

2M11.2986.6

Université de Montréal

Interprétation automatisée des spectres de masse
de protéines obtenus par ionisation Electrospray

Par
Martin Ethier

Département de Chimie
Faculté des Arts et Sciences

Mémoire présenté à la Faculté des études supérieures
En vue de l'obtention du grade de Maîtrise ès Sciences (M. Sc.)
En Chimie

Décembre 2001



©Martin Ethier, 2001

QD

3

U54

2002

V.023

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé :
Interprétation automatisée des spectres de masse
de protéines obtenus par ionisation Electrospray

présenté par :

Martin Ethier

A été évalué par un jury composé des personnes suivantes :

Thomas Ellis

président-rapporteur

Michel J. Bertrand

directeur de recherche

Karen Waldron

membre du jury

Résumé

L'ionisation Electrospray en spectrométrie de masse permet l'analyse des protéines. Par contre, les spectres de masse obtenus sont compliqués par le fait que chaque protéine possède une distribution d'ions multichargés. Il est alors nécessaire d'utiliser des méthodes de traitement automatisées appelées méthodes de déconvolution afin d'éliminer cette dégénérescence. Plusieurs groupes ont déjà proposé de tels algorithmes, mais tous présentent des limitations importantes. La présente étude suggère deux nouveaux algorithmes de déconvolution qui tentent d'éviter ces limitations.

Le premier algorithme détermine la distribution d'ions multichargés la plus probable en testant une à une chacune des distributions possibles. Deux façons d'utiliser cet algorithme sont proposées.

Le deuxième algorithme effectue la multiplication du spectre de masse par une série d'entiers pour obtenir un seul spectre multiplié. L'algorithme détermine la position dans ce spectre où les pics se regroupent, ce qui indique les charges de la distribution d'ions multichargés et permet la simplification du spectre de masse.

Les deux algorithmes sont des processus itératifs qui permettent de traiter une à la fois les protéines présentes dans l'échantillon, rendant par le fait même l'analyse des mélanges possible. Les deux algorithmes font appel à un prétraitement des données par une étude statistique du signal de fond ce qui améliore la déconvolution.

Il est montré que les algorithmes offrent des caractéristiques complémentaires et que l'utilisation en parallèle des algorithmes est donc souhaitable.

Mots clés

spectrométrie de masse, déconvolution, algorithme, filtration, protéomique, itération.

Abstract

Electrospray ionisation in mass spectrometry enables protein analysis. But mass spectra obtained are complicated since each protein has a distribution of charged states. Methods for automatic treatment of data, called deconvolution, are necessary. Several research groups have proposed algorithms, but all of them have important limitations. This study proposes new algorithms of deconvolution, which try to avoid these limitations.

The first algorithm finds the most probable distributions of charged states by testing one by one all possible distributions. Two ways of using the algorithm are proposed.

The second algorithm multiplies the mass spectrum by a series of integers to obtain one multiplied spectrum. The algorithm finds the position in it where all peaks regroup, which indicates the charges of the distribution and enables the simplification of the mass spectrum.

Both algorithms are iterative processes that treat each protein one at a time, making the analysis of mixtures possible. Both algorithms use a statistical study of the noise to improve the deconvolution.

It is shown that both algorithms offers complementary characteristics, thus they should be used in parallel to give best results.

Keywords

Mass spectrometry, deconvolution, algorithm, filtration, proteomics, iterative process.

Table des Matières

<i>Résumé</i>	<i>ii</i>
<i>Abstract</i>	<i>iii</i>
<i>Table des Matières</i>	<i>iv</i>
<i>Listes des Figures</i>	<i>vi</i>
<i>Liste des Tableaux</i>	<i>ix</i>
<i>Liste des Abréviations</i>	<i>x</i>
<i>Dédicace</i>	<i>xi</i>
<i>Remerciements</i>	<i>xii</i>
<i>Chapitre I</i>	<i>1</i>
Introduction	<i>2</i>
<i>Chapitre II</i>	<i>5</i>
Revue de la littérature	<i>6</i>
Historique de l'ESI	<i>6</i>
Allure du spectre de masse des protéines en ESI	<i>6</i>
Interprétation manuelle des spectres de masse	<i>8</i>
Déconvolution des spectres de masse (approches existantes)	<i>10</i>
Algorithme de détermination statistique de la masse de la protéine	<i>10</i>
Algorithme de déconvolution (Mann et Fenn)	<i>12</i>
Entropie maximale.....	<i>14</i>
Déconvolution tridimensionnelle	<i>16</i>
Corrélation multiplicative.....	<i>17</i>
Zscore	<i>18</i>
Algorithmes suggérés	<i>19</i>
<i>Chapitre III</i>	<i>22</i>
Partie Théorique	<i>23</i>
Principes de base.....	<i>23</i>
Principe de l'ionisation Electrospray.....	<i>24</i>
Mécanisme de l'ionisation Electrospray.....	<i>26</i>
Facteurs influençant la protonation des protéines	<i>27</i>
Perturbation du spectre par les mélanges et le signal de fond	<i>29</i>
<i>Chapitre IV</i>	<i>31</i>
Partie Expérimentale	<i>32</i>
Appareils utilisés.....	<i>32</i>
Protéines utilisées.....	<i>32</i>
Langage de programmation	<i>33</i>
<i>Chapitre V</i>	<i>34</i>
Résultats	<i>35</i>
Traitement du signal de fond	<i>35</i>
Déconvolution	<i>38</i>
Détermination des distributions d'ions multichargés	<i>41</i>
Détermination des distributions les plus probables	<i>41</i>

Détermination des positions de haute densité de masse	46
Processus itératif	51
Utilisation directe vs inversée.....	52
Chapitre VI	53
Applications	54
Évaluation de l'erreur sur les résultats	54
Simulation du spectre d'une protéine pure.....	55
Insuline	57
Ubiquitine.....	63
Cytochrome C	67
Simulation d'un mélange binaire de protéines	72
Mélange binaire	75
Mélange tertiaire	79
Mélange quaternaire	85
Échantillon inconnu	90
Discussion des algorithmes	96
Chapitre VII	100
Conclusion Générale	101
Références	104

Listes des Figures

Figure 1 : Spectre de masse non filtré de la myoglobine (MM = 16953 Da) obtenu par ESI.....	7
Figure 2 : Exemple d'utilisation de l'algorithme de détermination statistique de la masse de la protéine. La courbe continue représente les valeurs théoriques. (tiré de la référence 8).....	11
Figure 3 : Exemple de déconvolution par l'algorithme de Mann et Fenn. Haut) Spectre de masse simulé (10000 Da). Bas) Spectre déconvolué. (tiré de la référence 8)	13
Figure 4 : Schéma du montage permettant l'ionisation Electrospray.	24
Figure 5 : Exemple d'utilisation de l'algorithme de filtration pour trouver l'intensité la plus fréquente d'un intervalle de m/z.....	36
Figure 6 : Exemple de filtration d'un spectre de masse simulé (haut) . Le signal de fond continu a été enlevé avec succès dans le spectre de masse filtré (bas).....	39
Figure 7 : Exemple d'une distribution d'ions multichargés qui n'a pas la forme d'une distribution de Poisson.	41
Figure 8 : Organigramme de l'algorithme de détermination des distributions les plus probables.	42
Figure 9 : Exemple du traitement d'une distribution d'ions multichargés avec l'algorithme Probablion. a) Spectre de masse. b) Calcul de la masse déconvoluée proposée. c) et d) Calculs des m/z des trois prochains pics attendus d'après la masse déconvoluée proposée.....	44
Figure 10 : Exemple de traitement d'une distribution par l'algorithme.	46
Figure 11 : Organigramme de l'algorithme de détermination des positions de haute densité de masse.	47
Figure 12 : Exemple de déconvolution avec l'algorithme Dension. a) Spectre de masse simulé. b) Spectre multiplié.	49
Figure 13 : Exemple d'harmoniques créés lors de la multiplication des pics.....	50
Figure 14 : Spectre de masse simulé d'une protéine de masse de 15000 Da sans signal de fond continu et une précision infinie.	56
Figure 15 : Spectre déconvolué de la Figure 14 selon l'algorithme de détermination des distributions les plus probables avec l'intervalle de confiance entre parenthèses.	56
Figure 16 : Spectre déconvolué de la Figure 14 selon l'algorithme de détermination des positions de haute densité de masse avec l'intervalle de confiance entre parenthèses.	57
Figure 17 : Spectres de masse accumulés (77 spectres) de l'insuline pure (5734 Da). a) Spectre non filtré. b) Spectre filtré.....	58
Figure 18 : Spectres déconvolués de l'insuline pure par l'algorithme Probablion. a) À partir du spectre non filtré. b) À partir du spectre filtré.	60
Figure 19 : Spectre déconvolué pour l'insuline par l'utilisation inverse de l'algorithme Probablion. Encadré : Agrandissement autour du pic 5733 Da. ...	61
Figure 20 : Spectre déconvolué de l'insuline par l'utilisation de l'algorithme Dension.	62

Figure 21 : Spectres de masse accumulés (75 spectres) de l'ubiquitine pure (8565 Da). a) non filtré. b) filtré.....	64
Figure 22 : Spectres déconvolués de l'ubiquitine pure par l'utilisation directe de l'algorithme Probablion. a) À partir du spectre non filtré. b) À partir du spectre filtré.....	65
Figure 23 : Spectre déconvolué par l'utilisation inverse de l'algorithme Probablion sur les 75 spectres de masse de l'ubiquitine.	66
Figure 24 : Spectre déconvolué obtenu avec l'algorithme Dension pour l'ubiquitine.	67
Figure 25 : Spectres de masse accumulés (75 spectres) du cytochrome C pur (12230 Da). a) Spectre non filtré. b) Spectre filtré.....	68
Figure 26 : Spectres déconvolués du cytochrome C pur par l'utilisation directe de l'algorithme Probablion. a) À partir du spectre non filtré. b) À partir du spectre filtré.....	69
Figure 27 : Spectre déconvolué par l'utilisation inverse de l'algorithme Probablion pour les 75 spectres de masse du Cytochrome C.	70
Figure 28 : Spectre déconvolué obtenu à partir du spectre de masse filtré du cytochrome C avec l'algorithme Dension.....	71
Figure 29 : Spectres de masse simulés d'un mélange d'insuline et de cytochrome C. a) Spectre non filtré. b) Spectre filtré.....	73
Figure 30 : Spectres déconvolués par l'utilisation directe de l'algorithme Probablion d'un mélange simulé d'insuline et de Cytochrome C. a) À partir du spectre non filtré. b) À partir du spectre filtré.....	74
Figure 31 : Spectre déconvolué obtenu par l'algorithme Dension pour le spectre de masse filtré du mélange simulé de l'insuline et du cytochrome C.....	75
Figure 32 : Spectres de masse accumulés (80 spectres) d'un mélange de l'ubiquitine (8565 Da) et du cytochrome C (12230 Da). a) Spectre non filtré. b) Spectre filtré.....	76
Figure 33 : Spectres déconvolués par utilisation directe de l'algorithme Probablion d'un mélange d'ubiquitine et du Cytochrome C. À partir du spectre non filtré (a) et filtré (b).	77
Figure 34 : Spectre déconvolué du mélange d'ubiquitine et du cytochrome C par l'utilisation inverse de l'algorithme Probablion sur 80 spectres de masse.	78
Figure 35 : Spectre déconvolué par l'algorithme Dension du spectre de masse accumulé et filtré du mélange d'ubiquitine et du cytochrome C.	79
Figure 36 : Spectres de masse accumulés (76 spectres) d'un mélange de l'ubiquitine (8565 Da), du cytochrome C (12230 Da) et de la myoglobine (16952 Da). a) Spectre non filtré. b) Spectre filtré.....	80
Figure 37 : Spectres déconvolués par l'utilisation directe de l'algorithme Probablion d'un mélange d'ubiquitine, du cytochrome C et de myoglobine. a) À partir du spectre non filtré. b) À partir du spectre filtré.....	82
Figure 38 : Spectre déconvolué du mélange d'ubiquitine, du cytochrome C et de la myoglobine par l'utilisation inverse de l'algorithme Probablion.	83
Figure 39 : Spectre déconvolué par l'algorithme Dension du spectre de masse accumulé et filtré du mélange d'ubiquitine, du cytochrome C et de la myoglobine.....	84

Figure 40 : Spectres de masse accumulés (83 spectres) d'un mélange de l'insuline (5734 Da), d'ubiquitine (8565 Da), du cytochrome C (12230 Da) et de la myoglobine (16952 Da). a) Spectre non filtré. b) Spectre filtré.	86
Figure 41 : Spectres déconvolués par l'utilisation directe de l'algorithme Probablion d'un mélange d'insuline, d'ubiquitine, du cytochrome C et de myoglobine. a) À partir du spectre non filtré. b) À partir du spectre filtré.	88
Figure 42 : Spectre déconvolué du mélange d'insuline, d'ubiquitine, du cytochrome C et de la myoglobine par l'utilisation inverse de l'algorithme Probablion au 83 spectres de masse.	89
Figure 43 : Spectre déconvolué par l'algorithme Dension du spectre de masse accumulé et filtré du mélange d'insuline, d'ubiquitine, du cytochrome C et de la myoglobine.	90
Figure 44 : Spectres de masse de l'échantillon inconnu en format continu. a) spectre non filtré. b) spectre filtré.	92
Figure 45 : Spectres de masse de l'échantillon inconnu après la reconnaissance de pic. a) Spectre obtenu à partir du spectre de masse en continu filtré. b) Spectre de masse après filtration de a.	93
Figure 46 : Spectres déconvolués des spectres de masse de la Figure 45 avec l'algorithme Probablion. a) à partir du spectre de masse non filtré. b) à partir du spectre de masse filtré.	94
Figure 47 : Spectre déconvolué du spectre de masse filtré de la Figure 45b avec l'algorithme Dension.	95

Liste des Tableaux

Tableau I : Masse des protéines utilisées lors de l'analyse.....	33
--	----

Liste des Abréviations

Da :	Dalton
EI :	Electron impact (ionisation par impact électronique)
ESI:	ElectroSpray Ionization (ionisation ElectroSpray)
FAB:	Fast Atom Bombardment (Bombardement d'atomes rapides)
I.c. :	Intervalle de confiance
Int. Abs. :	Intensité absolue
MALDI :	Matrix Assisted Laser Desorption Ionisation
MM :	Masse moléculaire de la protéine
MS :	Spectrométrie de masse
MS/MS :	Spectrométrie de masse en tandem
m/z:	Rapport de la masse sur la charge
m/z max :	Valeur maximale de m/z pour une distribution d'ions multichargés
rms :	root mean square (racine carrée de la moyenne des carrés)
Π :	Multiplication successive

Dédicace

À ma femme Nadine qui a toujours su me supporter dans les bons et les mauvais moments.

Remerciements

J'aimerais tout d'abord remercier mon directeur de recherche Michel J. Bertrand de m'avoir accepté dans son groupe. Ces conseils et apprentissages se sont révélés très utiles à ma formation. J'aimerais aussi remercier Dimo Zidarov pour l'aide exceptionnelle qu'il m'a généreusement accordée sur le plan informatique. L'aide de Gabriel Sanchez et Daniel Carazzato a aussi été vitale puisqu'ils ont aidé à obtenir les spectres de masse expérimentaux. J'aimerais remercier Theodore Krastev pour ses suggestions par rapport à la simulation des spectres de masse. Finalement, j'aimerais remercier le Conseil de recherche en sciences naturelles et en génie (CRSNG) pour le financement du projet.

Chapitre I

Introduction

La spectrométrie de masse est une technique importante en chimie analytique moderne. Celle-ci permet la caractérisation et la quantification d'espèces inconnues dans des échantillons complexes. L'analyse de molécules non volatiles a par contre amené des problèmes depuis les débuts de la spectrométrie de masse. Plusieurs approches ont été tentées afin de contourner ces problèmes et de pouvoir étudier les molécules d'intérêt biologique qui sont généralement assez grosses. Les protéines sont des molécules de ce type et peuvent être ionisées sans amener une fragmentation excessive.

Il a tout d'abord été tenté de dériver les protéines afin de les rendre suffisamment volatiles pour être amenées à l'état gazeux sans trop d'efforts¹ avec la source d'ionisation à impact électronique (EI). Cette méthode se limite aux petits peptides seulement² et demande des manipulations supplémentaires.

Les limitations de la technique de dérivation ont mené à l'ionisation par bombardement d'atomes rapides (FAB)³. La technique permet l'analyse de plus gros peptides, mais provoque aussi une fragmentation variable⁴.

La fragmentation intense empêchait l'utilisation de la technique pour les mélanges de protéines. La technique de MS/MS a donc été utilisée pour analyser un seul ion du spectre de masse⁵.

Il y avait encore un intérêt pour une source d'ionisation qui pourrait ioniser de grosses protéines sans amener de fragmentation. Deux techniques modernes

répondirent à ces attentes : MALDI (Matrix assisted laser desorption ionisation)⁶ et ESI (Electrospray ionisation)⁷.

La source MALDI amène une ionisation douce en laissant une matrice répartir à l'échantillon l'énergie absorbée du laser.

La source ESI amène la nébulisation d'une solution contenant la protéine sous pression atmosphérique. Ce type d'ionisation douce amène la formation d'une distribution de pics dans les spectres de masse due aux différents états de charge que la protéine peut porter. L'interprétation en est donc compliquée et il est nécessaire d'utiliser un ordinateur pour les traiter. Ces méthodes de traitement automatique s'appellent méthodes de déconvolution ou simplement déconvolution.

Plusieurs groupes ont conçu des algorithmes pour y arriver. Mann et Fenn⁸ ont proposé un algorithme pour la détermination statistique de la masse de la protéine et un algorithme de déconvolution. Il y eut ensuite la conception d'un algorithme de déconvolution par entropie maximale⁹. Labowsky et Fenn¹⁰ ont ensuite publié une méthode d'analyse tridimensionnelle où la masse de l'adduit (espèce ionique ajoutée pour amener l'ionisation) est aussi incluse dans les calculs. Hagen et Monnig¹¹ ont proposé quant à eux un algorithme de corrélation multiplicative. Zhang et Marshall¹² ont conçu l'algorithme Zscore qui détermine la meilleure enveloppe de pics selon les rapports isotopiques.

Les algorithmes proposés dans la présente étude consistent tout d'abord à prétraiter les spectres afin d'enlever le maximum de faux pics par une détermination statistique du signal de fond¹³. Les algorithmes utilisés sont des processus itératifs où une seule des protéines de l'échantillon est détectée à la fois et ensuite soustraite du

spectre de masse afin de pouvoir recommencer le processus. Le premier algorithme, appelé Probablion, est une modification de l'algorithme de Zhang et Marshall¹², sauf que la meilleure enveloppe est déterminée par la somme des intensités des pics qui la composent. Le deuxième algorithme, appelé Dension, consiste à multiplier tous les pics par une série d'entiers et à déterminer la meilleure masse parmi les masses de haute densité de pics.

Les deux algorithmes sont ensuite testés et comparés en utilisant des spectres de masse simulés et des spectres de masse de vrais échantillons.

Chapitre II

Revue de la littérature

Historique de l'ESI

La source d'ionisation Electrospray couplée à la spectrométrie de masse a été proposée par Dole¹⁴ en 1968. Le but de l'étude présentée par Dole était de pouvoir ioniser des espèces polymériques présentes dans des peintures afin de les caractériser par spectrométrie de masse. Il proposa quand même un mécanisme pour l'ionisation qui impliquait l'évaporation complète du solvant d'une gouttelette. Dans la fin des années 70, Iribarne et Thomson^{15,16,17,18} proposèrent un nouveau mécanisme pour expliquer l'ionisation, celui-ci impliquait l'émission d'ions directement à partir de la gouttelette. L'ionisation Electrospray passa par contre inaperçue jusqu'à ce que Yamashita et Fenn¹⁹ ainsi que Aleksandrov et ses collaborateurs²⁰ en 1984 présentent un développement plus général de la technique. Fenn et al.²¹ élaborèrent sur le mécanisme de Iribarne et Thomson en 1990. Ils ajoutèrent la notion que le taux de désorption des ions dépend de la densité de charge sur la gouttelette. Fenn²² approfondit encore plus le modèle lorsqu'il publia que la forme de la protéine influence aussi le mécanisme d'ionisation.

Allure du spectre de masse des protéines en ESI

Les protéines se comportent de façon particulière lorsqu'elles sont ionisées par ESI. En effet, l'ESI amène pour les protéines la formation d'ions portant plusieurs charges et ce sans induire la fragmentation. Les ions sont formés par l'ajout d'une ou plusieurs espèces ioniques provenant du solvant (par exemple H^+ , Na^+ , ...),

appelées adduits, à la molécule de protéine neutre. Comme l'ajout d'un H^+ peut s'effectuer à plusieurs reprises sur la protéine, il existe une certaine distribution d'ions multichargés. Le spectre de masse qui en résulte présente alors une distribution de pics qui correspondent aux différentes charges que l'ion peut porter. La Figure 1 montre un exemple d'un spectre de masse d'une protéine obtenu par ESI où il est possible de voir la distribution d'ions multichargés.

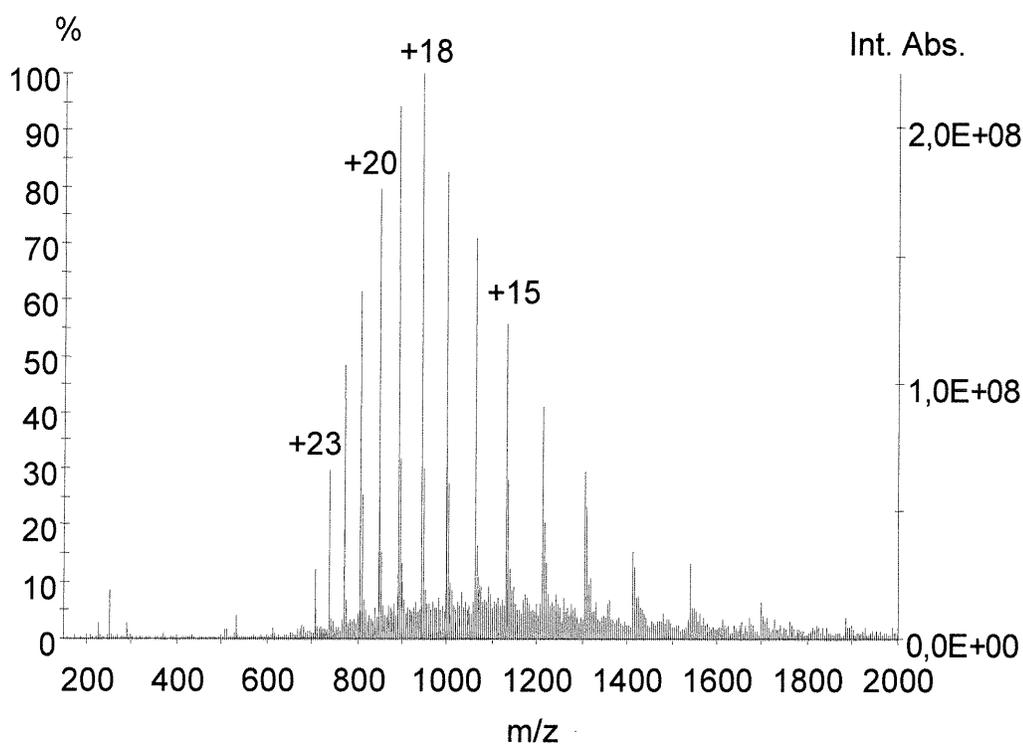


Figure 1 : Spectre de masse non filtré de la myoglobine (MM = 16953 Da) obtenu par ESI.

La distribution de pics en forme de cloche observée dans la Figure 1 est due au fait que l'ionisation produit une certaine distribution statistique des états de charge autour d'un état préféré.

Comme il y a plusieurs pics pour une même protéine, la détermination de la masse de cette protéine (MM) est plus précise. En effet, seulement deux pics sont

nécessaires pour déterminer la charge de ceux-ci donc la masse de la protéine. Les pics supplémentaires ne servent donc qu'à augmenter la précision de la mesure.

De plus, le fait que les protéines portent plusieurs charges réduit la gamme de masse à analyser, puisque la mesure effectuée par le spectromètre de masse est le rapport de m/z . Si la charge z augmente alors le rapport m/z diminue. Comme les m/z à mesurer sont plus petits, cela simplifie les analyses des protéines puisque les analyseurs de masse ont souvent des gammes de m/z limitées. Par contre, la dégénérescence des pics complique l'allure des spectres de masse des mélanges puisqu'il peut y avoir chevauchement des distributions de pics. L'analyse manuelle de ces spectres devient alors difficile. Il serait donc utile de développer une méthode automatisée pour extraire la masse des protéines d'un spectre de masse afin d'en simplifier l'allure tout en tirant profit de la dégénérescence des pics pour augmenter la précision de la mesure. Certaines de ces méthodes existent déjà et sont généralement appelées méthodes de déconvolution.

Interprétation manuelle des spectres de masse

Comme chaque pic de la distribution d'ions multichargés correspond à la même protéine, le rapport m/z d'un pic observé dans le spectre de masse répond à l'équation suivante pour le mode d'ionisation positif (le développement mathématique est similaire pour le mode négatif):

$$\frac{m}{z} = \frac{MM + nC^+}{n} \quad (\text{II.1})$$

où MM est la masse moléculaire de la protéine,

n est la charge portée par l'ion et

C^+ est la masse de l'adduit (mode positif).

En réarrangeant l'équation II.1, il est possible d'obtenir la relation qui donne la masse moléculaire de la protéine en fonction des autres paramètres.

$$MM = n \left[\frac{m}{z} - C^+ \right] \quad (\text{II.2})$$

Puisque n'importe quel couple de pics (i et j) de la distribution d'ions multichargés représente la même protéine, l'équation II.2 doit donner le même résultat pour les deux pics.

$$MM = n_k \left[\left(\frac{m}{z} \right)_k - C^+ \right] = n_j \left[\left(\frac{m}{z} \right)_j - C^+ \right] \quad (\text{II.3})$$

Or comme deux pics adjacents de la distribution ont une différence de charges d'une unité, l'équation II.3 peut être réarrangée pour donner l'équation suivante :

$$n_j = \frac{m_k - C^+}{m_j - m_k} \quad (\text{II.4})$$

où l'indice j correspond au pic ayant la charge la plus petite et le m/z le plus grand et

m_x est le rapport m/z observé.

Une fois la charge d'un pic déterminée, il est facile de trouver la masse moléculaire de la protéine en utilisant l'équation II.2.

Les distributions d'ions multichargés peuvent donc être interprétées manuellement, mais cette tâche devient rapidement difficile pour les mélanges de protéines où les distributions d'ions multichargés peuvent se chevaucher et où il est difficile de déterminer à quelle protéine appartient chaque pic.

Déconvolution des spectres de masse (approches existantes)

Les deux premiers algorithmes de déconvolution ont été conçus par Mann et Fenn⁸ en 1989. Le premier se nomme « Averaging algorithm » qui sera traduit par algorithme de détermination statistique de la masse de la protéine. Le deuxième a été nommé simplement algorithme de déconvolution.

Algorithme de détermination statistique de la masse de la protéine

Cet algorithme se classe plutôt parmi les méthodes de déconvolution semi-automatisées. En effet, les pics de la distribution d'ions multichargés doivent être extraits manuellement du spectre de masse et la charge de chaque ion doit être déterminée par l'utilisateur. La partie automatisée consiste donc à déterminer la masse de la protéine en utilisant toutes les combinaisons possibles de deux pics de la distribution d'ions multichargés. L'équation II.5 peut être obtenue en réarrangeant l'équation II.3.

$$\frac{K'_i}{K'_{i+b}} = 1 + \frac{b}{i} \quad (\text{II.5})$$

où K' sont les rapports m/z moins la masse de l'adduit des pics choisis,

i la charge du premier pic et

b la différence de charge entre les deux pics.

En réarrangeant l'équation II.5, il est possible d'obtenir une relation linéaire qui relie la différence de charge de deux ions à la charge d'un de ces ions.

$$\frac{1}{i} = \frac{1}{b} \left(\frac{K'_i}{K'_{i+b}} - 1 \right) \quad (\text{II.6})$$

Cette relation est utilisée pour tracer le graphique de la Figure 2. Les valeurs théoriques, c'est-à-dire les valeurs de K' qui devraient être obtenues pour répondre à l'équation II.6, sont représentées par la courbe continue. Les losanges représentent les valeurs expérimentales mesurées. Il est possible de juger de la pertinence d'un pic par l'étendue de la dispersion verticale des points autour de la valeur théorique d'une charge donnée. Plus les points pour une charge sont dispersés, moins le pic correspondant à cette charge est pertinent pour le calcul de la masse de la protéine. Un poids statistique est donc assigné à chacun des pics selon cette dispersion afin que ceux-ci aient une importance plus ou moins grande dans la détermination de la masse de la protéine.

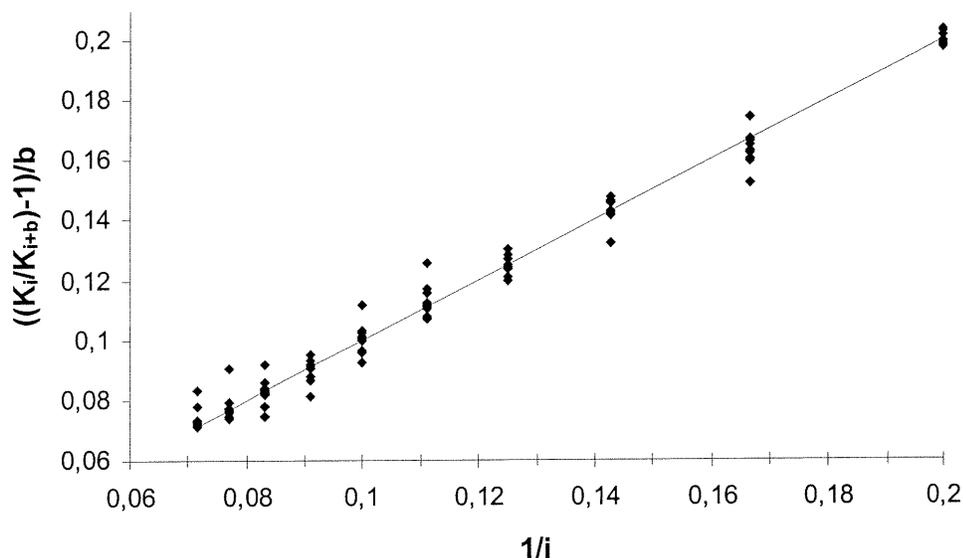


Figure 2 : Exemple d'utilisation de l'algorithme de détermination statistique de la masse de la protéine. La courbe continue représente les valeurs théoriques. (tiré de la référence 8)

Cet algorithme n'est pas une technique automatisée pour déconvoluer les spectres de masse de protéines. Il peut être utile dans le cas où la détermination des

pics appartenant à une protéine est simple, c'est-à-dire le cas où il y a un nombre réduit de protéines dans l'échantillon.

Algorithme de déconvolution (Mann et Fenn)

Cette méthode consiste à transformer le spectre de masse afin de rendre la détermination de la masse de la protéine plus facile. Le spectre transformé, appelé spectre déconvolué, consiste à porter en graphique l'intensité des molécules neutres en fonction de leur masse moléculaire. L'axe des x n'est donc plus le rapport m/z mais bien la masse moléculaire seulement qui se mesure en Da.

L'algorithme commence par tester une série de masses moléculaires déterminée par l'utilisateur. Lors du test d'une masse donnée, l'algorithme cherche dans le spectre de masse tous les pics qui pourraient porter une charge entière pour la masse moléculaire testée. La somme des intensités de tous les pics trouvés pour cette masse moléculaire est ensuite calculée. Cette intensité est reportée dans le spectre déconvolué et une nouvelle masse moléculaire est testée. Le pic du spectre déconvolué le plus intense devrait donc se situer à la masse moléculaire de la protéine. La Figure 3(haut) montre un spectre de masse simulé d'une protéine de 10000 Da où toutes les intensités sont égales et ce pour simplifier l'exemple. La Figure 3(bas) montre le spectre déconvolué obtenu grâce à l'algorithme de déconvolution développé par Fenn et al.

Cette méthode présente quelques limitations. En effet, comme le montre la Figure 3(bas), la méthode amène inévitablement la formation d'harmoniques, c'est à dire de faux pics à des valeurs multiples ou des fractions entières d'un vrai pic. Ceci

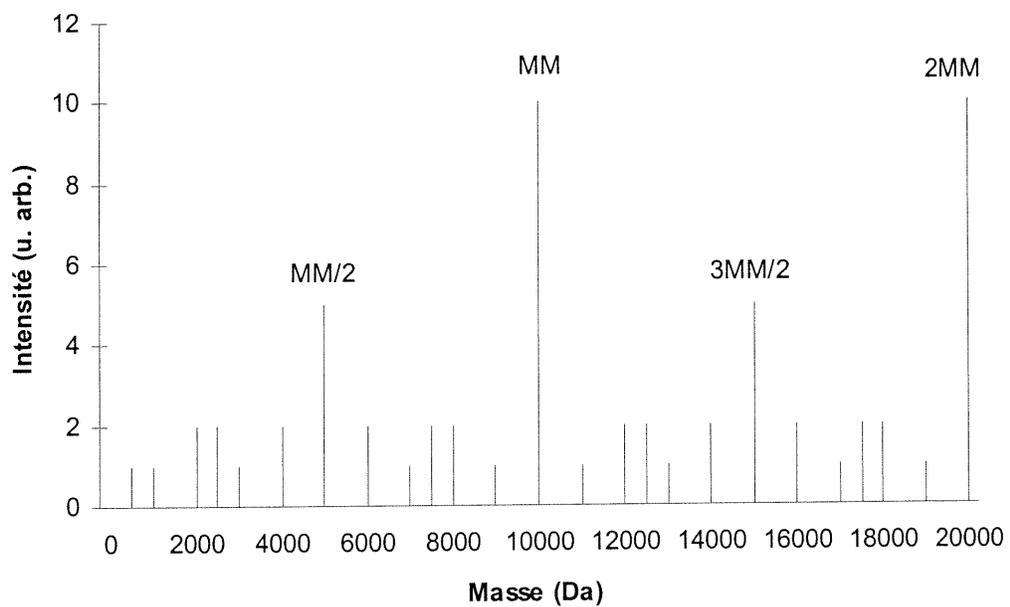
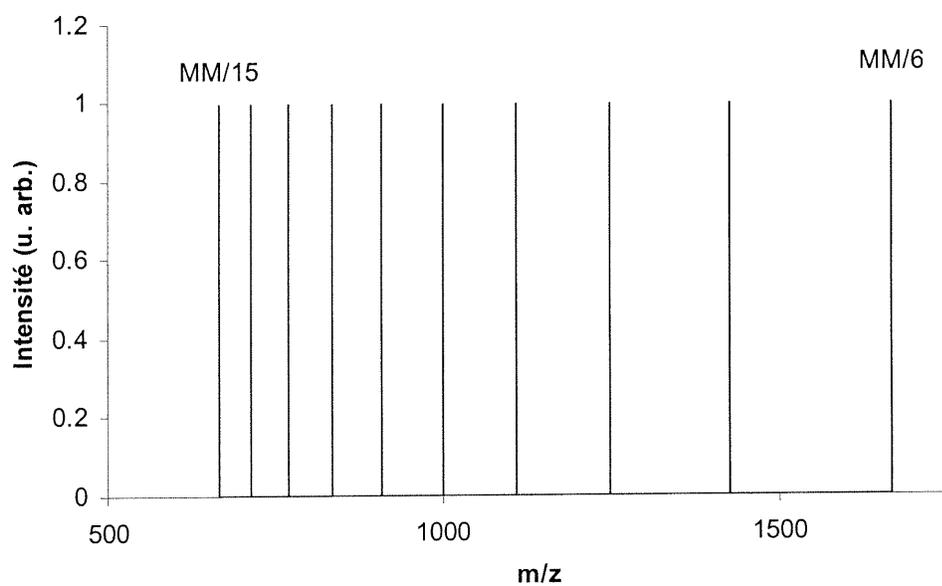


Figure 3 : Exemple de déconvolution par l'algorithme de Mann et Fenn. Haut) Spectre de masse simulé (10000 Da). Bas) Spectre déconvolué. (tiré de la référence 8)

est dû au fait qu'une distribution d'ions multichargés peut se voir assigner une masse deux fois plus grande que sa vraie masse en portant deux fois plus de charges.

