Reconstruction de profils moléculaires : modélisation et inversion d'une chaîne de mesure protéomique

Grégory Strubel

INSTITUT POLYTECHNIQUE DE GRENOBLE

N° attribué par la bibliothèque
|__|__|__|__|__|__|__|__|__|__|

T H E S E

pour obtenir le grade de

DOCTEUR DE L'Institut polytechnique de Grenoble

Spécialité : Signal, Image, Parole et Télécoms

préparée au laboratoire

Electronique et Systèmes pour la Santé du Commissariat à l'Energie Atomique de Grenoble

dans le cadre de **l'Ecole Doctorale**

Electronique, Electrotechnique, Automatique, Télécommunications et Signal

présentée et soutenue publiquement

par

Grégory STRUBEL

le 1er décembre 2008

Reconstruction de profils moléculaires : modélisation et inversion d'une chaîne de mesure protéomique

Directeur de thèse : Jean-François Giovannelli
Co-directeur de thèse : Pierre Grangeat

JURY

M. Jérôme	MARS	, Président
M. Jérôme	IDIER	, Rapporteur
M. Jérôme	LEMOINE	, Rapporteur
M. Jean-François	GIOVANNELLI	, Directeur de thèse
M. Pierre	GRANGEAT	, Co-encadrant
M. Alain	VIARI	, Examinateur
M. Christophe	MASSELON	, Examinateur

A Ali Marie-Pier

Remerciements

Voici venu le temps pour moi d'écrire les dernières pages de ma thèse et ainsi clore une aventure passionnante. Cher lecteur, les pages qui suivent vont te[1] raconter une histoire scientifique captivante, celle d'une rencontre entre deux domaines de pointe riches en promesses. J'espère qu'elle te plaira.

Toutefois, tu ne trouveras pas l'autre histoire, plus humaine, qui raconte mes trois dernières années. Pourtant, celle là non plus ne manque pas de charme. Parfois drôle, parfois haletante. Elle est tissée de courses contre la montre, de mariage au bout du monde, d'espoirs, de déceptions, de joies intenses et de rebonds divers[2]. Mais, avant tout, elle est remplie de personnages exceptionnels que tu aimerais, j'en suis sûr, rencontrer. C'est eux qui m'ont permis d'en arriver là et j'aimerais les remercier.

Mes premiers remerciements vont à mes directeurs de thèse Jean-François Giovannelli et Pierre Grangeat. Il est dur de résumer en peu de mots tout ce qu'ils m'ont apporté. J'ai eu la chance d'avoir pour encadrant de très grands scientifiques qui n'ont pas été avares de leur temps pour me guider avec intelligence et disponibilité. Et tout ça sans compter leur infatigable patience devant mes nombreuses fautes d'orthographe, elle doit être louée. Merci à vous.

Je remercie Jean Chabbal, Olivier Peyret, Philippe Rizo et Régis Guillemaud pour m'avoir fait confiance en me donnant ce sujet. Merci aussi d'avoir accepté mes nombreux voyages parisiens, malgré ces périodes de tensions fortes exercées sur la recherche. Je les remercie pour leur attention et leur soutien constant.

J'adresse mes profonds remerciements à Jérôme Idier et Jérôme Lemoine pour m'avoir fait l'honneur d'expertiser mon travail en acceptant le rôle de rapporteur. Et, de façon plus générale, je remercie Jérôme Mars, Alain Viari et Christophe Masselon pour avoir bien voulu faire partie de mon jury mais également pour avoir analysé mon manuscrit dans ses moindres détails.

Je remercie Guy Demoment, Ali Mohammad-Djafari et Thomas Rodet qui m'ont chaleureusement accueilli dans leur équipe lors de mes séjours au LSS. Leurs conseils ont été précieux pour avancer, mais plus que leur expertise technique, ce sont les conviviales conversations au « tripot » que je retiendrais. Merci pour ce café corsé et pour les pâtisseries exotiques que vous avez généreusement partagés avec moi.

Un grand merci à Laurent Gefault et Caroline Paulus. Merci d'avoir quotidiennement été en première ligne à mes côtés. Merci d'avoir supporté mes délires, d'avoir accepté que je couvre vos tableaux d'équations et de schémas obscurs. Merci simplement d'avoir été là, de m'avoir écouté. Merci de m'avoir aidé, je n'aurais pas pu en arriver là sans vous.

Je remercie Patrick Hugonnard et Alain Nocca pour m'avoir chaleureusement accueilli dans leurs bureaux respectifs. Merci de m'avoir considéré comme des vôtres immédiatement. Merci pour vos récits d'avant. Merci pour vos conseils qui m'ont permis de comprendre la jungle administrative du

[1] Je me permets de te tutoyer, comme nous allons passer un peu de temps ensemble, j'espère que tu ne t'en offusqueras pas.

[2] Certes, il n'y a aucun dragon, mais il y a eu quelques duels à l'épée et surtout le terrible labyrinthe de Minatec avec, caché en son sein, le secrétariat SIPT : la solution était pourtant simple, il suffisait de passer par le parking !

CEA. Et merci d'avoir partagé votre sagesse pour mes nombreux problèmes de bricolage domestique. Merci à toi Fabien Vrillon, grâce à toi cela a été un plaisir de venir travailler.

Merci à mes frères de thèse Alexandre Chibane et François Orieux. Alexandre, toi qui était toujours là pour répondre à mes appels à l'aide mathématiques quand tout le monde avait renoncé. Merci de m'avoir fait rêver avec les magnifiques images que tu ramenais du toit du monde. Merci à toi, François, toi qui m'a fraternellement accueilli au LSS. Grâce à toi, je me suis senti là-bas aussi, chez moi.

Merci à tous les thésards et post-doc du DTBS. Et, en particulier, merci à Agnès Fonverne pour m'avoir patiemment appris tout ce que je sais sur la chromatographie. Ta présence m'a manqué quand tu as été exilée à l'autre bout du centre. Un grand merci à Jean Rinkel et à Thomas Lamotte pour avoir supporté le petit nouveau alors que vous étiez dans votre douloureuse dernière année. Merci à toi Ricardo Escola pour avoir partagé avec moi la boisson magique des argentins qui vous permet d'être aussi fort au football et au rugby. Merci à vous, Timothée Levi et Philippe-Antoine David pour avoir été des soutiens quotidiens lors de ma dernière année. Merci à Christopher Coello et à Anne Frassati, vous qui avez été là du début à la fin, partageant tout avec moi. Et enfin, mes respectueuses salutations à monsieur Ernest Galbrun.

Je voudrais exprimer ma gratitude aux clefs de voute du département, Michelle Chambaz et Véronique Birkenheier, pour avoir réglé tant de problèmes insolubles immédiatement et, en particulier, les miens. Merci pour votre bonne humeur constante.

Un grand merci à Régis Guillemaud, Alain Bourgerette, Caroline Paulus, Eric Charierre, Jean-François Bêche, Michel Antonakios, Guillaume Charvet, Pascale Pham, Philippe-Antoine David, Sadok Gharbi, Stéphane Bonnet, Venceslass Rat, Pierre Grangeat, Ricardo Escola, Tetiana Aksenova, Thierry Flaven, Michael Palmieri, Alain Nocca, Laurent Gerfault, Pierre Jallon, Antoine Defontaine et Fabien Vrillon, j'ai été honoré de participer avec vous à la création du laboratoire LE2S. Je suis fier d'avoir fait partie de votre équipe. Merci à vous Jean-François, Alain et Fabien pour m'avoir initié aux subtilités du travail d'électronicien. Pascale, tu avais raison, comme pour tes courses de montagne, on ne regrette rien une fois la ligne d'arrivée franchie.

Je tiens à remercier l'ensemble des professeurs qui m'ont formé. Vous avez tous contribué à ce travail. Je vous en remercie. La première d'entre elle fut ma mère, sans qui bien sûr je ne serais rien. Je remercie mon père qui m'a donné le goût des sciences. Merci à vous deux, vous avez tellement fait pour moi. Et je remercie mon grand père et mon frère qui seront toujours pour moi des modèles à qui j'aimerais ressembler.

Et surtout, merci à toi, Ali Marie-Pier, mon âme sœur, qui partage mes jours et mes nuits. Tu es vraiment ma moitié et c'est au moins la part de ce travail qui te revient. Merci Ali.

Table des matières

Conventions et notations

Conventions

Nomenclature des éléments du système de mesure

Afin d'éviter les confusions, voici les conventions adoptées dans ce document concernant les termes mesure, chaîne d'analyse, instrument et système de mesure.

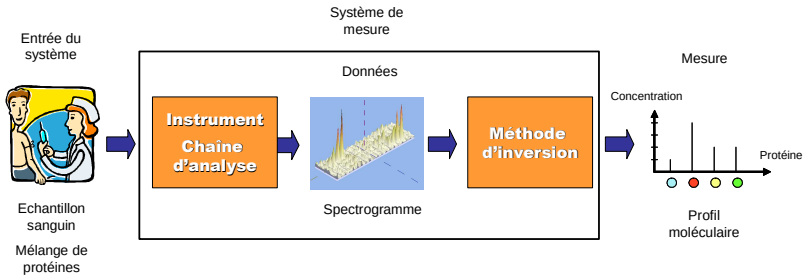


Figure 1 : nomenclature des éléments du système de mesure.

L'instrument et la chaîne d'analyse décrivent le transducteur qui transforme l'échantillon prélevé en signal correspondant à la grandeur physique que nous souhaitons mesurer. Cependant l'instrument n'étant pas parfait, une étape de traitement peut être nécessaire avant d'accéder à la mesure souhaitée. Le système de mesure englobe donc instrument et méthode d'inversion numérique.

Loi normale et fonction gaussienne

La loi normale est utilisée abondamment dans les applications de traitement du signal, notamment pour modéliser le bruit. Dans cette thèse, nous essaierons également d'ajuster des fonctions gaussiennes sur le signal. Nous distinguerons au niveau du vocabulaire l'utilisation en statistique des autres utilisations. Pour une utilisation statistique, nous utiliserons le mot « loi normale » et pour les autres utilisations, nous utiliserons le mot « fonction gaussienne ».

Paramétrage d'une fonction gaussienne

Plusieurs façons de paramétrer une fonction gaussienne existent, notamment pour caractériser la « largeur » de ces fonctions : largeur à mi-hauteur, écart type, variance. Dans cette thèse nous utiliserons pour caractériser la largeur, l'inverse variance appelée parfois paramètre de précision. Cette notation de la largeur, même si elle est moins utilisée et plus difficile à lire sur une courbe, permet cependant d'alléger de façon conséquente certains calculs.

La gaussienne centrée en μ de largeur γ appliquée à la variable t sera notée

$$N(t;\mu,\gamma)=(2\pi)^{-1/2}\gamma^{1/2}\exp\left(-\frac{1}{2}\gamma(t-\mu)^2\right)$$

Notons qu'il s'agit d'une fonction normalisée dont l'intégrale est égale à 1.

Nous pouvons également étendre cette notation aux fonctions gaussiennes multivariées, la gaussienne centrée sur le vecteur $\boldsymbol{\mu}$ et de matrice de largeur $\boldsymbol{\Gamma}$ appliquée à la variable $\boldsymbol{x}$ sera notée

$$N(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Gamma}) = (2\pi)^{-1/2}\left|\boldsymbol{\Gamma}\right|^{1/2}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^t\boldsymbol{\Gamma}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

La matrice de largeur est l'inverse de la matrice de covariance.

Notations

$\boldsymbol{H}$	matrice
$\boldsymbol{x}$	vecteur
h_{ij} ou $h_{i,j}$	élément de la ligne i et la colonne j de la matrice H
x_i	$i^{\text{ème}}$ élément du vecteur $\boldsymbol{x}$
$\boldsymbol{x}_{-i}$	vecteur $\boldsymbol{x}$ sans l'élément x_i
$\text{diag}(\boldsymbol{x})$	matrice diagonale dont les éléments diagonaux sont les éléments du vecteur $\boldsymbol{x}$
$\propto$	proportionnel à
$p(a\|b)$	densité de probabilité de a sachant b
$\sim$	distribué selon
$\hat{a}$	estimation de la variable a

1. Introduction

A travers le projet européen LOCCANDIA et le projet transverse du CEA CAPSI, le département des micro-technologies pour la biologie et la santé (DTBS) développe une chaîne d'analyse pour la détection précoce du cancer. Cette dernière repose sur la quantification de protéines caractéristiques de cette maladie dans le sang. Dans cette optique, le DTBS élabore un laboratoire sur puce dédié à cette analyse. Parallèlement au développement de ces composants miniaturisés, le laboratoire développe des méthodes de traitement numérique adaptées afin d'améliorer la mesure.

L'objet de cette thèse est d'appliquer à ce problème les outils du traitement du signal et plus spécifiquement la démarche des problèmes inverses. Elle a été réalisée en collaboration avec le laboratoire des signaux et systèmes (LSS), unité mixte de recherche CNRS–Supélec–Université Paris Sud.

Le caractère multidisciplinaire de cette thèse amènera le lecteur à affronter un contexte biologique, une instrumentation de chimie analytique complexe, et les outils du traitement du signal. Ce chapitre a pour but d'éclaircir ces différentes notions quelle que soit la communauté d'origine du lecteur.

1.1. De l'ADN à la protéine

La découverte de la structure en double hélice de l'ADN en 1953 par Watson, Crick, Wilkins et Franklin est l'événement fondateur de la biologie moléculaire. Il a sonné le début de la compréhension des causes moléculaires des mécanismes biologiques.

Le décryptage du génome a permis de grandes avancées en médecine. Il a notamment permis de détecter les prédispositions des individus aux maladies. Cependant, le génome contient bien moins de gènes que prévu, 25 000 selon les estimations actuelles, bien peu comparé au nombre de fonctions codées. Une explication de ce décalage est que l'ADN n'est que le début d'une chaîne. Chaque gène peut produire plusieurs brins d'ARN messagers différents et chaque ARN messager peut produire plusieurs sortes de protéines. Ces mécanismes permettent au final de produire un grand éventail de protéines différentes (peut-être un million) pour réaliser les différentes fonctions de l'organisme. Par exemple, elles catalysent les réactions chimiques, servent de messager chimique à travers l'organisme ou forment des moteurs moléculaires complexes (pour faire bouger les muscles par exemple). Elles servent même parfois de matériaux de base.

Les protéines remplissent de nombreux rôles dans l'organisme, mais les mécanismes qui permettent d'en générer un grand nombre contrôlent également leur quantité au cours du temps. ADN, ARN messagers et protéines interagissent et répondent aux stimuli de l'environnement. La quantité de chaque protéine est régulée pour s'adapter suivant les situations. Ainsi, on obtient des organismes aussi différents qu'un têtard et une grenouille à partir d'un unique patrimoine génétique.

L'étude du génome seul ne suffit donc pas pour comprendre les mécanismes régissant la vie cellulaire. Afin de compléter les informations de la génomique, la protéomique se donne pour but d'étudier la nature et la quantité des protéines d'une cible biologique à un moment donné. Elle s'intéresse de plus à la fonction de chaque protéine et à ses interactions avec les autres molécules et en particulier les relations inter-protéines.

Dans le but de comprendre les causes moléculaires des maladies, les protéines sont donc d'excellents sujets d'étude. Par la suite nous allons nous intéresser essentiellement aux techniques d'analyse quantitative des protéines. Pour des informations générales sur l'ADN, l'ARN, les protéines ou le fonctionnement d'une cellule, le lecteur pourra consulter les ouvrages de biochimie et de biologie cellulaire [1, 2].

1.2. Protéomique clinique et profils moléculaires

Dans une perspective médicale, la connaissance des mécanismes moléculaires promet le développement de nouveaux médicaments. Les protéines sont également étudiées afin de servir de biomarqueurs, notamment dans le but d'un diagnostic précoce de maladies comme le cancer.

Le cancer est causé par la combinaison de plusieurs altérations de l'ADN d'une cellule entraînant un comportement anormal. Ces dysfonctions produisent une division incontrôlée de la cellule initiatrice et l'altération des fonctions vitales de l'organisme [3]. De façon générale, le pronostic est d'autant meilleur que le cancer est détecté précocement.

Même si elles sont encore en développement, les méthodes de dépistage basées sur la protéomique visent à identifier les protéines produites par ces modifications génétiques dans les fluides biologiques : sang, urine, salive, *etc.* Des variations dans les niveaux de protéines normales seraient également révélatrices d'un dysfonctionnement. Le but est de déceler les premiers signes des modifications de la cellule avant les premiers symptômes morphologiques. Le biomarqueur idéal est présent en quantité relativement importante, il est très spécifique et n'entraîne ni faux positif, ni faux négatif. Cependant un tel cas se présente rarement et les protéines potentiellement intéressantes sont présentes en concentration extrêmement faible [4-6]. Quelques protéines sont actuellement utilisées en routine clinique en tant que biomarqueur du cancer : AFP, CEA, PSA, CA125, *etc.*, mais aucune n'a les caractéristiques désirées pour constituer un diagnostic suffisamment spécifique à elle seule [7]. C'est pourquoi les recherches s'orientent vers l'étude simultanée d'un panel de protéines (appelé profil moléculaire) associé à la maladie.

D'un point de vue analytique, le défi est de taille : un échantillon sanguin standard contient environ 100 mg/ml de protéines, alors que la concentration des protéines d'intérêt est de l'ordre de quelques ng/ml [7], soit 8 ordres de grandeur. De plus, la composition de l'échantillon est également complexe, on dénombre actuellement plusieurs milliers de protéines dans le plasma [8]. Des problèmes similaires apparaissent dans les autres types d'échantillons biologiques.

1.3. Instrumentation

Dans cette thèse, nous nous sommes concentrés sur les traitements numériques des données issues d'une chaîne d'analyse basée sur la chromatographie et la spectrométrie de masse. Pour cela, nous avons utilisé les données issues d'expériences antérieures réalisées par les expérimentateurs du CEA. Nous n'avons pas contribué à la définition des protocoles, ni cherché à améliorer la chaîne de mesure utilisée. Si les instruments employés ont été choisis précédemment, nous donnons dans cette section quelques éléments d'information concernant les choix technologiques effectués.

L'analyse des protéines, surtout dans des fluides biologiques, n'a jamais été une chose aisée. Les avancées dans ce domaine sont intimement reliées aux progrès des méthodes séparatives. Les protéines sont d'ailleurs souvent les premières molécules analysées [8].

Parmi les méthodes les plus utilisées en protéomique nous pouvons citer les suivantes.

- *Les puces à anticorps et méthodes ELISA*. Elles sont basées d'une part sur l'association spécifique de la protéine cible et d'un anticorps spécialement choisi, d'autre part sur une méthode de détection généralement optique reconnaissant cette association. En ce qui concerne la quantification des marqueurs cancéreux, elles sont considérées comme le standard de référence, notamment en raison de leur sensibilité [9]. Cependant, la réussite de cette

méthode dépend de l'anticorps utilisé, il faut que la complémentarité protéine-anticorps soit la plus spécifique possible.

- *Les méthodes par électrophorèse* (SDS-PAGE, gels à deux dimension 2-DE). Pendant plusieurs dizaines d'années, cette méthode a été la technique de séparation analytique la plus utilisée en protéomique. Les gels obtenus étant numérisés, leurs images sont ensuite comparées à une référence. Le principal défaut de cette méthode est son manque de sensibilité.
- *Les méthodes par spectrométrie de masse.* Elles suscitent un intérêt croissant [10]. Leur principal atout est qu'elles permettent une identification précise des protéines impliquées dans le mélange. Cette faculté d'identification en fait le principal outil de la recherche de biomarqueurs. La spectrométrie de masse a souvent été critiquée pour ne pas être suffisamment quantitative et sensible pour les applications cliniques [7], mais ses performances sont en progression constante.

La principale limite des puces à anticorps concerne la spécificité de la liaison protéine-anticorps dans les milieux complexes comme le sang. Dans ces situations, elles auront tendance à être sujettes aux faux positifs et négatifs. A l'opposée, les méthodes issues de la chimie analytique comme la spectrométrie de masse, viseront à séparer chacun des constituants du mélange. Pour dépasser les problèmes quantitatifs, de nombreuses techniques d'étalonnage utilisant des isotopes lourds ont été développées. Leur sensibilité dépend beaucoup des étapes de préparation de l'échantillon et du type de spectromètre utilisé. La technologie évolue rapidement, les spectromètres de masse MRM par exemple, affichent déjà des performances équivalentes aux méthodes ELISA (section 2.4.2.1).

Suivant les travaux précurseurs de Petricoin *et al.* [11], les méthodes par spectrométrie de masse dépassent le cadre de la recherche de biomarqueurs pour se rapprocher de la protéomique clinique. C'est également le mode de mesure choisi dans notre approche. Bien sûr, nous n'avons fait qu'effleurer la description de la chaîne de mesure et la présentation des approches concurrentes. Tout d'abord, un spectromètre de masse n'accepte en entrée qu'un gaz d'ions. En protéomique, le spectromètre de masse est donc indissociable d'une technologie permettant d'ioniser le mélange à analyser. Ensuite, on lui adjoint généralement une ou plusieurs méthodes de séparation permettant de simplifier le mélange. Le monde de la protéomique par spectrométrie de masse est donc riche en technologies différentes. Le lecteur pourra se référer à [10, 12] pour comparer les avantages et les défauts de chaque technique. Nous nous concentrerons dans nos travaux sur la chaîne de mesure qui associe une colonne de chromatographie liquide, un electrospray et une trappe ionique linéaire. Toutefois, si les technologies concurrentes n'utilisent pas les mêmes principes physiques, elles produisent des données assez similaires. La méthode présentée dans ce document pourra en grande partie se généraliser à d'autres chaînes.

1.4. Approche de la thèse

Nous avons décrit notre but qui est l'estimation des profils moléculaires et évoqué l'instrumentation utilisée. Quelle place cela laisse-t-il au traitement du signal ?

La plupart du temps, pour effectuer une mesure, on essaie de se ramener à une relation biunivoque, et, si possible, linéaire entre la grandeur recherchée et une grandeur facilement observable. Par exemple pour un thermomètre à mercure, on s'appuie sur une théorie physique qui stipule que la variation de la hauteur de mercure est proportionnelle à la variation de la température. Le problème « difficile » d'estimation de la température a été ramené à une mesure « facile » de distance. Dans certains cas plus complexes, il devient impossible de se ramener à une telle relation : non-linéarités, limite de sensibilité, données lacunaires, données à concilier… Parallèlement à l'amélioration de l'instrument, l'amélioration du traitement des signaux observés gagne à être étudiée.

Dans cette thèse, nous développons une méthode d'estimation des concentrations dont l'objectif est de dépasser les performances des méthodes traditionnelles. Dans une perspective d'utilisation clinique, seul le problème de la quantification des protéines sera traité. En particulier, nous n'évoquerons pas le problème de la découverte de nouveaux biomarqueurs, ni de celui du séquençage des protéines

(identification de sa structure moléculaire). Nous développons notre méthode dans l'objectif de traiter les signaux suivants :

1) les signaux faibles par rapport au bruit de la chaîne de mesure,
2) les signaux insuffisamment séparés par la chaîne de mesure.

Nous plaçons cette étude dans le cadre des problèmes inverses. Ce type d'approche est basé sur deux étapes principales. Dans la première, on établit l'ensemble des équations mathématiques qui relient les données observées aux concentrations recherchées. Dans la seconde, on cherche les concentrations et les paramètres du modèle. Pour prendre en compte la méconnaissance de ces paramètres, nous utilisons une approche probabiliste issue des statistiques bayésiennes.

La résolution de notre problème est présentée avec le découpage suivant.

- *Chapitre 2*. Dans ce chapitre, nous expliquerons le fonctionnement de l'instrument de mesure, et nous le modéliserons. La chaîne d'analyse est constituée de plusieurs modules classiquement utilisés en chimie analytique : colonne de chromatographie, electrospray, spectromètre de masse. Ces appareils existent depuis de nombreuses années et une littérature importante leur a été consacrée. Dans cette thèse, nous présentons une synthèse de la littérature existante et produisons un modèle global adapté au traitement du signal. Nous mettrons notamment en évidence les sources de variations et les dépendances des différentes variables.
- *Chapitre 3*. Le modèle étant posé, ce chapitre présente les différentes méthodes utilisées en protéomique pour estimer les concentrations, ainsi que leurs limites. Nous détaillerons en particulier les méthodes utilisées pour confronter nos résultats. En plus des méthodes spécifiques à la protéomique, nous donnerons une synthèse des différentes méthodes bayésiennes adaptées aux données spectrométriques en général.
- *Chapitre 4*. Ce chapitre expose notre méthode. Elle estime les concentrations et les paramètres du modèle étant donné un modèle et des données observées. En effet, la principale difficulté de l'estimation des concentrations provient de l'imperfection de notre modèle et la méconnaissance de ses paramètres. Notre méthode prendra notamment en compte les fluctuations des positions des pics et du gain du système. Plus précisément, nous modéliserons l'information dont nous disposons sous forme de lois de probabilité *a priori*. Puis, nous calculerons la loi de probabilité *a posteriori* permettant d'associer à chaque valeur possible des paramètres recherchés une probabilité. Finalement, la valeur fournie par notre méthode sera l'estimateur de la moyenne. Pour calculer cet estimateur nous utilisons les techniques de Monte Carlo par chaîne de Markov (MCMC).
- *Chapitre 5*. Dans ce chapitre nous validerons notre méthode à l'aide de données simulées et de données réelles. Tout d'abord, nous utiliserons notre méthode pour estimer les concentrations de morceaux de protéines issus d'un digestat de cytochrome C. Ensuite, nous estimerons la concentration de toxines dans de l'urine.
- *Chapitre 6.* Dans ce chapitre nous établirons le bilan de notre travail. Puis nous évoquerons les différentes perspectives qui se dessinent.

2. Principes et modélisation

Dans ce chapitre, nous présentons la chaîne d'analyse utilisée pour reconstruire les profils moléculaires. Elle a pour but de traiter un échantillon biologique dont l'exemple le plus caractéristique est le sang. Elle fournit en sortie un signal à deux dimensions que nous nommerons spectrogramme (Figure 2).

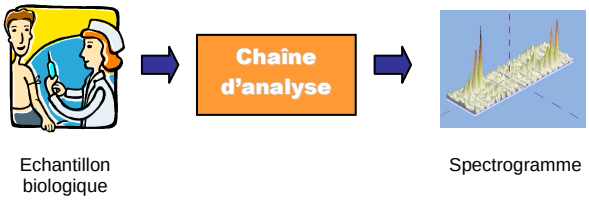


Figure 2 : chaîne d'analyse.

Chaque module de cette chaîne (Figure 3) sera décrit et modélisé, afin d'obtenir un jeu d'équations reliant les inconnues recherchées aux données.

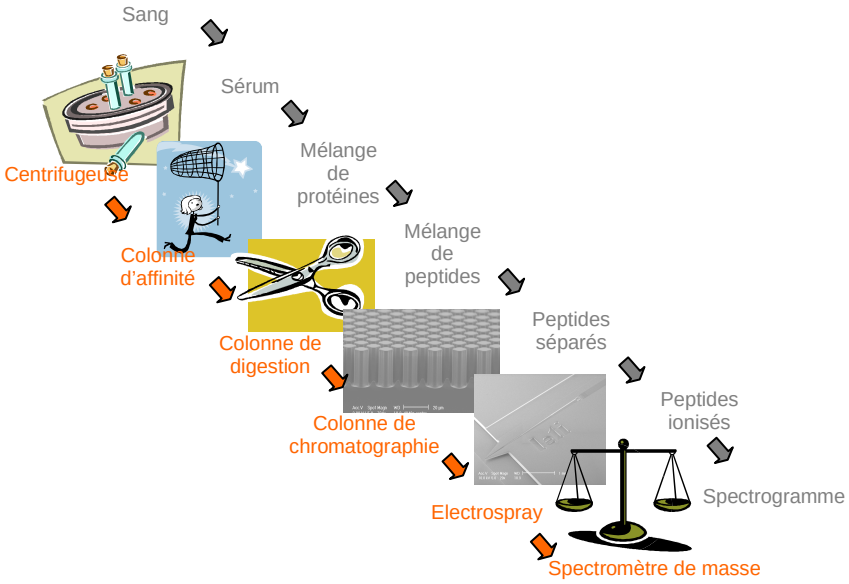


Figure 3 : les différents modules de la chaîne d'analyse.

La chaîne d'analyse utilisée est principalement basée sur deux technologies analytiques : un spectromètre de masse et une colonne de chromatographique liquide. Cependant, un spectromètre de masse analyse un mélange gazeux d'ions et la chromatographie manipule des liquides. Un electrospray

nébulisant le mélange en sortie de colonne de chromatographie est donc utilisé afin de coupler les deux modules utilisant deux phases distinctes de la matière.

De plus, les protéines étant des macromolécules, leurs masses ne sont pas compatibles avec la gamme de mesure d'un spectromètre de masse normal. Afin de dépasser cette limitation, nous n'analyserons pas directement le mélange de protéines, mais chaque protéine subira une étape de digestion, où elle sera découpée en plusieurs morceaux appelés peptides. Ce sont ces peptides qui seront analysés par chromatographie et spectrométrie. Enfin, deux étapes de séparation et de concentration sont ajoutées afin d'améliorer les performances : centrifugation et colonne d'affinité. Ces trois étapes, situé en début de chaîne d'analyse, sont appelées étapes de préparation de l'échantillon.

2.1. Etapes de préparation

Parmi les étapes de préparation, l'étape de digestion joue un rôle central. Les deux autres étapes de préparation, utilisant une centrifugeuse et une colonne d'affinité, simplifient l'échantillon en le purifiant. Nous supposerons que dans ces étapes il n'y a pas de perte de protéines d'intérêt.

Pendant l'étape de digestion, chaque protéine est découpée en plusieurs morceaux, appelées peptides, en des endroits prédéterminés. Une protéine est formée par une succession de petites molécules organiques appelés acides aminés. Il en existe 20 différents, chacun étant noté par une lettre de l'alphabet[1]. Dans notre système, la digestion est réalisée par la trypsine. Cette enzyme découpe les protéines après l'arginine et la lysine notées respectivement R et K. Chaque protéine produira des peptides aux séquences en acides aminés bien déterminées (Figure 4).

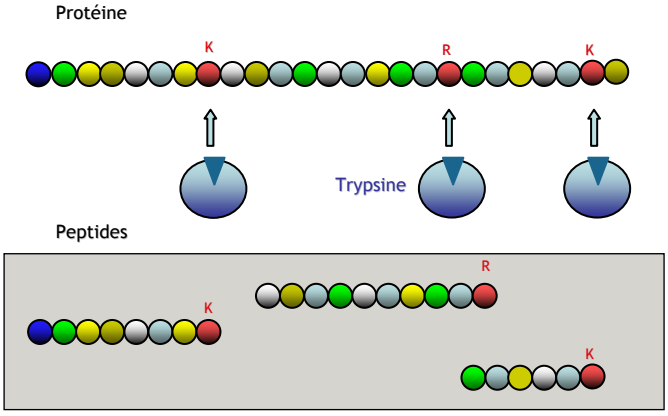


Figure 4 : digestion par la trypsine.
Chaque cercle représente un acide aminé. La trypsine coupe après la lysine et l'arginine. Figure fournie par J. Garin CEA, IRTSV, Laboratoire d'Etude de la Dynamique des Protéomes, Grenoble.

Dans une protéine, des séquences identiques d'acides aminés peuvent apparaître car leur nombre est limité. Il est donc possible qu'une protéine unique génère plusieurs exemplaires d'un peptide et plusieurs protéines différentes peuvent produire un peptide identique (Figure 5).

Soit d_{ip} le nombre de peptides i générés par une protéine p et $\boldsymbol{D}$, la matrice de digestion telle que $(\boldsymbol{D})_{ip} = d_{ip}$. Prenons le cas présenté Figure 5, nous avons

$$\boldsymbol{D} = \begin{bmatrix} 2 & 0 \\ 1 & 1 \\ 0 & 1 \end{bmatrix}$$

La première colonne se lit de la façon suivante : la protéine 1 produit deux exemplaires du peptide 1 et un peptide 2, mais pas de peptide 3.

[1] Voir pages 74 et 75 de [1] pour une classification des différents acides aminés

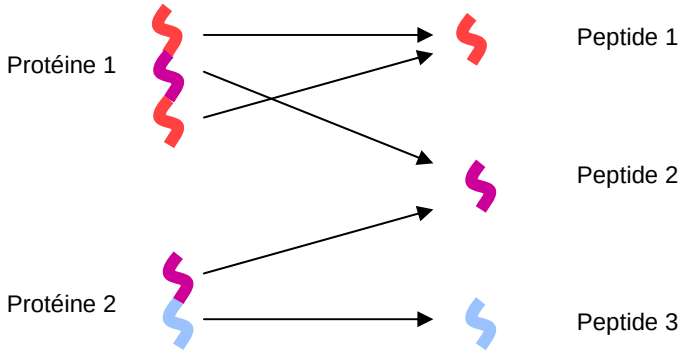


Figure 5 : cas possibles de création de peptides.
Un peptide peut apparaître plusieurs fois dans une protéine ou être issu de plusieurs protéines différentes.

Nous supposerons que nous ne perdons aucune protéine d'intérêt ou peptide d'intérêt dans ces étapes de préparation. La matrice de digestion permet donc de passer de la quantité des protéines en entrée de la chaîne d'analyse, à la quantité des peptides à la fin des étapes de préparation. Nous recherchons à reconstruire un profil moléculaire, nos variables d'intérêt sont donc les concentrations x_p des protéines d'intérêt p. Soit V le volume de l'échantillon biologique injecté dans le système. La quantité de protéine p à l'entrée du système est donc $x_p V$. La quantité η_i de peptide i à la sortie des étapes de préparation est donnée par

$$\boldsymbol{\eta} = V\boldsymbol{D}\boldsymbol{x} \text{ ou } \eta_i = V\sum_{p=1}^{P} d_{ip}x_p$$

avec $\boldsymbol{\eta} = [\eta_1 \quad \dots \quad \eta_I]^t$, $\boldsymbol{x} = [x_1 \quad \dots \quad x_P]^t$, où I et P représentent respectivement les nombres de peptides et de protéines considérés.

Le système prend en entrée un mélange de protéines mais toutes les étapes suivantes de la chaîne d'analyse manipulent des peptides.

2.2. Chromatographie liquide

Une colonne de chromatographie a pour but de séparer les constituants d'un mélange. Pour cela, elle utilise les différences d'affinité des constituants pour deux phases physiques différentes. Dans notre chaîne protéomique, son but est de séparer les peptides en utilisant leur hydrophobicité[1] caractéristique.

Ce sont les travaux du russe Tswett en 1900 qui marquent le début de la chromatographie. Son étude portait sur la chlorophylle et les différents pigments des plantes. Les spéculations sont toujours d'actualité sur l'origine du nom de la discipline : vient-elle du nom de l'auteur (Tswett signifie couleur en russe) ou de son sujet d'étude ? Quel que soit ses origines, ce terme regroupe aujourd'hui une grande famille de techniques de séparation engendrant quelques problèmes de nomenclature. L'organisme de standardisation IUPAC a donc proposé une classification de ces différentes méthodes dans [13]. Selon cette classification, nous utilisons dans notre système une colonne de chromatographie liquide à haute performance avec l'utilisation d'une phase inverse en mode gradient… quelques explications s'imposent.

En chromatographie liquide, le dispositif est le plus souvent constitué d'une colonne remplie de billes. La taille des billes conditionne les performances de séparation. Depuis les années 70, l'utilisation de billes dont la taille est de l'ordre de la dizaine de µm et les grandes performances qui en ont résulté a conduit à caractériser les colonnes comme étant à « haute performance ». Dans les laboratoires sur puces étudiés au DTBS, ces billes sont avantageusement remplacées par des structures

[1] L'hydrophobicité est la capacité d'une molécule à ne pas se mélanger avec l'eau. L'huile, par exemple, est formée de molécules très hydrophobes.

en forme de piliers que l'on grave à l'intérieur du silicium par des techniques de lithographie, mais cela ne change pas les principes du système. Colonne, billes et piliers constituent la phase stationnaire. Un solvant circule de façon continue dans la colonne, il constitue la phase mobile.

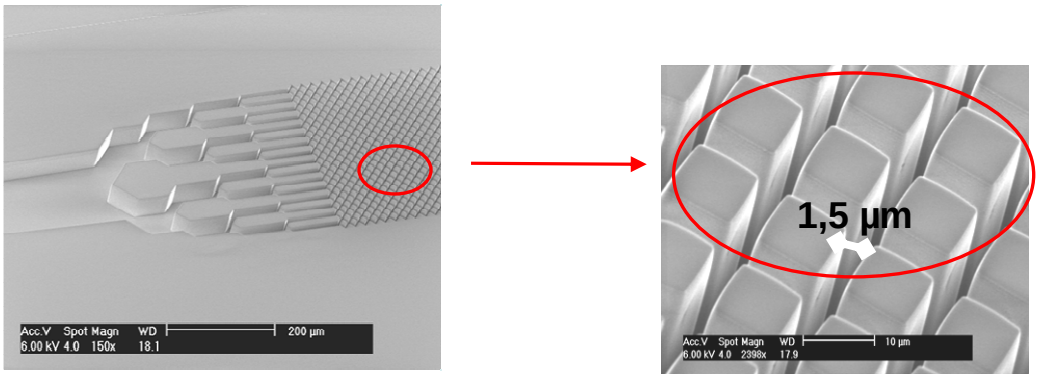


Figure 6 : colonne de chromatographie développée au DTBS.

Dans le cas où la phase mobile est plus polaire que la phase stationnaire, on parle de fonctionnement en phase inverse. Si la composition du solvant utilisé reste la même pendant toute la durée d'utilisation, il s'agit d'un fonctionnement en mode isocratique, si la composition du solvant varie de façon à changer sa polarité, on parle de mode gradient. Le mode gradient permet de séparer en une seule analyse un mélange constitué d'éléments peu hydrophobes et très hydrophobes.

Les peptides sont injectés en début de colonne, puis ils sont entraînés par la phase mobile. Durant son trajet dans la colonne, le peptide interagit de nombreuses fois avec la phase stationnaire. Lorsqu'un peptide se trouve dans la phase mobile, il est porté par le solvant et circule à sa vitesse. Lorsqu'il est fixé à la phase stationnaire, sa vitesse devient nulle. Au bout d'un certain temps, le peptide fixé sera relâché dans la phase mobile. Les phénomènes d'adsorption et de désorption qui contrôlent cette fixation produisent un ralentissement de la vitesse moyenne de propagation du peptide. Plus un peptide aura d'affinité avec la phase stationnaire, plus il sera retardé, plus son temps d'arrivée en sortie de colonne t_i sera grand (Figure 7).

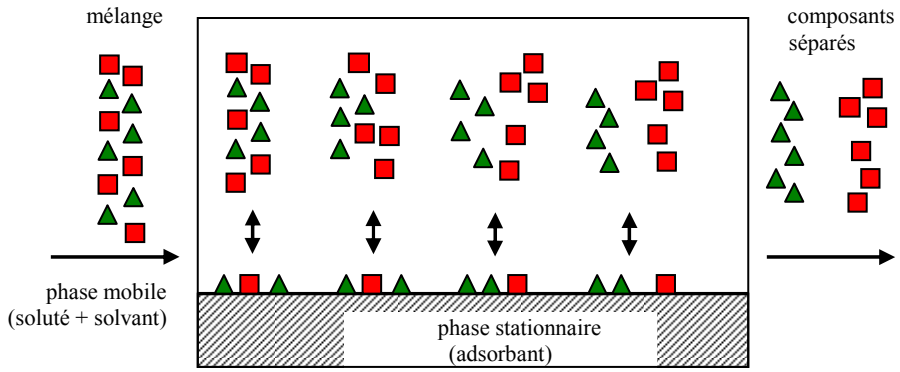


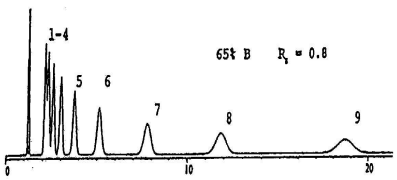
Figure 7 : principe de la chromatographie.

Intéressons nous à la quantité de peptide i sortant de la colonne entre l'instant t et l'instant $t+\mathrm{d}t$, le débit molaire de i, exprimé en mol/s est noté $c_i(t)$. Idéalement, tous les peptides identiques sortent de la colonne au même moment, appelé temps de rétention. Le débit molaire de i se modélise donc sous la forme d'une fonction de Dirac. Toute la matière injectée dans la colonne sort de la colonne, nous pouvons donc écrire

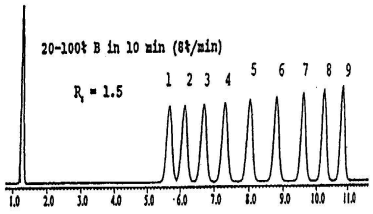
$$c_i(t) = \eta_i \delta(t - t_i)$$

avec η_i la quantité de peptide i injectée dans la colonne de chromatographie, t_i le temps de rétention du peptide i, δ la fonction de Dirac centrée en 0.

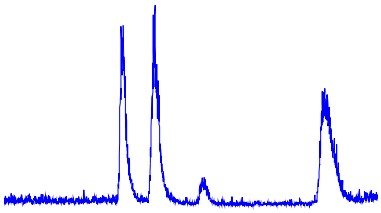
Pour une analyse plus précise du cas idéal, le lecteur pourra se référer au chapitre 2 de [14]. En pratique, plusieurs causes de dispersions interviennent et étalent un peu les temps d'arrivée des peptides créant un pic chromatographique d'une largeur non nulle. Nous pouvons citer l'influence de la diffusion, des effets de cinétique de réaction ou la dispersion de la vitesse du solvant due à la viscosité du liquide. Chacune de ces causes a été étudiée dans [15-17]. Deux stratégies duales existent pour modéliser ces phénomènes : proposer des équations différentielles locales et les résoudre ou directement proposer une modélisation de la réponse globale. Ainsi plusieurs modèles locaux basés sur l'équation de convection diffusion ont été proposés [14], ainsi que de nombreuses fonctions [18]. Cependant la fonction la plus généralement employée reste la fonction gaussienne [13].



Une analyse en mode isocratique. Figure extraite de [19].



Une analyse en mode gradient. Figure extraite de [19].



Un chromatogramme en mode gradient obtenu sur nos systèmes.

Figure 8 : comparaison des modes isocratique et gradient.
Un chromatogramme représente une somme pondérée des débits de chaque peptide à la sortie de la colonne de chromatographie.

Ces publications concernent de façon générale le mode isocratique où la composition du solvant ne change pas au cours de l'expérience. Dans notre chaîne d'analyse, nous utilisons le mode gradient. Dans ce mode d'analyse, la forme des pics change peu et l'hypothèse gaussienne est toujours employée. Par contre, la largeur des pics reste constante quelle que soit la molécule concernée [19, 20]. La Figure 8 compare des expériences réalisées en mode isocratique et gradient, et présente un chromatogramme issu de nos systèmes.

Finalement le modèle retenu est :

$$c_i(t) = \eta_i g_i(t) \text{ avec } g_i(t) = N(t; t_i, \gamma_c)$$

Avec η_i la quantité de peptide i injectée dans la colonne de chromatographie, t_i le temps de rétention du peptide i et γ_c la largeur des pics chromatographiques et $N(t\ ;\ \mu, \gamma)$ la fonction gaussienne centrée en μ et de largeur γ.

Si la largeur des pics est relativement stable, des variations des temps de rétention de l'ordre de la minute sont observées d'une expérience à l'autre [21] (Figure 9). Plusieurs raisons ont été avancées [22], mais ce phénomène n'est pas encore totalement compris ni contrôlé.

Dans cette section nous avons décrit le fonctionnement des colonnes de chromatographie. Nous avons modélisé les pics chromatographiques par des fonctions gaussiennes de même largeur. Cependant, nous avons également relevé que la position des pics n'est pas entièrement prévisible. La section suivante décrit le phénomène de nébulisation du liquide.

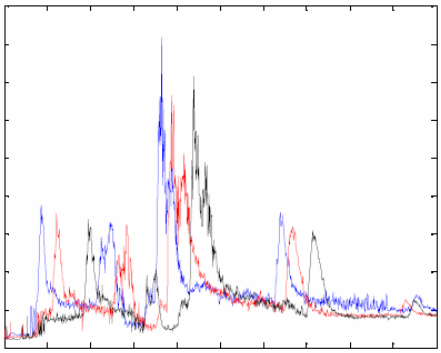


Figure 9 : répétition d'expériences chromatographiques.
Trois expériences sont réalisées dans des conditions similaires. La superposition de leurs chromatogrammes révèle des variations du temps de rétention de l'ordre de la minute.

2.3. Electrospray

L'electrospray effectue le couplage entre le spectromètre de masse et la colonne de chromatographie. Le spectromètre de masse ne manipule qu'un gaz d'ions. Pour être analysés, les peptides doivent donc être chargés et extraits de leur phase liquide.

L'electrospray est une technique d'ionisation. Cependant, cette appellation est trompeuse car les ions ne sont pas créés par l'electrospray, mais ils sont déjà présents en solution dans la colonne de chromatographie. Comme dans toute phase principalement constituée d'eau, nous retrouvons notamment une grande quantité d'ions H_3O^+ et OH^- et quelques sels (Na^+, Cl^-). Dans ce milieu, les peptides sont chargés. En effet, ils sont constitués d'acides aminés et ont donc au moins une fonction acide carboxylique (qui attire les charges négatives), et au moins une fonction amine (qui attire les charges positives). Ils peuvent donc selon les conditions chimiques être des ions positifs ou négatifs. Dans les expériences considérées, les solvants chromatographiques sont choisis de façon à obtenir majoritairement des peptides chargés positivement.

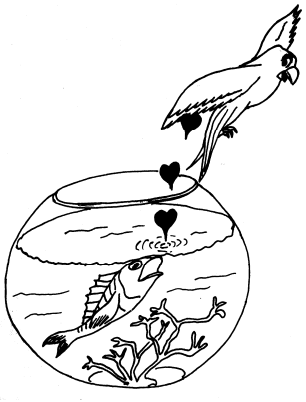


Figure 10 : métaphore de l'electrospray d'Arpino.
Arpino illustre la difficulté du changement de phase par l'image de l'amour impossible entre un poisson et un perroquet.
Figure extraite de [23].

L'electrospray a pour principal but d'ioniser les protéines. Pour cela, il sépare toutes les molécules sortant de la colonne de chromatographie. Cette séparation doit être extrêmement poussée : aucune microgouttelette d'eau, ou plus généralement, aucun agrégat de plusieurs molécules ne doit persister. En effet, la masse mesurée par le spectromètre de masse est alors celle de l'agrégat et non plus celle de la molécule seule. Sans cette séparation, les peptides ne pourraient être détectés et quantifiés par le spectromètre de masse.

Plusieurs modèles concernant l'electrospray ont été réalisés ces dernières années. Ces travaux sont synthétisés dans les documents [24-26]. Nous proposons dans cette section un modèle simplifié adapté au traitement des spectrogrammes.

2.3.1. Historique

Les racines de l'electrospray remontent aux premiers traités sur l'électromagnétisme. Une des expériences de Gilbert publiée en 1600 semblent avoir produit un electrospray [27]. Le phénomène fut plus précisément décrit au vingtième siècle et sera notamment utilisé comme technique de peinture. Dole propose de l'utiliser comme source d'ions pour la spectrométrie de masse dans les années 60. Dans les années 80, Fenn l'utilise pour analyser les molécules biologiques. Il reçu le prix Nobel de chimie en 2002 pour ces travaux. Dans les années 90, Wilm et Mann ont amélioré considérablement les performances du système en le miniaturisant : leur technique prendra le nom de nanospray. Plus de détails peuvent être trouvés dans la thèse de Bökman [24].

2.3.2. Le principe

2.3.2.1. Appareillage

Comme le montre la précocité de l'expérience de Gilbert, l'appareillage nécessaire pour produire un phénomène électrospray est minime. En effet, il suffit d'appliquer un fort champ électrique (quelques kilo Volts) entre un capillaire (par exemple la sortie de la colonne de chromatographie) et une contre électrode (l'entrée du spectromètre). Cette simplicité trompeuse cache la multiplicité et de la diversité des phénomènes multi-échelles qui soutiennent son principe.

2.3.2.2. Echelle macroscopique – Le cône de Taylor

La cohésion d'un liquide est assurée par les forces intermoléculaires comme les liaisons hydrogène. Au niveau macroscopique, on parle de tension de surface. Elle a tendance à maintenir le fluide dans un volume propre et à minimiser l'interface avec les autres phases. En l'absence d'autres forces, le liquide à la sortie de la colonne de chromatographie a une interface de forme sphérique.

L'augmentation de la tension modifie l'interface jusqu'à démarrer le phénomène electrospray. Soumis à un fort champ électrique, le liquide se déforme (Figure 11). Les charges électriques positives migrent vers un pôle, les charges négatives vers l'autre pôle. Le champ électrique tend ainsi à étirer la goutte. Le liquide est donc soumis aux influences antagonistes du champ électrique qui a tendance à l'étirer et de la tension de surface qui maintient sa cohésion. Avec l'augmentation du champ électrique, l'interface s'allonge de plus en plus jusqu'à avoir une forme d'amande, pour tendre vers un cône. Cette déformation fait suite à l'accumulation d'ions de même charge due à l'influence du champ électrique. Cette déformation continue jusqu'au point où les forces de cohésion ne suffisent plus. Dès lors, des micro-gouttelettes sont projetées depuis la pointe du cône sous forme d'un spray.

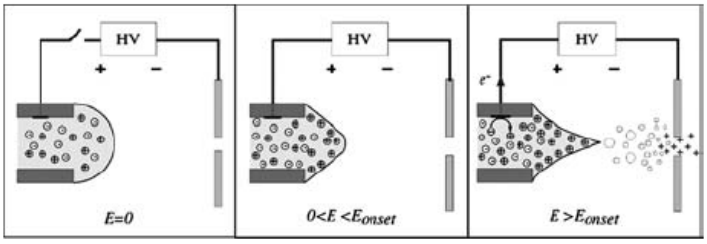


Figure 11 : formation du cône de Taylor.
Figure extraite de [28].

Si le fluide est mis en mouvement, l'étude de ces cônes et de leur comportement est du domaine de l'électro-hydrodynamique (EHD). La résolution de ces équations est complexe et peut donner des solutions très différentes en fonction des différents paramètres impliqués [29]. Coupleau et ses

successeurs ont proposé une classification des différents modes de l'electrospray suivant leurs différents comportements morphologiques (Figure 12).

