Technique objective de traitement des spectres de masse protéomiques basée sur le seuillage flou multi-échelle

An objective reduction technique of proteomic mass spectra based on multi-scale fuzzy thresholding

N. M. Nafati, J.-M. Guigonis, B. Rossi et M. Samson

IFR50, Plate-Forme Protéomique Nice-Pasteur, Faculté de Médecine, Nice, France
nafati@unice.fr

Manuscrit reçu le 1er septembre 2005

Résumé et mots clés

La protéomique offre une approche puissante et complémentaire à la génomique. Elle permet de répertorier et caractériser les protéines, de comparer leur niveau d'expression entre un état physiologique sain et malade par exemple. L'analyse protéomique se fait essentiellement par l'utilisation de la technique d'électrophorèse bidimensionnelle couplée à la technique d'analyse par Spectrométrie de Masse (SM). La première, aidée par l'imagerie protéomique, conduit à la localisation des protéines candidates à une analyse par SM. La comparaison des spectres de masses obtenus à des bases de données protéiques, conduit à l'identification des protéines d'intérêt en terme de peptides. Le problème qui se pose souvent est que les spectres sont bruités et pauvres en masses. En effet, le bruit du détecteur, le bruit électronique et chimique, la présence de peu de matériel protéique et enfin le bruit de la réduction des spectres (mauvais filtrage et/ou seuillage), tous ces bruits peuvent induire des Pics de Masses Parasites (PMP) et/ou supprimer des Pics de Masses Utiles (PMU) de faible intensité. La conséquence immédiate est que la présence des PMP et l'absence des PMU seront utilisées au dépens de la qualité d'identification de la protéine.

Dans cet article, nous proposons un algorithme original éliminant les PMP, détectant et amplifiant ceux utiles. Le principe du pré-traitement utilise une Analyse Multirésolution (AM) couplée à un seuillage basé sur la logique floue (seuillage flou multi-échelle), une amplification locale des PMU, et enfin une correction adaptative de la Ligne de Base (LB). Les fréquences associées aux PMP sont réparties sur toute la bande passante du spectre, ce qui nous conduit à une AM dite en arbre. Le principe consiste à découper la bande passante fréquentielle de chaque spectre de masses en deux sous-bandes, une Basse Fréquence (BF), l'autre Haute Fréquence (HF), ensuite chaque sous-bande est à son tour découpée en deux sous-bandes etc. Les sous-bandes HF sont seuillées selon le critère de minimisation de l'entropie floue de Shannon et amplifiées localement, la ligne de base est calculée automatiquement et soustraite du spectre reconstruit. Pour évaluer la qualité de cet algorithme, nous présentons une comparaison des résultats obtenus par notre algorithme, et ceux fournis par le spectromètre MALDI-TOF (Matrix Assisted Laser Desorption/Ionisation-Time Of Flight), qui utilise le logiciel « DataExplorer » comme logiciel de réduction.

Protéomique, Electrophorèse bidimensionnelle, Spectrométrie de Masse (MS), Bases de données protéiques, Seuillage flou multi-échelle, Amplification locale, Correction adaptative de la ligne de base, Entropie floue de Shannon, MALDI-TOF, DataExplorer.

Abstract and key words

A proteomic approach offers a powerful and complementary tool to genomics. It allows to index and characterize proteins, and, for example, to compare their levels of expression between healthy and pathological states. Proteomic analyses are mainly based on the separation of proteins by two-dimensional gel electrophoresis and their subsequent identification by comparing the data from Mass Spectrometry (SM) analyses to the theoretical ones contained in databases.

In mass spectrometry, the detector noise, the electronic and chemical noise, sometimes the small amount of peptides that has to be treated and finally the spectrum reduction noise (due to bad filtering and/or thresholding), can induce Parasitic Mass Peaks (PMP) and/or hide some Useful Mass Peaks (UMP) of low intensities. The immediate consequence is that the presence of the PMP and the absence of the UMP will be detrimental to the protein identification quality. In this article, we propose an original algorithm eliminating the PMP, detecting and amplifying those which are useful. The preprocessing principle uses a multi-scale analysis technique coupled to a fuzzy thresholding (multi-scale fuzzy thresholding), a local amplification of the UMP, and finally an adaptive Base Line Correction.

The associated frequencies with the PMP are distributed on all the spectrum pass bandwidth. This leads us to a dyadic tree structure subband decomposition. The algorithm principle consists of dividing the frequential pass bandwidth of each masses spectrum into two subbands, a Low and High Frequency (LF,HF) subband, then each subband is in turn divided into two subbands etc. The HF subbands are then thresholded according to the minimization criterion of the Shannon fuzzy entropy, and then amplified locally; the base line is calculated in an adaptive way and subtracted from reconstructed spectrum. To evaluate the quality of this algorithm, we present a comparison of the results obtained by our algorithm, and those obtained by the DataExplorer software. The latter is a reduction software provided within the MALDI-TOF spectrometer software package.

Proteomic approch, Two-dimensional electrophoresis, Mass Spectrometry, Proteinic databases, Multi-scale Fuzzy Thresholding, Adaptive Base Line Correction, Shannon Fuzzy Entropy, DataExplorer Software, MALDI-TOF.

1. Introduction

La protéomique est un domaine qui permet de mettre en relation la séquence du génome et le comportement cellulaire. Son but est l'étude des produits protéiques dynamiques exprimés à partir du génome et leurs interactions à un moment donné ou sous certaines conditions environnementales [3,4]. La spectrométrie de masse représente un maillon important de la chaîne d'analyse protéomique, elle permet de transformer les macromolécules dans leur état naturel en ions dans l'état gazeux et donc d'obtenir leur spectre de masse [5] [8] [14]. Ces spectres sont souvent bruités et présentent des pics de masse utiles noyés dans le bruit. Il est généralement admis que la qualité des spectres de masse dépend de certains facteurs liés à la sélection et la nature de la matrice, aux caractéristiques intrinsèques de la bio molécule, aux solvants utilisés etc. La figure de l'Annexe II venant de la référence [6], indique que le bruit chimique engendre des hautes fréquences, et que le bruit dû à la matrice fait augmenter le bruit de fond. Ces bruits d'origine chimique s'ajoutent au bruit instrumental (du détecteur et de l'électronique de conditionnement). Celui-ci est un bruit blanc réparti sur toutes les fréquences comme cela est mentionné dans la référence [2]. La conséquence immédiate, est que ces facteurs parasites influencent directement les intensités des différentes masses ioniques mesurées, et remettent en cause la fiabilité d'identification des protéines sur les banques de données. D'où l'importance de se doter de moyens de traitement robustes en terme d'objectivité et de fiabilité. Malgré les progrès technologiques récents sur l'enregistrement des spectres, la réduction de ces derniers pour extraire l'information utile, reste un challenge en terme de diagnostic et de discrimination.

Dans ce document, nous proposons un algorithme de réduction des spectres de masse. Cet algorithme utilise la technique multirésolution couplée à un seuillage basé sur la logique floue. L'idée est de séparer les pics de masse en groupes de Coefficients Ondelettes (CO) selon des échelles dyadiques, de seuiller les COs de détails en analysant l'entropie floue de Shannon, de les amplifier, et enfin de soustraire la ligne de base de manière adaptative une fois le spectre reconstruit.

2. Approche protéomique

L'approche protéomique globale permet de réaliser un inventaire du contenu protéique. Elle permet en particulier une mesure directe de la signature de certaines protéines et promet des avancées considérables pour le diagnostic et le traitement de maladies. Avant d'aborder le cœur du sujet, rappelons quelques définitions essentielles à la compréhension de notre travail. Rappelons qu'une protéine est une macromolécule constituée d'une ou plusieurs chaînes peptidiques, les plus courtes font une cinquantaine d'acides aminés, les plus longues pouvant atteindre plusieurs milliers. La masse moléculaire d'une protéine est de plusieurs milliers de Daltons. Le Dalton (Da) est l'unité utilisée pour décrire la masse d'une molécule, Il correspond à la masse d'un atome d'hydrogène, soit 1,66 10^{-24} g) [3,4]. Rappelons également que le protocole d'analyse protéomique mentionné dans l'Annexe I, est en général le suivant : recueil et solubilisation des protéines, séparation des protéines par la Technique d'Electrophorèse Bidimensionnelle (TEB), analyse d'image des gels 2D obtenus par la TEB et localisation des Protéines d'Intérêt (PrI), découpage et digestion des PrI, analyse par spectrométrie de masse de l'extrait en vue d'obtenir la carte peptidique, et enfin interrogation des banques de donnée pour identifier ces PrI. Chaque étape de ce protocole joue un rôle très important, celles se trouvant en amont de l'analyse par spectrométrie de masse, ont un rôle de localisation des tâches protéiques candidates pour une digestion. Celle-ci fait intervenir une enzyme spécifique, le plus souvent la trypsine dont les sites de coupure sont localisés au niveau de la lysine et de l'arginine. Le but de cette digestion est d'obtenir des peptides dont la taille soit compatible avec la gamme de masse des spectromètres utilisés [14,15] [18] [20,21].