De plus, la technique s'applique difficilement aux mélanges à cause de la grande quantité d'harmoniques créées dans le spectre déconvolué. En effet, il est difficile d'y différencier un vrai pic d'un faux pic.

La technique est aussi sensible au signal de fond contenu dans le spectre de masse. En effet, l'algorithme peut inclure par hasard les pics provenant du signal de fond dans de fausses distributions d'ions multichargés. Ceci a pour effet de créer des pics de protéines à de fausses masses.

Finalement, quoique la technique soit automatisée, l'utilisateur doit donner une gamme de masse dans laquelle l'algorithme doit chercher pour réduire la présence des harmoniques, la quantité de signal de fond traité et pour optimiser la vitesse de l'algorithme.

Entropie maximale

Reinhold et Reinhold ont proposé un algorithme d'entropie maximale⁹ en 1991. Cet algorithme teste aussi toutes les masses dans un intervalle donné. Pour chacune des masses de l'intervalle, l'algorithme produit une distribution modèle qu'il compare aux données expérimentales. La probabilité que tous les pics d'une distribution modèle correspondent aléatoirement aux pics du spectre de masse expérimental est faible. Quand une bonne correspondance se produit réellement, il est considéré que le système a une entropie faible (organisation élevée). L'entropie du système est mesurée par l'équation suivante :

$$I_p^2(v) = -\sum_{i=1}^N v_i \log\left(\frac{v_i}{p_i}\right) \quad (\text{II.7})$$

où $I_p^2(v)$ est l'entropie du système,
i est l'indice représentant chaque *m/z* du spectre,
v est la distribution modèle et
p est le spectre expérimental.

Une grande entropie est donnée au système si un pic présent dans la distribution modèle n'est pas dans le spectre expérimental. Ceci a pour effet, de réduire l'effet d'harmoniques observé avec l'algorithme de Mann et Fenn. En effet, une harmonique au double de la vraie masse sera représentée dans le spectre expérimental par seulement un pic sur deux ce qui lui donnera une grande entropie. Pour obtenir le spectre déconvolué, l'algorithme trace ensuite l'exponentielle de l'entropie du système par rapport à la masse testée.

Cet algorithme offre l'avantage de réduire les harmoniques et de pouvoir traiter les échantillons contenant plus d'une protéine si le comportement de celles-ci est comparable. En effet, l'algorithme crée une distribution modèle qu'il essaie de faire correspondre au spectre expérimental. Or, l'allure des distributions expérimentales dépend grandement des conditions expérimentales et de chaque protéine. Les distributions expérimentales ne sauraient donc être comparées à un modèle unique. Une protéine possédant plus ou moins d'états de charge que la norme réagira moins bien face à cet algorithme.

De plus, l'algorithme trace une mesure de la correspondance du spectre de masse avec une distribution modèle. L'algorithme ne peut donc pas être utilisé pour

la quantification mais seulement pour l'identification. En effet, la mesure d'entropie effectuée n'est aucunement une mesure de la quantité de protéines en présence dans l'échantillon.

Déconvolution tridimensionnelle

Labowsky et Fenn ont développé en 1992 une méthode de déconvolution tridimensionnelle¹⁰ qui est en fait une extension de l'algorithme de déconvolution de Mann et Fenn. En effet, l'algorithme de déconvolution est le même sauf qu'il n'y est pas supposé que l'adduit soit toujours un proton. Même si dans la majorité des cas le proton est l'adduit de la protéine, il est possible qu'il y en ait d'autres comme le sodium ou même un mélange de plusieurs adduits. L'algorithme effectue donc les calculs nécessaires pour tester toutes les combinaisons possibles de masses de la protéine et de l'adduit dans une certaine gamme de masse. Cette méthode permet donc de déterminer avec exactitude la masse de l'adduit et de mettre en évidence les erreurs possibles de calibration car la masse de l'adduit est connue. La masse de l'adduit étant déterminée plus exactement, la masse de la protéine mesurée sera elle aussi plus exacte.

Puisque l'algorithme utilisé pour la déconvolution est le même que celui de Mann et Fenn, il présente les mêmes limitations. C'est-à-dire la difficulté de traiter des mélanges de protéines et la présence d'harmoniques. De plus, le choix des gammes de masse est crucial à l'élimination du signal de fond et des harmoniques, ce qui rend cette partie de la méthode difficile à automatiser.

Corrélation multiplicative

Hagen et Monnig ont développé l'algorithme de corrélation multiplicative en 1994¹¹. Cet algorithme fait appel au principe que la multiplication d'une série de nombres donne un plus grand résultat que la somme de ceux-ci si les nombres sont également répartis. La somme d'une série de nombres ne donne aucune information sur la distribution de ceux-ci, mais la multiplication le peut. Par exemple, les nombres 1, 2 et 97 donnent une somme de 100 et un produit de 194. Or, les nombres 33, 33 et 34 donnent toujours la somme de 100 mais le produit est alors de 37026. L'algorithme tire donc profit du fait que les spectres de protéines en ESI sont répartis sur plusieurs états de charge. Les faux pics quant à eux seront la plupart du temps distribués majoritairement dans un état de charge. L'algorithme teste donc toutes les masses d'un intervalle donné en multipliant les intensités de tous les pics qui portent une charge entière pour une masse testée. L'équation II.8 montre le fonctionnement de l'algorithme.

$$D(MM) = \frac{\prod_{i_{\min}}^{i_{\max}} I\left(\frac{MM}{i} + C^+\right)}{rms^{nec}} \quad (\text{II.8})$$

où MM est la masse moléculaire testée,

D est l'intensité de cette masse dans le spectre déconvolué,

i_{\min} et i_{\max} sont les charges minimales et maximales,

C^+ est la masse de l'adduit,

I est l'intensité du pic dans le spectre de masse,

nec est le nombre d'états de charge trouvés et

rms est la racine carrée de la moyenne du carré des intensités du spectre de masse.

Le dénominateur de l'équation II.8 sert à annuler le fait que plus la masse testée est grande plus il est facile de trouver des pics dans le spectre.

Cet algorithme permet d'augmenter dramatiquement le rapport signal sur signal de fond du spectre déconvolué. De plus, il peut analyser des mélanges complexes de protéines. Par contre, l'intensité dans le spectre déconvolué varie exponentiellement avec la concentration de la protéine dans l'échantillon. Cette distorsion des pics a pour effet d'empêcher l'analyse des mélanges de protéines dispersées sur plusieurs ordres de grandeur de concentration. En effet, les protéines de faibles concentrations seront noyées par celles de grandes concentrations à cause de la dépendance exponentielle. L'algorithme amène aussi la création d'harmoniques mais dans de plus faibles quantités que l'algorithme de Mann et Fenn.

Zscore

Zhang et Marshall ont développé Zscore en 1997¹². Cet algorithme est le premier qui utilise un processus itératif. C'est-à-dire qu'il analyse les pics d'une protéine à la fois et les soustrait ensuite du spectre de masse avant de recommencer le processus. ZScore détermine le maximum du spectre et lui assigne successivement toutes les charges dans un certain intervalle. Pour chaque charge, il trouve tous les pics que la distribution contiendrait si la charge assignée était la bonne. Plus il y a de pics dans la distribution, plus le pointage de cette charge est grand. Le pointage est aussi basé sur la découverte de distributions isotopiques autour de chacun des pics.

La charge ayant le plus grand pointage est considérée comme étant celle faisant partie de la bonne distribution.

Si la résolution n'est pas suffisante pour permettre de distinguer les distributions isotopiques, l'algorithme donne un pointage à la distribution selon la découverte d'autres pics dans la distribution seulement. Une fois la meilleure distribution d'ions multichargés déterminée, celle-ci est soustraite du spectre et le processus est recommencé.

Cet algorithme peut, grâce au processus itératif, traiter plus facilement les mélanges complexes de protéines. L'algorithme donne de meilleurs résultats dans le mode à haute résolution. Cela fait par contre appel à des appareils à haute résolution qui sont généralement dispendieux et peu accessibles.

De plus, aucun traitement préalable n'est effectué pour enlever le signal de fond. Ce qui fait que l'algorithme doit donc demander à l'utilisateur de choisir un certain seuil pour l'empêcher de chercher inutilement des pics dans le signal de fond. Ceci rend l'automatisation de la technique assez difficile et le choix de l'utilisateur est crucial à la déconvolution.

Algorithmes suggérés

Deux algorithmes ont été conçus pour résoudre le problème analytique mis en évidence dans cette étude tout en essayant d'éviter les limitations des algorithmes existants.

Les deux algorithmes de déconvolution proposés font appel à un algorithme de prétraitement des données afin d'enlever le maximum de signal de fond des spectres de masse. La filtration des données est basée sur le fait que l'intensité du

signal de fond dans des spectres de masse accumulés est l'intensité ayant la fréquence d'apparition la plus élevée. En déterminant l'intensité ayant la plus grande fréquence, il est possible de le soustraire du spectre de masse et de réduire par le fait même la quantité de signal de fond. Cette étape devrait permettre de réduire le temps de calcul puisqu'une plus petite partie du spectre est traitée. Elle devrait aussi réduire la quantité de faux pics puisque les algorithmes n'essaieront pas de trouver des structures à des endroits où il n'y en a pas.

Les deux algorithmes de déconvolution sont des processus itératifs, c'est-à-dire qu'ils effectuent une série d'opérations qu'ils répètent à plusieurs reprises. Chaque itération correspond au traitement des pics d'une protéine qui sont ensuite soustraits du spectre de masse avant que la prochaine itération ne commence. Le fait que les algorithmes soient des processus itératifs devrait permettre de traiter les mélanges efficacement.

Les deux algorithmes sont automatisés faisant appel à un minimum de manipulation par l'utilisateur.

Le premier algorithme est appelé algorithme de détermination des distributions les plus probables ou plus simplement Probablion. Celui-ci est une modification de l'algorithme de Zhang et Marshall. La première différence est que la meilleure enveloppe de pics est déterminée par la somme des intensités des pics qui la composent et non la distribution isotopique autour des pics. Ceci a pour effet de permettre l'analyse à la fois de spectres de masse de faible résolution ou de haute résolution. La deuxième différence est la façon d'utiliser l'algorithme. Zhang et Marshall utilisent leur algorithme sur un spectre de masse unique seulement.

L'algorithme proposé peut soit traiter un spectre de masse moyen ou soit traiter les spectres de masse individuellement et faire la moyenne des spectres déconvolués. Ceci a pour effet de réduire l'intensité du signal de fond dans le spectre déconvolué puisque le signal de fond étant aléatoire celui-ci n'est pas accumulé aussi efficacement que les vrais signaux.

Le deuxième algorithme proposé est appelé algorithme de détermination des positions de haute densité de masse ou plus simplement Dension. Celui-ci consiste à multiplier tous les pics du spectre de masse par une série d'entiers pour obtenir un spectre similaire à ceux obtenus par Mann et Fenn (Figure 3). Les pics d'une même distribution devraient alors se regrouper à la masse de la protéine et former une zone de haute densité de masse. La masse de haute densité ayant l'intensité la plus grande est alors reportée dans le spectre déconvolué et tous les pics qui y correspondent sont soustraits du spectre avant que l'algorithme ne commence une autre itération. Cet algorithme offre l'avantage de ne rien assumer. En effet, toutes les données sont traitées en même temps afin de les laisser s'organiser elles-mêmes. La méthode de Zhang et Marshall assigne une charge au pic maximum du spectre de masse et suppose que ce pic est bon. Le reste de la distribution est ensuite déterminé grâce à cette supposition initiale. Si le pic n'est pas un vrai pic, cela peut causer des problèmes.

Chapitre III

Partie Théorique

Principes de base

La spectrométrie de masse consiste à ioniser les espèces composant un échantillon et à mesurer, sous vide, la masse caractéristique des ions formés.

Lors de l'ionisation d'une espèce, une certaine quantité d'énergie lui est attribuée. Celle-ci peut provenir de collisions, de l'effet d'un champ électrique ou même de l'absorption d'énergie d'une source lumineuse. L'ionisation d'une espèce donnée exige une quantité d'énergie qui lui est propre. L'énergie excédentaire à la suite de l'ionisation est alors redistribuée dans le reste de la structure de la molécule. Si cette énergie est trop grande, cela peut amener la molécule ionisée à se fragmenter sous l'effet du stress supplémentaire qui lui est imposé. Ce qui cause la formation d'ions de plus faible masse. Ces fragments peuvent être utiles à la détermination de structure d'espèces inconnues. Par contre, cela peut nuire à l'analyse de mélange puisque les spectres en sont grandement compliqués. Il peut donc être utile d'utiliser une méthode d'ionisation douce qui fournit juste assez d'énergie pour induire l'ionisation mais pas la fragmentation.

La détection de la masse des ions se fait grâce à l'analyseur de masse. Le fait que les ions portent une charge permet de les différencier des espèces neutres. Pour ce faire, il est nécessaire de tirer profit de l'interaction de la charge des ions avec les champs électriques ou magnétiques. En effet, ces champs agissent sur les ions tout en ignorant les espèces neutres. Comme l'effet des champs électriques et magnétiques sur les ions dépend à la fois de la charge(z) et de la masse(m) des ions, il est

impossible de dissocier complètement ces deux paramètres. Il est par contre possible de mesurer le rapport de m sur z . En balayant les champs électriques et magnétiques, il est possible de mesurer successivement toutes les valeurs de m/z et d'obtenir un spectre de masse.

Principe de l'ionisation Electrospray

Il existe plusieurs méthodes d'ionisation utilisées en spectrométrie de masse. L'ionisation Electrospray (ESI) est une des méthodes d'ionisation douces. Contrairement à plusieurs des autres méthodes, l'ESI se produit à pression atmosphérique et l'analyte doit être en solution. La Figure 4 montre le processus menant à l'ionisation des espèces en solution.

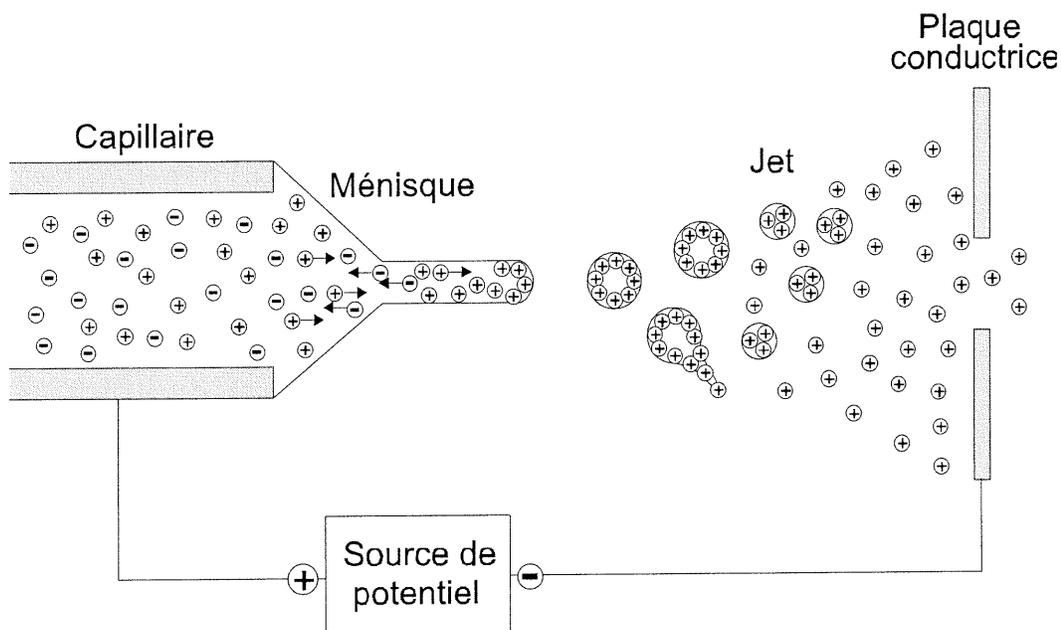


Figure 4 : Schéma du montage permettant l'ionisation Electrospray²³.

Le montage consiste en un capillaire de faible diamètre et d'une plaque conductrice placée à faible distance du capillaire. Une différence de potentiel est

appliquée entre le capillaire et la plaque conductrice. Le capillaire sert à introduire l'échantillon dans la source.

Pour obtenir des ions positifs en phase gazeuse, il faut appliquer un potentiel positif au capillaire et négatif à la plaque conductrice. Ceci a pour effet d'attirer les cations vers la plaque, c'est-à-dire le bout du capillaire où ceux-ci s'accumulent. La répulsion des charges positives ainsi que la tension de surface du solvant amènent alors la formation d'un ménisque dont la surface est chargée positivement. Lorsque la répulsion des charges devient trop importante, la tension de surface n'est plus suffisante pour retenir le solvant au bout du capillaire. Il y a alors nébulisation c'est-à-dire éjection de gouttelettes de solvant sous forme de jet. La surface des gouttelettes formées est alors chargée positivement. L'évaporation du solvant amène les gouttelettes à devenir plus petites. Lorsque la répulsion des charges devient encore une fois trop grande pour être contre balancée par la tension de surface, les gouttelettes émettent un nouveau jet de gouttelettes plus petites. Ce processus continue jusqu'à ce que l'émission d'espèces à l'état gazeux se produise.

Le mécanisme qui permet de faire passer une espèce à l'état gazeux à partir des gouttelettes est par contre peu connu et il existe plusieurs hypothèses. Deux mécanismes ont tout d'abord été proposés^{14,15}, mais aucun des deux n'a pu être démontré expérimentalement sans équivoque. Le premier mécanisme impliquerait l'évaporation complète du solvant ce qui donnerait une protéine ionisée en phase gazeuse. L'autre mécanisme proposé par Iribarne et Thomson impliquerait une émission de la protéine ionisée à l'état gazeux directement à partir d'une petite gouttelette surchargée. Ces deux mécanismes ont de la difficulté à expliquer les

différents états de charge observés. Fenn et al.²¹ élaborèrent alors sur le modèle de Iribarne et Thomson en 1990. Fenn²² suggéra finalement un autre mécanisme où la charge d'un ion dépend de la densité de charge sur la gouttelette et la surface qu'il traverse quand il est émis. Une fois à l'état gazeux, les espèces sont introduites dans l'analyseur de masse par le trou de la plaque conductrice.

Mécanisme de l'ionisation Electrospray

Dole et al.¹⁴ ont suggéré que l'émission de la protéine à l'état gazeux se ferait par l'évaporation complète du solvant d'une gouttelette comprenant une seule protéine. Ce mécanisme peut difficilement expliquer le fait que certaines protéines possèdent des charges unitaires. Or le mécanisme suggère que l'évaporation amènerait une concentration des charges sur la gouttelette et donc sur la protéine. Thomson et Iribarne¹⁵ ont suggéré qu'une fois le rayon de la gouttelette suffisamment faible, celle-ci pouvait émettre directement les ions à l'état gazeux. Ce modèle ne pouvait pas, tel quel, expliquer les différents états multichargés observés expérimentalement.

Fenn et al.²¹ ont alors continué le développement de ce mécanisme. Ceux-ci en conclurent que le taux de désorption des ions de la gouttelette et la charge qu'ils portent dépendent de la densité de charge sur celle-ci. Au début, la densité est faible et seuls les ions de faibles charges sont émis à un taux faible. À mesure que la densité de charge devient de plus en plus grande, les ions sont émis de plus en plus rapidement et avec des charges de plus en plus grandes. Comme en se faisant, la gouttelette s'appauvrit en protéine, il arrive un moment où il devient de plus en plus difficile d'émettre des ions. Par contre, la charge portée par les ions émis continuera

à augmenter puisque la densité de charge continue à augmenter. Ce modèle réussit donc à expliquer la distribution d'ions multichargés observée expérimentalement.

Après cette étude, Fenn²² en est venu à suggérer un autre mécanisme où la distance entre les charges sur la protéine dépend de la distance entre les charges sur la gouttelette lorsque la protéine traverse sa surface. Ce qui veut dire que non seulement la densité de charge de la gouttelette détermine la charge des ions, mais aussi la forme de la protéine. En effet, la forme de la protéine change la superficie de la surface de la gouttelette qu'elle traverse lors de son émission.

Facteurs influençant la protonation des protéines

Il existe plusieurs paramètres expérimentaux qui peuvent modifier la distribution de pics en ESI. Ils se divisent en deux catégories : paramètres de la solution et paramètres instrumentaux.

Un changement du pH de la solution avant le passage dans l'ESI peut avoir des effets importants. En effet, en changeant le pH, il est possible de dénaturer la protéine. Cette dénaturation peut amener des sites à devenir plus accessibles à la protonation. En effet, la surface effective de la protéine devenant plus grande, elle peut accumuler plus de charges lors de l'ionisation. Le pH semble ne pas avoir d'autres effets sur l'ionisation comme le montre une étude²⁴. En effet, l'ionisation de petites protéines y est peu affectée par le pH de la solution. Comme ces petites protéines sont trop petites pour être repliées, la dénaturation n'est pas possible et le pH n'a aucun effet important. Deux autres études^{25,26} montrent même que l'obtention d'ions positifs est possible à de hauts pH où la protéine porte une charge globalement négative.

Le solvant utilisé pour faire la solution a aussi un effet sur la distribution des pics multichargés. En effet, plus la constante diélectrique du solvant est grande, plus le solvant peut stabiliser des analytes possédant une grande quantité de charges. Ces solvants amènent donc un déplacement des distributions vers les hautes charges.

La concentration de l'analyte a aussi un effet sur les distributions. L'abondance d'un ion varie linéairement avec la concentration pour les faibles concentrations d'analyte. Le signal de l'analyte plafonne ensuite et peut même diminuer à de hautes concentrations. Une étude²⁷ a montré que les sensibilités des ions de différents états chargés ne sont pas toutes les mêmes. L'augmentation de concentration amène un déplacement des distributions vers de plus faibles charges. En effet, les ions avec beaucoup de charges ont une plus grande sensibilité à la concentration et ils plafonnent plus rapidement. Si la concentration augmente trop, ces états chargés diminuent d'intensité jusqu'à disparaître pour laisser la place aux ions avec de faibles charges qui eux n'ont pas encore plafonné.

La température a aussi un effet sur les distributions d'ions multichargés. En effet, celle-ci agit de la même façon que le changement de pH, c'est-à-dire que l'augmentation de température dénature graduellement la protéine et les sites de protonation cachés à l'intérieur de la protéine deviennent de plus en plus accessibles.

Les paramètres instrumentaux peuvent aussi avoir une influence sur la distribution d'ions multichargés. L'ESI est par contre indépendante de l'analyseur de masse auquel elle est couplée.

Le flux de gaz circulant autour du capillaire influence la position des distributions d'ions multichargés. En effet, une étude²² a montré que l'augmentation

du flux de gaz permettait d'obtenir des distributions d'ions multichargés déplacées vers les bas m/z . Ce paramètre favoriserait la désorption des ions et augmenterait donc le taux de désorption des ions portant de faibles charges. L'appauvrissement de la gouttelette se produit ainsi plus tôt et la désorption des ions portant de fortes charges est réduite.

Le fait d'appliquer un potentiel à la sortie de la source peut aussi influencer la charge des ions. En effet, une étude²⁸ a montré que ceci amène les ions à perdre des charges, ce qui déplace la distribution d'ions multichargés vers les plus hauts m/z .

Perturbation du spectre par les mélanges et le signal de fond

L'interprétation des spectres de masse de protéines est compliquée par le signal de fond et la présence de plusieurs protéines à la fois.

Le signal de fond peut nuire de deux façons à l'interprétation des spectres de masse. Pendant l'acquisition des données de cette étude, l'ordinateur a simplifié les données brutes en effectuant une détection de pics. La détection de pics calcule l'aire des pics et leur m/z pondéré. Ce sont ces résultats qui sont observés dans le spectre de masse. Cette méthode trouve inévitablement des pics dans le signal de fond. Ceci amène donc la création de faux pics qui se présentent à des positions aléatoires d'un spectre de masse à l'autre. Les faux pics obtenus sont non négligeables et peuvent se confondre aux pics d'une distribution d'ions multichargés réelle.

L'accumulation est une pratique courante en spectrométrie de masse et consiste en l'obtention de la moyenne de plusieurs spectres de masse. Les faux pics provenant du signal de fond des spectres de masse individuels compliquent les spectres de masse accumulés. En effet, l'accumulation amène la formation d'un

signal de fond continu. Ceci est dû au fait que les faux pics se trouvent à des valeurs de m/z aléatoires. Dans ces spectres de masse accumulés, l'intensité du signal de fond a habituellement une distribution plus ou moins normale autour d'une intensité moyenne, c'est-à-dire que cette intensité moyenne est généralement l'intensité la plus fréquente. La variation aléatoire du signal de fond autour de l'intensité moyenne est appelée le bruit. La valeur moyenne de ce signal de fond continu est aussi dépendante du rapport m/z et change avec celui-ci. Le signal de fond complique donc l'interprétation puisque la quantité de faux pics est beaucoup plus grande. Ceci a pour effet d'augmenter la quantité de pics à traiter lors de la déconvolution.

Le fait d'avoir plusieurs protéines complique grandement les spectres de masse. En effet, il est alors possible que deux protéines aient des distributions d'ions multichargés qui se chevauchent. Il est alors difficile de différencier visuellement les deux distributions et d'assigner les bonnes charges aux pics. Plus le nombre de protéines présentes dans un échantillon est grand, plus la probabilité que des distributions d'ions multichargés se chevauchent est grande.

Si les spectres de masse ne présentaient aucun signal de fond et que l'échantillon était composé d'une seule protéine, il serait alors facile d'attribuer les charges aux pics de la distribution d'ions multichargés. Comme la réalité est différente de la théorie, il est nécessaire de contourner ces deux problèmes qui nuisent à l'interprétation des spectres de masse de protéines afin d'obtenir de bons résultats.

Chapitre IV

Partie Expérimentale

Appareils utilisés

Deux spectromètres de masse ont été utilisés pour prendre les spectres expérimentaux qui seront traités lors de la présente étude. Le premier appareil est le modèle LCQ de la compagnie Finnigan. Il consiste en une source d'ionisation de type Electrospray couplée à une trappe d'ions comme analyseur de masse. L'introduction de l'échantillon a été faite par infusion directe dans la source à un débit de 5 $\mu\text{L}/\text{min}$. La phase mobile était composée à 60% (v/v) d'acétonitrile et 40% (v/v) d'eau. De l'acide trifluoroacétique (0,1% v/v) a aussi été ajouté à la phase mobile.

Le deuxième appareil utilisé est le modèle TSQ 7000 de Finnigan. Celui-ci consiste en une source d'ionisation de type Electrospray couplée à un triple quadrupole comme analyseur de masse. L'introduction de l'échantillon a été faite par infusion directe dans la source à un débit de 10 $\mu\text{L}/\text{min}$. Deux phases mobiles ont été utilisées lors de ces analyses. Une des phases mobiles était composée à 60% (v/v) d'eau et 40% (v/v) d'acétonitrile en plus de 0,1% (v/v) d'acide trifluoroacétique. L'autre phase mobile était composée à 50% (v/v) d'eau et 50% (v/v) de méthanol. De l'acide acétique (0,5% p/v) a été ajouté à cette phase mobile.

Protéines utilisées

Les protéines utilisées pour préparer les échantillons proviennent de la compagnie Sigma-Aldrich. Leur nom et masse sont montrés dans le Tableau I.

Tableau I : Masse des protéines utilisées lors de l'analyse.

Protéines	Masse (Da)
Insuline (pancréas boeuf)	5734
Ubiquitine (globule rouge boeuf)	8565
Cytochrome C (coeur boeuf)	12230
Myoglobine (muscle squelettique cheval)	16952

Langage de programmation

Le langage utilisé pour programmer les algorithmes proposés est le Visual Basic 6.0. Certaines des bibliothèques utilisées ont été programmées en Power Basic 6.0. Les spectres de masse simulés ont été créés avec le logiciel MatLab 5.3 version étudiante grâce à un programme développé dans le laboratoire. L'ordinateur utilisé pour la programmation est un Intel Pentium III 700 MHz avec 256 mégaoctets de mémoire vive.

Chapitre V

Résultats

Traitement du signal de fond

Une analyse statistique du signal peut permettre de filtrer un spectre pour enlever le maximum de faux pics et d'enlever le maximum de signal de fond. Pour pouvoir faire une bonne analyse statistique, le spectre doit avoir des pics à presque toutes les valeurs de m/z . Cette méthode de filtration n'est donc utile que dans ce type de situations. Le phénomène de pics à chaque m/z est observé normalement avec l'ionisation FAB à cause de la contamination du spectre par la matrice utilisée pour faciliter l'ionisation. L'application de l'algorithme de filtration a été réussie pour ce type d'ionisation¹³. Dans le cas de l'ESI, ce phénomène est observé s'il y a accumulation de spectres de masse, c'est-à-dire si la moyenne de plusieurs spectres de masse est effectuée.

Il est impossible pour les spectres de masse ayant une grande variation du signal de fond d'utiliser un seuil fixe pour l'enlever. En effet, le seuil nécessaire pour enlever la totalité du signal de fond consisterait à la valeur maximale de celle-ci. Ceci amènerait automatiquement une sur-correction du signal de fond dans les parties du spectre où il est moins important. Une partie du signal d'importance pourrait alors être irrémédiablement perdue. L'algorithme de filtration doit donc tenir compte de cette variation du signal de fond.

L'algorithme utilisé pour effectuer un filtrage par analyse statistique du spectre de masse fait la supposition que le signal de fond présente toujours l'intensité

la plus fréquente puisque celui-ci a une distribution plus ou moins normale. L'algorithme consiste à séparer le spectre de masse en intervalles de m/z de longueur fixe. Un exemple de détermination de l'intensité la plus fréquente d'un intervalle de d'un spectre de masse simulé est montré dans la Figure 5.

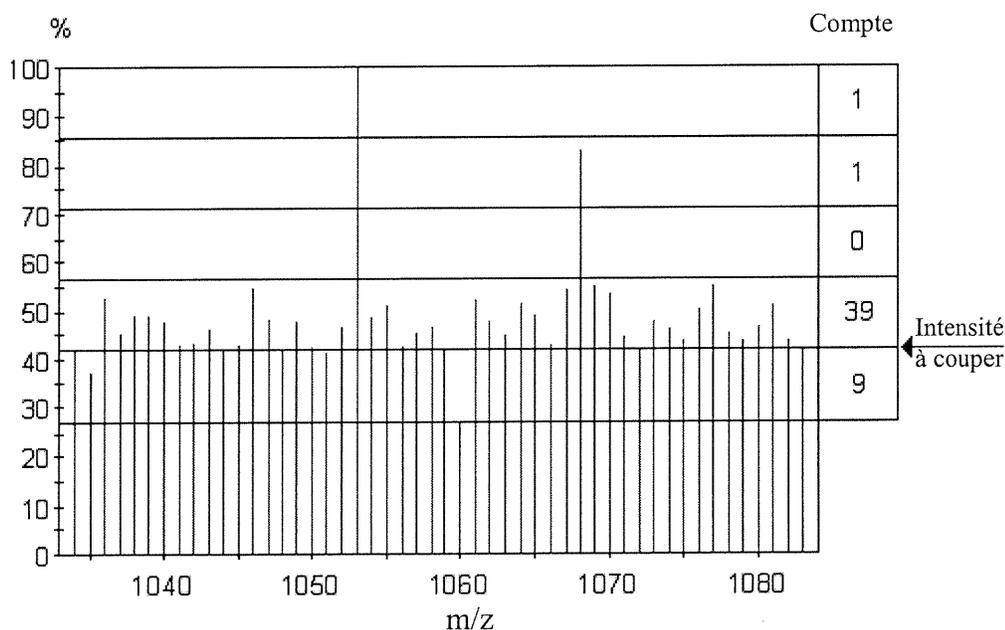


Figure 5 : Exemple d'utilisation de l'algorithme de filtration pour trouver l'intensité la plus fréquente d'un intervalle de m/z

Pour chacun de ces intervalles, l'algorithme classe chaque intensité des pics de l'intervalle afin de trouver l'envergure de l'intensité la plus fréquente. Cette valeur est alors enlevée à l'intensité de chacun des pics de cet intervalle. Dans le cas où il y aurait dégénérescence de l'intensité la plus fréquente, seule l'intensité ayant la valeur la plus faible est soustraite afin de s'assurer de toujours enlever le minimum de signal. Pour empêcher des problèmes liés aux frontières des intervalles, l'algorithme effectue l'analyse à plusieurs reprises en utilisant différentes longueurs d'intervalle. La valeur à soustraire à l'intensité d'une valeur de m/z donnée est

obtenue en calculant la moyenne des intensités à soustraire pour ce m/z à travers les différentes analyses effectuées.