Le nombre d'ions créés dépendra fortement du comportement de ce spray. Pour nos expériences, les paramètres sont réglés de façon à se placer dans le mode stable conejet, permettant d'avoir un comportement constant au cours du temps. Maintenir ce mode est d'autant plus difficile que le mode gradient chromatographique change non seulement la polarité du solvant, mais influe également sur ses propriétés électriques et donc sur la stabilité du jet.

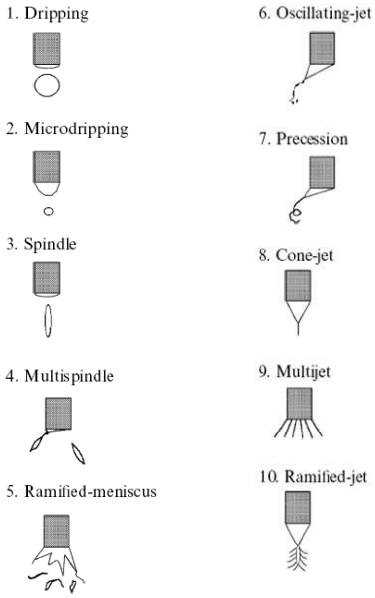


Figure 12 : modes possibles de l'electrospray
Le comportement de l'electrospray peut être stable (8 Cone-jet mode) ou instable. Figure extraite de [29].

2.3.2.3. Echelle microscopique – Comportement des gouttes

Les gouttes formées par l'electrospray comportent des charges positives et négatives. Le cône de Taylor est dû à un excès de charges près de l'interface. Dans nos expériences, c'est un excès de charges positives. Les gouttes formées à partir de l'interface ont elles aussi un excès de charges positives. Elles sont donc soumises à la force électrique et se déplacent vers la contre électrode du spectromètre de masse.

Durant son vol, la goutte ne reste pas statique. Tout d'abord la goutte subit une évaporation. Les molécules présentes à sa surface, chargées ou non, passent en phase gazeuse, et la goutte perd de son volume. Parallèlement, deux phénomènes décrits ci-dessous tendent à fractionner les gouttes.

- *Les explosions coulombiques.* Les charges positives en excès se repartissent à la surface de la goutte de façon à minimiser leur énergie. L'évaporation entraîne une perte de volume de la goutte. Si le volume est trop faible pour que les charges de même signe soient suffisamment éloignées, on atteint la limite de Rayleigh. La goutte se scinde alors en plusieurs gouttes filles.
- *Les cônes de Taylor*. Soumis à leur influence propre, l'excès d'ions positifs a tendance à se répartir sur toute la surface de la goutte. Mais ils sont sous l'influence du champ électrique, donc une population plus importante de charges positives a tendance à se concentrer sur la partie de la goutte faisant face à la contre-électrode. L'interface réagit à cet excès de charges positives, en se déformant jusqu'à former un cône de Taylor à l'échelle de la goutte.

La différence entre ces deux phénomènes est minime. Les deux sont dus à une densité de charges électriques importante et ils amènent à la fission de la gouttelette. Cependant, le premier à une influence globale sur la goutte et le second concernera principalement sa partie superficielle. De plus, ces phénomènes peuvent être concurrents : le phénomène des cônes de Taylor extrait peu de masse de la goutte principale, mais un grand nombre de ses charges. Le cône de Taylor peut ainsi empêcher la

goutte d'atteindre la limite de Rayleigh pendant le temps de son vol. Les discussions sont encore importantes pour savoir quel phénomène est prépondérant.

2.3.2.4. Echelle moléculaire – passage des peptides en phase gazeuse

Si deux théories de fractionnement des gouttes existent, deux théories de séparation des ions et des gouttes existent également : la théorie de l'évaporation des ions (ion evaporation mecanism, IEM) et la théorie du fractionnement jusqu'à l'ion unique (charged residue mechanism , CRM). La première indique que les ions sont formés par évaporation, la seconde que les ions seront le résultat ultime d'une succession de fissions, qu'elles aient eu lieues à la suite d'explosions coulombiques ou de cônes de Taylor. Aucun consensus ne semble avoir été trouvé pour savoir quel est le phénomène prépondérant, en particulier pour les peptides.

Durant le temps de vol, toutes les molécules formant la goutte ne semblent pas avoir le temps de passer en phase gazeuse. Ainsi quelle que soit la théorie de formation des ions prépondérante, la position des peptides dans la partie la plus externe de la goutte semble être un avantage déterminant pour pouvoir être analysé par le spectromètre de masse. Les paramètres chimiques des peptides rentrent donc en jeu pour modéliser l'influence de l'electrospray.

En effet, seul l'excès de charges positives se répartit à la surface, et un nombre important de charges reste confiné à l'intérieur de la goutte avec les charges négatives. Tous les ions positifs sont alors en compétition pour accéder à la surface de la goutte. Cette dernière peut être considérée comme une phase au même titre que l'intérieur de la goutte et que l'extérieur de la goutte. Le temps de migration à l'intérieur de la goutte étant très inférieur au temps entre deux fissions, un équilibre entre les diverses réactions de changement de phase a le temps de se réaliser.

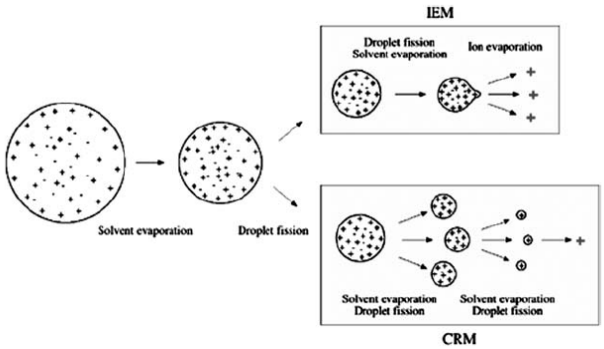


Figure 13 : deux théories concurrentes pour la formation des ions.
L'ion est formé soit par évaporation (IEM), soit il est le résultat ultime d'une série de fragmentations. Figure extraite de [28].

2.3.3. Modélisation

Nous considérons la relation liant $\varepsilon_i(t)$, le débit de peptide i à l'entrée du spectromètre, et $c_i(t)$, son débit à la sortie de la colonne de chromatographie. Si des phénomènes retardant existent pendant l'electrospray, ils sont négligeables par rapport à l'influence de la colonne de chromatographie. Nous ne modélisons donc que les pertes. L'electrospray agit donc comme un gain multiplicatif entre les deux fonctions citées ci-dessus.

Nous pouvons décomposer ce gain entre les différents phénomènes en jeu dans l'electrospray :

- P_1, proportion du liquide nébulisé par un cône de Taylor (formation des gouttelettes),
- P_2, proportion des gouttelettes présentes sur une trajectoire qui amènera les peptides à l'entrée du spectromètre et non pas sur les électrodes,
- P_3, proportion des peptides ionisés dans les gouttelettes.

Nous supposerons que le spray est en fonctionnement « stable conejet » pendant toute la durée de l'expérience, que tout le liquide sera nébulisé par le cône de Taylor (P_1 = 1) et que toutes les gouttelettes ont la même taille.

La deuxième proportion dépend du comportement macroscopique du spray qui dépend des caractéristiques physiques de l'appareil (débits, solvants employés, dimensions, tension utilisée). Ils ne dépendent pas des concentrations des peptides ni de leur nature. Ainsi ces proportions pourront varier au court du temps puisque nous feront varier la composition des solvants au court du temps, mais cette composition variera lentement au court de l'expérience. Le gain variera également de façon lente.

La troisième proportion dépendra des caractéristiques chimiques des ions à l'intérieur de la goutte, et dépendra donc des différentes concentrations des peptides. On suppose également que le gain du peptide *i* dépendra majoritairement de la relation entre le peptide *i*, et les ions du solvant, leur concentration et leur influence varie de façon très lente.

L'electrospray pourra être modélisé de la façon suivante

$$\varepsilon_i(t) = P_1 P_2(t) P_3^i(t) c_i(t)$$

De plus, cette relation peut se simplifier. Le signal $c_i(t)$ provenant de la colonne de chromatographie a été modélisé par une gaussienne centrée en t_i et de largeur γ_c et dont l'amplitude est proportionnelle à la concentration x_i du peptide *i*

$$c_i(t) = \eta_i g_i(t) \text{ avec } g_i(t) = \left(2\pi\gamma_c^{-1}\right)^{-1/2} \exp\left(-0.5\gamma_c (t - t_i)^2\right)$$

L'écart type de ce signal est de l'ordre de la minute, soit bien inférieure à l'échelle caractéristique du changement de composition en mode gradient. Durant la durée du signal chromatographique, les proportions étudiées varient peu. Nous pouvons donc remplacer ces proportions par un gain K_i qui dépend du peptide considéré, mais ne dépend pas du temps.

Le modèle retenu sera donc

$$\begin{aligned} \varepsilon_i(t) &= P_1 P_2(t) P_3^i(t) c_i(t) \\ &\approx P_2(t_i) P_3^i(t_i) \eta_i g_i(t) \\ &= K_i \eta_i g_i(t) \end{aligned}$$

avec $K_i = P_2(t_i) P_3^i(t_i)$.

Nous avons donc modélisé dans cette section l'electrospray qui assure le couplage entre la chromatographie et la spectrométrie de masse. Il nous reste à modéliser la dernière étape de la chaîne d'analyse lorsque le flux d'ion séparé est converti en données numériques.

2.4. Spectrométrie de masse

La spectrométrie de masse est une méthode d'analyse permettant d'identifier et de quantifier les espèces d'un mélange d'ions. Un spectromètre de masse prend en entrée une certaine quantité de matière et fournit en sortie un spectre donnant la quantité de molécules en fonction de leur rapport masse sur charge.

Un spectromètre de masse est constitué de trois éléments.

- La source, qui génère des ions.
- L'analyseur, qui sépare les différents ions du mélange, en leur donnant une trajectoire spécifique.
- Le détecteur, qui convertit les ions en signal mesurable.

Dans cette section, nous ne traitons que de l'analyseur, assimilant analyseur et spectromètre. Les autres éléments seront traités dans des parties dédiées.

Le spectromètre de masse utilisé dans notre chaîne de mesure est le LTQ développé par la société Thermo. L'analyseur qui le compose est une trappe linéaire. Il s'agit d'un analyseur relativement récent dont le concept de base a été breveté en 1995 [30], puis en 2003 en ce qui concerne la méthode d'éjection radiale [31, 32]. Une présentation synthétique de l'instrumentation et de ses modes de fonctionnement a été proposée par l'équipe de Douglas [33]. Dans ce chapitre, nous compléterons cet article par un modèle adapté au traitement du signal. Nous mettrons notamment en évidence la relation entrée-sortie ainsi que les sources de perturbation d'une trappe linéaire.

La trappe linéaire emprunte sa géométrie à l'analyseur quadripolaire, mais elle s'utilise comme une trappe à ion de Paul. Les différences entre ces appareils et la trappe linéaire sont faibles. Nous utiliserons pour réaliser notre modèle les nombreux travaux portant sur ces appareils proches et nous les étendrons à notre cas. Dans le paragraphe 2.4.2, nous présentons une synthèse bibliographique de certains de ces articles [34-38] qui nous permettra dans le paragraphe 2.4.3 de modéliser la trappe linéaire par un retard et d'une série d'imperfections. Mais tout d'abord, replaçons la trappe linéaire dans la famille des analyseurs disponibles.

2.4.1. Analyseurs existants

L'analyseur sépare les ions en utilisant la force électromagnétique. Elle induit des trajectoires spécifiques pour chaque ion qui dépendent du rapport entre leur masse et leur charge. Les principaux analyseurs existants sont les suivants.

- *Les spectromètres à déflection magnétique*. Les ions sont accélérés puis soumis à un champ magnétique orthogonal qui dévie leur trajectoire selon leur masse et leur charge. C'est ce principe qui a été utilisé pour les premiers spectromètres de masse [39, 40].
- *Le spectromètre de masse à temps de vol*. Les ions sont soumis à un champ électrique uniforme dans une chambre d'accélération. Le temps de parcours de cette chambre dépendra de leur masse et de leur charge [41].
- *Les spectromètres de masse à résonance cyclotronique ionique et à transformée de Fourier*. Les ions sont soumis à un fort champ magnétique. Leur fréquence de rotation autour des lignes de champ dépend de leur rapport masse sur charge. Les mouvements des ions sont détectés et l'on remonte à leur fréquence de rotation par transformée de Fourier. Plus le temps d'acquisition est important, plus la résolution est importante rendant sa résolution inégalée. Mais sa cadence d'analyse (nombre de spectres à la seconde) est limitée.
- *Les quadripôles et les trappes ioniques*. Ces deux appareils sont basés sur des technologies assez proches. Certains ions sont confinés par un champ électrique suivant leur rapport masse sur charge et les tensions appliquées. En jouant sur ces tensions, un spectre de masse peut être obtenu. Nous détaillerons leur principe de fonctionnement au paragraphe 2.4.2.

2.4.2. Les pièges ioniques

2.4.2.1. Historique

En 1950, l'équipe de Wolfgang Paul invente deux appareils pouvant servir de spectromètre de masse : l'analyseur quadripolaire et la trappe ionique. Alors que le premier appareil est très vite utilisé en chimie analytique pour obtenir des spectres de masse, le second est surtout utilisé par la communauté des physiciens, notamment par Hans Dehmelt. Ces travaux furent couronnés par le prix Nobel de physique en 1989 pour leurs travaux sur le « piégeage » des particules atomiques.

La trappe ionique fut tout d'abord considérée comme aussi performante que l'analyseur quadripolaire, mais de nombreuses améliorations lui ont été apportées dans la deuxième moitié du vingtième siècle, notamment grâce à l'équipe de Cooks (Purdue university). Ces améliorations se traduisent par divers modes d'utilisation qui confèrent à l'appareil une excellente adaptabilité tout en gardant un coût très modeste par rapport aux autres systèmes. Dernièrement, les trappes ioniques

linéaires ont été développées. Les ions sont confinés dans un volume plus grand ce qui diminue l'influence perturbatrice des ions entre eux.

Grâce à leur sensibilité, les quadripôles reviennent actuellement sur le devant de la scène pour les applications quantitatives, notamment dans le mode dit MRM (Multiple Reaction Monitoring). Dans ce mode, les quadripôles visent un ion bien particulier et ne fonctionnent pas en balayant une gamme de masse ce qui permet d'augmenter leur sensibilité.

2.4.2.2. Principe

Les quadripôles et les trappes ioniques jouent sur la stabilité d'un piège utilisant la force électrique afin d'obtenir un spectre de masse. Nous verrons dans cette section sur quel principe repose ce confinement. Puis nous étudierons certaines géométries permettant de construire un tel piège ainsi que les conditions de stabilité des ions dans ces géométries.

Dans un spectromètre de masse, on doit absolument éviter que les ions se neutralisent, car seules des particules chargées sont manipulables par des forces électromagnétiques. Tout d'abord, un vide poussé est créé dans l'enceinte du spectromètre de masse pour éviter les collisions avec les molécules d'air. De plus, les spectromètres sont conçus de manière à confiner les ions loin des parois.

On souhaite créer une force électrique de rappel de type ressort dont la raideur est proportionnelle à la charge de l'ion et à la tension appliquée aux électrodes Φ_0. Cette force aura tendance à maintenir les ions au centre de l'appareil.

On va créer une telle force à l'aide d'un potentiel électrique parabolique de la forme

$$\Phi(x,y,z)=\frac{\Phi_0}{r_0^2}\left(\lambda x^2+\sigma y^2+\gamma z^2\right)$$

Φ : potentiel électrique

r_0 : distance caractéristique de la trappe (Figure 14)

Φ_0 : valeur absolue de la tension appliquée sur les électrodes (Figure 14)

La condition de Laplace $\Delta\Phi = 0$, impose la condition $\lambda+\sigma+\gamma=0$. Elle empêche d'avoir un potentiel attractif dans toutes les directions de l'espace en même temps. Cet inconvénient est contourné en appliquant une tension Φ_0 alternative aux électrodes, ce qui permet d'inverser le rôle attractif et répulsif de chaque direction de l'espace. L'inertie de l'ion évite jusqu'à un certain point que les trajectoires des ions divergent pendant la phase répulsive.

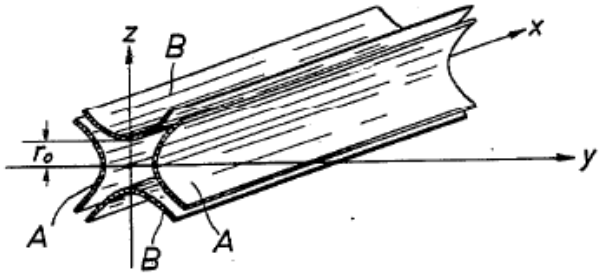


Figure 14 : quadripôle
Cette figure est extraite du brevet de Paul [42]. Les ions sélectionnés parcourent le dispositif en restant au centre des électrodes notées A et B. Les ions rejetés sont éjectés et neutralisés sur les électrodes. On applique la tension Φ_0 à l'électrode A, $-\Phi_0$ à l'électrode B.

Nous allons maintenant étudier quelles géométries d'électrodes permet d'avoir un tel potentiel. Il y a deux solutions évidentes pour satisfaire la condition de Laplace.

- Si $\lambda = -\sigma = 1$ et $\gamma = 0$ (invariance par translation selon z), il s'agit d'un analyseur quadripolaire, ou d'une trappe ionique linéaire (Figure 14).
- Si les électrodes sont construites afin d'avoir $\lambda = \sigma = 1$ et $\gamma = -2$ (géométrie cylindrique, invariance par rotation autour de z), il s'agit d'une trappe ionique de Paul (Figure 15).

La forme des électrodes produisant un tel potentiel est donnée par les équipotentielles [33]. En pratique, les électrodes ont des formes paraboliques ou hyperboliques.

Nous avons défini les formes d'électrodes et le potentiel électrique résultant pour chaque analyseur étudié (quadripôle, trappe linéaire et trappe de Paul). Etudions le comportement des ions dans le champ électrique résultant.

Les trois électrodes de la trappe ionique

Les trois électrodes montées (vue en coupe)

Figure 15 : électrodes de la trappe ionique de Paul.
Figures extraites de [35].

La force électrique qui dérive de ce potentiel est égale à $\vec{F}_e = -e.\vec{\text{grad}}\,\Phi$, où e est la charge de l'ion

$$\vec{F}_e = -\frac{2e\Phi_0}{r_0^2}\begin{pmatrix}\lambda x\\ \sigma y\\ \gamma z\end{pmatrix}$$

Nous avons vu que λ, σ et γ ne peuvent pas tous être positifs et qu'il nous faut donc appliquer aux électrodes une tension alternative pour pouvoir confiner les ions :

$$\Phi_0(t) = U + V\cos\Omega t$$

U : amplitude de la composante continue de la tension

V : amplitude de la composante alternative de la tension

Ω : fréquence de la composante alternative

Appliquons le principe fondamental de la dynamique, nous obtenons le système d'équations

$$\begin{cases} m\dfrac{\mathrm{d}^2x}{\mathrm{d}t^2} = \dfrac{-2e\lambda}{r_0^2}\left(U + V\cos(\Omega t)\right)x \\ m\dfrac{\mathrm{d}^2y}{\mathrm{d}t^2} = \dfrac{-2e\sigma}{r_0^2}\left(U + V\cos(\Omega t)\right)y \\ m\dfrac{\mathrm{d}^2z}{\mathrm{d}t^2} = \dfrac{-2e\gamma}{r_0^2}\left(U + V\cos(\Omega t)\right)z \end{cases} \qquad (2.4\text{-}1)$$

En 1868, Mathieu étudia les équations différentielles de forme similaire [43, 44]

$$\frac{\mathrm{d}^2u}{\mathrm{d}\xi^2} + \left(a_u + 2q_u\cos(2\xi)\right)u = 0$$

où a_u *et* q_u sont les paramètres caractéristiques de l'équation. Sur un domaine infini $\left(\xi \in R\right)$, cette équation différentielle admet une solution de la forme

$$u(\xi) = A\sum_{n=-\infty}^{+\infty} C_{2n}\cos\left[(2n+\beta_u)\xi\right] + B\sum_{n=-\infty}^{+\infty} C_{2n}\sin\left[(2n+\beta_u)\xi\right] \qquad (2.4\text{-}2)$$

A et B sont des constantes dépendant des conditions initiales, C_{2n} et β_u sont des constantes dépendant de a_u *et* q_u.

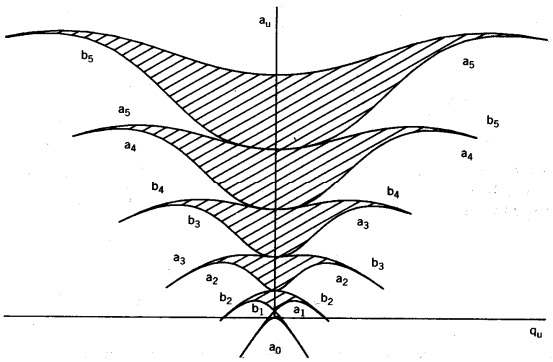


Figure 16 : diagramme de stabilité des équations de Mathieu.
Suivant les valeurs de a_u et q_u, les solutions sont stables ou instables quelles que soient les conditions initiales. Les zones stables sont hachurées, le reste représente les zones instables. Figure extraite de [36].

Ces sommes infinies peuvent conduire à des solutions stables, c'est-à-dire bornées ou conduire à des solutions instables. Suivant les valeurs de a_u *et* q_u, on peut construire un diagramme de stabilité (Figure 16).

L'équation du principe fondamental de la dynamique (2.4-1) peut être écrite sous la forme de l'équation de Mathieu ci-dessus en choisissant les valeurs du Tableau 1 pour paramètres caractéristiques.

Pour que l'ion ait une trajectoire stable, il faut que les 3 équations différentielles de son mouvement aient des solutions stables. Nous pouvons faire en sorte de tracer les 3 diagrammes de stabilité sur la même figure (Figure 17 à Figure 19) à l'aide des changements de variable appropriés décrits ci-dessous.

$a_x = \dfrac{8e\lambda U}{mr_0^2\Omega^2}$	$q_x = \dfrac{4e\lambda V}{mr_0^2\Omega^2}$
$a_y = \dfrac{8e\sigma U}{mr_0^2\Omega^2}$	$q_y = \dfrac{4e\sigma V}{mr_0^2\Omega^2}$
$a_z = \dfrac{8e\gamma U}{mr_0^2\Omega^2}$	$q_z = \dfrac{4e\gamma V}{mr_0^2\Omega^2}$

Tableau 1 : définitions des coordonnées de Mathieu pour chaque axe.

Le coefficient $4/\Omega^2$ *apparaît à la suite du changement de variable* $2\xi = \Omega t \Rightarrow 2\mathrm{d}\xi = \Omega\,\mathrm{d}t \Rightarrow 4\mathrm{d}\xi^2 = \Omega^2\,\mathrm{d}t^2$.

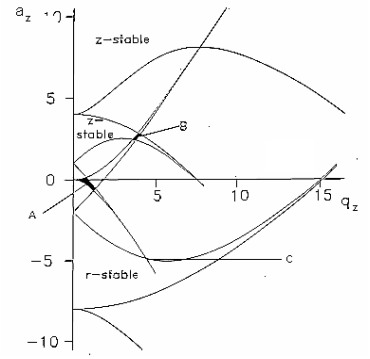
Considérons tout d'abord le cas de la trappe à ions de Paul. Dans ce cas là, la géométrie est invariante par rotation autour de l'axe z ($\lambda = \sigma = 1$). Les diagrammes de la direction x et y sont donc identiques. Dans ce cas là on préfère d'ailleurs utiliser les coordonnées cylindriques (r,θ,z) et ne représenter que les directions r et z. On a donc les changements de variables

$$\begin{cases}(q_x, a_x) \to (q_r, a_r)\\ (q_y, a_y) \to (q_r, a_r)\end{cases} \text{ avec } \begin{cases}a_r = a_x = a_y\\ q_r = q_x = q_y\end{cases}$$

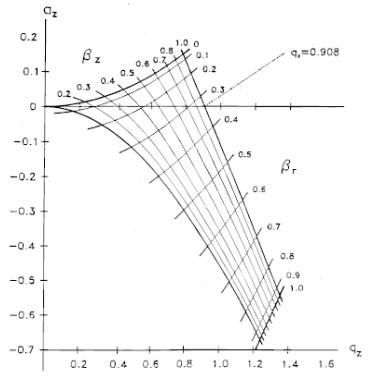
De plus, pour satisfaire la condition de Laplace, γ = -2. On peut donc déduire du Tableau 1 le changement de variable suivant

$$(q_r, a_r) \rightarrow (q_z, a_z) \text{ avec } \begin{cases} a_z = -2a_r \\ q_z = -2q_r \end{cases}$$

Le diagramme de stabilité de la direction *r* peut donc être représenté en coordonnées (q_z, a_z) après une homothétie centrée à l'origine de rapport -2. Si nous superposons le diagramme de la direction *r* et *z* nous obtenons la Figure 17. Les zones où les ions auront des trajectoires stables seront situées aux endroits où les deux diagrammes se superposent (indiqués A, B et C sur la figure). Chacune de ces zones peut être utilisée pour effectuer un spectre de masse, mais seule la zone A est utilisée en pratique.



Superposition des diagrammes dans les directions *r* et *z*. Les zones où les ions ont des trajectoires stables sont indiquées par les lettres A, B et C.



Agrandissement de la zone A, majoritairement utilisée en spectrométrie de masse. Les coordonnées (q_z, a_z) dépendent de la tension appliquée à la trappe, de la masse et de la charge de l'ion. Les coordonées (β_r, β_z) décrivent les fréquences propres du mouvement des ions,

Figure 17 : diagramme de stabilité de la trappe à ions de Paul
Issu de [35].

Un diagramme similaire pourra être construit pour la géométrie des quadripôles et les trappes linéaires. Dans ce cas, il y a invariance par translation suivant l'axe z, γ = 0. La force électrique est donc nulle suivant cet axe, il n'y a pas de diagramme de stabilité associé. On a de plus λ = 1 et $\sigma = -1$. Le diagramme de stabilité de la direction *y* peut être représenté suivant les coordonnées (q_x, a_x) après le changement de variable déduit du Tableau 1 suivant

$$(q_y, a_y) \rightarrow (q_x, a_x) \text{ avec } \begin{cases} a_x = -a_y \\ q_x = -q_y \end{cases}$$

soit après une symétrie centrée à l'origine. Le diagramme résultant de la superposition des diagrammes de toutes les directions est représenté Figure 18. La zone d'intérêt principale est représentée Figure 19.

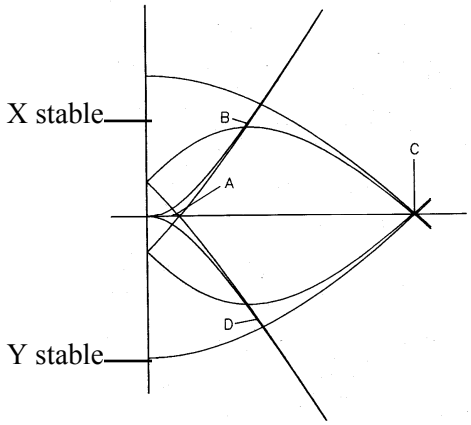


Figure 18 : diagramme des quadripôles et des trappes linéaires.
Ce diagramme est formé de l'intersection des diagrammes de chaque coordonnée. Les seules zones stables sont les zones A, B, C, D. Le diagramme est symétrique suivant les axes a_x et q_x. Seule la partie concernant les q_x positifs est représentée ici. Figure extraite de [36].

Sur ces diagrammes, chaque ion est représenté par un point dont les coordonnées dépendent de sa masse et de sa charge mais également de la tension appliquée (Tableau 1). Leurs abscisses augmenteront linéairement avec V, leurs ordonnées avec U. De plus, tous les points représentant les ions sont situés sur la droite d'équation

$$a_u = 2\frac{U}{V}q_u$$

Les coordonnées β_u représentées sur la Figure 17 et la Figure 19 permettent de calculer les fréquences propres du mouvement dans la direction u comme l'indique l'équation (2.4-2). Elles ne dépendent que de a_u et q_u. Le changement de variable est donné par la formule :

$$\beta_u^2 = a_u + \frac{q_u^2}{(\beta_u+2)^2 - a_u - \dfrac{q_u^2}{(\beta_u+4)^2 - a_u - \dfrac{q_u^2}{(\beta_u+6)^2 - a_u - \cdots}}} + \frac{q_u^2}{(\beta_u-2)^2 - a_u - \dfrac{q_u^2}{(\beta_u-4)^2 - a_u - \dfrac{q_u^2}{(\beta_u-6)^2 - a_u - \cdots}}}$$

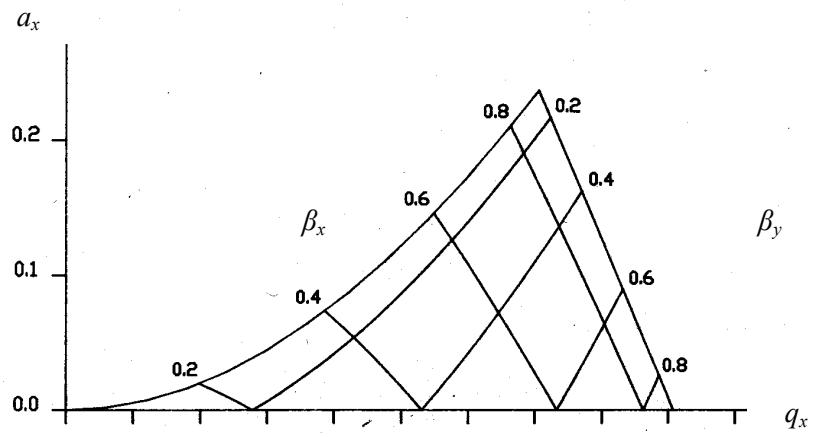


Figure 19 : zone principale du diagramme de stabilité des quadripôles et des trappes linéaires.
Il s'agit de la zone A présentée à la Figure 18. Seule la partie a_x et q_x positifs est représentée ici. Figure extraite de [36].

Nous avons présenté le lien entre la géométrie et le diagramme de stabilité. Nous allons voir dans la section suivante comment, à travers les modes opératoires, ces diagrammes sont exploités de façon à obtenir des spectres de masse.

2.4.2.3. Modes opératoires

La trappe ionique et l'analyseur quadripolaire permettent de confiner ou d'éjecter des ions en fonction de leur position sur le diagramme de stabilité. Celle-ci ne dépend que de la masse de l'ion, de sa charge, ainsi que de la tension appliquée aux électrodes. Dès lors, plusieurs modes opératoires peuvent être envisagés pour utiliser un piège ionique comme un spectromètre de masse. Ces modes jouent sur la tension appliquée et nécessitent une position des détecteurs spécifique.

Confinement sélectif en masse (mass-selective stability scan)

On choisit le rapport *U/V* de manière à ce que la droite portant les ions dans le diagramme de stabilité coupe un des bords de la zone de stabilité (Figure 20). Ainsi seul un intervalle de masse sur charge extrêmement réduit sera dans la zone stable du diagramme. Cet intervalle peut être choisi en réglant la valeur de U tout en maintenant le rapport *U/V* constant. Seul l'ion inclus dans cet intervalle restera confiné. Les autres seront éjectés vers les électrodes.

C'est la technique utilisée dans les analyseurs quadripolaires (Figure 14 page 28). Le mélange d'ions est injecté d'un côté de l'analyseur, un détecteur est placé de l'autre côté. Seul l'ion inclus dans l'intervalle de stabilité pourra atteindre le détecteur. Si l'injection du mélange est continue, un spectre de masse peut être réalisé en faisant en sorte que la faible zone de stabilité balaye toute la gamme de masse. Ceci peut être réalisé en balayant les tensions *U* et *V*, le rapport *U/V* restant constant (la droite portant les ions ne bouge pas pendant l'analyse).

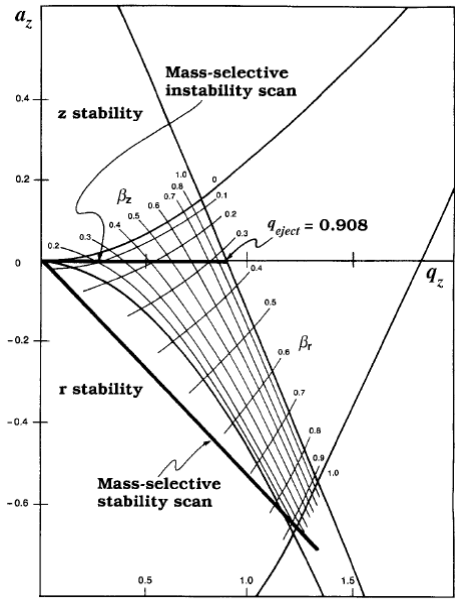


Figure 20 : modes d'utilisation en spectrométrie de masse.
Les droites en gras donnent les positions possibles des points représentant les ions. Leur position dépend de la tension appliquée aux électrodes. Les deux droites tracées représentent un mode opératoire pour un piège à ion (section 2.4.2.3). Figure extraite de [34].

Ejection sélective en masse (mass-selective instability scan)

La tension *U* est réglée à zéro. La droite portant les ions se confond donc avec l'axe des abscisses. *V* et Ω sont réglés de façon à ce que la totalité des ions d'intérêt soit située dans la zone stable. Une quantité déterminée d'un mélange d'ions est injectée au centre de la trappe. Une fente est positionnée dans les électrodes de façon à y positionner un détecteur (Figure 21 et Figure 22). La tension V est augmentée progressivement. Les ions se déplacent vers la droite du diagramme de stabilité jusqu'à atteindre la limite du diagramme de stabilité au point $q_E \approx 0.908$. Leur trajectoire devient instable et ils viennent frapper les électrodes. Une partie atteindra le détecteur en traversant les fentes pratiquées. Tout se passe comme si la limite d'instabilité q_E balayait la gamme de masse.

C'est généralement le mode principal des trappes de Paul ainsi que de notre trappe linéaire.

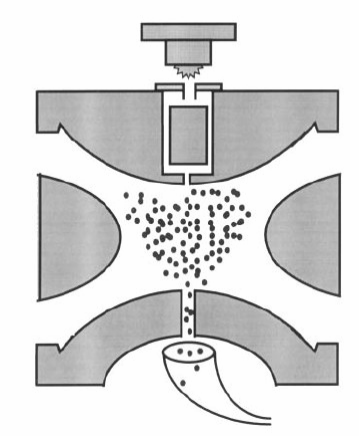


Figure 21 : disposition du détecteur dans une trappe de Paul.
Les ions sont confinés au centre de la trappe. Lorsque leur trajectoire devient instable certains vont frapper les parois, d'autres atteindre le détecteur par la fente pratiquée, ici un multiplicateur d'électrons (cornet en bas du schéma). Figure extraite de [35].

Absorption résonante

Cette technique est souvent employée en complément de la méthode précédente pour accélérer la sortie des ions atteignant q_E. Elle peut également être employée pour isoler un ensemble d'ions en provoquant l'éjection des autres.

On excite les ions que l'on veut éjecter du piège avec un champ électrique extérieur supplémentaire dont la fréquence correspond aux fréquences propres du mouvement des ions choisis. Nous rappelons que les fréquences propres sont indiquées dans les diagrammes par les coordonnées β_u. Tant que les ions sont excités de cette façon, l'amplitude de leur mouvement augmente jusqu'à aller frapper le détecteur.

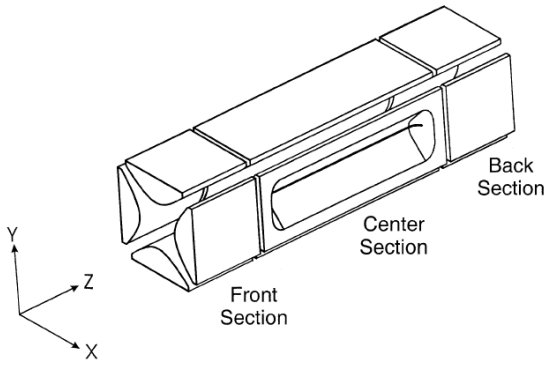


Figure 22 : trappe linéaire.
La section centrale est un quadripôle. On distingue la fente d'où sortiront les ions éjectés pour aller frapper le détecteur. Les sections avant et arrière sont alimentées de façon à maintenir les ions dans la partie centrale. Figure extraite de [31].

Mode MS^n

Ce mode est essentiel pour identifier la formule peptidique des protéines ou des peptides car plusieurs peptides différents peuvent avoir la même masse, par exemple en permutant les acides aminés dans la séquence. L'identification de l'ion peut être affinée par analyse de ses fragments. L'ion est successivement isolé, fragmenté par collision avec de l'hélium, puis ses fragments subissent une analyse de masse. Les fragments peuvent également être fragmentés et analysés, puis ainsi de suite. Le mode MS^n indique que nous réalisons n analyses de masse successives sur des fragments de plus en plus petits. Ces étapes sont réalisées par un enchaînement des techniques expliquées dans les paragraphes précédents.

2.4.3. Modélisation du LTQ de Thermo Scientific

Le LTQ est un spectromètre de masse dont le schéma général est présenté ci-dessous à la Figure 23. Il est employé dans le mode éjection sélective en masse en conjonction avec une absorption résonante en q_E. Le mode MS^n est possible avec cet appareil, mais il ne sera pas employé dans les expériences effectuées dans le cadre de cette thèse.

Le spectromètre de masse analyse le flux d'ions provenant de l'electrospray. Pour cela, il alterne deux phases, une phase de stockage où la trappe se remplit d'ions, suivie d'une phase de « lecture » où le contenu de la trappe est déterminé, puis il recommence un cycle. Ainsi il réalise « l'échantillonnage » du flux de l'electrospray à la période T_e^c.

La première phase joue un rôle d'intégrateur afin d'augmenter le signal sur le détecteur, la seconde utilise le principe de la trappe linéaire pour éjecter séquentiellement les ions en fonction de leur rapport masse sur charge.

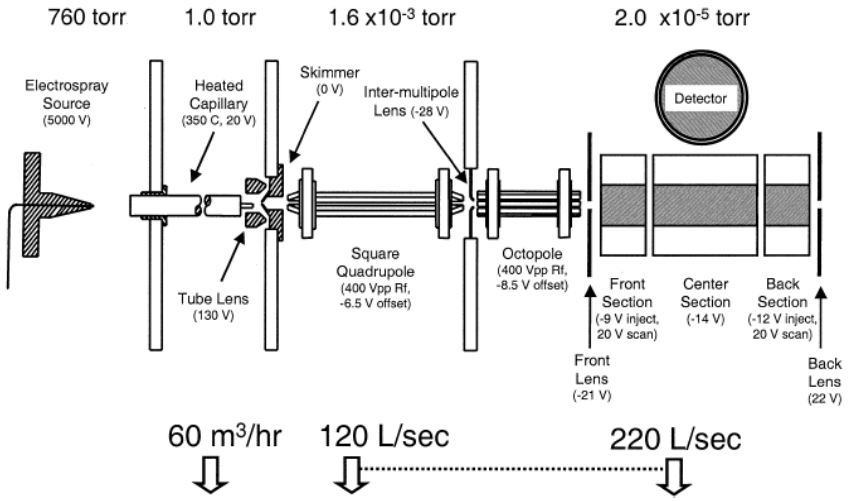


Figure 23 : schéma général du LTQ
Les ions sont injectés à gauche par un electrospray, ils seront analysés par la trappe linéaire située à droite du schéma. On reconnaît les sections centrales, avant et arrière de la Figure 22 et le cercle représente l'ouverture d'un multiplicateur d'électrons. Le dispositif au centre du schéma sert à idéaliser le spectromètre de masse. Le détecteur représente l'entrée d'un multiplicateur d'électrons (Figure 28). Notons l'indication des tensions de la trappe pour les deux phases, stockage et lecture. Figure extraite de [31].

2.4.3.1. Modèle de la phase de stockage

Pendant la phase de stockage ou d'injection les tensions appliquées aux sections avant et arrière sont ajustées afin que les ions provenant de l'electrospray et ayant passé à travers le quadripôle et l'octopôle puissent remplir la section centrale de la trappe ionique (Figure 23). Au bout de la période d'intégration T_{int}, les tensions des sections avant et arrière sont modifiées afin de piéger les ions dans la partie centrale. Si l'on s'intéresse au nombre υ_u^i d'ions i piégés dans la trappe à l'instant d'échantillonnage u, cela revient à intégrer le flux provenant de l'electrospray pendant une durée T_{int}.

$$\upsilon_u^i = \int_{t=uT_e^c-T\,\text{int}}^{uT_e^c} \varepsilon_i(t)\mathrm{d}t$$

Cependant, le flux $\varepsilon_i(t)$ varie peu pendant la période d'intégration. Nous réalisons donc l'approximation de l'intégrale suivante

$$\upsilon_u^i \approx T_{\text{int}}\varepsilon_i(uT_e^c)$$

2.4.3.2. Modèle de la phase de lecture

Une trappe à ion linéaire idéale peut être modélisée par une fonction retard. Ce retard d'une durée T_E correspond à l'instant où l'ion arrive à la coordonnée q_E, limite de la région de stabilité du diagramme, qui dépend de la masse et de la charge de l'ion, mais aussi de la tension appliquée à la trappe. Dans notre cas, l'amplitude *V(t)* de la tension appliquée augmente linéairement avec le temps, $V(t) = at + b$. D'après le Tableau 1 page 30 :

$$q_x(t) = \frac{4eV(t)}{mr_0^2\Omega^2} \text{ d'où } q_E = \frac{4eV(T_E)}{mr_0^2\Omega^2}$$

$$T_E = \frac{q_E r_0^2 \Omega^2}{4a}\frac{m}{e} - \frac{b}{a} \qquad (2.4\text{-}3)$$

Il y a donc une transformation affine qui associe le temps d'éjection et le rapport masse sur charge de l'ion. Dans les paragraphes suivants, nous utiliserons cette relation pour associer une masse aux temps intervenant dans les équations et les illustrations.

De plus, notons que cette transformation affine est identifiée par une procédure d'étalonnage standard conçue par le fabriquant où l'on mesure le spectre d'une substance connue. Ainsi l'appareil exprime les mesures effectuées en masse sur charge et non en temps. Nous ne remettrons pas en cause l'estimation de cette transformation dans ce document.

Ce modèle idéal est perturbé par plusieurs phénomènes.

- *Influence des conditions initiales*. Le modèle précédent considère que l'ion atteindra le détecteur au moment où sa trajectoire sera instable. Or les ions peuvent l'atteindre avant : la stabilité de la trajectoire n'empêche pas que son amplitude ne sera pas suffisante pour atteindre le détecteur. Ce phénomène dépendra des conditions initiales à travers les constantes *A* et *B* de l'équation (2.4-2).

  Inversement, quand la trajectoire d'un ion instable ne sera plus bornée, l'ion mettra un certain temps avant d'atteindre le détecteur. Mais ce temps est raccourci grâce à la présence d'un champ électrique supplémentaire excitant les ions dont la fréquence correspond à q_E. Les conditions initiales interviennent également car plus les constantes *A* et *B* seront importantes, plus l'ion atteindra le détecteur rapidement.

  En pratique, les conditions initiales sont fixées de la façon suivante. Les ions dissipent la plus grande partie de leur énergie cinétique par collision avec de l'hélium. De plus, l'angle du faisceau d'ions et sa position sont contrôlés et focalisés par les lentilles électrostatiques, le quadripôle et l'octopôle situé au centre de l'appareil (Figure 23). Malgré tout, les ions conserveront toujours une position et une vitesse initiale qui seront légèrement différentes d'un ion à l'autre, provoquant pour les ions d'un même rapport masse sur charge des temps d'arrivée légèrement différents.

- *Influence de la forme des électrodes*. Les électrodes ne sont pas parfaites et le potentiel électrique induit n'est donc pas parfait non plus. Il est donc légèrement différent à chaque coordonnée *z*. Deux ions situés à des positions *z* différentes auront des temps d'arrivée légèrement différents.

- *Influence du champ électrique induit par les ions de la trappe*. Chaque ion produit lui-même un champ électrique qui influencera les ions alentours. Ces champs parasites modifient le temps d'arrivée théorique des ions sur la trappe. Plus il y a d'ions dans la trappe, plus ce phénomène est important.

  Pour limiter ce problème, on s'arrange pour que le nombre d'ions dans la trappe soit à peu près le même, en jouant légèrement sur le temps d'intégration.

- *Pertes*. Tous les ions n'atteindront pas le détecteur. Certains seront neutralisés par des ions résidents, d'autres atteindront les électrodes et seront également neutralisés. Cependant ces pertes peuvent être négligées car un vide puissant est effectué, et l'amplitude du

mouvement sur l'axe y est faible, grâce à l'influence de l'absorption résonante. Tous les ions injectés peuvent ainsi être considérés comme arrivant sur le détecteur.

L'ensemble des phénomènes détaillés au paragraphe précédent a une conséquence sur le temps d'éjection de l'ion ζ. Nous allons considérer ce temps t_E^ζ comme une variable aléatoire dont la densité de probabilité f_ζ est une gaussienne centrée autour de T_E^ζ, donné par l'équation (2.4-3), et d'inverse variance γ_s (Figure 24) :

$$f_\zeta(t_E^\zeta) = \left(2\pi\gamma_s^{-1}\right)^{-1/2} \exp\left(-0.5\gamma_s\left(t - T_E^\zeta\right)^2\right)$$

γ_s peut dépendre de l'ion sur un spectromètre de masse quelconque, mais nous avons constaté que celui-ci variait peu sur le LTQ. Nous considérons donc γ_s comme étant indépendant de l'ion mesuré.

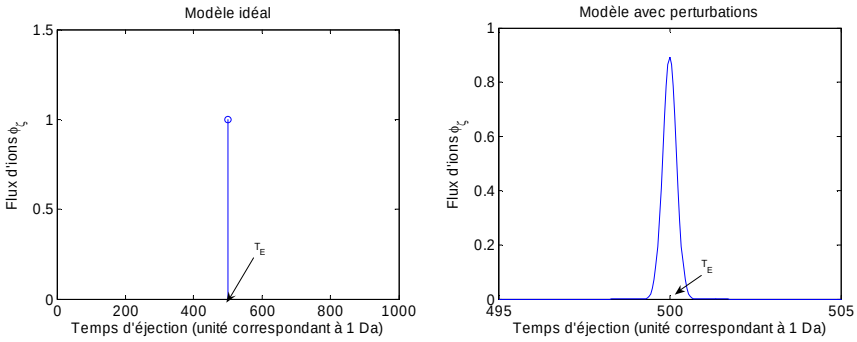


Figure 24 : densité de probabilité du temps d'éjection d'un ion.

Nous venons de déterminer la densité de probabilité du temps d'éjection d'un ion ζ d'une masse et d'une charge données. Examinons maintenant la densité de probabilité h_i de temps d'éjection t_E^i d'un peptide i.

Il faut tout d'abord noter qu'un spectromètre de masse est suffisamment sensible pour mesurer les variations de masse entre deux isotopes. Ainsi nous verrons l'influence sur le signal des neutrons supplémentaires portés par les peptides.

Chaque neutron supplémentaire augmente la masse du peptide d'environ 1 Dalton (Da), et donc son temps d'éjection de la trappe va également augmenter suivant la formule (2.4-3). La densité de probabilité h_i du temps d'éjection d'un peptide i se décompose suivant les masses possibles du peptide (Figure 14). La hauteur relative de chaque pic est donnée par la probabilité qu'un peptide ait un neutron supplémentaire. Elle peut être calculée à partir des proportions connues des isotopes de chaque élément chimique. Cet ensemble de pics représentant un même peptide mais ayant un nombre de neutrons légèrement différent est appelé massif isotopique.

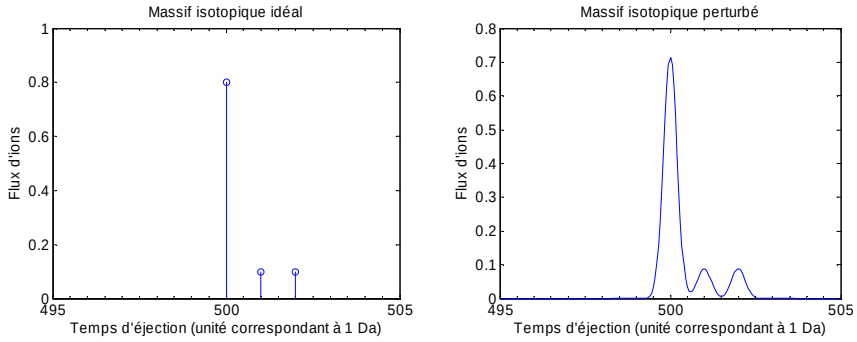


Figure 25 : massif isotopique

De même, comme nous l'avons vu au chapitre 2.3, le peptide peut être chargé une ou plusieurs fois. Le moment d'éjection de l'ion est proportionnel au rapport de sa masse et de sa charge (2.4-3). L'espacement entre deux pics du massif isotopique est donc de $1/e$ Da. Les ions du massif isotopique

de la Figure 25 situés aux alentours de 500 Da sont donc chargés une fois car leurs pics sont effectivement séparés de 1 Da.

Un nombre de charges différent modifie grandement le temps d'éjection du peptide. Un ion chargé deux fois plus sera éjecté deux fois plus vite. Nous retrouverons donc un massif d'ions chargés deux fois aux alentours de 250 Da et trois fois à 167 Da (Figure 26). La hauteur relative de ces massifs dépend de la probabilité de porter le nombre de charges correspondant.

Classiquement, un peptide pourra avoir jusqu'à 3 neutrons et jusqu'à 3 charges supplémentaires. Chaque ion ζ sera désigné dorénavant par le triplet (i,j,k) où chacun des indices représente respectivement la formule peptidique de l'ion, son nombre de charges et son nombre de neutrons supplémentaires.

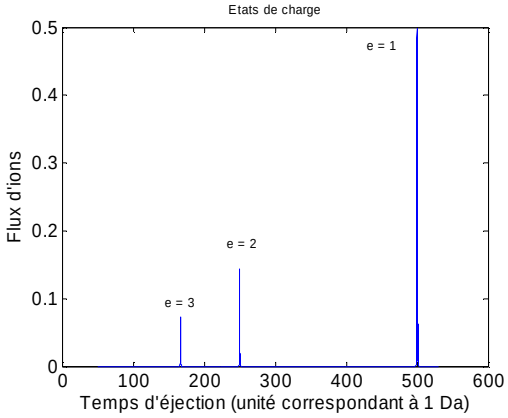


Figure 26 : états de charge

La densité de probabilité h_i du temps d'éjection t_E^i d'un peptide i se décompose finalement en plusieurs gaussiennes suivant les masses et les charges possibles du peptide dont les probabilités sont données respectivement par π_{ij}, probabilité du peptide i de porter j charges, et π'_{ijk}, probabilité du peptide i portant j charges de porter k neutrons supplémentaires par rapport à sa configuration la plus stable

$$h_i(t_E^i) = \sum_{j=1}^{J} \sum_{k=0}^{K} \pi_{ij} \pi'_{ijk} f_{ijk}(t_E^i)$$

J étant le nombre de charges maximales observées, K le nombre maximal de neutrons supplémentaires.

