oryza sativa*, des statistiques ont été réalisées au TIGR (http://www.tigr.org/tdb/e2k1/osa1/para.family/index.shtml) et comme l'illustre la figure 3, le génome du riz a connu de nombreuses duplications de gènes qui ont donné lieu à de grandes familles multigéniques.

Figure 3 : Statistiques établies par le TIGR (http://www.tigr.org/tdb/e2k1/osa1/para.family/index.shtml) pour les familles de protéines chez le riz. Un total de 3865 familles de protéines paralogues contenant 21998 protéines ont été identifiée

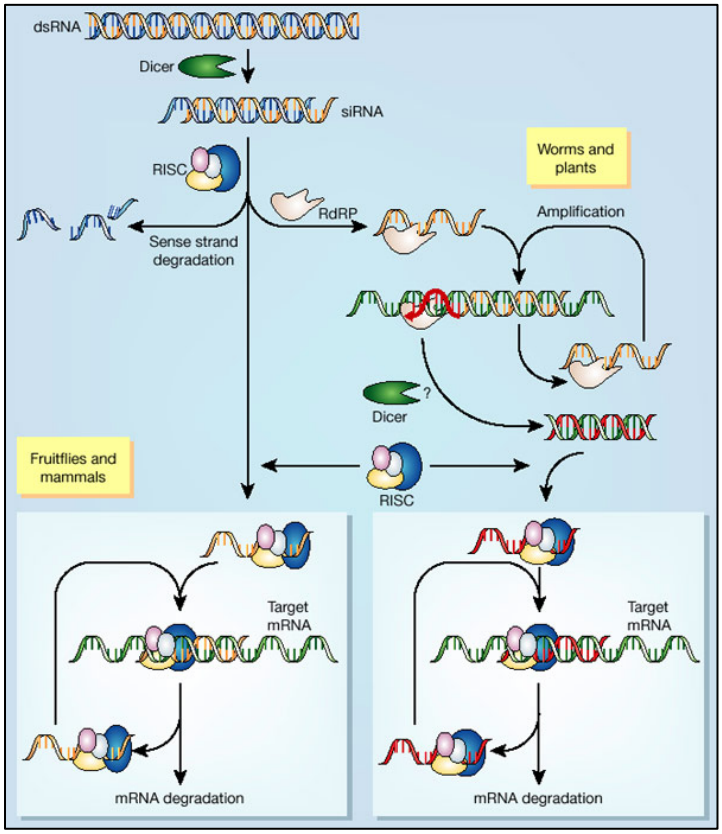


1.2 La problématique de l'infection du riz par le RYMV

La thématique biologique de ce projet a déjà été décrite dans le chapitre III de la partie II des résultats. Dans ce chapitre sera décrit la stratégie qui a été mise en place par la suite afin de pouvoir répondre précisément à une exigence particulière soulevée suite à l'étude de l'interactome virus-protéines hôtes du riz.

En effet, consécutivement à l'établissement de la liste des protéines hôtes recrutées par le virus à son entrée dans les cellules, l'objectif était d'aller perturber la formation de ces complexes en éteignant spécifiquement un ou plusieurs partenaires du complexe. A cet effet une série de candidats ont été retenus pour la mise en place d'expériences d'extinction post-transcriptionnelle de gène (technique de PTGS=post-transcriptional gene silencing chez les plantes et de RNAi=RNA interference chez les animaux). Apparaissant comme très utiles en génomique fonctionnelle, ces techniques ont connu des progrès significatifs ces dernières années [Wang et coll., 2002 ; Brummall et coll., 2003 ; Wesley et coll., 2001 ; Baulcombe, 1999 ; Wassenegger et coll., 1998 ; Depicker et coll., 1997 ; Noniva et coll., 2004]. La figure 4 donne une représentation schématique du principe de ces techniques [Noniva et coll., 2004].

Figure 4 : Short interfering RNAs-post-transcriptional gene silencing., d'après Novina et coll., Nature 2004. Représentation schématique de la stratégie d'extinction post-transcriptionnelle de gènes.



Dans le cas de l'infection d'*Oryza sativa* par le RYMV, l'extinction très spécifique d'un ou de plusieurs gènes paralogues recrutés par le virus devrait ainsi entraîner la perturbation de la formation du complexe.

Cependant, pour réaliser l'extinction très spécifique du ou des gènes paralogues recrutés, la connaissance de la séquence complète et exacte du gène paralogue en question est indispensable [Stam et coll., 1998 ; Van der Krol et coll., 1990]. Or, la plupart des protéines précédemment identifiées dans les complexes appartiennent à des grandes familles multigéniques regroupant de 2 à 10 membres, présentant une grande diversité de séquences [Wolfe et coll., 1997 ; Semple et coll., 1999 ; Zhang, 2003] et dont tous les gènes paralogues ne sont pas présents dans les banques de données protéiques.

Dans ce cadre, une stratégie utilisant le génome complet d'*Oryza sativa*, contenant toute l'information donc l'ensemble des gènes paralogues, pour réaliser les recherches avec les données de spectrométrie de masse a été développée afin de pouvoir discriminer le ou les gènes paralogues spécifiquement exprimés et recrutés par le virus.