Cet algorithme permet d'enlever un très grand nombre de pics provenant du signal de fond. Ceci accélère la déconvolution du spectre de masse puisqu'une grande partie du signal n'aura pas besoin d'être traitée.

Le signal de fond amène la formation de pics non négligeables dans les spectres de masse individuels à cause de la détection automatique des pics lors de l'acquisition. Ceux-ci sont impossibles à différencier des vrais signaux. L'accumulation est donc nécessaire pour pouvoir les éliminer. En effet, l'accumulation amène l'apparition d'un signal de fond continu à cause de la distribution aléatoire des faux pics, provenant du signal de fond des spectres de masse individuels. Ce signal de fond continu peut alors être différencié des vrais signaux à cause de sa continuité et peut être éliminé. Il est donc possible d'enlever le signal de fond d'un spectre de masse, mais seulement de façon indirecte.

En effet, la filtration peut être effectuée sur le spectre de masse accumulé seulement. L'accumulation des spectres de masse offre par contre un avantage majeur. Comme elle consiste en la moyenne de plusieurs spectres de masse, elle amène les pics provenant du signal de fond à diminuer d'intensité puisqu'ils ne sont pas accumulés aussi efficacement que les signaux d'intérêts. Ceci est dû au fait que les faux pics changent de m/z d'un spectre de masse à l'autre, mais les vrais pics quant à eux restent sensiblement aux mêmes m/z .

La Figure 6(haut) montre un spectre de masse simulé. La simulation reproduit un spectre de masse accumulé de la myoglobine (MM = 16953 Da) d'où l'addition d'un signal de fond continu. Par contre, aucun bruit n'a été ajouté à ce signal de fond continu. Ce spectre a été traité avec l'algorithme de filtration et la Figure 6(bas) montre le résultat obtenu. Il est possible d'y voir que le signal de fond continu en forme de distribution de Poisson est presque complètement disparu. Bien sûr, les spectres de masse expérimentaux n'ont pas des signaux de fond aussi lisses car il y a toujours un bruit présent. Mais, il est quand même possible de voir que l'algorithme peut faire automatiquement ce que l'oeil aurait permis de faire manuellement.

Déconvolution

Il a été montré que la déconvolution des spectres de masse de protéines peut être faite de plusieurs façons. Les méthodes utilisées au cours de cette étude ont été inspirées par d'autres algorithmes. La filtration automatisée des spectres de masses n'a par contre été utilisée dans aucune de ces méthodes jusqu'à présent. Or celle-ci peut probablement améliorer grandement les résultats obtenus puisqu'une partie non significative du signal a été enlevée et ne sera pas traitée. Dans les autres méthodes, le signal de fond est traité en même temps que le reste, ce qui contamine le spectre déconvolué résultant ou complique l'analyse. Même si la filtration aide à enlever les faux pics et aussi les harmoniques de vrais pics, il reste toujours une partie du signal de fond qui ne peut pas être enlevée.

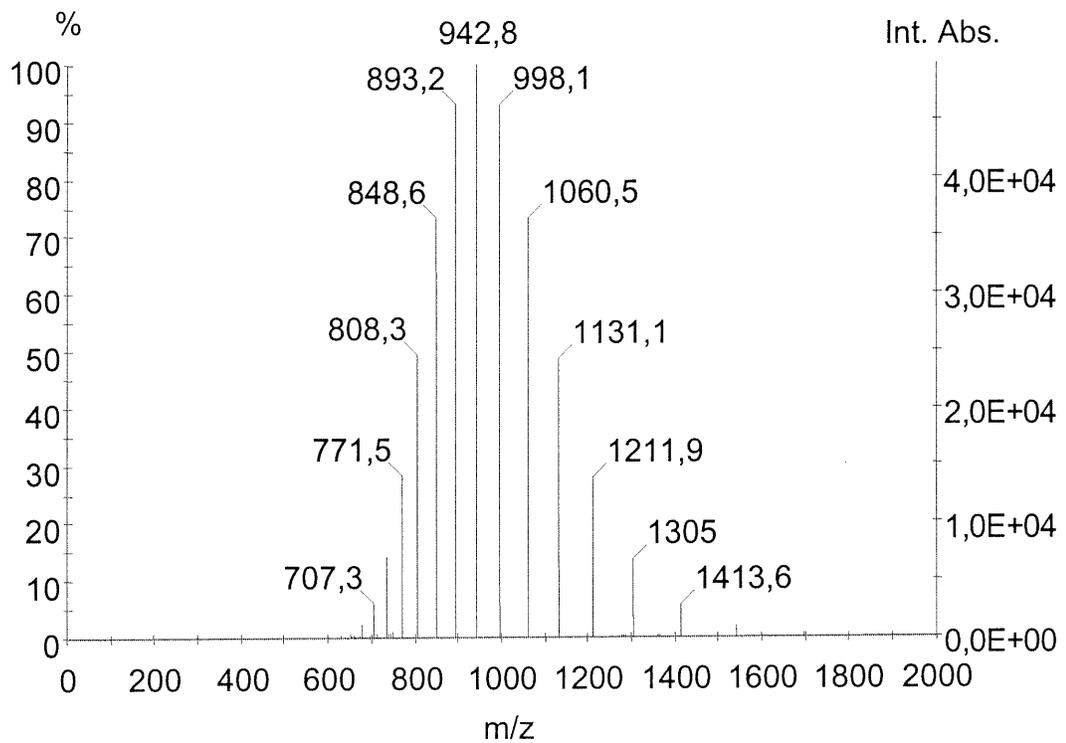
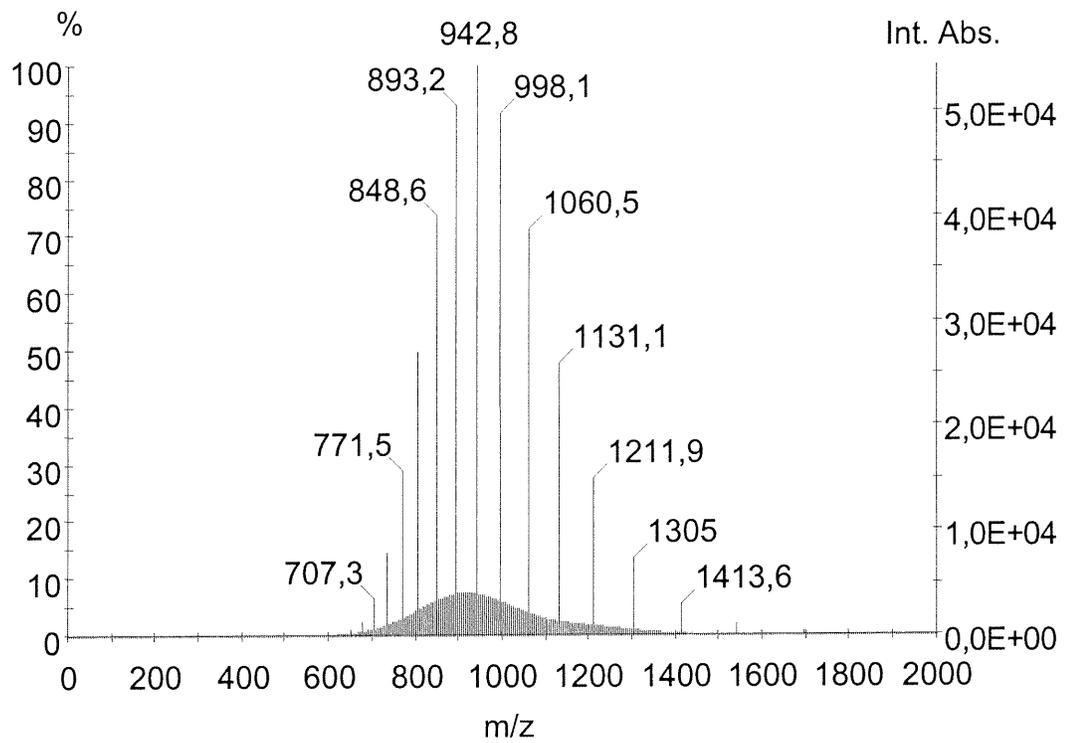


Figure 6 : Exemple de filtration d'un spectre de masse simulé (haut) . Le signal de fond continu a été enlevé avec succès dans le spectre de masse filtré (bas).

Quelques expériences ont été tentées pour enlever ces faux pics du spectre déconvolué. Il a tout d'abord été tenté d'utiliser la forme caractéristique des distributions d'ions multichargés pour discriminer les bonnes distributions des mauvaises. En effet, les intensités des distributions d'ions multichargés ont presque toujours l'apparence d'une distribution de Poisson (gaussienne non symétrique). La technique consiste donc à essayer de faire correspondre une fonction de Poisson (gaussienne modifiée) aux intensités de la distribution en déterminant les paramètres de la fonction testée par une optimisation de type simplex.

Cette technique donne généralement de bons résultats et peut différencier efficacement les distributions dues au signal de fond et aux harmoniques de celles dues à un vrai signal. Par contre, il a été observé que dans les cas où les distributions d'ions multichargés n'avaient pas la forme d'une distribution de Poisson, la technique réagissait de façon inadéquate. En effet, lors de ces situations, la distribution était simplement éliminée puisque la technique n'était pas suffisamment flexible. La Figure 7 montre un exemple d'une telle situation. Les pics de la distribution d'ions multichargés de l'ubiquitine semblent plutôt être une combinaison de distributions de Poisson non résolues.

Dans ce cas, la technique ne peut pas faire correspondre la fonction test à la distribution et le pic est rejeté du spectre déconvolué même s'il devait être le plus important. Cette limitation majeure de la technique a causé le rejet de la technique à cause du risque trop élevé de rejeter de bons pics. Mieux vaut conserver quelques mauvais pics que de rejeter des pics importants. La filtration reste donc le moyen le plus sûr de réduire la quantité de faux pics dans le spectre.

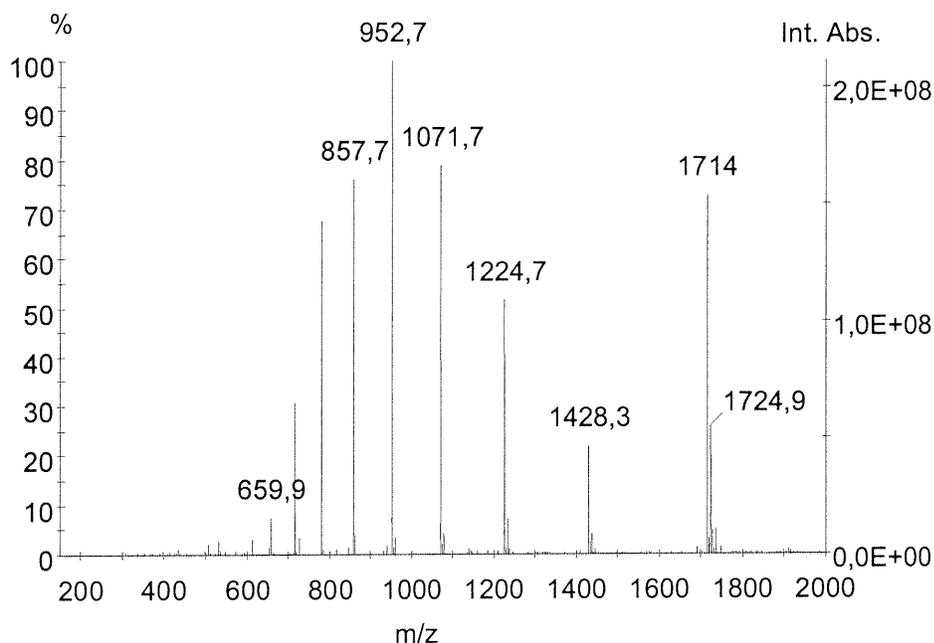


Figure 7 : Exemple d'une distribution d'ions multichargés qui n'a pas la forme d'une distribution de Poisson.

Détermination des distributions d'ions multichargés

Ce qui différencie les deux algorithmes de déconvolution conçus lors de la présente étude est la stratégie pour repérer les distributions d'ions multichargés dans les spectres de masse. Le premier algorithme consiste à déterminer la distribution d'ions multichargés la plus probable pour un pic donné. Celle-ci est la distribution qui présente la somme des intensités la plus grande. Le deuxième algorithme consiste à repérer la masse de haute densité de pics ayant l'intensité la plus grande.

Détermination des distributions les plus probables

La déconvolution par détermination des distributions les plus probables consiste en un algorithme qui a été inspiré par l'algorithme de Zhang et Marshall¹². L'algorithme, appelé Probablion, est résumé par l'organigramme présenté à la Figure 8.

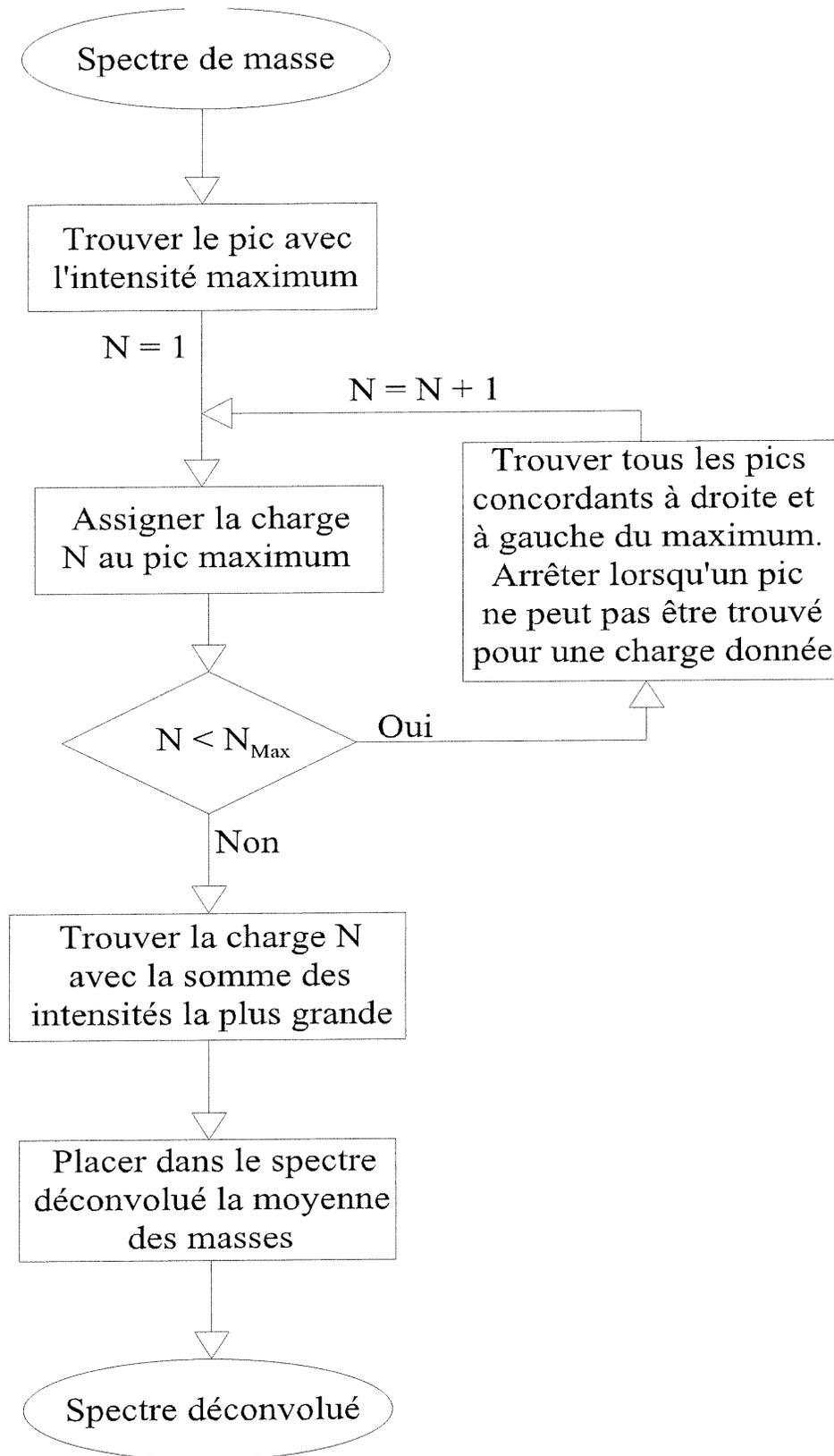
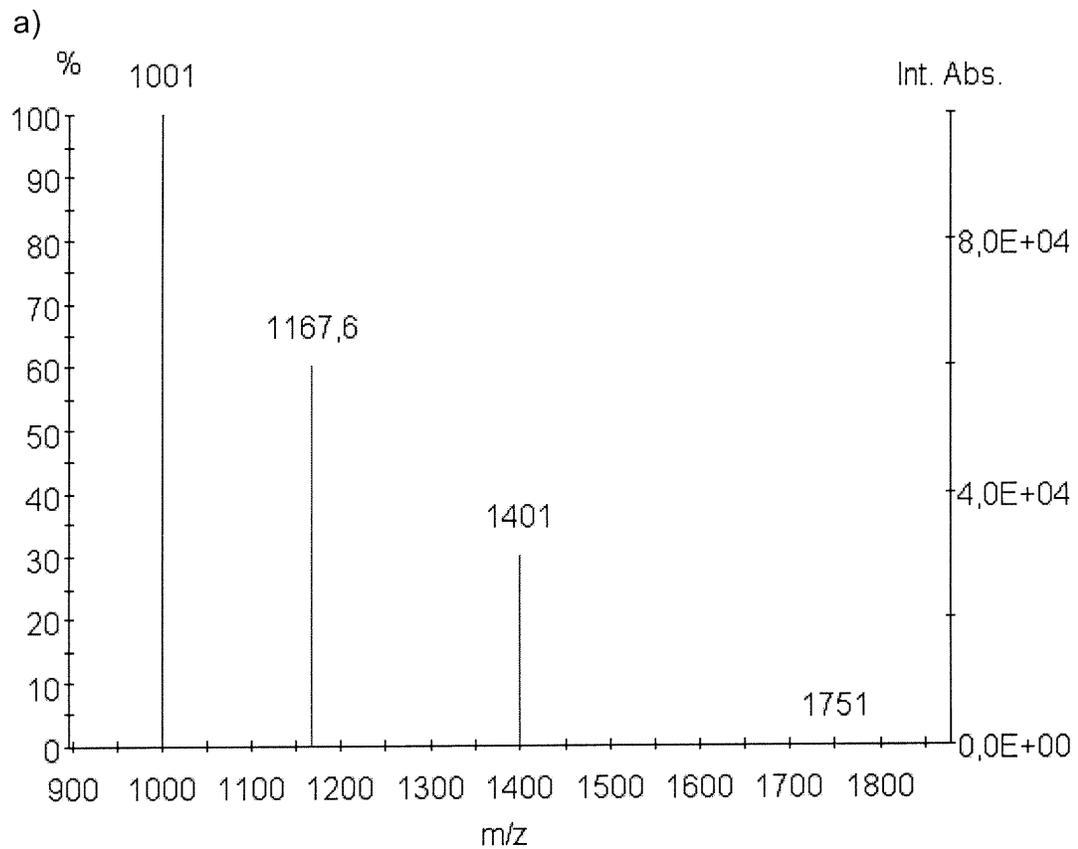


Figure 8 : Organigramme de l'algorithme de détermination des distributions les plus probables.

Probablion commence par trouver dans le spectre de masse le pic avec l'intensité maximale et ce dans le but de traiter en premier la partie la plus importante du signal. Il assigne ensuite à ce pic, une à la fois, une série de charges possibles. À chaque fois qu'une charge est assignée au pic, l'algorithme détermine tous les pics du spectre de masse qui feraient partie de la même distribution d'ions multichargés si ce pic avait cette charge. Pour ce faire, il calcule la masse de la protéine représentée par la charge testée en utilisant l'équation II.2. Il est alors possible, en utilisant l'équation II.1, de déterminer la valeur de m/z où devrait se trouver le prochain pic de la distribution testée.

La Figure 9 montre le traitement effectué par l'algorithme Probablion lors de la déconvolution. La Figure 9a montre un spectre de masse simulé. L'algorithme commence par trouver le maximum qui est le pic à la valeur de m/z de 1001. Il teste ensuite la charge $1+$ et essaye de trouver d'autres pics qui appartiendraient à cette distribution. Une fois qu'il a terminé avec cette charge, il passe à la suivante jusqu'à ce qu'il ait testé toutes les charges possibles. La Figure 9b montre en détail comment l'algorithme Probablion effectue le traitement d'une de ces charges, dans ce cas la charge $7+$. Si le pic à 1001 m/z possédait la charge de $7+$, la masse de la protéine correspondante serait de 7000 Da. Comme les autres pics de la même distribution d'ions multichargés doivent représenter la même protéine, la masse déconvoluée de la protéine devrait être sensiblement la même. Probablion peut donc utiliser cette masse déconvoluée pour prédire l'emplacement des prochains pics de la distribution, c'est-à-dire les pics ayant des charges différentes. La Figure 9c et d montre les calculs pour les charges 4, 5 et $6+$. Le calcul pour les pics de la distribution d'ions multichargés ayant des charges supérieures à la charge testée se fait de façon similaire.



a) Maximum du spectre de masse
 Charge Testée
 $(1001 - 1) * 7 = 7000$
 Masse déconvoluée
 Masse d'un proton

b) Masse déconvoluée
 Masse d'un proton
 $7000 / 6 + 1 = 1167,66$
 m/z où devrait de trouver le prochain pic
 Charge du prochain pic attendu

c) $7000 / 5 + 1 = 1401$
 $7000 / 4 + 1 = 1751$

Figure 9 : Exemple du traitement d'une distribution d'ions multichargés avec l'algorithme Probablion. a) Spectre de masse. b) Calcul de la masse déconvoluée proposée. c) et d) Calculs des m/z des trois prochains pics attendus d'après la masse déconvoluée proposée.

L'algorithme commence donc par chercher un pic correspondant à la charge $6+$ autour de la valeur attendue. S'il n'en trouve aucun ou si l'intensité du pic est trop faible, comme dans le cas du pic à 1751, l'algorithme arrête de chercher et passe à l'autre charge à tester sur le maximum du spectre de masse.

Une fois que toutes les charges possibles ont été testées, l'algorithme possède une distribution d'ions multichargés pour chacune des charges testées sur le maximum. Pour déterminer quelle distribution est la plus probable, il additionne l'intensité de tous les pics des distributions et garde seulement celle possédant la plus grande intensité. Comme l'algorithme arrête de chercher des pics quand un des pics attendus est manquant, une bonne distribution devrait donner un signal intense et les corrélations aléatoires devraient donner un signal plus faible. L'algorithme peut ensuite calculer la masse déconvoluée moyenne représentée par cette distribution et placer un pic y correspondant dans le spectre déconvolué. L'intensité utilisée pour ce pic déconvolué est la somme de tous les pics de la distribution d'ions multichargés.

La Figure 10 montre la réaction de l'algorithme Probablion face à la situation où deux charges testées sont en compétition pour un même pic. Il est possible de voir que l'algorithme trouve tout d'abord la possibilité d'une distribution d'ions multichargés en assignant la charge $4+$ au pic d'intensité maximale. En effet, l'algorithme a réussi à trouver des pics possédant les charges 2 , 3 et $5+$ qui sont en accord avec la masse de la protéine telle que déterminée grâce à la charge $4+$. La masse de la protéine assignée à cette distribution d'ions multichargés est de 4000 Da.

L'algorithme a par contre trouvé une autre distribution d'ions multichargés lorsque la charge $7+$ est assignée au pic d'intensité maximale. En effet, Probablion a réussi à trouver des pics en accord avec la masse déconvoluée de 7000 Da calculée grâce au pic d'intensité maximale.

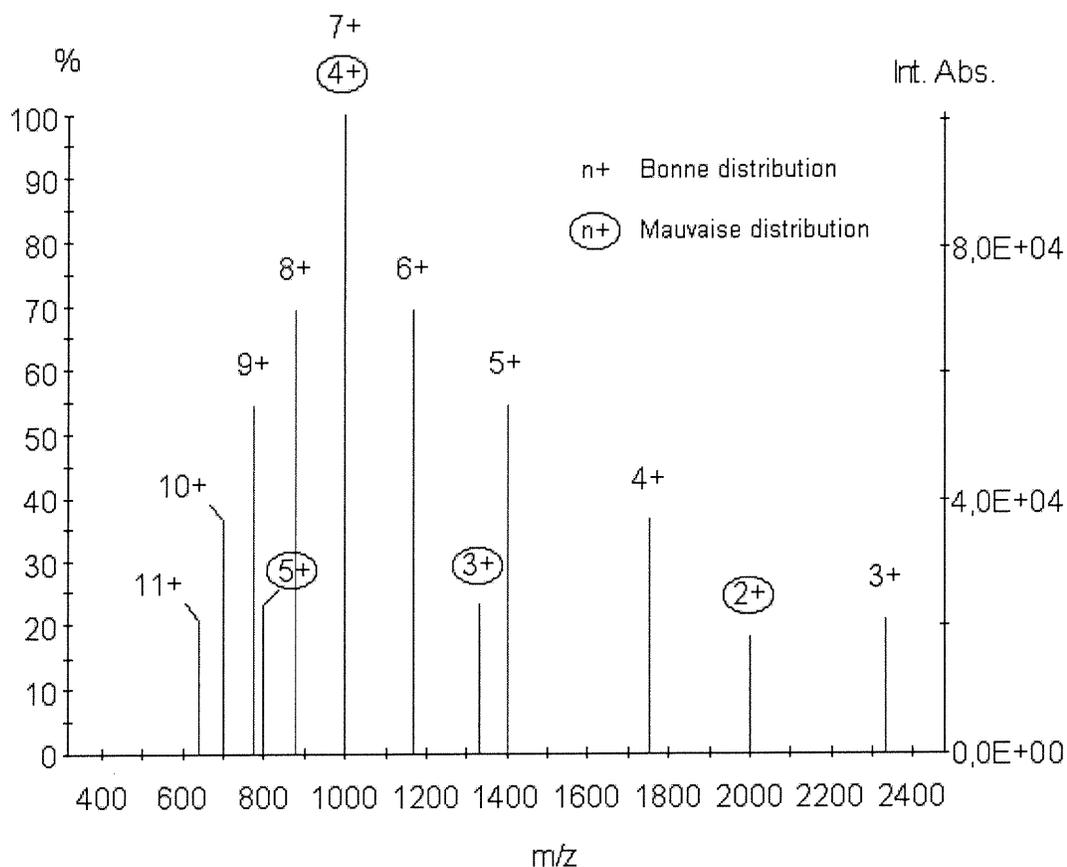


Figure 10 : Exemple de traitement d'une distribution par l'algorithme.

Comme la somme des intensités de ces distributions d'ions multichargés est plus importante pour la charge 7+, ce sera cette charge qui sera choisie comme la base de la distribution d'ions multichargés à placer dans le spectre déconvolué.

Détermination des positions de haute densité de masse

L'algorithme de détermination des positions de haute densité de masse, appelé Dension, est résumé par l'organigramme de la Figure 11. L'algorithme Dension consiste à multiplier les m/z corrigés de tous les pics du spectre de masse (soustraction de la masse d'un proton) par une série d'entiers représentant les différentes charges possibles pour les pics. Tous les pics multipliés sont placés dans

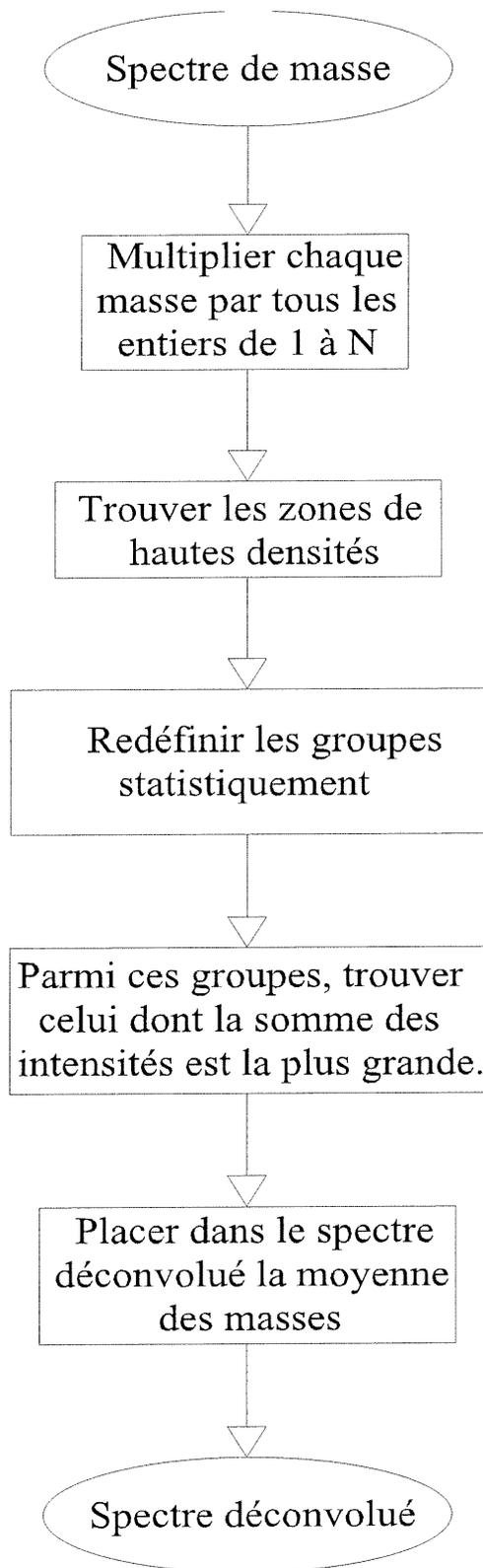


Figure 11 : Organigramme de l'algorithme de détermination des positions de haute densité de masse.

un même spectre *multiplié*. Les pics provenant d'une même distribution d'ions multichargés devraient tous se regrouper à la même valeur qui est la masse de la protéine.

La Figure 12a montre un spectre de masse simulé. La Figure 12b montre le spectre multiplié obtenu une fois que les pics ont été multipliés. Il est possible d'y voir que les deux pics se regroupent à 5000 Da. Ces deux pics se regroupent parce qu'ils font partie de la même distribution d'ions multichargés. C'est-à-dire quand le m/z corrigé du pic à 1001 m/z est multipliée par 5 et celui du pic à 1251 m/z est multipliée par 4. Ceci signifie que la charge du pic à 1001 m/z est de 5+ et celle à 1251 m/z est de 4+.

Les masses ayant les plus grandes densités de pics sont les plus aptes à représenter une protéine. L'algorithme doit donc déterminer où sont ces zones de hautes densités. Dans l'exemple de la Figure 12, l'intensité peut être utilisée comme mesure de la densité de pics et il serait déterminé que la masse de 5000 Da est une position de hautes densités de masse. Pour palier à la variation naturelle dans les valeurs de m/z à l'intérieur d'une distribution d'ions multichargés, l'algorithme prend tous les pics à l'intérieur d'un certain intervalle autour de cette zone de hautes densités. Par exemple dans la Figure 12, l'algorithme prendrait tous les pics contenus dans l'intervalle 5000 ± 1 Da. Une analyse statistique de ces groupes de pics permet de déterminer leur largeur plus exactement. En effet, pour chaque groupe, la moyenne des pics est calculée et seuls les pics à l'intérieur d'un écart type autour de la moyenne sont conservés. La moyenne est ensuite calculée à nouveau et le groupe redéfini. Il est supposé que tous les pics se trouvant dans le groupe de pics ayant l'intensité la plus élevée proviennent de la protéine. La masse qu'ils représentent est

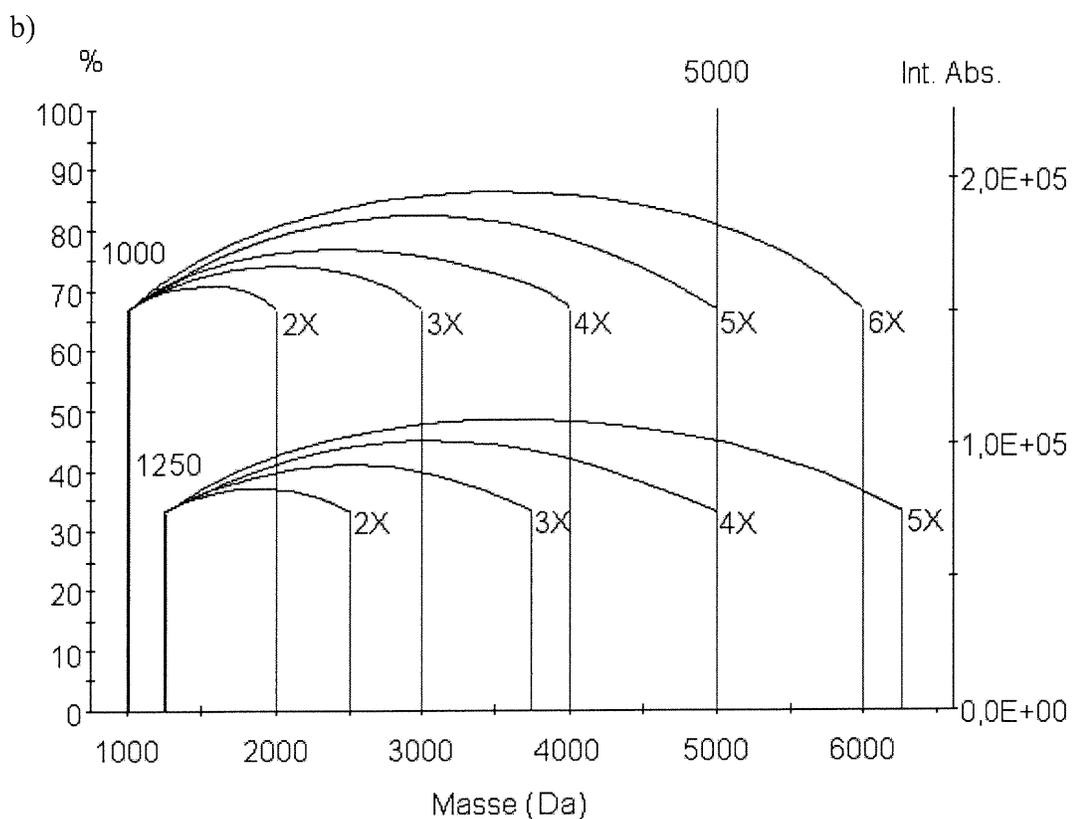
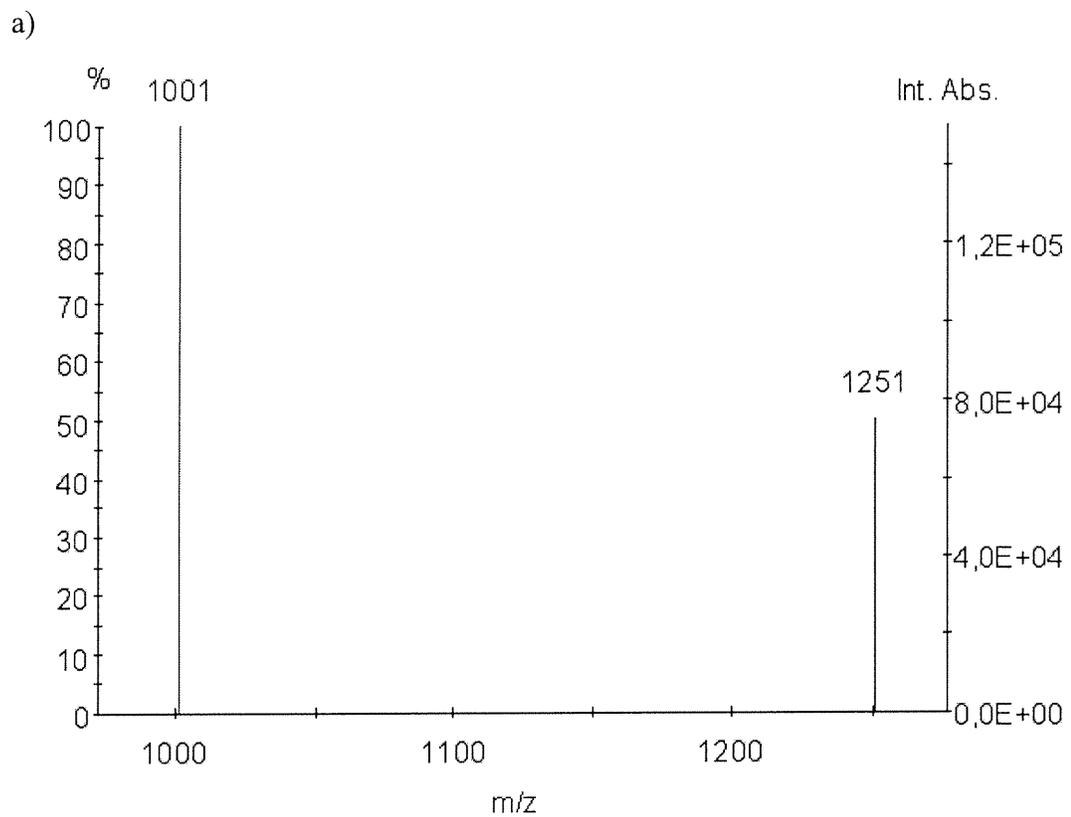


Figure 12 : Exemple de déconvolution avec l'algorithme Dension. a) Spectre de masse simulé. b) Spectre multiplié.

reportée dans le spectre déconvolué et la somme des intensités de tous les pics du groupe est l'intensité du pic déconvolué. Avec cet algorithme, les harmoniques sont inévitables dans le spectre multiplié. En effet, la Figure 13 montre le spectre multiplié obtenu à partir du spectre de masse de la Figure 12a. Il est possible de voir qu'il y a aussi un pic à 10000 Da.