Examinons maintenant la grandeur de sortie de l'analyseur, le flux $\varphi_i(t)$ du peptide i sortant de la trappe linéaire. Celui ci dépend du nombre d'ions stockés dans la trappe. Si ce nombre est suffisamment important, l'histogramme des temps d'éjections et donc le signal $\varphi_i(t)$ correspondra à la densité de probabilité h_i multipliée par le nombre d'ions v_u^i stockés dans la trappe à l'instant d'échantillonnage u.

Ce phénomène sera répété à chaque instant d'échantillonnage (Figure 27). Ainsi $\varphi_i(t)$ le flux du peptide i sortant de la trappe ionique s'écrit sous la forme

$$\varphi_i(t) = \sum_{u=-\infty}^{+\infty} v_u^i h_i\left(t - uT_e^c\right)$$

De plus, à la section précédente, nous avons obtenu la formule suivante

$$v_u^i \approx T_{\text{int}} \varepsilon_i(uT_e^c)$$

En combinant les deux relations précédentes, nous obtenons

$$\varphi_i(t) \approx \sum_{u=-\infty}^{+\infty} T_{\text{int}} \varepsilon_i(uT_e^c) h_i\left(t - uT_e^c\right)$$

Nous avons donc relié le flux de peptides rentrant dans l'analyseur $\varepsilon_i(t)$ au flux de peptides sortant de l'analyseur $\varphi_i(t)$. Notons que cette relation peut s'écrire sous la forme d'une convolution entre h_i et le signal échantillonné de ε_i, mais la forme présentée nous semble plus adaptée pour les calculs du chapitre suivant.

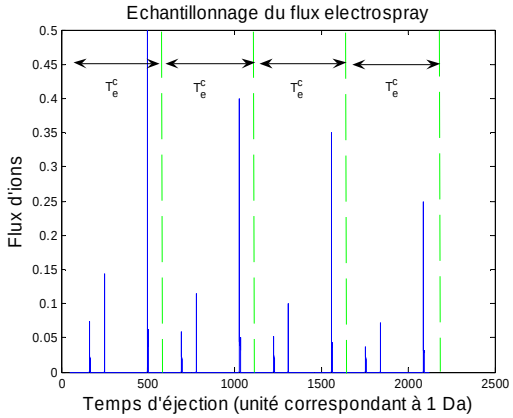


Figure 27 : flux d'ions sortant de la trappe.
Le signal correspondant à chaque ion i est répété à chaque période d'échantillonnage, modulé par la quantité de peptides i stockés dans la trappe.

2.5. Détecteur

Cette partie est dédiée à la dernière étape de la chaîne d'analyse. Jusqu'à présent les différents flux de molécules ont été séparés. Pour cela on a appliqué un retard spécifique à chaque peptide dans la colonne de chromatographie puis un retard à chaque ion ayant une masse et une charge données dans le spectromètre de masse. Il nous reste à convertir ce flux de matière en données numériques. Ces données seront traitées par la suite afin de reconstruire les concentrations de chaque peptide (chapitre 4).

Cette conversion peut être séparée en deux étapes :

- conversion du flux de matière en signal électrique effectué par le multiplicateur d'électrons,
- échantillonnage de ce signal électrique et présentation des données sous forme d'une image.

2.5.1. Multiplicateur d'électrons

Une description des détecteurs utilisés en spectrométrie de masse a été réalisée par Bolbach [45].

Le détecteur employé sur le LTQ est un multiplicateur d'électrons à dynode continue ou channeltron. Ce détecteur fonctionne de la façon suivante. L'impact entre les ions et le détecteur libère des électrons. Ces électrons sont accélérés par la dynode jusqu'à ce que leur impact contre l'électrode crée de nouveaux électrons plus nombreux. Le processus se répète en cascade et à la fin de la dynode on récupère un courant important (Figure 28).

Bolbach propose de modéliser ce détecteur comme un gain reliant le flux d'ions $\varphi_i(t)$ entrant au signal électrique produit par le peptide i, $y_i(t)$. Ce gain dépend de l'énergie cinétique du peptide lors du premier choc. C'est pourquoi nous le modéliserons par le gain G_i dépendant du peptide i.

$$y_i(t) = G_i \varphi_i(t)$$

Le signal électrique final est produit par la somme de tous ces signaux. En effet, deux électrons produits par des ions différents ne sont pas discernables les uns des autres.

$$y(t) = \sum_{i=1}^{I} y_i(t)$$

Le signal de flux de matière a été transformé en un signal électrique. L'échantillonnage de ce dernier sera étudié dans la section suivante.

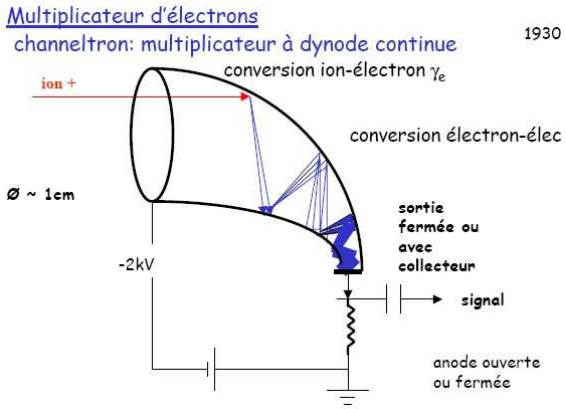


Figure 28 : channeltron.
L'ion en rouge frappe l'électrode. L'impact va créer plusieurs électrons par ions. Le champ électrique crée par les électrodes dirige les électrons créés de l'autre coté de la dynode. Les trajectoires des électrons sont représentées en bleu. Chaque choc avec la paroi crée plusieurs électrons pour chaque électron incident. Issu de [45].

2.5.2. Echantillonneur et présentation sous forme d'image

Pour être transformé sous forme numérique, le signal électrique du paragraphe précédent est échantillonné à une période T_e^s.

Il s'agit du deuxième échantillonnage intervenant dans la chaîne de mesure. En effet, le flux de peptides issu de la colonne de chromatographie avait été échantillonné à la période T_e^c par la trappe linéaire. Nous pouvons proposer une représentation des données mettant en valeur ces deux échantillonnages. On peut interpréter le signal électrique que nous échantillonnons actuellement comme la succession de spectres de masse occupant une période T_e^c sur le signal (Figure 27, page 39). Si nous nous arrangeons pour que T_e^c soit un multiple de T_e^s, nous pouvons réorganiser les échantillons sous forme d'image de façon à ce que tous les échantillons représentant le même rapport masse sur charge soient situés sur une même ligne. Chaque colonne correspondrait à un spectre de masse donné ou un temps de rétention chromatographique donné.

Chaque échantillon sera dorénavant représenté par deux indices, le premier représentant une masse sur charge, le second un temps de rétention chromatographique. Ces deux indices peuvent être obtenus par division euclidienne. En notant *v* l'ancien numéro de l'échantillon,

$$v = nN_s + m \text{ avec } T_e^c = N_s T_e^s$$

m est le numéro de l'échantillon correspondant à une masse donnée, *n* est le numéro de l'échantillon correspondant à un temps de rétention chromatographique donné. De plus, N_s est le nombre d'échantillons par spectre de masse, qui correspond également au rapport entre les deux périodes d'échantillonnage.

Nous avons vu que le double échantillonnage se prêtait à une représentation des données sous forme d'image avec un axe représentant l'aspect spectrométrique et l'autre représentant l'aspect chromatographique. Nous pouvons nous demander si le signal a une structure répondant à ce découpage. En particulier nous allons montrer que le signal d'un peptide est séparable suivant ces deux axes.

Nous nous intéressons au signal échantillonné de chaque peptide. Rappelons que le signal échantillonné total est la somme du signal de chaque peptide *i*.

$$y(vT_e^S) = \sum_{i=1}^{I} y_i(vT_e^s)$$

Synthétisons les formules issues des sections précédentes de manière à mettre en évidence les signaux chromatographiques et spectrométriques normalisés du peptide *i*.

$$\begin{aligned} y_i(vT_e^s) &= G_i \varphi_i(vT_e^s) \\ &= \eta_i T_{\text{int}} K_i G_i \sum_{u=-\infty}^{+\infty} g_i(uT_e^c)\, h_i(vT_e^s - uT_e^c) \end{aligned}$$

g_i étant le chromatogramme normalisé du peptide *i*, h_i le spectre de masse normalisé de *i*. Réunissons les gains autres que la quantité de peptide sous la variable $\xi_i = T_{\text{int}} K_i G_i V$, V étant le volume d'échantillon injecté en début de système. L'utilisation de cette variable simplifiera les expressions finales faisant intervenir les concentrations des protéines d'intérêt (sections 2.1 page 18 et 2.6 page 42). De plus, si on applique le changement d'indice correspondant aux coordonnées de l'image, on obtient

$$\begin{aligned} y_i\left([nN_s + m]T_e^S\right) &= \frac{\eta_i \xi_i}{V} \sum_{u=-\infty}^{+\infty} g_i\left(uT_e^c\right) h_i\left([nN_s + m]T_e^s - uT_e^c\right) \\ &= \frac{\eta_i \xi_i}{V} \sum_{u=-\infty}^{+\infty} g_i\left(uT_e^c\right) h_i\left([n-u]T_e^c + mT_e^s\right) \end{aligned}$$

Or $h_i(t)$ est négligeable hors du domaine $t \in \left[0; T_e^c\right]$, qui dans notre modèle correspond au domaine du rapport masse sur charge de ce spectromètre. Donc le seul terme de la somme potentiellement non négligeable est celui correspondant à $u = n$.

$$y_i\left([nN_s + m]T_e^s\right) \approx \frac{\eta_i \xi_i}{V} g_i\left(nT_e^c\right) h_i\left(mT_e^s\right)$$

Représentons ces échantillons sous forme matricielle. Soit $\boldsymbol{Y}_i$ la matrice formée par ces échantillons et représentant le signal final d'un peptide *i*.

$$\left(\boldsymbol{Y}_i\right)_{m,n} = \frac{\eta_i \xi_i}{V} g_i\left(nT_e^c\right) h_i\left(mT_e^s\right)$$

Dans les pages précédentes nous avons insisté sur la dualité existant entre le temps d'éjection de la trappe linéaire et les rapports masse sur charge correspondant. La séparation entre les aspects chromatographique et spectrométrique dans le signal final permet de remplacer la période d'échantillonnage T_e^s par la période d'échantillonnage en masse M_e^s équivalente.

$$\left(\boldsymbol{Y}_i\right)_{m,n} = \frac{\eta_i \xi_i}{V} g_i\left(nT_e^c\right) h_i\left(m_0^s + mM_e^s\right)$$

m_0^s correspond à la masse du premier échantillon.

Ces formules montrent que la réorganisation des échantillons rend le signal séparable entre les aspects chromatographiques et spectrométriques (g_i et h_i désignent les signaux normalisés issus de la chromatographie et de la spectrométrie). Une dimension des images issue de la chaîne d'analyse ne concernera que l'aspect chromatographique, l'autre ne concernera que l'aspect spectrométrique. Cette étude nous a permis de comprendre par quels mécanismes une chaîne de mesure protéomique produit une image.

2.6. Modèle retenu

Dans les sections précédentes, nous avons proposé des modèles pour chaque module de la chaîne d'analyse. Nous les synthétisons dans cette section de façon à relier les variables d'intérêt aux données.

Le signal final est formé de la somme des signaux de chaque peptide. De plus, ces signaux ont deux dimensions séparables décrivant les séparations chromatographique et spectrométrique décrites par les fonctions g_i et h_i.

$$(\boldsymbol{Y})_{m,n} = \sum_{i=1}^{I} (\boldsymbol{Y}_i)_{m,n} = \sum_{i=1}^{I} \frac{\eta_i \xi_i}{V} g_i\left(nT_e^c\right) h_i\left(m_0^s + mM_e^s\right)$$

Chaque signal spectrométrique h_i est formé de plusieurs fonctions gaussiennes f_{ijk} décrivant chacune un ion de charge et de nombre de neutrons donné.

$$(\boldsymbol{Y})_{m,n} = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=0}^{K} \frac{\eta_i \xi_i}{V} \pi_{ij} \pi'_{ijk} \, f_{ijk}\left(m_0^s + mM_e^s\right) g_i\left(nT_e^c\right)$$

Enfin, il faut relier la quantité de peptides η_i aux concentrations des protéines x_i à travers la formule de digestion $\eta_i = V \sum_{p=1}^{P} d_{ip} x_p$.

$$(\boldsymbol{Y})_{m,n} = \sum_{p=1}^{P} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=0}^{K} x_p d_{ip} \xi_i \pi_{ij} \pi'_{ijk} \, f_{ijk}\left(m_0^s + mM_e^s\right) g_i\left(nT_e^c\right)$$

Rappelons que f_{ijk} et g_i sont deux fonctions gaussiennes monodimensionnelles.

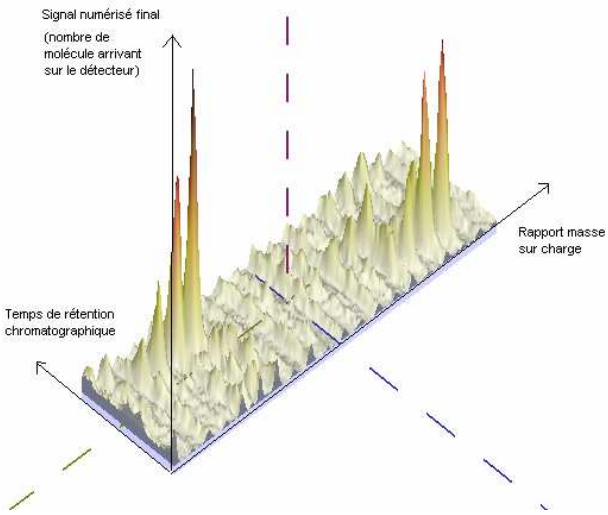


Figure 29 : exemple de signal bidimensionnel obtenu.
Une fois les échantillons réorganisés, on distingue que le signal 2D est formé d'une somme de pics modélisables par des gaussiennes bi-dimensionnelles. Il s'agit que d'une petite partie du signal total. Cette image a été obtenue à partir de données réelles en utilisant le logiciel de visualisation MSight du Swiss Institute of Bioinformatics.

Le signal est donc formé d'une somme de gaussiennes bidimensionnelles, modèle confirmé par les données réelles (Figure 29). Chaque gaussienne est reliée à une protéine d'intérêt. Cette relation met également en évidence la présence d'une redondance d'information, plusieurs gaussiennes pouvant être utilisées pour retrouver la concentration des protéines d'intérêt.

Afin de simplifier les futurs calculs, nous pouvons également introduire les vecteurs $\boldsymbol{s}_{ijk}$ et $\boldsymbol{c}_i$

$$\boldsymbol{Y} = \sum_{p=1}^{P} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=0}^{K} x_p d_{ip} \xi_i \pi_{ij} \pi'_{ijk} \boldsymbol{s}_{ijk} \boldsymbol{c}_i^t$$

avec

$$\boldsymbol{s}_{ijk} = \left[f_{ijk}\left(m_0^s\right) \quad f_{ijk}\left(m_0^s + M_e^s\right) \quad \ldots \quad f_{ijk}\left(m_0^s + (N_s - 1)M_e^s\right) \right]^t$$

$$\boldsymbol{c}_i = \left[g_i(0) \quad g_i\left(T_e^c\right) \quad \ldots \quad g_i\left((N_c - 1)T_e^c\right) \right]^t$$

$$f_{ijk}(m) = \left(2\pi\gamma_s^{-1}\right)^{-1/2} \exp\left(-\frac{1}{2}\gamma_s\left(m - m_{ijk}\right)^2 \right)$$

$$g_i(t) = \left(2\pi\gamma_c^{-1}\right)^{-1/2} \exp\left(-\frac{1}{2}\gamma_c\left(t - t_i\right)^2 \right)$$

$\boldsymbol{c}_i$: chromatogramme théorique du peptide i discrétisé

$\boldsymbol{s}_{ijk}$: spectre théorique discrétisé du peptide i , portant j charges et k neutrons supplémentaires

f_{ijk} : spectre théorique de l'ion provenant du peptide i, portant j charges et k neutrons

g_i : chromatogramme du peptide i

$\boldsymbol{Y}$: données

x_p : concentration de la protéine p

d_{ip} : nombre de peptides i générés par la protéine p

ξ_i : gain du système pour le peptide i

γ_b : inverse variance du bruit,

t_i : position de la gaussienne chromatographique i

m_{ijk} : position de la gaussienne spectrométrique de l'ion (i, j, k)

γ_c : largeur des gaussiennes chromatographiques

γ_s : largeur des gaussiennes spectrométriques

π_{ij} : proportion du peptide i portant j charges

π'_{ijk} : proportion du peptide i ayant j charges portant k neutrons supplémentaires

N_s : nombre de points correspondant à la dimension de la spectrométrie de masse

N_c : nombre de points correspondant à la dimension de la chromatographie

Le modèle produit est formé d'une somme de gaussiennes et nous conseillons une écriture matricielle mettant en valeur la structure séparable du signal et facilitant les futurs calculs. Nous pouvons également ajouter quelques remarques sur le modèle retenu.

Premièrement, parmi toutes les gaussiennes correspondant à un peptide i, une des gaussiennes est généralement beaucoup plus importante que les autres. Cette gaussienne correspond à l'ion $\left(i, \bar{j}_i, 0\right)$ composé des isotopes les plus stables et les plus répandus (H^1, C^{12}, N^{14}, O^{16}, S^{32}), c'est-à-dire ne portant pas de neutrons supplémentaires ($k = 0$). De plus, nous avons observé dans nos expériences que les conditions expérimentales favorisaient un nombre de charges optimal pour chaque peptide que nous appellerons $\bar{j}_i$. Si nous souhaitons nous intéresser à ces gaussiennes là et négliger les autres, nous pouvons utiliser les notations suivantes :

$$\pi_{ij} = \delta_{j;\bar{j}_i} \text{ et } \pi'_{ijk} = \delta_{k;0}$$

où $\delta_{a;b}$ est le symbole de Kronecker.

Deuxièmement, nous pouvons calculer la valeur des positions m_{ijk} des gaussiennes spectrométriques pour chaque ion (i, j, k). Cette position dépend de la masse de l'ion divisée par son nombre j de charges exprimé en Dalton. Dans le cas standard, les charges supplémentaires sont constituées de protons. Ces protons et les neutrons supplémentaires modifieront d'autant la masse de l'ion. La masse d'un proton et d'un neutron est proche d'un Dalton. Soit $\overline{M}_i$ la masse du peptide i sans neutron ni proton supplémentaire. Elle peut être calculée à partir de la formule peptidique de i en utilisant par exemple le PIR Molecular weight calculator [46]. La position des gaussiennes spectrométriques m_{ijk} peut être trouvée grâce à la formule suivante :

$$m_{ijk} = \frac{\overline{M}_i + j\,m_{\mathrm{p}} + k\,m_{\mathrm{n}}}{j} \approx \frac{\overline{M}_i + j + k}{j}$$

avec m_{p} la masse d'un neutron et m_{n} la masse d'un neutron dont la valeur est proche d'un Dalton.

Enfin, notons que pour des raisons de simplicité, nous avons supposé que les deux échantillonnages chromatographique et spectrométrique étaient réguliers. Afin d'améliorer les performances certains instruments pourront proposer des échantillonnages non réguliers plus complexes. Dans un premier temps, on pourra interpoler les données obtenues afin d'obtenir un échantillonnage régulier.

Dans cette section nous avons donc proposé un modèle de la chaîne de mesure reliant les variables d'intérêt aux données, il reste à modéliser les erreurs de ce modèle.

2.7. Modélisation du bruit

La modélisation retenue n'est bien entendu pas parfaite, elle ne rend pas compte de la totalité des phénomènes ayant lieu dans l'appareil. Ainsi, il existera une erreur de modélisation $b_{n,m}$ entre les données réelles et les données issues du modèle pour chaque échantillon mesuré. Ces différentes erreurs sont regroupées dans la matrice $\boldsymbol{B}$.

$$\boldsymbol{Y} = \sum_{p=1}^{P}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=0}^{K} x_p d_{ip} \xi_i \pi_{ij} \pi'_{ijk}\, \boldsymbol{s}_{ijk} \boldsymbol{c}_i^t + \boldsymbol{B} \text{ avec } (\boldsymbol{B})_{m,n} = b_{m,n}$$

Nous disposons de peu d'informations sur ces erreurs. Toutefois, nous pouvons modéliser sa densité de probabilité *a priori* à partir des hypothèses suivantes.

- Le modèle rend suffisamment bien compte des données. Il n'y a pas d'erreur systématique évidente. Nous supposerons donc que la moyenne de la distribution $p(b_{m,n})$ est nulle.
- Les erreurs ont une échelle caractéristique. Nous supposerons que la variance de la distribution $p(b_{m,n})$ existe et vaut γ_b^{-1}.
- Il n'y a pas de corrélation évidente entre les erreurs. Les erreurs $b_{m,n}$ sont indépendantes.
- Nous n'avons aucune information permettant de dire que les propriétés statistiques des erreurs sont *a priori* différentes. Nous supposerons donc que le bruit est stationnaire.

Ces considérations nous amènent à modéliser les erreurs par la densité de probabilité gaussienne

$$p(b_{m,n}|\gamma_b) = (2\pi\gamma_b^{-1})^{\frac{-1}{2}} \exp\left(-\frac{1}{2}\gamma_b b_{m,n}^2\right)$$

Nous pouvons également obtenir la densité de probabilité de la matrice $\boldsymbol{B}$ à partir des règles de calcul des probabilités [47, 48]. Les erreurs étant indépendantes

$$
\begin{aligned}
p(\boldsymbol{B}|\gamma_b) &= p(b_{1,1}\dots b_{N_s,N_c}|\gamma_b) \\
&= \prod_{m=1}^{N_s}\prod_{n=1}^{N_c} p(b_{m,n}|\gamma_b) \\
&= \prod_{m=1}^{N_s}\prod_{n=1}^{N_c} \left(2\pi\gamma_b^{-1}\right)^{\frac{-1}{2}} \exp\left(-\frac{1}{2}\gamma_b b_{m,n}^2\right)
\end{aligned}
$$

Cette distribution peut être reformulée en utilisant la norme matricielle de Frobenius [49]

$$
\begin{aligned}
p(\boldsymbol{B}|\gamma_b) &= \prod_{m=1}^{N_s}\prod_{n=1}^{N_c} \left(2\pi\gamma_b^{-1}\right)^{\frac{-1}{2}} \exp\left(-\frac{1}{2}\gamma_b b_{m,n}^2\right) \\
&= \left(2\pi\gamma_b^{-1}\right)^{\frac{-N_cN_s}{2}} \exp\left(-\frac{1}{2}\gamma_b \sum_{m=1}^{N_s}\sum_{n=1}^{N_c} b_{m,n}^2\right) \\
&= \left(2\pi\gamma_b^{-1}\right)^{\frac{-N_cN_s}{2}} \exp\left(-\frac{1}{2}\gamma_b \|\boldsymbol{B}\|^2\right)
\end{aligned}
$$

avec $\|\boldsymbol{B}\|$ la norme de Frobenius

$$
\|\boldsymbol{B}\|^2 = \mathrm{Tr}\left(\boldsymbol{B}^t\boldsymbol{B}\right) = \sum_{m,n} b_{m,n}^2 \ .
$$

De plus, remarquons qu'il y a une équivalence entre la norme matricielle de Frobenius et la norme vectorielle euclidienne, $\|\boldsymbol{B}\| = \|\boldsymbol{b}\|$ si $\boldsymbol{b}$ est le vecteur formé par la concaténation des colonnes de $\boldsymbol{B}$.

Ce modèle du bruit est le premier élément probabiliste intervenant dans cette thèse. C'est également la base sur laquelle est construite l'estimation bayésienne des concentrations, comme nous le verrons dans le chapitre 4.

2.8. Conclusion

Dans ce chapitre, nous avons présenté le fonctionnement d'une chaîne de mesure protéomique. Les principes et les modèles associés à chaque module ont été décrits. Nous avons notamment étudié la colonne de chromatographie et le spectromètre de masse. Le premier obéit à une équation de convection-diffusion et le second à une équation de Matthieu. En première approximation, les solutions de ces équations différentielles sont approchées par une superposition de fonctions gaussiennes monodimensionnelle. Nous nous sommes particulièrement intéressés aux éléments utilisés dans le projet européen LOCCANDIA et le projet transverse CEA CAPSI, à savoir une colonne de chromatographie liquide utilisant une phase inverse hydrophobe et une trappe linéaire. Nous avons étudié leur spécificité, mais les équations et les mécanismes exposés ont également un caractère général et sont utilisés dans la plupart des autres systèmes chromatographiques et spectrométriques. Nous avons également pu comprendre comment cet instrument combine ces deux systèmes et produit une image à deux dimensions composée de gaussiennes bidimensionnelles.

L'étude détaillée de chaque module a permis de comprendre sur quelle base physique ils fonctionnent et de proposer un modèle global de la chaîne d'analyse. Il s'agit d'un premier modèle qui pourra être enrichi dans des travaux ultérieurs. En effet, pour chaque module, les travaux que nous avons cités proposent des pistes pour améliorer la précision du modèle. Cependant, un modèle plus complexe ne nous permettra pas à coup sûr d'obtenir une meilleure estimation des concentrations. En effet, chaque modèle fait intervenir d'autres variables que nous appellerons paramètres instrument. Nous verrons dans le chapitre suivant que dans les cas complexes, ces variables influent sur la mesure des variables d'intérêt. De plus, leur valeur est rarement connue et elles nécessitent d'être également estimées. Certaines de ces variables pourront être estimées préalablement, d'autres devront être estimées conjointement : ce sera le sujet du chapitre 4. Bien sûr, plus le nombre de ces variables instrument est grand, et moins leurs valeurs sont connues, plus l'estimation des concentrations sera délicate.

Dans notre méthode, nous nous concentrerons sur ce modèle composé d'une somme de fonctions gaussiennes. En limite de résolution et de sensibilité, ces gaussiennes se rapprocheront jusqu'à se confondre. Notre méthode d'estimation devra donc pouvoir séparer les différents signaux. De plus, le modèle nous indique que l'information de concentration est répartie sur plusieurs gaussiennes. Cette redondance devra être utilisée afin d'obtenir les meilleures performances.

Nous verrons dans le chapitre suivant les méthodes actuellement utilisées pour quantifier les données protéomiques et leurs limites.

3. Etat de l'art

Dans le chapitre précédent, nous décrivons la chaîne de mesure comme étant formée d'une série de modules. Chacun de ces modules a été associé à un ensemble d'équations. Après quelques simplifications, nous avons abouti à un modèle global de cette chaîne d'analyse, produisant des images formées de pics gaussiens 2D. Dans ce chapitre, nous décrirons le problème de quantification des espèces dans le cas de données spectrométriques. Bien sûr, nous nous intéresserons particulièrement au domaine de la protéomique, mais nous évoquerons également les développements réalisés pour des domaines voisins.

La section 3.1 sera consacrée au pipeline classique de traitement de données LC/MS[1]. Nous décrirons notamment le problème de l'étalonnage isotopique et son utilisation pour la quantification absolue des protéines. Puis, à la section 3.2, nous élargirons notre étude en nous intéressant aux méthodes d'analyse factorielle souvent utilisées en chimiométrie[2] [50, 51]. Enfin à la section 3.3, nous introduirons l'approche bayésienne pour le traitement de données spectrométriques en présentant quelques travaux caractéristiques.

3.1. Traitement de données classique en protéomique

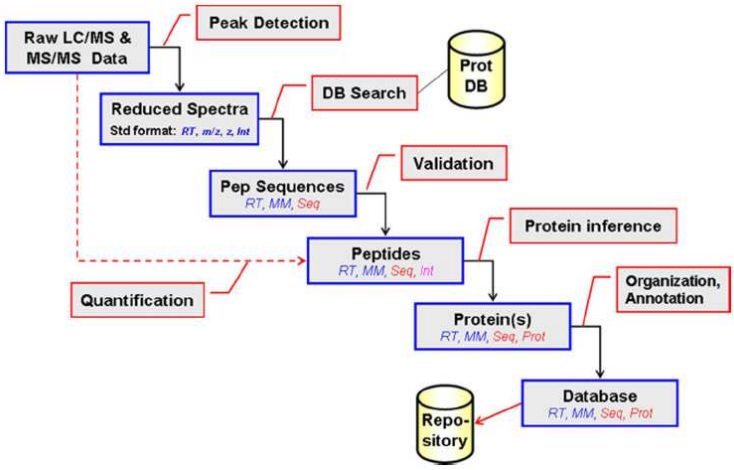


Figure 30 : le pipeline de traitement classique en protéomique.
Figure extraite de [52].

En ce qui concerne la protéomique, le traitement de données est réalisé par un logiciel ou un ensemble de logiciels formant généralement un pipeline de traitement (Figure 30) [52-55]. Ce pipeline résulte d'un découpage analytique du problème :

1) détection des pics,

[1] Données issues de chaînes associant la chromatographie liquide (LC) et la spectrométrie de masse (MS).

[2] La chimiométrie rassemble les méthodes de traitement de données utilisées en chimie.

2) identification des peptides,

3) quantification des peptides,

4) identification et quantification des protéines,

5) gestion des données recueillies.

Du fait de la complexité des échantillons et de la quantité de données à analyser, le problème de détection-estimation est découpé en plusieurs étapes. Tout d'abord, on détecte les pics présents dans les données. Cette étape de décimation permet de réduire les données brutes à un ensemble de fonctions de Dirac, la « liste de pics » résumant la position des pics et leur intensité. Cette liste est ensuite confrontée à une base de données listant une bibliothèque de pics possibles. Cette dernière permet de proposer des identifications de « candidats peptides » associés à un score mesurant la fiabilité de la reconnaissance. Cette première étape d'inférence est suivie d'une seconde étape d'inférence où les différentes protéines sont déduites à partir des différents peptides candidats. L'étape de quantification peut donner une information supplémentaire pour identifier certaines protéines. Ce découpage permet donc de simplifier le problème de traitement en ne conservant que la quantité minimale d'information nécessaire. Notons que le problème de gestion de données est une préoccupation importante. Les laboratoires doivent conserver une trace de leurs expériences et des conditions expérimentales, mais la taille des données générées par l'expérience et la diversité des instruments rend ce problème non trivial. Plusieurs infrastructures ont été proposées pour le résoudre.

La quantification des protéines s'inscrit dans un processus général utilisé par la majeure partie des logiciels de traitement de données LC-MS. Les sections suivantes se concentreront sur l'étape de quantification qui est le centre de la problématique de cette thèse. Avant de décrire les techniques utilisées pour estimer l'intensité d'un pic et relier cette intensité à une concentration à l'aide d'un marquage isotopique, nous listerons les prétraitements généralement employés.

3.1.1. Marquage isotopique

3.1.1.1. Introduction

Le passage de la protéomique analytique, qui a pour but l'identification des protéines, à la protéomique quantitative, qui doit estimer la concentration de ces protéines, pose la question suivante. Comment comparer les résultats d'expériences réalisées sur des systèmes sensibles aux perturbations ? En effet, les paramètres des systèmes LC-MS subissent de nombreuses variations, dont la principale est celle du gain du système pour chaque peptide qui varie d'une façon importe d'une expérience à une autre. Si des techniques « sans marquage » sont développées (label free quantification) [56-58], les réponses à ce problème sont majoritairement basées sur un marquage isotopique. Cette technique consiste à marquer les échantillons d'intérêt à l'aide de molécules plus ou moins alourdies, puis d'analyser conjointement leur mélange. Le spectromètre de masse sera capable de séparer leurs signaux, mais le reste de l'analyse se fera dans des conditions identiques, ce qui permettra de s'affranchir des variations du gain et de comparer les deux expériences.

La plupart de la littérature se concentre sur les techniques de synthèse permettant d'incorporer les molécules marquées aux molécule d'intérêt [59-62]. La Figure 31 propose une classification de ces différentes techniques. La plupart d'entre elles visent à réaliser une quantification relative des différentes protéines en réponse à un stimulus par exemple. Certains marquages sont réalisés *in vivo* par les processus biochimiques de la cellule, d'autres sont réalisés *in vitro* par différents procédés de synthèse chimique. Parallèlement à ces techniques de quantification relative, des techniques de quantification absolue se développent. Ici, le but n'est plus de comparer deux échantillons voisins, mais d'estimer la quantité de protéine d'intérêt. Pour cela, on incorpore des peptides ou des protéines synthétiques servant d'étalon dans l'échantillon. Ces étalons seront de concentration connue et permettront d'affecter à chaque protéine d'intérêt leur concentration.

Plusieurs techniques de quantification absolue existent, mais leurs performances dépendent de leur endroit d'injection dans la chaîne d'analyse. L'étalon idéal doit subir les mêmes transformations que la protéine d'intérêt : étapes de préparation, de digestion, *etc*. Nous nous intéresserons en particulier à la

méthode PSAQ (Protein Standard Absolute Quantification) développée récemment qui a l'avantage de pouvoir être injectée en début de chaîne d'analyse.

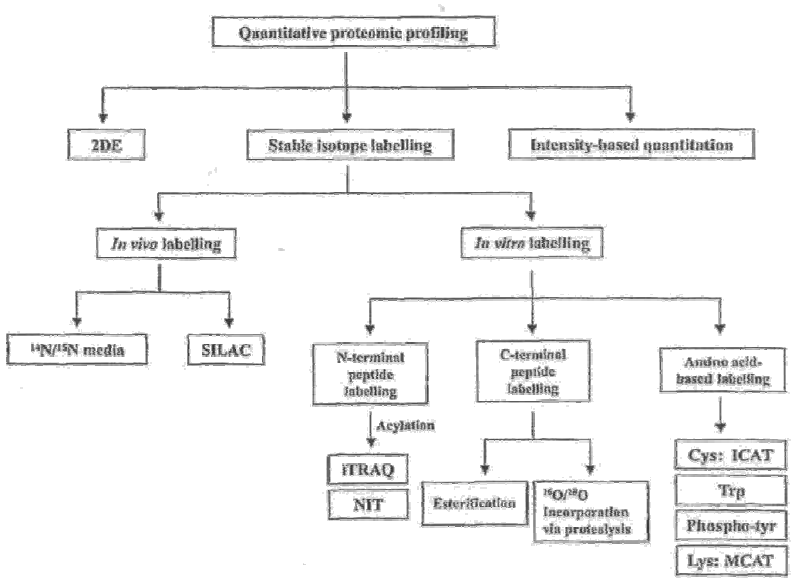


Figure 31 : diagramme représentant les différentes stratégies en analyse quantitative.
Figure extraite de [61].

3.1.1.2. Méthode d'étalonnage PSAQ (Protein Standard Absolute Quantification)

Le laboratoire EDyP de l'iRTSV[1] a développé une technique d'étalonnage particulière pour mesurer par exemple la concentration absolue en toxines du staphylocoque doré. Cette quantification exacte est rendue difficile par la nature complexe des échantillons à analyser qui demande l'adjonction d'étapes supplémentaires de traitements des échantillons.

L'étude publiée par V. Brun *et al*. [63] a montré les limites des méthodes de quantification absolue actuelles : AQUA (absolute quantification) et QconCAT (quantification concatamers) d'où la nécessité d'élaborer une autre technique. En effet ces méthodes ne prennent pas en compte les rendements de toutes les étapes biochimiques nécessaires à la préparation de l'échantillon et effectuent donc une erreur dans leur quantification (Figure 32).

Par son adjonction dans l'échantillon biologique, les échantillons PSAQ subissent toutes les modifications et pertes induites par la totalité de la chaîne analytique. De plus, la biosynthèse en système « cell-free » permet de mieux contrôler la qualité, la pureté, de l'étalon produit.

[1] Institut de Recherche en Technologies et en Sciences du Vivant du Commissariat à l'Energie Atomique de Grenoble.

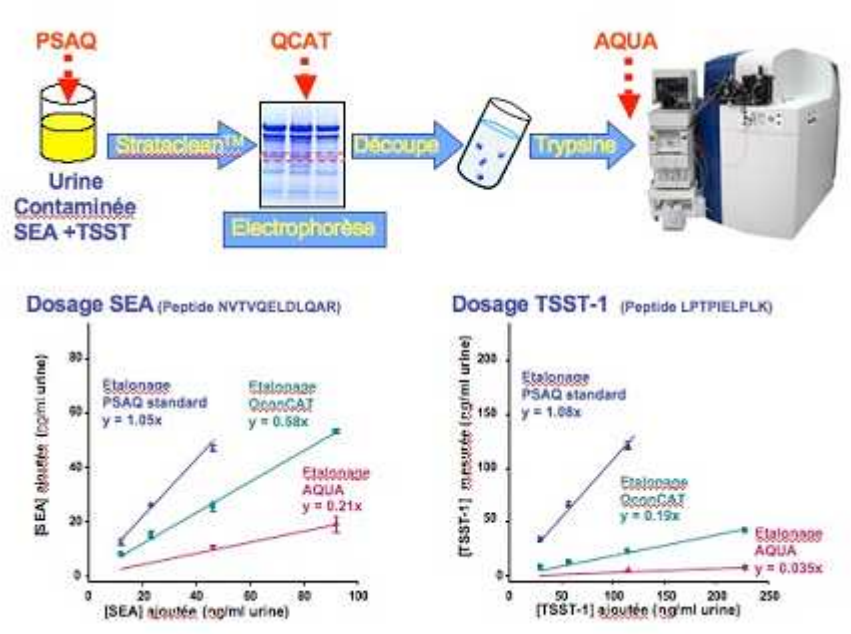


Figure 32 : comparaison des méthodes de quantification pour les toxines SEA et TSST.
Un échantillon d'urine est contaminé avec des quantités connues d'entérotoxines SEA et TSST (toxines naturelles). Chaque échantillon est divisé en trois, enrichi en toxine cible par concentration sur Strataclean™ et électrophorèse, et quantifié de manière comparative par les 3 méthodes. Les trois types d'étalon utilisés sont ajoutés comme indiqué par les flèches. Pour chaque toxine, la titration d'un même peptide protéotypique est représentée pour les trois méthodes (y = pente de la droite de titration). Figure extraite de [64].

3.1.2. Quantification et intensité des pics

Nous avons décrit l'importance de l'étalonnage isotopique pour la mesure d'une concentration absolue. En effet, grâce aux techniques de marquage, la quantification se réduit à l'application d'une règle de trois. Cependant, sur quel élément appliquer cette règle de proportionnalité ?

En effet, plusieurs techniques existent pour estimer l'intensité d'un pic ; par exemple, nous pouvons estimer la valeur de son maximum, l'aire d'une coupe suivant l'axe chromatographique ou spectrographique ou encore estimer son volume. De façon générale, plus le paramètre du pic choisi utilise de données, plus l'influence du bruit diminue. Les méthodes estimant le volume sont donc la plupart du temps préférées par rapport à celles estimant le maximum du pic. Nous présentons dans les paragraphes suivant les détails de la méthode utilisée dans [63] qui nous servira de méthode de référence dans la suite du document.

Tout d'abord, à partir du spectrogramme nous calculons un chromatogramme extrait (extracted ion current ou XIC) centré sur le rapport masse sur charge du pic d'intérêt comme représenté sur le schéma suivant. Il s'agit pour chaque temps de rétention de cumuler les valeurs dans la zone définie. Typiquement, ce chromatogramme extrait considère le premier pic du massif isotopique. Ensuite, selon l'expérimentateur, l'aire de la courbe ou le maximum de la courbe constitue la mesure. On réalise de la même manière le calcul pour le peptide marqué. Finalement, on obtient la quantification du peptide par une règle de trois à l'aide de la connaissance de sa concentration de protéines marquées injectées.

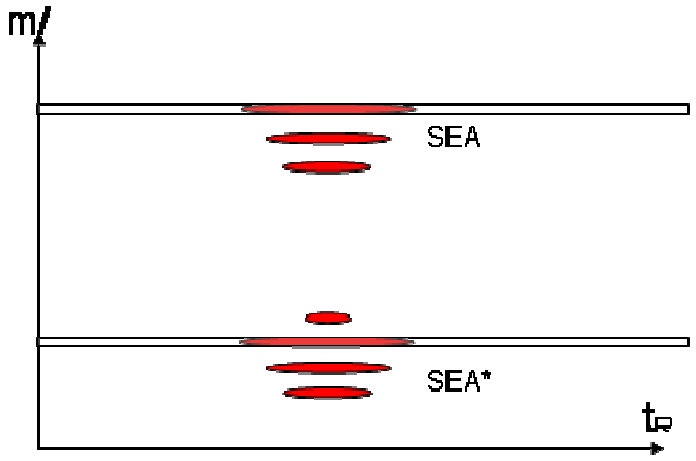


Figure 33 : calcul de XIC à partir de spectrogramme.

Nous avons listé quelques techniques permettant d'estimer l'intensité des pics. Bien sûr, de nombreuses autres méthodes sont également possibles. Le lecteur pourra par exemple se référer à [65] pour une revue plus complète.

3.1.3. Prétraitements

Une autre étape importante du processus de quantification des peptides concerne les prétraitements réalisés sur les données brutes. Hilario *et al.* proposent une revue de ces différents traitements [66]. Ils distinguent les étapes suivantes :

- élimination de la ligne de base,
- estimation du bruit,
- lissage,
- détection des pics surnageant,
- détection du massif isotopique,
- simplification du spectre de masse d'un peptide en un pic unique (de-isotoping, charge deconvolution),
- alignement de ces pics en vue d'une comparaison de plusieurs spectrogrammes.

Certains de ces prétraitements visent à éliminer le bruit chimique ou permettent de détecter les massifs isotopiques dans le signal, d'autres permettent de décomplexifier le signal. Ces différents traitements sont destinés à diminuer l'influence des différentes perturbations du système sur l'estimation des concentrations, ils sont donc essentiels. Mais les solutions proposées sont souvent développées de façon *ad hoc* et nécessitent le réglage de nombreux paramètres.

3.1.4. Conclusion

Nous avons effectué un panorama des méthodes de traitement dédiées aux données LC-MS, en nous intéressant plus particulièrement au domaine de la quantification. Nous avons donc évoqué les prétraitements utilisés, les techniques d'estimation de l'intensité et l'utilisation de molécules étalon.

La caractéristique de ces méthodes est de découper le traitement de données en plusieurs étapes (prétraitements, identification en plusieurs étapes, quantifications, …). Cependant, l'ordre dans lequel effectuer ces étapes n'est pas évident. Listgarten remarque, par exemple, que pour effectuer une quantification relative, il lui faut d'abord réaligner les différents signaux, mais pour effectuer un bon alignement, il lui faut connaître l'intensité des pics [67]. C'est souvent le cas pour de nombreux autres traitements. Par exemple, l'estimation du bruit fait généralement partie des premières étapes

effectuées, mais idéalement, pour bien estimer le bruit il faut connaître toutes les caractéristiques des autres signaux présents dans les données et donc attendre la fin du traitement.

De plus, notons que ces méthodes sont basées sur la détection de pics bien séparés du reste de leur environnement, elles trouveront leurs limites dans le cas où les signaux se superposent, mais également dans le cas de signaux d'intérêt faibles ou de bruit élevé. En effet, il deviendra difficile de distinguer les pics.

Pour répondre à ces problèmes, nous devons utiliser des méthodes qui cherchent à séparer le signal d'intérêt des autres signaux, mais qui permettent également d'utiliser le maximum de signal possible pour effectuer la quantification. Nous décrirons dans la section suivante des méthodes issues de l'analyse factorielle qui répondent à ce cahier des charges.

3.2. Analyse factorielle

Si l'on suppose le système linéaire et reproductible, une famille d'approches possibles pour réaliser la reconstruction de profils moléculaires est basée sur l'analyse factorielle. Le but est d'extraire l'information pertinente de données physico-chimiques mesurées en construisant et en exploitant un modèle multivariable. Celui-ci relie les variables X, dont l'estimation est délicate (dans notre cas, les concentrations des protéines d'intérêt) aux variables Y facilement mesurables (dans notre cas, les données en sortie du système, les spectrogrammes). Dans ce cas, l'analyse statistique est réalisée sur l'ensemble des points du spectre MS et non sur une liste de pics extraits et aucune information *a priori* n'est introduite sur le modèle.

Deux étapes sont nécessaires. La première, appelée étalonnage, correspond à la construction du modèle mathématique à partir de mélanges d'étalonnage (X et Y sont connues). La deuxième étape, appelée prédiction, permet pour des mélanges de composition inconnue d'estimer les concentrations $\boldsymbol{X}$ à partir des données mesurées $\boldsymbol{Y}$ et du modèle construit à l'étape d'étalonnage. Les variables sont liées par la relation tensorielle suivante :

$$\underline{\boldsymbol{Y}} = \underline{\boldsymbol{H}} \times_1 \boldsymbol{X} \tag{3.2-1}$$

où $\underline{\boldsymbol{Y}}$ est le tenseur des mesures de taille (M, J, K), $\boldsymbol{X}$ la matrice des concentrations de taille (M, N), $\underline{\boldsymbol{H}}$ le tenseur d'interaction contenant les spectrogrammes théoriques des protéines cibles en concentration unitaire de taille (N, J, K) et $\times_1$ le produit n-mode [68] selon la première dimension du tenseur. J et K correspondent aux dimensions d'un spectrogramme, M au nombre de mélanges et N au nombre de protéines cibles. $\underline{\boldsymbol{H}}$ est estimé à l'étape d'étalonnage. Ce modèle peut aussi s'écrire sous forme 2D (les spectrogrammes sont dépliés en longs vecteurs) :

$$\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{X} \tag{3.2-2}$$

où $\boldsymbol{Y}$, $\boldsymbol{H}$ et $\boldsymbol{X}$ sont respectivement des matrices de taille (JK, M), (JK, N) et (N, M). En particulier, nous pouvons noter que si nous ne considérons qu'un seul mélange, c'est-à-dire que nous ne réalisons qu'une seule expérience, la relation précédente se réécrit

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} \tag{3.2-3}$$

où $\boldsymbol{y}$ est le vecteur de taille JK représentant les données d'une expérience et $\boldsymbol{x}$ est le vecteur de taille N représentant les concentrations.

3.2.1. Méthodes de prédiction par analyse factorielle

L'analyse factorielle cherche à modéliser les données par une combinaison linéaire de signaux appelés facteurs pondérés par les variables X. Plusieurs méthodes sont utilisables dans notre cas.

3.2.1.1. Unfold-Partial Least Square (Unfold-PLS)

La méthode PLS a été développée et popularisée en science analytique par Wold [69]. Unfold-PLS est identique à la régression PLS mais est appliquée à des données tensorielles réarrangées sous forme vectorielle (*cf.* équation (3.2-2)). Cette technique, proche de la méthode PCR (Principal Component

Regression), prend en compte à la fois l'information sur les concentrations $\boldsymbol{X}$ et les spectrogrammes $\boldsymbol{Y}$ lors de l'étape d'étalonnage. La méthode PLS réalise donc une extraction des vecteurs propres à partir des matrices de concentration $\boldsymbol{X}$ et des matrices de mesures $\boldsymbol{Y}$ des mélanges d'étalonnage.

Il existe plusieurs types de régression PLS. La plus simple, appelé PLS1, s'applique au cas de l'équation (3.2-3) où $\boldsymbol{x}$ est un simple vecteur (une seule protéine cible par exemple). Dans le cas de l'équation (3.2-2) où l'on recherche plusieurs protéines cibles, $\boldsymbol{X}$ correspond à une matrice et on utilise alors l'algorithme PLS2.

L'utilisation simultanée de l'information sur $\boldsymbol{X}$ et $\boldsymbol{Y}$ à l'étape d'étalonnage permet d'obtenir de meilleurs résultats de prédiction qu'avec la méthode PCR qui n'utilise que l'information sur $\boldsymbol{Y}$. Considérons des mélanges complexes de protéines pour l'étalonnage dont la matrice de concentration ne contient que l'information de concentration des protéines cibles. Si l'on réalise l'analyse en composantes principales de ce mélange, alors les composantes obtenues peuvent contenir que très peu d'information relative aux protéines cibles si celles-ci sont en minorité dans le mélange. Étant donné que la PLS recherche l'espace des facteurs le plus conforme aux variables $\boldsymbol{X}$ et $\boldsymbol{Y}$, sa prédiction est meilleure.

3.2.1.2. Méthodes multidimensionnelles

Les méthodes multidimensionnelles [70] permettent de garder la cohérence tridimensionnelle des données car elles traitent les données tensorielles sous leur forme initiale (*cf.* équation (3.2-1)). Dans notre cas, $\underline{\boldsymbol{Y}}$ est un tenseur d'ordre 3 (1ère dimension : mélange, 2ème dimension : rapport masse sur charge, 3ème dimension : temps de rétention).

N-way Partial Least Square (N-PLS)

La méthode N-PLS [71] est une extension de la méthode PLS aux tableaux d'ordre supérieur. Elle correspond à une décomposition du tenseur $\underline{\boldsymbol{Y}}$ d'étalonnage en un ensemble de tenseurs de rang 1. Les modèles obtenus sont généralement beaucoup plus simples et robustes et permettent de réduire le risque de surmodélisation par rapport aux approches de type unfold-PLS.

Parallel Factor Analysis (PARAFAC)

La décomposition PARAFAC [72] correspond à une généralisation de l'analyse en composantes principales (PCA) aux tenseurs d'ordre N. Elle fut initiée par Harshman en 1970 [73]. Cependant, certaines caractéristiques diffèrent de la décomposition à l'ordre deux. Tout d'abord, la décomposition PARAFAC est unique. De plus, à la différence de la PCA, les facteurs ne sont pas estimés successivement. Enfin, la décomposition utilise à la fois les mélanges d'étalonnage et de prédiction et ne tient pas compte de l'information contenue dans $\boldsymbol{X}$ (à la différence des méthodes PLS et N-PLS). L'avantage de ce type de méthode est qu'elle permet d'obtenir des modèles simples, robustes et facilement interprétables. Un modèle de décomposition PARAFAC d'un tenseur $\underline{\boldsymbol{Y}}$ d'ordre 3 est donné à partir de trois matrices de décomposition : $\boldsymbol{A}$, $\boldsymbol{B}$ et $\boldsymbol{C}$. Si F est le nombre de facteurs, le modèle trilinéaire minimise la somme des carrés des résidus e_{ijk} du modèle suivant :

$$y_{ijk} = \sum_{f=1}^{F} a_{ij} b_{jf} c_{kf} + e_{ijk} \qquad (4)$$

3.2.2. Conclusion

L'approche par analyse factorielle permet d'effectuer la quantification des protéines d'intérêt en cherchant à utiliser le maximum de signal possible. Pour cela, la méthode cherche préalablement à séparer les différents signaux. Cependant il s'agit d'une approche type boîte noire [74]. En effet, il s'agit d'une méthode générique qui permet de s'adapter à de nombreuses situations. Le revers de cette médaille est que la méthode ne prend pas en compte l'information dont on dispose. Les seules informations utilisées concernent la linéarité du système et l'utilisation d'expériences étalon de concentration connue. La méthode n'utilise pas de modèle direct spécifique et n'inclut pas d'information sur les paramètres recherchés.