3. Principe d'identification protéomique par spectrométrie de masse

Le principe de la spectrométrie de masse MALDI-TOF (Matrix Assisted Laser Desorption/ Ionisation-Time Of Flight), consiste à irradier un échantillon co-cristallisé dans une matrice de type acide 2,5-DiHydroxyBenzoïque (DHB), absorbant à la longueur d'onde UV (UltraViolet à 337nm) du rayonnement laser. L'irradiation provoque l'éjection des Molécules d'échantillon (M) et de matrice en phase gazeuse. L'échantillon est donc ionisé majoritairement par transfert de protons (H^+) pour former des ions mono ou multichargés de type [M+nH^+]. Les ions monochargés [M+H^+] sont majoritaires dans les spectres MALDI. Ces particules ioniques sont ensuite accélérées, puis évoluent selon leur masse dans une zone de vide poussé (le tube de vol), avant d'être analysées par le détecteur. Les masses des peptides mesurées constituent ainsi l'empreinte peptidique de la protéine analysée [5] [8] [25] [27]. La Figure 1 montre un spectre obtenu par MALDI-TOF (Voyager DE Pro, PerSeptive Biosystems) qui correspond à une protéine identifiée chez le rat comme étant l'Acyl-CoA Déshydrogenase. Sur l'axe des abscisses (X), les valeurs m/z lues correspondent aux masses véritables des peptides en Daltons plus un proton [M^+H^+], sur l'axe des ordonnées (Y), les intensités normalisées (en %) sont calculées en fonction de la représentation statistique des ions sur le détecteur.

La sélection des pics de masse sur le spectre en vue d'une présentation à des moteurs de recherche pour identifier la protéine, se fait par le programme DataExplorer du MALDI-TOF. En pratique, l'utilisateur automatise la réduction des spectres, en établissant un programme de réduction intégrant des outils de prétraitement, à savoir le filtrage du bruit, la correction de la Ligne de Base, et enfin la détection des pics de masses utiles. Ces outils sont paramétrés de manière subjective, leur choix reste à l'appréciation de l'utilisateur qui est souvent un biologiste, un médecin ou un chimiste. Le programme DataExplorer du logiciel ProteinProspector propose un filtrage bass-bas gaussien non optimal, une correction non adaptative de la LB, et enfin un seuillage statique de détection des pic de masses. Il est donc nécessaire et important d'avoir des outils de réduction fiables en terme de discrimination. Le spectre brut de la Figure 1 une fois réduit, conduit au spectre résultat de la Figure 2. Sur ce dernier, on constate que les seuls pics conservés sont ceux qui sont vus comme significatifs. On verra plus loin dans le paragraphe résultat, que d'autres pics utiles peuvent être détectés.

L'identification de la protéine se fait en comparant l'empreinte peptidique de la Figure 2 à celle issue de la digestion théorique (in silico) des protéines répertoriées dans les banques protéiques. Les masses des peptides sont présentées au moteur de recherche MS-FIT du logiciel Protein Prospector (http://prospector.ucsf.edu). Les paramètres de sortie sont le nom de la protéine trouvée et également d'autres paramètres d'appréciation et de précision. Parmi ces paramètres, on trouve le « Mowse Score Parameter » qui indique le degré de confiance pour que l'événement de corrélation entre masses expérimentales et théoriques soit un événement non aléatoire. Le « Cov Parameter en % » indique le taux de recouvrement en terme de peptides de la protéine candidate de la banque, et enfin le « Mean Err Parameter », qui donne la précision moyenne sur la masse qui s'exprime en ppm (point par million) [17]. Dans le cas du spectre précédent, les résultats obtenus par comparaison avec la banque SwissProt sont donnés dans le Tableau 1 ci-après :

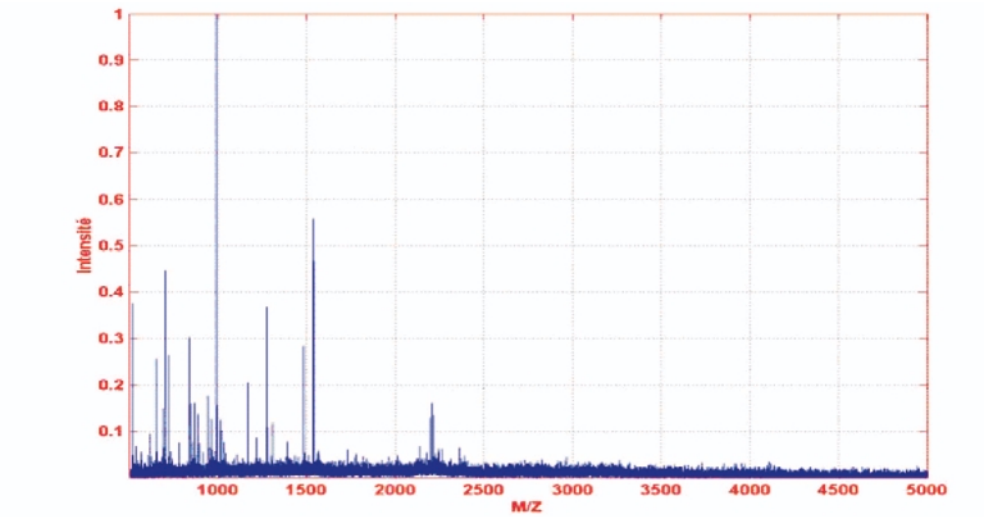


Figure 1. Spectre brut de masse de la protéine Acyl-CoA du RAT. En abscisse, les pics de masses peptidiques. En ordonné, les intensités normalisées données par le détecteur.

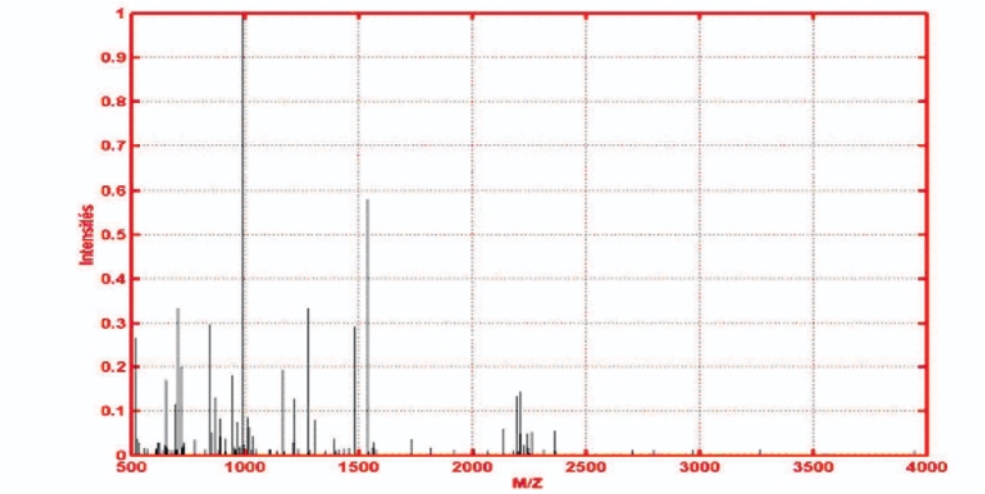


Figure 2. Pics de masse détectés par le programme DataExplorer.

Tableau 1. Résultats d'identification par comparaison des pics du spectre traité de la Figure 2 avec la banque SwissProt. La protéine attendue sort au rang 1 et 11, avec des degrés de confiance respectifs de 3.586 10^4 chez l'espèce « RAT » et 2215 chez l'espèce « SOURIS ».

Rang	Mowse Score	Cov %	Mean Err (ppm)	Species	Protein Name
1	**3.586 10^4**	**36.0**	**29.7**	**RAT**	**Acyl-CoA Dehydrogenase**
2	1.598 10^4	12.0	–15.2	CHICK	Cytochrome
3	9997	13	0.200	MOUSE	Bcl-2 related
...	...	...	...	...	...
10	2435	13	–9.57	MOUSE	Brain-specific serine
11	**2215**	**24**	**24.0**	**MOUSE**	**Acyl-CoA Dehydrogenase**

Différents essais ont été réalisés pour optimiser l'exploitation d'un ou de plusieurs spectres. En effet, « DataExplorer » permet l'ajustement manuel des paramètres des processus de réduction. « DataEplorer User Guides » [1] donne des détails sur ces nombreux paramètres et donne également leur paramétrage par défaut. Les meilleurs résultats (présentés dans le Tableau 1.) ont été obtenus en appliquant le paramétrage donné dans l'Annexe IV. Celui-ci a fait l'objet de tests expérimentaux visant à améliorer au maximum la qualité d'identification de la protéine. On a constaté qu'un paramétrage près de celui pris par défaut n'influence que peu le résultat.
Conclusion : la protéine utilisée expérimentalement a été identifiée en première position chez le rat avec un taux de recouvrement de 36%. Cette valeur, un peu faible (< 50%), peut s'expliquer par l'absence de pics utiles et/ou la présence de pics parasites.