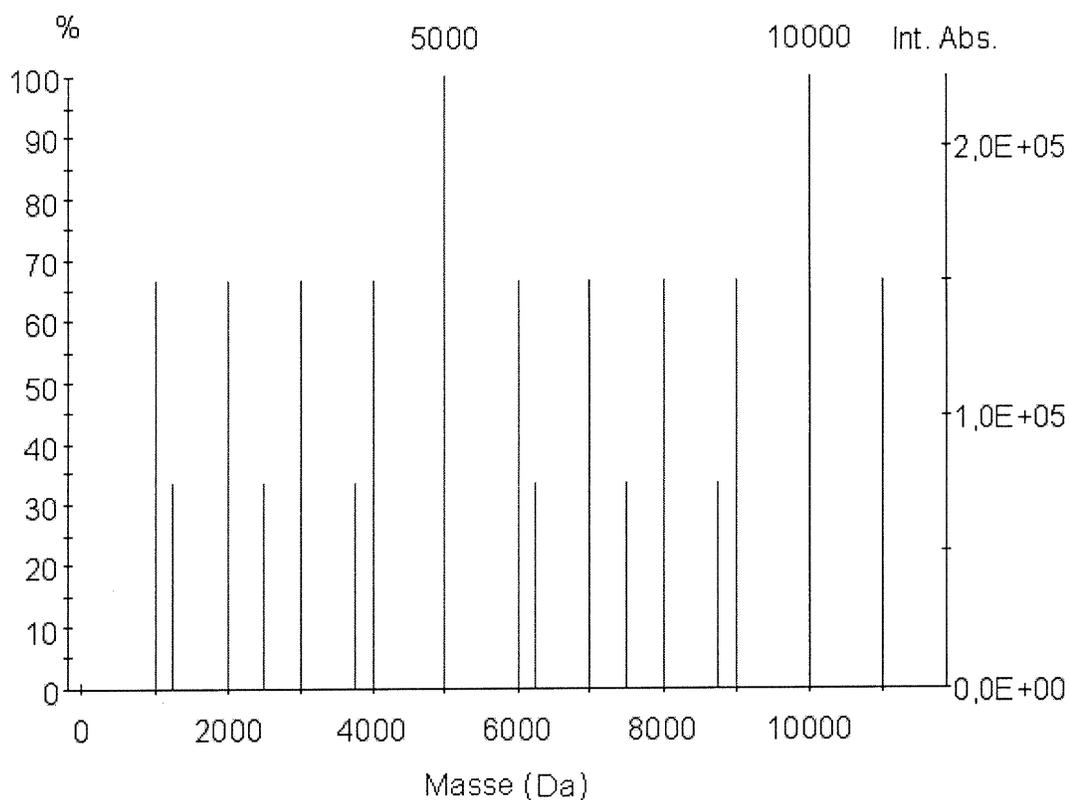


Figure 13 : Exemple d'harmoniques créés lors de la multiplication des pics.

Un problème peut alors survenir si une de ces harmoniques est plus intense que le vrai pic. En effet, l'algorithme détecterait alors une fausse masse pour la protéine. Pour palier à ce problème, l'algorithme Dension doit chercher tous les pics multiples ou représentant une fraction entière d'une zone de hautes densités qui ont sensiblement la même intensité. Il garde ensuite seulement la zone de hautes densités qui a la plus faible masse. Dans l'exemple de la Figure 13, si le pic à 10000 Da était

légèrement plus intense que celui à 5000 Da à cause d'une contamination par le signal de fond, l'algorithme détecterait alors 10000 comme la masse de la protéine. Mais comme il cherche aussi les multiples ou fractions entières d'intensités similaires, il trouverait aussi le pic à 5000 Da comme la moitié de 10000 Da. Or, comme la masse 5000 est plus petite que 10000, l'algorithme déterminerait que le pic à 5000 Da est la vraie zone de hautes densités.

Processus itératif

Ces deux algorithmes seraient peu utiles tels qu'ils sont représentés dans le cas de mélanges de protéines. Il faut donc qu'ils soient améliorés pour tenir compte de ces situations. Ceci est fait par un processus itératif, c'est-à-dire qu'ils effectuent une fois le processus de déconvolution sur les données expérimentales afin de trouver une des protéines. Une fois que ceci est terminé, la partie déjà traitée du spectre de masse est éliminée pour ensuite recommencer le processus de déconvolution jusqu'à ce que tout le spectre de masse soit traité.

La première itération de l'algorithme de détermination de la distribution la plus probable, Probablion, utilise le maximum global du spectre de masse. Les autres itérations utilisent le maximum du spectre de masse qui reste après la soustraction des pics appartenant à la distribution d'ions multichargés détectée.

L'algorithme de détermination des positions de hautes densités de masse, Dension, utilise le spectre multiplié obtenu à partir du spectre de masse original comme point de départ. Les autres itérations utilisent le spectre multiplié obtenu à partir du spectre de masse résiduel de l'itération précédente. Avec cet algorithme, tous les pics appartenant au groupe de haute densité le plus intense sont soustraits du spectre de masse afin qu'ils ne soient plus traités lors des prochaines itérations.

À la fin d'une itération, l'algorithme prend la décision de commencer une nouvelle itération si le nombre de pics dans le spectre résiduel est significatif. Le processus itératif permet de traiter une seule partie du signal à la fois, ce qui permet de traiter les mélanges de protéines plus efficacement.

Utilisation directe vs inversée

L'algorithme de détermination des distributions les plus probables, Probablion, peut être utilisé de deux façons. Celles-ci sont appelées utilisation directe et inversée en fonction du moment où l'accumulation de spectres est effectuée. L'utilisation directe consiste à déconvoluer un spectre de masse obtenu en accumulant plusieurs spectres de masse d'un même échantillon. L'utilisation inversée quant à elle consiste à déconvoluer chaque spectre de masse d'un échantillon pour ensuite accumuler les spectres déconvolués obtenus. Même si ces deux utilisations semblent équivalentes, il sera montré que ce n'est pas le cas. Dans le cas de l'utilisation inversée, la filtration des spectres de masse individuels n'est pas possible à cause du fait qu'il n'y a pas de pics à chaque m/z comme dans le cas des spectres accumulés. La filtration permet de réduire la quantité d'harmoniques observées ainsi que la quantité de faux pics aléatoires. L'utilisation inversée n'utilisant pas la filtration ne peut pas se départir des harmoniques aussi facilement, mais l'accumulation permet de réduire grandement l'effet des faux pics aléatoires car ceux-ci ne sont pas accumulés aussi efficacement que les vrais signaux. Les harmoniques qui restent sont plus faciles à reconnaître que les faux pics aléatoires puisqu'elles se trouvent à des multiples ou des fractions entières de vrais pics.

Chapitre VI

Applications

Afin de tester les algorithmes de déconvolution conçus, une série de protéines ont été utilisées. Plusieurs protéines pures ont été traitées ainsi que quelques mélanges de quelques-unes d'entre elles. De plus, des spectres de masse simulés ont aussi été utilisés afin de déterminer les limitations théoriques de la méthode.

Évaluation de l'erreur sur les résultats

Afin de donner une estimation de l'erreur sur les valeurs des masses obtenues, le calcul d'un intervalle de confiance a été effectué. Cet intervalle de confiance est calculé grâce à la table du test de Student. Pour chaque pic du spectre déconvolué, il est possible de calculer une masse moyenne et un écart type à cause des différents pics des distributions d'ions multichargés. L'écart type sert à calculer l'intervalle de confiance (I.c.) autour de la moyenne et l'équation VI.1 donne la façon de le calculer.

$$I.c. = \frac{t_{\%,n-1} * S_d}{\sqrt{n}} \quad (VI.1)$$

où $t_{\%, n-1}$ est la valeur de référence de la table de Student pour un niveau de confiance % et un nombre de degrés de liberté de n-1,

s_d est l'écart type sur les masses de la protéine et

n est le nombre de pics dans la distribution d'ions multichargés.

Pour cette étude le niveau de confiance utilisé est de 90% ce qui signifie que la masse de la protéine devrait se situer dans l'intervalle de confiance (entre

parenthèses dans les spectres) 9 fois sur 10. Ce niveau a été jugé suffisant pour cette étude.

Pour l'utilisation inverse de l'algorithme de détermination des distributions les plus probables, un intervalle de confiance a été calculé pour tous les pics des spectres déconvolués individuels. Il est aussi important d'avoir une mesure de la qualité de la déconvolution pour l'utilisation inverse de l'algorithme Probablion. Cette technique fait la moyenne de tous les spectres déconvolués obtenus pour qui des intervalles de confiance ont déjà été calculés. Il a donc été jugé suffisant d'avoir, pour une masse donnée dans le spectre déconvolué final, la moyenne des intervalles de confiance de cette masse dans les spectres déconvolués individuels comme mesure de la qualité de cette déconvolution.

Simulation du spectre d'une protéine pure

Le premier spectre de masse traité est un spectre simulé (Figure 14) où il n'y a pas de faux pics ou de signal de fond continu et où il n'y a pas de variation autour des vraies valeurs de m/z . La masse de la protéine telle que simulée est de 15000 Da.

Comme le montre la Figure 15 et la Figure 16, la déconvolution par l'algorithme Probablion et l'algorithme Dension donne de bons résultats. Un petit intervalle de confiance est calculé dans les deux cas, celui-ci étant dû aux erreurs amenées par les arrondissements des masses. En effet, l'ordinateur conserve les valeurs de m/z et d'intensité dans des variables qui ne peuvent conserver qu'un nombre déterminé de décimales. Cet effet est négligeable par rapport à la variation due à la calibration de l'appareil qui sera observé dans les spectres expérimentaux.

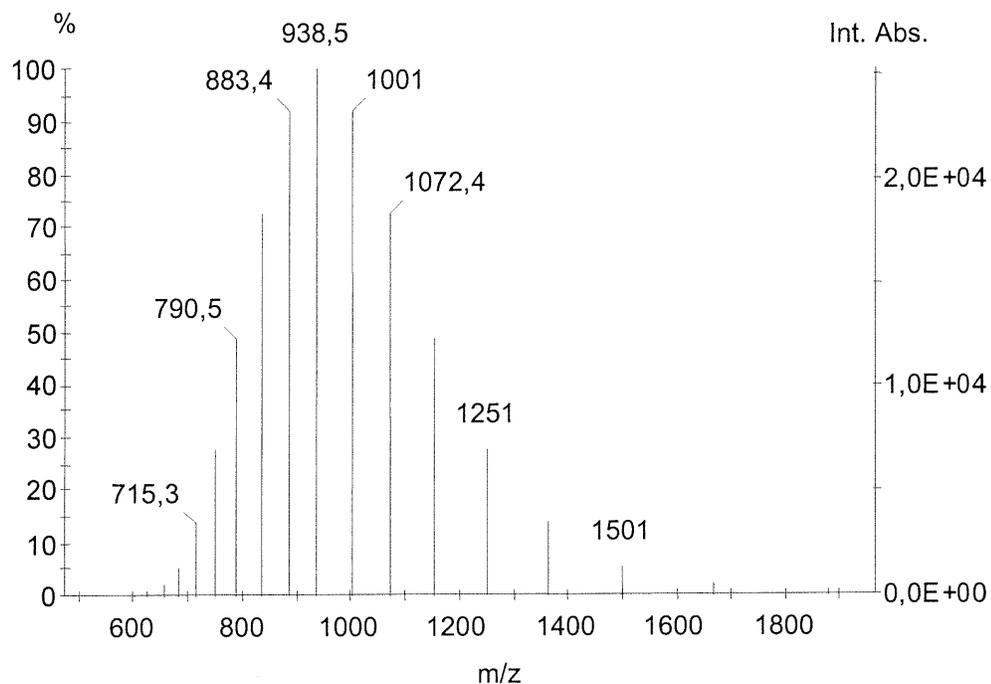


Figure 14 : Spectre de masse simulé d'une protéine de masse de 15000 Da sans signal de fond continu et une précision infinie.

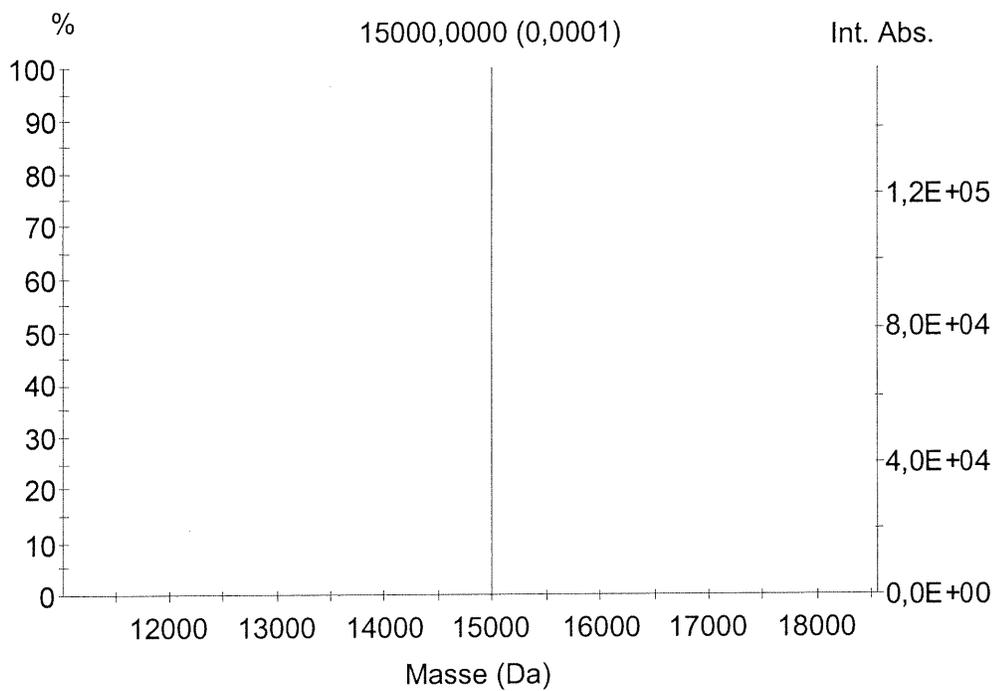


Figure 15 : Spectre déconvolué de la Figure 14 selon l'algorithme de détermination des distributions les plus probables avec l'intervalle de confiance entre parenthèses.

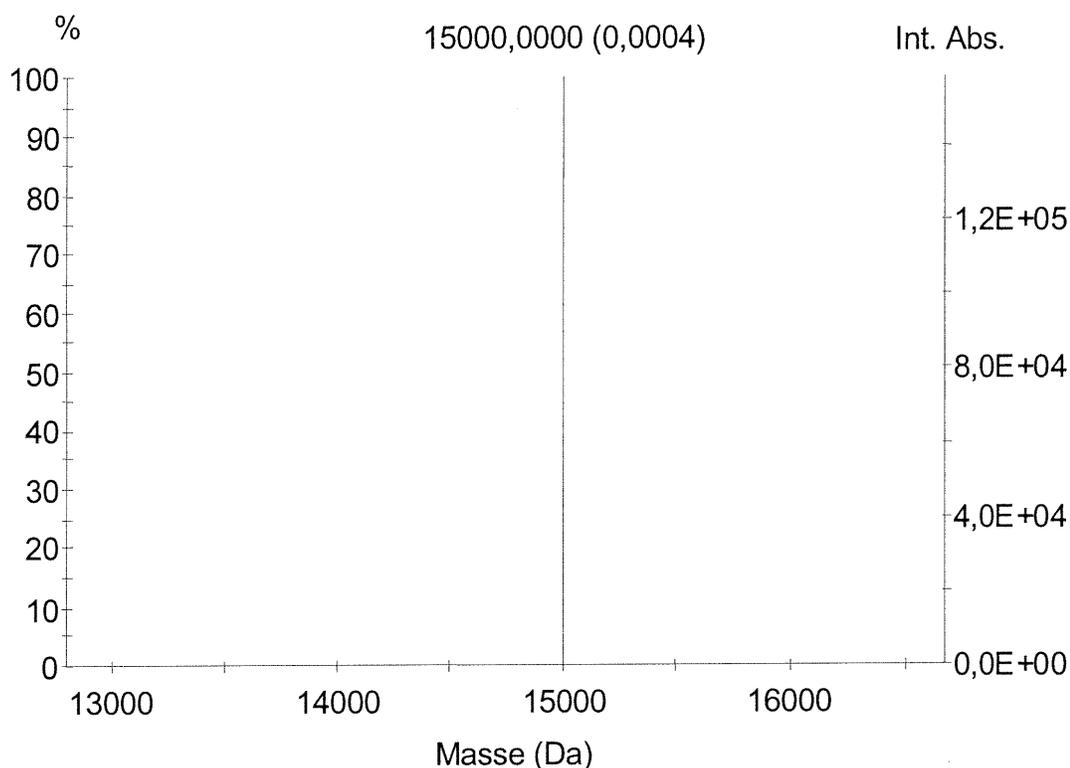


Figure 16 : Spectre déconvolué de la Figure 14 selon l'algorithme de détermination des positions de hautes densité de masse avec l'intervalle de confiance entre parenthèses.

Insuline

La première vraie protéine étudiée, l'insuline, a une masse de 5734 Da. La Figure 17a montre le spectre de masse original pour cette protéine. La Figure 17b montre le même spectre de masse après la filtration. La sensibilité de l'appareil face aux ions de l'insuline est tellement importante que la contribution du signal de fond est négligeable, ce qu'il est possible d'observer visuellement. La filtration ne devrait donc pas avoir d'effet trop important sur la déconvolution, mais il sera montré le contraire.

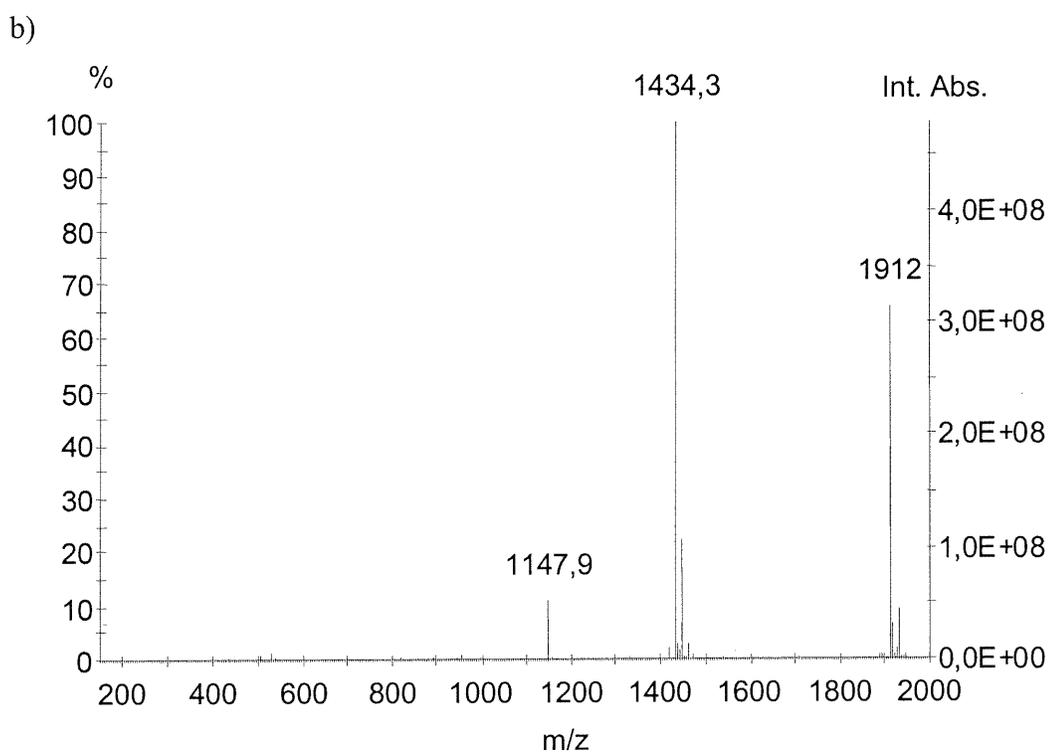
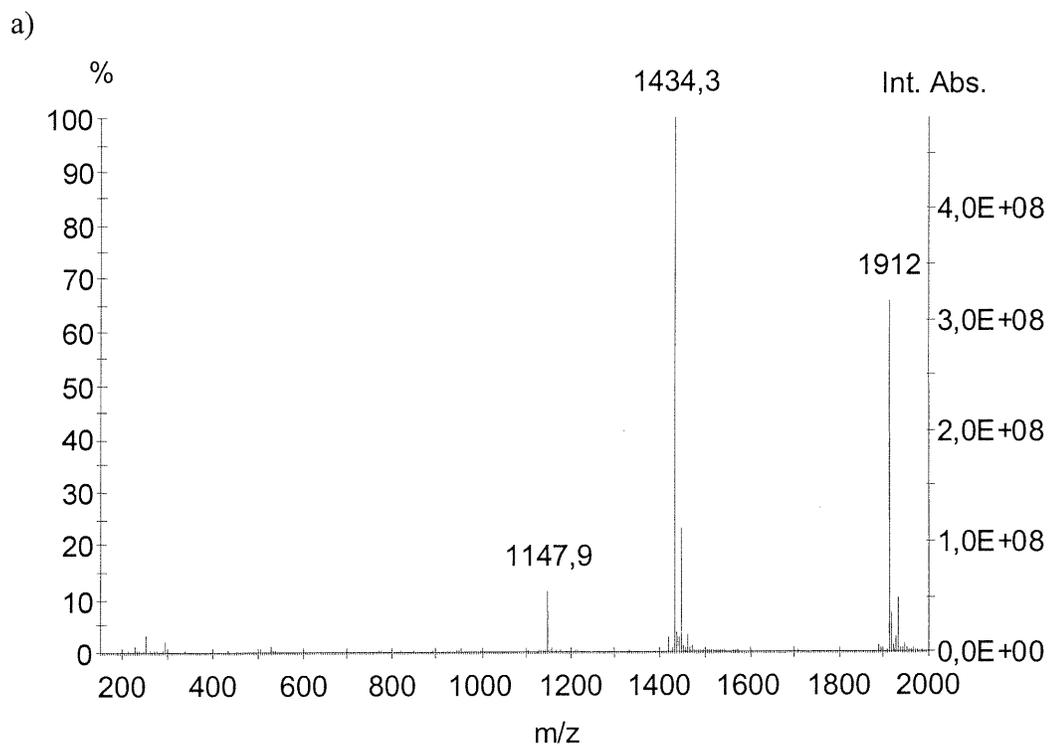


Figure 17 : Spectres de masse accumulés (77 spectres) de l'insuline pure (5734 Da). a) Spectre non filtré. b) Spectre filtré.

Le spectre déconvolué de l'insuline obtenu grâce à l'utilisation directe de l'algorithme de détermination des distributions les plus probables sur l'accumulation de 77 spectres de masse est montré dans la Figure 18a. La Figure 18b est le spectre déconvolué obtenu à partir du spectre de masse accumulé et filtré.

Il est possible de voir que même si le filtrage ne semble pas enlever beaucoup de signal de fond du spectre de masse original, ce traitement change quand même l'allure du spectre déconvolué. Le signal de fond peut donc interférer avec le processus de déconvolution et amener le déplacement de pics. Ceci complique l'interprétation des spectres de masse même dans le cas où le signal de fond continu semble négligeable. Ceci montre donc une autre bonne raison, en plus de réduire la quantité de faux pics aléatoires, de filtrer les spectres de masse avant d'effectuer la déconvolution.

L'utilisation inverse de l'algorithme de détermination des distributions les plus probables est montrée dans la Figure 19 pour l'accumulation des 77 spectres déconvolués de l'insuline. Avec cette méthode, le spectre déconvolué est presque exempt d'harmoniques ou de faux pics aléatoires car le rapport signal sur signal de fond a été augmenté à un point tel que les faux pics sont devenus négligeables face au pic de la vraie masse. Cette utilisation de l'algorithme Probablion demande par contre un plus grand temps de calcul puisqu'il doit effectuer 77 fois le calcul normalement nécessaire pour traiter un spectre de masse non filtré. Le temps de calcul supplémentaire peut être nécessaire dans les cas où l'utilisation directe amène la formation d'une grande quantité de faux pics dans le spectre déconvolué.

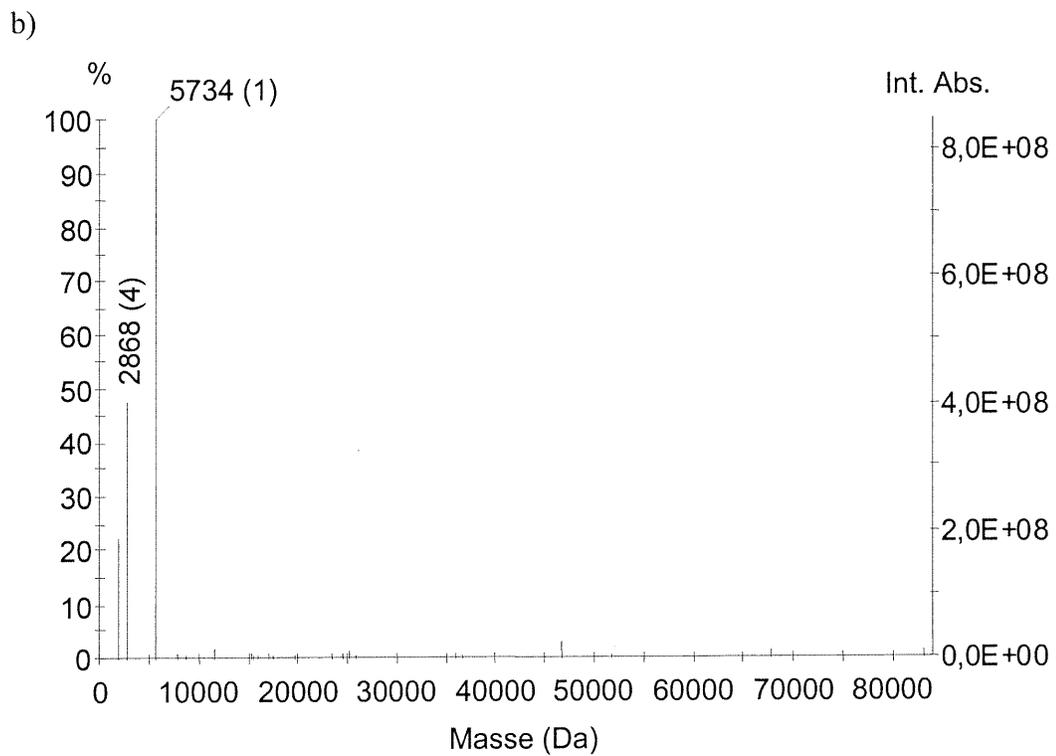
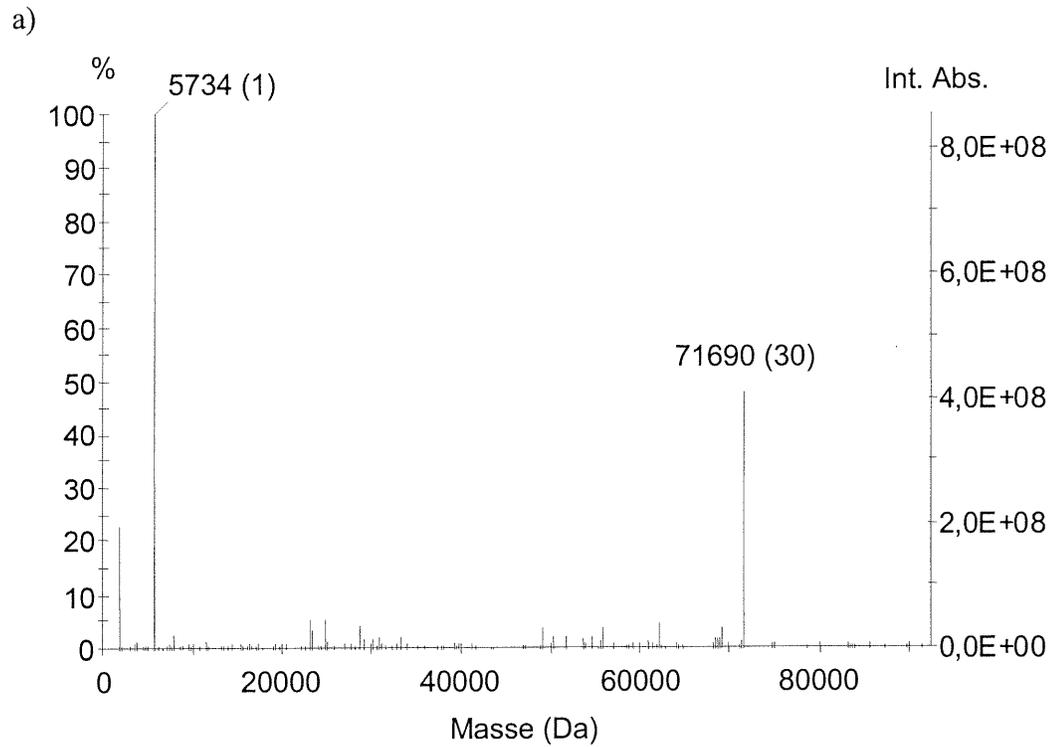


Figure 18 : Spectres déconvolués de l'insuline pure par l'algorithme Probablion. a) À partir du spectre non filtré. b) À partir du spectre filtré.

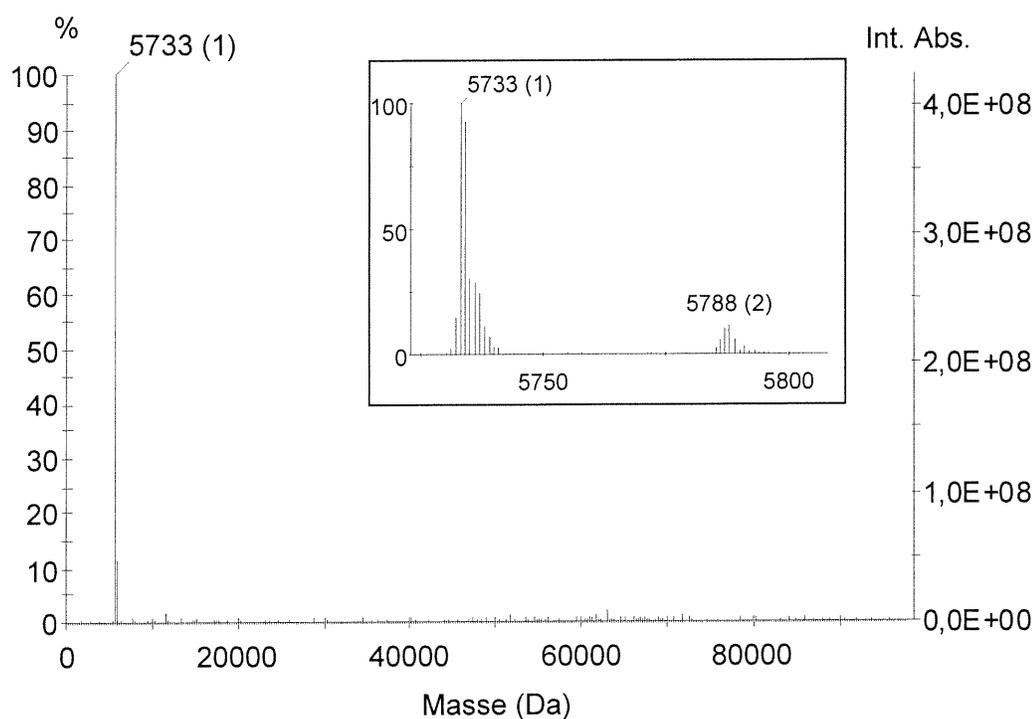


Figure 19 : Spectre déconvolué pour l'insuline par l'utilisation inverse de l'algorithme Probablion. Encadré : Agrandissement autour du pic 5733 Da.