2 La stratégie de recherche dans le génome d'*Oryza sativa*

2.1 Les échantillons et les données nanoLC-MS/MS

Les échantillons utilisés ici sont les mêmes que ceux précédemment décrits dans le chapitre III de la partie II des résultats. Les données de nanoLC-MS/MS générées pour ces échantillons ont été utilisées pour des recherches dans le génome du riz en complément des recherches dans les banques protéiques qui avaient été réalisées auparavant.

2.2 Recherche dans le génome d'*Oryza sativa* et visualisation des résultats

2.2.1 La banque de données génomiques

Les données génomiques utilisées dans ce travail ont été téléchargées du site du TIGR (htpp://www.tigr.org/tdb/e2k1/osa1/) : ces données contenaient 3755 BACs chevauchants au format fasta avec une longueur moyenne de 136 kbps.

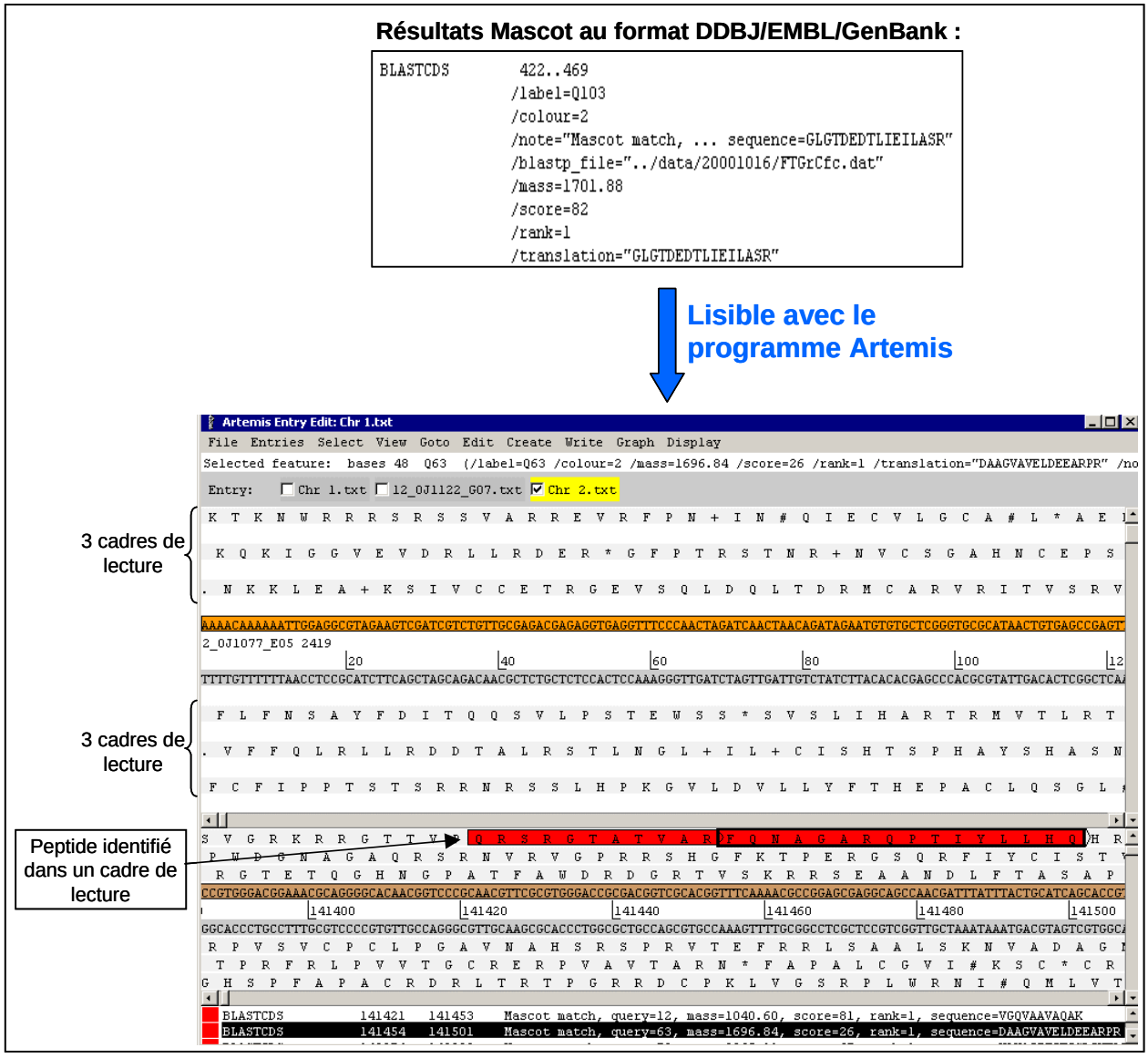
2.2.2 Visualisation des résultats d'identification

La stratégie de recherche dans les génomes est la même que celle décrite dans le premier chapitre de cette partie des résultats.

Cependant, dans un premier temps, les BACs ont été importés dans Mascot sans les fragmenter au préalable en segments. La plupart des BACs dépassant la limite de 30000 acides nucléiques, la visualisation des résultats classique par Mascot n'est plus possible et les résultats sont donnés sous la forme d'un fichier au format DDBJ/EMBL/GenBank. Ce fichier peut être enregistré au format texte et être lu par un logiciel de visualisation de séquences d'ADN et même de génome entier, comme Artemis, gratuitement téléchargeable au Sanger Institute (http://www.sanger.ac.uk/Software/Artemis/) (Figure 5). La visualisation des résultats de cette manière permet d'avoir une vue plus globale des différents peptides qui ont été identifiés. En effet, chez les eucaryotes, certains gènes possèdent des séquences introniques très longues et le fait de visualiser les peptides

identifiés sur des séquences plus longues permet d'avoir une vision plus globale de la structure des gènes identifiés.