4. Algorithme proposé

Le principe de notre algorithme, comme l'indique la Figure 3 ci-dessous, se trouve localisé au niveau de l'utilisation de la technique multirésolution, couplée à un seuillage flou optimal basé sur la minimisation de l'entropie floue de Shannon. Avant de rentrer dans les détails de ce processus de prétraitement, nous rappelons le principe de l'analyse multirésolution et le choix du banc de filtres ondelettes, nous exposerons le principe du seuillage flou des hautes fréquences et le choix de la fonction d'appartenance. Nous indiquerons également, la méthode d'amplification des coefficients ondelettes seuillés et celle de la soustraction de la ligne de base après reconstruction du spectre.

4.1 Technique multirésolution adoptée

Le concept d'Analyse Multirésolution (AM) qui est sous-jacent à la Transformée en Ondelettes (TO), offre une étude pyramidale multi-échelle. L'analyse se fait à partir de dilatées et de translatées de l'ondelette mère. La TO décrit les détails d'un signal pour chaque niveau de résolution. Ces détails correspondent à la différence d'informations entre deux niveaux de résolution successifs. Dans notre étude, on s'est limité à l'utilisation d'une analyse dyadique, c'est-à-dire à des facteurs d'échelle égaux à 2. L'opération de base de l'AM est la décomposition du signal en deux parties, une approximation et les détails du signal. L'approximation est obtenue en projetant le signal sur les translatées d'une fonction basse fréquence appelée, fonction échelle. Cette projection isole les variations lentes par un filtrage passe-bas. Les détails (variations rapides) du signal sont obtenus par projection sur les translatées d'une fonction haute fréquence, appelée fonction ondelette [10] [15,16] [23].
L'algorithme de décomposition démarre avec le spectre de masse du départ Y, d'où l'on calcule les coefficients Y_{iL} d'approximation et Y_{iH} des détails. Les signaux Y_{iL} et Y_{iH} sont obtenus en convoluant Y avec un banc de filtres (LowPass Filter, HighPass Filter). La décomposition peut être classique, itérée sur les basses fréquences (Figure 1 de l'Annexe III) ou en arbre (Figure 2 de l'Annexe III). Cette dernière est une variante qui effectue la décomposition non seulement sur les coefficients d'approximation, mais également sur les coefficients de détails. Auparavant, on avait précisé que le bruit que l'on veut traiter se trouve dans toute la bande passante du spectre, c'est pourquoi, la décomposition dite en arbre a été utilisée dans notre approche.

4.2 Choix de l'ondelette pour l'application

Rappelons que notre objectif principal dans cette étude est le débruitage des spectres de masses sans perte ni modification de l'information. L'analyse sous-bandes et la reconstruction ne doivent en aucun cas altérer la position des pics de masse utiles. La conséquence immédiate de cette contrainte conduit au choix du banc de filtres RIF (Réponse Impulsionnelle Finie) à phase linéaire. En effet, il existe deux types d'ondelettes, les ondelettes dites orthogonales et bi-orthogonales. Les ondelettes orthogonales présentent un grand intérêt, car elles bâtissent une base orthonormée de l'espace des signaux, ce qui facilite l'inversion de la transformation. La famille des ondelettes orthogonales à support compact est la plus intéressante, les filtres associés étant à réponse impulsionnelle finie, ce qui facilite leur implémentation. L'orthogonalité permet d'obtenir une bonne qualité du signal lors de la reconstruction. Cependant, les ondelettes orthogonales ne sont pas symétriques et induisent des distorsions lors du changement de base. Pour pallier ce problème,

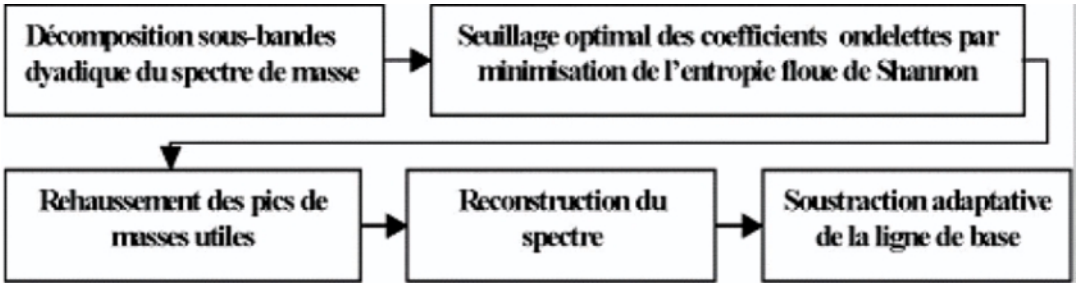


Figure 3. Architecture globale de l'algorithme proposé.

les ondelettes bi-orthogonales ont été proposées. Sans être orthogonales, elles possèdent les caractéristiques des ondelettes orthogonales. En plus, elles sont symétriques. Les bancs de filtres associés sont des filtres RIF à phase linéaire, donc symétriques. Les résultats donnés dans cet article sont obtenus en utilisant un système de Dec/Rec (Décomposition/Reconstruction) bi-orthogonal de type B-spline cubique [10] [15,16] [19]. L'intérêt pour les ondelettes B-spline est dû au compromis satisfaisant entre les localisations temporelle et fréquentielle.

4.3 Débruitage par seuillage optimal flou

Les systèmes physiques travaillent parfois avec des données incertaines et incomplètes, tels que les systèmes d'aide à la décision, les systèmes d'exploitation des bases de connaissances etc. Ces types d'informations sont représentées et traitées grâce à la théorie des sous-ensembles flous [11] [22] [24]. Soit Ω un ensemble de N éléments tel que : $\Omega = \{Y_1, Y_2, ..., Y_N\}$. Supposons que l'on doive chercher des éléments répondants à une propriété quelconque α. L'ensemble Ω se divise alors en deux sous-ensembles A et B. A contient des éléments possédants α, tandis que les autres appartiennent au sous-ensemble B, le complément de A dans Ω. En vue de la logique classique, un élément n'appartient qu'à un sous-ensemble A ou B. Ceci veut dire qu'un élément n'a que deux possibilités : soit il a cette propriété, soit il ne l'a pas. Cependant, il est possible qu'il existe dans Ω des éléments qui ne possèdent qu'avec un certain degré. Dans ce cas, il vaut mieux prendre le sous-ensemble flou pour représenter ces informations.

4.3.1. Sous-ensemble flou

Un sous-ensemble flou A de l'espace observé Ω, est caractérisé par une fonction d'appartenance μ_A, qui associe un élément Y de Ω à un nombre réel $\mu_A(Y)$ dans l'intervalle [0,1], et qui quantifie le degré d'appartenance de l'élément Y au sous-ensemble A. Généralement, un sous-ensemble flou est défini comme une collection de paires en ordre $(Y, \mu_A(Y))$. La notation normalement adoptée pour représenter le sous-ensemble flou A de Ω est :

$$A = \sum_{Y \in \Omega} \frac{\mu_A(Y)}{Y} \text{ si } A \text{ est fini}$$

4.3.2. Fonction d'appartenance

Chaque élément dans un sous-ensemble A possède un degré qui indique dans quelle mesure l'appartenance de l'élément dans A. Ce degré est déterminé par la Fonction d'Appartenance (FA) μ_A, telle que :

$$\mu_A : Y \in \Omega \longrightarrow \mu_A(Y) \in [0,1]$$

Il existe différentes FA. La plus utilisée et la plus connue est la fonction linéaire. La sélection de la fonction d'appartenance dépend de chaque application. L'idée de base de notre approche est d'exploiter d'une façon optimale et hiérarchique tout l'espace fréquentiel. Ceci permet d'éliminer les pics parasites de masse tout en préservant les pics utiles. La condition possible pour que la FA soit appropriée est que : plus petite est la distance entre le niveau du signal et la moyenne du sous-ensemble, plus grande est la valeur de la fonction d'appartenance. Dans notre algorithme, la fonction choisie respectant cette condition est :

$$\mu(Y) = \begin{cases} \dfrac{1}{1 + \left| \dfrac{Y - \mu_1(t)}{c} \right|^2} & \text{si } Y \leq t \\ \dfrac{1}{1 + \left| \dfrac{Y - \mu_2(t)}{c} \right|^2} & \text{si } Y > t \end{cases}$$

$$\text{Avec } \mu_1(t) = \frac{\sum_{i=Y_{\min}}^{t} Y_i \cdot h(Y_i)}{\sum_{i=Y_{\min}}^{t} h(Y_i)} \text{ et } \mu_2(t) = \frac{\sum_{i=t+1}^{Y_{\max}} Y_i \cdot h(Y_i)}{\sum_{i=t+1}^{Y_{\max}} h(Y_i)}$$

Où t signifie le coefficient ondelette seuil choisi, h l'histogramme des coefficients ondelettes de détails, C est une constante représentant la différence entre le CO maximum et CO minimum (Y_{max}-Y_{min}) à une échelle donnée. μ_1, μ_2 sont respectivement les valeurs moyennes des sous-ensembles flous A et B.