L'agrandissement de la Figure 19 montre que le pic correspondant à la protéine est mieux représenté grâce à cette utilisation de l'algorithme. En effet, le pic est constitué d'une distribution statistique des masses obtenues pour la protéine à travers les déconvolutions des spectres de masse. De plus, cette méthode permet d'observer que la protéine est aussi présente avec d'autres adduits. Le pic à 5788 Da est causé par ce phénomène. Cette masse est une valeur apparente car l'algorithme suppose que les adduits sont des ions H^+ seulement. Il serait donc nécessaire d'utiliser la méthode tridimensionnelle de Labowsky et al.¹⁰, si une étude plus approfondie était souhaitable. Dans la grande majorité des cas, l'adduit H^+ est majoritaire et il n'y a pas d'ambiguïté. Il est aussi possible que l'échantillon de protéine pure contenait des impuretés.

L'utilisation de l'algorithme de détermination des positions de haute densité de masse sur le spectre de masse de l'insuline filtré est montré dans la Figure 20.

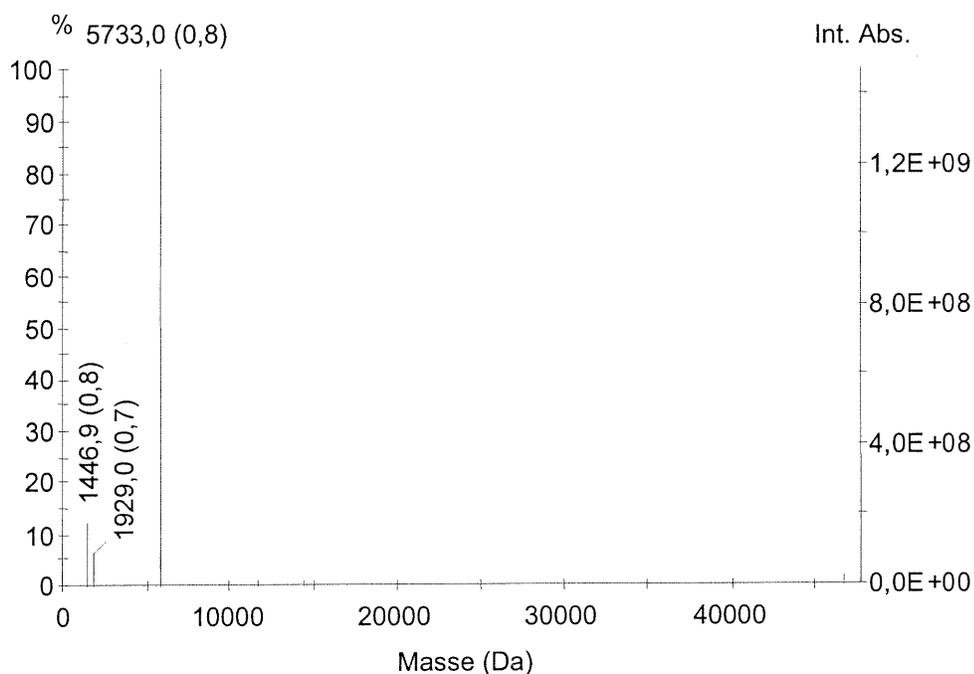


Figure 20 : Spectre déconvolué de l'insuline par l'utilisation de l'algorithme Dension.

Comme dans le cas de l'utilisation directe de l'algorithme Probablion, le spectre déconvolué obtenu par l'algorithme Dension possède aussi deux pics autres que le pic principal. Dans les deux cas, ce ne sont pas des faux pics. Ces pics proviennent de pics présents dans le spectre original qui correspondent à la protéine possédant d'autres adduits que l'ion d'hydrogène. Ils ont été reportés dans le spectre déconvolué comme des ions monochargés d'une protéine. Ceci peut en effet poser des problèmes lors de la déconvolution et donner l'impression qu'il y a des contaminants dans le spectre. La difficulté à traiter les adduits autres que l'ion

d'hydrogène est donc une limitation de ces algorithmes. Si l'observation de ces autres adduits est nécessaire, il est alors préférable d'utiliser une méthode de déconvolution tridimensionnelle comme celle développée par Labowsky et al.¹⁰ En effet, l'algorithme de déconvolution de Labowsky et al. effectue la déconvolution du spectre de masse pour tous les adduits possibles.

Ubiquitine

La deuxième protéine étudiée est l'ubiquitine dont la masse est 8565 Da. Les spectres de masse accumulés, non filtrés et filtrés, sont montrés dans la Figure 21a et la Figure 21b. La forme de la distribution des ions multichargés est différente de celles des autres protéines étudiées. En effet, la distribution a plutôt la forme de deux distributions de Poisson non résolues. Ceci est probablement dû au fait que deux états chargés sont favorisés pour cette protéine. Ceci peut être dû à une dénaturation partielle de la protéine.

Le signal de fond continu du spectre de masse n'est pas vraiment intense et la filtration ne semble pas être vraiment nécessaire. Or, comme le montre la Figure 22 la filtration montre encore une fois qu'elle peut réduire la quantité de pics qu'il serait difficile de qualifier de contaminant sans son intervention dans l'utilisation directe de l'algorithme de détermination des distributions les plus probables.

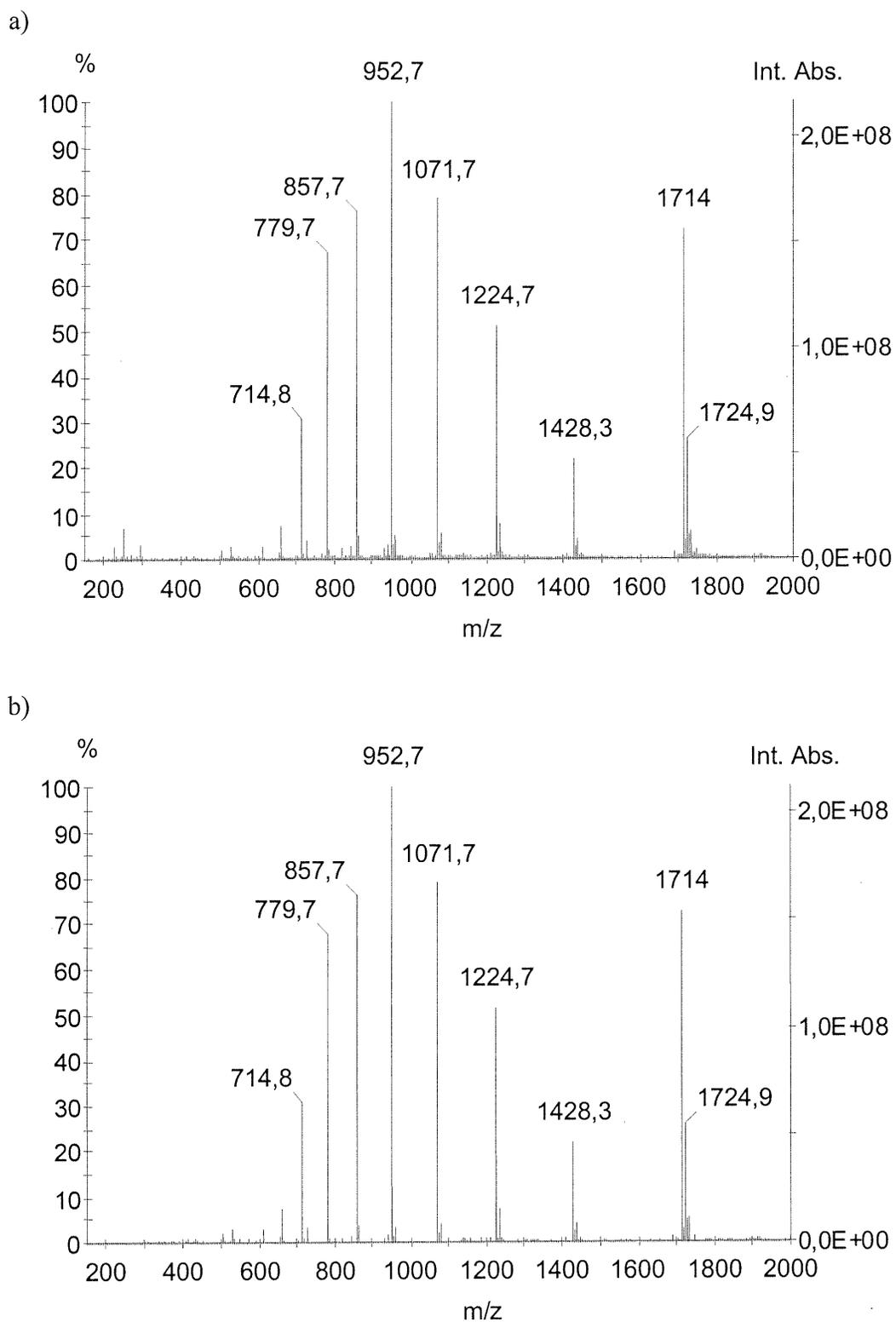


Figure 21 : Spectres de masse accumulés (75 spectres) de l'ubiquitine pure (8565 Da). a) non filtré. b) filtré.

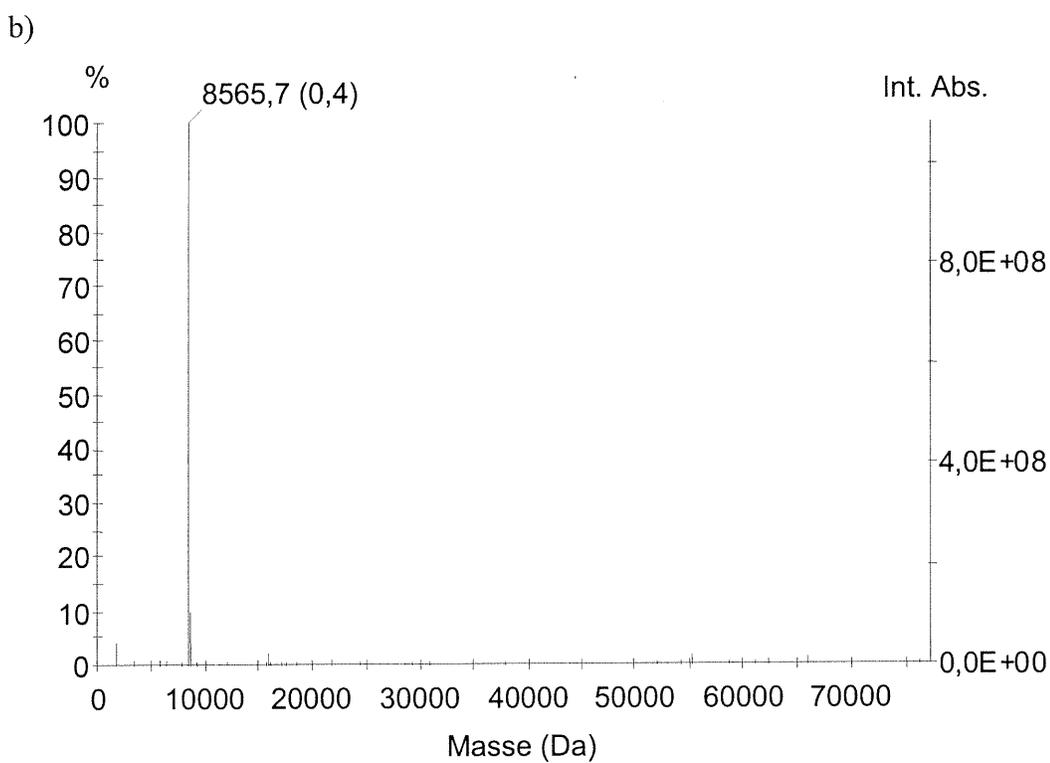
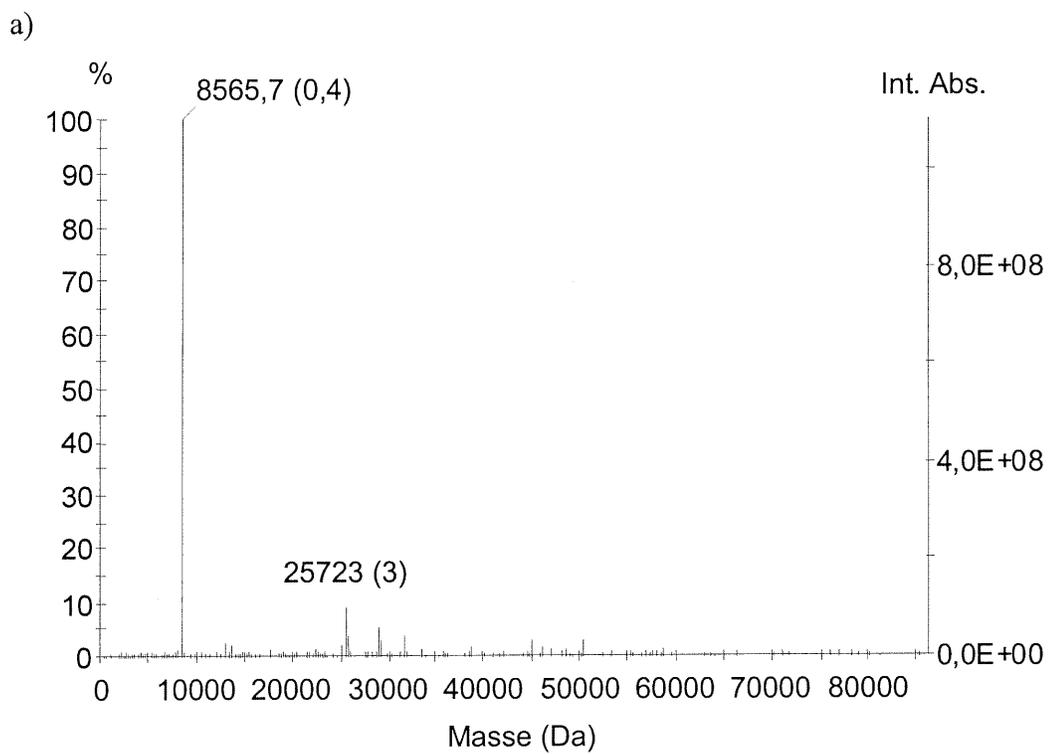


Figure 22 : Spectres déconvolués de l'ubiquitine pure par l'utilisation directe de l'algorithme Probablion. a) À partir du spectre non filtré. b) À partir du spectre filtré.

L'utilisation inverse de l'algorithme de détermination des distributions les plus probables sur les spectres de masse non filtrés donne le spectre déconvolué de la Figure 23. Le signal de fond y est presque indiscernable face au pic de l'ubiquitine.

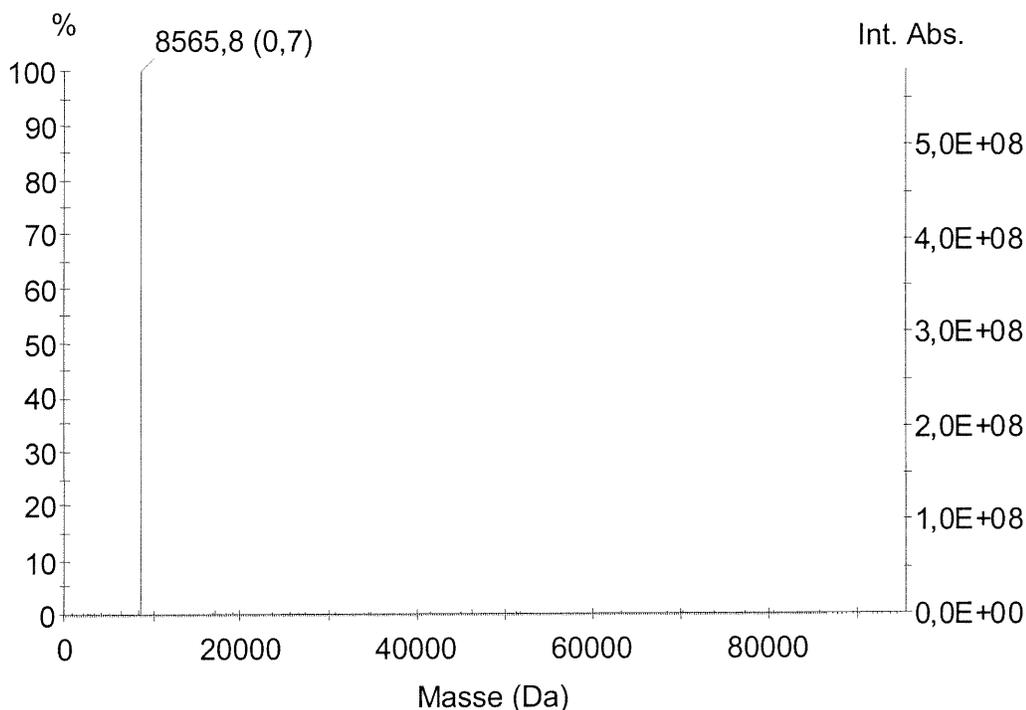


Figure 23 : Spectre déconvolué par l'utilisation inverse de l'algorithme Probablion sur les 75 spectres de masse de l'ubiquitine.

L'utilisation de l'algorithme Dension sur le spectre de masse filtré de l'ubiquitine donne le spectre déconvolué montré dans la Figure 24.

Le spectre obtenu est aussi satisfaisant que ceux obtenus par l'utilisation directe de l'algorithme Probablion. En effet, il y a un seul pic contaminant, mais il n'est pas seulement dû au signal de fond puisque ce pic était aussi présent dans le spectre de masse original. Il doit apparaître à cause qu'il a été jumelé à un faux pic provenant du signal de fond.

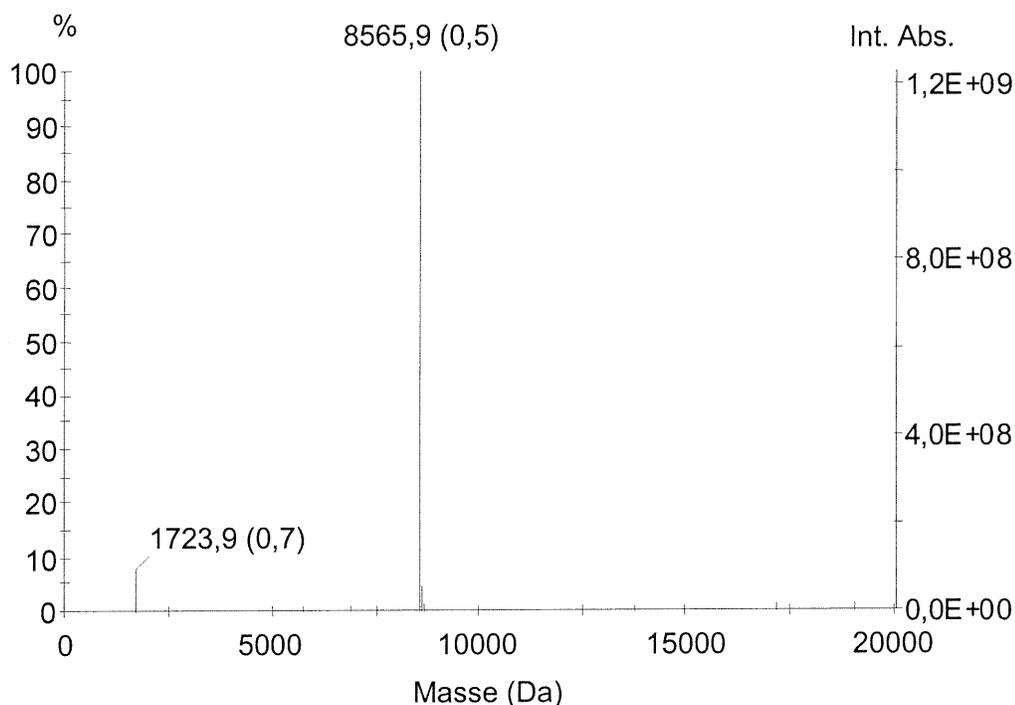


Figure 24 : Spectre déconvolué obtenu avec l'algorithme Dension pour l'ubiquitine.

Cytochrome C

La troisième protéine étudiée est le cytochrome C dont la masse est 12230 Da. Les spectres de masse accumulés à partir de 75 spectres de masse sont montrés dans la Figure 25a et la Figure 25b. Il est possible d'y voir la forme caractéristique des distributions d'ions multichargés en ESI. Le signal de fond continu y est plus apparent mais sans pour autant sembler causer vraiment de problème.

La déconvolution par l'utilisation directe de l'algorithme de détermination des distributions les plus probables donne les spectres montrés dans la Figure 26a et la Figure 26b. Il est possible de voir que la filtration enlève un pic qui pourrait être confondu avec un vrai pic et enlève aussi une grande quantité d'autres faux pics aléatoires.

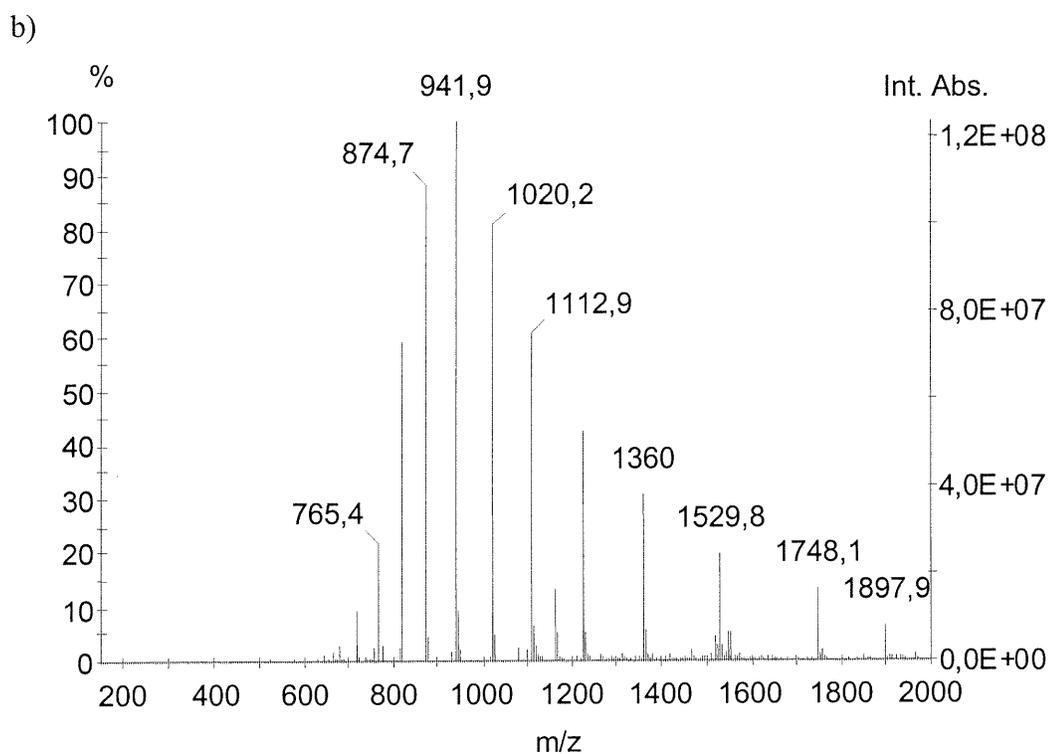
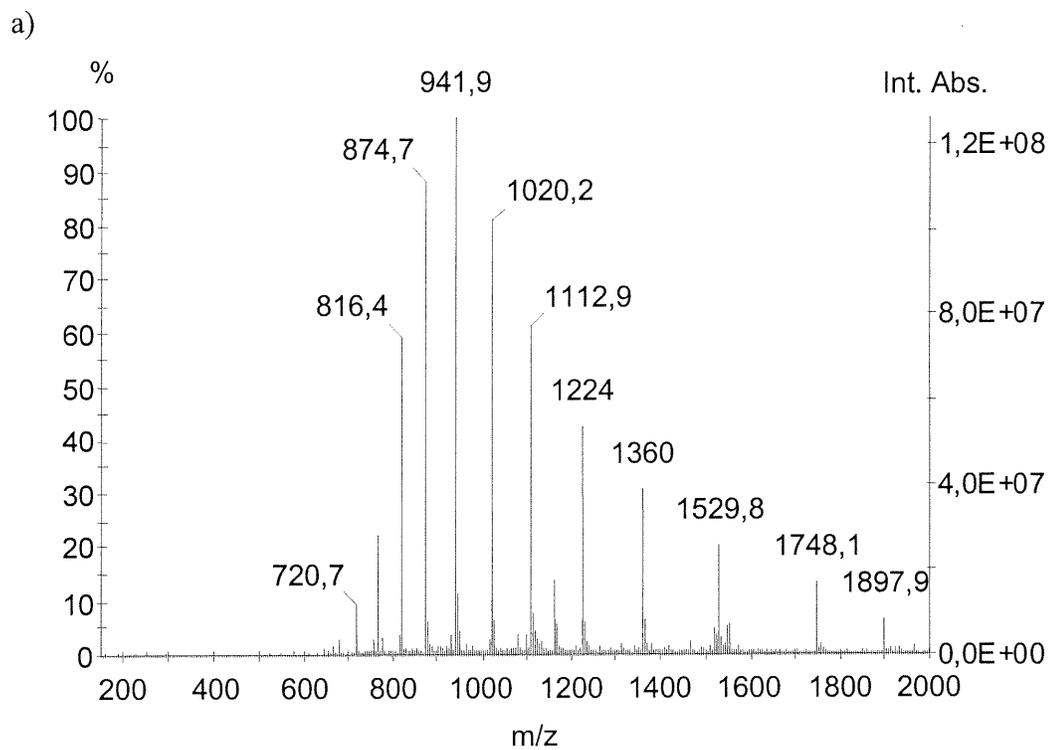


Figure 25 : Spectres de masse accumulés (75 spectres) du cytochrome C pur (12230 Da). a) Spectre non filtré. b) Spectre filtré.

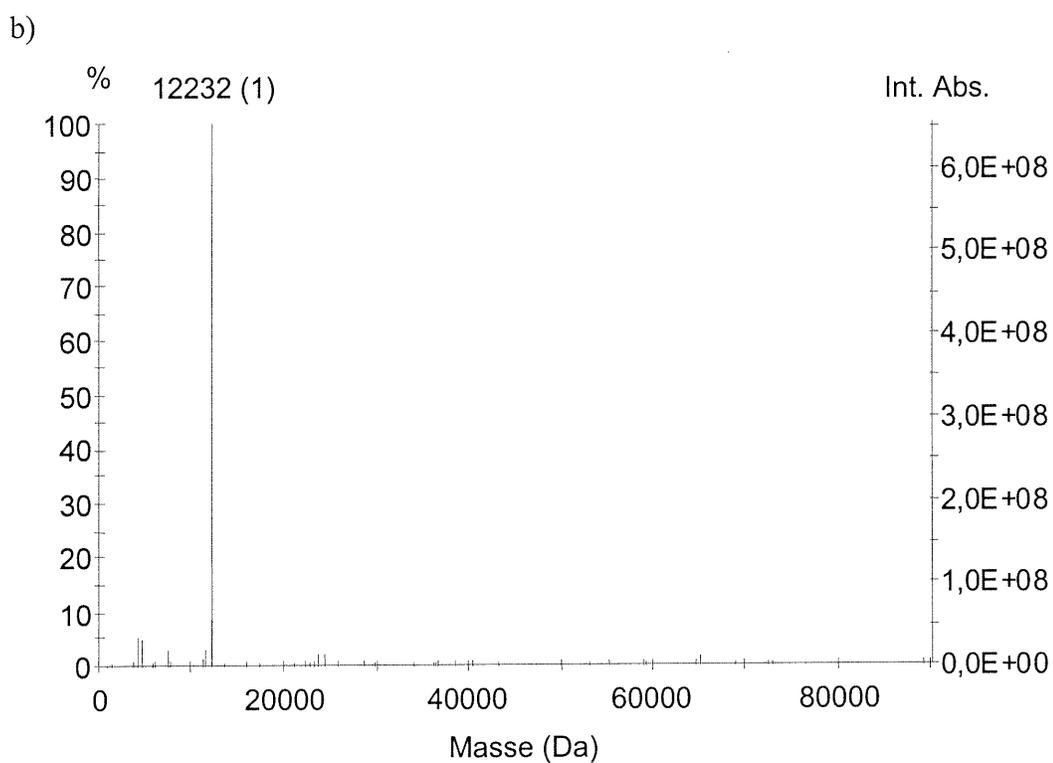
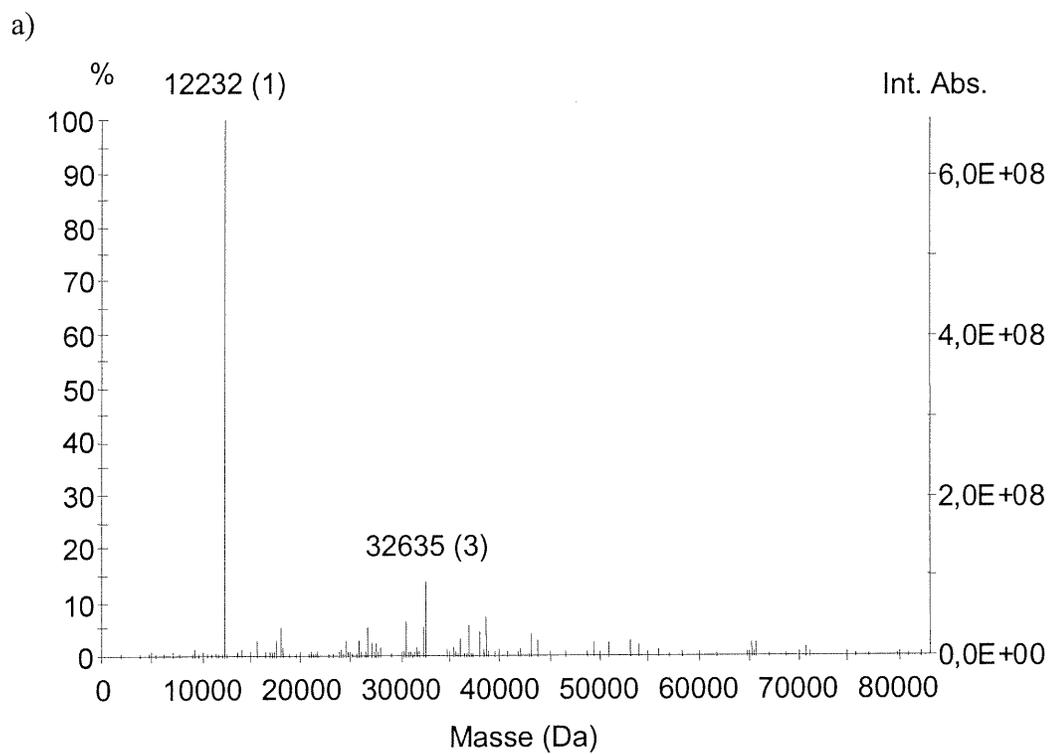


Figure 26 : Spectres déconvolués du cytochrome C pur par l'utilisation directe de l'algorithme Probablon. a) À partir du spectre non filtré. b) À partir du spectre filtré.

La déconvolution par l'utilisation inverse de l'algorithme Probablion donne le spectre déconvolué de la Figure 27. Il est possible d'y voir sans ambiguïté que seul le pic de la protéine est présent. Les autres pics sont presque négligeables face à ce pic.

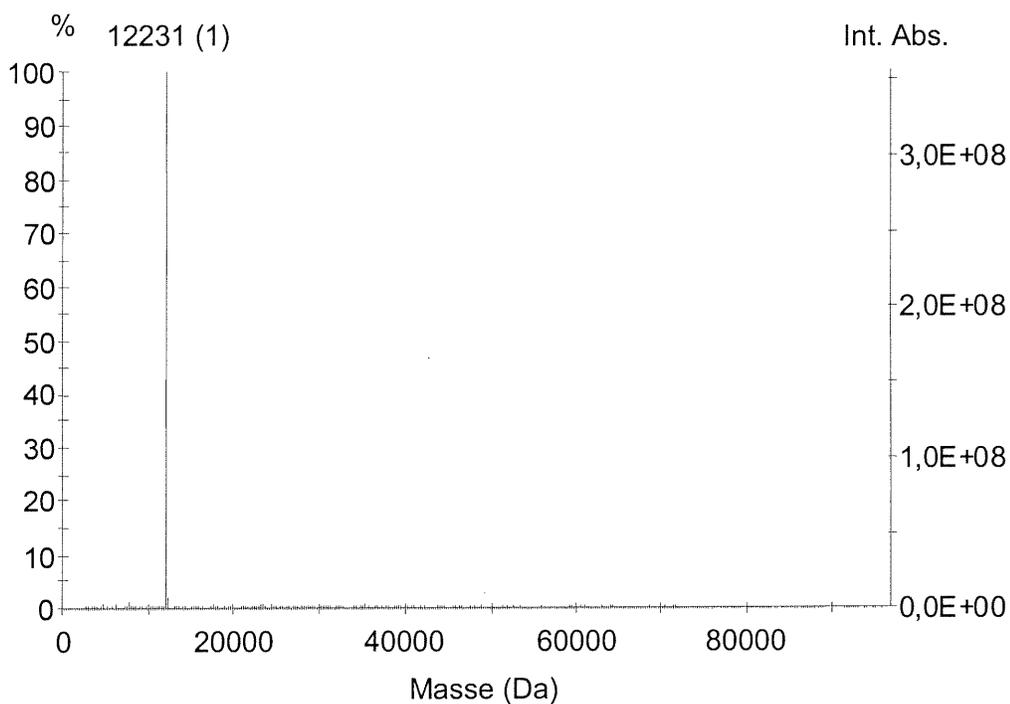


Figure 27 : Spectre déconvolué par l'utilisation inverse de l'algorithme Probablion pour les 75 spectres de masse du Cytochrome C.

L'utilisation de l'algorithme de détermination des positions de haute densité de masse sur le spectre de masse filtré du cytochrome C donne le spectre déconvolué de la Figure 28.