Figure 3 : Extrait d'un fichier de résultat Mascot au format DDBJ/EMBL/GenBank qui peut être visualisé grâce à un programme de visualisation de séquences d'ADN comme Artemis (http://www.sanger.ac.uk/Software/Artemis/)



3 Exemple de discrimination d'un gène paralogue au sein d'une famille multigénique, le cas de la phenylalanine ammonia-lyase (PAL)

3.1 Représentation de la fonction PAL dans les banques protéiques et dans le génome

Lors de cette étude 3 PAL d'*Oryza sativa* étaient présentes dans **la banque UniProtKB** :

-La PAL 1, P14717 -La PAL 2, P53443 -La PAL, Q84VE0

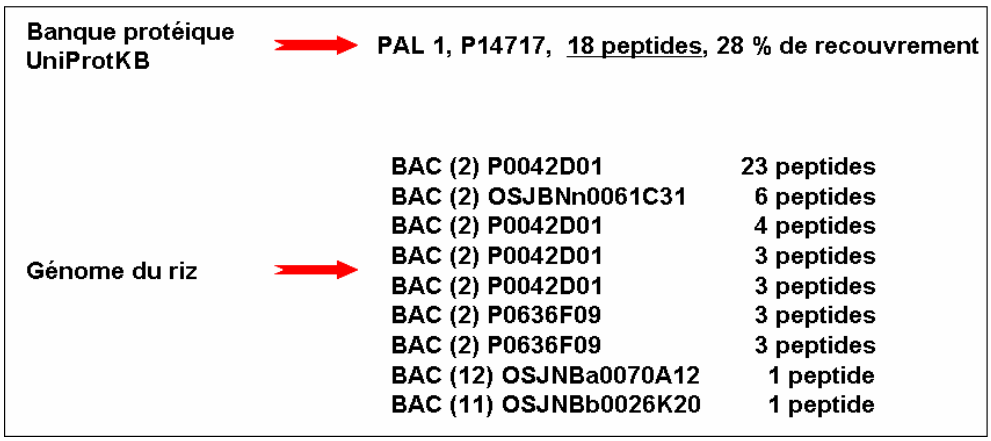
Alors que dans **le génome** d'*Oryza sativa*, 9 gènes codants pour la phénylalanine ammonia-lyase sont présents, répartis sur 5 chromosomes différents (2, 4, 5, 11, 12) et chaque séquence de gène est différente.

3.2 Résultats des recherches Mascot dans le génome et dans la banque protéique

Lors de la recherche Mascot dans la banque protéique UniProtKB, la PAL 1 est identifiée avec 18 peptides qui représentent 28% de couverture de séquence.

Lors de la recherche dans le génome, des peptides sont identifiés sur 9 BACs (Figure 4). La question qui se pose est alors de savoir lequel ou lesquels de ces gènes paralogues sont réellement exprimés.

Figure 4 : Résultats des recherches Mascot dans la banque protéique UniProtKB et dans le génome.