Le calcul de la FA, pour un coefficient ondelette donné, se déroule de la manière suivante : pour une sous-bande (Ω_{SB}) donnée de taille N ($N = N_0/2^j$, où $N_0 =$ longueur initiale du spectre et j étant le niveau de la pyramide de décomposition), et un seuil t donné :

I : On divise les coefficients de Ω_{SB} en deux sous-ensembles flous $\{A, B\}$, avec $A = \{Y \leq t | Y \in \Omega_{SB}\}$, $B = \{Y > t | Y \in \Omega_{SB}\}$, $N_1 = \dim(A)$, $N_2 = \dim(B)$ et $N = N_1 + N_2$. On peut supposer que A représente le fond, B le signal.

II : On calcule les moyennes μ_1 et μ_2 respectivement de A et B. Ces moyennes peuvent être écrites sous forme de probabilités telles que :

$$\mu_1(t) = \frac{\sum_{i=i_{\min}}^{t} i \cdot p_i}{\sum_{i=i_{\min}}^{t} p_i}, \quad \mu_2(t) = \frac{\sum_{i=t+1}^{i_{\max}} i \cdot p_i}{\sum_{i=t+1}^{i_{\max}} p_i} \text{ avec } p_i = \frac{h(i)}{N}$$

Où P_i est la probabilité d'occurrence du coefficient ondelette Y_i noté i.

III : On calcule la fonction $\mu_A(Y)$ comme suit :

Si $Y \in A$, alors :	Sinon, si $Y \in B$, alors :
$\mu(Y) = \dfrac{1}{1 + \left\| \dfrac{Y - \mu_1(t)}{c} \right\|^2}$	$\mu(Y) = \dfrac{1}{1 + \left\| \dfrac{Y - \mu_2(t)}{c} \right\|^2}$

4.3.3. Seuillage par minimisation de l'entropie floue de Shannon

Selon la théorie de l'information, l'entropie mesure la quantité d'information d'un système [12]. Dans notre approche, chaque sous-bande Ω_{SB}, est composée de coefficients considérés comme des événements indépendants, si p_i est la probabilité d'occurrence de chaque élément Y_i, alors l'entropie associée à Ω_{SB} est définie par :

$$H(p_1, p_2, \ldots, p_N) = -\sum_{i=1}^{N} p_i * \log_2(p_i)$$
$$\text{avec } \sum_{i=1}^{N} \log_2(p_i) = 1$$

L'entropie est maximale (égal à 0), si les données sont réparties dans une seule classe. Elle est minimale (dépendant du nombre de classes), si les données sont réparties uniformément dans toutes les classes. Plusieurs méthodes de seuillage et de segmentation l'utilisent dans le but de maximiser la quantité d'informations. L'incertitude de nature floue et non pas de nature aléatoire dans un sous-ensemble flou A, peut être évaluée par l'entropie floue de Shannon H_f. Contrairement à l'entropie, celle-ci est non probabiliste. Son expression est :

$$H_f(A) = \frac{-1}{N * \ln(2)} * \sum_{Y \in A} \left[\mu_A(Y) * \ln(\mu_A(Y)) + (1 - \mu_A(Y)) * \ln(1 - \mu_A(Y))\right]$$

L'entropie floue de Shannon [11] [22] [24] mesure le niveau de flou, elle est maximale quand $\mu_A = 0.5$, minimale quand $\mu_A = 0$ ou 1. Dans les méthodes de seuillage, le but est de trouver le meilleur seuil permettant d'extraire l'information utile. En raisonnant avec la logique floue, l'objectif est de trouver la valeur seuil qui minimise l'incertitude associée à un spectre. Cette incertitude peut être déterminée par l'indice de flou (ou bien le Degré d'Ambiguïté (DA)) donné par la formule suivante :

$$E(t) = \frac{-1}{N * \ln(2)} * \sum_{Y_i = Y_{\min}}^{Y_i = Y_{\max}} \left[\mu_t(Y_i) * \ln(\mu_t(Y_i)) + (1 - \mu_t(Y_i)) * \ln(1 - \mu_t(Y_i))\right] * h(Y_i)$$

Le seuil optimal t_{op} recherché distinguant les pics de masses utiles des pics parasites dans un spectre donné, est celui qui minimise le critère du degré d'ambiguïté indiqué ci-dessus. En effet, $E(t)$ est calculé pour chaque sous-bande HF et pour un seuil t donné, avec $Y_{\min} \leq t \leq Y_{\max}$. Il faut dire que les sous-bandes HFs sont divisées en blocs de 40 points. Cette valeur a été choisie de manière expérimentale, le critère de choix est le compromis entre la qualité des résultats et le temps de calcul. La sommation au sein $E(t)$ est assurée sur tous les COs à une échelle donnée et à l'intérieur d'un bloc. Par conséquent, les seuils qui minimisent l'entropie floue de Shannon, sont ceux obtenus par bloc, chaque seuil correspond au minimum local de l'entropie. La Figure 4 ci-dessous donne un exemple de spectre HF, niveau 2, et les seuils flous correspondants.

Les coefficients ondelettes, pris en considération et amplifiés, sont ceux qui sont supérieurs ou égaux à t_{op}. L'amplification se déroule en multipliant l'intensité de chaque pic détecté par le facteur G suivant :

$$G = \frac{\sigma_{\text{total}}}{\sigma_{\text{local}}}$$

Ce facteur représente le rapport de la déviation standard totale sur la déviation standard locale des COs de détails. Après avoir éliminé les COs parasites et amplifié les COs utiles intervient la reconstruction et enfin la correction de la ligne de base dont le principe est donné ci-dessous au paragraphe 4.4.

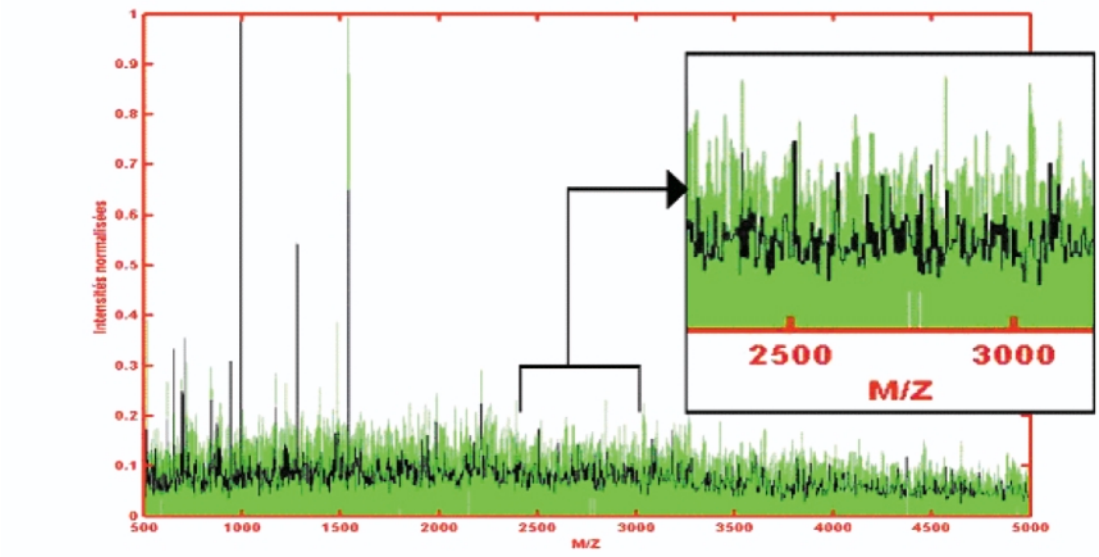


Figure 4. La courbe verte représente le module du spectre HF niveau 2 ; les seuils toppar bloc sont indiqués par la courbe noire. L'axe des abscisses indique la masse. Celui des ordonnées donne les intensités normalisées.

4.4. Méthode adaptative de la correction de la ligne de base

Le principe de la méthode réside dans l'établissement de deux étapes : la reconnaissance de la ligne de base et sa modélisation [9]. La reconnaissance se déroule de la manière suivante : pour savoir si la i-ème masse $X(i)$ appartient à la ligne de base, on la place au milieu d'une fenêtre de longueur $2*L+1$ points. Parmi ces points, l'intensité minimale $Y_i^{\min}$ et maximale $Y_i^{\max}$ sont calculées. Si leur différence n'excède pas l'écart type standard du bruit σ_{bruit} multiplié par un facteur défini F ($y_i^{\max} - y_i^{\min} < F * \sigma_{\text{bruit}}$), alors cette i-ème masse est considérée comme appartenant à la ligne de base. σ_{bruit} est estimé en divisant le spectre en 32 régions et en calculant l'écart type standard minimum par rapport à ceux intra-région. La modélisation, quant à elle, consiste à interpoler et générer la ligne de base par une méthode appropriée. En effet, une fois l'identification des masses $X_b(i)$, appartenant à la LB, est faite, on génère la LB par un interpolateur adéquat. L'interpolateur linéaire, utilisé ici, est la convolution du signal $Y_{\mathbf{Xb}}$ associé à $X_{\mathbf{b}}(i)$ sur une longueur de $2*L+1$ selon la formule :
Avec :

$$Y_b(i) = \frac{1}{2*L+1} * \sum_{k=-L}^{L} Y_{X_b}(i+\kappa) \text{ avec } i \in [1, N_0]$$

- $Y_b(i)$: l'intensité de la LB au point x_i (i-ème masse du spectre).
- $Y_{X_b}(i+k)$: l'intensité de la LB au point $x_b(i+k)$.