Ce spectre contient plus de faux pics de faible intensité que les autres spectres déconvolués, mais il reste assez facile de voir le composant majeur de l'échantillon. Il serait par contre plus difficile de juger si celui-ci est le seul composant de l'échantillon.

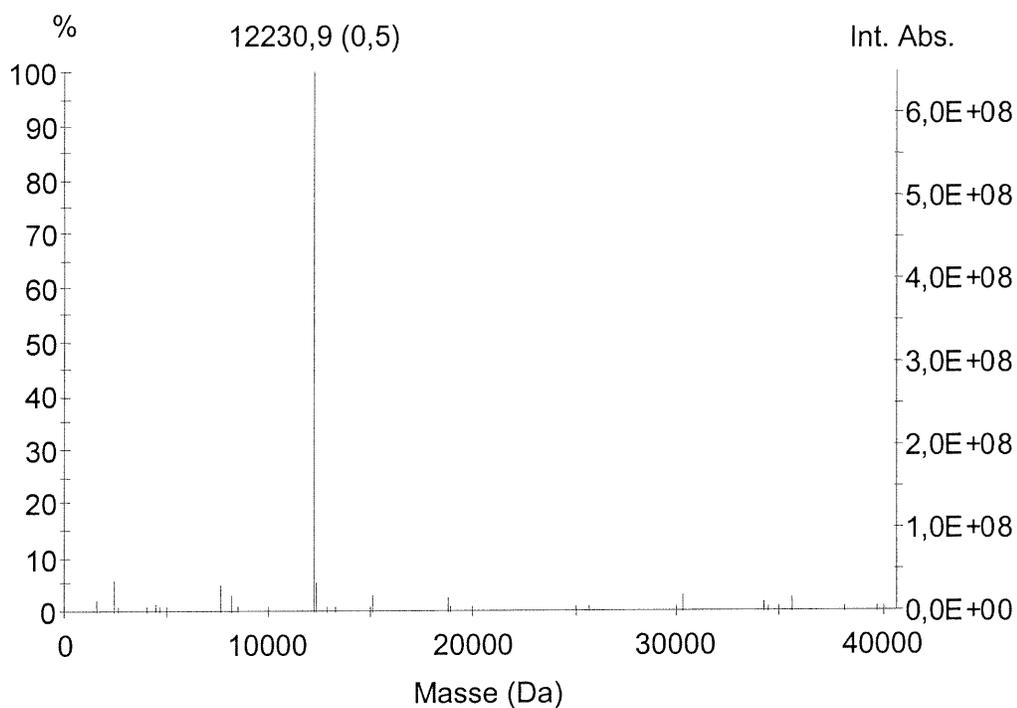


Figure 28 : Spectre déconvolué obtenu à partir du spectre de masse filtré du cytochrome C avec l'algorithme Dension.

L'étude des échantillons contenant une seule protéine est concluante. Il est possible de déterminer la masse d'une protéine et de simplifier le spectre avec les deux algorithmes conçus. Pour ces échantillons, le temps supplémentaire nécessaire pour l'utilisation inverse de l'algorithme Probablion n'est pas vraiment suggéré puisque l'utilisation directe donne d'assez bons résultats. Mais si un doute subsiste sur l'authenticité d'un pic, il peut être souhaitable de l'utiliser pour fins de confirmation.

L'étude d'échantillons contenant une seule protéine semblant assez bien fonctionner, quelques mélanges de protéines ont été étudiés.

Simulation d'un mélange binaire de protéines

Les spectres de masse simulés d'un échantillon contenant un mélange d'insuline (5734 Da) et de cytochrome C (12230 Da) sont montrés à la Figure 29. Un signal de fond continu de la forme d'une distribution de Poisson y a été ajouté pour représenter le comportement expérimental parfois observé. Un bruit sur le signal de fond continu a aussi été ajouté au spectre de masse sous forme de pics contaminants. Le nombre de pics ajoutés correspond à 25% du courant ionique total sans tenir compte du signal de fond continu. De plus, une faible variation autour des valeurs de m/z des pics a été induite. Ceci a été fait en créant une distribution normale de pics dont la valeur moyenne est la masse moléculaire de la protéine simulée et l'écart type est une mesure de la dispersion des pics. En augmentant la valeur de l'écart type de la distribution normale créée, il est possible de simuler une variation de plus en plus grande de la variation autour des valeurs de m/z .

La Figure 29b montre bien que le signal de fond continu est soustrait du spectre de masse grâce à la filtration et que le spectre de masse en est grandement simplifié. Le bruit sur le signal de fond continu n'a pas pu être soustrait puisque celui-ci est présent à des intensités aléatoires.

L'utilisation directe de l'algorithme de détermination des distributions les plus probables est montrée dans la Figure 30.

Il est évident dans la Figure 30a que le signal de fond continu fausse complètement les résultats de la déconvolution. Seule une protéine est apparente et ce à une fausse masse. Le seul fait d'enlever ce signal de fond continu permet à l'algorithme de déconvolution de trouver les deux protéines aux bonnes masses.

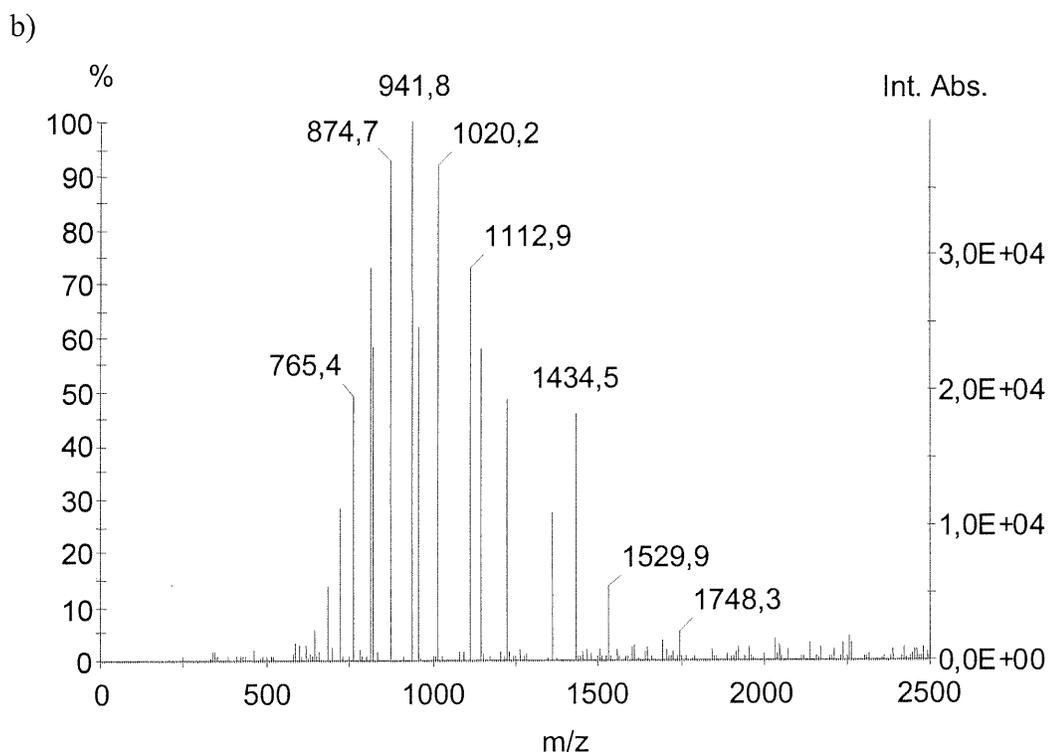
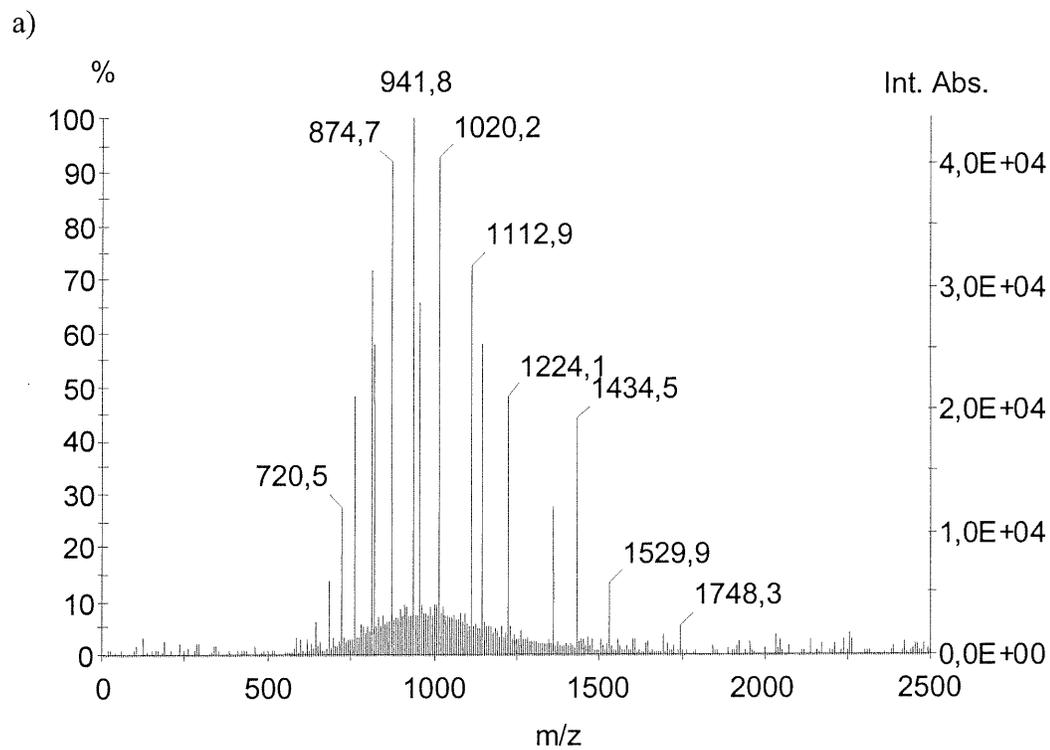


Figure 29 : Spectres de masse simulés d'un mélange d'insuline et de cytochrome C. a) Spectre non filtré. b) Spectre filtré.

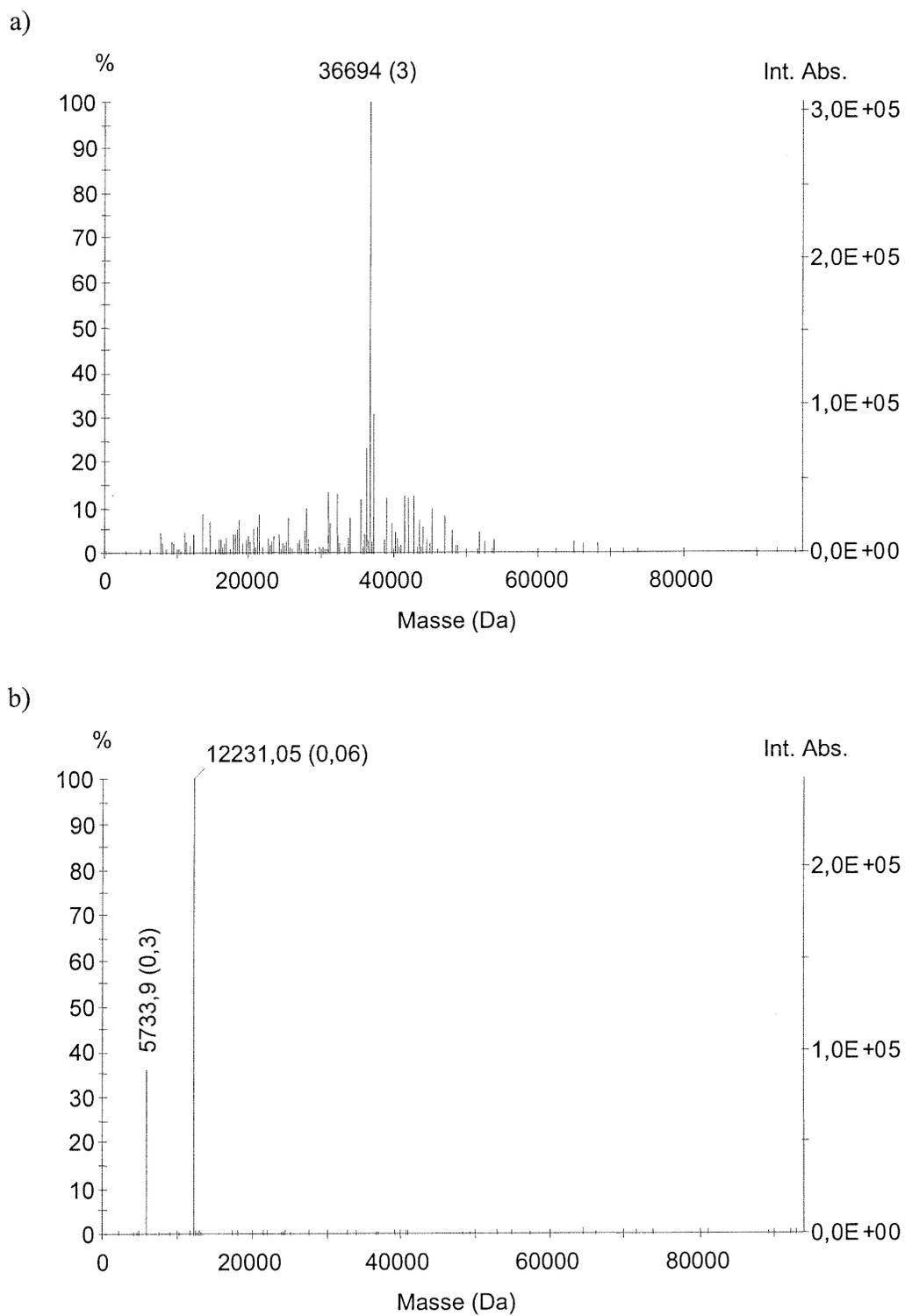


Figure 30 : Spectres déconvolués par l'utilisation directe de l'algorithme Probabilien d'un mélange simulé d'insuline et de Cytochrome C. a) À partir du spectre non filtré. b) À partir du spectre filtré.

L'algorithme de détermination des positions de haute densité de masse permet d'obtenir le spectre déconvolué de la Figure 31. Les résultats obtenus sont pratiquement les mêmes que pour l'algorithme Probablion. Les deux algorithmes réagissent donc bien face à des échantillons plus complexes.

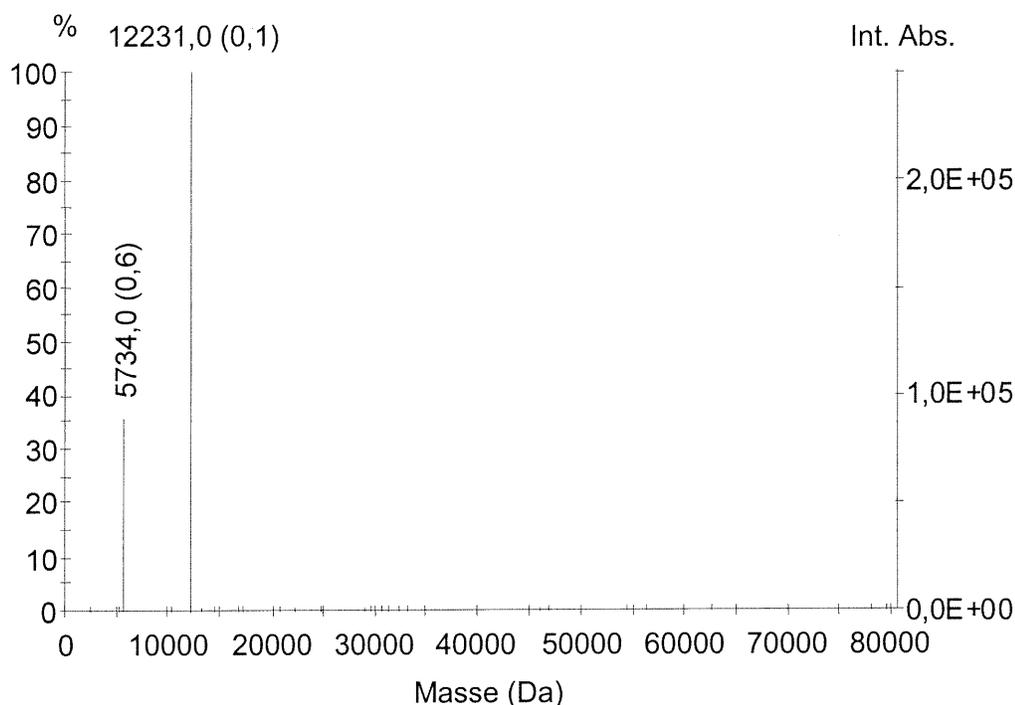


Figure 31 : Spectre déconvolué obtenu par l'algorithme Dension pour le spectre de masse filtré du mélange simulé de l'insuline et du cytochrome C.

Mélange binaire

Le premier des vrais mélanges est celui de l'ubiquitine (8565 Da) et du cytochrome C (12230 Da) dans des proportions de 30:70. Les spectres de masse accumulés avant et après la filtration sont montrés dans la Figure 32.

L'utilisation directe de l'algorithme de détermination des distributions les plus probables sur les spectres de masse du mélange donne les spectres déconvolués de la Figure 33.

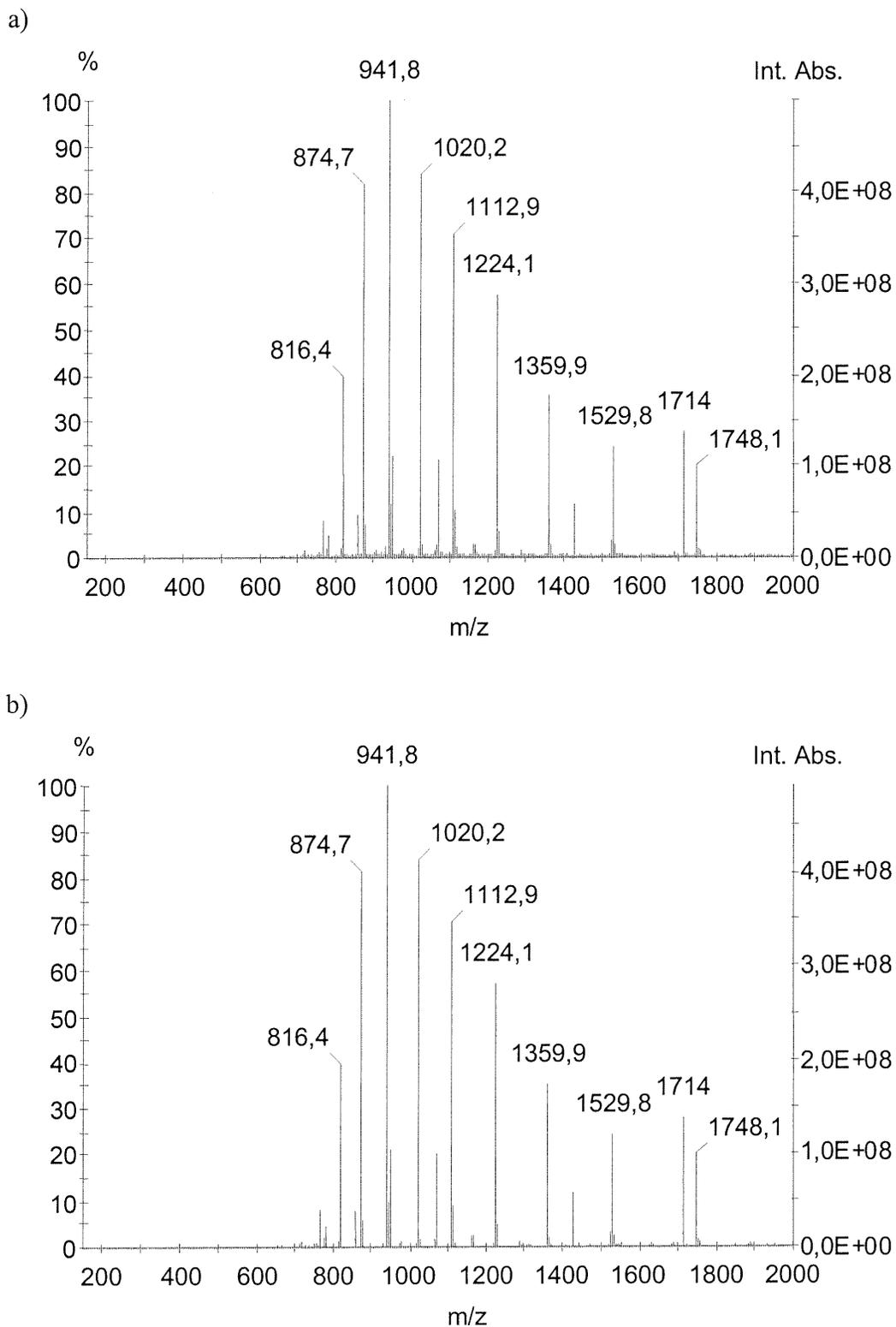


Figure 32 : Spectres de masse accumulés (80 spectres) d'un mélange de l'ubiquitine (8565 Da) et du cytochrome C (12230 Da). a) Spectre non filtré. b) Spectre filtré.

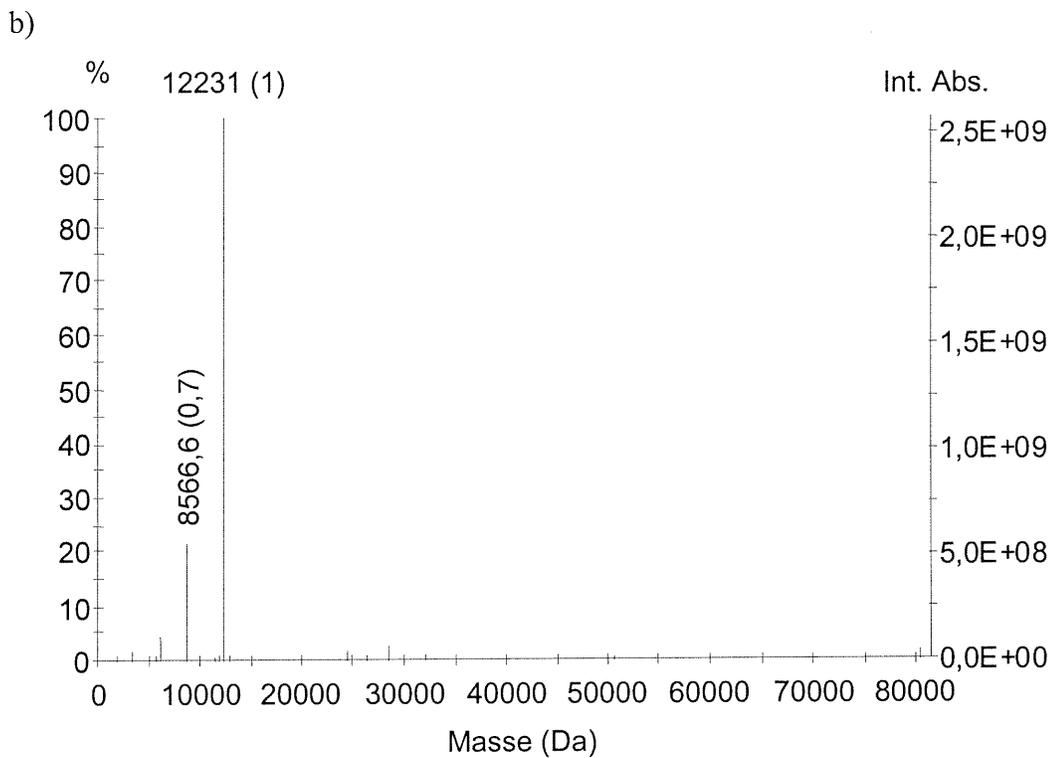
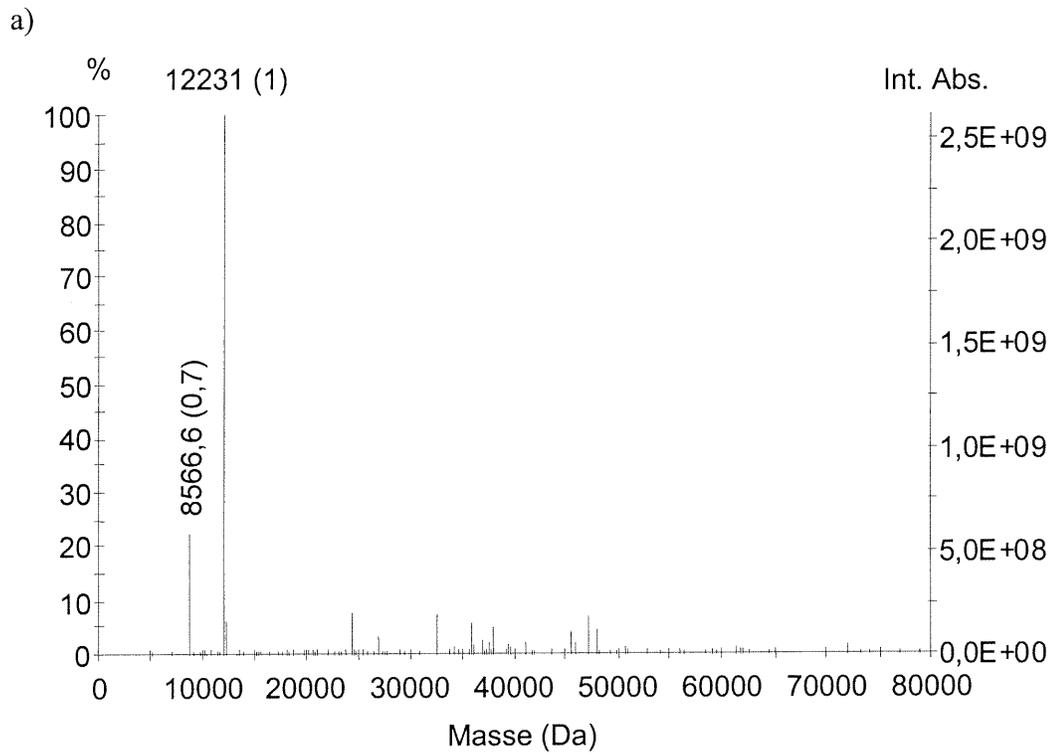


Figure 33 : Spectres déconvolués par utilisation directe de l'algorithme Probablion d'un mélange d'ubiquitine et du Cytochrome C. À partir du spectre non filtré (a) et filtré (b).

Dans le cas de cette protéine, la déconvolution est peu influencée par la filtration. En effet, l'utilisation directe de l'algorithme Probablion permet de trouver les deux protéines aux bonnes masses avec ou sans le filtrage. Le filtrage permet quand même de réduire la quantité de faux pics dans le spectre déconvolué même si au départ le filtrage semblait sans effet majeur sur le spectre de masse de l'échantillon.

L'utilisation inverse de l'algorithme Probablion donne le spectre déconvolué de la Figure 34. Le signal de fond présent dans ce spectre déconvolué est négligeable et il n'y a aucune ambiguïté sur le contenu de l'échantillon dans ce cas.

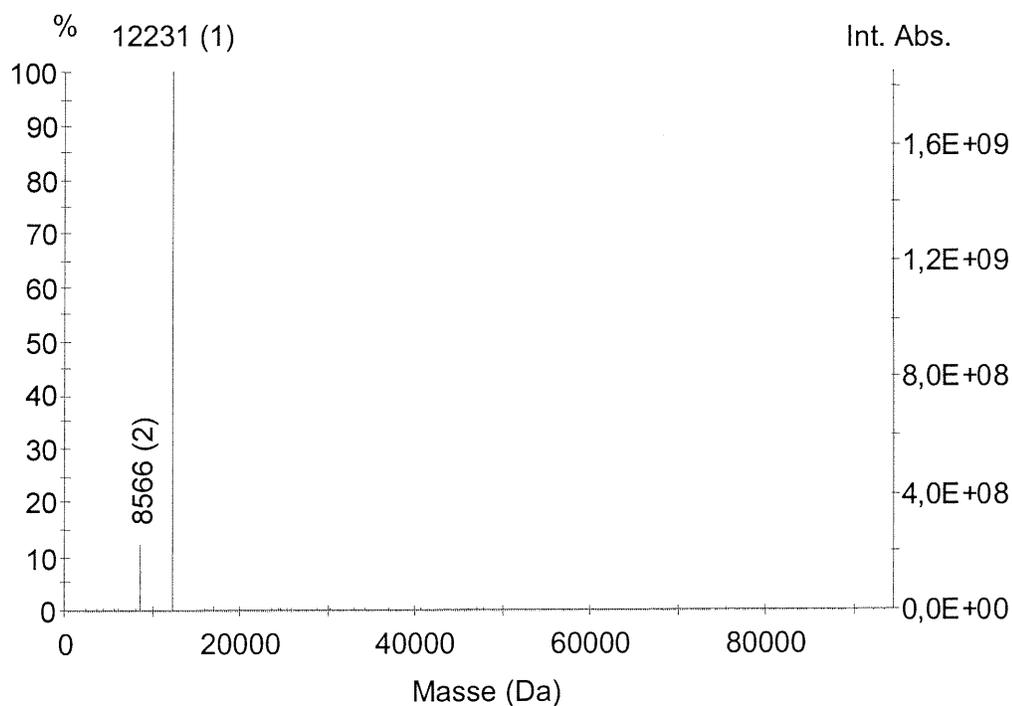


Figure 34 : Spectre déconvolué du mélange d'ubiquitine et du cytochrome C par l'utilisation inverse de l'algorithme Probablion sur 80 spectres de masse.

L'algorithme Dension quant à lui donne le spectre déconvolution montré dans la Figure 35.

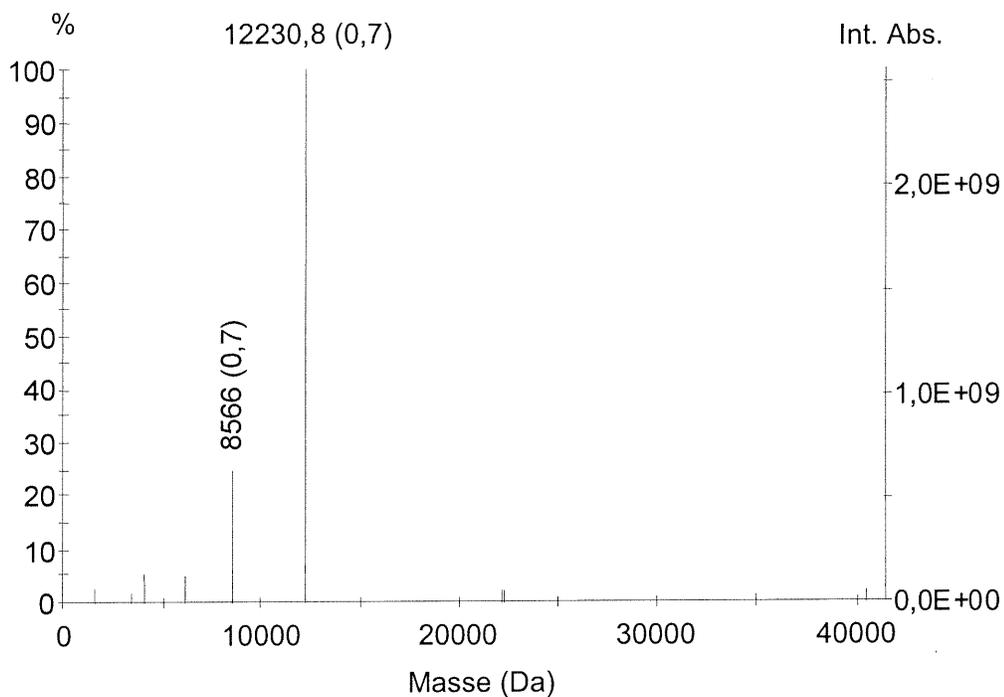


Figure 35 : Spectre déconvolué par l'algorithme Dension du spectre de masse accumulé et filtré du mélange d'ubiquitine et du cytochrome C.

Quelques pics provenant du signal de fond sont présents dans ce spectre déconvolué. L'intensité de certains d'entre eux est suffisamment grande pour que cela soit difficile de juger de leur authenticité avec une grande certitude dans le cas d'échantillons inconnus.

Mélange tertiaire

Le deuxième mélange étudié est celui de l'ubiquitine (8565 Da), du cytochrome C (12230 Da) et de la myoglobine (16952 Da) dans des proportions de 1 :1 :1. Les spectres de masse accumulés non filtrés et filtrés sont montrés dans la Figure 36.