De plus, cette méthode suppose que la matrice $\boldsymbol{H}$ du système est constante d'une expérience à l'autre. Cette hypothèse ne permet pas de prendre en compte la variation des temps de rétention chromatographiques et des gains système. La méthode nécessite donc l'application de prétraitements effectuant un recalage des différents pics et une normalisation des gains système. Ces prétraitements vont modifier artificiellement les données, ce qui peut entraîner une perte de performances.

Dans la section suivante, nous introduirons l'approche bayésienne qui permet de proposer des méthodes spécifiques à un problème d'intérêt. Nous décrirons notamment des travaux caractéristiques pour le traitement de données spectrométriques.

3.3. Inférence statistique et approche bayésienne

Toutes les méthodes précédentes utilisent la linéarité du modèle par rapport aux concentrations, mais notons qu'aucune de ces méthodes ne prend en compte de façon directe le bruit. C'est pourtant la principale source d'incertitude. Plus le bruit sera important par rapport au signal, plus l'estimation des concentrations demandera des méthodes utilisant toute l'information disponible.

Les méthodes statistiques bayésiennes sont particulièrement adaptées dans ce cas [75, 76]. D'une part, elles sont basées sur une modélisation du système et du bruit, d'autre part elles proposent un cadre formel qui permet d'inclure une grande variété d'informations dans la méthode de traitement.

La démarche employée est la suivante.

1) Le modèle direct associé à notre instrument est défini, puis le bruit est modélisé par une densité de probabilité.
2) Les informations *a priori* dont nous disposons sur les paramètres recherchés sont traduites sous forme de loi de probabilité. Il s'agit d'une étape délicate car le problème du choix des fonctions *a priori* est encore largement ouvert. Quelques règles formelles existent (principe d'invariance par reparamétrisation [77], utilisation de principes entropiques par exemple), mais les lois *a priori* sont souvent choisies de façon pragmatique de façon à coder de façon raisonnable l'information et de simplifier les calculs.
3) Nous utilisons la règle de Bayes pour calculer la probabilité *a posteriori* associée à chaque valeur des paramètres recherchés. Il ne s'agit ici que d'un simple calcul réalisé à partir du modèle direct, de la loi du bruit, et des lois *a priori* sur les paramètres. La réalisation du calcul utilise les règles de calculs sur les probabilités et, notamment, la règle du produit [48].
4) La loi de probabilité *a posteriori* résume l'information de notre problème, mais nous pouvons choisir de ne retenir que quelques valeurs représentatives de cette loi multidimensionnelle. La densité pourra être représentée par un estimateur ponctuel, représentant la meilleure valeur possible, associée à des marges d'erreurs pour nos paramètres.

Le lecteur désireux d'en savoir plus sur l'inférence bayésienne et le calcul des probabilités en général pourra consulter le cours [47] ou les livres [76, 78]. Dans un autre registre, citons également le livre d'histoire des sciences [79] qui relate la genèse du concept de probabilité au XVII^{e} siècle. Ce livre éclaire notamment les raisons de cette naissance tardive, montre les racines communes des écoles bayésiennes et des autres écoles probabilistes, et permet surtout de retrouver le visage humain qui se cache derrière le masque des formules absconses.

L'approche bayésienne a récemment été utilisée en protéomique dans le cadre de la recherche de biomarqueurs [67], mais à notre connaissance, elle n'a jamais été utilisée pour l'estimation de profils moléculaires dans des chaînes d'analyse basées sur la chromatographie et la spectrométrie de masse. Nous pouvons cependant rapprocher notre étude de travaux réalisés pour les données spectrométriques et en particulier pour la spectrométrie de masse. Toutes les méthodes que nous allons voir supposent un modèle linéaire et un bruit suivant une loi normale. Elles suivent donc le modèle suivant

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{b}$$

où le vecteur $\boldsymbol{y}$ désigne les données, $\boldsymbol{x}$ les concentrations, $\boldsymbol{b}$ le bruit et $\boldsymbol{H}$ la matrice instrument modélisant le système. Les approches suivantes peuvent être classées suivant la façon de traiter la matrice instrument.

3.3.1. Estimation linéaire des concentrations, la matrice instrument étant connue

La première approche consiste à ne rechercher que les concentrations $\boldsymbol{x}$ en considérerant que la matrice $\boldsymbol{H}$ est connue. C'est l'approche réalisée dans l'article [80]. Plusieurs lois *a priori* sont proposées afin de prendre en compte diverses informations sur les concentrations recherchées. Mais l'accent est mis sur la loi *a priori* normale. Le modèle est linéaire, le bruit suit une loi normale et l'information *a priori* sur les variables recherchées est modélisée par une loi normale. Il s'agit du cas « linéaire gaussien » classique en traitement du signal. A titre d'exemple, nous allons détailler ce cas dans les paragraphes suivants.

Dans un premier temps, la loi des données est calculée, tous les paramètres étant connus. Cette loi est obtenue grâce à un changement de variable à partir de la loi du bruit et du modèle direct.

$$p(\boldsymbol{y}|\boldsymbol{x}) \propto \exp\left(-\frac{1}{2}\gamma_b \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\|^2\right)$$

Deuxièmement, la loi *a priori* pour $\boldsymbol{x}$ est choisie : une loi normale centrée en $\bar{\boldsymbol{x}}$ et d'inverse variance γ_x.

$$p(\boldsymbol{x}) \propto \exp\left(-\frac{1}{2}\gamma_x \|\boldsymbol{x} - \bar{\boldsymbol{x}}\|^2\right)$$

Cette loi permet de traduire l'information suivante : les concentrations recherchées sont proches des valeurs nominales $\bar{\boldsymbol{x}}$ avec une certaine incertitude traduite par γ_x.

Ces deux dernières lois permettent de calculer la loi *a posteriori* grâce à la formule de Bayes.

$$p(\boldsymbol{x}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})}{p(\boldsymbol{y})} \propto \exp\left(-\frac{\gamma_b}{2}\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\|^2 - \frac{\gamma_x}{2}\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|^2\right)$$

Cette distribution étant gaussienne, l'estimateur de la moyenne, de la médiane et du maximum sont égaux, et sa valeur est donnée de façon explicite par

$$\hat{\boldsymbol{x}} = \left(\boldsymbol{H}^t\boldsymbol{H} + \frac{\gamma_x}{\gamma_b}\boldsymbol{I}\right)^{-1}\left(\boldsymbol{H}^t\boldsymbol{y} + \frac{\gamma_x}{\gamma_b}\bar{\boldsymbol{x}}\right)$$

Plus le bruit est élevé, plus nous ferons confiance à la valeur nominale. Etudions maintenant deux cas extrêmes pour l'information *a priori*. Si on ne veut pas injecter de connaissances antérieures sur $\boldsymbol{x}$, ou favoriser de valeurs particulières par rapport aux autres, l'information *a priori* sera traduite par une loi de probabilité à la variance grande. Son inverse tendra vers 0, l'estimateur sera alors égal à la solution des moindres carrés.

$$\hat{\boldsymbol{x}}_{\mathrm{MC}} = \left(\boldsymbol{H}^t\boldsymbol{H}\right)^{-1}\boldsymbol{H}^t\boldsymbol{y}$$

Si pour une raison ou une autre nos connaissances *a priori* sont extrêmement précises, nous voudrons coder cette information par une distribution rassemblée autour de la valeur nominale $\bar{\boldsymbol{x}}$. Dans le cas limite cette distribution tend vers une fonction de Dirac et la valeur γ_x tend alors vers $+\infty$. Cela revient à fixer la valeur de l'estimateur à la valeur nominale. Il s'agit d'un caractère général des méthodes bayésiennes. Elles permettent d'introduire une information *a priori* sur les paramètres recherchés. Mais nous pouvons utiliser la même méthode dans les deux cas extrêmes : celui où on ne souhaite pas introduire d'information et celui où on souhaite fixer un des paramètres.

Cet estimateur a l'avantage de se calculer facilement grâce à sa structure linéaire. De manière générale, il est donc avantageux de se placer dans le cas linéaire gaussien. Cependant, il s'agit d'une méthode un peu fruste : le modèle doit être linéaire, la matrice instrument entièrement connue et les informations *a priori* doivent se limiter aux caractéristiques statistiques d'ordre 2.

3.3.2. Estimation conjointe des concentrations et de la matrice instrument

L'approche décrite la section précédente nécessite deux étapes comme celle de l'analyse factorielle. Dans un premier temps, on estime la matrice instrument à l'aide d'expériences d'étalonnage où les concentrations sont supposées parfaitement connues et, dans un second temps, on estime les concentrations d'intérêt à l'aide de la matrice instrument obtenue à l'étape précédente. Cependant, les concentrations d'étalonnage sont rarement connues parfaitement. Une approche issue de la séparation de source est présentée dans [81, 82], elle propose d'estimer conjointement les concentrations et la matrice instrument à l'aide d'une approche bayésienne. Cette dernière permet notamment à l'auteur de prendre en compte la non-négativité des concentrations et des coefficients de la matrice.

3.3.3. Variation des paramètres instrument et marginalisation

Jusqu'à présent, nous avons supposé que la matrice instrument est constante d'une expérience à l'autre. Or la matrice $\boldsymbol{H}$ peut varier d'une expérience à l'autre. Dans notre cas ce sont les gains ξ_i et les positions des pics chromatographiques qui peuvent notamment varier [21]. Dans [83], les auteurs prennent en compte une gamme de variation possible pour les éléments de la matrice $\boldsymbol{H}$. Ils n'utilisent pas la piste d'une estimation conjointe, car seules les concentrations les intéressent. L'approche bayésienne propose dans ce cas d'écarter le problème de l'estimation de $\boldsymbol{H}$ en marginalisant ce paramètre. En pratique, cela revient à intégrer la loi *a posteriori* par rapport aux éléments de $\boldsymbol{H}$.

Si cette approche permet de prendre en compte la méconnaissance de la matrice instrument, l'information *a priori* utilisée n'est pas la plus adaptée à notre problème. Nous pouvons être plus spécifique et plus performant en plaçant l'incertitude sur la position des pics, et non sur la valeur des coefficients de la matrice qui est générée à partir de ces positions.

3.3.4. Approche paramétrique

Les méthodes précédentes ne prennent pas en compte de modèle direct explicite, il s'agit de méthodes non paramétriques. Elles ne cherchent pas à imposer une forme particulière à la matrice instrument. Une seconde approche consiste à diminuer le nombre d'inconnues en proposant une expression analytique de la matrice instrument. On peut ainsi espérer obtenir une méthode plus robuste car correspondant mieux au problème posé.

Dans sa thèse de doctorat [84], Vincent Mazet propose de modéliser le signal par une somme de fonctions lorentziennes. Son but n'est pas d'estimer la quantité d'espèces spécifiques, mais de reconstruire un spectre impulsionnel à partir des mesures. Dans un premier temps, il procède par déconvolution impulsionnelle myope où toutes les fonctions lorentziennes ont la même largeur. Dans un second temps, il décompose le signal sur une famille de fonctions (les largeurs des lorentziennes sont ajustées séparément). Ces méthodes fournissent en sortie la position et l'amplitude de chaque pic. L'identification des espèces présentes et de leurs concentrations n'est pas incluse dans le travail.

Cette approche se classe dans les approches d'inversion myope où l'on estime de façon conjointe les variables d'intérêt et les paramètres instrument. Cependant, l'approche se distingue car elle utilise une forme paramétrique du signal. D'autres informations sont injectées grâce au cadre probabiliste de l'approche bayésienne, notamment la parcimonie du signal et la non-négativité du spectre. Plus précisément, le signal sera expliqué avec un nombre de pics minimum et leurs amplitudes doivent être positives.

Si cette méthode utilise la forme des pics, les liens entre les différents pics d'un signal ne sont pas exploités. De plus, cette méthode, comme les précédentes, ne traite pas directement notre problème. Une méthode plus proche de notre problème doit donc être développée. Toutefois, notons que ce

travail a été une source importante d'inspiration pour la méthode d'inversion proposée dans cette thèse.

Notons qu'il existe une distinction importante entre le problème traité par V. Mazet et le notre. Comme nous l'avons dit plus haut, la méthode de V. Mazet cherchera à expliquer le signal par un nombre de pics minimum. Le nombre « d'objets » que l'on doit traiter est inconnu. Dans notre cas, le nombre de protéines recherchées est connu et ne doit pas être estimé. Bien sûr, sur le spectrogramme total, des signaux de contaminants se superposent aux signaux d'intérêt. Comme nous connaissons les positions des pics d'intérêt, nous pouvons en première approximation, conserver uniquement la zone du spectrogramme contenant le signal d'intérêt, et ignorer les signaux des contaminants. C'est l'approche réalisée dans cette thèse. Dans une version ultérieure, pour gagner en performance, les signaux des contaminants devront être détectés et traités. Les travaux de V. Mazet pourront servir de base à ce travail.

3.3.5. Conclusion

Nous avons présenté dans cette section la démarche bayésienne et décrit des méthodes l'appliquant à l'analyse de spectres. La démarche bayésienne permet la prise en compte du modèle direct et du bruit. De plus, elle permet d'introduire un grand nombre d'informations *via* les lois *a priori*. De nombreuses approches ont été proposées. Notons toutefois que les approches sont souvent génériques et n'utilisent pas toute l'information que nous avons mise à jour pour notre application. De nombreux travaux supposent que les paramètres instruments sont constants d'une expérience à l'autre. Ils ne peuvent donc pas prendre en compte la variation des paramètres de la matrice instrument. Si certains proposent d'estimer ces paramètres à l'aide d'une démarche d'inversion myope, ils ne prennent pas en compte la forme exacte du problème.

3.4. Bilan

Nous avons évoqué dans ce chapitre les différentes méthodes existantes pour quantifier les protéines dans des données LC-MS. Si de nombreuses méthodes utilisant un étalonnage isotopique ont été développées, ces méthodes sont la plupart du temps bâties sur l'hypothèse de la séparation des pics par rapport à leur environnement. Cette hypothèse ne sera plus valide en présence d'un faible rapport signal sur bruit ou de contaminant proche entraînant une superposition des signaux.

D'autres méthodes doivent donc être développées afin de dépasser cette limitation. Une des approches principales consiste à ne plus découper les différentes étapes du traitement. Une première solution est proposée par l'emploi de méthodes issues de l'analyse factorielle. Les méthodes PLS et N-PLS par exemple effectuent de façon conjointe d'une par la séparation et la détection dans l'étape d'étalonnage et d'autre part, la séparation et la quantification des signaux dans l'étape de prédiction. Si cette méthode permet en théorie d'utiliser la totalité du signal utile pour effectuer la quantification, elle suppose que les paramètres du système ne varient pas d'une expérience à l'autre. La méthode doit donc utiliser un ensemble de prétraitement afin « d'idéaliser » les données.

Pour aller plus loin dans cette optique, l'approche bayésienne semble tout adaptée, elle permet en effet de traiter les variations des différents paramètres, mais également de s'affranchir de l'aspect « boîte noire » des méthodes précédentes. Cette dernière permet donc de développer des méthodes dédiées à notre problème. Si des méthodes utilisant cette approche ont déjà été développées pour traiter les données spectrométriques, aucune ne prend en compte le modèle que nous avons mis à jour.

Nous proposons dans le chapitre suivant une méthode utilisant le modèle que nous avons développé dans le dernier chapitre. Elle supposera l'utilisation de protéines étalon, et sera robuste aux variations des gains et des temps de rétention. Le cadre bayésien dans lequel elle sera développée permettra de prendre en compte de l'information *a priori* supplémentaire sur les paramètres recherchés.

4. Inversion

Ce chapitre présente notre méthode d'estimation des concentrations, variables d'entrée du système à partir des variables de sortie. Nous résolvons ici le problème inverse, c'est-à-dire nous caractérisons les causes à partir de leurs effets. Par opposition, les équations du modèle direct présentées au chapitre 2 permettent d'obtenir les effets lorsque les causes sont connues.

Afin de résoudre au mieux ce problème d'inversion, la méthode utilisera le modèle direct et le modèle du bruit. De plus, elle prend en compte les incertitudes de certains paramètres instrument pour améliorer l'estimée des variables d'entrée. Pour cela, nous estimons ces paramètres instrument conjointement aux variables d'entrée.

Dans les paragraphes suivants, les paramètres du modèle seront listés. Nous indiquerons, parmi ces paramètres, ceux dont l'estimation nous semble nécessiter les efforts les plus grands et ceux dont l'estimation peut être obtenue de manière préalable. Après avoir fixé la valeur de ces paramètres plus secondaires, nous présenterons notre méthode d'estimation conjointe.

4.1. Modèle

Le modèle retenu au chapitre 2 pour la chaîne de mesure s'écrit

$$\boldsymbol{Y} = \sum_{p=1}^{P}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=0}^{K} x_p d_{ip} \xi_i \pi_{ij} \pi'_{ijk} \, \boldsymbol{s}_{ijk} \boldsymbol{c}_i^t + \boldsymbol{B}$$

Cette relation indique que l'image en sortie de système $\boldsymbol{Y}$ est formée par la somme de signaux élémentaires décrivant un ion. Cette somme est indicée par p, i, j et k, décrivant respectivement la protéine parente, la formule peptidique de l'ion, le nombre de charges portées par l'ion et le nombre de neutrons supplémentaires par rapport à la configuration la plus stable. Chaque signal bidimensionnel élémentaire se décompose comme le produit de deux vecteurs $\boldsymbol{s}_{ijk}$ et $\boldsymbol{c}_i$ décrivant respectivement l'influence du spectromètre de masse et de la colonne de chromatographie.

$$\boldsymbol{s}_{ijk} = \left[h_{ijk}\left(m_0^s\right) \quad h_{ijk}\left(m_0^s + M_e^s\right) \quad \ldots \quad h_{ijk}\left(m_0^s + (N_s - 1)M_e^s\right) \right]^t$$
$$\boldsymbol{c}_i = \left[g_i(0) \quad g_i\left(T_e^c\right) \quad \ldots \quad g_i\left((N_c - 1)T_e^c\right) \right]^t$$

Ces deux vecteurs représentent les gaussiennes suivantes

$$h_{ijk}(m) = \left(2\pi\gamma_s^{-1}\right)^{-1/2} \exp\left(-\frac{1}{2}\gamma_s \left(m - m_{ijk}\right)^2 \right)$$

$$g_i(t) = \left(2\pi\gamma_c^{-1}\right)^{-1/2} \exp\left(-\frac{1}{2}\gamma_c \left(t - t_i\right)^2 \right)$$

Le modèle est complété par un bruit additif $\boldsymbol{B}$ de densité de probabilité

$$p\left(\boldsymbol{B}|\gamma_b\right) = \left(2\pi\gamma_b^{-1}\right)^{\frac{-N_c N_s}{2}} \exp\left(-\frac{1}{2}\gamma_b \|\boldsymbol{B}\|^2 \right)$$

Nous complétons ce modèle en incluant le système d'étalonnage interne par protéines marquées isotopiquement présenté au chapitre 3. Ces protéines marquées ne diffèrent des protéines d'intérêt que par leur nombre de neutrons, ce qui ne modifie pas les phénomènes physico-chimiques étudiés au chapitre 2. Les seules variables du modèle impactées par ce nombre de neutrons sont les positions des gaussiennes spectrométriques, les proportions de neutrons supplémentaires et les concentrations de protéines marquées, notées respectivement m^*_{ijk}, π'^*_{ijk} et x^*_p.

En incluant les molécules marquées, le modèle se réécrit

$$\boldsymbol{Y} = \sum_{p=1}^{P}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=0}^{K} d_{ip}\xi_i\pi_{ij}\left(x_p\pi'_{ijk}\boldsymbol{s}_{ijk} + x^*_p\pi'^*_{ijk}\boldsymbol{s}^*_{ijk}\right)\boldsymbol{c}_i^t + \boldsymbol{B}$$

avec $\boldsymbol{s}^*_{ijk}$ les vecteurs spectrométriques générés à partir des nouvelles positions m^*_{ijk}. Finalement, cet ensemble d'équations dépend des paramètres suivants :

- les paramètres d'intérêt : x_p, les concentrations des protéines, leur nombre est connu,
- les paramètre d'étalonnage : x^*_p, les concentrations des protéines marquées connues,
- les paramètres instrument
  - les paramètres du bruit : γ_b, l'inverse variance du bruit,
  - les paramètres des gaussiennes décrivant le signal
    - t_i et γ_c, les positions et la largeur des gaussiennes chromatographiques,
    - m_{ijk}, m^*_{ijk} et γ_s, les positions et la largeur des gaussiennes spectrométriques,
  - les gains
    - d_{ip}, le gain de digestion ou le nombre de peptide *i* généré par la protéine *p*,
    - ξ_i, le gain du système pour le peptide *i*,
    - π_{ij}, la proportion du peptide *i* ayant *j* charges,
    - π'_{ijk} et π'^*_{ijk}, les proportions du peptide *i* ayant *j* charges et *k* neutrons supplémentaires.

Tous ces paramètres instrumentaux influencent la valeur des paramètres d'intérêt. De plus, ils sont tous connus de façon plus ou moins précise. Dans un premier temps, nous nous concentrerons sur les paramètres dont l'estimation nous semble la plus cruciale : x_p, ξ_i, t_i, γ_b. Nous utiliserons pour cela le cadre de l'inférence bayésienne. Les autres paramètres nécessitent moins d'attention. Nous les obtenons au moyen de méthodes plus empiriques détaillées dans la section 4.2. Ce choix n'est pas imposé par la méthodologie, mais pour des raisons plus pratiques. Augmenter le nombre de paramètres ou leur incertitude, sans rajouter d'information par ailleurs, diminue la qualité des estimées en augmentant leurs marges d'erreur. De plus, il est normal de commencer l'étude sur un petit nombre de paramètres, puis d'augmenter leur nombre, accroissant ainsi progressivement la difficulté du problème.

Les paramètres sur lesquels nous allons nous concentrer étant déterminés, commençons par les mettre en évidence par une notation adéquate

$$\boldsymbol{Y} = \Theta\left(\boldsymbol{x}, \boldsymbol{x}^*, \boldsymbol{\xi}, \boldsymbol{t}\right) + \boldsymbol{B}$$

avec

$$\Theta\left(\boldsymbol{x}, \boldsymbol{x}^*, \boldsymbol{\xi}, \boldsymbol{t}\right) = \sum_{p=1}^{P}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=0}^{K} d_{ip}\xi_i\pi_{ij}\left(x_p\pi'_{ijk}\boldsymbol{s}_{ijk} + x^*_p\pi'^*_{ijk}\boldsymbol{s}^*_{ijk}\right)\boldsymbol{c}_i^t$$

et les variables x_p, x_p^*, ξ_i et t_i respectivement regroupées dans les vecteurs $\boldsymbol{x}$, $\boldsymbol{x}^*$, $\boldsymbol{\xi}$ et $\boldsymbol{t}$. Rappelons que le vecteur $\boldsymbol{c}_i$ dépend de t_i et que la densité de probabilité de $\boldsymbol{B}$ dépend de γ_b. De plus, notons que $\boldsymbol{Y}$ et $\boldsymbol{B}$ sont de dimension $N_s \times N_c$, $\boldsymbol{x}$ et $\boldsymbol{x}^*$ sont de dimension P et $\boldsymbol{\xi}$ et $\boldsymbol{t}$ sont de dimension I.

La fonction Θ nous permet d'alléger les notations. De plus, nous pouvons souligner les relations linéaires liant $\boldsymbol{\xi}$, $\boldsymbol{x}$ et $\boldsymbol{x}^*$ aux données $\boldsymbol{Y}$ sous forme matricielle. Pour cela, nous introduisons la matrice $\boldsymbol{G}$ de dimension $N_sN_c \times I$ et les matrices $\boldsymbol{H}$ et $\boldsymbol{H}^*$ de dimension $N_sN_c \times P$. Ces matrices nous permettent de réécrire le modèle précédent sous la forme suivante afin d'isoler la contribution de la variable d'intérêt

$$\boldsymbol{y} = \boldsymbol{G}\boldsymbol{\xi} + \boldsymbol{b} \quad \text{et} \quad \boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{H}^*\boldsymbol{x}^* + \boldsymbol{b}$$

avec les vecteurs $\boldsymbol{y}$ et $\boldsymbol{b}$ de dimension N_sN_c respectivement formés par la concaténation des vecteurs colonnes des matrices $\boldsymbol{Y}$ et $\boldsymbol{B}$. Les éléments des matrices $\boldsymbol{G}$, $\boldsymbol{H}$ et $\boldsymbol{H}^*$ dépendent des paramètres du modèle. Or les expressions décrivant les éléments de ces matrices sont en grande partie similaires. Ces éléments communs seront décrits à travers de nouvelles matrices $\boldsymbol{F}$ et $\boldsymbol{F}^*$ de dimension $N_sN_c \times I$. $\boldsymbol{F}$ et $\boldsymbol{F}^*$ correspond à la matrice système des signatures des peptides d'intérêt et des peptides marqués. Les colonnes numéro i des matrices $\boldsymbol{F}$ et $\boldsymbol{F}^*$ sont formées respectivement par la concaténation des vecteurs colonnes des matrices $\boldsymbol{F}_i$ et $\boldsymbol{F}_i^*$. Ces dernières correspondent à la signature de chaque peptide i, sont de dimension $N_s \times N_c$ et sont définies par

$$\boldsymbol{F}_i = \sum_{j=1}^{J}\sum_{k=0}^{K} \pi_{ij}\pi'_{ijk}\boldsymbol{s}_{ijk}\boldsymbol{c}_i^t \quad \text{et} \quad \boldsymbol{F}_i^* = \sum_{j=1}^{J}\sum_{k=0}^{K} \pi_{ij}\pi'^*_{ijk}\boldsymbol{s}^*_{ijk}\boldsymbol{c}_i^t$$

Les expressions des matrices $\boldsymbol{G}$, $\boldsymbol{H}$ et $\boldsymbol{H}^*$ faisant intervenir les matrices $\boldsymbol{F}$ et $\boldsymbol{F}^*$ se déduisent de l'expression de Θ, en réorganisant les différentes sommes

$$\Theta(\boldsymbol{x}, \boldsymbol{x}^*, \boldsymbol{\xi}, \boldsymbol{t}) = \underbrace{\sum_{p=1}^{P}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=0}^{K} x_p d_{ip}\xi_i\pi_{ij}\pi'_{ijk}\boldsymbol{s}_{ijk}\boldsymbol{c}_i^t}_{\boldsymbol{E}_1} + \underbrace{\sum_{p=1}^{P}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=0}^{K} x^*_p d_{ip}\xi_i\pi_{ij}\pi'^*_{ijk}\boldsymbol{s}^*_{ijk}\boldsymbol{c}_i^t}_{\boldsymbol{E}_2}$$

avec

$$\boldsymbol{E}_1 = \sum_{i=1}^{I} \underbrace{\left(\sum_{j=1}^{J}\sum_{k=0}^{K} \pi_{ij}\pi'_{ijk}\boldsymbol{s}_{ijk}\boldsymbol{c}_i^t\right)}_{\boldsymbol{F}_i} \underbrace{\left(\sum_{p=1}^{P} d_{ip}x_p\right)}_{(\mathrm{diag}(\boldsymbol{D}\boldsymbol{x}))_{i,i}} \xi_i$$

$$\boldsymbol{E}_2 = \sum_{i=1}^{I} \underbrace{\left(\sum_{j=1}^{J}\sum_{k=0}^{K} \pi_{ij}\pi'^*_{ijk}\boldsymbol{s}^*_{ijk}\boldsymbol{c}_i^t\right)}_{\boldsymbol{F}_i^*} \underbrace{\left(\sum_{p=1}^{P} d_{ip}x^*_p\right)}_{\left(\mathrm{diag}\left(\boldsymbol{D}\boldsymbol{x}^*\right)\right)_{i,i}} \xi_i$$

D'où

$$\boldsymbol{G} = \boldsymbol{F}\,\mathrm{diag}(\boldsymbol{D}\boldsymbol{x}) + \boldsymbol{F}^*\,\mathrm{diag}\left(\boldsymbol{D}\boldsymbol{x}^*\right)$$

De même, en récrivant ces expressions pour mettre en valeur les concentrations, on a

$$\boldsymbol{E}_1 = \sum_{p=1}^{P}\left(\sum_{i=1}^{I} \underbrace{\left(\sum_{j=1}^{J}\sum_{k=0}^{K} \pi_{ij}\pi'_{ijk}\boldsymbol{s}_{ijk}\boldsymbol{c}_i^t\right)}_{\boldsymbol{F}_i} \underbrace{\left(\xi_i d_{ip}\right)}_{(\mathrm{diag}(\xi)\boldsymbol{D})_{i,p}} x_p\right)$$

$$\boldsymbol{E}_2 = \sum_{p=1}^{P}\left(\sum_{i=1}^{I} \underbrace{\left(\sum_{j=1}^{J}\sum_{k=0}^{K} \pi_{ij}\pi'^*_{ijk}\boldsymbol{s}^*_{ijk}\boldsymbol{c}_i^t\right)}_{\boldsymbol{F}_i^*} \underbrace{\left(\xi_i d_{ip}\right)}_{(\mathrm{diag}(\xi)\boldsymbol{D})_{i,p}} x^*_p\right)$$

D'où

$$\boldsymbol{H} = \boldsymbol{F}\,\mathrm{diag}(\boldsymbol{\xi})\boldsymbol{D} \quad \text{et} \quad \boldsymbol{H}^* = \boldsymbol{F}^*\,\mathrm{diag}(\boldsymbol{\xi})\boldsymbol{D}$$

où diag($\boldsymbol{v}$) est la matrice diagonale dont les éléments diagonaux sont les éléments du vecteurs $\boldsymbol{v}$ et $\boldsymbol{D}$ la matrice de digestion de dimension $I \times P$ formée des éléments d_{ip}.

Avant de décrire la méthode, examinons notre problème. Les matrices $\boldsymbol{G}$ et $\boldsymbol{H}$ expriment les relations linéaires entre $\boldsymbol{\xi}$ et $\boldsymbol{Y}$ d'une part et $\boldsymbol{x}$ et $\boldsymbol{Y}$ d'autre part. Fixons les paramètres instrument, et considérons l'estimation de $\boldsymbol{x}$. La loi associée à $\boldsymbol{b}$ est une loi normale et si nous associons à $\boldsymbol{x}$ une loi *a priori* normale, nous pouvons donc nous placer dans le cas « linéaire gaussien » classique en traitement du signal pour estimer nos inconnues (section 3.3.1 page 55). Fixons maintenant $\boldsymbol{x}$ et tous les paramètres instrument excepté $\boldsymbol{\xi}$. Comme le montre la relation utilisant la matrice $\boldsymbol{G}$, ce problème peut également se ramener au cas « linéaire gaussien ». Notre cas est cependant plus complexe car nous estimons conjointement $\boldsymbol{x}$ et $\boldsymbol{\xi}$. Nous nous retrouvons vis-à-vis de ces deux paramètres, dans un cas bilinéaire plus complexe à traiter. Et de plus, le problème est non linéaire vis-à-vis des paramètres $\boldsymbol{t}$ et γ_b. L'estimation conjointe de ces paramètres est difficile, et les méthodes standard d'estimation ne peuvent pas être utilisées.

Intéressons nous maintenant à la norme des erreurs de modélisation. En effet, quelque soit la méthode utilisée pour estimer les paramètres, cette dernière joue un rôle important car elle permet de mesurer l'éloignement entre le modèle et les données. Son calcul interviendra de nombreuses fois dans notre méthode. Nous nous intéressons ici à la meilleure façon de la calculer. Pour exprimer cette norme, nous proposons la notation suivante

$$\chi(\boldsymbol{x}, \boldsymbol{x}^*, \boldsymbol{\xi}, \boldsymbol{t}) = \left\| \boldsymbol{Y} - \Theta(\boldsymbol{x}, \boldsymbol{x}^*, \boldsymbol{\xi}, \boldsymbol{t}) \right\|^2$$

Remarquons qu'il s'agit également de la norme du bruit. Or à la section 2.7 page 44, nous avions remarqué que $\|\boldsymbol{B}\| = \|\boldsymbol{b}\|$. Nous pouvons donc également utiliser les matrices $\boldsymbol{G}$, $\boldsymbol{H}$ et $\boldsymbol{H}^*$ définies ci-dessus pour calculer χ.

$$\begin{aligned} \chi(\boldsymbol{x}, \boldsymbol{x}^*, \boldsymbol{\xi}, \boldsymbol{t}) &= \left\| \boldsymbol{y} - \boldsymbol{H}\boldsymbol{x} - \boldsymbol{H}^*\boldsymbol{x}^* \right\|^2 \\ &= \left\| \boldsymbol{y} - \boldsymbol{G}\boldsymbol{\xi} \right\|^2 \end{aligned}$$

En raison de la taille conséquente des données, et ceci pour les trois expressions précédentes, le calcul direct de la norme de l'erreur de modélisation peut être coûteux en espace mémoire et en temps de calcul. En effet les matrices considérées occupent plusieurs centaines de mégaoctets. Nous proposons donc une méthode de calcul rapide de la norme faisant intervenir les matrices de petite taille $\boldsymbol{G}^t\boldsymbol{G}$, $\boldsymbol{G}^t\boldsymbol{y}$, $\boldsymbol{H}^t\boldsymbol{H}$, $\boldsymbol{H}^t\boldsymbol{y}$, $\boldsymbol{H}^{*t}\boldsymbol{H}^*$, $\boldsymbol{H}^t\boldsymbol{H}^*$, $\boldsymbol{H}^{*t}\boldsymbol{H}^*$ et $\boldsymbol{H}^{*t}\boldsymbol{y}$ (*cf.* section 7.2 p. 109).

Ces matrices interviennent en développant le calcul de la norme, car pour un vecteur $\boldsymbol{v}$, $\|\boldsymbol{v}\|^2 = \boldsymbol{v}^t\boldsymbol{v}$.

$$\begin{aligned} \chi(\boldsymbol{x}, \boldsymbol{x}^*, \boldsymbol{\xi}, \boldsymbol{t}) &= \boldsymbol{y}^t\boldsymbol{y} - 2\boldsymbol{\xi}^t\boldsymbol{G}^t\boldsymbol{y} + \boldsymbol{\xi}^t\boldsymbol{G}^t\boldsymbol{G}\boldsymbol{\xi} \\ \chi(\boldsymbol{x}, \boldsymbol{x}^*, \boldsymbol{\xi}, \boldsymbol{t}) &= \left(\boldsymbol{y}^t\boldsymbol{y} - 2\boldsymbol{x}^{*t}\boldsymbol{H}^{*t}\boldsymbol{y} + \boldsymbol{x}^{*t}\boldsymbol{H}^{*t}\boldsymbol{H}^*\boldsymbol{x}^*\right) - 2\boldsymbol{x}^t\left(\boldsymbol{H}^t\boldsymbol{y} - \boldsymbol{H}^t\boldsymbol{H}^*\boldsymbol{x}^*\right) + \boldsymbol{x}^t\boldsymbol{H}^t\boldsymbol{H}\boldsymbol{x} \end{aligned}$$

Nous pouvons utiliser l'une ou l'autre de ces formes pour calculer efficacement cette norme, en choisissant la solution la moins coûteuse algorithmiquement. Pour nos algorithmes, nous avons utilisé l'expression faisant intervenir la matrice $\boldsymbol{H}$.

Nous avons développé une méthode de calcul rapide des matrices $\boldsymbol{G}^t\boldsymbol{G}$, $\boldsymbol{G}^t\boldsymbol{y}$, $\boldsymbol{H}^t\boldsymbol{H}$, $\boldsymbol{H}^t\boldsymbol{y}$, $\boldsymbol{H}^{*t}\boldsymbol{H}^*$, $\boldsymbol{H}^t\boldsymbol{H}^*$, $\boldsymbol{H}^{*t}\boldsymbol{H}^*$ et $\boldsymbol{H}^{*t}\boldsymbol{y}$. Pour cela, nous nous appuyons dans un premier temps sur la structure séparable du modèle puis dans un second temps sur la forme gaussienne des pics. Cette méthode est détaillée à l'annexe 7.2 page 109.

Dans cette section nous avons proposé plusieurs réécritures du modèle. Avant de rentrer dans les détails de cette méthode, indiquons comment nous avons estimé les paramètres secondaires.

4.2. Estimation des paramètres secondaires

Les paramètres secondaires sont choisis de la façon suivante. Il s'agit soit des paramètres dont la valeur est connue, soit des paramètres dont la valeur est stable et qui peuvent donc être estimés par une expérience d'étalonnage préalable. Nous décrirons dans les paragraphes suivants les méthodes utilisées pour les obtenir.

- *Gain de digestion* d_{ip}. Il s'agit du nombre de peptides générés pour chaque protéine (section 2.1 page 18). Pour l'obtenir, nous comptons le nombre d'occurrence de la séquence d'acides aminés du peptide i dans la séquence de la protéine p. Nous avons utilisé la base UniProt [85] pour obtenir les séquences des protéines.

- *Position des gaussiennes spectrométriques* m_{ijk}. Nous avons vu au paragraphe 2.6 page 42 que ces positions sont obtenues grâce à la formule suivante

$$m_{ijk} = \frac{\overline{M}_i + j + k}{j}$$

où $\overline{M}_i$ est la masse du peptide i sans neutron ni proton supplémentaire. Elle se calcule à partir de la formule du peptide. Dans cette étude, nous avons utilisé le logiciel « PIR Molecular weight calculator » [46].

- *Proportions*. Comme nous l'avons vu précédemment, le signal spectrométrique d'un peptide i est formé de plusieurs gaussiennes. Dans un premier temps, nous nous intéresserons uniquement à la gaussienne la plus importante, négligeant ainsi l'influence des autres gaussiennes. Nous utiliserons l'hypothèse proposée à la section 2.6 page 42.

$$\pi_{ij} = \delta_{j;\bar{j}_i} \text{ et } \pi'_{ijk} = \delta_{k;0}$$

où $\delta_{a;b}$ est le symbole de Kronecker. La valeur $\bar{j}_i$ sera identifiée sur les données d'une expérience. Pour cela, nous utilisons le fait que $\bar{j}_i$ désigne le plus grand pic d'un peptide i. Dans les expériences que nous avons réalisées, nous avons calculé les m_{ijk} possibles pour j allant de 1 à 4, puis confronté les valeurs aux positions des pics les plus importants, $\bar{j}_i$ est obtenu par simple identification. D'autres démarches sont possibles, en utilisant le fait que l'espace entre deux pics d'un massif isotopique est égal à $1/j$ Da (section 2.4.3.2 page 36). Dans un second temps, nous pourrons mesurer les hauteurs ou les volumes relatifs qui correspondent directement aux proportions. Nous avons mesuré la hauteur maximum a_{ijk} de chaque pic et a^*_{ijk} le maximum du pic calibrant correspondant.

Les proportions de neutrons supplémentaires s'obtiennent de la façon suivante

$$\pi'_{ijk} = \frac{a_{ijk}}{\sum_{k=1}^{K} a_{ijk}} \text{ et } \pi'^*_{ijk} = \frac{a^*_{ijk}}{\sum_{k=1}^{K} a^*_{ijk}}$$

D'autres pistes sont envisageables. En effet, ces proportions seraient calculables à partir des proportions des isotopes de chaque élément (voir page 95 de [53]), mais nous n'avons pas employé ces méthodes dans le cadre de la thèse.

- *Largeurs chromatographiques et spectrométriques* γ_c *et* γ_s. Elles seront mesurées de façon classique à partir de la largeur à mi-hauteur mesurées sur les courbes. Plus précisément, nous avons recherché le maximum de chaque pic, puis travaillé sur les coupes parallèles à chaque axe passant par ce maximum. Après avoir estimé la hauteur a_{ijk} des pics, nous avons tracé sur chaque coupe la droite d'équation $y=a_{ijk}/2$, relevé l'abscisse des points d'intersection de cette droite et des pics. S'il existe plus de deux points d'intersection, un choix arbitraire est réalisé. Puis nous avons calculé la moyenne des largeurs de chaque pic.

Les paramètres secondaires étant estimés, leur valeur ne sera plus remise en cause dans les sections suivantes. Nous allons maintenant décrire le cœur de la méthode, l'estimation des variables $\boldsymbol{x}$, $\boldsymbol{\xi}$, $\boldsymbol{t}$ et γ_b.

4.3. Estimation paramétrique bayésienne

Nous utilisons l'inférence bayésienne pour estimer ces paramètres. A partir d'un jeu de données et de nos connaissances sur l'expérience, elle permet d'estimer le degré de confiance associé à chaque valeur sous forme de densité de probabilité *a posteriori*. Pour cela, commençons par calculer la loi de probabilité associée aux données et à l'ensemble des paramètres recherchés.

4.3.1. Loi jointe et calcul des autres lois

La densité jointe totale est la loi centrale qui permet d'effectuer les calculs de probabilité. Elle résume l'information disponible sur notre problème en traduisant la connaissance ou la méconnaissance des paramètres étudiés. Son expression est obtenue en séparant les informations données par l'instrument des informations connues par ailleurs. Le calcul correspondant est obtenu par application de la règle du produit

$$p\left(\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{t},\gamma_b,\boldsymbol{Y}\right)= p\left(\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{t},\gamma_b\right)p\left(\boldsymbol{Y}\middle|\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{t},\gamma_b\right)$$

La signification de ces différentes densités de probabilité est donnée ci-dessous.

- $p\left(\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{t},\gamma_b,\boldsymbol{Y}\right)$ la densité jointe des paramètres et des données.
- $p\left(\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{t},\gamma_b\right)$ est la densité de probabilité conjointe *a priori*. Elle traduit la confiance que l'on associe à chacune des valeurs possibles des paramètres et ceci avant que l'expérience soit réalisée.
- $p\left(\boldsymbol{Y}\middle|\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{t},\gamma_b\right)$ est la densité de probabilité des données, les paramètres étant fixés, si on la considère comme une fonction des données ou la fonction de vraisemblance si on la considère comme une fonction des paramètres. Elle traduit la part d'information apportée par les données dans la loi jointe.

Ces deux dernières densités peuvent être obtenues directement et seront précisées dans les deux prochaines sections. La section 4.3.2 indique comment trouver la fonction de vraisemblance à partir de la loi du bruit. La section 4.3.3 indique comment nous pouvons traduire nos informations *a priori* en lois de probabilité.

Une fois l'expression de cette loi connue, nous pouvons déduire toutes les autres lois par marginalisation ou en appliquant la règle du produit. A titre d'exemple voici comment est obtenue la loi des données $p(\boldsymbol{Y})$. Cette loi interviendra dans plusieurs expressions en tant que coefficient de normalisation. Nous la calculons à partir de la loi jointe par marginalisation des autres paramètres.

$$\begin{aligned} p(\boldsymbol{Y}) &= \int p\left(\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{t},\gamma_b,\boldsymbol{Y}\right)\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{\xi}\,\mathrm{d}\boldsymbol{t}\,\mathrm{d}\gamma_b \\ &= \int p\left(\boldsymbol{Y}\middle|\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{t},\gamma_b\right)p\left(\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{t},\gamma_b\right)\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{\xi}\,\mathrm{d}\boldsymbol{t}\,\mathrm{d}\gamma_b \end{aligned}$$

Notons que la loi des données ne dépend d'aucun paramètre.

Les principes évoqués ci-dessus ont un caractère général et peuvent s'appliquer à d'autres problèmes d'estimation et plus généralement à tout problème faisant intervenir les probabilités. Les sections suivantes seront spécifiques à notre problème. Nous y décrirons comment nous avons traduit notre connaissance du problème en loi de probabilité.

4.3.2. Vraisemblance

Cette fonction est issue de la densité des données, les paramètres étant fixés. Elle est définie par changement de variable à partir de la densité de probabilité du bruit présenté à la section 2.7 page 44 et du modèle.

$$L(\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{t},\gamma_b) = p(\boldsymbol{Y}|\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{t},\gamma_b) = (2\pi\gamma_b^{-1})^{-N_c N_s/2} \exp\left(-\frac{1}{2}\gamma_b \left\|\boldsymbol{Y} - \Theta(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t})\right\|^2\right)$$
$$= (2\pi\gamma_b^{-1})^{-N_c N_s/2} \exp\left(-\frac{\gamma_b}{2}\chi(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t})\right)$$

Cette loi est essentiellement formée autour d'une norme traduisant un terme d'adéquation du modèle aux données. Plus le modèle est proche des données, plus la valeur de la fonction de vraisemblance est élevée. Cette fonction fait le lien avec d'autres méthodes d'estimation paramétrique. Citons par exemple l'estimation au sens des moindres carrés et le maximum de vraisemblance. Dans la première, on minimise la norme, dans la seconde on cherche le maximum de cette loi. Par rapport à ces méthodes notre méthode propose à l'utilisateur d'intégrer les informations dont il dispose *a priori*, cette information supplémentaire sera décrite par la loi suivante.

4.3.3. Loi a priori

La loi *a priori* décrit l'information disponible sur les paramètres avant que l'expérience soit réalisée. Pour définir cette loi de probabilité, nous pouvons encore une fois utiliser la règle du produit et ainsi introduire des liens de dépendance entre les différents paramètres. Par exemple, si la présence de la protéine 2 est reliée à celle de la protéine 1, nous pouvons prendre en compte cette information à travers la loi *a priori* de x_2 conditionnellement à x_1. En l'absence d'information particulière sur ces liens, nous choisissons de considérer les paramètres comme indépendants *a priori*.

Dans ce cas, la règle du produit indique que la loi *a priori* recherchée est obtenue à partir du produit des lois *a priori* de chaque paramètre

$$p(\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{t},\gamma_b) = p(\gamma_b)\prod_{p=1}^{P} p(x_p)\prod_{i=1}^{I} p(\xi_i)p(t_i)$$

Ces lois sont définies dans les paragraphes suivants. Précisons que les liens de dépendance pourront toujours apparaître dans la loi jointe *via* la vraisemblance. De plus, la loi *a priori* modélise l'information disponible à un certain moment. Ainsi, si des informations ultérieures permettent de modéliser des liens entre ces paramètres, une méthode plus adaptée pourra être développée utilisant une nouvelle loi *a priori*.

D'une façon générale, le choix des lois *a priori* dépend de deux critères :

1) la traduction des informations *a priori* disponibles,
2) la possibilité de faire des calculs simples.

Nous décrivons dans les sections suivantes les lois *a priori* de chaque paramètre. Nous avons choisi de coder le domaine de recherche des paramètres en loi de probabilité. Une fois que l'ensemble de ces lois sera fixé, la loi jointe sera définie et l'ensemble de notre problème aura été décrit de façon mathématique. Les autres lois de probabilité et les estimateurs pourront alors être déduits.

4.3.3.1. Loi a priori pour les concentrations

Cette densité modélise l'information disponible sur les concentrations. Nous prenons en compte le fait que les concentrations se situent dans une gamme de valeurs fixée. La densité choisie doit attribuer une forte probabilité aux valeurs appartenant à cet intervalle et une probabilité plus faible aux autres. Parmi les densités généralement employées pour traduire cette information, nous considérons la loi uniforme et la loi normale.

L'information *a priori* « x_i appartient à la gamme de concentrations $\left[x_p^m; x_p^M\right]$ » pourra donc être traduite de deux façons. Nous pourrons choisir la loi uniforme entre x_p^m et x_p^M, de moyenne $\left(x_p^M + x_p^m\right)/2$ et d'inverse variance $12/\left(x_p^M - x_p^m\right)^2$ ou la loi normale de même moyenne et de même variance.

La densité uniforme attribue un poids égal à toutes les valeurs de l'intervalle et la densité gaussienne favorise les valeurs centrales par rapport aux extrémités. La loi uniforme interdit les valeurs en dehors de l'intervalle, la loi gaussienne les autorise avec une faible probabilité. En toute exactitude, ces deux densités produisent des méthodes différentes aux résultats différents. Toutefois, ces différences seront généralement minimes car l'information traduite est la même. De plus, les résultats de ces méthodes seront encore plus proches si l'intervalle de recherche est grand ou au contraire que la loi *a priori* ou la fonction de vraisemblance sont suffisamment piquées.

Il nous reste à choisir la distribution permettant les calculs les plus simples. Il est difficile de réaliser ce choix sans connaître ni l'estimateur, ni l'algorithme utilisé, car la simplicité de la méthode dépendra grandement de ces facteurs. De manière générale, il vaut mieux essayer de ne manipuler que des lois de probabilité usuelles. Parmi les deux *a priori* sélectionnées, la distribution gaussienne semble alors le meilleur choix car il s'agit de l'*a priori* conjugué[1] de la vraisemblance [86]. Dans notre cas, la vraisemblance est une fonction gaussienne à un coefficient multiplicatif près

$$L\left(x_p\right) \propto N\left(x_p; \mu_{\mathbf{Y}|x_p}, \gamma_{\mathbf{Y}|x_p}\right)$$

Le centre de cette distribution correspond à la solution des moindres carrés (*cf.* l'annexe 7.3 page 116). La vraisemblance étant gaussienne, le choix d'un *a priori* gaussien nous donne une loi jointe des concentrations et des données gaussiennes. En résumé, l'information *a priori* « x_i appartient à la gamme de concentrations $\left[x_p^m; x_p^M\right]$ » sera donc traduite par la densité de probabilité normale

$$N\left(x_p; \bar{x}_p, \gamma_x^p\right) \text{ avec } \bar{x}_p = \frac{1}{2}\left(x_p^m + x_p^M\right) \text{ et } \gamma_x^p = 12/\left(x_p^M - x_p^m\right)^2$$

Dans le cas limite où l'utilisateur ne souhaite favoriser aucune valeur, où x_p peut prendre n'importe quelle valeur *a priori*, le paramètre γ_x^p doit tendre vers 0. La loi *a priori* tendra alors vers une loi uniforme quelle que soit la valeur de $\bar{x}_p$. En pratique, c'est le choix effectué dans nos expérimentations.