La valeur de F, trouvée expérimentalement, est de 100, celle de L est fixée à 20. Une variation modeste de L n'influence que peu le résultat. Le nombre de régions fixé à 32 permet un compromis entre le temps de calcul et la précision de la correction. La Figure 5 ci-dessous, montre le résultat intéressant de cette approche, simple et efficace.

4.5. Algorithme global de réduction proposé

L'algorithme proposé se déroule comme suit :

Étape 1

-Décomposition en sous-bandes de 2 niveaux par un banc de filtres associé aux ondelettes bi-orthogonales de type B-spline cubique.

-Les filtres RIF BF et HF ont respectivement la taille 6 et 8.

-La décomposition dyadique en arbre, donne une pyramide de 2 niveaux, soit 4 Sous-Bandes (SB), une basse fréquence, les autres hautes fréquences. Soit Ω_{HF} l'ensemble des sous-bandes HF, tel que : $\Omega_{HF} = \{\Omega_{HF,1}, \Omega_{HF,2}, \Omega_{HF,3}\}$

Étape 2

- Pour Chaque sous-bande $\Omega_{HF,i}$ faire :
- Diviser $\Omega_{HF,i}$ en blocs de 40 points
- Pour chaque bloc faire :
- Calculer l'histogramme h.
- Diviser le bloc en deux sous-ensembles flous A et B.
- Calculer les moyennes μ_1 et μ_2.
- Calculer les indices flous E(t).
- Calculer t_{op} qui minimise E.
- Seuiller et amplifier les Y_i.
- Fin
- Fin

Étape 3 : Reconstruction du spectre.

Étape 4 : Soustraction de la ligne de base.

Étape 5 : Détection des pics de masse par le programme DataExplorer du logiciel ProteinProspector.

Étape 6 : Identification de la protéine par corrélation des pics obtenus avec la base de données SwissProt.

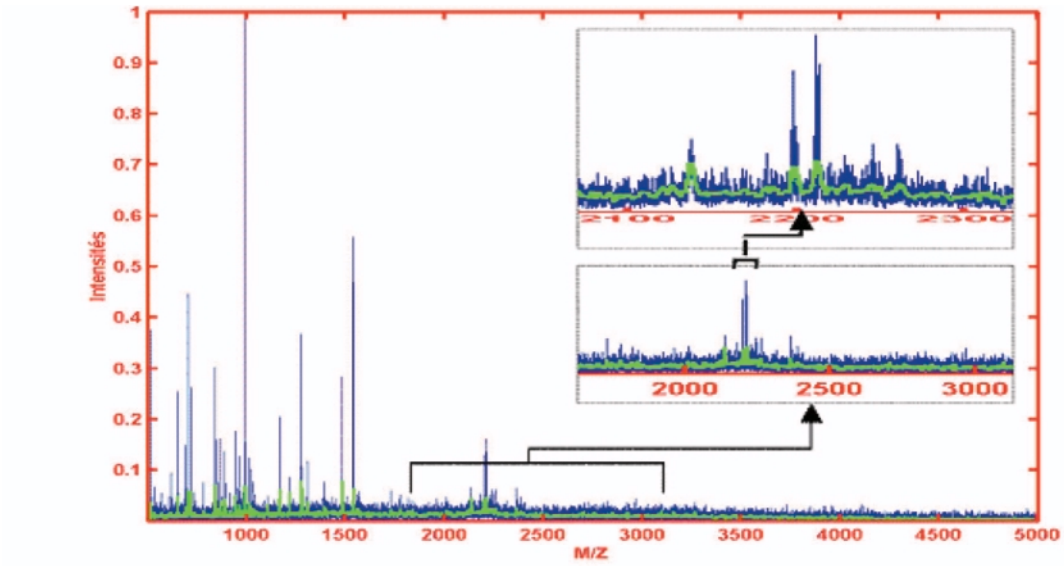


Figure 5. Spectre avec sa ligne de base verte calculée adaptativement.

5. Résultats

Ce paragraphe exposera quelques résultats obtenus avec l'algorithme proposé. Tout d'abord, les résultats obtenus avec le noyau de l'algorithme seul, c'est à dire sans rehaussement, ni correction de la ligne de base, ensuite avec le noyau plus l'amplification et la correction. Ce choix est fait dans le but d'apprécier l'apport positif ou négatif du Seuillage Flou Multi-échelle et la correction adaptative de la LB.

5.1 Application du noyau de l'algorithme

Pour mettre en évidence l'intérêt de la méthode adaptative de la correction de la LB, le spectre initial (Figure 1) est traité selon l'algorithme proposé, et sa LB a été corrigée selon la méthode du logiciel DataExplorer. Le résultat obtenu est à voir sur la Figure 1.
La corrélation de la liste des masses obtenues avec la DataBase SwissProt donne les résultats ci-après :

Tableau 2. Protéines identifiées dans la base SwissProt.

Rang	Mowse Score	Cov %	Mean Err (ppm)	Species	Protein Names
1	**$1.630\ 10^9$**	**51.0**	**22.4**	**RAT**	**Acyl-CoA Dehydrogenase**
2	$7.950\ 10^7$	20	1.74	CAEEL	Hypothetical protein zk632.5
3	$5.840\ 10^7$	26	–10.0	HUMAN	Tyrosine-protein kinase ZAP-70
4	**$2.346\ 10^7$**	**33**	**19.5**	**MOUSE**	**Acyl-CoA Dehydrogenase**

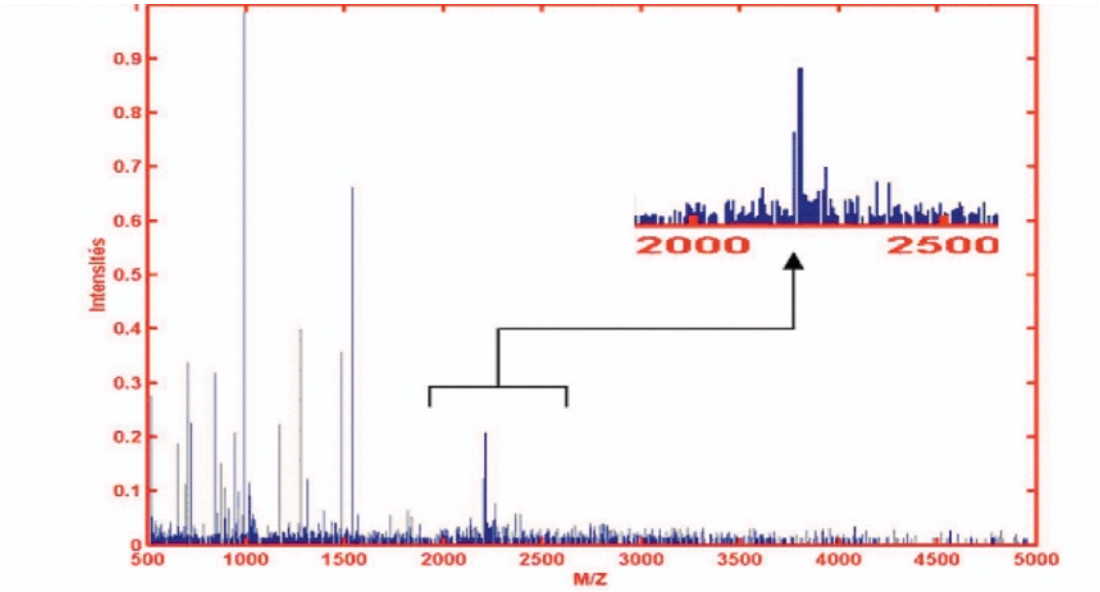


Figure 6. Spectre obtenu avec notre algorithme. La ligne de base est corrigée selon l'approche de DataExplorer. Le seuillage flou a permis de ressortir des pics de masse utiles.

Cette-fois ci, on remarque que la protéine est identifiée avec un score de $1.63\ 10^9$ et un taux de recouvrement de 51% chez le rat. La même protéine sort au rang 4 avec $2.34\ 10^7$ comme score et un taux de 33% chez la souris. Les pics de masses récupérés ont permis d'augmenter le taux de recouvrement. On peut dire que la quantité d'information utile en terme de masse a été améliorée.