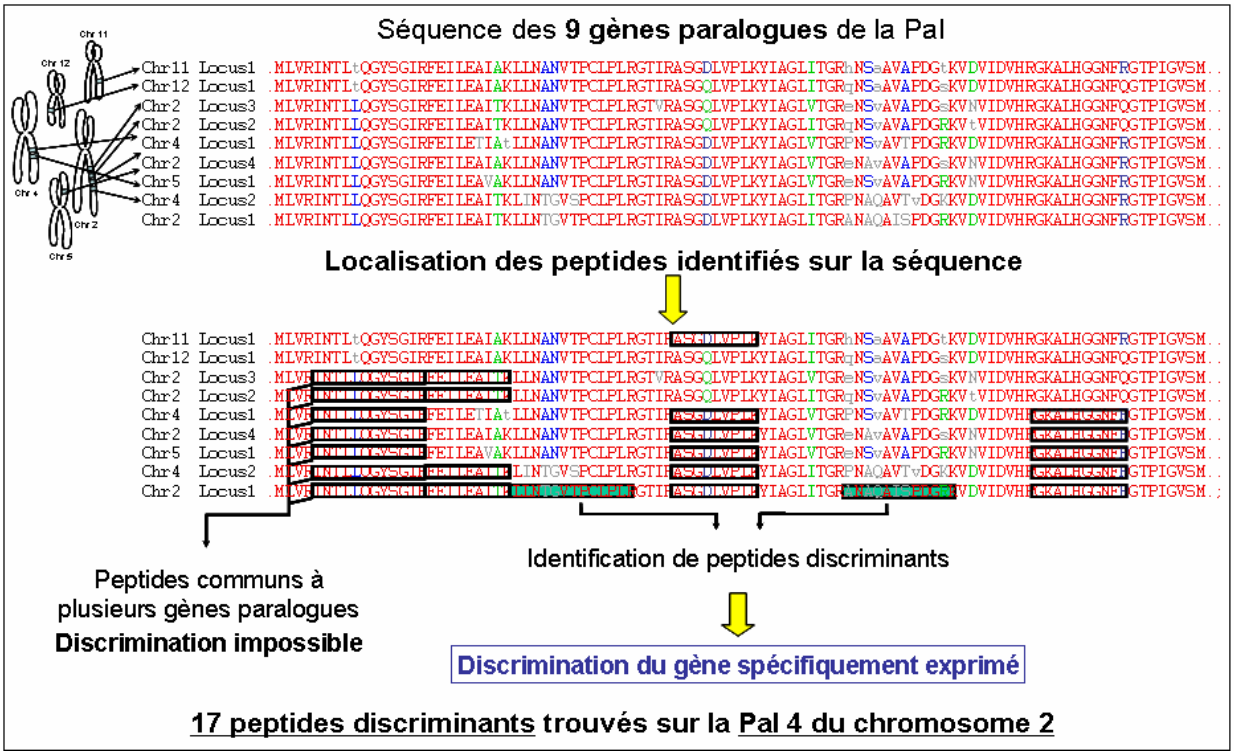


3.3 Discrimination du/des gène(s) paralogue(s) spécifiquement exprimé(s)

Pour répondre à cette question, un alignement d'un fragment des 9 gènes paralogues codant pour la PAL a été réalisé (Figure 5). Les peptides identifiés par MS et recherche dans le génome sur ces fragments ont été encadrés. Des peptides en commun ont été identifiés sur plusieurs gènes paralogues. En parallèle, des peptides n'ont été identifiés que sur un seul des gènes paralogues. Ce sont précisément ces peptides qui permettent de discriminer un gène paralogue spécifiquement exprimé. Ainsi, dans ce cas de la PAL, 17 peptides ont pu être identifiés uniquement sur le gène paralogue de la PAL sur le chromosome 2 : ces peptides discriminants permettent d'attester de l'expression de ce gène.

Par contre, des conclusions ne peuvent être tirées que lorsque des peptides discriminants sont identifiés car pour ce qui est des autres gènes paralogues, ces expériences ne permettent pas d'affirmer que ces gènes ne sont pas exprimés. En effet, lorsque des peptides discriminants sont identifiés, l'expression du gène en question est confirmée. En revanche il se peut que les peptides qui auraient pu servir à discriminer les autres gènes paralogues soient trop faiblement présents et que ces peptides n'aient donc de ce fait pas pu être sélectionnés et analysés en nanoLC-MS/MS. La non-expression des autres gènes paralogues ne peut donc pas être attestée.

Figure 5 : Exemple de la discrimination d'un gène paralogue spécifiquement exprimé dans la famille des PAL.



Plusieurs conditions favorisent la possibilité de discrimination des gènes paralogues exprimés au sein d'une famille multigénique :

- La qualité des données de spectrométrie de masse est déterminante car plus le nombre de peptides trypsiques sélectionnés et fragmentés est grand, plus la chance d'identifier des peptides discriminants est grande.
- Des degrés d'homologies entre gènes paralogues pas trop importants sont favorables car plus de peptides discriminants vont pouvoir potentiellement être identifiés
- Les divergences de séquences bien réparties sur toute la séquence du gène sont également favorables pour la même raison. Des divergences de séquence regroupées dans une zone ne vont pas donner lieu à un grand nombre de peptides discriminants potentiels.

3.4 Explication des peptides différents identifiés dans les recherches dans le génome et les recherches en banques protéiques

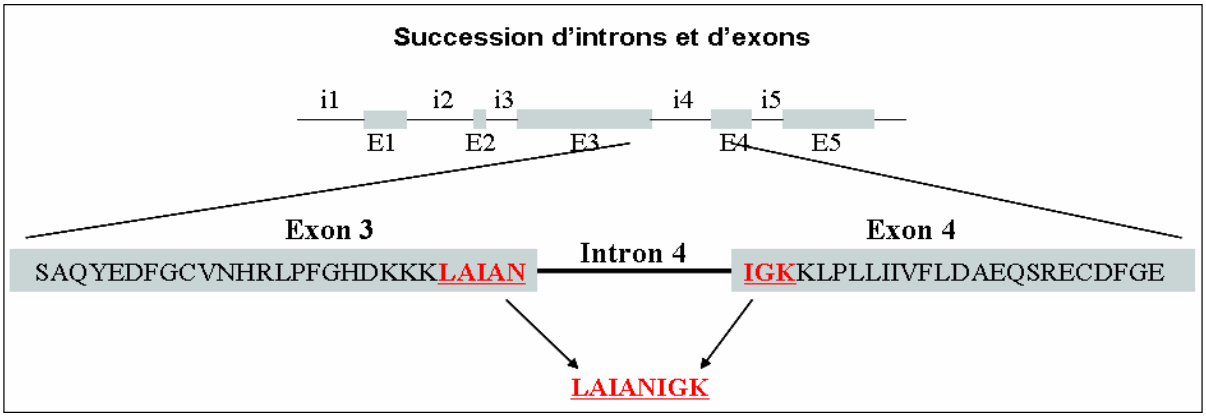
3.4.1 Identification de peptides supplémentaires dans les recherches dans le génome

L'identification de peptides supplémentaires dans les recherches dans le génome peut s'expliquer soit par l'absence de la protéine dans les banques de données protéiques soit par des erreurs dans la séquence présente dans la banque protéique.