4.3.3.2. Loi a priori pour les gains

La loi *a priori* des gains est obtenue de façon similaire. En effet, les gains possèdent les caractéristiques suivantes. D'un point de vue déterministe, le modèle est linéaire par rapport aux gains, si l'on fixe les autres paramètres (section 4.1). D'un point de vue statistique, la loi affectée au bruit est normale, la vraisemblance des gains est donc gaussienne. Pour coder notre domaine de recherche, nous choisissons la loi normale

$$p(\xi_i) = N\left(\xi_i; \bar{\xi}_i, \gamma_\xi^i\right)$$

Cette loi est la loi conjuguée à la vraisemblance des gains. Les paramètres de cette loi sont obtenus de façon similaire à la section précédente à partir des valeurs extrêmes du domaine de recherche.

Deux cas limites seront particulièrement étudiés dans nos expérimentations. Dans un premier temps, nous fixons la valeur de ξ_i. La loi *a priori* tend alors vers une distribution de Dirac positionnée sur la valeur désirée. Dans un second temps, nous ne souhaitons pas introduire d'information *a priori* particulière et la loi tend vers la loi uniforme entre $-\infty$ et $+\infty$.

[1] Pour une vraisemblance donnée, une loi *a priori* conjuguée est une loi de probabilité qui produit une loi jointe de la même forme.

4.3.3.3. Loi a priori pour les positions

D'une expérience à l'autre, la position des pics chromatographiques peut varier de quelques minutes [21] pour des valeurs nominales de l'ordre de quelques dizaines de minutes. De plus, nous pouvons définir une position moyenne pour chaque peptide, à l'aide de quelques expériences d'étalonnage. L'intervalle de recherche pour la position des gaussiennes chromatographiques est donc fixé.

Nous avons le choix entre les lois uniforme et gaussienne, cependant le choix de l'*a priori* conjugué n'est pas réalisable ici car la vraisemblance des positions n'est pas une fonction usuelle. En effet, les positions interviennent dans l'argument d'une exponentielle à travers la norme des erreurs et surtout la relation qu'elles entretiennent avec les données est non linéaire. Les calculs de l'annexe 7.4 page 118 montrent qu'à un coefficient additif près cette log-vraisemblance est une somme pondérée de fonctions gaussiennes.

$$L(t_{i_0}) \propto \exp\left\{ \sum_{n=1}^{N_c} \beta'_{i_0 n} \gamma_b N(nT_e^c ; t_{i_0}, \gamma_c) - \sum_{\substack{u=1 \\ u \neq i_0}}^{I} \frac{\alpha'_{i_0 u} \gamma_b}{T_e^c} N\left(t_u ; t_{i_0}, \frac{\gamma_c}{2} \right) \right\}$$

Avec

$$\alpha'_{iu} = \sum_{pjk} \sum_{qvw} d_{ip} d_{uq} \xi_i \xi_u \pi_{ij} \pi_{uv} \left(x_p x_q \pi'_{ijk} \pi'_{uvw} \boldsymbol{s}^t_{ijk} \boldsymbol{s}_{uvw} + 2 x_p x^*_q \pi'_{ijk} \pi'^*_{uvw} \boldsymbol{s}^t_{ijk} \boldsymbol{s}^*_{uvw} + x^*_p x^*_q \pi'^*_{ijk} \pi'^*_{uvw} \boldsymbol{s}^{*t}_{ijk} \boldsymbol{s}^*_{uvw} \right)$$

$$\boldsymbol{\beta}'^t_i = \sum_{pjk} d_{ip} \xi_i \pi_{ij} \left(x_p \pi'_{ijk} \boldsymbol{s}^t_{ijk} + x^*_p \pi'^*_{ijk} \boldsymbol{s}^{*t}_{ijk} \right) \boldsymbol{Y} = \begin{bmatrix} \beta'_{i1} & \dots & \beta'_{iN_c} \end{bmatrix}$$

Les coefficients pondérateurs dépendent des données, cette loi peut donc prendre de nombreuses formes. Comme la loi pour les positions n'a pas de forme particulière, le choix entre les deux distributions candidates est alors arbitraire : nous choisissons la distribution uniforme. L'information *a priori* « t_i appartient à la gamme de temps $[t_i^m ; t_i^M]$ » sera traduite par la densité de probabilité

$$p(t_i) = U(t_i ; t_i^m, t_i^M) = \begin{cases} 1 / \left| t_i^M - t_i^m \right| & \text{si } t_i \in [t_i^m ; t_i^M] \\ 0 & \text{sinon} \end{cases}$$

Nous choisissons donc de ne pas privilégier certaines valeurs de position par rapport à d'autres.

4.3.3.4. Loi a priori pour l'inverse puissance du bruit

Examinons maintenant l'inverse puissance du bruit. Une réécriture de sa vraisemblance nous montre que son expression est identique à celle d'une distribution gamma, à un coefficient multiplicatif près.

$$\begin{aligned} L(\gamma_b) &= (2\pi\gamma_b^{-1})^{-N_c N_s / 2} \exp\left(-\frac{1}{2} \gamma_b \chi(\boldsymbol{x}, \boldsymbol{x}^*, \boldsymbol{\xi}, \boldsymbol{t}) \right) \\ &\propto G\left(\gamma_b ; \frac{N_c N_s}{2} + 1, \frac{2}{\chi(\boldsymbol{x}, \boldsymbol{x}^*, \boldsymbol{\xi}, \boldsymbol{t})} \right) \\ &= G(\gamma_b ; \alpha_{b|Y}, \beta_{b|Y}) \end{aligned}$$

avec $G(\gamma ; \alpha, \beta) = \frac{\gamma^{\alpha-1}}{\beta^\alpha \Gamma(\alpha)} \exp\left(-\frac{\gamma}{\beta} \right)$ la distribution gamma de paramètres α et β et Γ la fonction gamma [87]. Nous choisissons pour la distribution $p(\gamma_b)$ l'*a priori* conjugué de cette vraisemblance qui est aussi une loi gamma.

$$p(\gamma_b) = G(\gamma_b ; \alpha_b, \beta_b)$$

La moyenne de cette distribution est $\alpha_b\beta_b$, sa variance $\alpha_b\beta_b^2$. L'utilisateur disposant d'informations sur le bruit pourra les injecter dans notre méthode en fixant ces paramètres. Cependant, comme ces caractéristiques dépendent fortement du modèle utilisé pour décrire les données, les utilisateurs peuvent souhaiter disposer de valeurs par défaut, pour utiliser la distribution la moins informative possible.

La distribution de Jeffreys est généralement choisie pour ses propriétés d'invariance par reparamétrisation. Il s'agit d'un cas limite de la famille proposée, obtenu quand le paramètre α_b tend vers 0 et β_b tend vers $+\infty$.

Nous avons maintenant défini toutes les lois *a priori* et la vraisemblance, la loi jointe est maintenant complètement explicitée. Nous pouvons donc calculer toutes les autres lois associées.

4.3.4. Loi a posteriori

Dans cette section, nous calculons la densité de probabilité *a posteriori*. La loi *a posteriori* est la loi de probabilité des paramètres, sachant que l'on a acquis les données particulières $\boldsymbol{Y}$. C'est donc la loi associée à la mesure de nos paramètres. Nous la faisons apparaitre après une nouvelle application de la règle du produit sur la loi jointe.

$$p(\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{t},\gamma_b,\boldsymbol{Y}) = p(\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{t},\gamma_b|\boldsymbol{Y})p(\boldsymbol{Y})$$

En combinant la formule précédente et l'expression de la loi jointe, nous obtenons l'expression de la loi *a posteriori*.

$$p(\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{t},\gamma_b|\boldsymbol{Y}) = \frac{p(\boldsymbol{Y}|\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{t},\gamma_b)p(\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{t},\gamma_b)}{p(\boldsymbol{Y})}$$

Il s'agit de la formule de Bayes permettant de passer de la loi *a priori* à la loi *a posteriori via* la fonction de vraisemblance. La loi des données y joue le rôle de facteur de normalisation. En détaillant l'expression de chaque probabilité, nous obtenons

$$\begin{aligned} p(\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{t},\gamma_b|\boldsymbol{Y}) \propto & \exp\left(-\frac{1}{2}\gamma_b\left\|\boldsymbol{Y}-\Theta(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t})\right\|^2\right) \\ & \times\frac{\gamma_b^{\alpha_b-1}}{\beta_b^{\alpha_b}\Gamma(\alpha_b)}\exp\left(-\frac{\gamma_b}{\beta_b}\right)\times\prod_{i=1}^{P}\exp\left(-\frac{1}{2}\gamma_x^p\left(x_p-\bar{x}_p\right)^2\right) \\ & \times\prod_{i=1}^{I}\exp\left(-\frac{1}{2}\gamma_\xi^i\left(\xi_i-\bar{\xi}_i\right)^2\right)\times\prod_{i=1}^{I}U\left(t_i;t_i^m,t_i^M\right) \end{aligned}$$

Si l'on excepte le terme de normalisation, cette loi est composée de cinq facteurs : la fonction de vraisemblance avec son terme d'attache aux données et les lois *a priori* pour chacun des quatre paramètres recherchés. Nous avons donc obtenu la loi de probabilité permettant l'estimation de nos paramètres. Il nous reste à définir un estimateur (estimateur du maximum, de la moyenne, …) ainsi qu'un algorithme pour le calculer. Ces étapes seront discutées à la section 4.4, mais avant, étudions quelques lois conditionnelles *a posteriori*.

4.3.5. Lois conditionnelles a posteriori

La densité *a posteriori* est la loi permettant de résoudre notre problème d'estimation. Pour l'obtenir, nous avons fixé un paramètre de la loi jointe, les données. Il est également possible de fixer d'autres paramètres ou plus généralement, tout ensemble de paramètres. Les lois ainsi obtenues décrivent des problèmes plus contraints. Dans cette section, nous étudions les distributions conditionnelles *a posteriori* pour chaque paramètre, les autres étant fixés. Notons que nous avons déjà examiné la loi conditionnelle où seules les données varient à la section 4.3.2, puisqu'il s'agit de la

vraisemblance. Toutes les lois conditionnelles restantes seront à données fixées, nous les qualifions donc de lois conditionnelles *a posteriori*.

Les lois conditionnelles sont toutes obtenues de la même façon. Nous utilisons le symbole ω pour désigner successivement $\boldsymbol{x}$, $\boldsymbol{\xi}$, $\boldsymbol{t}$ et γ_b. Le vecteur $\boldsymbol{\theta}$ regroupe les autres paramètres. La loi conditionnelle de ω est obtenue à partir de la loi jointe.

$$p(\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{t},\gamma_b,\boldsymbol{Y}) = p(\omega,\boldsymbol{\theta},\boldsymbol{Y})$$

Par application de la règle du produit nous obtenons une expression de la loi conditionnelle *a posteriori* faisant intervenir la loi jointe.

$$p(\omega|\boldsymbol{\theta},\boldsymbol{Y}) = \frac{p(\omega,\boldsymbol{\theta},\boldsymbol{Y})}{p(\boldsymbol{\theta},\boldsymbol{Y})} \propto p(\omega,\boldsymbol{\theta},\boldsymbol{Y})$$

La loi conditionnelle *a posteriori* est donc égale à la loi jointe à un coefficient multiplicatif près. Elle constitue donc une coupe de la loi jointe. De plus par une nouvelle application de la règle du produit sur la formule précédente, nous constatons qu'il s'agit également d'une coupe de la loi *a posteriori*.

$$p(\omega|\boldsymbol{\theta},\boldsymbol{Y}) = \frac{p(\omega,\boldsymbol{\theta}|\boldsymbol{Y})p(\boldsymbol{Y})}{p(\boldsymbol{\theta},\boldsymbol{Y})} \propto p(\omega,\boldsymbol{\theta}|\boldsymbol{Y})$$

Les lois conditionnelles *a posteriori* permettent donc de mieux étudier le problème, de visualiser la forme des lois et de noter les éventuelles indéterminations de notre problème. En outre, nous verrons plus tard qu'elles serviront à calculer notre estimateur.

L'expression de la loi conditionnelle *a posteriori* peut se déduire de la vraisemblance et de la loi *a priori*. En effet par une nouvelle application de la règle du produit nous obtenons

$$p(\omega|\boldsymbol{\theta},\boldsymbol{Y}) = \frac{p(\boldsymbol{Y}|\omega,\boldsymbol{\theta})p(\omega,\boldsymbol{\theta})}{p(\boldsymbol{\theta},\boldsymbol{Y})}$$

Nous avons modélisé nos paramètres comme étant indépendants. La loi *a priori* des paramètres $p(\omega, \boldsymbol{\theta})$ peut donc se factoriser.

$$p(\omega|\boldsymbol{\theta},\boldsymbol{Y}) = \frac{p(\boldsymbol{Y}|\omega,\boldsymbol{\theta})p(\omega)p(\boldsymbol{\theta})}{p(\boldsymbol{\theta},\boldsymbol{Y})}$$

Les lois $p(\boldsymbol{\theta})$ et $p(\boldsymbol{\theta}, \boldsymbol{Y})$ ne dépendent pas de ω. La loi conditionnelle *a posteriori* du paramètre ω est donc le produit de sa fonction de vraisemblance et de sa loi *a priori*.

$$p(\omega|\boldsymbol{\theta},\boldsymbol{Y}) \propto p(\boldsymbol{Y}|\omega,\boldsymbol{\theta})p(\omega) = L(\omega)p(\omega)$$

Cette formule sera donc utilisée dans tous les paragraphes suivants.

4.3.5.1. Loi conditionnelle a posteriori des concentrations et des gains

Cette loi conditionnelle est le produit de deux fonctions gaussiennes décrivant respectivement l'*a priori* et la vraisemblance. Nous obtenons donc une loi gaussienne

$$\begin{aligned} p(x_p|\boldsymbol{Y},\boldsymbol{x}_{-p},\boldsymbol{\xi},\boldsymbol{t},\gamma_b) &\propto N(x_p;\bar{x}_p,\gamma_x^p)\times N\left(x_p;\mu_{\boldsymbol{Y}|x_p},\gamma_{\boldsymbol{Y}|x_p}\right) \\ &\propto N\left(x_p;\frac{\gamma_x^p\bar{x}_p+\gamma_{\boldsymbol{Y}|x_p}\mu_{\boldsymbol{Y}|x_p}}{\gamma_x^p+\gamma_{\boldsymbol{Y}|x_p}},\gamma_x^p+\gamma_{\boldsymbol{Y}|x_p}\right) \end{aligned}$$

La gaussienne résultante est centrée sur la moyenne des centres de la vraisemblance et de la loi *a priori* pondérée par leur largeur. La largeur résultante correspond à la somme de leurs deux largeurs.

On peut également s'intéresser à la loi conditionnelle de toutes les concentrations conjointement. C'est le produit de la vraisemblance et de la loi *a priori* de toutes les positions.

$$p(\boldsymbol{x}|\boldsymbol{Y},\boldsymbol{\xi},\boldsymbol{t},\gamma_b) \propto L(\boldsymbol{x}) \times p(\boldsymbol{x})$$

La loi *a priori* $p(\boldsymbol{x})$ est obtenue en multipliant les différentes lois normales $p(x_i)$. Nous obtenons une loi normale multivariée

$$\begin{aligned} p(\boldsymbol{x}) &= \prod_{p=1}^{P} p(x_p) \\ &= N(\boldsymbol{x};\bar{\boldsymbol{x}},\boldsymbol{\Gamma}_x) \end{aligned}$$

avec les scalaires $\bar{x}_p$ regroupés dans le vecteur $\bar{\boldsymbol{x}}$ et les γ_x^p dans la matrice diagonale $\boldsymbol{\Gamma}_x$.

De plus, la vraisemblance de $\boldsymbol{x}$ est gaussienne (*cf.* annexe 7.3 page 116), la conditionnelle *a posteriori* est donc gaussienne

$$\begin{aligned} p(\boldsymbol{x}|\boldsymbol{Y},\boldsymbol{\xi},\boldsymbol{t},\gamma_b) &\propto L(\boldsymbol{x}) \times p(\boldsymbol{x}) \\ &\propto N(\boldsymbol{x};\boldsymbol{\mu}_{Y|x},\boldsymbol{\Gamma}_{Y|x}) \times N(\boldsymbol{x};\bar{\boldsymbol{x}},\boldsymbol{\Gamma}_x) \\ &\propto N(\boldsymbol{x};\boldsymbol{\mu}_{x|Y},\boldsymbol{\Gamma}_{x|Y}) \end{aligned}$$

avec

$$\begin{cases} \boldsymbol{\mu}_{x|Y} = (\boldsymbol{\Gamma}_{Y|x} + \boldsymbol{\Gamma}_x)^{-1}(\boldsymbol{\Gamma}_{Y|x}\boldsymbol{\mu}_{Y|x} + \boldsymbol{\Gamma}_x\bar{\boldsymbol{x}}) \\ \boldsymbol{\Gamma}_{x|Y} = \boldsymbol{\Gamma}_{Y|x} + \boldsymbol{\Gamma}_x \end{cases}$$

De plus, la moyenne $\boldsymbol{\mu}_{Y|x}$ et la matrice de largeur $\boldsymbol{\Gamma}_{Y|x}$ peuvent se calculer à partir de la matrice $\boldsymbol{H}$.

$$\begin{cases} \boldsymbol{\mu}_{Y|x} = (\boldsymbol{H}^t\boldsymbol{H})^{-1}(\boldsymbol{H}^t\boldsymbol{y} - \boldsymbol{H}^t\boldsymbol{H}^*\boldsymbol{x}^*) \\ \boldsymbol{\Gamma}_{Y|x} = \gamma_b\boldsymbol{H}^t\boldsymbol{H} \end{cases}$$

Nous retrouvons le même genre de structure que précédemment. La gaussienne résultante est centrée sur la moyenne des centres de la vraisemblance et de la loi *a priori* pondérée par leur largeur. La largeur résultante correspond à la somme de leurs deux matrices de largeur $\boldsymbol{\Gamma}_{Y|x}$ et $\boldsymbol{\Gamma}_x$.

Les concentrations et les gains ont des comportements similaires dans cette méthode. Premièrement, la relation qui lie ces deux variables aux données est linéaire. Deuxièmement, leur a priori est modélisé par une loi normale. Les relations concernant les concentrations calculées ci-dessus peuvent donc se transposer aux gains. Nous obtenons

$$\begin{aligned} p(\boldsymbol{\xi}|\boldsymbol{Y},\boldsymbol{x},\boldsymbol{t},\gamma_b) &\propto N\left(\boldsymbol{\xi};(\boldsymbol{\Gamma}_{Y|\xi} + \boldsymbol{\Gamma}_\xi)^{-1}(\boldsymbol{\Gamma}_{Y|\xi}\boldsymbol{\mu}_{Y|\xi} + \boldsymbol{\Gamma}_\xi\bar{\boldsymbol{\xi}}), \boldsymbol{\Gamma}_{Y|\xi} + \boldsymbol{\Gamma}_\xi\right) \\ &= N(\boldsymbol{x};\boldsymbol{\mu}_{\xi|Y},\boldsymbol{\Gamma}_{\xi|Y}) \end{aligned}$$

avec $\boldsymbol{\mu}_{Y|\xi} = (\boldsymbol{G}^t\boldsymbol{G})^{-1}\boldsymbol{G}^t\boldsymbol{y}$ et $\boldsymbol{\Gamma}_{Y|\xi} = \gamma_b\boldsymbol{G}^t\boldsymbol{G}$.

Dans le cas où l'utilisateur ne souhaite pas introduire d'information *a priori*, les centres et largeurs de ces distributions correspondent aux centres et largeurs des fonctions de vraisemblance. Pour des raisons de simplicité, c'est le choix proposé dans la suite, sauf mention contraire.

4.3.5.2. Loi conditionnelle a posteriori pour l'inverse variance du bruit

Comme nous l'avions prévu en prenant une loi *a priori* conjugué, la loi *a posteriori* conditionnelle pour γ_b est une loi gamma

$$\begin{aligned} p(\gamma_b|\boldsymbol{Y},\boldsymbol{x},\boldsymbol{t}) &\propto L(\gamma_b)\times p(\gamma_b) \\ &= G\left(\gamma_b;\alpha_{b|\boldsymbol{Y}},\beta_{b|\boldsymbol{Y}}\right)\times G(\gamma_b;\alpha_b,\beta_b) \\ &\propto G\left(\gamma_b;\alpha_{b|Y}+\alpha_b-1,\left(\beta_{b|\boldsymbol{Y}}^{-1}+\beta_b^{-1}\right)^{-1}\right) \\ &= G(\gamma_b;\alpha,\beta) \end{aligned}$$

Etudions le cas où la loi *a priori* tend vers une loi de Jeffreys, car nous ne souhaitons pas introduire d'information *a priori*. Le paramètre de la loi *a priori* α_b tend vers 0 et β_b tend vers $+\infty$. La loi *a posteriori* conditionnelle est toujours une loi gamma de paramètres

$$\alpha=\frac{N_c N_s}{2} \text{ et } \beta=\frac{2}{\chi\left(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t}\right)}$$

Calculons la moyenne de cette loi. Elle est obtenue en réalisant le produit de ces paramètres.

$$\alpha\beta=\frac{N_c N_s}{\chi\left(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t}\right)}$$

Cette expression peut sembler étrange au premier abord, mais n'oublions pas que nous estimons l'inverse de la variance. Nous constatons donc que l'inverse de cette expression est la moyenne empirique des carrés des erreurs. La loi est donc centrée autour d'une solution empirique permettant d'estimer l'inverse variance du bruit.

4.3.5.3. Loi conditionnelle a posteriori pour les positions

La loi conditionnelle pour les positions est issue du produit de la vraisemblance et de la loi *a priori*

$$p\left(t_i|\boldsymbol{Y},\boldsymbol{x},\boldsymbol{t}_{-i},\gamma_b\right)\propto\left(2\pi\gamma_b^{-1}\right)^{-N_c N_s/2}\exp\left(-\frac{\gamma_b}{2}\chi\left(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t}\right)\right)\times U\left(t_i;t_i^m,t_i^M\right)$$

Si nous regardons la loi de densité conditionnelle pour toutes les positions conjointement nous obtenons

$$p\left(\boldsymbol{t}|\boldsymbol{Y},\boldsymbol{x},\gamma_b\right)\propto\exp\left(-\frac{\gamma_b}{2}\chi\left(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t}\right)\right)\times\prod_{i=1}^{I}U\left(t_i;t_i^m,t_i^M\right)$$

Nous avions remarqué à la section 4.3.3.3 page 67 que la vraisemblance par rapport aux positions n'a pas une forme usuelle, c'est également le cas de la loi conditionnelle *a posteriori* qui est principalement formée par cette vraisemblance. Nous verrons lors du calcul de l'estimateur, que ce sont les positions qui vont causer le plus de difficultés à cause de cette forme non standard.

4.4. Estimateur de la moyenne

La loi *a posteriori* a été calculée à la section précédente. Elle associe une probabilité à chaque valeur. Cependant, on souhaite généralement aller plus loin en choisissant une valeur parmi toutes les valeurs candidates. Selon la théorie bayésienne, le choix de l'estimateur dépend du coût associé aux erreurs d'estimation. Plusieurs estimateurs sont classiquement utilisés (maximum, moyenne, médiane, …), chacun ayant des qualités et des défauts. Nous étudierons dans les paragraphes suivants l'estimateur de la moyenne qui correspond à un coût quadratique. De plus, cet estimateur a pour propriété de posséder un biais moyen nul.

L'estimateur de la moyenne est obtenu en calculant l'espérance de la loi *a posteriori*.

$$\left[\hat{\boldsymbol{x}},\hat{\boldsymbol{\xi}},\hat{\boldsymbol{t}},\hat{\gamma}_b\right]=\int\left[\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{t},\gamma_b\right]p\left(\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{t},\gamma_b|\boldsymbol{Y}\right)\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{\xi}\,\mathrm{d}\boldsymbol{t}\,\mathrm{d}\gamma_b$$

A notre connaissance, cette intégrale ne peut se calculer directement. Si le calcul direct n'est pas possible, plusieurs solutions existent pour la calculer de façon approchée.

- *Méthodes de Riemann, des trapèzes, de Simpson,…* Il s'agit des méthodes les plus classiques pour calculer numériquement une intégrale. Elles sont toutes basées sur une approximation de la fonction à intégrer par une fonction polynomiale par morceaux. Toutefois, ces méthodes nécessitent en pratique de quadriller l'espace par un maillage fin et d'évaluer la fonction en chaque point. Le coût de calcul pour les intégrales à plusieurs dimensions est rapidement rédhibitoire.
- *Méthodes variationnelles.* Dans ces méthodes, la fonction à intégrer est approchée par une fonction sélectionnée de la façon suivante. Elle doit être intégrable analytiquement et la formule obtenue doit nécessiter peu de calculs. Si l'intégrale est calculée plus rapidement que précédemment, la fonction approchante doit être choisie avec le plus grand soin pour éviter que l'approximation soit trop grossière.
- *Méthodes de Monte Carlo*. Les méthodes de Monte Carlo désignent les techniques utilisant le hasard et les probabilités. Dans notre cas, elles se basent sur la constatation suivante. La valeur de la densité *a posteriori* est quasi nulle sur une grande partie de l'espace. Il est donc inefficace d'utiliser un quadrillage régulier du domaine d'intégration. Les techniques de Monte Carlo visent à concentrer les efforts aux endroits de forte probabilité. Ces derniers sont obtenus à l'aide d'un générateur aléatoire simulant cette loi. En effet, les échantillons produits se concentrent autour des valeurs les plus probables.

Nous utiliserons les méthodes de Monte Carlo, car nous pensons qu'elles présentent un bon compromis entre performances, généricité et temps de calcul. Plus précisément, notre intégrale est l'espérance d'une densité de probabilité. Dans ce cas, le résultat s'approche en calculant la moyenne empirique d'échantillons de cette loi. Nous obtenons donc

$$\left[\hat{\boldsymbol{x}} \quad \hat{\boldsymbol{\xi}} \quad \hat{\boldsymbol{t}} \quad \hat{\gamma}_b\right] \approx \frac{1}{K}\sum_{k=K_0}^{K+K_0-1}\left[\boldsymbol{x}^{(k)} \quad \boldsymbol{\xi}^{(k)} \quad \boldsymbol{t}^{(k)} \quad \gamma_b^{(k)}\right]$$

où $\left[\boldsymbol{x}^{(k)} \quad \boldsymbol{\xi}^{(k)} \quad \boldsymbol{t}^{(k)} \quad \gamma_b^{(k)}\right] \sim p\left(\boldsymbol{x},\boldsymbol{t},\gamma_b|\boldsymbol{Y}\right)$. Précisons que la qualité de l'approximation dépend du nombre d'échantillons utilisés K. De plus, une partie des échantillons générés peut être naturellement ignorée, ici les $K_0 - 1$ premiers échantillons. Plus de détails concernant cette approximation pourront être obtenus page 83 de [88].

En résumé, l'estimateur de la moyenne fait intervenir une intégrale qui n'est pas calculable analytiquement. Plusieurs méthodes peuvent être utilisées pour approcher sa valeur, nous avons choisi d'utiliser une technique de Monte Carlo faisant intervenir un générateur aléatoire. En effet, cette dernière correspond à l'espérance de la loi *a posteriori*. Elle peut donc être obtenue en calculant la moyenne d'échantillons tirés sous cette loi. Cependant, rappelons que cette loi n'est pas usuelle, il n'existe donc pas de générateur standard permettant de l'échantillonner directement. Finalement, nous avons reporté la difficulté du problème d'intégration sur la construction du générateur aléatoire. Nous allons décrire dans la section suivante la construction d'un générateur simulant cette loi à l'aide des techniques de Monte Carlo par chaîne de Markov.

4.5. Echantillonneur de Gibbs

La construction d'un générateur de variables aléatoires quelconque peut se faire grâce aux techniques de Monte Carlo par chaîne de Markov (MCMC). Il s'agit de construire une procédure itérative fournissant après un certain temps de chauffe des échantillons de la loi souhaitée. La construction *ex nihilo* d'une telle procédure n'est pas une entreprise facile, mais nous disposons de deux algorithmes standard ayant les propriétés requises, les échantillonneurs[1] de Gibbs et de Metropolis-Hastings. Pour plus d'informations générales sur les méthodes MCMC, le lecteur pourra se référer à [88].

[1] Le terme échantillonnage est ici utilisé dans son sens statistique. Synonyme : tirage aléatoire.

- Initialiser $\boldsymbol{x}^{(1)}, \boldsymbol{\xi}^{(1)}, \gamma_c^{(1)}), \boldsymbol{t}^{(1)}$.
- Pour k=1 à K+K_0-1

  Echantillonner $\boldsymbol{x}^{(k+1)} \sim p\left(\boldsymbol{x}\middle|\boldsymbol{Y}, \boldsymbol{\xi}^{(k)}, \boldsymbol{t}^{(k)}, \gamma_b^{(k)}\right)$

  Echantillonner $\boldsymbol{\xi}^{(k+1)} \sim p\left(\boldsymbol{\xi}\middle|\boldsymbol{Y}, \boldsymbol{x}^{(k+1)}, \boldsymbol{t}^{(k)}, \gamma_b^{(k)}\right)$

  Echantillonner $\gamma_b^{(k+1)} \sim p\left(\gamma_b\middle|\boldsymbol{Y}, \boldsymbol{x}^{(k+1)}, \boldsymbol{\xi}^{(k+1)}, \boldsymbol{t}^{(k)}\right)$

  Echantillonner $\boldsymbol{t}^{(k+1)} \sim p\left(\boldsymbol{t}\middle|\boldsymbol{Y}, \boldsymbol{x}^{(k+1)}, \boldsymbol{\xi}^{(k+1)}, \gamma_b^{(k+1)}\right)$
- Fin Pour.

Algorithme 1 : échantillonneur de Gibbs.

Pour notre générateur aléatoire, nous utiliserons un échantillonneur de Gibbs, qui permet de transformer le problème d'échantillonnage d'une loi multivariée complexe en un problème d'échantillonnage de lois plus simples (lois monovariées ou multivariées standard). L'algorithme de Gibbs est très simple. Pour échantillonner la loi multivariée, il suffit d'échantillonner successivement les lois conditionnelles en mettant successivement les paramètres à jour, puis d'itérer cette procédure jusqu'à obtenir le nombre d'échantillons voulu (Algorithme 1). Nous avons déjà défini au paragraphe 4.3.5 les différentes lois conditionnelles *a posteriori*. Puisque ces dernières sont connues, il nous reste à décrire le fonctionnement des générateurs qui simulent ces lois plus simples. La plupart sont des lois standard et pourront être échantillonnées facilement mais d'autres comme la loi conditionnelle des positions demanderont plus d'efforts.

4.5.1. Echantillonnage des concentrations et des gains

Les variables $\boldsymbol{x}$ et $\boldsymbol{\xi}$ sont distribuées selon deux lois normales multivariées

$$p\left(\boldsymbol{x}\middle|\boldsymbol{Y}, \boldsymbol{\xi}, \boldsymbol{t}, \gamma_b\right) = N\left(\boldsymbol{x}; \boldsymbol{\mu}_{x|Y}, \boldsymbol{\Gamma}_{x|Y}\right)$$
$$p\left(\boldsymbol{\xi}\middle|\boldsymbol{Y}, \boldsymbol{x}, \boldsymbol{t}, \gamma_b\right) = N\left(\boldsymbol{\xi}; \boldsymbol{\mu}_{\xi|Y}, \boldsymbol{\Gamma}_{\xi|Y}\right)$$

Ce sont des lois de probabilité standard. Nous avons développé des échantillonneurs rapides à partir d'un générateur de loi normale monovariée. Pour cela, nous appliquons à des échantillons indépendants suivant une loi normale centrée réduite[1] une transformation affine. Les paramètres de cette transformation sont choisis de telle façon que les échantillons obtenus aient la moyenne et la matrice de covariance souhaitée. En pratique, la matrice de la transformation est obtenue en factorisant la matrice de covariance par la transformation de Cholesky (Algorithme 2).

- Calculer la matrice de covariance $\boldsymbol{R} = \boldsymbol{\Gamma}^{-1}$
- Calculer sa décomposition de Cholesky de $\boldsymbol{R} = \boldsymbol{\Lambda}^t \boldsymbol{\Lambda}$.
- Générer $\boldsymbol{g}$ un vecteur de N variables indépendantes distribuées suivant une loi normale centrée réduite.
- Calculer l'échantillon $\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{\Lambda}^t \boldsymbol{g}$.

Algorithme 2 : échantillonneur de loi normale N($\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Gamma}$) .

L'échantillonnage de ces lois normales multivariées ne demande pas en soit un effort calculatoire important car les vecteurs considérés sont de petites tailles, typiquement une dizaine de paramètres. Cependant, le calcul de la moyenne et de la matrice de covariance font intervenir les matrices de grande taille $\boldsymbol{G}$, $\boldsymbol{H}$ et $\boldsymbol{H}^*$ pouvant occuper quelques centaines de mégaoctets. En effet,

$$\boldsymbol{\mu}_{x|Y} = \left(\boldsymbol{H}^t \boldsymbol{H}\right)^{-1}\left(\boldsymbol{H}^t \boldsymbol{y} - \boldsymbol{H}^t \boldsymbol{H}^* \boldsymbol{x}^*\right) \quad \boldsymbol{\Gamma}_{x|Y} = \gamma_b \boldsymbol{H}^t \boldsymbol{H}$$
$$\boldsymbol{\mu}_{\xi|Y} = \left(\boldsymbol{G}^t \boldsymbol{G}\right)^{-1} \boldsymbol{G}^t \boldsymbol{y} \qquad \boldsymbol{\Gamma}_{\xi|Y} = \gamma_b \boldsymbol{G}^t \boldsymbol{G}$$

[1] Nous avons utilisé la routine standard de MATLAB *randn*.

dans le cas où nous utilisons des lois *a priori* non informatives.

Si les matrices $\boldsymbol{G}$, $\boldsymbol{H}$ et $\boldsymbol{H}^*$ ont une taille importante, les matrices $\boldsymbol{G}^t\boldsymbol{G}$, $\boldsymbol{G}^t\boldsymbol{y}$, $\boldsymbol{H}^t\boldsymbol{H}$, $\boldsymbol{H}^t\boldsymbol{H}^*$ et $\boldsymbol{H}^t\boldsymbol{y}$ sont des matrices et des vecteurs de petite taille. De plus, nous disposons d'une méthode permettant de les calculer rapidement. Cette méthode utilise le caractère séparable du modèle et la forme gaussienne des pics. Notons que nous avons déjà évoqué le calcul de ces matrices pour évaluer rapidement la norme des erreurs (section 4.1 page 59).

4.5.2. Echantillonnage de l'inverse puissance du bruit

La variable γ_b est distribuée sous une loi gamma pouvant être échantillonnée facilement[1]

$$p\left(\gamma_b|\boldsymbol{Y},\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{t}\right)=\frac{\gamma_b^{\alpha-1}}{\beta^{\alpha}\Gamma(\alpha)}\exp\left(-\frac{\gamma_b}{\beta}\right)$$

avec $\alpha = N_c N_s/2$, $1/\beta = \chi\left(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t}\right)/2$, dans le cas où nous utilisons des lois *a priori* non informatives.

Notons que le calcul du paramètre β peut poser des difficultés. En effet, en raison de la taille conséquente des données, le calcul direct de la norme de l'erreur de modélisation peut être coûteux en espace mémoire et en temps de calcul. Ici aussi, nous pouvons utiliser la méthode de calcul rapide de la norme utilisant les matrices de petite taille $\boldsymbol{G}^t\boldsymbol{G}$, $\boldsymbol{G}^t\boldsymbol{y}$, $\boldsymbol{H}^t\boldsymbol{H}$, $\boldsymbol{H}^t\boldsymbol{y}$, $\boldsymbol{H}^{*t}\boldsymbol{H}^*$, $\boldsymbol{H}^t\boldsymbol{H}^*$, $\boldsymbol{H}^{*t}\boldsymbol{H}^*$ et $\boldsymbol{H}^{*t}\boldsymbol{y}$ présentée à la section 4.1 page 59.

4.5.3. Echantillonnage des positions chromatographiques

La distribution conditionnelle des positions n'est pas une distribution classique à cause de la dépendance complexe du modèle vis à vis des positions (section 4.3.3.3 page 67).

$$p\left(\boldsymbol{t}|\boldsymbol{Y},\boldsymbol{x},\boldsymbol{\xi},\gamma_b\right)\propto\exp\left(-\frac{\gamma_b}{2}\chi\left(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t}\right)\right)\times\prod_{i=1}^{I}U\left(t_i;t_i^m,t_i^M\right)$$

Il n'existe pas d'échantillonneur standard pour cette distribution. Nous avons développé deux échantillonneurs basés sur l'algorithme de Metropolis-Hastings. Le premier générant toutes les positions en même temps, le second une à une. Ces deux échantillonneurs sont décrits précisément dans les sections suivantes, mais tout d'abord nous allons expliquer les caractéristiques communes de ces échantillonneurs. Nous utiliserons dans ce paragraphe le paramètre $\boldsymbol{\theta}$ qui représente les paramètres $\boldsymbol{t}$ et t_i.

L'échantillonneur de Metropolis-Hastings permet de générer des échantillons d'une distribution cible φ à l'aide d'une distribution instrumentale ψ pour laquelle nous disposons d'un générateur aléatoire. La loi instrumentale peut être quelconque sous réserve de respecter les conditions suivantes :

- le support de la loi instrumentale doit contenir le support de la loi cible,
- la queue de la loi cible doit être plus courte que la queue de la loi instrumentale.

Tout comme pour l'échantillonneur de Gibbs, l'algorithme utilise une structure itérative et ne fournit des échantillons sous la loi cible φ qu'après un certains nombre d'itérations. Cette période est qualifiée de temps de chauffe (Algorithme 3). A chaque itération, nous testons un nouvel échantillon de la loi instrumentale. Soit il est accepté, et nous l'utilisons comme échantillon de la loi cible, soit il est refusé, et nous conservons l'échantillon précédent. Ce choix est réalisé de façon aléatoire suivant la probabilité δ

$$\delta=\min\left(1;\frac{\varphi(\boldsymbol{\theta}')\,\psi\left(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}'\right)}{\varphi\left(\boldsymbol{\theta}^{(k)}\right)\psi\left(\boldsymbol{\theta}';\boldsymbol{\theta}^{(k)}\right)}\right)$$

[1] Nous avons utilisé la routine standard de la « statistical toolbox » de MATLAB *gamrnd*.

Cette dernière dépend des probabilités des échantillons proposés $\boldsymbol{\theta}'$ et courant $\boldsymbol{\theta}^{(k)}$ suivant la loi cible φ et la loi instrumentale ψ.

- Initialisation $\boldsymbol{\theta}^{(1)}$
- Pour k = 1 à K
  - Simuler $\boldsymbol{\theta}'$ sous la loi $\psi\left(\boldsymbol{\theta}';\boldsymbol{\theta}^{(k)}\right)$
  - Calculer la probabilité δ
  - Acceptation / Conservation

$$\boldsymbol{\theta}^{(k+1)} = \begin{cases} \boldsymbol{\theta}' & \text{avec la probabilité} \quad \delta \\ \boldsymbol{\theta}^{(k)} & \text{avec la probabilité} \quad 1-\delta \end{cases}$$

- Fin Pour

Algorithme 3 : forme générale de l'échantillonneur de Metropolis-Hastings.

Notons que, sous la forme la plus générale de l'algorithme, la loi instrumentale peut être paramétrée par l'échantillon fourni à l'itération précédente. Dans ce travail, nous employons l'échantillonneur de Metropolis-Hastings indépendant, qui utilise à chaque itération la même loi instrumentale. Plus exactement, nous utiliserons la loi *a priori* comme loi instrumentale. Cette loi instrumentale, associée au fait que nous échantillonnons une loi conditionnelle *a posteriori*, permet de simplifier le calcul de δ. En effet, les lois conditionnelles *a posteriori* sont égales au produit d'une fonction de vraisemblance, et d'une loi *a priori* à un coefficient multiplicatif près (section 4.3.5). Nous avons donc

$$\varphi(\boldsymbol{\theta}) \propto L(\boldsymbol{\theta}) p(\boldsymbol{\theta})$$
$$\psi(\boldsymbol{\theta}) = p(\boldsymbol{\theta})$$

Nous pouvons opérer les simplifications suivantes

$$\delta = \min\left(1; \frac{L(\boldsymbol{\theta}')\, p(\boldsymbol{\theta}')\, p\left(\boldsymbol{\theta}^{(k)}\right)}{L\left(\boldsymbol{\theta}^{(k)}\right) p\left(\boldsymbol{\theta}^{(k)}\right) p(\boldsymbol{\theta}')}\right)$$
$$= \min\left(1; \frac{L(\boldsymbol{\theta}')}{L\left(\boldsymbol{\theta}^{(k)}\right)}\right)$$

Le calcul de δ fait intervenir le rapport de fonctions de vraisemblance. D'autres simplifications sont possibles. En effet, nous pouvons déduire de l'expression de la vraisemblance

$$L(\boldsymbol{\theta}) \propto \exp\left(-\frac{\gamma_b}{2}\chi(\boldsymbol{\theta})\right)$$

D'où la simplification

$$\delta = \min\left(1; \frac{L(\boldsymbol{\theta}')}{L\left(\boldsymbol{\theta}^{(k)}\right)}\right)$$
$$= \min\left(1; \exp\left(-\frac{1}{2}\gamma_b\left(\chi(\boldsymbol{\theta}') - \chi\left(\boldsymbol{\theta}^{(k)}\right)\right)\right)\right)$$

avec $\chi(\boldsymbol{\theta})$ la norme de l'erreur de modélisation.

Finalement le calcul de δ est basé sur la différence de deux normes de l'erreur de modélisation : la première utilise l'échantillon proposé, et la deuxième utilise l'échantillon courant. Notons que dans le cas où la proposition améliore la fidélité aux données, l'exponentielle est supérieure à 1. Dans ce cas, nous acceptons toujours l'échantillon proposé. Dans le cas contraire, plus l'aggravation est importante, plus la probabilité d'acceptation est faible. De plus, l'échantillonneur ne sera pas utilisé seul, mais sera

placé à l'intérieur d'une boucle de Gibbs étudié à la section 4.5 page 72. La création d'une seconde boucle n'est donc pas nécessaire.

Suite aux simplifications précédentes, l'algorithme obtenu (Algorithme 4) fonctionne de la façon suivante. On propose une valeur de $\boldsymbol{\theta}$ sous la loi *a priori*. Si la proposition réduit l'erreur de modélisation, elle est acceptée. Dans le cas contraire, on peut toujours accepter ces positions moins efficaces, et donc revenir en arrière, avec la probabilité δ.

Dans cette section nous avons étudié les éléments communs des échantillonneurs des positions. Etudions maintenant leurs spécificités.

- Simuler $\boldsymbol{\theta}'$ sous la loi $\psi(\boldsymbol{\theta}')$
- Si $\chi(\boldsymbol{\theta}') \leq \chi(\boldsymbol{\theta}^{(k)})$

  Alors $\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}'$

  Sinon $\boldsymbol{\theta}^{(k+1)} = \begin{cases} \boldsymbol{\theta}' & \text{avec la probabilité } \delta \\ \boldsymbol{\theta}^{(k)} & \text{avec la probabilité } 1-\delta \end{cases}$

Algorithme 4 : simplifications de l'algorithme de Metropolis-Hastings

4.5.3.1. Algorithme « Toutes les positions en même temps »

Dans cet algorithme, nous utilisons directement l'échantillonneur de Metropolis-Hastings pour simuler les positions. La loi cible est la conditionnelle *a posteriori* des positions et la loi instrumentale est la loi *a priori* des positions.

$$\begin{cases} \varphi_1(\boldsymbol{t}) = p\left(\boldsymbol{t} \middle| \boldsymbol{Y}, \boldsymbol{x}^{(k+1)}, \boldsymbol{\xi}^{(k+1)}, \gamma_b^{(k+1)}\right) \\ \psi_1(\boldsymbol{t}) = p(\boldsymbol{t}) = \prod_{i=1}^{I} U\left(t_i; t_i^m, t_i^M\right) \end{cases}$$

Le générateur de la loi ψ_1 est obtenu directement en tirant successivement les différents t_i uniformément entre t_i^m et t_i^M.

- $\forall i$, Echantillonner $t'_i \sim \mathcal{U}_{[t_i^m ; t_i^M]}$
- $\boldsymbol{t}' \leftarrow [t'_{1..I}]^t$
- Si $\chi(\boldsymbol{x}^{(k+1)}, \boldsymbol{x}^*, \boldsymbol{\xi}^{(k+1)}, \boldsymbol{t}') \leq \chi(\boldsymbol{x}^{(k+1)}, \boldsymbol{x}^*, \boldsymbol{\xi}^{(k+1)}, \boldsymbol{t}^{(k)})$

  Alors $\boldsymbol{t}^{(k+1)} = \boldsymbol{t}'$

  Sinon $\boldsymbol{t}^{(k+1)} = \begin{cases} \boldsymbol{t}' & \text{avec la probabilité } \delta \\ \boldsymbol{t}^{(k)} & \text{avec la probabilité } 1-\delta \end{cases}$

Algorithme 5 : échantillonneur des positions « toutes les positions en même temps »

L'échantillonneur obtenu fonctionne de la façon suivante (Algorithme 5). On propose un nouvel ensemble de positions parmi toutes les positions possibles. Puis on accepte la nouvelle proposition suivant une certaine probabilité dépendant de l'erreur de modélisation. Si la proposition a été refusée, on conserve les positions courantes.

Notons que dans cet algorithme toutes les positions évoluent en même temps ou alors pas du tout. Pendant un certain nombre d'itérations, l'algorithme peut conserver les mêmes positions provoquant un palier. La longueur de ce palier dépend de la probabilité d'acceptation. Si la loi *a priori* utilisée pour proposer les positions est suffisamment resserrée autour des valeurs les plus probables, l'algorithme va accepter souvent. Dans le cas contraire, beaucoup de positions seront refusées et l'algorithme stagnera sur une valeur.

Dans la section suivante, nous allons étudier un algorithme permettant d'accepter plus souvent en faisant évoluer séparément chaque position.

4.5.3.2. Algorithme « une position à la fois »

Dans cet algorithme, nous ne traitons pas directement l'échantillonnage des positions, nous allons simplifier le problème en utilisant une seconde fois l'échantillonneur de Gibbs. Chaque position est simulée séquentiellement. Ici aussi, nous pouvons profiter de la boucle itérative de l'échantillonneur de Gibbs principal, sans avoir à en créer une nouvelle[1]. L'Algorithme 6 décrit le nouvel échantillonneur.

- Pour $i = 1$ à I
  Echantillonner $t_i^{(k+1)} \sim p\left(t_i \left| \boldsymbol{Y}, \boldsymbol{x}^{(k+1)}, \boldsymbol{\xi}^{(k+1)}, \gamma_b^{(k+1)}, t_1^{(k+1)}, \ldots, t_{i-1}^{(k+1)}, t_{i+1}^{(k)}, \ldots, t_I^{(k)} \right.\right)$
- Fin Pour

Algorithme 6 : échantillonneur de Gibbs des positions

Comme pour tout échantillonneur de Gibbs, l'échantillonneur de la position t_i a besoin de connaître la valeur des autres positions, il se sert des positions déjà générées dans cette itération, et de celles générées à l'itération précédente. Nous structurons ce dernier échantillonneur autour de l'algorithme de Metropolis-Hastings. La loi cible est la conditionnelle *a posteriori* de la position t_i, et nous choisissons comme loi instrumentale la loi *a priori*.

$$\begin{cases} \varphi_2(t_i) = p\left(t_i \left| \boldsymbol{Y}, \boldsymbol{x}^{(k+1)}, \boldsymbol{\xi}^{(k+1)}, \gamma_b^{(k+1)}, t_1^{(k+1)}, \ldots, t_{i-1}^{(k+1)}, t_{i+1}^{(k)}, \ldots, t_I^{(k)} \right.\right) \\ \psi_2(t_i) = U\left(t_i; t_i^m, t_i^M\right) \end{cases}$$

On tire donc la nouvelle proposition de t_i sous la loi uniforme entre t_i^m et t_i^M. La probabilité d'acceptation δ est calculée de la façon suivante

$$\delta = \min\left(1; \exp\left(-\frac{1}{2}\gamma_b^{(k+1)}\left(\chi\left(\boldsymbol{x}^{(k+1)}, \boldsymbol{x}^*, \boldsymbol{\xi}^{(k+1)}, \boldsymbol{t}'\right) - \chi\left(\boldsymbol{x}^{(k+1)}, \boldsymbol{x}^*, \boldsymbol{\xi}^{(k+1)}, \boldsymbol{t}^{\#}\right)\right)\right)\right)$$

avec $\boldsymbol{t}' = \left[t_1^{(k+1)} \quad \ldots \quad t_{i-1}^{(k+1)} \quad t_i' \quad t_{i+1}^{(k)} \quad \ldots \quad t_I^{(k)}\right]^t$ et $\boldsymbol{t}^{\#} = \left[t_1^{(k+1)} \quad \ldots \quad t_{i-1}^{(k+1)} \quad t_i^{(k)} \quad t_{i+1}^{(k)} \quad \ldots \quad t_I^{(k)}\right]^t$

Les simplifications font apparaître la différence de deux normes : l'une fait intervenir la nouvelle position t_i', l'autre utilise la position courante $t_i^{(k)}$. Elles sont respectivement regroupées avec les autres positions retenues précédemment dans le vecteur candidat $\boldsymbol{t}'$ et le vecteur de référence $\boldsymbol{t}^{\#}$. Finalement, nous obtenons l'Algorithme 7. A l'intérieur de la boucle de Gibbs sur les positions, nous obtenons une structure de Metropolis-Hastings :

- échantillonnage de la nouvelle position candidate sous la loi instrumentale,
- calcul de la probabilité δ,
- acceptation / rejet suivant la probabilité δ.

[1] Notons que nous serions arrivé à la même méthode en découpant directement le problème suivant les paramètres $\boldsymbol{x}$, $\boldsymbol{\xi}$, γ_b, t_1, … t_I à la section 4.5.

- Pour $i = 1$ à I

  Echantillonner $t'_i \sim \mathcal{U}_{[t_i^m ; t_i^M]}$

  Affecter les variables $\begin{cases} \boldsymbol{t}' \leftarrow \left[t_1^{(k+1)} \quad \dots \quad t_{i-1}^{(k+1)} \quad t'_i \quad t_{i+1}^{(k)} \quad \dots \quad t_I^{(k)}\right]^t \\ \boldsymbol{t}^{\#} \leftarrow \left[t_1^{(k+1)} \quad \dots \quad t_{i-1}^{(k+1)} \quad t_i^{(k)} \quad t_{i+1}^{(k)} \quad \dots \quad t_I^{(k)}\right]^t \end{cases}$

  Si $\chi\left(\boldsymbol{x}^{(k+1)}, \boldsymbol{x}^*, \boldsymbol{\xi}^{(k+1)}, \boldsymbol{t}'\right) \leq \chi\left(\boldsymbol{x}^{(k+1)}, \boldsymbol{x}^*, \boldsymbol{\xi}^{(k+1)}, \boldsymbol{t}^{\#}\right)$

  Alors $t_i^{(k+1)} = t'_i$

  Sinon $t_i^{(k+1)} = \begin{cases} t'_i & \text{avec la probabilité} \quad \delta \\ t_i^{(k)} & \text{avec la probabilité} \quad 1-\delta \end{cases}$

- Fin Pour

Algorithme 7 : échantillonneur des positions « une position à la fois ».