5.2 Résultat de l'algorithme proposé

Cette-fois ci, le spectre initial (Figure 1) est traité selon l'algorithme proposé, c'est-à-dire, le SFM, l'amplification et la correction adaptative de la LB. Le résultat obtenu est à voir sur la Figure 7.
La corrélation avec la base de données SwissProt donne les résultats du Tableau 3.
Les résultats obtenus sont encore plus intéressants : la protéine est identifiée avec un score de $2.871\ 10^{10}$ et un taux de recou-

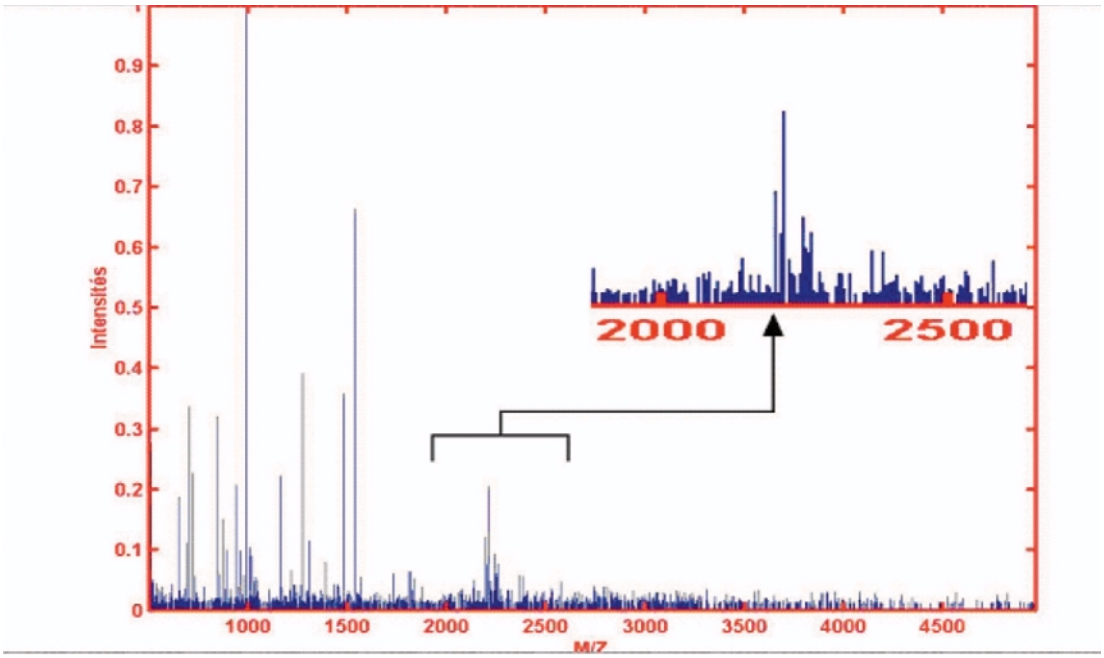


Figure 7. Spectre initial réduit par l'algorithme proposé (SFM+Amplification+Correction de la LB).

Table 3. Protéines identifiées dans la base SwissProt avec l'algorithme complet.

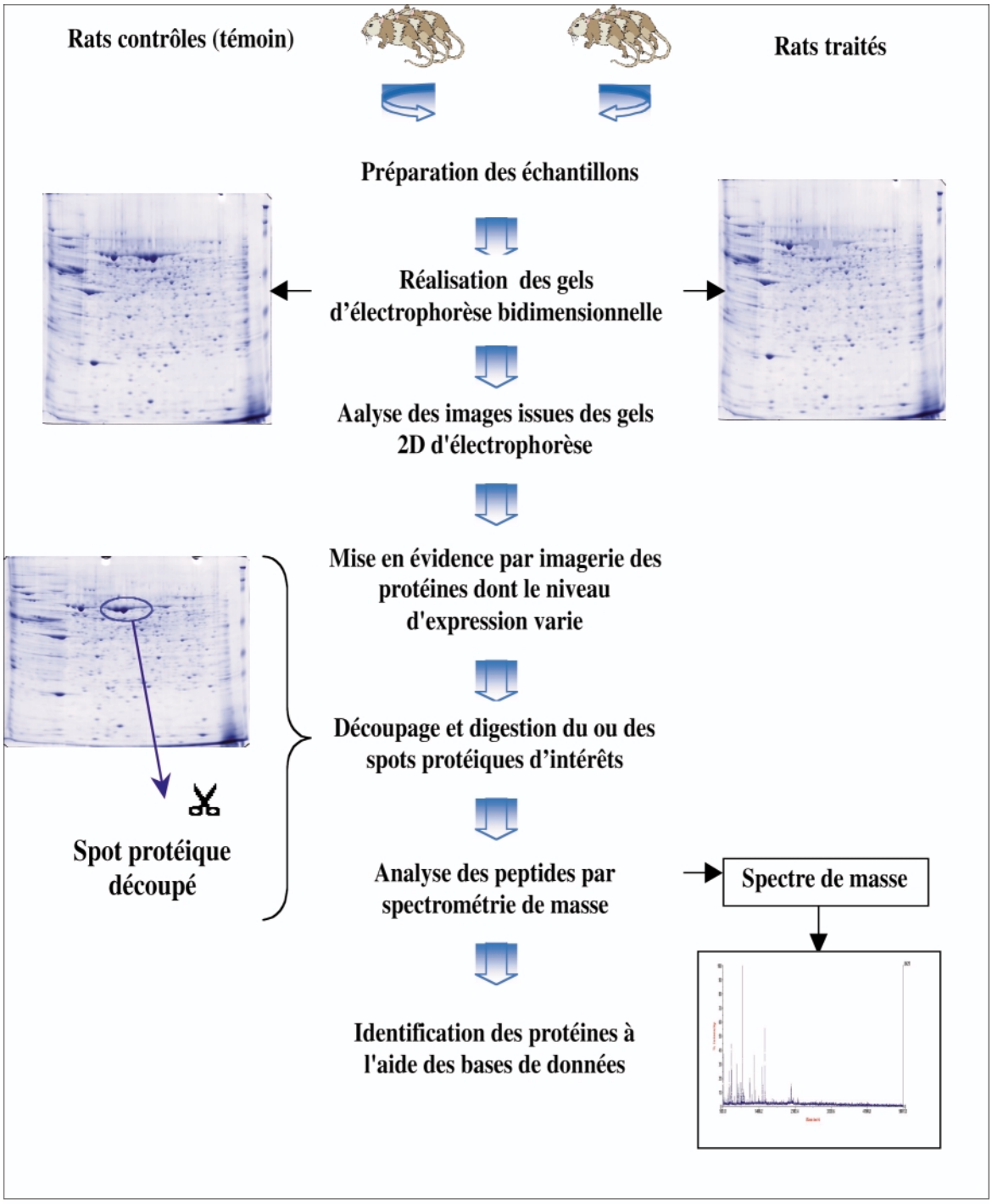
Rang	Mowse Score	Cov %	Mean Err (ppm)	Species	Protein Names
1	**$2.871\ 10^{10}$**	**56**	**21.8**	**RAT**	**Acyl-CoA Dehydrogenase**
2	**$2.196\ 10^{8}$**	**38**	**18.6**	**MOUSE**	**Acyl-CoA Dehydrogenase**
.....	....	.....	.....	.....	.....
6	**$1.818\ 10^{7}$**	**40**	**20.2**	**HUMAN**	**Acyl-CoA Dehydrogenase**

vrement de 56% chez le rat. La même protéine sort au rang 2 chez la souris, avec $2.196\ 10^{8}$ comme score et un taux de 38%. Elle sort également chez l'homme en position 6, avec $1.818\ 10^{7}$ et avec un taux de recouvrement de 40%. Les Acyl-CoA déshydrogenases de souris et d'homme possèdent respectivement 96 et 86% d'identité de séquence avec celle du rat, ce qui explique qu'en augmentant les taux de recouvrement, elles apparaissent dans le tableau avec un rang de plus en plus proche de celui du rat. On remarque également une nette amélioration de la quantité d'information en terme de masses.