3.4.2 Identification de peptides supplémentaires dans les recherches dans les banques protéiques

La figure 6 illustre le cas où un peptide peut être identifié dans la banque protéique mais ne le sera pas lors de la recherche dans le génome. En effet, si le peptide se trouve à la jonction intron-exon du gène, il ne pourra pas être identifié dans le génome. Néanmoins, l'identification de ces peptides-jonction est particulièrement informative puisqu'elle permet de localiser et de valider précisément la jonction intron-exon. Cette possibilité de validation de la localisation des introns-exons exprimés sur le gène est donc un autre avantage de la recherche dans les génomes pour les organismes eucaryotes pour lesquels l'élucidation des processus d'épissage alternatif est particulièrement intéressante.

Figure 6 : L'identification directe des peptides à la jonction d'un intron et d'un exon ne sera pas possible lors des recherches dans le génome.



4 Résultats publiés

Les résultats obtenus dans ce travail ont fait l'objet d'une publication acceptée dans *Proteomics* en juillet 2004.

5 Conclusion et perspectives

Dans ce travail, un intérêt plus spécifique de la recherche dans les génomes complets a pu être mis en évidence. Le génome du riz est largement connu et sa carte génétique est l'une des plus complète pour les végétaux, après celle d'*Arabidopsis thaliana* [Yuan et coll., 2001]. Pourtant, malgré cela, procéder aux recherches directement dans le génome complet a permis d'apporter des réponses qui n'auraient pu être apportées par des recherches classiques dans les banques de données protéiques.

Les recherches dans les génomes ne sont donc pas seulement une alternative lorsque l'on dispose d'un génome qui n'est encore que peu ou mal annoté, elles sont aussi nécessaires, même pour des organismes bien connus, lorsque la connaissance de la séquence précise du gène codant pour la protéine identifiée est nécessaire. Même si l'interprétation des recherches dans le génome chez les organismes eucaryotes n'est pas évidente (du à la structure complexe des gènes), ces recherches apportent des informations précieuses lorsque l'on tente d'aller plus loin que les classiques identifications de fonction des protéines.

Multigenic families and proteomics: Extended protein characterization as a tool for paralog gene identification

*François Delalande, **Christine Carapito**, Jean-Paul Brizard, Christophe Brugidou et Alain Van Dorsselaer*

Proteomics, 2005, Vol. 5, Pages 450-460

Pages 450 à 460 :

Bibliographie

Baginsky S. and Gruissem W.

Arabidopsis thaliana proteomics: from proteome to genome. *J Exp Bot*, **2006**, 21, 1-7.

Baranov P. V., Gesteland R. F. and Atkins J. F.

Recoding: translational bifurcations in gene expression. *Gene*, **2002**, 286, 187-201.

Baulcombe D. C.

Gene silencing: RNA makes RNA makes no protein. *Curr Opin Plant Biol*, **1999**, 2, 109-113.

Bianchetti L., Thompson J.D., Lecompte O., Plewniak F. and Poch O.

vALID: validation of protein sequence quality based on multiple alignment data. *J. Bioinform. Comput. Biol.*, **2005**, 3, 1-19.

Brent M.R.

Genome annotation past, present, and future : How to define an ORF at each locus. *Genome Res*. **2005**, 15 (12), 1777-1786.

Brummell D. A., Balint-Kurti P. J., Harpster M. H., Palys J. M., Oeller P.W. and Gutterson N.

Inverted repeat of a heterologous 3'-untranslated region for high-efficiency, high-throughput gene silencing. *Plant J*, **2003**, 33(4), 793-800.

Carapito C., Muller D., Turlin E., Koechler S., Danchin A., Van Dorsselaer A., Leize-Wagner E., Bertin P. N. and Lett M. C.

Identification of genes and proteins involved in the pleiotropic response to arsenic stress in *Caenibacter arsenoxydans*, a metalloresistant beta-proteobacterium with an unsequenced genome. *Biochimie*, **2006**, 88, 595-606.

Choudhary J. S., Blackstock W. P., Creasy D. M. and Cottrell J. S.

Interrogating the human genome using uninterpreted mass spectrometry data. *Proteomics*, **2001**, 1, 651-667.

Depicker A. and Montagu M. V.

Post-transcriptional gene silencing in plants. *Curr Opin Cell Biol*, **1997**, 9, 373-382.

Desiere F., Deutsch E. W., King N. L., Nesvizhskii A. I., Mallick P., Eng J., Chen S., Eddes J., Loevenich S. N. and Aebersold R.

The PeptideAtlas project. *Nucleic Acids Res.* **2006**, 34, D655-D658.

Desiere F., Deutsch E. W., Nesvizhskii A. I., Mallick P., King N. L., Eng J. K., Aderem A., Boyle R., Brunner E., Donohoe S., Fausto N., Hafen E., Hood L., Katze M. G., Kennedy K. A., Kregenow F., Lee H., Lin B., Martin D., Ranish J. A., Rawlings D. J., Samelson L. E., Shiio Y., Watts J. D., Wollscheid B., Wright M. E., Yan W., Yang L., Yi E. C., Zhang H. and Aebersold R.

Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol*, **2005**, 6, R9.

Doerks T., Bairoch A. and Bork P.

Protein annotation: detective work for function prediction. *Trends Genet.*, **1998**, 14, 248-250.

Farabaugh P.J.

Programmed translational frameshifting. *Annu. Rev. Genet*., **1996**, 30, 507-528.

Jaffe J. D., Berg H. C. and Church G. M.

Proteogenomic mapping as a complementary method to perform genome annotation. *Proteomics*, **2004**, 4, 59-77.