Les performances constatées de ce nouvel algorithme sur un cas typique sont les suivantes : son exécution demande 4 fois plus de temps à nombre d'itération constant, mais nécessite 10 fois moins d'itérations pour converger. Le nouvel algorithme est donc globalement 2.5 fois plus rapide (21 secondes de calculs à la place de 56 secondes).

La principale raison de ces performances est un taux d'acceptation beaucoup plus grand. En effet, l'espace de recherche d'une proposition convenable est monodimensionnel dans le cas de l'algorithme « une position à la fois » alors que pour le précédent algorithme il était multidimensionnel. Comme l'algorithme accepte plus souvent, il a moins tendance à produire des paliers et à explorer plus facilement la loi cible. De plus, le nouvel algorithme est 4 fois moins rapide à nombre d'itérations constant. Cela correspond au nombre de normes calculées. L'algorithme précédent ne calculait qu'une nouvelle norme par appel, l'échantillonneur « une position à la fois » calcule I nouvelles normes.

4.6. Conclusion

Dans ce chapitre nous avons proposé une méthode d'estimation conjointe des paramètres d'intérêt et des paramètres instrument utilisant une approche bayésienne. Dans le cadre de cette thèse, nous nous sommes focalisés sur l'estimation des concentrations, des gains, des positions des pics chromatographiques et du paramètre de bruit.

L'approche bayésienne propose un cadre formel adapté à l'estimation paramétrique dans des cas complexes. Elle permet notamment d'introduire de l'information *a priori* sur les paramètres recherchés. Ainsi, pour améliorer les performances de la méthode, l'utilisateur a la possibilité de préciser l'intervalle de recherche des paramètres. Cette information est codée sous la forme de lois de probabilité uniforme, normale et gamma. A partir de ces lois *a priori* et de la modélisation du bruit, la formule de Bayes permet de calculer la loi *a posteriori* des paramètres. Cette dernière associe à chaque valeur des paramètres une probabilité.

Après avoir calculé la probabilité *a posteriori*, nous ne retenons qu'une valeur en utilisant l'estimateur de la moyenne *a posteriori*. Cet estimateur, activement étudié ces dernières années en traitement du signal, possède un biais moyen nul et permet d'introduire une pénalisation quadratique des erreurs d'estimation.

Nous avons choisi une mise en œuvre de cet estimateur basée sur les méthodes de Monte-Carlo par chaîne de Markov. Nous avons donc conçu un générateur aléatoire simulant la loi *a posteriori*. Ce dernier utilise une structure itérative de Gibbs dont le corps est composé de générateurs aléatoires de lois classiques (uniforme, normale et gamma). A l'intérieur de chaque itération, le coût calculatoire est concentré sur les paramètres de ces lois. Cet effort est d'autant plus grand que leur calcul fait intervenir des matrices de grande taille coûteuse en espace mémoire et en temps de calcul. Cependant, nous avons indiqué des simplifications utilisant la structure particulière de ces matrices.

Parmi les différents paramètres estimés, c'est l'échantillonnage des positions qui est le plus délicat à mettre en œuvre, en raison de la complexité de la relation qui lie les positions aux données. Nous avons proposé deux échantillonneurs des positions qui utilisent l'algorithme de Metropolis-Hastings. Si le premier est celui qui demande le moins de calculs par itération, le second permet d'explorer la loi cible plus efficacement.

La méthode développée ne se restreint pas aux paramètres listés, car une même démarche pourrait facilement être adaptée aux autres paramètres. Distinguons les paramètres linéaires, comme les gains de digestion et les proportions, des paramètres non linéaires comme les autres paramètres des pics. Les premiers pourront être estimés d'une façon similaire à la méthode présentée pour les concentrations. Pour les seconds, on pourra utiliser la technique utilisée pour estimer les positions. De même, la méthode ne se résume pas aux pics gaussiens, elle peut être adaptée à d'autres formes de pics plus réalistes en utilisant la même méthode. Notons tout de même que les fonctions gaussiennes permettent une accélération supplémentaire des calculs (voir 7.2 page 109).

Enfin nous avons proposé deux algorithmes d'échantillonnage de position basés sur deux découpages différents de l'algorithme de Gibbs. Une convergence plus rapide peut être obtenue en étudiant d'autres échantillonneurs plus complexes basés sur une loi de proposition plus proche de la loi cible.

5. Evaluation de la méthode

Dans le chapitre précédent, nous avons développé une méthode de quantification basée sur les statistiques bayésiennes et sur un modèle direct développé chapitre 2. Dans ce chapitre, nous la validerons à l'aide de données simulées proches de cas réels ce qui nous permettra de comparer nos estimations aux vraies valeurs. Ensuite, nous la testons sur plusieurs séries de données réelles issues d'expériences réalisées par des spécialistes sur les plateformes protéomiques du département DTBS et de l'institut iRTSV. Une première série d'expériences consiste en l'analyse de peptides de cytochrome C dilués dans de l'eau. Il s'agit d'une première confrontation de la méthode avec des données réelles dans un cas relativement simple. Puis, nous testons la méthode sur un cas plus complexe qui contient un nombre important de protéines, l'analyse d'une toxine du staphylocoque doré dans de l'urine humaine. Enfin, nous comparerons nos résultats à ceux produits par des méthodes présentées dans l'état de l'art.

5.1. Analyse du cytochrome C dans de l'eau

Les données simulées que nous utilisons cherchent à être proche d'un cas réel. Commençons donc par décrire ce cas réel. L'expérience d'intérêt a pour but d'analyser un mélange de peptides issus de la digestion par la trypsine du cytochrome C. Le cytochrome C est une protéine intervenant dans la respiration. Elle s'agit également d'une protéine type, souvent utilisé comme cas d'école.

Pour ces expériences, nous ne disposons pas de protéines marquées isotopiquement évoquées à la section 3.1.1 page 48. La fabrication de ces marqueurs est difficile et aurait demandé trop de temps. Cependant, sans l'utilisation de telles protéines d'étalonnage, il y a une indétermination entre les concentrations et les gains. Afin de lever cette indétermination, nous fixons arbitrairement tous les gains ξ_i à 1. Cette action peut être interprétée de trois façons indiquées ci-après.

1. Le gain est effectivement égal à 1. Il n'y a pas de perte de matière dans la chaîne de mesure et l'électronique du détecteur est réglée de façon à avoir un gain unitaire.
2. Le gain est toujours égal à 1, mais il y a des pertes dans la chaîne. Toutefois ces pertes sont identiques pour tous les peptides et elles sont compensées par l'électronique du détecteur.
3. L'hypothèse de gain unitaire est invalide. Dans ce cas, la valeur renvoyée par la méthode n'est pas l'expression directe de la concentration, mais le produit des concentrations et des gains. De plus, le reste du signal étant normalisé, la variable $\boldsymbol{x}$ renvoyée correspond au volume des pics. Ainsi il y aura une valeur renvoyée par peptide et non pas une seule valeur par protéine.

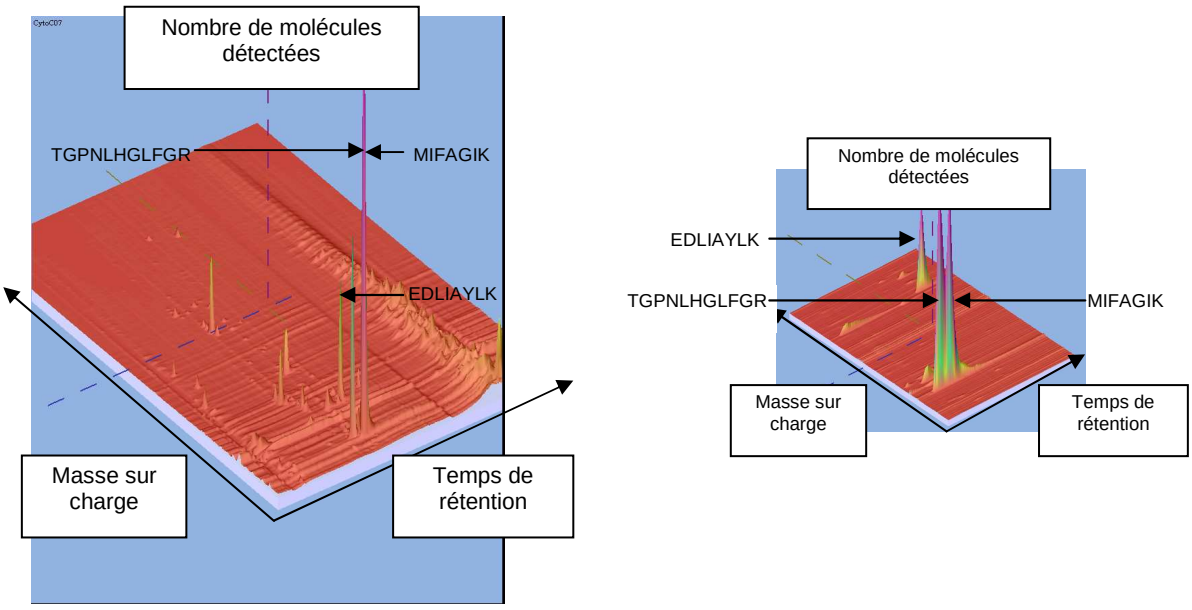
Dans les sections suivantes, nous estimerons donc les concentrations, leur position et l'écart type du bruit. L'estimation des concentrations des protéines et du gain sera l'objet de la section 5.2.

5.1.1. Protocole des expériences

Le mélange est analysé par une chaîne commerciale standard utilisant une colonne de chromatographie Dionex, une aiguille electrospray New Objective et une trappe linéaire LTQ de Thermo Scientific.

Plus précisément, il s'agit d'une nano-colonne Dionex Pepmap avec un diamètre interne de 75 µm et 15 cm de long. L'échantillon est directement injecté. Il n'y a pas de colonne de préconcentration dans le circuit. On utilise un gradient de solvant binaire linéaire. Le premier solvant (A) est formé de 99.9% H_2O et 0.1% AF, le second solvant (B) est formé de 19.9% H_2O, 80% ACN et de 0.1% AF (ACN : acétonitrile ; AF : acide formique). Le gradient dure 43 min (2580s). Le débit de la pompe est de 0.2 µl/min sous une pression de 120 bar. L'expérience est suivie de 27 min de lavage de la colonne au solvant A.

L'aiguille electrospray est placée à 1 mm du spectromètre de masse puis alimentée sous une tension de 1.7kV. Le capillaire est chauffé à 200°C.



Spectrogramme du fichier total *Zone d'intérêt*

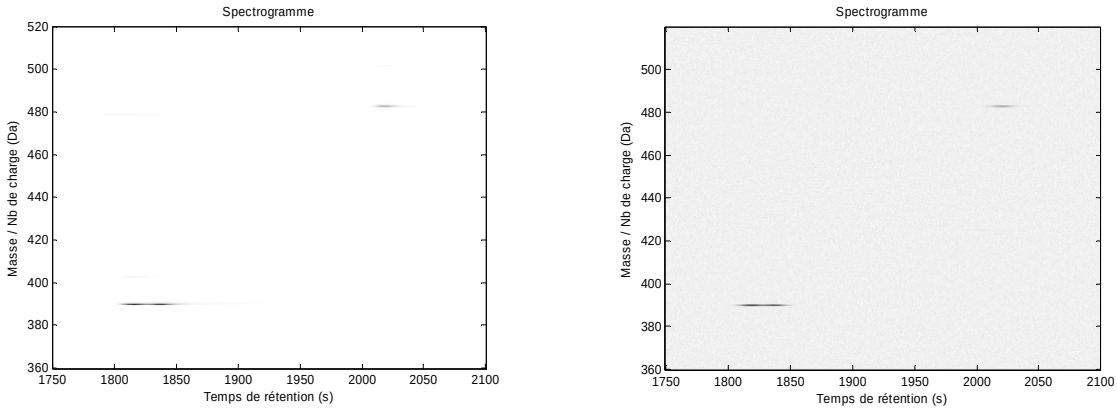
Figure 34 : spectrogramme des peptides du cytochrome C dilués à une concentration de 0.2 µmol/l. La figure de gauche représente le spectrogramme du fichier total. La figure de droite présente la zone d'intérêt correspondant au zoom sur le pic rose le plus intense de la figure de gauche, et à la somme des pics TGPNLHGLFGR et MIFAGIK.

Le spectromètre de masse produit des spectres de masse allant de 350.09 Da à 1100 Da toutes les 0.48 s en moyenne. Le spectre de masse est échantillonné régulièrement tous les 0.18 Da. Sous ces conditions, la taille du fichier brut décrivant l'analyse est d'environ 300 Mo. Une représentation du fichier obtenu pour des peptides concentrés à 0.2 µmol/l est donnée à la Figure 34.

Afin d'alléger les moyens informatiques nécessaires pour traiter cette image, nous nous intéresserons à un fragment de l'image totale entre 1750s et 2100s et 359.9091 Da et 519.9091 Da, soit environ 8‰ du fichier total. Cette extraction est réalisée à l'aide d'un logiciel développé au laboratoire, et comporte deux phases de transcription : la première passe du fichier au format propriétaire au format générique mzXML [89], la seconde convertit ce fichier en une matrice MATLAB de taille 723×881 après une étape d'interpolation suivant la direction chromatographique. En effet, les données sont échantillonnées régulièrement suivant la dimension spectrométrique mais irrégulièrement suivant la dimension chromatographique. Idéalement, notre modèle et notre méthode devrait prendre en compte cet échantillonnage irrégulier. Les prétraitements introduisent des modifications dans les données dont les conséquences sur les performances sont difficiles à prévoir. Cette image extraite est représentée sur la Figure 35. Sur cette image 3 peptides des 11 présents dans le mélange sont visibles. Leurs formules peptidiques sont TGPNLHGLFGR, MIFAGIK et EDLIAYLK.

Ces données ont l'avantage de présenter les pics bien visibles de 3 peptides sur une surface réduite. De plus, deux des pics se chevauchent partiellement, permettant de tester le comportement de l'algorithme dans cette situation.

Afin de simuler ces données, nous avons reporté diverses informations sur les pics de ces peptides dans le Tableau 2. De plus, notons que les écarts types chromatographiques et spectrométriques constatés sont respectivement de 6.4 s et 0.19 Da. Ces valeurs sont mesurées à partir de la largeur à mi-hauteur des pics sur des coupes de l'image.



Spectrogramme réel

Spectrogramme simulé

Figure 35 : détail du spectrogramme des peptides du cytochrome C.

Formule peptidique	Masse théorique (Da)	Nombre de charges observées	Rapport masse sur charge observée (Da)	Temps de rétention observés (s)
TGPNLHGLFGR	1168.32	3+	390.44	1817.6
MIFAGIK	779.00	2+	390.50	1837.5
EDLIAYLK	964.11	2+	483.06	2018.7

Tableau 2 : information sur les peptides étudiés.

5.1.2. Traitement de données simulées

5.1.2.1. Représentations possibles

Afin de pouvoir confronter les valeurs estimées à des valeurs théoriques, nous avons simulé des données reproduisant cette expérience. Chaque pic est représenté par une gaussienne. De plus, nous avons ajouté un bruit gaussien centré d'écart type 2000, soit d'inverse puissance $2.5.10^{-7}$ (RSB[1] 50 :1 ~ 34 dB). Les données simulées sont représentées sur la Figure 35.

Pour mieux visualiser les hauteurs des pics et le bruit rajouté, une projection du spectrogramme sur l'axe des temps de rétention et l'axe des masses est présentée Figure 36. Toutefois sur ces figures projetées, le niveau de bruit peut paraître plus important qu'il ne l'est réellement. Une vue en coupe peut donc avantageusement compléter les projections avec comme inconvénient de ne pouvoir visualiser qu'un peptide à la fois. Nous choisissons de mettre en valeur le pic du peptide TGPNLHGLFGR.

Une autre visualisation, couramment utilisée dans le domaine de la protéomique sous le nom de BPI (base peak intensity), permet de cumuler les avantages des deux visualisations précédentes. Il s'agit pour chaque temps de rétention, d'afficher la valeur maximale du spectre de masse (l'intensité du pic principal du spectre). Cette visualisation permet en général d'afficher tous les pics et d'afficher

[1] Rapport signal à bruit

un niveau « correct » de bruit (Figure 38). Cette visualisation n'est pas exempte de défauts, notamment pour les faibles valeurs où l'affichage est difficilement interprétable.

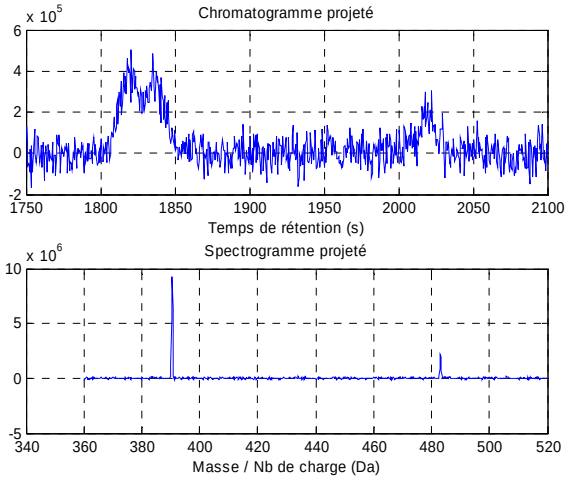


Figure 36 : projection des données simulées.

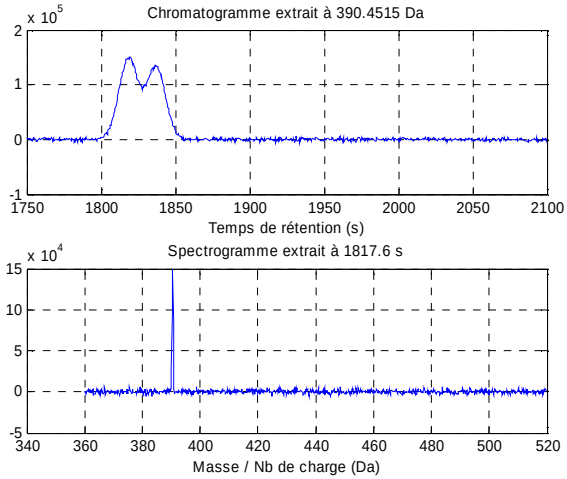


Figure 37 : coupe des données simulées.

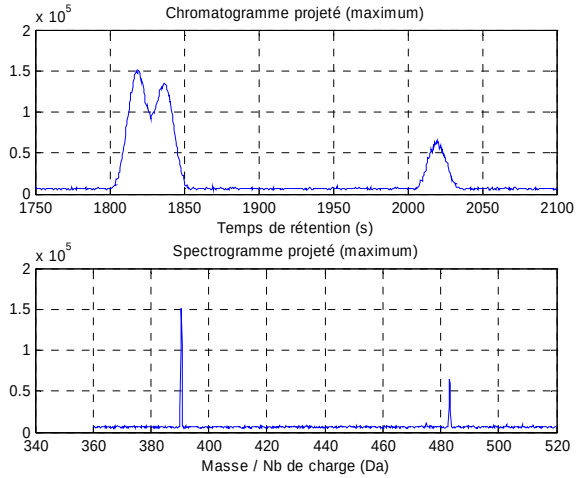


Figure 38 : base peak intensity, affichage de l'intensité maximale.

5.1.2.2. Déroulement de l'algorithme

Notre algorithme va estimer de façon fine les concentrations des peptides, la position des gaussiennes chromatographiques et l'inverse puissance du bruit. Comme tous les algorithmes, nous devons renseigner certains paramètres. Tout d'abord, nous devons indiquer les autres paramètres du modèle qui sont estimés par d'autres méthodes : les positions des gaussiennes spectrométriques, les largeurs des pics. Les gammes de temps et de masses considérées ainsi que les périodes d'échantillonnage correspondantes seront renseignées. Les valeurs numériques correspondantes ont été indiquées dans les deux parties précédentes.

Les paramètres secondaires du modèle étant fixés, nous indiquons à l'algorithme quelles informations *a priori* nous souhaitons injecter dans la méthode. Pour les concentrations et le bruit, nous choisissons de garder les distributions par défaut appelées distributions non informatives. Pour chaque position chromatographique, la distribution *a priori* est une distribution uniforme positionnée autour de la valeur théorique : 35 s avant et 17 s après. Nous utiliserons l'échantillonneur « toutes les positions en même temps » (section 4.5.3.1 page 76).

Enfin, l'échantillonneur de Gibbs a besoin de valeurs d'initialisation. Pour γ_b, nous choisissons un dixième de la valeur théorique. Les positions sont initialisées aux 2/3 de l'intervalle de la distribution *a priori*. Nous assignons la valeur des concentrations à 0.5.

La condition d'arrêt des échantillonneurs de Gibbs est un problème important et plusieurs solutions sont possibles. Nous pouvons par exemple faire tourner ces algorithmes jusqu'à ce que la valeur des estimateurs se stabilise. Pour des raisons de simplicité notre échantillonneur de Gibbs s'arrêtera au bout d'un nombre fixé d'itérations. Dans ces expériences nous avons obtenu de bons résultats avec 2000 itérations, soit une vingtaine de minutes de calcul sur un Pentium 4 cadencé à 3.2 GHz et possédant 2 Go de RAM. Le temps de calcul est très inférieur par rapport au temps de l'analyse (environ 1h pour l'analyse par LC-MS).

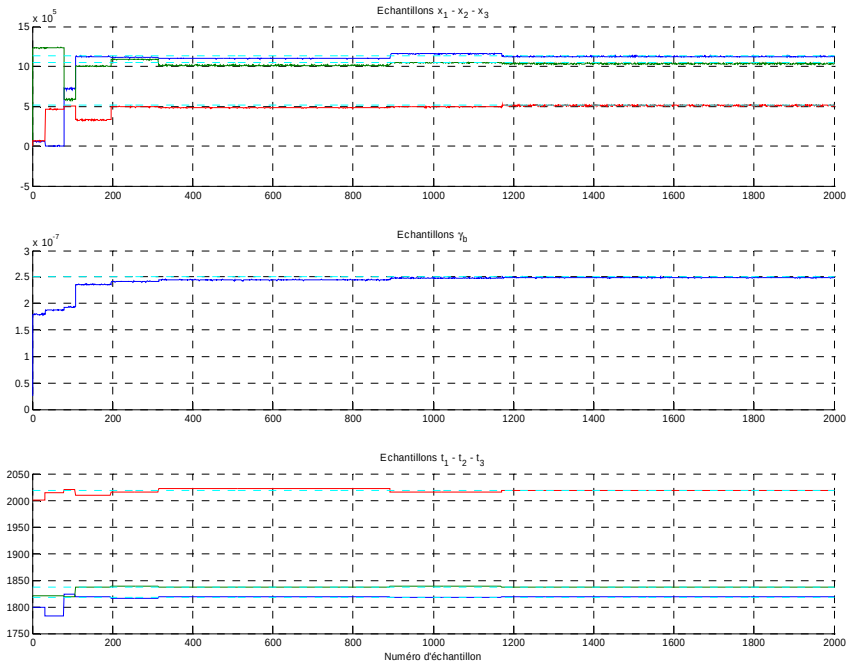


Figure 39 : trajectoires de l'échantillonneur de Gibbs.
Chaque trajectoire indique les valeurs proposées à chaque itération de l'algorithme. Pour les vecteurs estimés, leurs composantes sont représentées dans l'ordre par la courbe bleue, verte et rouge. Les droites cyans en pointillé indiquent les valeurs théoriques.

A chaque itération, l'échantillonneur de Gibbs propose une valeur pour les paramètres recherchés. Ces valeurs itératives forment des trajectoires représentées à la Figure 39. Plusieurs remarques peuvent être faites. Tout d'abord, ces trajectoires tendent vers les valeurs théoriques. Comme nous le verrons

plus loin, la théorie des méthodes MCMC nous indique que l'échantillonneur de Gibbs nous fournira des échantillons de la loi *a posteriori* au bout d'un certain temps. Sur cette figure, nous observons bien deux phases, les premiers échantillons correspondant aux échantillons de chauffe soit environ 400. Les derniers échantillons semblent mieux correspondre aux échantillons souhaités.

De plus, la Figure 39 présente des paliers. Mettons-les en évidence en confrontant sur la même figure les trajectoires des positions et des concentrations (Figure 40). A l'intérieur des paliers les valeurs des concentrations évoluent légèrement tandis que les valeurs des positions restent constantes. Lors d'une transition d'un palier à l'autre, toutes les valeurs évoluent. Ce comportement est dû à deux raisons : notre découpage de Gibbs, car nous tirons toutes les positions en même temps et à l'échantillonneur de Metropolis-Hastings utilisé pour tirer ces positions. Celui-ci peut soit accepter la nouvelle valeur des positions, soit la refuser et conserver la valeur précédente. Pendant un palier, l'échantillonneur de Metropolis-Hastings ne fait que refuser les échantillons, les positions n'évoluent donc pas. Par contre, les concentrations et le bruit continuent à évoluer car ils sont échantillonnés de façon directe sans phase d'acceptation. De plus notons qu'à l'intérieur d'un palier les échantillons de ces paramètres sont relativement proches les uns des autres, mais que de nouvelles positions font évoluer grandement ces paramètres. En effet, tant qu'il n'y a pas de changement de palier, la conditionnelle jointe des concentrations et du paramètre du bruit reste toujours la même. Les échantillons tirés sous ces lois ont donc les mêmes caractéristiques.

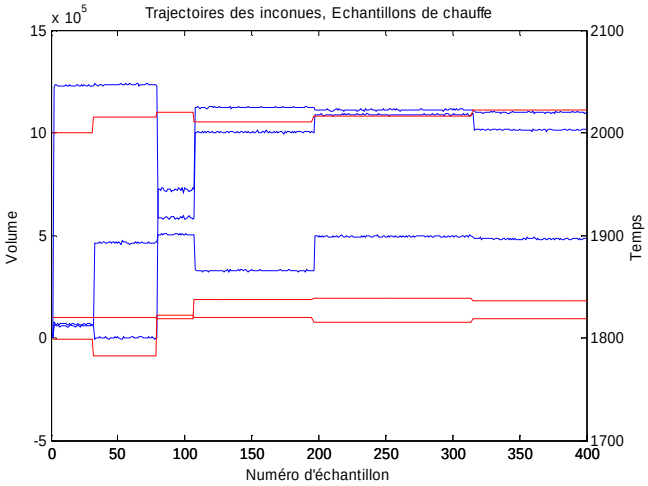


Figure 40 : échantillons de chauffe.
Les trajectoires des concentrations sont indiquées en bleu, les trajectoires des positions sont indiquées en rouge.

Nous avons choisi l'estimateur de la moyenne. Nous l'approchons par la moyenne empirique des échantillons produits. En théorie, les méthodes MCMC ne fournissent des échantillons de la loi désirée qu'après un nombre infini d'itérations. En pratique, elles fournissent de bons échantillons après un nombre fini d'itérations de chauffe. Ainsi, pour approcher l'estimateur de la moyenne, nous calculons la moyenne empirique des échantillons produits à chaque itération après avoir écarté un certain nombre d'échantillons de chauffe.

La question du nombre d'échantillons à écarter est un problème activement étudié. Cependant, en pratique, ce n'est pas un paramètre très sensible à régler. En effet, on pourrait ne pas écarter d'échantillons. En tirant suffisamment d'échantillons, l'influence des échantillons de chauffe dans le calcul de la moyenne tendra à être négligeable. Bien sur, il est plus efficace d'écarter les premiers échantillons pour obtenir plus rapidement une valeur utilisable. Nous pouvons donc soit écarter plus d'échantillons que nécessaire, soit prendre quelques échantillons de chauffe sans que leur influence soit significative. Bien sûr, il faudra garder suffisamment d'échantillons valables pour le calcul de la moyenne.

Sur les 2000 échantillons produits, nous choisissons de ne considérer que les 800 derniers échantillons. Nous avons obtenu de bons résultats sur toutes les données étudiées. Ceux-ci sont en parties représentés Figure 41. Les trajectoires fluctuent entre différentes valeurs admissibles, mais leur

comportement semble stationnaire. Notons que pendant cette période, l'échantillonneur de Metropolis-Hastings n'a pas accepté de nouvelle valeur. Nous savons qu'il a toujours une chance d'accepter une nouvelle valeur. En attendant suffisamment longtemps nous devrions pouvoir observer de nouvelles transitions.

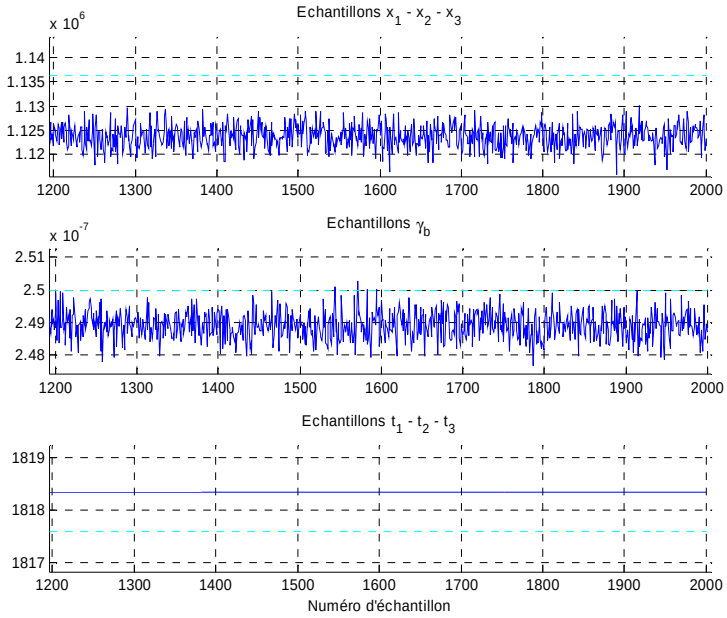


Figure 41 : échantillons utilisés pour calculer l'estimateur de la moyenne.
Détail de la Figure 39. Seules les trajectoires de x_1, γ_b et t_1 sont représentées.

	Valeur estimée	Valeur théorique	Erreur relative (%)
x_1	1.1238 E 6	1.1363 E 6	1.0999
x_2	1.0351 E 6	1.0544 E 6	1.8316
x_3	0.5110 E 6	0.5133 E 6	0.4556
t_1	1.8184 E 3	1.8176 E 3	0.0413
t_2	1.8370 E 3	1.8375 E 3	0.0257
t_3	2.0196 E 3	2.0187 E 3	0.0430
γ_b	2.489 E -7	2.500 E -7	0.4283

Tableau 3 : comparaison entre les valeurs estimées et les valeurs théoriques.

5.1.2.3. Valeurs estimées

Pour apprécier les performances de la méthode, nous pouvons soit comparer directement les valeurs estimées aux valeurs théoriques, soit comparer en sorties les données reconstruites aux données originales. Les valeurs estimées en écartant les 1200 premiers échantillons sont représentées au Tableau 3. Nous comparons dans les figures suivantes (Figure 42 à Figure 44) les données originales utilisées (en bleu) et reconstruites (en rouge). Pour générer les données reconstruites, nous avons utilisé le même simulateur nous ayant permis de simuler les données originales mais en utilisant les paramètres estimés. Les données reconstruites incluent une réalisation du bruit. Quelle que soit la représentation utilisée, les courbes s'ajustent parfaitement. Les seules légères disparités observées résultent des deux réalisations du bruit. Les performances sur données simulées sont donc excellentes. Nous allons maintenant tester l'algorithme sur des données réelles.

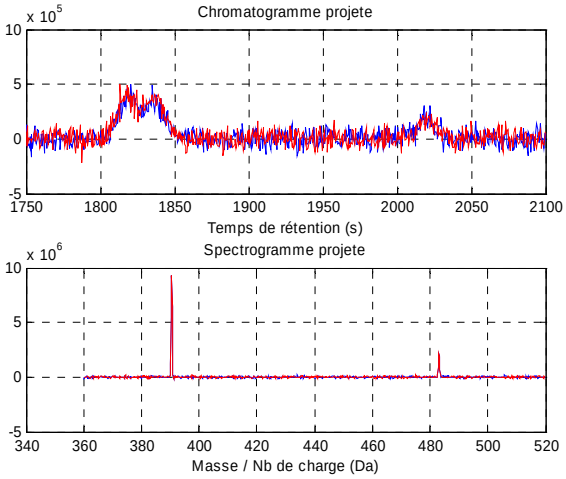


Figure 42 : comparaison des projections des données reconstruites et des données simulées.

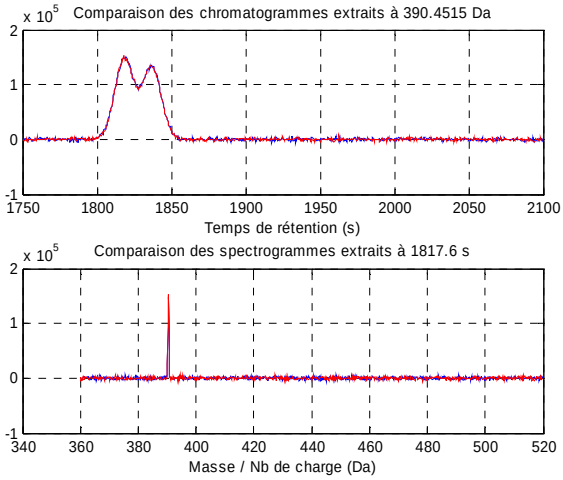


Figure 43 : comparaison des coupes des données reconstruites sur des données simulées.

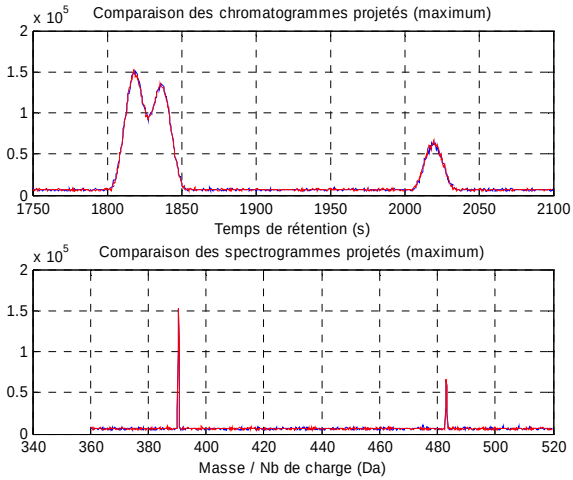


Figure 44 : comparaison des visualisations BPI des données reconstruites et des données simulées.

5.1.3. Traitement de données réelles

Nous disposons pour ces tests de plusieurs jeux de données. Ils ont été acquis par Emeline Mery à partir de plusieurs échantillons du digest de cytochrome C, plus ou moins dilué de façon à obtenir plusieurs concentrations : 0.2, 0.5 et 1 µmol/l. Tous les peptides ont la même concentration dans un échantillon. Nous disposons de 3 échantillons différents à 0.2 µmol/l. Nous avons fait tourner l'algorithme sur chacune de ces données. Nous examinerons en détail les trajectoires du premier échantillon à 0.2 µmol/l, appelé Yi7[1].

5.1.3.1. Déroulement de l'algorithme

Le comportement de l'algorithme sur données réelles est proche de celui observé sur les données simulées (Figure 45). Elles comportent une phase de chauffe puis une phase stationnaire. La phase de chauffe des trajectoires dure entre 300 et 1500 échantillons, selon notre degré de tolérance. Nous choisissons, comme pour les données simulées, de ne pas conserver les 1200 premiers échantillons et d'utiliser les autres pour estimer nos paramètres.

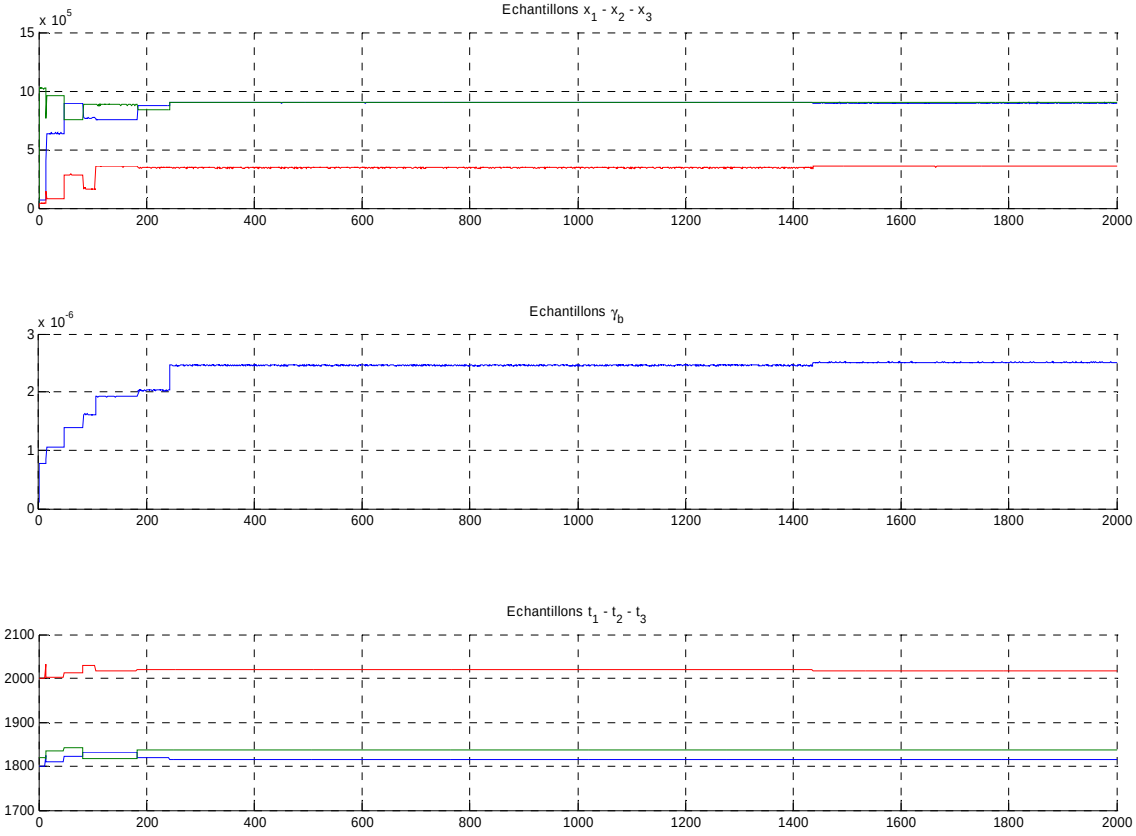


Figure 45 : trajectoires obtenues pour les données Yi7.

Pour estimer la performance de la méthode, nous ne disposons pas ici de valeur théorique des paramètres. Nous comparons donc sur les figures suivantes les données réelles (en bleu) et reconstruites (en rouge) à partir des paramètres estimés.

Les projections indiquent que les pics sont globalement aux bons endroits, mais les courbes ne correspondent pas aussi bien que pour les données simulées (Figure 46). En effet, on observe un offset sur les données réelles qui n'apparaissait pas dans les données simulées. Ceci est dû au fait que malgré le bruit, les données réelles ne sont jamais négatives, contrairement aux données simulées. Les valeurs négatives ne sont jamais compensées, ce qui provoque *in fine* un offset. Les visualisations suivantes (Figure 47 et Figure 48) permettent une meilleure représentation de l'adéquation de la solution aux

[1] Voir le Tableau 4 pour connaitre les concentrations des autres fichiers.

données. On constate que la solution reconstruite s'ajuste globalement bien aux données, ce qui laisse présager de bons résultats de quantification, malgré une légère dissymétrie des pics que les courbes gaussiennes ne peuvent retranscrire.

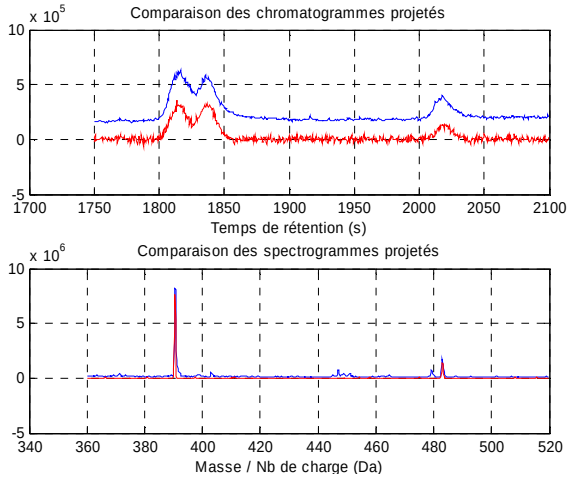


Figure 46 : projection des données réelles et reconstruites.

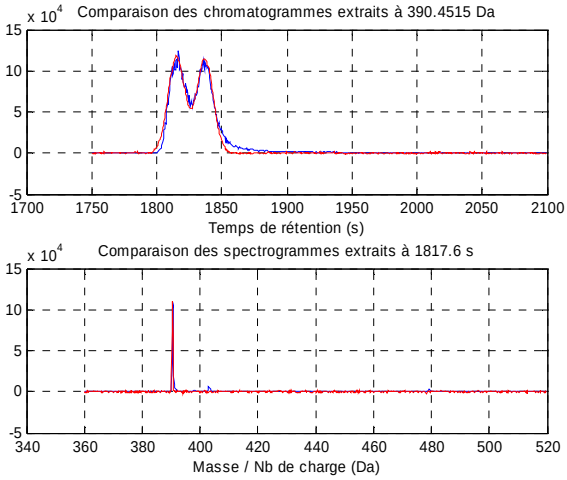


Figure 47 : coupe des données réelles et reconstruites.

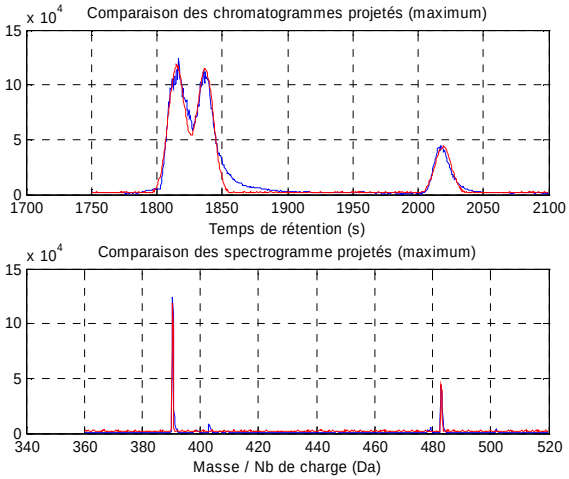


Figure 48 : visualisation BPI des données réelles et reconstruites.

Dans ce travail, nous n'avons pas cherché à estimer de façon particulière les paramètres des pics spectrométriques. Par rapport aux paramètres chromatographiques, ils sont réputés être bien plus stables. La Figure 49 agrandit ces pics que nous avons présentés sur les 3 dernières figures. Cette figure montre que nous devons rester vigilant et qu'une modélisation et une estimation plus poussée de ces pics spectrométriques peut aussi s'avérer nécessaire. Toutefois, en l'état, les performances sont très satisfaisantes comme nous le verrons à la section suivante.

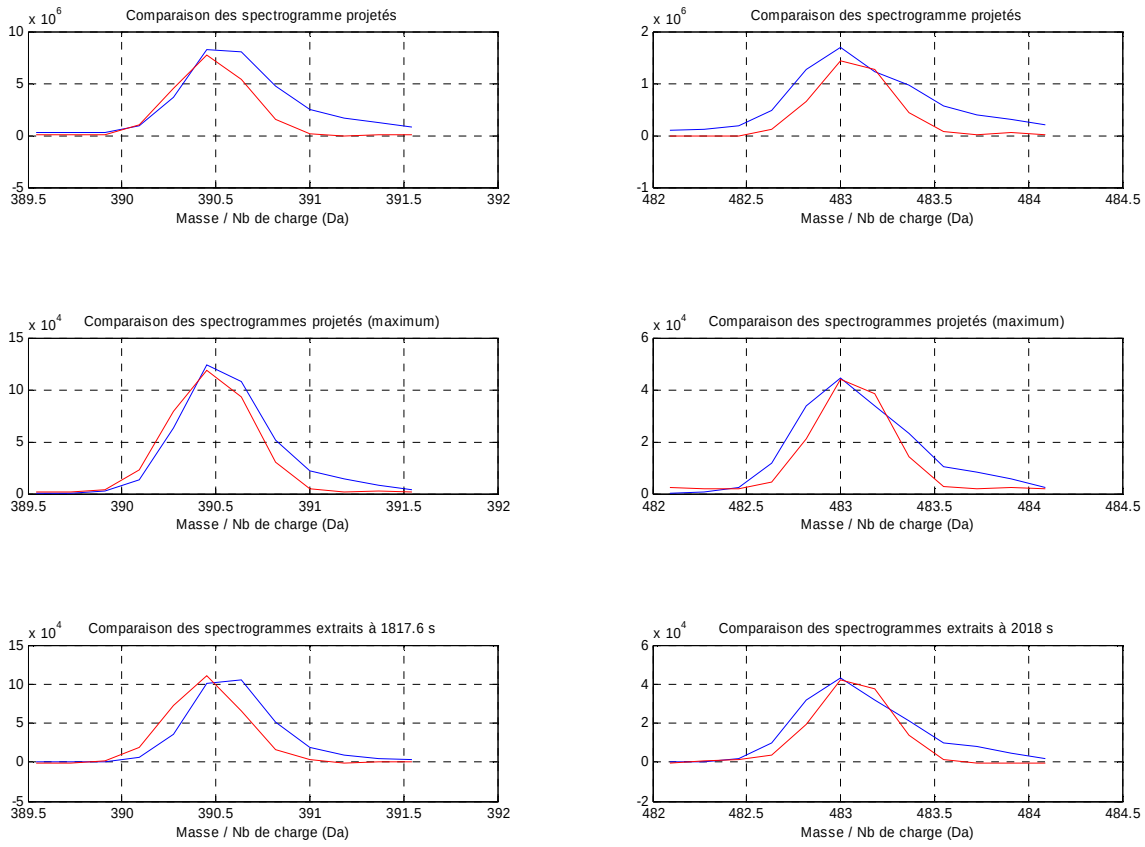


Figure 49 : zoom des deux pics spectrométriques suivant les 3 représentations.

5.1.3.2. Comparaison des expériences à différentes concentrations

Nous comparons maintenant les valeurs qui ont été estimées pour chacune des expériences. Sur chacune des données, les trajectoires se stabilisent. Les pics s'ajustent sur toutes les données de façon comparable aux figures de la section précédente. Les résultats du traitement sont représentés au Tableau 4.

Intéressons nous tout d'abord aux variables d'intérêt x_1, x_2 et x_3 (Figure 50). Les courbes représentées indiquent que ces variables sont linéaires par rapport aux concentrations. De même, les valeurs estimées sur des expériences de même concentration sont relativement stables (±5% de variation). Ces performances permettent déjà d'effectuer des mesures de concentration sur des échantillons inconnus. On pourra se servir de cette droite d'étalonnage pour associer une concentration aux valeurs estimées dans des expériences où nous ne connaîtrons pas la concentration des peptides.

Cependant, l'hypothèse simplificatrice réalisée qui assignait les gains ξ_i à 1, associant ainsi le volume des pics aux concentrations, n'est pas valide. En effet le volume est le produit du gain et de la concentration. La linéarité observée sur la Figure 50 impose que sur cette série d'expériences simples, le gain soit constant d'une expérience à l'autre. De plus, le gain ξ_i correspond aux coefficients directeurs de chacune des droites de régression. Notons que ce gain est spécifique à chaque peptide puisque nous observons un coefficient directeur différent pour chacun des peptides étudiés.

Echantillon	Yi7	Yi8	Yi9	Yi11	Yi14
concentration (µmol/l)	0,2	0,2	0,2	0,5	1
x_1	8,98E+05	1,01E+06	1,03E+06	2,49E+06	5,10E+06
x_2	9,03E+05	7,99E+05	8,40E+05	1,73E+06	3,22E+06
x_3	3,51E+05	3,33E+05	3,40E+05	8,78E+05	2,05E+06
γ_b	2,50E-06	2,55E-06	2,88E-06	5,11E-05	1,64E-07
t_1	1815,2	1812,9	1812,4	1807,6	1796,3
t_2	1837,1	1832,7	1832,2	1828,8	1815,4
t_3	2019,1	2010,2	2013,8	2008,4	1999,9

Tableau 4 : résultats de l'algorithme sur données réelles.

Concernant les autres paramètres, notons que la variation maximale de la position des pics n'est que de 22 s sur cette expérience, alors que nous avions prévu un intervalle de recherche de 52 s. Nous pourrions donc envisager d'utiliser un intervalle de recherche plus restreint, ce qui accélérerait la convergence de la chaîne de Markov.

Nous avons également observé dans certaines expériences un phénomène de permutation, les paramètres estimés portant le premier indice décrivaient le second pic, au lieu du premier. Pour prévenir ce risque nous disposons déjà des informations *a priori* permettant de situer les positions dans une certaine gamme de valeurs. Tant que celles-ci ne se superposent pas, nous n'observeront pas de problème de permutation. Si comme dans notre cas, les marges ne sont pas disjointes, il faut développer d'autres solutions. Vincent Mazet propose une série de solution pour résoudre ce problème [84]. On peut soit résoudre ce problème *a priori* en imposant une relation d'ordre aux temps de rétention, soit résoudre le problème *a posteriori* en réindexant les échantillons. Dans cette série d'expériences, ce phénomène n'est pas arrivé souvent et seulement dans la phase de chauffe. Nous avons appliqué à la fin de l'algorithme une simple permutation des indices de façon à ordonner les positions des peptides de façon croissante, ce qui a suffit à régler le problème.