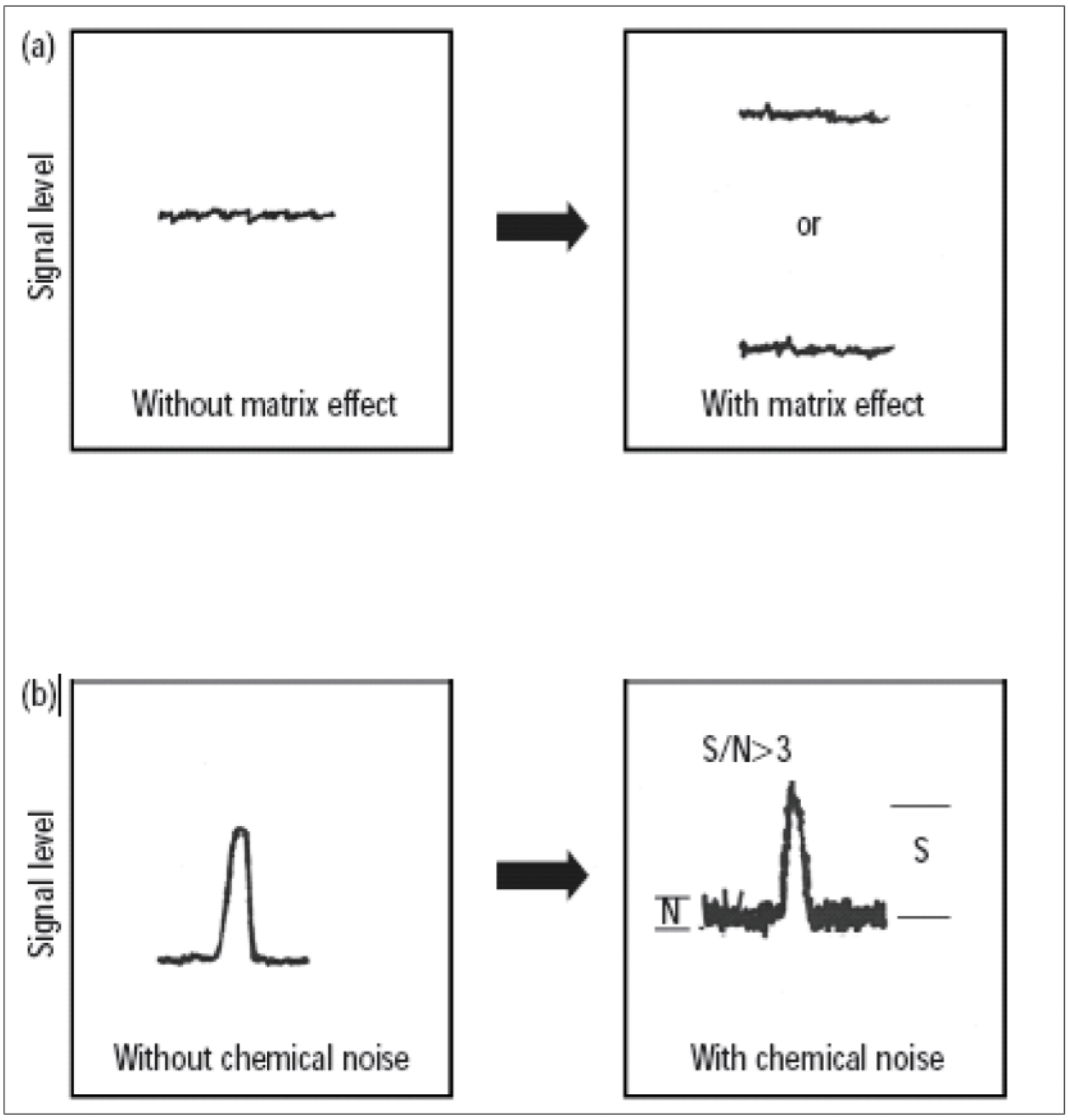
6. Conclusion

Dans le contexte de la protéomique, la spectrométrie de masse couplée à la bioinformatique joue un rôle fondamental vis-à-vis de l'identification des protéines. L'importance de la résolution et du rapport signal à bruit des spectres acquis, influencent les résultats d'identification de manière primordiale. Les logiciels de traitement des spectres, souvent fournis avec les instruments spectrométriques sont non objectifs. Dans l'algorithme proposé, le traitement multi-échelle des spectres de masses, se fait de manière objective et optimale, puisque le seuillage des coefficients ondelettes hautes fréquences se fait automatiquement, le seuil optimal étant calculé en minimisant le critère d'entropie de Shannon. D'autre part, la correction de la ligne de base s'applique de manière adaptative. Les résultats obtenus avec cet algorithme sont très intéressants, que ce soit au niveau du score de classification ou au niveau du taux de recouvrement. L'approche proposée a amélioré la quantité d'information du spectre en conduisant à l'identification de la protéine même chez d'autres espèces.

Annexe I


Rats contrôles (témoin)
Rats traités
Préparation des échantillons
Réalisation des gels d'électrophorèse bidimensionnelle
Aalyse des images issues des gels 2D d'électrophorèse
Mise en évidence par imagerie des protéines dont le niveau d'expression varie
Découpage et digestion du ou des spots protéiques d'intérêts
Spot protéique découpé
Analyse des peptides par spectrométrie de masse
Spectre de masse
Identification des protéines à l'aide des bases de données

Annexe II


(a)
Signal level
Without matrix effect
or
With matrix effect
(b)
Signal level
Without chemical noise
S/N>3
S
N
With chemical noise

Annexe III

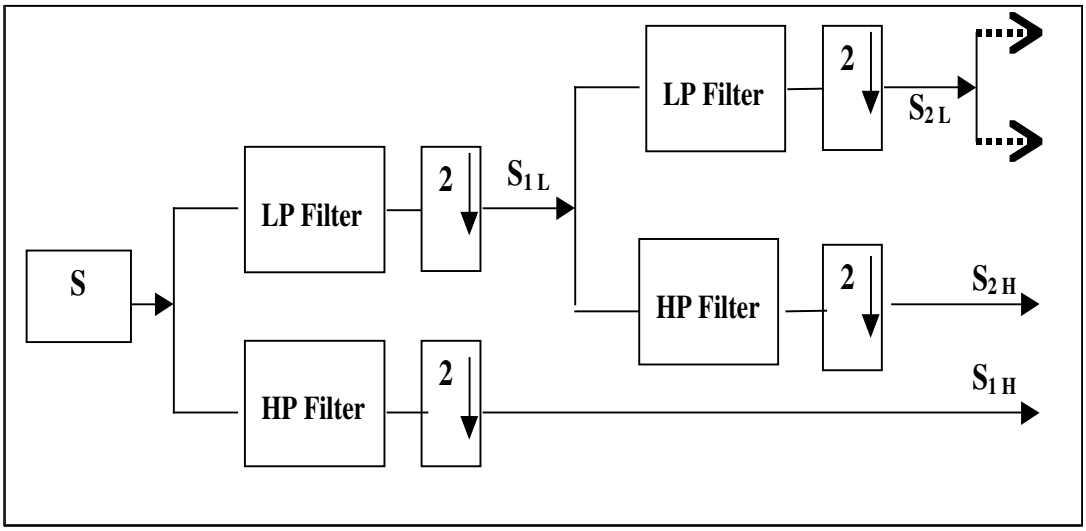


Figure 1. Décomposition dyadique classique. Un banc de filtres itéré en octaves. La partie analyse comporte les filtres passe-bas et passe-haut. L'itération a donc lieu sur les branches passe-bas, où les mêmes filtres sont utilisés à chaque étape (on a représenté ici deux itérations).

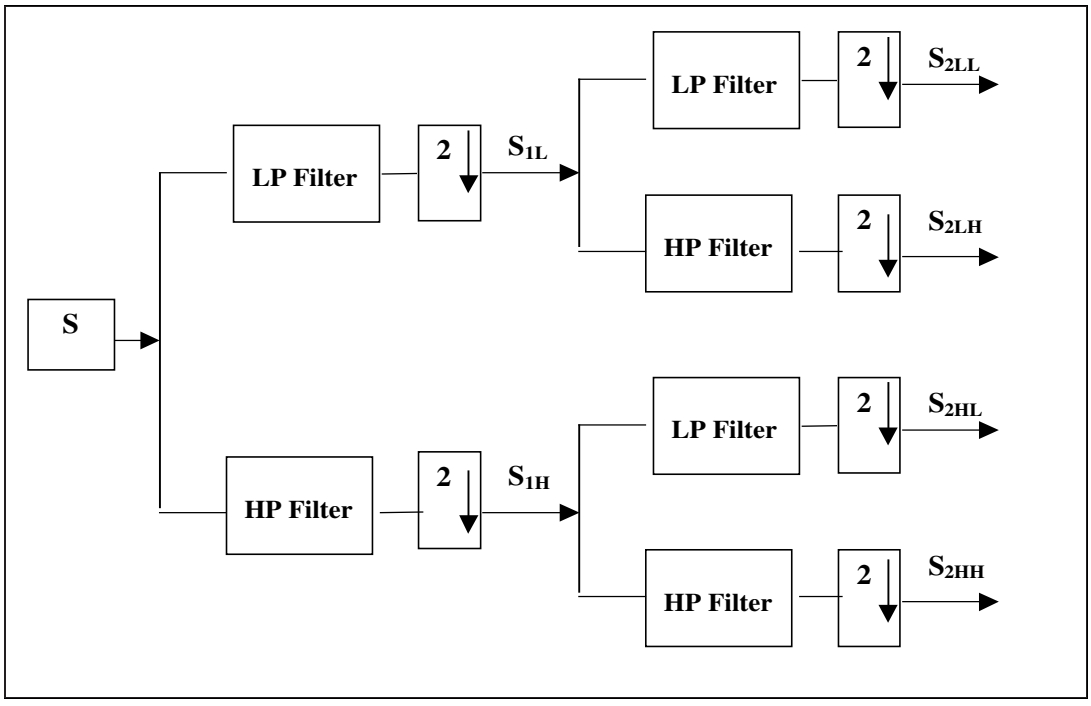


Figure 2. Décomposition en arbre dyadique. Un banc de filtres itéré en octaves. L'itération a donc lieu sur les branches passe-bas et haut, où les mêmes filtres sont utilisés à chaque étape (on a représenté ici deux itérations).

Annexe IV

Spectrum data Basic Settings

- **Base Peak Intensity = 0 %**
- **Max Peak Area = 2 %**
- **Use Resolution Dependent Settings : On**
- **Mass Resolution = 2000**

Peak Processing

- **Integration Baseline Setting : V to B**
- **Centroid = 56%**
- **Max. Charge State = 1**
- **Max. Isotopes = 5**
- **Min. Intensity = 10%**
- **Max. Intensity = 100%**

Advanced parameters

- **Detection Range Resolution Dependent**
- **Filter Width Resolution Dependent**
- **Increment = 1**
- **Noise Treshold = 0**
- **Base Peak Intensity = 0%**
- **Max Peak Intensity = 2%**
- **Min. Intensity = 0**
- **Min. Area = 0**

Références

[1] Applied Biosystems. Data Explorer™ Software. User Guide. Version 4.0 Software. Printed in the United States of America. Part Number 4317717 Rev. A, 08/2000.