Jaffe J. D., Stange-Thomann N., Smith C., DeCaprio D., Fisher S., Butler J., Calvo S., Elkins T., FitzGerald M. G., Hafez N., Kodira C. D., Major J., Wang S., Wilkinson J., Nicol R., Nusbaum C., Birren B., Berg H. C. and Church G. M.

The complete genome and proteome of Mycoplasma mobile. *Genome Res.*, **2004**, 14, 1447-1461.

Kuster B., Mortensen P., Andersen J. S. and Mann M.

Mass spectrometry allows direct identification of proteins in large genomes. *Proteomics*, **2001**, 1, 641-650.

Liolios K., Tavernarakis N., Hugenholtz P. and Kyrpides N.C.

The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. *Nucleic Acids Res.*, **2006**, 34, 332-334.

Nielsen P. and Krogh A.

Large-scale prokaryotic gene prediction and comparison to genome annotation. *Bioinformatics*, **2005**, 21, 4322-4329.

Noniva C. D. and Sharp P. A.

The RNAi revolution. *Nature*, **2004**, 430, 161-164.

Ou H. Y., Guo F. B. and Zhang C. T.

GS-Finder: a program to find bacterial gene start sites with a self-training method. *Int. J. Biochem. Cell Biol.*, **2004**, 36, 535-544.

Pennisi E.

Bioinformatics. Gene counters struggle to get the right answer. *Science*, **2003**, 301, 1040-1041.

Perrodou E., Deshayes C., Muller J., Schaeffer C., Van Dorsselaer A., Ripp R., Poch O., Reyrat J. M. and Lecompte O.

ICDS database: interrupted CoDing sequences in prokaryotic genomes. *Nucleic Acids Res.*, **2006**, 34, D338-343.

Semple C. and Wolfe K. H.

Gene duplication and gene conversion in the *Caenorhabditis elegans* genome. *J Mol Evol*, **1999**, 48, 555-564.

Smith J. C., Northey J. G., Garg J., Pearlman R. E. and Siu K. W.

Robust method for proteome analysis by MS/MS using an entire translated genome: demonstration on the ciliome of Tetrahymena thermophila. *J. Proteome Res.*, **2005**, 4, 909-919.

Stam M., Viterbo A., Mol J. N. M. and Kooter J. M.

Position-dependent methylation and transcriptional silencing of trasngenes in inverted T-DNA repeats: implications for posttranscriptional silencing of homologous host genes in plants. *Ann. Bot.*, **1997**, 79, 3-12.

Vallenet D., Labarre L., Rouy Z., Barbe V., Bocs S., Cruveiller S., Lajus A., Pascal G., Scarpelli C. and Medigue C.

MaGe : a microbial genome annotation system supported by synteny results. *Nucleic Acids Res.*, **2006**, 34, 53-65.

Van der Krol A. R., Mur L. A., de Lange P., Mol J. N. and Stuitje A.R.

Inhibition of flower pigmentation by antisense CHS genes: promoter and minimal sequence requirements for the antisense effect. *Plant Mol Biol*, **1990**, 14, 457-466.

Wang M. B. and Waterhouse P. M.

Application of gene silencing in plants. *Curr Opin Plant Biol*, **2002**, 5, 146-150.

Wassenegger M. and Pelissier T.

A model for RNA-mediated gene silencing in higher plants. *Plant Mol Biol*, **1998**, 37, 349-362.

Wesley S. V., Helliwell C. A., Smith N. A., Wang M. B., Rouse D. T., Liu Q., Gooding P. S., Singh S. P., Abbott D., Stoutjesdijk P. A., Robinson S. P., Gleave A. P., Green A. G. and Waterhouse P. M.

Construct design for efficient, effective and high-throughput gene silencing in plants. *Plant J*, **2001**, 27, 581-590.

Wolfe K. H. and Shields D. C.

Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, **1997**, 387, 708-713.

Yates J.R. III, Eng J.K. and McCormack A.L.

Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. *Anal Chem*, **1995**, 67, 3202-3210.

Yuan Q., Quackenbush J., Sultana R., Pertea M., Salzberg S.L. and Buell A.R.

Rice Bioinformatics. Analysis of Rice Sequence Data and Leveraging the Data to Other Plant Species. *Plant Physiology*, **2001**, 125, 1166-1174.

Zhang J.

Evolution by gene duplication: an update. *Trends in Ecology and Evolution*, **2003**, 18, 292-298.

Zhu H. Q., Hu G. Q., Ouyang Z. Q., Wang J. and She Z. S.

Accuracy improvement for identifying translation initiation sites in microbial genomes. *Bioinformatics*, **2004**, 20, 3308-3317.

CONCLUSION GENERALE

CONCLUSION GENERALE

La première partie de ce manuscrit a été consacrée à une étude bibliographique faisant le point sur l'état actuel de l'analyse protéomique par spectrométrie de masse. Les avancées dans les trois domaines clés (l'instrumentation MS, les techniques séparatives des peptides et protéines et les banques de données de séquence) ayant permis l'ascension de l'analyse protéomique y ont été décrites ainsi que les stratégies classiquement utilisées pour l'identification des protéines avec des données de spectrométrie de masse. Les nouveaux défis confiés aujourd'hui à l'analyse protéomique par spectrométrie de masse ayant pour objectifs la caractérisation complète des protéines et leur quantification ont également été abordés.