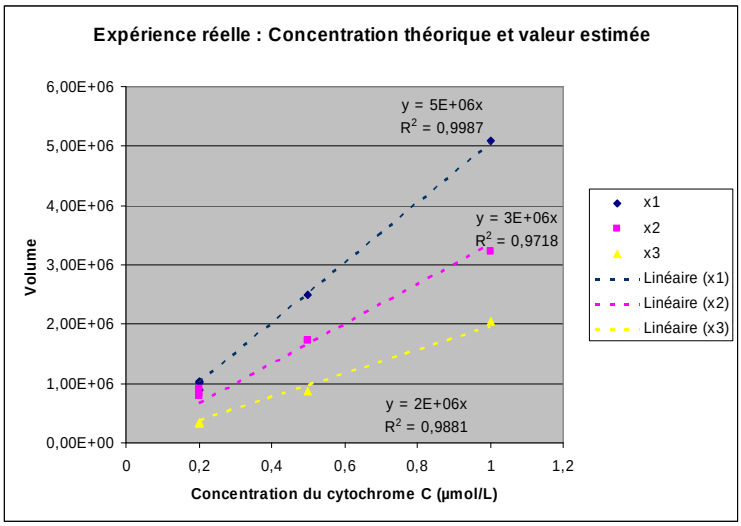


Figure 50 : volume des pics en fonction de la concentration des peptides du Cytochrome C.

5.1.4. Conclusion

La méthode développée permet d'estimer conjointement les volumes, les positions des pics et l'inverse puissance du bruit. Nous avons testé l'algorithme sur un ensemble de données simulées et réelles. Les performances observées sont très satisfaisantes dans les deux cas. Ces données décrivaient

un cas difficile où deux pics se chevauchent partiellement. Les performances observées ont été similaires à celles observées dans le cas où le pic est bien séparé de son voisinage.

Cependant, certaines limites de la méthode ont été observées. Les pics réels présentaient une dissymétrie dont le modèle ne rend pas compte. Pour gagner en précision, des modèles dissymétriques comme ceux utilisant l'exponentially modified Gaussian (EMG) [65] ou d'autres modèles plus proches de la physique gagneront donc à être étudiés. De plus, en l'absence de protéines d'étalonnage, nous n'avons pas pu estimer directement les concentrations souhaitées et nous avons dû nous restreindre à estimer les volumes des pics.

En réalisant l'estimation des volumes pour une série d'expériences simples, nous avons observé une relation linéaire entre les concentrations et les volumes. Ceci nous a permis de déduire le gain correspondant à chaque peptide. L'hypothèse du gain unitaire que nous avons formulée s'est alors révélée invalide, comme nous le prévoyions. Dans ce cas linéaire, l'estimation des concentrations est possible à l'aide d'une courbe d'étalonnage. Cependant, nous verrons dans la section suivante que le gain n'est pas toujours constant.

Nous avons observé des permutations d'indice et proposé une première méthode simple pour y remédier. Cependant, cette solution ne fonctionne pas dans le cas général. Les permutations d'indices peuvent advenir plusieurs fois à l'intérieur même de la chaîne de Markov. A la Figure 45, nous observons deux permutations aux alentours des échantillons 90 et 190. Si des permutations survenaient hors de la période de chauffe, nos estimées, qui moyennent ces échantillons, moyenneraient des échantillons correspondant à des pics différents, fournissant de mauvaises estimées. Dans notre cas, notre choix de découpage de l'algorithme de Gibbs et l'algorithme de Metropolis-Hastings font que la probabilité pour que cela arrive est faible. Des méthodes plus complexes pourront être utilisées dans un algorithme plus avancé. Par exemple, nous pourrons introduire dans la distribution *a priori* la relation d'ordre existante sur les positions des pics, un algorithme de réindexation [84] permettant d'associer à chaque échantillon tiré le bon indice, ou encore d'améliorer l'identification des pics, par exemple en utilisant des massifs isotopiques pour dissocier les pics ou les données issues d'un spectre MS/MS.

5.2. Analyse des toxines du Staphylocoque doré dans l'urine

Dans cette section, nous estimerons la concentration d'une toxine bactérienne dans de l'urine humaine. Nous utilisons les données de Virginie Brun *et al.* présentées dans [63]. Nous nous concentrerons sur l'Entérotoxine A du Staphylocoque doré (SEA) qui est une protéine à l'origine d'une pathologie sévère chez l'homme [90]. Nous disposons de protéines marquées ; nous pouvons donc estimer l'ensemble des paramètres étudiés au chapitre 4, la concentration x_p de la SEA, le gain ξ_i et la position chromatographique t_i de chaque peptide et le paramètre du bruit γ_b. De plus, nous utiliserons dans cette section l'échantillonneur « une position à la fois ». Nous avons testé cet algorithme sur données simulées et les performances ont été similaires à celles présentées à la section 5.1.2. Nous présentons dans cette section les résultats obtenus sur données réelles.

5.2.1. Protocole des expériences

L'expérience suit le protocole décrit dans [63]. L'échantillon analysé est constitué de l'urine d'une femme saine de 30 ans, à laquelle on ajoute de la SEA recombinante commercialisée par Toxin Technology (Sarasota, FL), et de la SEA marquée par le procédé PSAQ (section 3.1.1.2 page 49). Plusieurs concentrations de SEA sont étudiées.

La SEA marquée est produite de la façon suivante. Le gène de la SEA est répliqué à partir du génome du centre national de référence des staphylocoques par amplification PCR. La protéine correspondante est fabriquée *in vitro* à l'aide d'arginine et de lysine alourdies (constituées de carbone 13 et d'azote 15). Les protéines sont ensuite purifiées à l'aide d'une colonne d'affinité, puis quantifiées par une analyse des acides aminés.

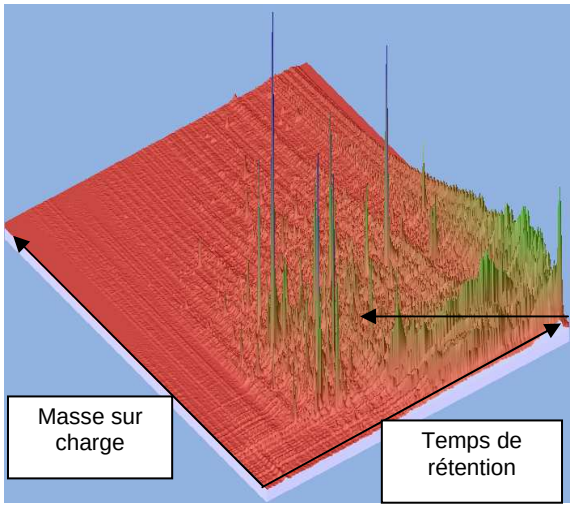


Figure 51 : exemple de spectrogramme obtenu.
Cette image a été obtenue à partir du fichier global UF7376 en utilisant le logiciel de visualisation MSight du Swiss Institute of Bioinformatics.

Des échantillons de 100 µl du mélange sont injectés dans la chaîne analytique. Celle-ci consiste en 3 étapes de préparation : préfractionement sur une résine Strataclean de Stratagene, purification par électrophorèse SDS-PAGE Invitrogen et digestion sur gel par la trypsine de Promega. Le gel obtenu est ensuite lavé dans une solution contenant 94.8% d'H_2O, 5% d'ACN et 0.2% AF. La solution est ensuite analysée à l'aide d'une colonne de chromatographie Dionex couplée par electrospray à un spectromètre de masse Waters.

Plus précisément, la solution est d'abord concentrée sur une précolonne Dionex Pepmap C18 avec un diamètre interne de 300 µm et une longueur de 5 cm. Puis, elle est analysée sur une nanocolonne Dionex Pepmap C18 avec un diamètre interne de 75 µm et une longueur de 15 cm. On utilise la colonne en mode gradient. Le premier solvant est formé de 10% d'ACN, 0.1% d'AF et de 89.9% d'H_2O. Le second solvant est formé de 80% d'ACN, 0.08% d'AF et de 19.92% d'H_2O. Le gradient dure 60 min et le débit de la pompe est de 200 nl/min.

Le spectromètre de masse est un QTOF Ultima de Waters. Il produit des spectres de 400 Da à 1600 Da toutes les secondes. Les spectromètres de masse sont échantillonnés en moyenne tout les 0.0085 Da. Sous ces conditions, la taille d'un fichier brut est de 1.7 Go. Une visualisation du spectrogramme total, obtenue à partir du logiciel MSight est présenté Figure 51. Le spectrogramme de l'urine contient un nombre de pics beaucoup plus important que le spectrogramme présenté Figure 34 page 82.

5.2.2. Prétraitements

Le premier de ces traitements consiste à convertir les fichiers bruts du constructeur en matrices exploitables sous MATLAB. Tout d'abord nous convertissons le fichier propriétaire d'origine en fichier générique mzXML [89]. Ensuite nous convertissons ce dernier fichier en matrice MATLAB. Les données sont échantillonnées régulièrement suivant la dimension chromatographique mais irrégulièrement suivant la dimension spectrométrique. Une étape d'interpolation suit l'étape de conversion de façon à obtenir une grille régulière pour la représentation matricielle. Encore une fois, idéalement cet échantillonnage irrégulier devrait être directement pris en compte dans le modèle et la méthode.

Afin d'alléger les moyens informatiques nécessaires pour traiter cette image, nous nous concentrons sur une zone comprise entre 1564.4 s (~26 min) et 1796.3 s (~30 min) et entre 688.35 Da et 705.35 Da, soit environ 1‰ du fichier total. Cette zone correspond au peptide NVTVQELDLQAR de la protéine SEA chargé deux fois. Une représentation de cette zone est donnée Figure 52. Les deux signaux les plus importants correspondent au peptide NVTVQELDLQAR et à son marqueur isotopique. Notons que ces deux massifs ont le même temps de rétention (Figure 55) et que la

résolution du spectromètre de masse QTOF est suffisante pour distinguer chaque élément du massif isotopique correspondant à chaque neutron supplémentaire (voir section 2.4.3.2 page 36). Les autres pics présents dans la zone correspondent à des contaminants inconnus probablement issus de l'urine.

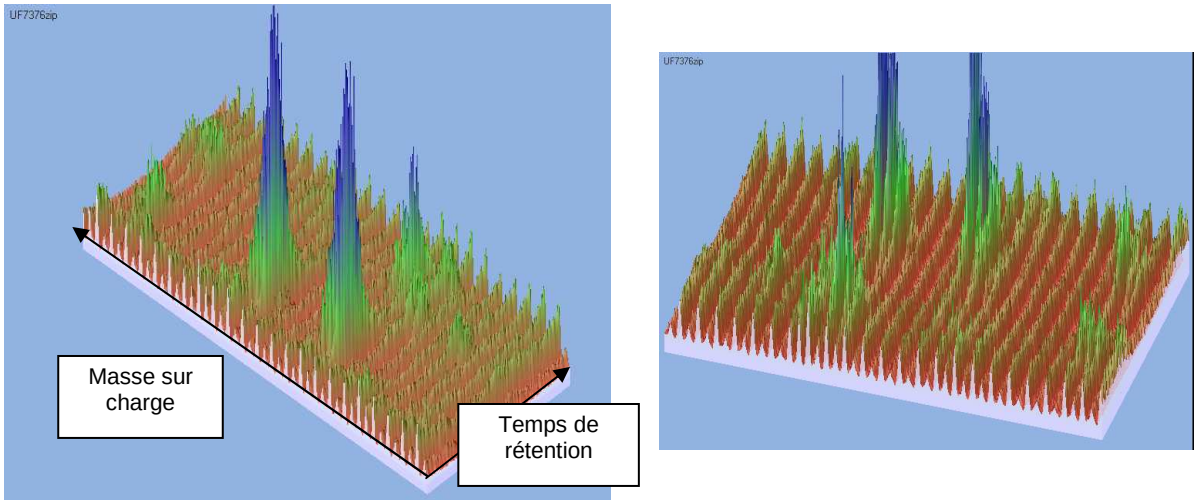


Figure 52 : zone d'étude.

De plus, cette figure fait apparaître un signal périodique qui s'ajoute aux signaux habituels des peptides. Il s'agit d'un signal, formant un ensemble de pics tous les demi Dalton, dans la dimension spectrométrique, mais constant dans la dimension chromatographique. Ce signal s'additionne au signal du peptide recherché. Il est parfois appelé bruit chimique et se comporte comme une ligne de base. Son origine semble lié a des complexes de molécules de solvant. Si ce signal peut être produit par les chaîne d'analyse avec un electrospray [91], des signaux similaires semblent être générés par les instruments avec ionisation MALDI [92]. Il apparaît sur toutes les expériences que nous avons réalisées avec de l'urine. Puisque notre modèle ne prend pas en compte ce signal supplémentaire, nous allons le supprimer sur nos données avant d'utiliser notre méthode de quantification.

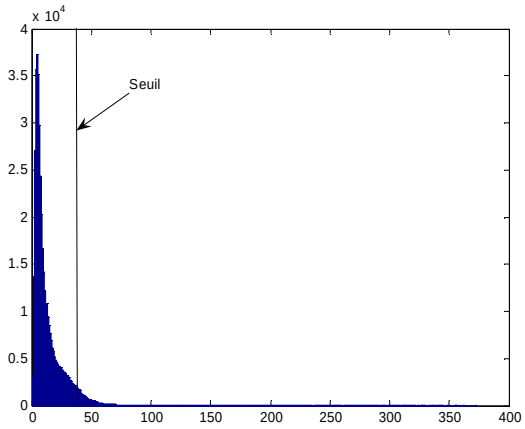


Figure 53 : histogramme des valeurs du spectrogramme et seuil choisi.

Nous constatons sur chaque fichier de données la présence d'une ligne de base qu'il est nécessaire de traiter. Nous allons pour cela utiliser une méthode simple en estimant une ligne de base moyenne. Cette ligne de base est obtenue en moyennant les spectres de masse du spectrogramme suivant la dimension chromatographique, après avoir exclu les zones contenant un pic. Ces zones sont déterminées par seuillage. Pour estimer le seuil, nous traçons l'histogramme des valeurs contenues dans le spectrogramme original (Figure 53). Le pic principal correspond aux amplitudes des pics du bruit. En prenant l'amplitude correspondant à 90% de la somme cumulative de cet histogramme, on obtient un seuil correct permettant de différencier le signal utile du bruit. Une fois la ligne de base moyenne obtenue, nous la soustrayons sur l'ensemble du spectrogramme. La dernière étape des prétraitements consiste à écrêter les valeurs négatives. L'effet de ces prétraitements est représenté

Figure 54 et Figure 55. Le prétraitement enlève donc la majorité du bruit chimique, mais conserve les pics d'intérêt.

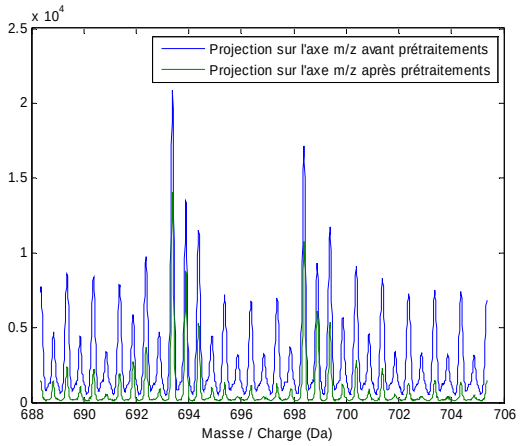


Figure 54 : comparaison des projections avant et après prétraitement.

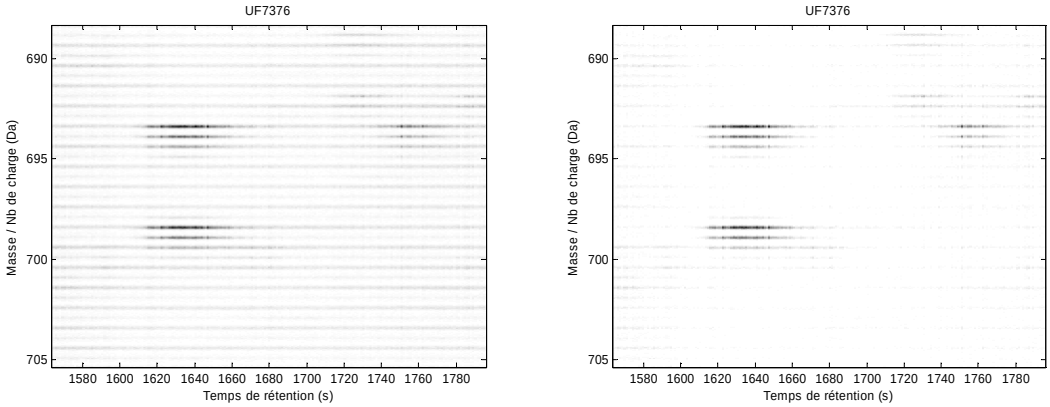


Figure 55 : comparaison des spectrogrammes avant et après prétraitement.

Protéine parente	Ref. Swiss Prot	Formule du peptide considéré	Masse théorique	Nombre de charges observées	Rapport masse sur charge observée	Temps de rétention observé
SEA	P0A0L2	NVTVQELDLQAR	1385.54	2+	693.35	~28 min

Tableau 5 : information sur le peptide de la SEA étudié.

5.2.3. Evaluation comparative des performances

5.2.3.1. Traitement de données expérimentales

Nous testons notre algorithme sur 12 expériences où nous injectons différentes quantités de SEA, mais toujours la même quantité de SEA marquée, que nous noterons ci-après SEA* (Tableau 6). Pour chaque concentration 3 expérimentations seront faites. L'algorithme sera lancé sur les 12 fichiers de données.

Expérience numéro	Nom expérience	Masse injectée de SEA (ng)	Masse injectée de SEA* (ng)
1	UF7425zip	1,2	5,0
2	UF7424zip	1,2	5,0
3	UF7423zip	1,2	5,0
4	UF7407zip	2,3	5,0
5	UF7408zip	2,3	5,0
6	UF7409zip	2,3	5,0
7	UF7365zip	4,6	5,0
8	UF7374zip	4,6	5,0
9	UF7376zip	4,6	5,0
10	UF7411zip	9,2	5,0
11	UF7414zip	9,2	5,0
12	UF7417zip	9,2	5,0

Tableau 6 : expérience et quantité de protéine injectée.

Les paramètres de notre méthode seront identiques pour tous les fichiers traités. La probabilité *a* priori de t_i sera choisie de façon à rechercher les positions entre 27 min (1620 s) et 29 min (1740 s). Nous choisissons de faire tourner l'algorithme pendant 200 itérations puis nous éliminons les 100 premières itérations en tant qu'échantillons de chauffe. En incluant les prétraitements, le calcul prend environ 1 min, mais seulement 3 s si on s'intéresse uniquement à la boucle principale de Gibbs. Pour cette étude, nous avons utilisé un Dual-Core Opteron 180 cadencé à 2.4 GHz et ayant 2 Go de RAM. Le temps de calcul est encore négligeable par rapport au temps d'analyse (environ 1h pour l'analyse et 1 journée si on inclut les prétraitements). Les paramètres secondaires du modèle sont estimés sur une expérience témoin de la façon décrite à la section page 4.2 page 63. L'initialisation de l'algorithme est quelconque.

Nous utilisons les quatre méthodes décrites au chapitre 3 pour comparer nos résultats. Les résultats obtenus sont présentés Tableau 7. Les expériences dont le texte est mis en gras seront commentées plus loin.

Comparons tout d'abord les résultats de notre méthode avec la méthode de la somme proposée par Virginie Brun dans [63]. Nous faisons ici référence au traitement numérique proposé dans l'article et non à la méthode de marquage utilisée dans tous les cas puisque nous utilisons les mêmes données. Pour comparer les deux méthodes, nous utilisons la même méthode que celle utilisée par V. Brun. Nous comparons donc les concentrations estimées par chaque méthode aux concentrations données par une méthode de référence (il s'agit de l'analyse des acides aminés *cf.* [63]). Pour juger de la proximité des méthodes testées avec la méthode de référence, nous effectuons une régression entre les valeurs estimées et les valeurs de références. La droite obtenue devra idéalement avoir une pente de 1 et un coefficient de détermination R^2 de 1[1]. Le coefficient directeur de la droite est un indicateur des différences systématiques entre la méthode testée et la méthode de référence. Le coefficient de détermination indique la dispersion des mesures. La Figure 56 donne une représentation graphique des valeurs estimées comparées à la mesure de référence. Pour chaque courbe, la droite de régression passant par l'origine est présentée. Au premier abord, les deux méthodes sont qualitativement semblables. Leur coefficient directeur est proche de l'unité et certains points de mesure semblent plutôt éloignés de la droite de régression. En regardant de plus près, la pente se rapproche de l'unité pour les méthodes de plus en plus performantes et on constate une amélioration du coefficient de détermination R^2. Notre méthode produit donc des résultats plus cohérents et moins dispersés.

[1] Le coefficient de détermination R^2 est le quotient de deux quantités. Le diviseur est la variance des ordonnées des points considérés, le dividende est la variance des ordonnées des projetés de ces points sur la droite de régression. Idéalement, ces deux quantités sont égales. Le coefficient de régression est toujours inférieur à 1.

Expérience numéro	Nom expérience	Concentration de référence	Méthode de la somme	PLS	NPLS	Bayésien
1	UF7425zip	1,2	1,09	1	0,99	1,11
2	UF7424zip	1,2	1,43	1,19	1,19	1,40
3	**UF7423zip**	**1,2**	**1,16**	**0,91**	**0,91**	**1,14**
4	UF7407zip	2,3	2,48	2,39	2,38	2,63
5	UF7408zip	2,3	2,67	2,59	2,59	2,76
6	UF7409zip	2,3	2,66	2,36	2,36	2,72
7	UF7365zip	4,6	4,71	4,97	4,98	5,27
8	UF7374zip	4,6	4,53	5,19	5,18	4,93
9	**UF7376zip**	**4,6**	**4,86**	**4,27**	**4,27**	**5,35**
10	**UF7411zip**	**9,2**	**7,69**	**8,61**	**8,63**	**8,54**
11	UF7414zip	9,2	7,98	10,15	10,15	8,74
12	UF7417zip	9,2	8,06	8,09	8,1	8,85
Ecart type moyen			0,16	0,45	0,85	0,15
Coefficient de variation moyen			6,2%	10,2%	10,3%	5,4%

Tableau 7 : résultats des différentes méthodes d'estimation.

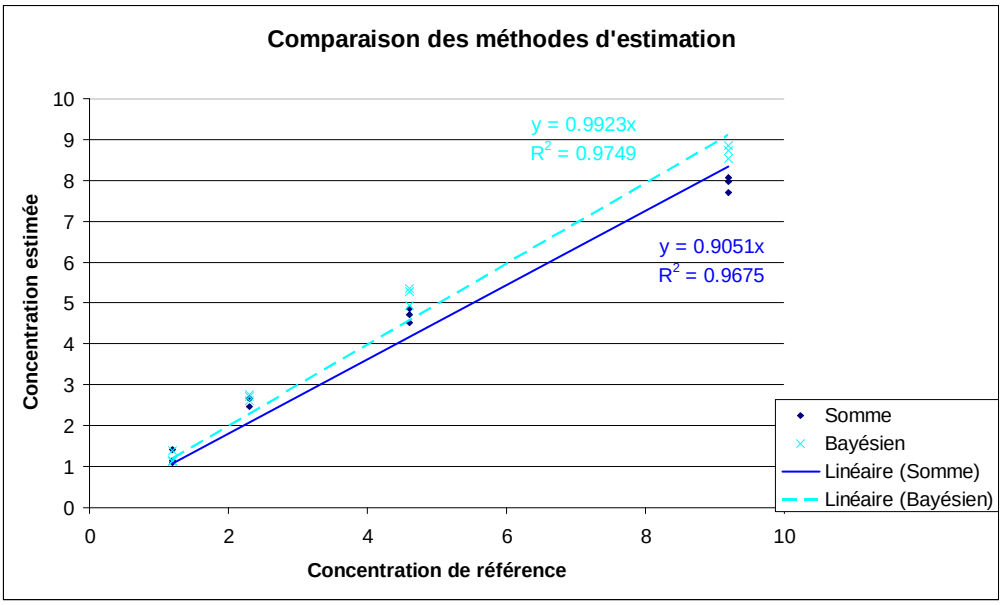


Figure 56 : comparaison de la méthode de la somme et de la méthode proposée.

Si nous traçons maintenant la droite de régression correspondant aux méthodes d'analyse factorielle, elles se confondent avec celle produite par notre méthode. Pour les dissocier, nous réalisons un zoom centré sur les derniers points de mesure (Figure 57). Le coefficient directeur et le coefficient de régression des méthodes PLS et N-PLS se révèlent être assez similaires. Nous notons une légère amélioration pour notre méthode. Si l'amélioration des performances est modeste sur ces données, notre méthode inclut directement des aspects auto-adaptatif et auto-calibrant, ce qui laisse espérer de meilleures performances dans des cas plus complexes.

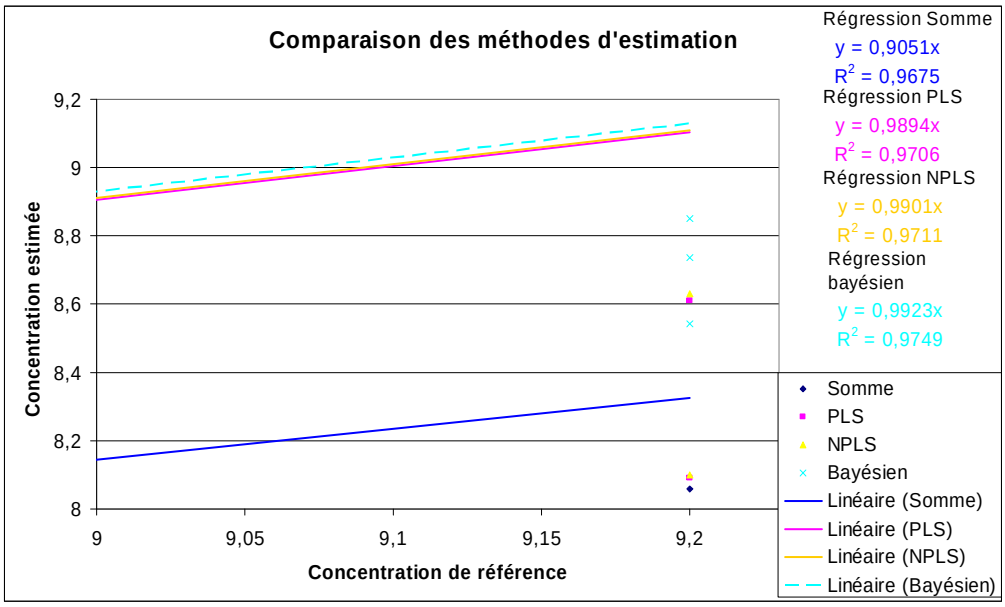


Figure 57 : comparaison de toutes les méthodes d'estimation.

Visualisons maintenant en détail les paramètres estimés pour quelques expériences correspondant à certains points caractéristiques de la courbe précédente. Nous nous intéresserons particulièrement aux expériences numéros 9 et 10 indiquées en gras dans le Tableau 7. Ces expériences désignent les points les plus éloignés de la droite de régression et correspondent respectivement à la quantité estimée la plus grande pour la série des expériences à 4.6 pg, et la quantité la plus faible pour la série d'expérience à 9.2 pg. Nous visualisons également l'expérience numéro 3 située sur la courbe de régression.

Pour apprécier la valeur des paramètres estimés, nous comparons les données reconstruites à partir des paramètres estimés et les données réelles. Plus précisément, nous visualiserons une série de coupes effectuées sur ces données (Figure 58 à Figure 60). Pour chaque expérience, nous visualiserons 3 coupes. La première et la dernière d'entre elles correspondent aux coupes suivant la direction chromatographique réalisées respectivement au point le plus haut du signal du peptide d'intérêt et du peptide marqué. La dernière a été effectuée suivant la direction des masses et correspond également au point le plus haut du massif (Figure 58). Pour les 3 expériences visualisées, nous observons un excellent ajustement du modèle aux données. De plus, les positions chromatographiques et les volumes des pics sont bien estimés.

Parallèlement aux estimations des concentrations, nous constatons les variations des positions et des gains sur notre série d'expériences (Figure 61). Ces deux paramètres varient en effet grandement d'une expérience à l'autre. Les expériences décrivant une même quantité de protéines ne se démarquent pas des expériences faites à des concentrations différentes, et nous ne voyons pas de tendance spécifique se dessiner avec l'augmentation de la quantité. La variation du gain est de l'ordre de sa valeur moyenne, et la variation de la position est de l'ordre de la minute.

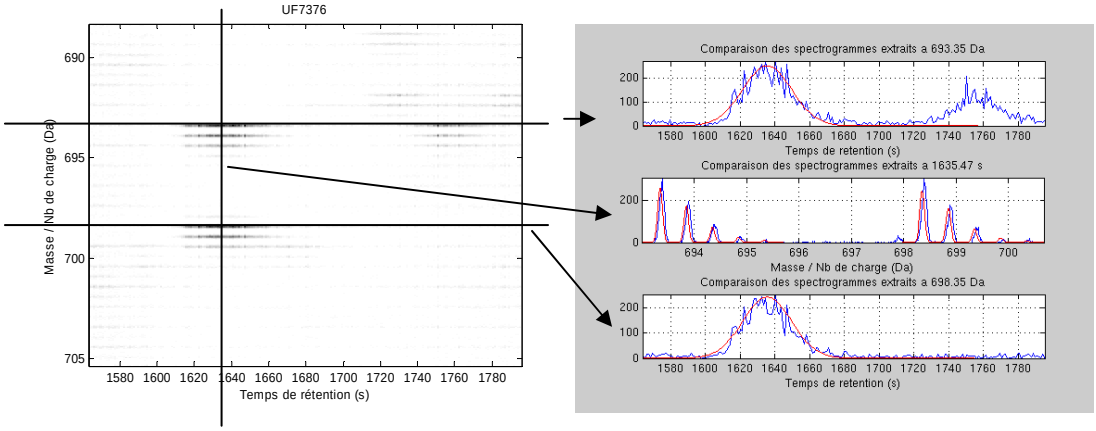


Figure 58 : données réelles et reconstruites de l'expérience numéro 9.

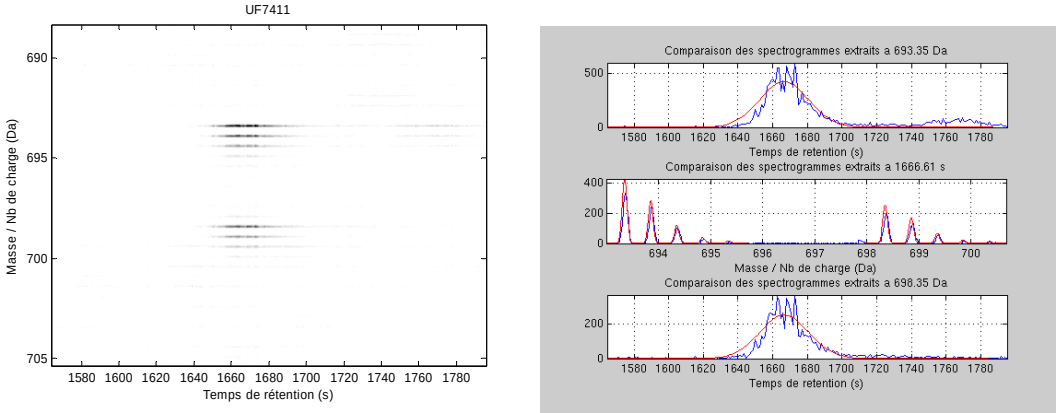


Figure 59 : données réelles et reconstruites de l'expérience numéro 10.

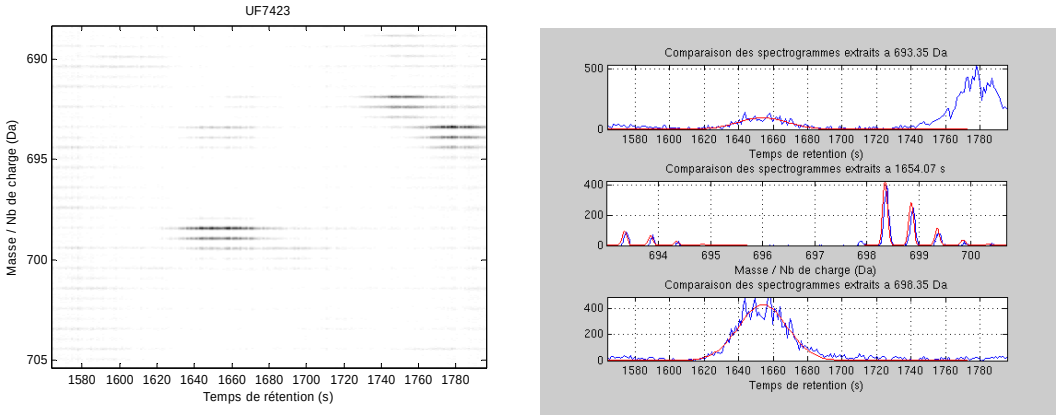


Figure 60 : données réelles et reconstruites de l'expérience numéro 3.

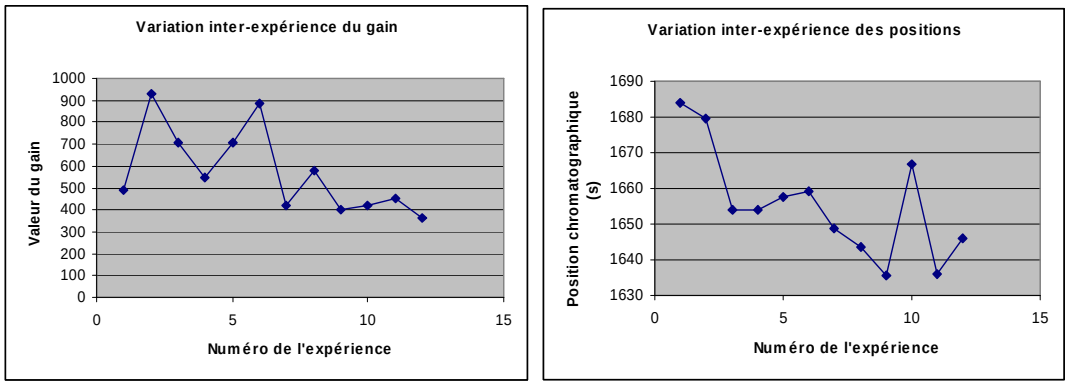


Figure 61 : variations inter-expériences du gain et de la position du pic.

5.2.3.2. Traitement de données simulées

Dans cette section, nous présentons des résultats complémentaires obtenus par Laurent Gerfault. Ces résultats ont pour but de montrer les différences entre les différentes méthodes de quantification, dans des situations difficiles. En effet, si les résultats obtenus par les différentes méthodes à la section 5.2.3.1 sont relativement proches les uns des autres, les différences entre méthodes seront plus marquées dans des situations plus difficiles (rapport bruits sur signal important, concentrations plus faibles, bruit plus important, chevauchements). A l'opposé, ces différences se réduiront dans les cas plus faciles (où il n'y a aucun bruit et aucun chevauchement par exemple). Les données présentées sont obtenues sur des données simulées représentant les données de la section 5.2. Nous reproduisons les expériences pour plusieurs valeurs de bruit. Les résultats obtenus sont représentés sur la Figure 62.

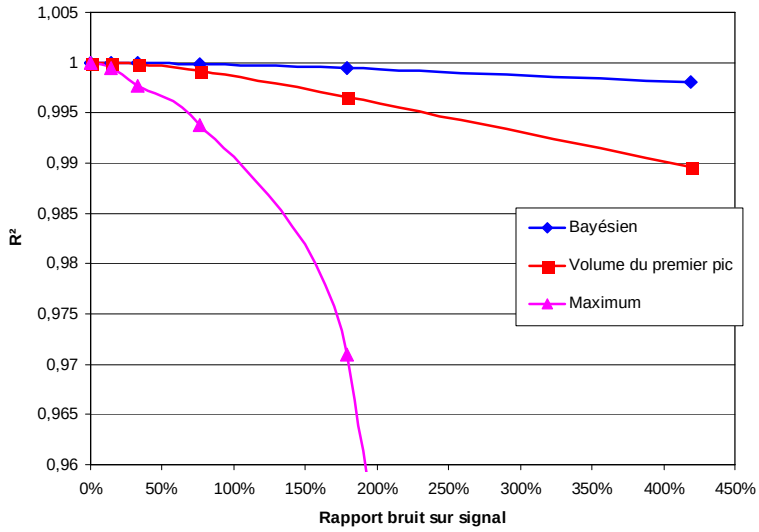


Figure 62 : performances des méthodes par rapport au bruit.

Cette figure visualise la précision de la méthode ou la dispersion des résultats à travers le coefficient de détermination R^2, de façon similaire à nos validations de la section précédente. Quand il n'y a pas de bruit, le coefficient de détermination vaut 1 et toutes les méthodes renvoient la bonne valeur sans aucune dispersion des résultats. Plus le bruit augmente, plus la dispersion entre les mesures augmente. Mais cette diminution des performances dépendra de la méthode de quantification utilisée. On voit ainsi que la méthode estimant les concentrations à l'aide du maximum des pics est la première à décrocher. La méthode du volume du premier pic qui moyenne les valeurs est plus robuste que la méthode du maximum des pics. C'est notre méthode qui résiste le mieux à l'augmentation du niveau de bruit. Ceci illustre le gain en robustesse apporté par notre méthode.

Par ailleurs, on peut constater que pour les valeurs de bruit sur signal testées jusqu'à présent, c'est-à-dire inférieures à 50 %, la différence entre les méthodes est faible. Il sera donc intéressant de voir le comportement de la méthode comparée à d'autres méthodes dans des cas réels plus difficiles à proximité de la limite de détection.

5.2.4. Conclusion

Dans cette section nous avons disposé de molécules d'étalonnage. Ainsi, nous avons pu tester la méthode du chapitre 4 dans son intégralité. Pour chaque expérience testée, et malgré une variation importante des paramètres, nous avons estimé par notre méthode la concentration de la SEA, le gain et la position chromatographique du peptide NVTVQELDLQAR et le paramètre du bruit. Les autres paramètres de la méthode, les paramètres secondaires du modèle et les paramètres de l'algorithme ont été réglés sur une expérience témoin, puis ce jeu de paramètres a été utilisé pour traiter les 12 expériences.

Nous avons du utiliser des prétraitements pour gérer des phénomènes non prévus dans notre modèle. Nous avons tout d'abord rencontré un problème d'échantillonnage irrégulier. De plus, contrairement à toutes nos expériences précédentes effectuées dans de l'eau, nous avons été confrontés à la présence d'un bruit chimique. Nous avons proposé des prétraitements simples avant l'application de notre méthode d'estimation. Des travaux ultérieurs pourront prendre en compte directement ces phénomènes.

Le temps de calcul nécessaire pour l'estimation de nos concentrations est très raisonnable. En effet, il demande moins d'une minute de calcul sur un ordinateur moderne, alors que le temps de traitement d'un échantillon biologique dans le reste de la chaîne peut aller d'une heure à une journée. Bien sûr, si le problème se complexifie, par augmentation du nombre de paramètres à estimer par exemple, le temps de calcul pourra devenir un problème beaucoup plus important. Dans ce cas, des améliorations pourront être apportées à l'algorithme afin d'améliorer sa vitesse. Une piste importante pour l'amélioration de la vitesse de la méthode est l'utilisation d'une meilleure loi de proposition. En effet, on utilise pour l'instant la loi *a priori* pour l'échantillonnage des positions, ce qui produit un taux d'acceptation assez bas. En utilisant une loi de proposition plus proche de la loi cible, on pourra obtenir en moins d'itérations une bonne estimée de la moyenne. Pour fournir une loi proche de la loi cible, nous pourrons partir des calculs effectués à l'annexe 7.4 page 118 qui montrent que la log-vraisemblance des positions est une somme de gaussienne.

Nous avons constaté une variation des positions de l'ordre de la minute et une variation du gain de l'ordre de sa valeur moyenne. De plus, le RSB constaté est d'environ 2:1. Malgré ces fluctuations importantes, les données reconstruites à l'aide des paramètres estimés s'ajustent aux données réelles pour chaque expérience.

Contrairement aux méthodes de l'analyse factorielle qui nécessitent plusieurs expériences d'étalonnage pour apprendre la forme de la réponse de chaque protéine, notre méthode n'a besoin que d'une expérience pour estimer les concentrations car elle utilise un modèle direct paramétrique. Notons que ces expériences d'étalonnage constituent une source intéressante d'informations qui pourra être utilisée par la suite.

Les performances de notre méthode ont été comparées à d'autres méthodes d'estimation présentées dans l'état de l'art. Nous avons ainsi comparé notre méthode à celle présentée par Virginie Brun dans son article. Nous avons constaté une amélioration de la mesure à travers la droite de régression comparant les concentrations estimées par rapport aux concentrations de référence. En effet, le coefficient directeur et le coefficient de détermination R^2 de notre droite sont plus proches de l'unité. Nous avons également comparé notre méthode aux résultats obtenus par les méthodes d'analyse factorielle PLS et N-PLS. Si les droites de régression de ces deux méthodes se confondent à notre droite de régression, nous observons un léger avantage de notre méthode.

Enfin nous avons observé, sur données simulées, que la différence entre les méthodes s'accentue quand le rapport bruit sur signal augmente. C'est notre méthode qui se révèle être la plus robuste au bruit de mesure.

La méthode que nous avons développée se révèle donc robuste aux variations des paramètres instrument et diminue également les variations observées entre les expériences. L'amélioration des performances est cependant relativement faible dans la situation étudiée, mais laisse espérer de bonnes performances dans des cas plus complexes comme celui de la mesure de marqueurs cancéreux dans le sang.

6. Conclusions et perspectives

6.1. Conclusions

Ce document traite d'une méthode d'estimation des concentrations de protéines à partir des données issues d'une chaîne protéomique basée sur une colonne de chromatographie et un spectromètre de masse. L'étude de la littérature a montré que, dans des cas de faible rapport signal sur bruit, de meilleures performances pouvaient être attendues par une approche probabiliste. Nous avons développé dans cette thèse une méthode relevant des statistiques bayésiennes et reposant sur deux éléments : un modèle de la chaîne de mesure et une méthode d'inversion.

Pour modéliser notre instrument, nous avons donc proposé des équations décrivant chaque module de la chaîne d'analyse. Après quelques simplifications, le modèle final décrit chaque pic des données par une gaussienne bi-dimensionelle. L'étape d'inversion consiste à estimer les paramètres de ce modèle. En effet, en plus d'estimer les concentrations, la méthode proposée permet de prendre en compte les variations de certains paramètres systèmes. Nous avons choisi de les estimer au sens de la moyenne *a posteriori* et pour cela nous avons implémenté notre méthode à l'aide de méthodes MCMC et, plus particulièrement, une boucle de Gibbs incluant un échantillonneur de Metropolis-Hastings.

Nous avons validé la méthode sur divers types de données. Nous avons testé notre algorithme sur données simulées et sur données réelles en obtenant à chaque fois un bon ajustement du modèle aux données. En particulier, nous nous sommes intéressé à l'estimation de la concentration de peptides du cytochrome C dans de l'eau et à celle d'une toxine dans de l'urine. Dans le premier cas, les gains des peptides étaient constants, ce qui nous a permis de nous passer de protéines marquées. Dans le second cas, la variation des gains a nécessité la mise en place d'un étalonnage interne utilisant la méthode PSAQ.

Par rapport à la méthode traditionnelle, qui consiste à intégrer les pics sur un certain domaine, notre méthode améliore deux aspects. Tout d'abord nous gagnons en précision, car notre méthode a diminué la dispersion entre les différentes valeurs estimées. De plus, notre méthode est plus automatique car nous pouvons utiliser les mêmes paramètres pour une série d'expériences, sans avoir à préciser pour chaque expérience l'emplacement du pic. En particulier, notre méthode a un caractère auto-adaptatif, car il s'ajuste aux variations des temps de rétention et un caractère auto-calibrant, car il prend en compte l'étalonnage isotopique.

6.2. Perspectives

Ce travail constitue une première approche de l'estimation de concentrations, à partir de données LC-MS, dans un cadre bayésien. Plusieurs voies d'amélioration sont envisageables.

Tout d'abord, nous pouvons tester les limites de l'algorithme. Notre méthode permet-elle d'abaisser la limite de quantification des biomarqueurs ? Dans quelles conditions les hypothèses que nous avons formulées pour notre modèle ne sont plus valides ? Nous avons testé notre méthode sur un spectromètre de masse de type « trappe linaire » et sur un spectromètre de masse à temps de vol. Peut-

on utiliser la méthode sur d'autres spectromètres de masse et d'autres méthodes de marquage isotopique ? Dans le cas contraire, quelles modifications de notre méthode sont nécessaires ?

La réponse à ces questions nous amènera sûrement à perfectionner le modèle présenté. Premièrement, nous pourrons mieux modéliser la réponse de la colonne de chromatographie en prenant en compte la dissymétrie de certains pics. Cette amélioration pourra se faire soit en utilisant des modèles *ad hoc* comme l'exponentially modified gaussian (EMG) ou une résolution exacte des équations différentielles modélisant les phénomènes locaux. Deuxièmement, on pourra travailler sur la modélisation de la fonction de réponse spectrométrique. Par exemple, on peut envisager des Lorentziennes ou des fonctions de Voigt pour la modélisation des pics. Ensuite, nous avons centré notre modélisation sur nos protéines d'intérêt. Nous pouvons améliorer notre modèle en prenant en compte divers éléments supplémentaires. Le premier de ces éléments est le bruit chimique que nous avons rencontré dans nos expériences sur la SEA. Nous avons proposé pour traiter ce problème une méthode simple, mais il pourra également être intégré au sein d'une nouvelle méthode bayésienne. De même nous pourrons prendre en compte dans notre modèle les différents contaminants dont la proximité pourrait interférer avec la mesure de nos protéines. Plusieurs voies sont envisageables. Nous pouvons les traiter comme nos protéines d'intérêt, avec la particularité qu'ils ne disposent pas de protéine étalon. Cette démarche nécessitera de renseigner différents paramètres correspondant à ces pics parasites (proportions de charges, proportions de neutrons, …). Une autre solution serait de conserver notre méthode pour les protéines d'intérêt et d'essayer d'ajuster au mieux sur ces contaminants un ensemble de modèles de pics génériques. Troisièmement, nous pouvons nous intéresser à la modélisation du bruit. On pourra prendre en compte que les données ne peuvent pas être négatives ou que l'amplitude du bruit semble corrélée à l'amplitude du signal. Enfin, nous pouvons améliorer les informations *a priori* injectées. Nous pourrons mieux modéliser la gamme de concentration attendue, ou modéliser des informations plus complexes par des lois multimodales par exemple. Nous avons aussi envisagé de prendre en compte un ordre dans les différents temps de rétention, il faudra s'assurer de la validité de cette hypothèse.

D'autres améliorations concernent la méthodologie et la stratégie de mesure choisie. Nous n'avons pour l'instant utilisé qu'un peptide pour effectuer la quantification, or la méthode PSAQ et notre méthode de traitement permettent d'utiliser tous les peptides de la protéine. Notons que des aménagements de notre implémentation seront peut-être nécessaires, pour pouvoir en pratique utiliser l'intégralité de l'information à notre disposition. Par exemple, notre algorithme devra pouvoir traiter simultanément plusieurs zones extraites du spectrogramme. De même, nous pouvons envisager de traiter en une fois plusieurs expériences. Ces expériences pourront traiter des fractions différentes du même échantillon d'intérêt pour améliorer l'estimation des concentrations. Pour cela, les liens liant différentes expériences devront être mieux définis. De plus, nous pouvons noter que la quantification n'est habituellement pas le but final du traitement. En effet, ces concentrations seront utilisées pour découvrir de nouveaux biomarqueurs ou pour réaliser un diagnostic. On peut donc envisager de créer des méthodes bayésiennes globales dédiées à ces deux tâches. Ces méthodes pourront être bâties à partir de notre méthode de quantification actuelle. Enfin notre méthode estime les valeurs des concentrations au sens de la moyenne. Les méthodes bayésiennes permettent également de fournir d'autres renseignements sur notre mesure. Nous pouvons ainsi obtenir les marges d'erreur de notre mesure et la probabilité associée à ces marges. De même, nous pouvons envisager de fournir conjointement à l'estimée au sens de la moyenne, l'estimée au sens du maximum. Un écart important entre leurs deux valeurs permettrait de détecter un problème d'indétermination entre plusieurs valeurs possibles. Pour fournir rapidement une valeur précise de l'estimateur de la moyenne et des marges d'erreurs nous devrons peut-être étudier des échantillonneurs plus performants. Nous pourrions ainsi nous intéresser encore plus à l'échantillonneur des positions. Nous pourrons entre autre proposer des lois de proposition plus proches de la loi cible ou essayer un échantillonneur de Metropolis-Hastings à marche aléatoire.

7. Annexe : Calculs

Dans cette thèse, nous utilisons de nombreuses lois de probabilité. Certaines sont définies, d'autres se déduisent par calcul, notamment les fonctions de vraisemblance et les conditionnelles *a posteriori*. Un peu de remise en forme des expressions est souvent nécessaire pour savoir quel type de loi se cache derrière ces expressions.

Une fois que le type de lois a été déterminé, nous pouvons remarquer que les échantillonneurs qui simulent ces lois nécessitent le calcul de paramètres qui font intervenir des matrices de grandes dimensions. Il est alors avantageux d'utiliser la structure particulière de ces matrices pour accélérer les calculs. Mais avant, commençons par calculer le produit de deux fonctions gaussiennes qui nous servira par la suite.

7.1. Produit de deux fonctions gaussiennes

Le produit de deux fonctions gaussiennes est une gaussienne. Il s'agit d'une phrase bien connue, mais nous pouvons nous poser la question suivante, quels sont les paramètres de cette gaussienne ? Considérons les deux fonctions gaussiennes suivantes

$$\begin{cases} f_1(t) = (2\pi)^{-1/2} \gamma_1^{1/2} \exp\left(-\frac{1}{2} \gamma_1 (t - t_1)^2 \right) \\ f_2(t) = (2\pi)^{-1/2} \gamma_2^{1/2} \exp\left(-\frac{1}{2} \gamma_2 (t - t_2)^2 \right) \end{cases}$$

Calculons le produit de ces deux fonctions

$$f_1(t) f_2(t) = (2\pi)^{-1} (\gamma_1 \gamma_2)^{1/2} \exp\left\{ -\frac{1}{2} \left(\gamma_1 (t - t_1)^2 + \gamma_2 (t - t_2)^2 \right) \right\}$$

On reconnait la forme d'une gaussienne, l'essentiel du travail se concentre sur l'argument de l'exponentielle. Soit $Q = \gamma_1 (t - t_1)^2 + \gamma_2 (t - t_2)^2$.