[2] Applied Biosystem. Removing Chemical Background from QSTAR® XL System Spectra Using a Fast Fourier Transformation Filter: Technical Note 115155.

[3] A. ABBOTT A Post-Genomic Challenge: Learning to Read Patterns of Protein Synthesis. Nature 402, 1999, pp. 715-720.

[4] NL. ANDERSON. Proteome and Proteomics. New Technologies, new concepts, and new words. Electrophoresis. 1998, pp. 1853-1861.

[5] P. BERNDT, U. HOBOHM , H. LANGEN. Reliable Automatic Protein Identification From Matrix-Assisted Laser Desorption/Ionization Mass Spectrometric Peptide Fingerprints. Electrophoresis 20, 1999, pp. 3521-3526.

[6] K.L. BUSCH. Chemical noise in mass spectrometry - Part I. Mass Spectrometry Forum Issue of Spectroscopy. October, 2002.

[7] G. CHANG, B. YU, and M. VETTERLI. Adaptive wavelet thresholding for image denoising and compression. IEEE Trans. Image Processing 9: pp. 1532-1546, 2000.

[8] J. GODOVAC-ZIMMERMANN, L. R. BROWN. Perspectives For Mass Spectrometry and Functional Proteomics. Mass Spectrometry Reviews, 2001, Rev. 20, pp. 1-57.

[9] S. GOLOTVIN, A. WILLIAMS, Improved Baseline Correction of FT NMR Spectra. Advanced Chemistry Development, NMR Newsletter Advanced Chemistry Development, 1999.

[10] S. GRACE, B. YU, M. VETTERLI, Adaptive wavelet Thresholding for Denoising and Compression, IEEE Transactions on image processing, Vol. 9, NO.9, 2000, pp. 1532-1546.

[11] H. HAUSSECKER, H. R. TIZHOOSH. Fuzzy Image Processing. I. Handbook of Computer Vision Application. Edited by B. Jagne, H. Haussecker, and P. Geisster, Academic Press 1999.

[12] E. D. JANSING, T. A. ALBERT, D. L. CHENOWETH, Two Dimensional Entropy Segmentation. Pattern Recognition Letters 20, Letters 20, 1999, pp. 329-336.

[13] H. LIM, J. ENG, J.R. YATES, S.L TOLLAKSEN, C.S. GIOMETTI, J.F HOLDEN, M.W.W. ADAMS, C.I. REICH, G.J. OLSEN, L.G. HAYS. Identification of 2D-gel proteins: A comparison of MALDI/TOF peptide mass mapping to LC-ESI tandem mass spectrometry. Journal of the American Society for Mass Spectrometry, Volume 14, Issue 9, 2003, pp. 957-970.

[14] A. LINDEGREN, Analysis of Proteomic Patterns for Detection of Prostate Cancer. Master Thesis. 2004.

[15] P. LIO, Wavelets in bioinformatics and computational biology: state of and perspectives, Bioinformatics, Vol. 19, 2003, pp. 2-9.

[16] B. LIU, Y. SERA, N. MATSUBARA, K. OTSUKA, S. TERABE. Signal Denoising by Wavelets for Microchip Electrophoresis, Tokyo-Society for Chromatograpghic Sciences, Vol. 23, 2002, pp. 59-60.

[17] D.I. MALYARENKO, W.E. COOKE, B.L. ADAM, G. MALIK, H. CHEN, E.R. TRACY, M.W. TROSSET, M. SASINOWSKI, O.J Semmes, and D. M. MANOS. Enhancement of sensitivity and resolution of Surface-Enhanced Laser Desorption/Ionisation Time-of-Flight Mass Spectrometric Records for Serum Peptides Using Times Series Analysis Technique, Proteomics and Protein Markers, Clinical Chemistry, 2005, pp. 65-74.

[18] M. MANN. Quantitative proteomics, Nat Biotechnol. 17, 1999, pp. 954-955.

[19] N. NAFATI. Synthèse itérative et simultanée de banc de filtres bi orthogonaux de reconstruction parfaite avec des critères adaptés aux applications de codage de la parole et de l'image, GRETSI Grenoble France. 1997, pp. 1085-1088. Septembre.

[20] P. NUGUES. Interprétation de Gels d'Electrophorèses 2D, Thèse de Doctorat, Université de Nancy, 1989.

[21] F. PARISI. Analysis of SELDI mass spectra of ovarian cancer blood serum samples. Thesis for the Degree of Master of Science. Master's Programme in Bioinformatics. Chalmers University of Technology and Göteborg University. Sweden. March 2004.

[22] T. D. PHAM. A New Approach for Calculating Implications of Fuzzy Rules. IEEE International Conference on Artificial Intelligence Systems, 2002, pp. 71.

[23] J. POLEC, J. PAVLOVIEOVA, T. KARLUBIKOVA. Application Of Shape-independent Orthogonal Transform For Image Inerpolation, Radio-Engineering, Vol. 11, No. 1, April 2002.

[24] E. G. SÁNCHEZ, Y.A. DIMITRIADIS, M. SANCHEZ-REYES MAS, P. S. GARCÍA, J.M. CANO IZQUIERDO, J. LOPEZ CORONADO, On-Line Character Analysis and Recognition With Fuzzy Neural Networks, Intelligent Automation and Soft Computing, Vol. 7, No. 3,1998, pp. 161-162.

[25] M. SCHREINER, K. STRUPAT, F. LOTTSPEICH, C. ECKERSKORN. Ultraviolet matrix assisted laser desorption ionization mass spectrometry of electroblotted proteins. Electrophoresis 17, 1996, pp. 954-961.

[26] John R. YATES. Mass Spectrometry and the Age of the Proteome. Journal of Mass Spectrometry, Vol. 33, pp. 1-19,1998.

[27] R. ZENOBI, R. KNOCHENMUSS, Ion formation in MALDI Mass Spectrometry. Mass Spectrometry 1998, Rev. 17, pp. 337-66.

N. M. **Nafati**

Nicolas Mohammed Nafati. Né en 1960 à Boulanouar (Maroc). Actuellement, ingénieur de recherche à l'INSERM. Responsable de la bio-informatique de la plate-forme Protéomique-Pasteur. Faculté de Médecine. Nice.

-Docteur en traitement d'Images et Signal, Master II en Systèmes Physiques et Métrologie au CNAM de Paris. -Ingénieur de l'ENSIETA de Brest en Electronique/Informatique. -Master I obtenu à la faculté des sciences de Tours en Physique. Mention Energétique/Informatique.

Ses principaux centre d'intérêt et travaux de recherche portent sur l'imagerie protéomique, le traitement d'image protéomique et le traitement du signal appliqué à la réduction des spectres de masse. Email : nafati@unice.fr. Web : http://www.ifr50.fr

Au nom de l'équipe de la plate forme protéomique et au nom du secrétaire général de l'IFR50, nous dédions ce travail à Monsieur Bernard Rossi. Ce travail, dont il n'a pas pu voir l'aboutissement, a été initié grâce à lui.

B. **Rossi**

Dr. Bernard Rossi. Né en 1949 à Monaco (France), décédé le 6 Mai 2006 à Nice (France). Directeur de Recherche à l'INSERM. Directeur de l'IFR50 de 1999 à 2006. Directeur de l'Unité 638 INSERM depuis 1990.

Ses principaux centres d'intérêt portent sur les mécanismes impliqués dans l'activation et la prolifération des cellules immuno-compétentes. Mr Bernard Rossi était membre des Sociétés Scientifiques suivantes : Expert auprès de la National Science Foundation. Membre de l'American Association of Immunologists. Membre de l'American Society for Micro-biology. Membre de l'American Association for the Advancement of Science. Membre de la International Cytokine Society.

J.M. **Guigonis**

Jean Marie Guigonis. Né 1957 à Nice (France). Ingénieur de recherche universitaire à la faculté de Médecine Nice-Pasteur. Ingénieur DPE et Européen.

De 1986 à 2001. Responsable technique du service de spectrométrie de masse de la Faculté des Sciences de Nice Sophia-Antipolis.

Son domaine d'analyse est à cette époque les petites molécules organiques issues de la synthèse. Depuis 2001 a rejoint la plate-forme Protéomique Pasteur de l'IFR50 de la Faculté de Médecine Nice. Il a en charge la réalisation et l'interprétation des analyses réalisées en spectrométrie de masse. Il intervient également comme enseignant au sein du Master-Protéomique. Email : guigonis@unice.fr.

M. **Samson**

Chercheur à l'INSERM U. 638. Né le 11 septembre 1954 à Rabat (Maroc). Responsable scientifique de la plate-forme protéomique-Pasteur. IFR50. Faculté de Médecine Pasteur. Nice.

Ses principaux centres d'intérêt portent sur l'analyse du protéome et du métabolome des tumeurs coliques dans le but de rechercher des marqueurs à usages diagnostiques pronostiques ou thérapeutiques. Email : samson@unice.fr