Dans ce contexte actuel de l'analyse protéomique par MS, les objectifs de ce travail de thèse se sont inscrits à plusieurs niveaux :

- L'amélioration de la prise de données MS/MS par la mise en place et l'optimisation instrumentale des couplages de type nanoHPLC-trappe ionique.
- Les développements méthodologiques liés à l'interprétation des données MS/MS afin d'élargir le champ d'application des organismes étudiés.
- L'application de ces développements dans divers contextes biologiques pour lesquels la contribution de l'analyse protéomique par MS a été sollicitée afin d'apporter des réponses à des questions précises de collaborateurs biologistes.

En réponse à ces objectifs, les moyens mis en œuvre durant ce travail ont permis d'aboutir à une série de conclusions.

Dans la première partie des résultats sont présentées les optimisations instrumentales réalisées pour la mise en place du couplage nanoHPLC-trappe ionique (nanoHPLC Series 1100, Agilent Technologies / HCTPlus, Bruker Daltonics). Les optimisations réalisées à la fois pour la séparation chromatographique des peptides et les paramètres d'acquisition des données MS/MS ont permis d'établir des méthodes d'acquisition adaptées pour l'analyse protéomique avec ce type de couplage. L'objectif de faire de ce système nanoLC-trappe ionique un couplage complémentaire au système nanoLC-Q-TOF pour l'acquisition de données nanoLC-MS/MS interprétables par séquençage *de novo* a été atteint. Enfin, des nouveautés instrumentales ont été installées sur ce couplage, le système nanoHPLC-Chip Cube (Agilent Technologies) ainsi que la fragmentation ETD (Bruker Daltonics) et des premiers résultats ont pu être obtenus et discutés.

Les développements instrumentaux réalisés ont été mis en application pour la résolution de questions biologiques portant, dans la seconde partie des résultats, sur des organismes séquencés et dont les génomes ont été bien annotés. Une carte protéomique de référence de la levure *Schizosaccharomyces pombe*, organisme modèle en plein essor, a pu être réalisée de même qu'une étude de l'effet d'un minéralocorticoïde, l'aldostérone, par une approche différentielle par gel 2D sur ce même organisme. De la même manière, l'exoprotéome d'un champignon pathogène du houblon, *Fusarium graminearum* a été étudié ainsi que l'interactome virus-protéines hôtes recrutées lors de l'infection du riz par le RYMV. Ces travaux ont été concrétisés par la publication de quatre articles.

L'analyse protéomique par MS a longtemps été réservée aux organismes « modèles », bien étudiés et de génome séquencé, les approches classiques d'identification des protéines nécessitant la présence de la protéine dans la banque de données utilisée pour les recherches. Néanmoins, nombreux sont les biologistes à s'intéresser à des organismes plus « exotiques » dont le génome n'a pas été séquencé. Pour répondre à ces besoins, l'interprétation par séquençage *de novo* et les identifications des protéines par homologies de séquence, tolérantes aux erreurs, sont la solution. Des développements réalisés dans cette voie durant ce travail ont permis la réalisation d'études protéomiques sur des organismes comme *Vitis vinifera*, la vigne, *Dicentrarchus labrax* , le poisson bar ou encore la bactérie *Caenibacter arsenoxydans*. Un chapitre méthodologique est suivi de deux applications ayant donné lieu à des publications, l'une concernant la bactérie *Caenibacter arsenoxydans* et l'autre la vigne *Vitis vinifera*.

Enfin, une stratégie permettant la recherche avec des données MS/MS directement dans les données génomiques non interprétées a été mise en place. Cette stratégie nous a permis de mettre en évidence l'intérêt de l'analyse protéomique par MS comme outil de validation des prédictions de gènes réalisées *in silico*. Cela particulièrement dans le cadre d'études portant sur des bactéries appartenant à la famille des génomes riches en GC présentant des difficultés pour la prédiction *in silico* des gènes. Cette approche nous a également permis de discriminer des gènes paralogues au sein de familles multigéniques dans le cadre de l'étude de l'interactome riz-RYMV.

Finalement, il m'a paru intéressant de soulever dans cette conclusion l'importance de la réflexion portant sur le choix de la stratégie à mettre en œuvre pour répondre aux différentes problématiques biologiques. Les questions biologiques variées abordées durant ce travail ont nécessité des choix stratégiques précis à plusieurs niveaux :

- Au niveau de la séparation des peptides et protéines
- Au niveau de l'instrumentation MS utilisée
- Au niveau de la banque de données utilisée pour les requêtes
- Au niveau de la stratégie d'interprétation des données pour l'identification des protéines

L'étude bibliographique menée en première partie de ce manuscrit m'a permis de dresser un tableau récapitulatif contenant les principaux avantages et inconvénients des différentes techniques et approches disponibles pour la séparation, l'instrumentation MS et l'identification des protéines. J'ai ici complété ce tableau en y ajoutant les apports et réflexions de ce travail de thèse concernant notamment :

- les nouveautés instrumentales : le nanoHPLC-Chip Cube et la fragmentation ETD,
- quelques éléments conditionnant le choix de la banque de données utilisée issus des réflexions menées sur les recherches avec des données MS/MS dans les données génomiques non interprétées,
- les approches d'identification par séquençage *de novo* ayant permis l'élargissement des études protéomiques à des organismes non séquencés.