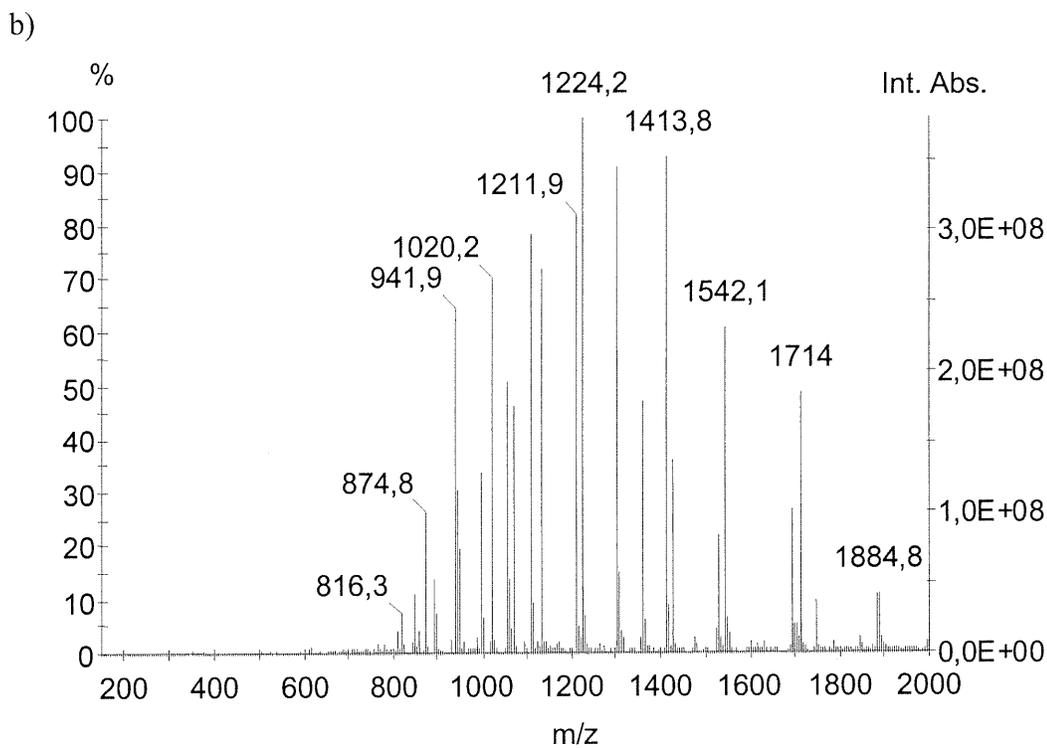
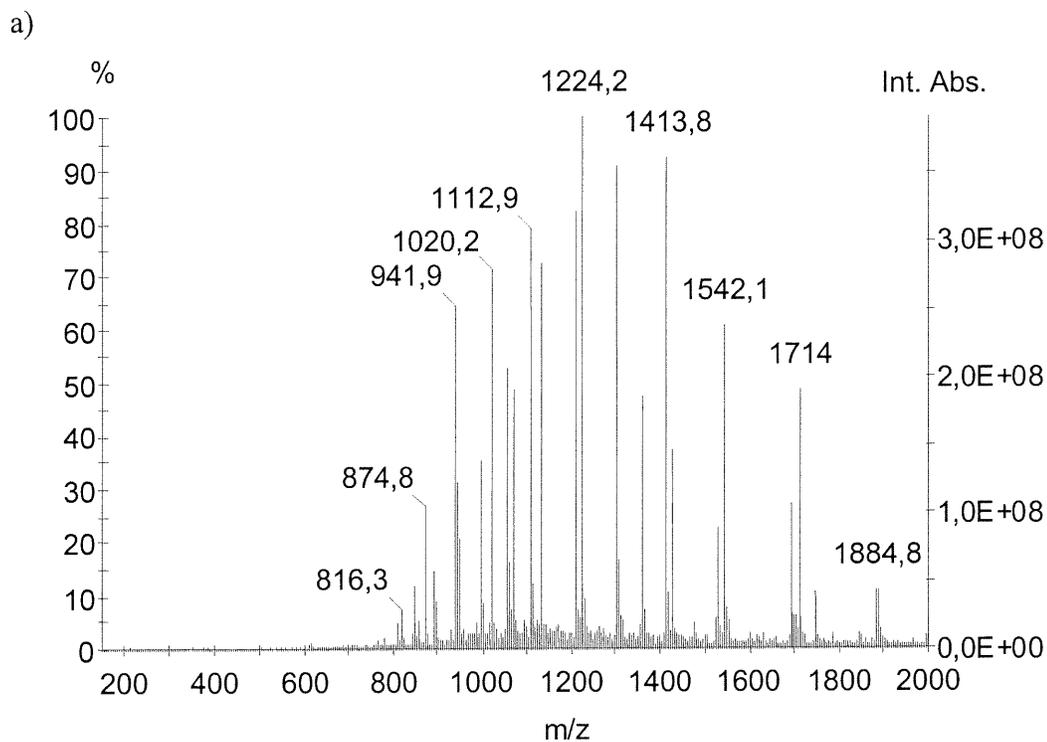


Figure 36 : Spectres de masse accumulés (76 spectres) d'un mélange de l'ubiquitine (8565 Da), du cytochrome C (12230 Da) et de la myoglobine (16952 Da). a) Spectre non filtré. b) Spectre filtré.

Il est possible de voir que le signal de fond continu est assez intense dans le spectre de masse de cet échantillon. En général, l'intensité du signal de fond continu autour des distributions d'ions multichargés dépend du nombre de pics qu'elles contiennent. Plus il y a de pics dans les distributions d'ions multichargés et plus l'intensité du signal de fond continu autour de celle-ci tend à être grande. En effet, si le nombre d'états chargés augmente, un même courant ionique devra être réparti entre plus de pics. Chaque pic aura donc une plus petite part du courant ionique et l'intensité apparente de la distribution diminuera. Ceci donne l'impression que le signal de fond continu augmente, mais ce sont les vrais pics qui diminuent d'intensité en réalité. Bien sûr cette variation du signal de fond continu n'est pas seulement due à ce phénomène. En effet, il y a aussi la sensibilité de la technique face à la protéine qui varie d'une protéine à l'autre. L'ionisation de certaines protéines est plus efficace que pour d'autres.

Lors de l'accumulation, il se développe une certaine enveloppe autour de chacun de ces pics à cause des variations des valeurs de m/z des pics d'un spectre à l'autre. Plus il y a de pics, plus ceux-ci sont rapprochés. La variation des valeurs de m/z peut alors contribuer à la création du signal de fond continu. Le fait d'avoir des mélanges de protéines complique alors les spectres de masse puisque cela augmente aussi le nombre de pics présents. Le filtrage devient donc de plus en plus nécessaire pour les mélanges de protéines afin de faciliter la déconvolution.

La Figure 37 montre la déconvolution par l'utilisation directe de l'algorithme de détermination des distributions les plus probables.

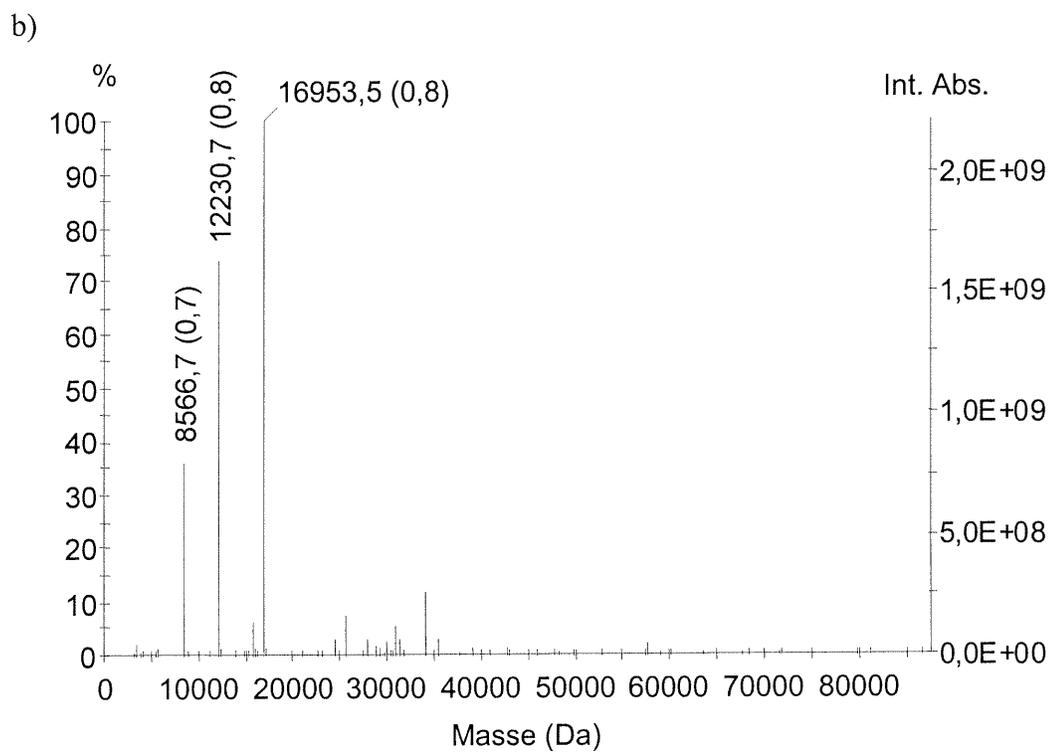
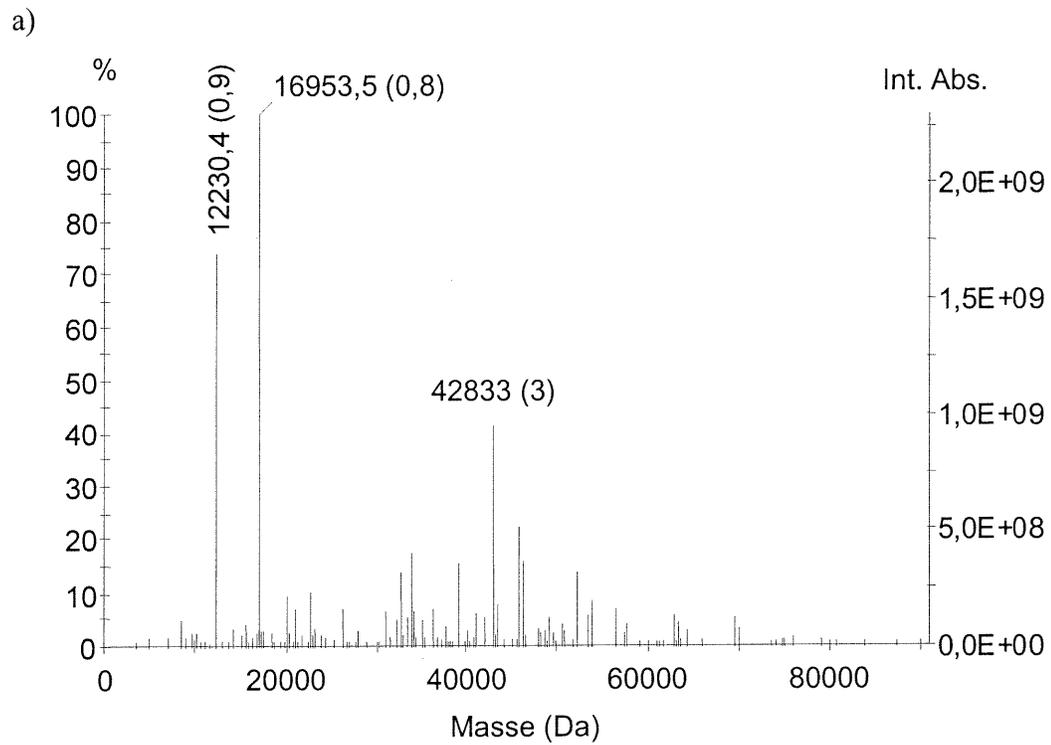


Figure 37 : Spectres déconvolués par l'utilisation directe de l'algorithme Probablion d'un mélange d'ubiquitine, du cytochrome C et de myoglobine. a) À partir du spectre non filtré. b) À partir du spectre filtré.

Il est évident que la filtration change drastiquement les résultats obtenus. Dans le cas du spectre de masse non filtré, la déconvolution trouve une masse deux fois trop grande pour une des protéines. Par contre, avec le spectre de masse filtré, la déconvolution trouve les bonnes masses et le nombre de faux pics est grandement réduit.

L'utilisation inverse de l'algorithme de détermination des distributions les plus probables mène au spectre déconvolué de la Figure 38.

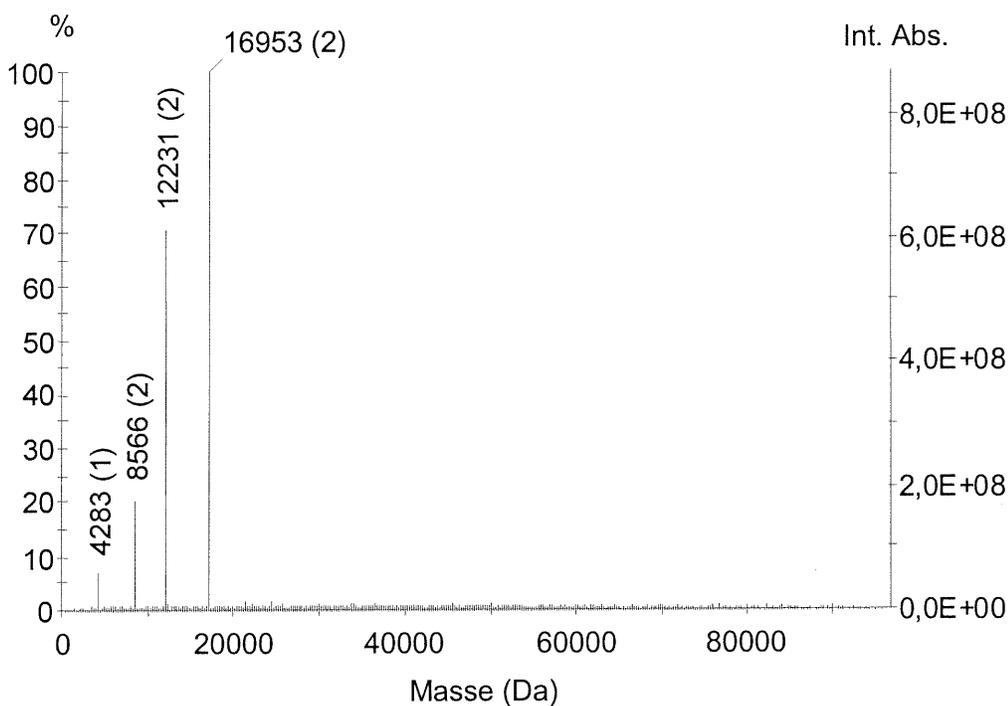


Figure 38 : Spectre déconvolué du mélange d'ubiquitine, du cytochrome C et de la myoglobine par l'utilisation inverse de l'algorithme Probablion.

Le spectre déconvolué obtenu par cette méthode possède une grande quantité de faibles pics dus au signal de fond. Ceux-ci formant un signal de fond continu simple, il est facile de les enlever avec n'importe quel type de filtration. Le fait que

cette utilisation de l'algorithme Probablion n'utilise pas la filtration amène la création d'une harmonique (4283 Da) à la moitié de la masse de l'ubiquitine (8566 Da). Le spectre est quand même assez facile à interpréter puisque le faux pic se trouve exactement à une fraction entière d'un vrai pic.

L'utilisation de l'algorithme de détermination des positions de haute densité de masse sur le spectre de masse accumulé et filtré du mélange tertiaire donne le spectre déconvolué de la Figure 39.

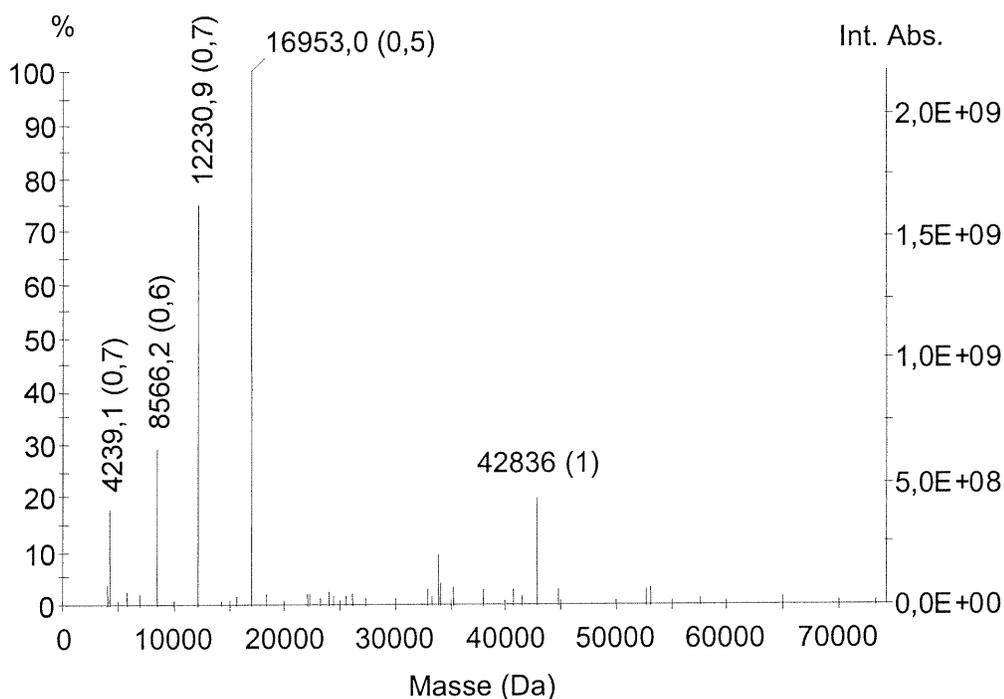


Figure 39 : Spectre déconvolué par l'algorithme Dension du spectre de masse accumulé et filtré du mélange d'ubiquitine, du cytochrome C et de la myoglobine.

Le spectre déconvolué obtenu par l'algorithme Dension possède plus de faux pics que les spectres déconvolués obtenus avec les autres algorithmes. Les pics principaux sont discernables mais il serait difficile d'affirmer que les autres pics sont

des faux pics dans un échantillon inconnu puisque leurs intensités sont du même ordre de grandeur que celle du composant mineur du mélange. Ceci est probablement dû au fait que l'algorithme n'inclut pas toujours tous les pics d'une distribution d'ions multichargés dans une zone de hautes densités de masse et que les pics restants peuvent être inclus dans d'autres zones de hautes densités provenant du signal de fond. En effet, certains pics d'une distribution d'ions multichargés ont un grand écart à la moyenne et ils seront juste à l'extérieur de la zone de hautes densités où ils devraient se situer.

Mélange quaternaire

Le troisième mélange étudié est celui de l'insuline (5734 Da), de l'ubiquitine (8565 Da), du cytochrome C (12230 Da) et de la myoglobine (16952 Da) dans les proportions relatives de 20 : 1 : 1 : 1 respectivement. Les spectres de masse obtenus par l'accumulation des 83 spectres de masse sont montrés dans la Figure 40 avant et après la filtration. Dans cet échantillon, l'insuline était beaucoup plus concentrée que les trois autres composants du mélange et ceci est évident lorsque le spectre de masse est observé. En effet, la distribution d'ions multichargés de l'insuline est majoritaire et les autres distributions sont presque indiscernables par rapport à celle-ci. Il est donc intéressant de voir si le processus itératif permettra de distinguer les distributions d'ions multichargés de protéines lorsqu'une autre protéine est grandement majoritaire.

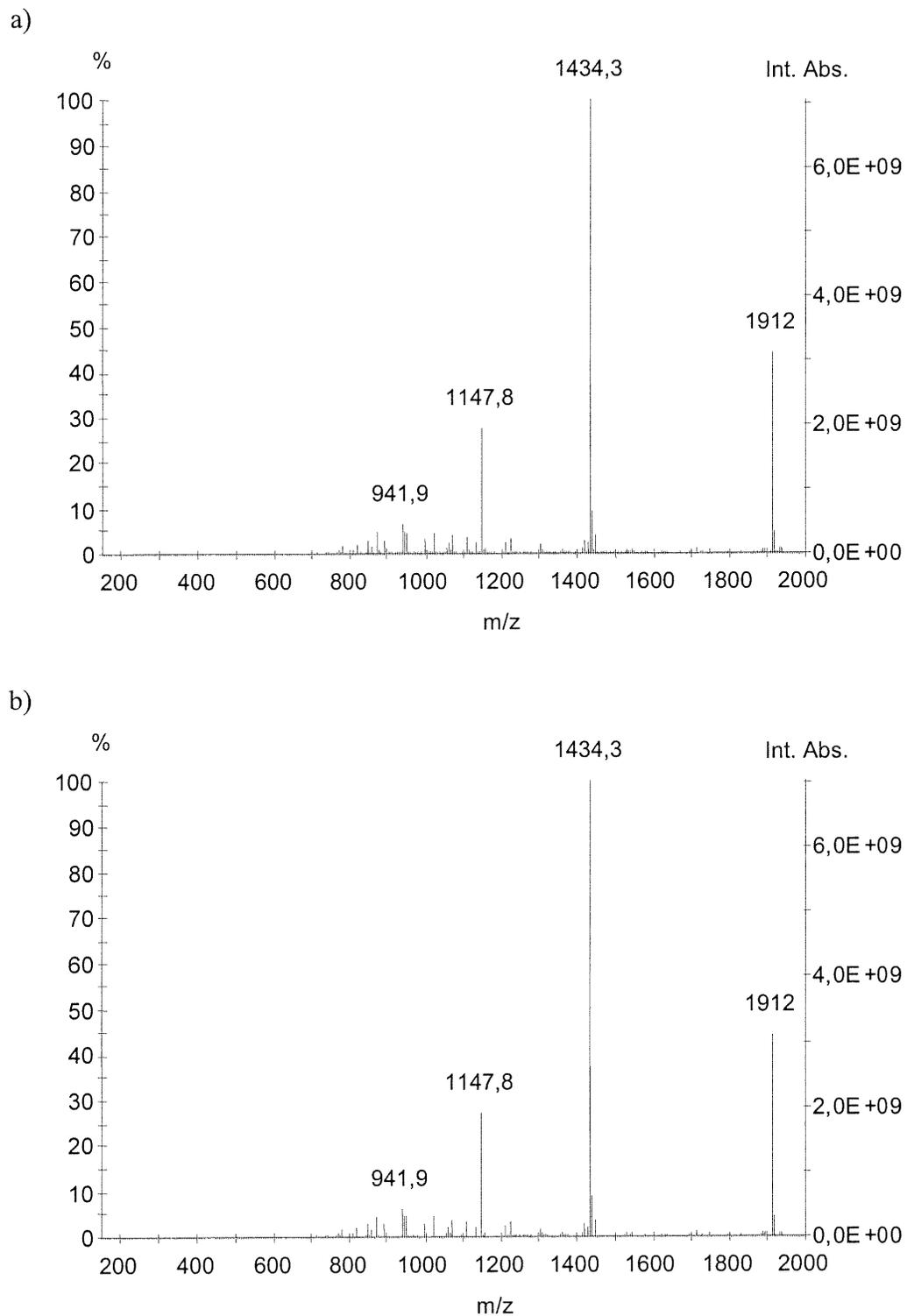


Figure 40 : Spectres de masse accumulés (83 spectres) d'un mélange de l'insuline (5734 Da), d'ubiquitine (8565 Da), du cytochrome C (12230 Da) et de la myoglobine (16952 Da). a) Spectre non filtré. b) Spectre filtré.

La déconvolution de ces spectres de masse par l'utilisation directe de l'algorithme de détermination des distributions les plus probables est montrée dans la Figure 41.

Les spectres déconvolués obtenus à partir des spectres filtrés et non filtrés sont semblables, mais une des protéines n'a pu être placée à la bonne masse dans le spectre déconvolué du spectre de masse non filtré.

Le filtrage a donc permis de réduire la quantité de faux pics présents dans le spectre déconvolué et de permettre de placer les protéines aux bonnes masses. Le fait qu'une des protéines soit grandement majoritaire ne semble pas affecter beaucoup la déconvolution. En effet, le processus itératif a permis de détecter toutes les protéines une à la fois.

La justesse de la détermination des masses des composants mineurs semble par contre avoir été affectée. En effet, les pics doivent être légèrement déformés par la grande quantité d'ions provenant de l'insuline puisque les masses moléculaires déterminées sont déplacées d'un à deux Da par rapport aux masses des déconvolutions d'autres échantillons.

L'utilisation inverse de l'algorithme Probablion a permis d'obtenir le spectre déconvolué de la Figure 42. Les quatre protéines y sont présentes. Les faux pics présents dans le spectre déconvolué ne sont pas suffisamment intenses pour qu'il puisse y avoir confusion quant à la composition du mélange.

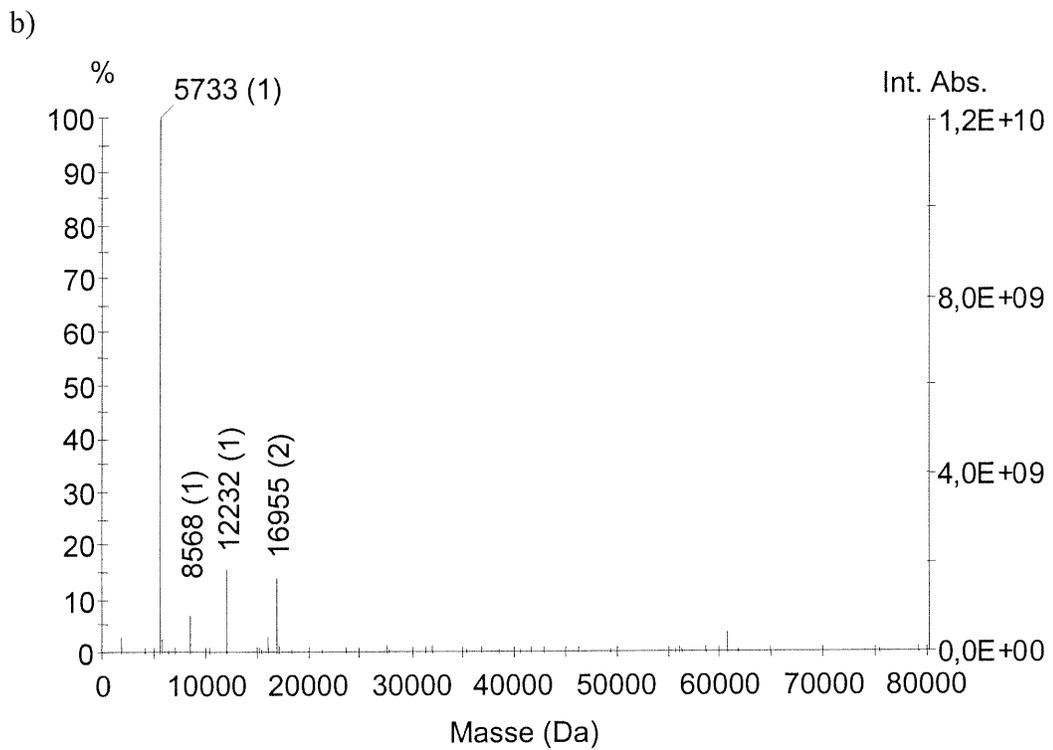
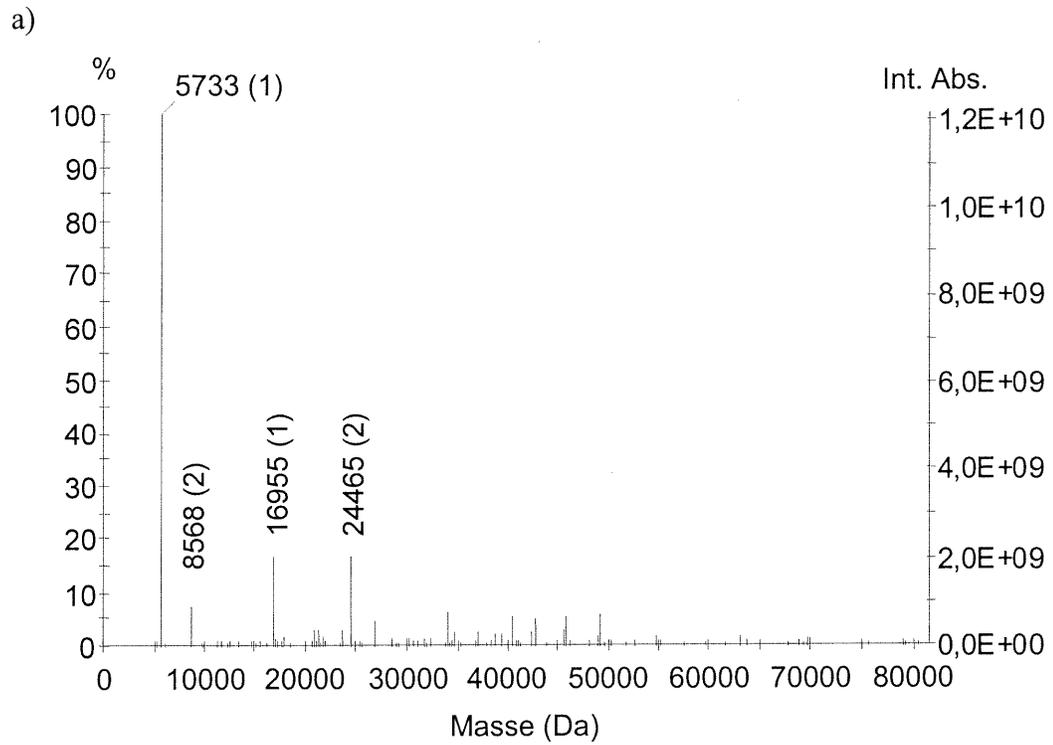


Figure 41 : Spectres déconvolués par l'utilisation directe de l'algorithme Probabilion d'un mélange d'insuline, d'ubiquitine, du cytochrome C et de myoglobine. a) À partir du spectre non filtré. b) À partir du spectre filtré.

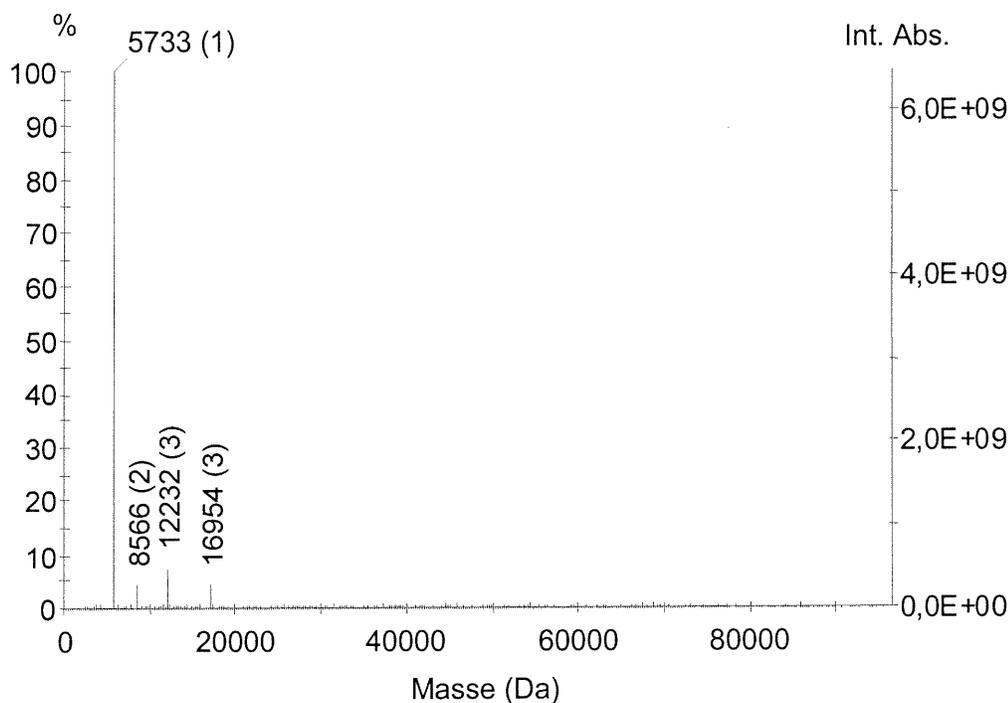


Figure 42 : Spectre déconvolué du mélange d'insuline, d'ubiquitine, du cytochrome C et de la myoglobine par l'utilisation inverse de l'algorithme Probablion au 83 spectres de masse.

L'algorithme de détermination des positions de haute densité de masse a été utilisé pour produire le spectre déconvolué de la Figure 43. Le spectre déconvolué obtenu est simple et ne présente aucune ambiguïté majeure ce qui est différent du spectre déconvolué de la Figure 39 qui possédait plus de faux pics. Ceci est probablement dû à une plus petite variation des valeurs de m/z autour des valeurs attendues pour les pics des distributions d'ions multichargées du mélange quaternaire.

Tous les algorithmes conçus semblent donc aptes à traiter des mélanges dont les concentrations des protéines sont dispersées sur différents ordres de grandeur. La capacité de détecter des composants mineurs devrait permettre d'éviter un certain nombre de séparations avant les analyses des échantillons.

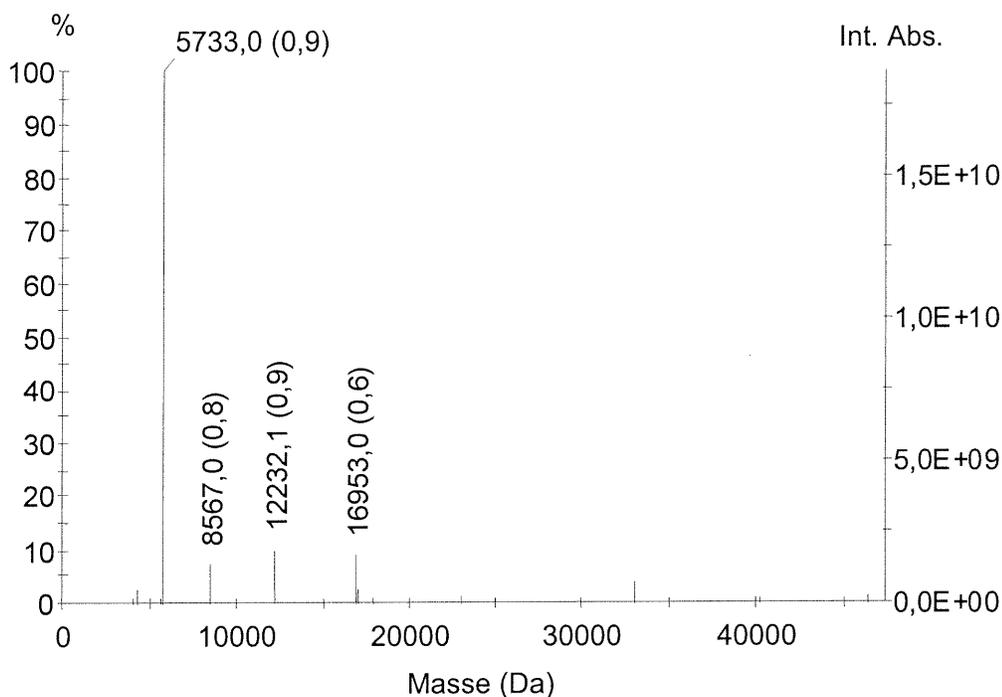


Figure 43 : Spectre déconvolué par l'algorithme Dension du spectre de masse accumulé et filtré du mélange d'insuline, d'ubiquitine, du cytochrome C et de la myoglobine.

Échantillon inconnu

Suite à une présentation à la conférence de l'American Society for Mass Spectrometry (ASMS) à Chicago, Dr. Neil L. Kelleher, professeur à l'Université d'Illinois, a approché notre laboratoire pour avoir accès au logiciel de déconvolution. Il nous a donc fait parvenir un spectre de masse d'un échantillon traité dans son laboratoire. Le spectre de masse obtenu a été pris avec un appareil maison de résonance cyclotronique avec ESI. Il a spécifié que l'échantillon devrait contenir deux protéines de masse 16071,7 et 36845,0 Da selon l'analyse qu'il avait effectuée avec l'algorithme de Zhang et Marshall¹².