La stratégie de séparation des protéines

❑ Le gel d'électrophorèse 2D

-Image globale des protéines avec une relativement bonne résolution (2000 protéines séparées en routine)
-Quantification relative possible par comparaison d'images de gel correspondant à deux états

-Problème des protéines hydrophobes
-Faible gamme dynamique
-Automatisation difficile
-Multiplicité des protéines par spot et multiplicité des spots par protéines rend la quantification difficile
-Mauvais rendement de l'extraction des peptides après digestion *in-gel*

❑ Le gel d'électrophorèse 1D

-Pas de problème de sous-représentation des protéines membranaires
-Facile à mettre en œuvre
-Couplage avec la nanoLC

-Faible pouvoir résolutif, utilisé en pré-fractionnement uniquement
-Mauvais rendement de l'extraction des peptides après digestion *in-gel*
-Ne permet pas de quantification

❑ Les stratégies par chromatographies multiples

-Combinaisons de plusieurs types de chromatographie possible (phase inverse, SCX, échangeuse d'anions, gel filtration, …)
-Automatisation possible
-Analyse de mélanges très complexes, pouvoir résolutif élevé
-Compatible avec les approches quantitatives ICAT, SILAC, ...

-Montage des chromatographies multiples en série délicats
-Problèmes de compatibilité entre éluants et analyse ESI
-Problèmes de reproductibilité et de validation des identifications pour les mélanges peptidiques très complexes

L'instrumentation utilisée

❑ Approche par MALDI-TOF-MS (ou LC-MS)

-Analyse rapide (1min pour l'acquisition d'un spectre)
-Bonne précision de mesure (<20ppm)
-Bonne sensibilité (femtomolaire)
-Bonne tolérance aux sels

-Problème de suppression de signal donc inadaptée pour les mélanges complexes
-Inadaptée pour les petites protéines car peu de peptides trypsiques
-Pas d'informations de séquence
-Pas de quantification

❑ Approches nanoLC-MS/MS

-Etape de séparation des peptides donc adaptée pour les mélanges plus complexes
-Détermination de séquence grâce à la MS/MS
-Spécificité et pouvoir discriminant apportés par la MS/MS
-Identification de modifications post-traductionnelles
-Quantification possible

-Analyse plus longue (environ 1h/analyse)
-Gestion informatique de fichiers d'analyses très volumineux
-Couplage nanoLC délicat : problèmes fréquents de bouchages, surpressions, fuites, …
-Reproductibilité difficile dans le cas de mélanges très complexes

❑Nouvelles composantes du couplage nanoLC-MS/MS (NanoHPLC-Chip Cube + Fragmentation ETD)

-Durée de la chromatographie réduite (environ 20min/analyse)
-Meilleure résolution chromatographique et sensibilité grâce au système nanoChip-Cube
-Des séquences plus complètes, moins de coupures préférentielles, maintien des modifications post-traductionnelles en fragmentation ETD

-Maintenance de l'instrumentation délicate (arrêt des pompes nano sur le système nanoHPLC-Chip Cube, nettoyage de la source ETD, …)
-Perte de sensibilité en fragmentation ETD et difficultés pour la fragmentation des grands peptides et protéines

Choix de la banque de données utilisée

❑Banques de données de protéines

-Adaptées pour les organismes séquencés et bien annotés comme *E. coli*, l'humain, la souris, …
-Identification de protéines avec bonne qualité d'annotation dans UniProtKB/SwissProt

-Problème de la non-validation des séquences présentes dans les banques protéiques issues de la traduction automatique *in silico* des génomes
-Présence d'erreurs de séquences et d'erreurs d'annotations
-Les banques protéiques sont incomplètes pour la plupart des organismes

❑Banques d'ESTs

-Permettent potentiellement l'identification de nouvelles protéines ou de variants d'épissage de protéines connues
-Données de nanoLC-MS/MS uniquement

-Pour être significatives, elles doivent contenir beaucoup d'ESTs, donc la banque devient volumineuse et les temps de recherche longs
-Les séquences sont très courtes, et parfois de mauvaise qualité

❑Banques génomiques

-Contiennent l'ensemble de l'information
-Taux d'erreurs plus faible
-Permettent la validation des prédictions et l'identification de nouvelles protéines non prédites et non annotées
-Données de nanoLC-MS/MS

-Grandes tailles des banques donc recherches longues selon la taille du génome
-Difficulté d'interprétation des résultats pour les génomes eucaryotes avec les structures intron/exon complexes

La stratégie d'identification

❑PMF (Empreinte peptidique massique)

-Identification rapide pour des protéines uniques ou des mélanges de faible complexité
-Adaptée pour l'identification de protéines présentes dans les banques de données

-Identifications difficiles lorsque plusieurs PMF sont superposées sur un même spectre
-Approche non tolérante aux erreurs
-Faible pouvoir discriminant

❑PFF (Empreinte de fragmentation peptidique)

-Spécificité et pouvoir discriminant élevés grâce à l'information de MS/MS
-Information de séquence
-Identification des modifications post-traductionnelles
-Adaptée pour l'identification de protéines présentes dans les banques de données

-Approche non tolérante aux erreurs
-Recherches en banque de données longues, en fonction de la banque de données choisie, car fichiers volumineux

❑Approches par séquençage *de novo*

-Approche tolérante aux erreurs
-Etudes sur des organismes **non séquencés** possible
-Approche complémentaire au PFF pour la détection de mutations, …

-Traitement des données long car nécessite une vérification manuelle des séquences