Le fichier envoyé était brut c'est-à-dire qu'aucune reconnaissance de pics n'avait été effectuée automatiquement par le spectromètre de masse comme dans le

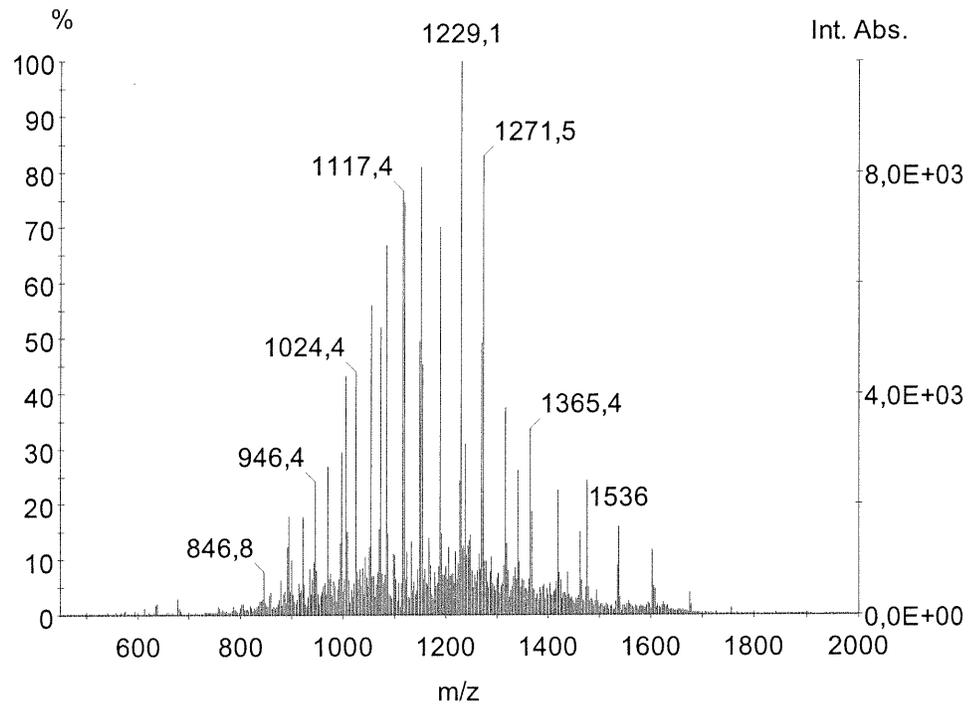
cas de notre instrument. Leur appareil mesure donc l'intensité des ions pour tous les centièmes de m/z ce qui donne un spectre de masse en mode continu. La démarche habituelle dans de telles situations est d'utiliser immédiatement un logiciel de reconnaissance de pics afin de transformer le fichier en une liste de pics contenant l'aire des pics et la valeur de m/z moyenne. Or, pour la présente étude, il a été tenté, compte tenu du signal de fond évident dans le spectre de masse en continu, de filtrer le spectre avant la reconnaissance de pic. La Figure 44 montre donc les spectres de masse non filtré et filtré pour l'échantillon inconnu avant la reconnaissance de pics. La filtration augmente la vitesse de traitement de la reconnaissance de pics puisque ceux-ci sont plus facile à trouver.

La Figure 45a montre le spectre de masse de l'échantillon inconnu après la reconnaissance de pics. Ce spectre de masse a aussi été filtré afin d'enlever le maximum du signal de fond continu et le résultat est montré dans la Figure 45b.

Les deux spectres de la Figure 45 ont été traité avec l'utilisation directe de l'algorithme de détermination des distributions les plus probables et les résultats obtenus sont montrés dans la Figure 46.

Les deux protéines que Dr. Kelleher avait mentionnées sont détectées dans les deux spectres déconvolués. Le spectre déconvolué obtenu à partir du spectre de masse non filtré contient évidemment une grande quantité de faux pics comme il est normal de s'y attendre. Le spectre déconvolué obtenu à partir du spectre filtré a aussi quelques faux pics ce qui n'est pas trop anormal. Par contre, le fait que les pics soient présents dans les spectres déconvolués du spectre non filtré et filtré est plus étonnant. Cela suggère que ces pics ne sont pas des faux pics aléatoires car la filtration est normalement suffisante pour les changer de masse dans les spectres déconvolués.

a)



b)

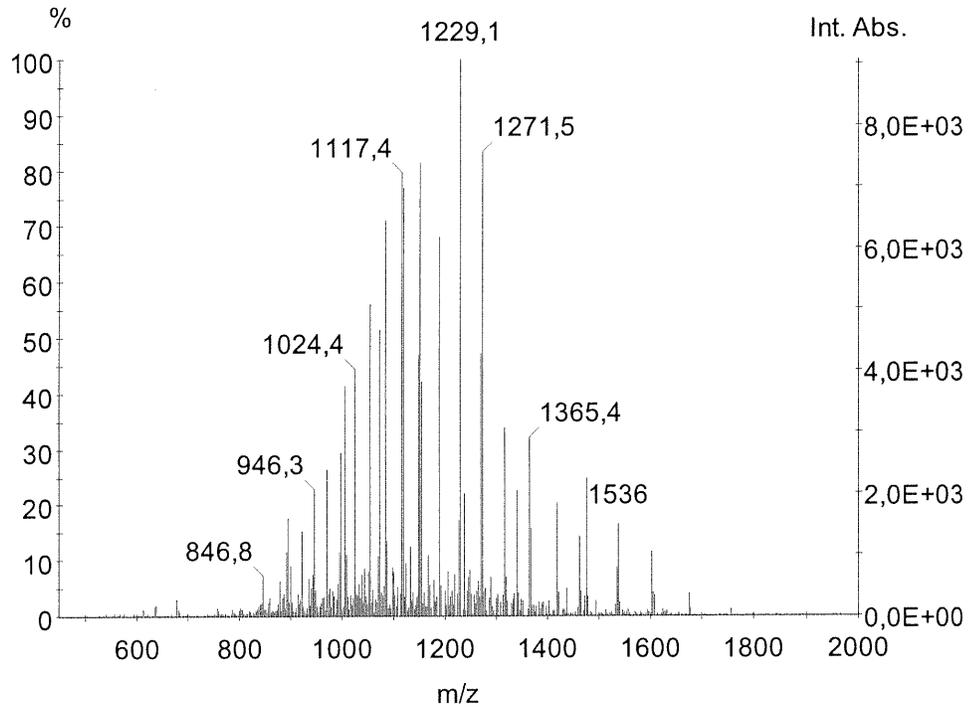


Figure 44 : Spectres de masse de l'échantillon inconnu en format continu. a) spectre non filtré. b) spectre filtré.

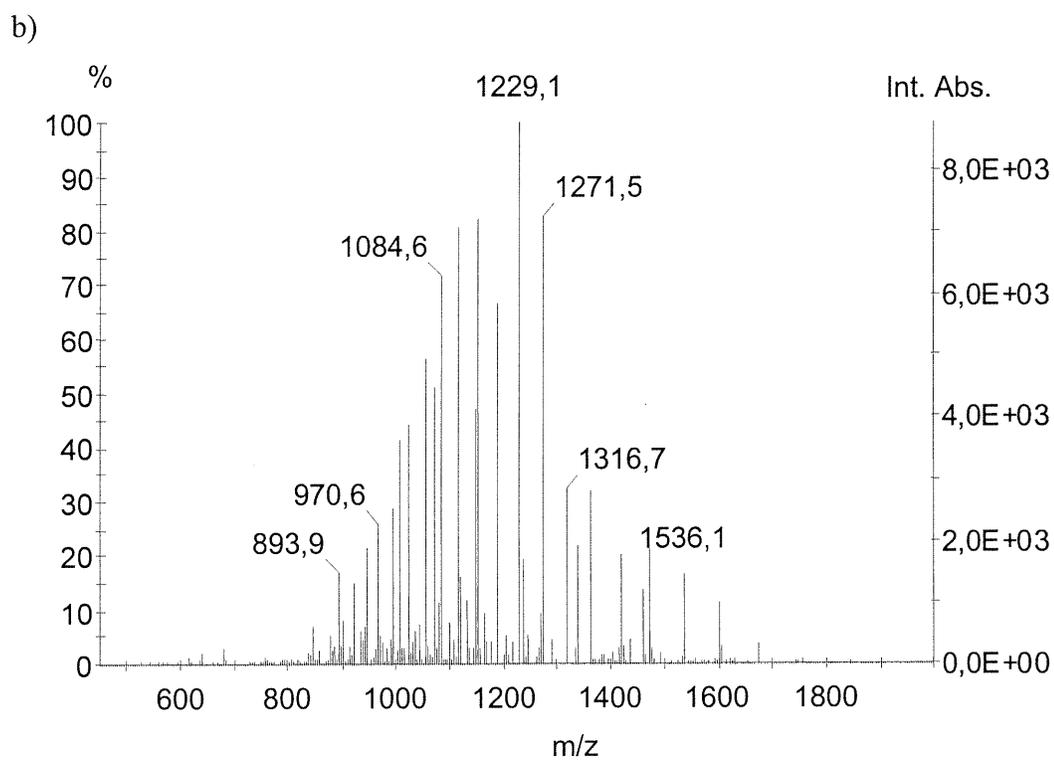
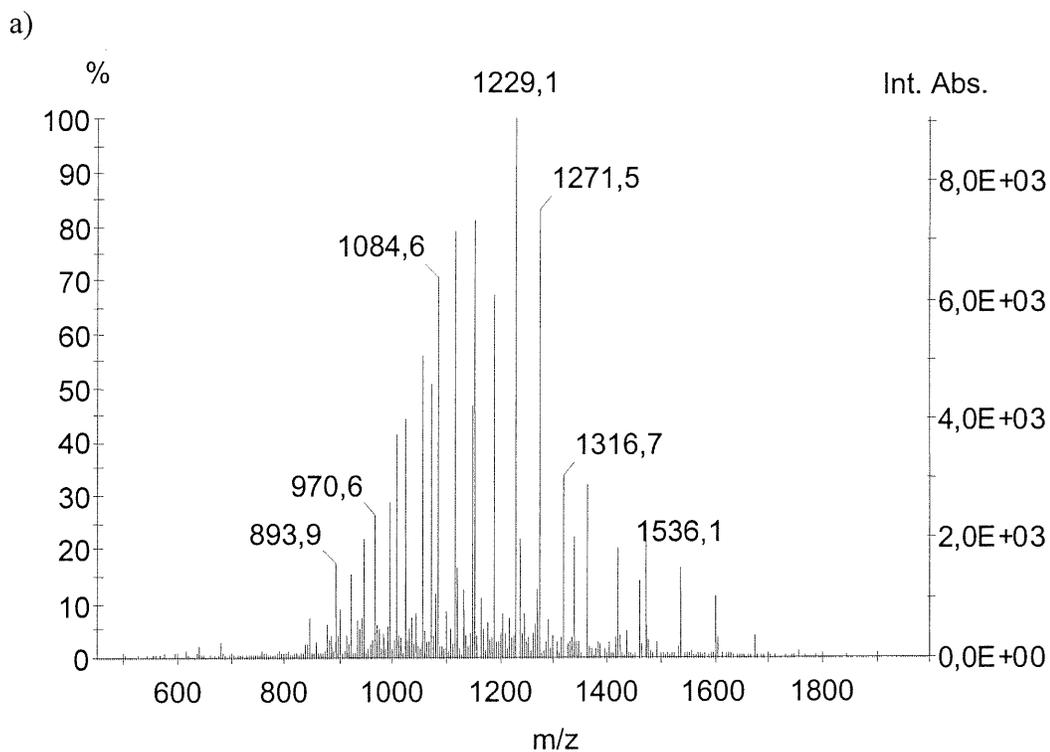


Figure 45 : Spectres de masse de l'échantillon inconnu après la reconnaissance de pic. a) Spectre obtenu à partir du spectre de masse en continu filtré. b) Spectre de masse après filtration de a.

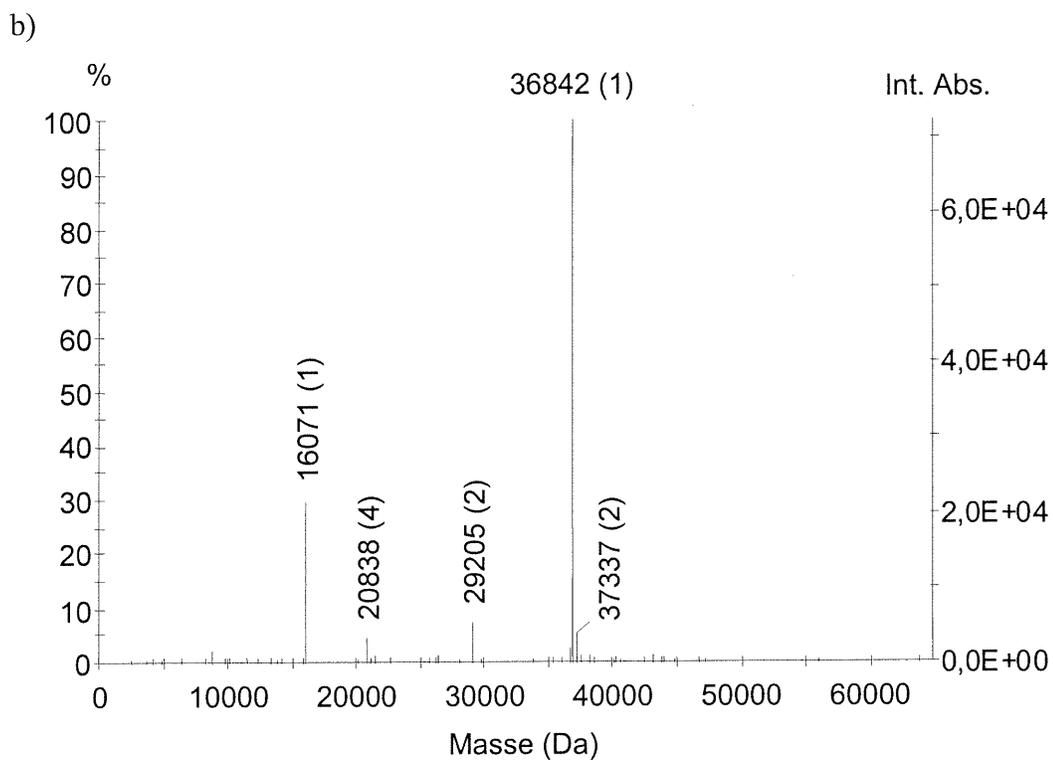
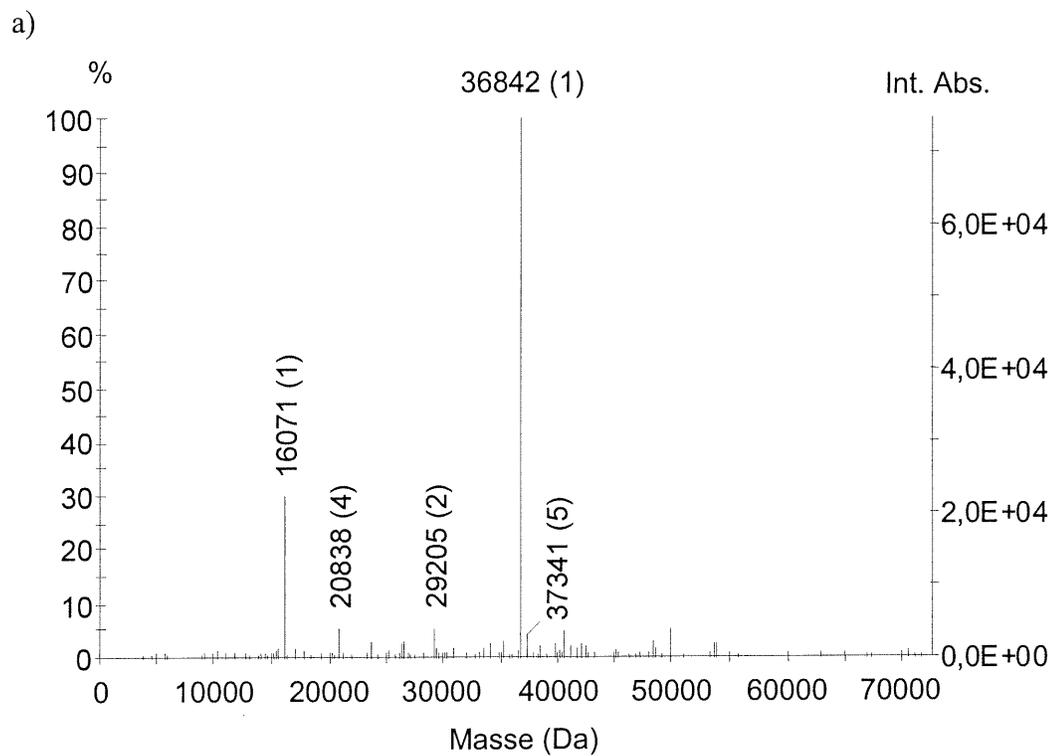


Figure 46 : Spectres déconvolués des spectres de masse de la Figure 45 avec l'algorithme Probablion. a) à partir du spectre de masse non filtré. b) à partir du spectre de masse filtré.

Comme ces pics ne sont pas des valeurs multiples ou des fractions entières de vrais pics, ce ne sont pas des harmoniques non plus. Il reste donc l'hypothèse que ce sont des pics dus à des espèces présentes dans l'échantillon.

Pour tester cette hypothèse, le spectre de masse filtré a aussi été traité par l'algorithme de détermination des positions de haute densité de masse. Le spectre déconvolué obtenu est montré dans la Figure 47.

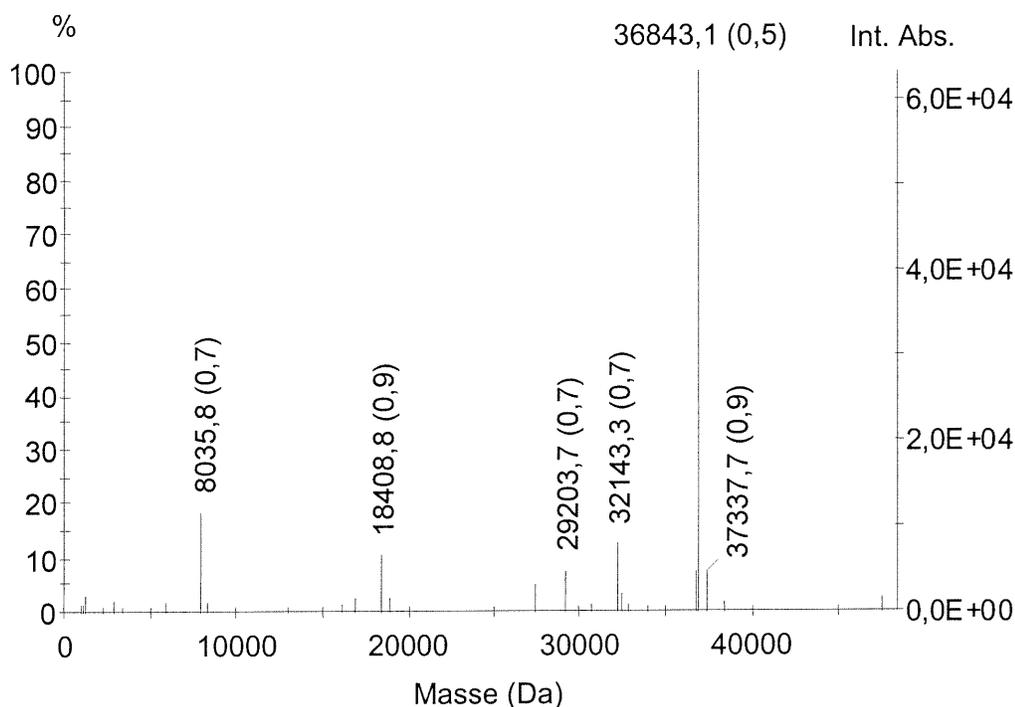


Figure 47 : Spectre déconvolué du spectre de masse filtré de la Figure 45b avec l'algorithme Dension.

Le spectre déconvolué montre une des protéines principales, mais pas la deuxième. En fait, celle-ci est présente, mais à la moitié de la masse sous forme d'harmonique. Deux des autres pics inexplicés du spectre déconvolué par l'algorithme Probablion sont aussi présents. Cela renforce l'hypothèse que ces deux pics sont dus à des espèces dans l'échantillon puisqu'ils sont aussi montrés par un

processus indépendant. Pour ce qui est du troisième pic, il est impossible d'affirmer qu'il provient d'une protéine mais il est aussi impossible d'affirmer le contraire.

Le spectre déconvolué de la Figure 46b a donc été envoyé à Dr. Kelleher en lui expliquant que les pics supplémentaires étaient probablement des espèces de plus dans l'échantillon, mais qu'il était possible qu'un des pics soit dû au signal de fond. Dr. Kelleher a répondu en expliquant qu'en effet, l'échantillon contenait deux composants mineurs et qu'il était étonné qu'il soit possible de les détecter.

Discussion des algorithmes

Les algorithmes développés offrent chacun des avantages et des limitations. L'utilisation directe de l'algorithme de détermination des distributions les plus probables, Probablion, offre l'avantage d'être rapide. En effet, l'accumulation, le filtrage et la déconvolution d'un spectre de masse typique prennent environ 4 s sur un ordinateur Pentium III 750 MHz. L'utilisation inverse de l'algorithme Probablion quant à elle est plus lente puisque chaque spectre de masse doit être déconvolué individuellement. De plus, cette méthode déconvolue les spectres de masse non filtrés. Le temps de calcul dépend donc du nombre de spectres de masse. La moyenne est d'environ 3 s par spectre de masse sur le même ordinateur. L'algorithme de détermination des positions de haute densité de masse, Dension, est lent aussi, bien que celui-ci n'ait pas été optimisé pour la vitesse. Le temps moyen de calcul est d'environ 40 s pour l'accumulation, le filtrage et la déconvolution. Dans les deux cas, le filtrage diminue le temps de calcul d'environ la moitié du temps nécessaire.

L'utilisation inverse de l'algorithme Probablion n'utilisant pas la filtration amène parfois la création d'harmoniques *en plus* des vrais pics. L'algorithme

Densition et l'utilisation directe de l'algorithme Probablion font appel à la filtration. Ceci amène la réduction de l'occurrence des harmoniques. Par contre, dans le cas de ces deux méthodes, la détection d'harmoniques se fait au détriment des vrais pics. Ceci rend la détermination des masses des protéines délicate puisqu'il est difficile d'affirmer qu'un pic est une harmonique si le vrai pic est absent. Pour l'utilisation inverse de l'algorithme Probablion, les harmoniques sont souvent de faibles intensités puisqu'elles ne sont pas accumulées aussi efficacement que les vrais pics. Ceci rend donc leur détection plus aisée.

La quantité de faux pics aléatoires dus au signal de fond dépend aussi grandement de la filtration des spectres de masse. Encore une fois, l'utilisation inverse de l'algorithme Probablion n'utilisant pas la filtration, la quantité de faux pics y est grandement augmentée. De plus, l'accumulation ajoute au spectre déconvolué moyen les faux pics de tous les spectres déconvolués individuels. Par contre, puisque l'accumulation des pics aléatoires n'est pas efficace, le rapport du signal sur le signal de fond augmente. Ceci amène donc la formation d'un signal de fond continu qu'il est ensuite relativement facile d'enlever pour obtenir un spectre déconvolué filtré. Pour les autres algorithmes, les faux pics provenant du signal de fond, bien que plus faibles, sont difficiles à enlever puisqu'il n'y a rien qui les différencie des vrais pics de façon certaine à 100%.

Les valeurs de la masse des protéines déterminées proviennent de plusieurs pics d'une distribution d'ions multichargés pour l'algorithme Densition et l'utilisation directe de l'algorithme Probablion. Ceci rend donc la valeur plus juste. L'utilisation inverse de l'algorithme Probablion use aussi de tous ces pics, mais la répétition de la

déconvolution montre aussi la distribution des valeurs de masse de la protéine autour de la valeur calculée. La valeur obtenue provenant d'un plus grand échantillonnage rend la mesure statistiquement plus fiable.

Le traitement des mélanges est possible dans le cas de tous les algorithmes. L'algorithme Probablion est par contre plus apte à bien déconvoluer les spectres de masse de mélanges. En effet, l'algorithme Dension est plus sensible au bon déroulement de la déconvolution. C'est-à-dire qu'il est possible qu'un des pics d'une protéine interfère avec la distribution d'ions multichargés d'une autre protéine. Si cela se produit au début de la déconvolution, celle-ci peut bien en être ruinée. Cette possibilité est plus probable dans le cas de l'algorithme Dension, puisque aucun contrôle de la distribution d'ions multichargés n'est effectué. En effet, l'algorithme Probablion ajoute des pics à la distribution d'ions multichargés potentielle de façon séquentielle et arrête lors d'une incongruité.

L'utilisation inverse de l'algorithme Probablion nécessite un grand nombre de spectres de masse. Or, ceci n'est pas toujours possible dans le cas où la quantité d'échantillon est limitée, par exemple, lors de l'analyse par spectrométrie de masse de séparations chromatographiques en temps réel. Les deux autres méthodes quant à elles peuvent être utilisées sur des spectres de masse uniques ou un petit nombre de spectres. La filtration risque alors d'être moins efficace et la déconvolution peut en souffrir, mais elle est toujours possible. Lors de ces situations, il est alors souhaitable d'obtenir les spectres de masse en mode continu et de filtrer ceux-ci avant la reconnaissance de pics comme dans le cas de l'échantillon de Dr. Kelleher. Ceci offre l'avantage d'aider la reconnaissance de pics et de réduire la quantité de faux

pics provenant du signal de fond dans les spectres de masse obtenus. Le temps de calcul en souffrira par contre beaucoup, puisque les spectres de masse en mode continu sont longs à traiter et très volumineux. Lorsque les données expérimentales sont très limitées, l'automatisation de la déconvolution est plus difficile et le jugement de l'utilisateur est de plus en plus nécessaire.

L'utilisation en parallèle des algorithmes de déconvolution semble être utile pour la confirmation des résultats obtenus et la différenciation de faux pics dus au signal de fond et de vrais pics. Ceci a été démontré dans le cas de l'échantillon du Dr. Kelleher dans lequel deux composants mineurs ont été détectés. L'automatisation de cette combinaison d'algorithmes n'a par contre pas été effectuée et reste difficile, mais ceci peut représenter un bon point de départ pour des futures recherches.

Chapitre VII

Conclusion Générale

La présente étude a permis de concevoir et de tester deux nouveaux algorithmes. Les deux algorithmes font appel à un algorithme de prétraitement (filtrage) appliqué pour la première fois aux spectres de masse de protéines obtenus avec la source d'ionisation Electrospray. Celui-ci permet de réduire la quantité d'harmoniques dans les spectres déconvolués et d'y réduire la quantité de faux pics en enlevant le signal de fond continu présent dans les spectres de masse. Il permet aussi de réduire le temps de calcul nécessaire pour la déconvolution.

L'algorithme de détermination des distributions les plus probables, Probablion, peut être utilisé de deux façons. L'utilisation directe est celle utilisée par tous les algorithmes conçus à ce jour. L'utilisation inverse, par contre, a servi pour la première fois et peut donner des résultats très satisfaisants si plusieurs spectres de masse sont disponibles et que le temps supplémentaire nécessaire est acceptable. En effet, cette technique peut réduire grandement le rapport signal sur signal de fond observé dans les spectres déconvolués. Les faux pics se présentent alors comme un signal de fond continu qu'il est relativement facile d'enlever pour obtenir un spectre déconvolué filtré. Les masses des protéines obtenues sont statistiquement plus fiables compte tenu du plus grand échantillonnage effectué. De plus, cet algorithme peut traiter les mélanges efficacement sans trop de difficulté grâce au fait qu'il est un processus itératif.

L'algorithme de détermination des positions de haute densité de masse, Dension, donne aussi des résultats satisfaisants. Cet algorithme est un peu plus lent que l'utilisation directe de l'algorithme Probablion. Ceci est dû au fait que toutes les

possibilités sont testées à la fois afin de déterminer la meilleure. En un sens, avec cet algorithme, les données s'organisent d'elles-mêmes et l'algorithme ne fait que déterminer la masse où la densité de pics est la plus grande par la suite. Avec l'algorithme Probablion, les possibilités qui semblent incorrectes sont éliminées au départ avec les erreurs que cela peut causer si le choix est incorrect. L'algorithme Dension peut donc donner de bons résultats dans des situations où d'autres algorithmes ne le pourraient pas. La nouveauté de cet algorithme provient de l'incorporation d'un processus itératif à un algorithme modifié d'après celui de Fenn et Mann.

Cette étude semble montrer qu'aucun algorithme ne peut donner des résultats infaillibles dans toutes les situations. Mais, compte tenu du fait que les algorithmes conçus lors de cette étude semblent complémentaires, il peut être utile de les utiliser en parallèle et de tirer les conclusions nécessaires face aux résultats combinés obtenus. La déconvolution des spectres de masse de protéines peut donc être automatisée, mais il faut rester prudent et utiliser différentes façons indépendantes afin d'obtenir des résultats fiables.

La présente étude montre donc que le prétraitement des données est vital au bon fonctionnement d'une méthode automatisée de déconvolution. Ce qui pourrait faciliter la recherche en protéomique puisque jusqu'à présent aucun prétraitement des données n'était effectué et la déconvolution n'est pas toujours efficace.

La présente étude montre aussi qu'aucun algorithme n'est parfait et qu'aucun n'est utilisable pour toutes les situations. Il est donc préférable d'utiliser une panoplie

d'algorithmes différents et de les utiliser en concert afin d'optimiser la déconvolution et de vérifier les résultats obtenus.

Il est donc normal de s'attendre que la recherche sur l'automatisation de la déconvolution de spectres de masse de protéines en ESI s'orientera probablement vers quelque chose d'autre que le développement de nouveaux algorithmes de déconvolution. En effet, il serait préférable de chercher une façon efficace d'automatiser l'utilisation en parallèle des différents algorithmes existants afin d'obtenir des résultats plus fiables. Ceci permettra d'accélérer encore plus l'obtention de résultats fiables.

Il serait aussi utile de développer d'autres méthodes de filtration afin de réduire au maximum la quantité de pics dus au signal de fond. Cette filtration pourrait inclure d'autres méthodes qui peuvent différencier les vrais pics des faux pics dus au signal de fond dans le spectre déconvolué.

Références

-
- ¹ E. Vilkas, E. Lederer, *Tetrahedron Lett.*, 1968, pp. 3089-3092.
- ² A. Parente, M.N. Short, R. Self et K.R. Parsley, *Biomed. Mass Spectrom.*, 1982, 9, pp. 141-145.
- ³ M. Barber, R.S. Bordoli, R.D. Sedgwick et A.N. Tyler, *J. Chem. Soc. Chem. Commun.*, 1981, pp. 325-327.
- ⁴ H.R. Morris, M. Panico, M. Barber, R.S. Bordoli, R.D. Sedgwick et A.N. Tyler, *Biochem. Biophys. Res. Commun.*, 1981, 101, pp. 623-631.
- ⁵ D.F. Hunt, J.R. Yates III, J. Shabanowitz, S. Winston et C.R. Hauer, *Proc. Natl. Acad. Sci. USA*, 1986, 83, p. 6233-6237.
- ⁶ M. Karas et F. Hillenkamp, *Anal. Chem.*, 1988, 60, pp. 2299-2301.
- ⁷ J.B. Fenn, M. Mann, C.K. Meng, F. Wong et C.M. Whitehouse, *Science*, 1989, 246, pp. 64-71.
- ⁸ M. Mann, C.K. Meng et John B. Fenn, *Anal. Chem.*, 1989, 61, pp. 1702-1708.
- ⁹ B.B. Reinhold et V.N. Reinhold, *J. Am. Soc. Mass Spectrom.*, 1992, 3, pp. 207-215.
- ¹⁰ M. Labowsky, C. Whitehouse et J.B. Fenn, *Rapid Commun. Mass Spectrom.*, 1993, 7, pp. 71-84.
- ¹¹ J.J. Hagen et C.A. Monnig, *Anal. Chem.*, 1994, 66, pp. 1877-1883.
- ¹² Z. Zhang et A.G. Marshall, *J. Am. Soc. Mass Spectrom.*, 1998, 9, pp. 225-233.
- ¹³ D. Zidarov, D. Faubert, J. Visentini et M.J. Bertrand, *Proc. Of 39th Annual Conf. ASMS and Allied Topics*, Nashville, 1991, pp. 1123-1124.
- ¹⁴ M. Dole, L.L. Mack, R.L. Hines, R.C. Mobley, L.D. Ferguson et M.B. Alice, *J. Chem. Phys.*, 1968, 49, pp. 2240-2249.
- ¹⁵ J.V. Iribarne et B.A. Thomson, *J. Chem. Phys.*, 1976, 64, p. 2287-2294.
- ¹⁶ B.A. Thomson et J.V. Iribarne, *J. Chem. Phys.*, 1979, 71, p. 4451-4463.
- ¹⁷ B.A. Thomson, J.V. Iribarne et P.J. Dziedzic, *Anal. Chem.*, 1982, 54, pp. 2219-2224.

-
- ¹⁸ J.V. Iribarne, P.J. Dziedzic et B.A. Thomson, *Int. J. Mass Spectrom. Ion Phys.*, 1983, 50, pp. 331-334.
- ¹⁹ M. Yamashita et J.B. Fenn, *J. Phys. Chem.*, 1984, 88, p. 4451 et 4671.
- ²⁰ M.L. Aleksandrov, L.N. Gall, V.N. Krasnov, V.I. Nikolaev, V.A. Pavlenko et V.A. Shkurov, *Dokl. Akad. Nauk SSSR*, 1984, 277, p. 379.
- ²¹ J.B. Fenn, M. Mann, C.K. Meng, S.K. Wong et C. Whitehouse, *Mass Spectrom. Rev.*, 1990, 9, pp. 37-70.
- ²² J.B. Fenn, *J. Am. Soc. Mass Spectrom.*, 1993, 4, pp. 524-535.
- ²³ R.B. Cole, *Electrospray Ionization Mass Spectrometry: Fundamentals, Instrumentation, and Applications*, John Wiley & Sons, Inc., New York, 1997, p. 9.
- ²⁴ G. Wang et R.B. Cole, *Organic Mass Spectrom.*, 1994, 29, pp. 419-427.
- ²⁵ J.A. Loo, H.R. Udseth et R.D. Smith, *Rapid Commun. Mass Spectrom.*, 1988, 2, pp. 207-210.
- ²⁶ M.A. Kelly, M.M. Vestling et C.C. Fenselau, *Org. Mass Spectrom.*, 1992, 27, pp. 1143-1147.
- ²⁷ G. Wang et R.B. Cole, *Anal. Chem.*, 1995, 67, pp. 2892-2900.
- ²⁸ D.S. Ashton, C.R. Beddell, D.J. Cooper, B.N. Green et R.W.A. Oliver, *Org. Mass Spectrom.*, 1993, 28, pp. 721-728.