Ecrivons les coefficients du polynôme Q d'ordre 2, et mettons-le sous forme canonique.

$$\begin{aligned} Q &= \gamma_1 \left(t^2 - 2 t t_1 + t_1^2 \right) + \gamma_2 \left(t^2 - 2 t t_2 + t_2^2 \right) \\ &= (\gamma_1 + \gamma_2) t^2 - 2t (\gamma_1 t_1 + \gamma_2 t_2) + \gamma_1 t_1^2 + \gamma_2 t_2^2 \\ &= (\gamma_1 + \gamma_2) \left(t^2 - 2t \frac{\gamma_1 t_1 + \gamma_2 t_2}{\gamma_1 + \gamma_2} \right) + \gamma_1 t_1^2 + \gamma_2 t_2^2 \end{aligned}$$

Posons les variables suivantes

$$\bar{\gamma} = \gamma_1 + \gamma_2 \text{ et } \bar{t} = \frac{\gamma_1 t_1 + \gamma_2 t_2}{\gamma_1 + \gamma_2}$$

Nous obtenons la forme canonique suivante

$$\begin{aligned} Q &= \bar{\gamma}\left(t^2 - 2t\bar{t}\right) + \gamma_1 t_1^2 + \gamma_2 t_2^2 \\ &= \bar{\gamma}\left((t-\bar{t})^2 - (\bar{t})^2\right) + \gamma_1 t_1^2 + \gamma_2 t_2^2 \\ &= \bar{\gamma}(t-\bar{t})^2 - \bar{\gamma}(\bar{t})^2 + \gamma_1 t_1^2 + \gamma_2 t_2^2 \end{aligned}$$

Le gaussienne que nous recherchons est centrée $\bar{t}$ en et de largeur $\bar{\gamma}$

$$f_1(t) f_2(t) \propto \exp\left\{-\frac{1}{2}\bar{\gamma}(t-\bar{t})^2\right\}$$

Il nous reste à déterminer le coefficient multiplicateur du produit de ces deux gaussiennes. Pour cela, simplifions le terme constant *C* du polynôme *Q*.

$$\begin{aligned} C &= \gamma_1 t_1^2 + \gamma_2 t_2^2 - \bar{\gamma}(\bar{t})^2 \\ &= \gamma_1 t_1^2 + \gamma_2 t_2^2 - \bar{\gamma}\frac{(\gamma_1 t_1 + \gamma_2 t_2)^2}{(\bar{\gamma})^2} \\ &= (\bar{\gamma})^{-1}\left\{\bar{\gamma}\left(\gamma_1 t_1^2 + \gamma_2 t_2^2\right) - (\gamma_1 t_1 + \gamma_2 t_2)^2\right\} \\ &= (\bar{\gamma})^{-1}\left\{(\gamma_1 + \gamma_2)\left(\gamma_1 t_1^2 + \gamma_2 t_2^2\right) - (\gamma_1 t_1 + \gamma_2 t_2)^2\right\} \\ &= (\bar{\gamma})^{-1}\left\{\gamma_1^2 t_1^2 + \gamma_1\gamma_2 t_2^2 + \gamma_1\gamma_2 t_1^2 + \gamma_2^2 t_2^2 - \left(\gamma_1^2 t_1^2 + 2\gamma_1\gamma_2 t_1 t_2 + \gamma_2^2 t_2^2\right)\right\} \end{aligned}$$

Plusieurs termes se simplifient dans cette expression, nous obtenons donc

$$\begin{aligned} C &= \gamma_1\gamma_2(\bar{\gamma})^{-1}\left(t_1^2 + t_2^2 - 2t_1 t_2\right) \\ &= \frac{\gamma_1\gamma_2}{\gamma_1 + \gamma_2}(t_1 - t_2)^2 \end{aligned}$$

Revenons au produit des deux gaussiennes, les calculs précédents permettent d'obtenir l'expression

$$\begin{aligned} f_1(t) f_2(t) &= (2\pi)^{-1}(\gamma_1\gamma_2)^{1/2}\exp\left\{-\frac{1}{2}Q\right\} \\ &= (2\pi)^{-1}(\gamma_1\gamma_2)^{1/2}\exp\left\{-\frac{1}{2}\left(\frac{\gamma_1\gamma_2}{\gamma_1 + \gamma_2}(t_1 - t_2)^2 + \bar{\gamma}(t-\bar{t})^2\right)\right\} \\ &= (2\pi)^{-1}(\gamma_1\gamma_2)^{1/2}\exp\left\{-\frac{1}{2}\frac{\gamma_1\gamma_2}{\gamma_1 + \gamma_2}(t_1 - t_2)^2\right\}\exp\left\{-\frac{1}{2}\bar{\gamma}(t-\bar{t})^2\right\} \end{aligned}$$

Faisons intervenir le coefficient de normalisation de notre gaussienne, nous pouvons donc écrite la formule précédente sous la forme suivante

$$\begin{aligned} f_1(t) f_2(t) &= (2\pi)^{-1}(\gamma_1\gamma_2)^{1/2}\exp\left\{-\frac{1}{2}\frac{\gamma_1\gamma_2}{\gamma_1 + \gamma_2}(t_1 - t_2)^2\right\} \\ &\quad \times (2\pi)^{1/2}(\bar{\gamma})^{-1/2}(2\pi)^{-1/2}(\bar{\gamma})^{1/2}\exp\left\{-\frac{1}{2}\bar{\gamma}(t-\bar{t})^2\right\} \end{aligned}$$

Après quelques simplifications, nous obtenons

$$\begin{aligned} f_1(t) f_2(t) &= (2\pi)^{-1/2}\left(\frac{\gamma_1\gamma_2}{\gamma_1 + \gamma_2}\right)^{1/2}\exp\left\{-\frac{1}{2}\frac{\gamma_1\gamma_2}{\gamma_1 + \gamma_2}(t_1 - t_2)^2\right\} \\ &\quad \times (2\pi)^{-1/2}(\bar{\gamma})^{1/2}\exp\left\{-\frac{1}{2}\bar{\gamma}(t-\bar{t})^2\right\} \end{aligned}$$

Nous faisons apparaitre une seconde gaussienne. Au final, le produit de deux gaussiennes normée est de norme $\bar{\alpha}$ et c'est une gaussienne centrée en $\bar{t}$, de largeur $\bar{\gamma}$

$$f_1(t)f_2(t) = \overline{\alpha}.N(t\,;t,\overline{\gamma})$$

Avec

$$\overline{\alpha} = N\left(t_1\,;t_2,\frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2}\right), \overline{\gamma} = \gamma_1+\gamma_2 \text{ et } \overline{t} = \frac{\gamma_1 t_1+\gamma_2 t_2}{\gamma_1+\gamma_2}$$

7.2. Calcul rapide des matrices G^tG, G^ty, H^tH, H^ty, H^tH^*, $H^{*t}H^*$ et $H^{*t}y$

Nous allons dans cette section décrire une méthode de calcul rapide des matrices $\boldsymbol{G^tG}$, $\boldsymbol{G^ty}$, $\boldsymbol{H^tH}$, $\boldsymbol{H^ty}$, $\boldsymbol{H^{*t}H^*}$, $\boldsymbol{H^tH^*}$. Pour cela, nous utiliserons l'équivalence entre les normes matricielles et vectorielles que nous avons déjà mis à jour au chapitre 4.

$$\chi(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t}) = \left\|\boldsymbol{Y}-\boldsymbol{\Theta}(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t})\right\|^2 = \left\|\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x}-\boldsymbol{H}^*\boldsymbol{x}^*\right\|^2$$

Explicitons chacun de ces termes. Commençons par la norme vectorielle qui fait apparaître les matrices $\boldsymbol{H^tH}$, $\boldsymbol{H^ty}$, $\boldsymbol{H^{*t}H^*}$, $\boldsymbol{H^tH^*}$.

$$\chi(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t}) = \left(\boldsymbol{y}^t\boldsymbol{y}-2\boldsymbol{x}^{*t}\boldsymbol{H}^{*t}\boldsymbol{y}+\boldsymbol{x}^{*t}\boldsymbol{H}^{*t}\boldsymbol{H}^*\boldsymbol{x}^*\right)-2\boldsymbol{x}^t\left(\boldsymbol{H}^t\boldsymbol{y}-\boldsymbol{H}^t\boldsymbol{H}^*\boldsymbol{x}^*\right)+\boldsymbol{x}^t\boldsymbol{H}^t\boldsymbol{H}\boldsymbol{x}$$

Montrons que l'expression utilisant la norme matricielle est égale à

$$\chi(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t}) = \boldsymbol{y}^t\boldsymbol{y}-2\sum_p\left(x_pb_p+x_p^*b_p^*\right)+\sum_p\sum_q x_px_qa_{pq}+2\sum_p\sum_q x_px_q^*a_{pq}^*+\sum_p\sum_q x_p^*x_q^*a_{pq}^{**}$$

Avec

$$a_{pq} = \sum_{ijk}\sum_{uvw} d_{ip}d_{uq}\xi_i\xi_u\pi_{ij}\pi_{uv}\pi'_{ijk}\pi'_{uvw}\boldsymbol{s}_{ijk}^t\boldsymbol{s}_{uvw}\boldsymbol{c}_i^t\boldsymbol{c}_u$$

$$a_{pq}^* = \sum_{ijk}\sum_{uvw} d_{ip}d_{uq}\xi_i\xi_u\pi_{ij}\pi_{uv}\pi'_{ijk}\pi'^*_{uvw}\boldsymbol{s}_{ijk}^t\boldsymbol{s}_{uvw}^*\boldsymbol{c}_i^t\boldsymbol{c}_u$$

$$a_{pq}^{**} = \sum_{ijk}\sum_{uvw} d_{ip}d_{uq}\xi_i\xi_u\pi_{ij}\pi_{uv}\pi'^*_{ijk}\pi'^*_{uvw}\boldsymbol{s}_{ijk}^{*t}\boldsymbol{s}_{uvw}^*\boldsymbol{c}_i^t\boldsymbol{c}_u$$

$$b_p = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=0}^{K} d_{ip}\xi_i\pi_{ij}\pi'_{ijk}\boldsymbol{s}_{ijk}^t\boldsymbol{Y}\boldsymbol{c}_i$$

$$b_p^* = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=0}^{K} d_{ip}\xi_i\pi_{ij}\pi'^*_{ijk}\boldsymbol{s}_{ijk}^{*t}\,\boldsymbol{Y}\boldsymbol{c}$$

Pour cela, nous allons développer l'expression de la norme matricielle

7.2.1. Développement de la norme matricielle

Nous pouvons développer l'expression de la norme matricielle en utilisant l'opérateur trace

$$\begin{aligned}\chi(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t}) &= \left\|\boldsymbol{Y}-\boldsymbol{\Theta}(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t})\right\|^2\\ &= \mathrm{Tr}\left\{\left(\boldsymbol{Y}-\boldsymbol{\Theta}(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t})\right)^t\left(\boldsymbol{Y}-\boldsymbol{\Theta}(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t})\right)\right\}\\ &= \mathrm{Tr}\left\{\left(\boldsymbol{Y}^t-\boldsymbol{\Theta}(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t})^t\right)\left(\boldsymbol{Y}-\boldsymbol{\Theta}(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t})\right)\right\}\\ &= \mathrm{Tr}\left\{\boldsymbol{Y}^t\boldsymbol{Y}-\boldsymbol{Y}^t\boldsymbol{\Theta}(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t})-\boldsymbol{\Theta}(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t})^t\boldsymbol{Y}+\boldsymbol{\Theta}(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t})^t\boldsymbol{\Theta}(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t})\right\}\end{aligned}$$

Cette expression peut encore se simplifier. L'opérateur trace est linéaire, donc on en déduit

$$\begin{aligned}&\chi(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t})\\ &= \mathrm{Tr}\left\{\boldsymbol{Y}^t\boldsymbol{Y}\right\}-\mathrm{Tr}\left\{\boldsymbol{Y}^t\boldsymbol{\Theta}(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t})\right\}-\mathrm{Tr}\left\{\boldsymbol{\Theta}(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t})^t\boldsymbol{Y}\right\}+\mathrm{Tr}\left\{\boldsymbol{\Theta}(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t})^t\boldsymbol{\Theta}(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t})^t\right\}\end{aligned}$$

La trace d'une matrice et de sa transposée sont égales, donc on a

$$\begin{aligned}\mathrm{Tr}\left\{\boldsymbol{Y}^t\Theta\left(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t}\right)\right\} &= \mathrm{Tr}\left\{\left(\boldsymbol{Y}^t\Theta\left(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t}\right)\right)^t\right\}\\ &= \mathrm{Tr}\left\{\Theta\left(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t}\right)^t\boldsymbol{Y}\right\}\end{aligned}$$

D'où

$$\chi\left(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t}\right) = \underbrace{\mathrm{Tr}\left\{\boldsymbol{Y}^t\boldsymbol{Y}\right\}}_{A} - 2\underbrace{\mathrm{Tr}\left\{\Theta\left(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t}\right)^t\boldsymbol{Y}\right\}}_{B} + \underbrace{\mathrm{Tr}\left\{\Theta\left(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t}\right)^t\Theta\left(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t}\right)^t\right\}}_{C}$$

Nous allons détailler dans les sections suivantes l'expression de ces trois termes.

7.2.1.1. Expression du terme A

Le terme A peut s'écrire sous la forme d'un produit scalaire vectoriel

$$A = \mathrm{Tr}\left\{\boldsymbol{Y}^t\boldsymbol{Y}\right\} = \left\|\boldsymbol{Y}\right\|^2 = \left\|\boldsymbol{y}\right\|^2 = \boldsymbol{y}^t\boldsymbol{y}$$

Ce terme est le plus simple à calculer, les suivants demandent plus de calculs

7.2.1.2. Expression du terme B

Le second terme peut encore se développer en utilisant la structure du modèle

$$\Theta\left(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t}\right) = \sum_{p=1}^{P}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=0}^{K} d_{ip}\xi_i\pi_{ij}\left(x_p\pi'_{ijk}\boldsymbol{s}_{ijk} + x_p^*\pi'^*_{ijk}\boldsymbol{s}^*_{ijk}\right)\boldsymbol{c}_i^t$$

Nous obtenons donc pour le terme B

$$\begin{aligned}B &= \mathrm{Tr}\left\{\Theta\left(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t}\right)^t\boldsymbol{Y}\right\}\\ &= \mathrm{Tr}\left\{\left(\sum_{p=1}^{P}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=0}^{K}\underbrace{x_p d_{ip}\xi_i\pi_{ij}\pi'_{ijk}}_{e_{pijk}}\boldsymbol{s}_{ijk}\boldsymbol{c}_i^t + \underbrace{x_p^* d_{ip}\xi_i\pi_{ij}\pi'^*_{ijk}}_{e^*_{pijk}}\boldsymbol{s}^*_{ijk}\,\boldsymbol{c}_i^t\right)^t\boldsymbol{Y}\right\}\\ &= \mathrm{Tr}\left\{\sum_{p=1}^{P}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=0}^{K} e_{pijk}\boldsymbol{c}_i\boldsymbol{s}_{ijk}^t\boldsymbol{Y} + e^*_{pijk}\boldsymbol{c}_i\boldsymbol{s}^{*t}_{ijk}\boldsymbol{Y}\right\}\end{aligned}$$

La trace est linéaire, nous obtenons donc

$$B = \sum_{p=1}^{P}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=0}^{K} e_{pijk}\,\mathrm{Tr}\left\{\boldsymbol{c}_i\left(\boldsymbol{s}_{ijk}^t\boldsymbol{Y}\right)\right\} + e^*_{pijk}\,\mathrm{Tr}\left\{\boldsymbol{c}_i\left(\boldsymbol{s}^{*t}_{ijk}\boldsymbol{Y}\right)\right\}$$

D'autre part, on a Tr(***AB***) = Tr(***BA***), pour toutes les matrices où ces produits matriciels ont un sens. Nous avons donc

$$\begin{aligned}B &= \sum_{p=1}^{P}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=0}^{K} e_{pijk}\,\mathrm{Tr}\left\{\boldsymbol{c}_i\left(\boldsymbol{s}_{ijk}^t\boldsymbol{Y}\right)\right\} + e^*_{pijk}\,\mathrm{Tr}\left\{\boldsymbol{c}_i\left(\boldsymbol{s}^{*t}_{ijk}\boldsymbol{Y}\right)\right\}\\ &= \sum_{p=1}^{P}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=0}^{K} e_{pijk}\,\mathrm{Tr}\left\{\left(\boldsymbol{s}_{ijk}^t\boldsymbol{Y}\right)\boldsymbol{c}_i\right\} + e^*_{pijk}\,\mathrm{Tr}\left\{\left(\boldsymbol{s}^{*t}_{ijk}\boldsymbol{Y}\right)\boldsymbol{c}_i\right\}\end{aligned}$$

La trace d'un scalaire est égale à lui-même. L'expression du deuxième terme devient donc

$$\begin{aligned}B &= \sum_{p=1}^{P}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=0}^{K} e_{pijk}\boldsymbol{s}_{ijk}^t\boldsymbol{Y}\boldsymbol{c}_i + e^*_{pijk}\boldsymbol{s}^{*t}_{ijk}\boldsymbol{Y}\boldsymbol{c}_i\\ &= \sum_{p=1}^{P} x_p b_p + x_p^* b_p^*\end{aligned}$$

Avec

$$b_p = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=0}^{K} d_{ip}\xi_i\pi_{ij}\pi'_{ijk}\boldsymbol{s}_{ijk}^{t}\boldsymbol{Y}\boldsymbol{c}_i$$

$$b_p^* = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=0}^{K} d_{ip}\xi_i\pi_{ij}\pi'^{*}_{ijk}\boldsymbol{s}_{ijk}^{*t}\boldsymbol{Y}\boldsymbol{c}$$

Il nous reste à expliciter le 3e terme.

7.2.1.3. Expression du terme C

Avec les même conventions qu'à la section précédente ; nous obtenons

$$\begin{aligned}
C &= \mathrm{Tr}\left\{\Theta(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t})^t\,\Theta(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t})\right\}\\
&= \mathrm{Tr}\left\{\left(\sum_{pijk} e_{pijk}\boldsymbol{s}_{ijk}\boldsymbol{c}_i^t + e_{pijk}^*\boldsymbol{s}_{ijk}^*\boldsymbol{c}_i^t\right)^t\left(\sum_{quvw} e_{quvw}\boldsymbol{s}_{uvw}\boldsymbol{c}_u^t + e_{quvw}^*\boldsymbol{s}_{uvw}^*\boldsymbol{c}_u^t\right)\right\}\\
&= \mathrm{Tr}\left\{\left(\sum_{pijk} e_{pijk}\boldsymbol{c}_i\boldsymbol{s}_{ijk}^t + e_{pijk}^*\boldsymbol{c}_i\boldsymbol{s}_{ijk}^{*t}\right)\left(\sum_{quvw} e_{quvw}\boldsymbol{s}_{uvw}\boldsymbol{c}_u^t + e_{quvw}^*\boldsymbol{s}_{uvw}^*\boldsymbol{c}_u^t\right)\right\}\\
&= \mathrm{Tr}\left\{\sum_{pijk}\sum_{quvw} e_{pijk}e_{quvw}\boldsymbol{c}_i\boldsymbol{s}_{ijk}^t\boldsymbol{s}_{uvw}\boldsymbol{c}_u^t + e_{pijk}e_{quvw}^*\boldsymbol{c}_i\boldsymbol{s}_{ijk}^t\boldsymbol{s}_{uvw}^*\boldsymbol{c}_u^t + e_{pijk}^*e_{quvw}\boldsymbol{c}_i\boldsymbol{s}_{ijk}^{*t}\boldsymbol{s}_{uvw}\boldsymbol{c}_u^t + e_{pijk}^*e_{quvw}^*\boldsymbol{c}_i\boldsymbol{s}_{ijk}^{*t}\boldsymbol{s}_{uvw}^*\boldsymbol{c}_u^t\right\}
\end{aligned}$$

Cette expression fait apparaître plusieurs produits scalaires vectoriels avec les vecteurs spectrométriques. Ces derniers étant des scalaires ils peuvent se factoriser. De plus l'opérateur trace étant linéaire, nous obtenons

$$C = \sum_{pijk}\sum_{quvw}\left(e_{pijk}e_{quvw}\boldsymbol{s}_{ijk}^t\boldsymbol{s}_{uvw} + e_{pijk}e_{quvw}^*\boldsymbol{s}_{ijk}^t\boldsymbol{s}_{uvw}^* + e_{pijk}^*e_{quvw}\boldsymbol{s}_{ijk}^{*t}\boldsymbol{s}_{uvw} + e_{pijk}^*e_{quvw}^*\boldsymbol{s}_{ijk}^{*t}\boldsymbol{s}_{uvw}^*\right)\mathrm{Tr}\left\{\boldsymbol{c}_i\boldsymbol{c}_u^t\right\}$$

Intéressons nous à l'argument de l'opérateur trace

$$\mathrm{Tr}\left\{\boldsymbol{c}_i\boldsymbol{c}_u^t\right\} = \mathrm{Tr}\left\{(\boldsymbol{c}_i)(\boldsymbol{c}_u^t)\right\} = \mathrm{Tr}\left\{(\boldsymbol{c}_u^t)(\boldsymbol{c}_i)\right\} = \boldsymbol{c}_u^t\boldsymbol{c}_i = \boldsymbol{c}_i^t\boldsymbol{c}_u$$

Le terme C peut donc s'écrire

$$\begin{aligned}
C = &\sum_{pijk}\sum_{quvw} e_{pijk}e_{quvw}\boldsymbol{s}_{ijk}^t\boldsymbol{s}_{uvw}\boldsymbol{c}_i^t\boldsymbol{c}_u + \sum_{pijk}\sum_{quvw} e_{pijk}e_{quvw}^*\boldsymbol{s}_{ijk}^t\boldsymbol{s}_{uvw}^*\boldsymbol{c}_i^t\boldsymbol{c}_u\\
&+ \sum_{pijk}\sum_{quvw} e_{pijk}^*e_{quvw}\boldsymbol{s}_{ijk}^{*t}\boldsymbol{s}_{uvw}\boldsymbol{c}_i^t\boldsymbol{c}_u + \sum_{pijk}\sum_{quvw} e_{pijk}^*e_{quvw}^*\boldsymbol{s}_{ijk}^{*t}\boldsymbol{s}_{uvw}^*\boldsymbol{c}_i^t\boldsymbol{c}_u
\end{aligned}$$

Intéressons nous au 3e terme de cette expression et montrons qu'il est égal au 2e terme. Pour cela il suffit de permuter les indices (p,i,j,k) et (q,u,v,w)

$$\begin{aligned}
\sum_{pijk}\sum_{quvw} e_{pijk}^*e_{quvw}\boldsymbol{s}_{ijk}^{*t}\boldsymbol{s}_{uvw}\boldsymbol{c}_i^t\boldsymbol{c}_u &= \sum_{quvw}\sum_{pijk} e_{quvw}^*e_{pijk}\boldsymbol{s}_{uvw}^{*t}\boldsymbol{s}_{ijk}\boldsymbol{c}_u^t\boldsymbol{c}_i\\
&= \sum_{pijk}\sum_{quvw} e_{pijk}e_{quvw}^*\boldsymbol{s}_{ijk}^t\boldsymbol{s}_{uvw}^*\boldsymbol{c}_i^t\boldsymbol{c}_u
\end{aligned}$$

L'expression du terme C se simplifie donc

$$\begin{aligned}
C &= \sum_{pijk}\sum_{quvw} e_{pijk}e_{quvw}\boldsymbol{s}_{ijk}^t\boldsymbol{s}_{uvw}\boldsymbol{c}_i^t\boldsymbol{c}_u + 2\sum_{pijk}\sum_{quvw} e_{pijk}e_{quvw}^*\boldsymbol{s}_{ijk}^t\boldsymbol{s}_{uvw}^*\boldsymbol{c}_i^t\boldsymbol{c}_u + \sum_{pijk}\sum_{quvw} e_{pijk}^*e_{quvw}^*\boldsymbol{s}_{ijk}^{*t}\boldsymbol{s}_{uvw}^*\boldsymbol{c}_i^t\boldsymbol{c}_u\\
&= \sum_p\sum_q x_p x_q a_{pq} + 2\sum_p\sum_q x_p x_q^* a_{pq}^* + \sum_p\sum_q x_p^* x_q^* a_{pq}^{**}
\end{aligned}$$

Avec

$$a_{pq} = \sum_{ijk}\sum_{uvw} d_{ip}d_{uq}\xi_i\xi_u\pi_{ij}\pi_{uv}\pi'_{ijk}\pi'_{uvw}\mathbf{s}^t_{ijk}\mathbf{s}_{uvw}\mathbf{c}^t_i\mathbf{c}_u$$
$$a^*_{pq} = \sum_{ijk}\sum_{uvw} d_{ip}d_{uq}\xi_i\xi_u\pi_{ij}\pi_{uv}\pi'_{ijk}\pi'^*_{uvw}\mathbf{s}^t_{ijk}\mathbf{s}^*_{uvw}\mathbf{c}^t_i\mathbf{c}_u$$
$$a^{**}_{pq} = \sum_{ijk}\sum_{uvw} d_{ip}d_{uq}\xi_i\xi_u\pi_{ij}\pi_{uv}\pi'^*_{ijk}\pi'^*_{uvw}\mathbf{s}^{*t}_{ijk}\mathbf{s}^*_{uvw}\mathbf{c}^t_i\mathbf{c}_u$$

7.2.1.4. Conclusion

L'expression de la norme matricielle est donc

$$\chi(\mathbf{x},\mathbf{x}^*,\boldsymbol{\xi},\mathbf{t}) = A + B + C$$
$$= \mathbf{y}^t\mathbf{y} - 2\sum_p\left(x_p b_p + x^*_p b^*_p\right) + \sum_p\sum_q x_p x_q a_{pq} + 2\sum_p\sum_q x_p x^*_q a^*_{pq} + \sum_p\sum_q x^*_p x^*_q a^{**}_{pq}$$

Avec

$$a_{pq} = \sum_{ijk}\sum_{uvw} d_{ip}d_{uq}\xi_i\xi_u\pi_{ij}\pi_{uv}\pi'_{ijk}\pi'_{uvw}\mathbf{s}^t_{ijk}\mathbf{s}_{uvw}\mathbf{c}^t_i\mathbf{c}_u$$
$$a^*_{pq} = \sum_{ijk}\sum_{uvw} d_{ip}d_{uq}\xi_i\xi_u\pi_{ij}\pi_{uv}\pi'_{ijk}\pi'^*_{uvw}\mathbf{s}^t_{ijk}\mathbf{s}^*_{uvw}\mathbf{c}^t_i\mathbf{c}_u$$
$$a^{**}_{pq} = \sum_{ijk}\sum_{uvw} d_{ip}d_{uq}\xi_i\xi_u\pi_{ij}\pi_{uv}\pi'^*_{ijk}\pi'^*_{uvw}\mathbf{s}^{*t}_{ijk}\mathbf{s}^*_{uvw}\mathbf{c}^t_i\mathbf{c}_u$$
$$b_p = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=0}^{K} d_{ip}\xi_i\pi_{ij}\pi'_{ijk}\mathbf{s}^t_{ijk}\mathbf{Y}\mathbf{c}_i$$
$$b^*_p = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=0}^{K} d_{ip}\xi_i\pi_{ij}\pi'^*_{ijk}\mathbf{s}^{*t}_{ijk}\,\mathbf{Y}\mathbf{c}$$

7.2.2. Expression des matrices H^tH, H^ty, H^tH^*, $H^{*t}H^*$ et $H^{*t}y$

A la section précédente, nous avons démontré

$$\chi(\mathbf{x},\mathbf{x}^*,\boldsymbol{\xi},\mathbf{t}) = \left(\mathbf{y}^t\mathbf{y} - 2\mathbf{x}^{*t}\mathbf{H}^{*t}\mathbf{y} + \mathbf{x}^{*t}\mathbf{H}^{*t}\mathbf{H}^*\mathbf{x}^*\right) - 2\mathbf{x}^t\left(\mathbf{H}^t\mathbf{y} - \mathbf{H}^t\mathbf{H}^*\mathbf{x}^*\right) + \mathbf{x}^t\mathbf{H}^t\mathbf{H}\mathbf{x}$$
$$= \mathbf{y}^t\mathbf{y} - 2\sum_p\left(x_p b_p + x^*_p b^*_p\right) + \sum_p\sum_q x_p x_q a_{pq} + 2\sum_p\sum_q x_p x^*_q a^*_{pq} + \sum_p\sum_q x^*_p x^*_q a^{**}_{pq}$$

Nous en déduisons par identification polynomiale, les expressions des matrices d'intérêt

$$\left(\mathbf{H}^t\mathbf{y}\right)_p = b_p = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=0}^{K} d_{ip}\xi_i\pi_{ij}\pi'_{ijk}\mathbf{s}^t_{ijk}\mathbf{Y}\mathbf{c}_i$$
$$\left(\mathbf{H}^{*t}\mathbf{y}\right)_p = b^*_p = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=0}^{K} d_{ip}\xi_i\pi_{ij}\pi'^*_{ijk}\mathbf{s}^{*t}_{ijk}\,\mathbf{Y}\mathbf{c}$$
$$\left(\mathbf{H}^t\mathbf{H}\right)_{p,q} = a_{pq} = \sum_{ijk}\sum_{uvw} d_{ip}d_{uq}\xi_i\xi_u\pi_{ij}\pi_{uv}\pi'_{ijk}\pi'_{uvw}\mathbf{s}^t_{ijk}\mathbf{s}_{uvw}\mathbf{c}^t_i\mathbf{c}_u$$
$$\left(\mathbf{H}^t\mathbf{H}^*\right)_{p,q} = a^*_{pq} = \sum_{ijk}\sum_{uvw} d_{ip}d_{uq}\xi_i\xi_u\pi_{ij}\pi_{uv}\pi'_{ijk}\pi'^*_{uvw}\mathbf{s}^t_{ijk}\mathbf{s}^*_{uvw}\mathbf{c}^t_i\mathbf{c}_u$$
$$\left(\mathbf{H}^{*t}\mathbf{H}^*\right)_{p,q} = a^{**}_{pq} = \sum_{ijk}\sum_{uvw} d_{ip}d_{uq}\xi_i\xi_u\pi_{ij}\pi_{uv}\pi'^*_{ijk}\pi'^*_{uvw}\mathbf{s}^{*t}_{ijk}\mathbf{s}^*_{uvw}\mathbf{c}^t_i\mathbf{c}_u$$

7.2.3. Expression des matrices F^tF, F^ty, F^tF^*, $F^{*t}F^*$ et $F^{*t}y$

Montrons que

$$\left(\boldsymbol{F}^t\boldsymbol{y}\right)_i = \sum_{j=1}^{J}\sum_{k=0}^{K}\pi_{ij}\pi'_{ijk}\boldsymbol{s}_{ijk}^t\boldsymbol{Y}\boldsymbol{c}_i$$

$$\left(\boldsymbol{F}^{*t}\boldsymbol{y}\right)_i = \sum_{j=1}^{J}\sum_{k=0}^{K}\pi_{ij}\pi'^{*}_{ijk}\boldsymbol{s}_{ijk}^{*t}\boldsymbol{Y}\boldsymbol{c}_i$$

$$\left(\boldsymbol{F}^t\boldsymbol{F}\right)_{i,u} = \sum_{jk}\sum_{vw}\pi_{ij}\pi_{uv}\pi'_{ijk}\pi'_{uvw}\boldsymbol{s}_{ijk}^t\boldsymbol{s}_{uvw}\boldsymbol{c}_i^t\boldsymbol{c}_u$$

$$\left(\boldsymbol{F}^t\boldsymbol{F}^*\right)_{i,u} = \sum_{jk}\sum_{vw}\pi_{ij}\pi_{uv}\pi'_{ijk}\pi'^{*}_{uvw}\boldsymbol{s}_{ijk}^t\boldsymbol{s}_{uvw}^{*}\boldsymbol{c}_i^t\boldsymbol{c}_u$$

$$\left(\boldsymbol{F}^{*t}\boldsymbol{F}^*\right)_{i,u} = \sum_{jk}\sum_{vw}\pi_{ij}\pi_{uv}\pi'^{*}_{IJK}\pi'^{*}_{uvw}\boldsymbol{s}_{ijk}^{*t}\boldsymbol{s}_{uvw}^{*}\boldsymbol{c}_i^t\boldsymbol{c}_u$$

Nous allons détailler la démonstration pour les matrices $\boldsymbol{F}^t\boldsymbol{F}$ et $\boldsymbol{F}^t\boldsymbol{y}$. Au chapitre 4, nous avons montré le lien entre la matrice $\boldsymbol{F}$ et $\boldsymbol{H}$

$$\boldsymbol{H} = \boldsymbol{F}\,\mathrm{diag}(\boldsymbol{\xi})\boldsymbol{D}$$

D'où

$$\boldsymbol{H}^t\boldsymbol{y} = \boldsymbol{D}^t\,\mathrm{diag}(\boldsymbol{\xi})\boldsymbol{F}^t\boldsymbol{y}$$
$$\boldsymbol{H}^t\boldsymbol{H} = \boldsymbol{D}^t\,\mathrm{diag}(\boldsymbol{\xi})\boldsymbol{F}^t\boldsymbol{F}\,\mathrm{diag}(\boldsymbol{\xi})\boldsymbol{D}$$

Et

$$\left(\boldsymbol{H}^t\boldsymbol{y}\right)_p = \sum_{i=1}^{I}\left(\boldsymbol{D}^t\right)_{p,i}\left(\mathrm{diag}(\boldsymbol{\xi})\right)_{i,i}\left(\boldsymbol{F}^t\boldsymbol{y}\right)_i$$

$$\begin{aligned}\left(\boldsymbol{H}^t\boldsymbol{H}\right)_{p,q} &= \sum_{i=1}^{I}\left(\boldsymbol{D}^t\right)_{p,i}\left(\mathrm{diag}(\boldsymbol{\xi})\right)_{i,i}\left(\boldsymbol{F}^t\boldsymbol{F}\,\mathrm{diag}(\boldsymbol{\xi})\boldsymbol{D}\right)_{i,q}\\ &= \sum_{i=1}^{I}\left(\boldsymbol{D}^t\right)_{p,i}\left(\mathrm{diag}(\boldsymbol{\xi})\right)_{i,i}\left(\sum_{u=1}^{I}\left(\boldsymbol{F}^t\boldsymbol{F}\right)_{i,u}\left(\mathrm{diag}(\boldsymbol{\xi})\right)_{u,u}\left(\boldsymbol{D}\right)_{u,q}\right)\end{aligned}$$

Les expressions de $\boldsymbol{F}^t\boldsymbol{F}$ et $\boldsymbol{F}^t\boldsymbol{y}$ que nous avons supposer permettent de retrouver les expressions des matrices $\boldsymbol{H}^t\boldsymbol{H}$, $\boldsymbol{H}^t\boldsymbol{y}$ de la section précédente. En effet, nous avons

$$\left(\boldsymbol{F}^t\boldsymbol{y}\right)_i = \sum_{j=1}^{J}\sum_{k=0}^{K}\pi_{ij}\pi'_{ijk}\boldsymbol{s}_{ijk}^t\boldsymbol{Y}\boldsymbol{c}_i$$

$$\Leftrightarrow \left(\boldsymbol{H}^t\boldsymbol{y}\right)_p = \sum_{i=1}^{I}\left(\boldsymbol{D}^t\right)_{p,i}\left(\mathrm{diag}(\boldsymbol{\xi})\right)_{i,i}\left(\boldsymbol{F}^t\boldsymbol{y}\right)_i$$

$$\Leftrightarrow \left(\boldsymbol{H}^t\boldsymbol{y}\right)_p = \sum_{i=1}^{I}d_{ip}\xi_i\sum_{j=1}^{J}\sum_{k=0}^{K}\pi_{ij}\pi'_{ijk}\boldsymbol{s}_{ijk}^t\boldsymbol{Y}\boldsymbol{c}_i$$

De même, nous avons

$$\left(\boldsymbol{F}^t\boldsymbol{F}\right)_{i,u} = \sum_{jk}\sum_{vw}\pi_{ij}\pi_{uv}\pi'_{ijk}\pi'_{uvw}\boldsymbol{s}_{ijk}^t\boldsymbol{s}_{uvw}\boldsymbol{c}_i^t\boldsymbol{c}_u$$

$$\Leftrightarrow \left(\boldsymbol{H}^t\boldsymbol{H}\right)_{p,q} = \sum_{i=1}^{I}\left(\boldsymbol{D}^t\right)_{p,i}\left(\mathrm{diag}(\boldsymbol{\xi})\right)_{i,i}\left(\sum_{u=1}^{I}\left(\boldsymbol{F}^t\boldsymbol{F}\right)_{i,u}\left(\mathrm{diag}(\boldsymbol{\xi})\right)_{u,u}\left(\boldsymbol{D}\right)_{u,q}\right)$$

$$\Leftrightarrow \left(\boldsymbol{H}^t\boldsymbol{H}\right)_{p,q} = \sum_{i}d_{ip}\xi_i\left(\sum_{u}\left(\sum_{jk}\sum_{vw}\pi_{ij}\pi_{uv}\pi'_{ijk}\pi'_{uvw}\boldsymbol{s}_{ijk}^t\boldsymbol{s}_{uvw}\boldsymbol{c}_i^t\boldsymbol{c}_u\right)\xi_u d_{uq}\right)$$

La démonstration pour les autres matrices est similaire

7.2.4. Calcul rapide des matrices F^tF, F^ty, F^tF^*, $F^{*t}F^*$ et $F^{*t}y$

Ces matrices ne dépendent pas des variables $\boldsymbol{x}$, $\boldsymbol{\xi}$, et γ_b, mais elles dépendent de $\boldsymbol{t}$ via les vecteurs $\boldsymbol{c}_i$. Nous pouvons précalculer une grande partie de ces matrices en dehors de la boucle de Gibbs, il s'agit des scalaires α_{iu}^{s}, $\alpha_{iu}^{s^*}$ et $\alpha_{iu}^{s^{**}}$ et des vecteurs $\boldsymbol{\beta}_i^t$ et $\boldsymbol{\beta}_i^{*t}$.

$$\left(\boldsymbol{F}^t\boldsymbol{y}\right)_i = \underbrace{\sum_{j=1}^{J}\sum_{k=0}^{K}\pi_{ij}\pi'_{ijk}\boldsymbol{s}_{ijk}^t}_{\boldsymbol{\beta}_i^t}\boldsymbol{Y}\boldsymbol{c}_i$$

$$\left(\boldsymbol{F}^{*t}\boldsymbol{y}\right)_i = \underbrace{\sum_{j=1}^{J}\sum_{k=0}^{K}\pi_{ij}\pi'^{*}_{ijk}\boldsymbol{s}_{ijk}^{*t}}_{\boldsymbol{\beta}_i^{*t}}\boldsymbol{Y}\boldsymbol{c}_i$$

$$\left(\boldsymbol{F}^t\boldsymbol{F}\right)_{i,u} = \underbrace{\sum_{jk}\sum_{vw}\pi_{ij}\pi_{uv}\pi'_{ijk}\pi'_{uvw}\boldsymbol{s}_{ijk}^t\boldsymbol{s}_{uvw}}_{\alpha_{iu}^{s}}\underbrace{\boldsymbol{c}_i^t\boldsymbol{c}_u}_{\alpha_{iu}^{c}}$$

$$\left(\boldsymbol{F}^t\boldsymbol{F}^*\right)_{i,u} = \underbrace{\sum_{jk}\sum_{vw}\pi_{ij}\pi_{uv}\pi'_{ijk}\pi'^{*}_{uvw}\boldsymbol{s}_{ijk}^t\boldsymbol{s}_{uvw}^*}_{\alpha_{iu}^{s^*}}\underbrace{\boldsymbol{c}_i^t\boldsymbol{c}_u}_{\alpha_{iu}^{c}}$$

$$\left(\boldsymbol{F}^{*t}\boldsymbol{F}^*\right)_{i,u} = \underbrace{\sum_{jk}\sum_{vw}\pi_{ij}\pi_{uv}\pi'^{*}_{IJK}\pi'^{*}_{uvw}\boldsymbol{s}_{ijk}^{*t}\boldsymbol{s}_{uvw}^*}_{\alpha_{iu}^{s^{**}}}\underbrace{\boldsymbol{c}_i^t\boldsymbol{c}_u}_{\alpha_{iu}^{c}}$$

De plus, nous pouvons encore accélérer les calculs en profitant que les vecteurs $\boldsymbol{c}_i$ échantillonnent des gaussiennes. En effet pour deux vecteurs $\boldsymbol{u}$ et $\boldsymbol{v}$ tel que

$$\boldsymbol{u} = \left[f_1(\theta + T_e) \quad \ldots \quad f_1(\theta + NT_e)\right]^t$$

$$\boldsymbol{v} = \left[f_2(\theta + T_e) \quad \ldots \quad f_2(\theta + NT_e)\right]^t$$

$$f_1(t) = (2\pi)^{-1/2}\gamma_1^{1/2}\exp\left(-\frac{1}{2}\gamma_1(t-t_1)^2\right)$$

$$f_2(t) = (2\pi)^{-1/2}\gamma_2^{1/2}\exp\left(-\frac{1}{2}\gamma_2(t-t_2)^2\right)$$

Nous avons

$$\int_{-\infty}^{+\infty} f_1(t)f_2(t)\mathrm{d}t \approx T_e\boldsymbol{u}^t\boldsymbol{v}$$

D'où

$$\boldsymbol{u}^t\boldsymbol{v} \approx \frac{1}{T_e}\int_{-\infty}^{+\infty} f_1(t)f_2(t)\mathrm{d}t$$

Or nous avons déterminé à la section 7.1

$$\int_{-\infty}^{+\infty} f_1(t)f_2(t)\mathrm{d}t = N\left(t_1\,;t_2,\frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2}\right)$$

D'où

$$\begin{aligned}\boldsymbol{c}_i^t\boldsymbol{c}_u &\approx \frac{1}{T_e^c}N\left(t_i\,;t_u,\frac{\gamma_c\gamma_c}{\gamma_c+\gamma_c}\right)\\ &= \frac{1}{T_e^c}N\left(t_i\,;t_u,\frac{\gamma_c}{2}\right)\end{aligned}$$

La qualité de cette approximation peut être appréciée en comparant la période d'échantillonnage à l'écart type σ de la gaussienne résultante du produit des deux gaussiennes chromatographiques. Nous avons vu au paragraphe 7.1 que son inverse variance est la somme des inverses variances.

Nous avons donc, en prenant l'exemple de l'analyse des peptides du cytochrome C

$$\sigma^{-2} = 2\gamma_c$$
$$\sigma = \frac{\sqrt{2}}{2} \times 6.4 \approx 4.5\,\text{s}$$

Or $T_e^c = 0.48\,\text{s}$. L'approximation est donc valable.

7.2.5. Calcul rapide des matrices G^tG, G^ty, H^tH, H^ty, H^tH^*, $H^{*t}H^*$ et $H^{*t}y$

Après avoir calculé les matrices F^tF, F^ty, F^tF^*, $F^{*t}F^*$ et $F^{*t}y$ nous pouvons obtenir facilement les matrices G^tG, G^ty, H^tH, H^ty, H^tH^*, $H^{*t}H^*$ et $H^{*t}y$.

Nous avons

$$\boldsymbol{G} = \boldsymbol{F}\underbrace{\text{diag}(\boldsymbol{Dx})}_{\boldsymbol{Q}} + \boldsymbol{F}^*\underbrace{\text{diag}(\boldsymbol{Dx}^*)}_{\boldsymbol{Q}^*}$$

Avec $\boldsymbol{Q}$ et $\boldsymbol{Q}^*$ deux matrices diagonales donc symétriques.

$$\boldsymbol{Q}^t = \boldsymbol{Q}$$
$$\boldsymbol{Q}^{*t} = \boldsymbol{Q}^{*t}$$

Nous pouvons en déduire les matrices G^tG, G^ty

$$\begin{aligned}\boldsymbol{G}^t\boldsymbol{G} &= \left(\boldsymbol{F}\ \boldsymbol{Q}\ + \boldsymbol{F}^*\boldsymbol{Q}^*\right)^t\left(\boldsymbol{F}\ \boldsymbol{Q}\ + \boldsymbol{F}^*\boldsymbol{Q}^*\right)\\ &= \left(\boldsymbol{Q}\ \boldsymbol{F}^t + \boldsymbol{Q}^*\boldsymbol{F}^{*t}\right)\left(\boldsymbol{F}\ \boldsymbol{Q}\ + \boldsymbol{F}^*\boldsymbol{Q}^*\right)\\ &= \boldsymbol{Q}\ \boldsymbol{F}^t\boldsymbol{F}\ \boldsymbol{Q}\ + \underbrace{\boldsymbol{Q}\ \boldsymbol{F}^t\boldsymbol{F}^*\boldsymbol{Q}^*}_{\boldsymbol{W}} + \underbrace{\boldsymbol{Q}^*\boldsymbol{F}^{*t}\boldsymbol{F}\ \boldsymbol{Q}}_{\boldsymbol{W}^t} + \boldsymbol{Q}^*\boldsymbol{F}^{*t}\boldsymbol{F}^*\boldsymbol{Q}^*\end{aligned}$$

$$\boldsymbol{G}^t\boldsymbol{y} = \boldsymbol{Q}\ \boldsymbol{F}^t\boldsymbol{y} + \boldsymbol{Q}^*\boldsymbol{F}^{*t}\boldsymbol{y}$$

De même nous avons

$$\boldsymbol{H} = \boldsymbol{F}\underbrace{\text{diag}(\xi)\boldsymbol{D}}_{\boldsymbol{P}} \text{ et } \boldsymbol{H}^* = \boldsymbol{F}^*\underbrace{\text{diag}(\xi)\boldsymbol{D}}_{\boldsymbol{P}}$$

Nous pouvons déduire les matrices H^tH, H^ty, H^tH^*, $H^{*t}H^*$ et $H^{*t}y$. Nous avons donc

$$\begin{aligned}\boldsymbol{H}^t\boldsymbol{H} &= \boldsymbol{D}^t\,\text{diag}(\xi)\boldsymbol{F}^t\boldsymbol{F}\,\text{diag}(\xi)\boldsymbol{D}\\ &= \boldsymbol{P}^t\boldsymbol{F}^t\boldsymbol{F}\boldsymbol{P}\\ \boldsymbol{H}^t\boldsymbol{H}^* &= \boldsymbol{P}^t\boldsymbol{F}^t\boldsymbol{F}^*\boldsymbol{P}\\ \boldsymbol{H}^{*t}\boldsymbol{H}^* &= \boldsymbol{P}^t\boldsymbol{F}^{*t}\boldsymbol{F}^*\boldsymbol{P}\\ \boldsymbol{H}^t\boldsymbol{y} &= \boldsymbol{P}^t\boldsymbol{F}^t\boldsymbol{y}\\ \boldsymbol{H}^{*t}\boldsymbol{y} &= \boldsymbol{P}^t\boldsymbol{F}^{*t}\boldsymbol{y}\end{aligned}$$

Nous avons donc proposé une méthode de calcul rapide des matrices G^tG, G^ty, H^tH, H^ty, H^tH^*, $H^{*t}H^*$ et $H^{*t}y$.

7.3. Calcul de la vraisemblance des concentrations

Montrons que la vraisemblance des concentrations $L(x_i)$ est proportionnelle à une gaussienne. Elle s'obtient à partir de la loi des données

$$L(\boldsymbol{x}) = p(\boldsymbol{Y}|\boldsymbol{x},\boldsymbol{t},\gamma_b) = (2\pi\gamma_b^{-1})^{-N_c N_s/2} \exp\left(-\frac{1}{2}\gamma_b \chi(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t})\right)$$

Or rappelons que la norme de l'erreur peut se mettre sous forme matricielle ou vectorielle. En effet,

$$\chi(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t}) = \left\|\boldsymbol{Y} - \Theta(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t})\right\|^2 = \left\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x} - \boldsymbol{H}^*\boldsymbol{x}^*\right\|^2$$

Développons le calcul pour ces deux formulations

7.3.1. Démonstration utilisant la norme vectorielle

A la section 7.2, nous avons déterminé qu'en utilisant la norme vectorielle, nous pouvons développer la norme des erreurs de la façon suivante

$$\chi(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t}) = (\boldsymbol{y}^t\boldsymbol{y} - 2\boldsymbol{x}^{*t}\boldsymbol{H}^{*t}\boldsymbol{y} + \boldsymbol{x}^{*t}\boldsymbol{H}^{*t}\boldsymbol{H}^*\boldsymbol{x}^*) - 2\boldsymbol{x}^t(\boldsymbol{H}^t\boldsymbol{y} - \boldsymbol{H}^t\boldsymbol{H}^*\boldsymbol{x}^*) + \boldsymbol{x}^t\boldsymbol{H}^t\boldsymbol{H}\boldsymbol{x}$$

En utilisant la norme vectorielle, la vraisemblance peut s'écrire de la façon suivante :

$$\begin{aligned} L(\boldsymbol{x}) &= (2\pi\gamma_b^{-1})^{-N/2} \exp\left(-\frac{1}{2}\gamma_b \chi(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t})\right) \\ &\propto (2\pi\gamma_b^{-1})^{-N/2} \exp\left(-\frac{1}{2}\gamma_b\left(-2\boldsymbol{x}^t(\boldsymbol{H}^t\boldsymbol{y} - \boldsymbol{H}^t\boldsymbol{H}^*\boldsymbol{x}^*) + \boldsymbol{x}^t\boldsymbol{H}^t\boldsymbol{H}\boldsymbol{x}\right)\right) \end{aligned}$$

Montrons que $L(\boldsymbol{x})$ est proportionnelle à la gaussienne centrée en $\boldsymbol{\mu}_L = (\boldsymbol{H}^t\boldsymbol{H})^{-1}(\boldsymbol{H}^t\boldsymbol{y} - \boldsymbol{H}^t\boldsymbol{H}^*\boldsymbol{x}^*)$ et de matrice de covariance $\boldsymbol{R}_L = (\gamma_b\boldsymbol{H}^t\boldsymbol{H})^{-1}$.

Pour cela, montrons que l'expression $(\boldsymbol{x}-\boldsymbol{\mu}_L)^t\boldsymbol{R}_L^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_L)$ est égale, à une constante additive près, à $\gamma_b\left(-2\boldsymbol{x}^t(\boldsymbol{H}^t\boldsymbol{y} - \boldsymbol{H}^t\boldsymbol{H}^*\boldsymbol{x}^*) + \boldsymbol{x}^t\boldsymbol{H}^t\boldsymbol{H}\boldsymbol{x}\right)$.

Commençons par développer l'expression

$$(\boldsymbol{x}-\boldsymbol{\mu}_L)^t\boldsymbol{R}_L^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_L) = \boldsymbol{\mu}_L^t\boldsymbol{R}_L^{-1}\boldsymbol{\mu}_L - 2\boldsymbol{x}^t\boldsymbol{R}_L^{-1}\boldsymbol{\mu}_L + \boldsymbol{x}^t\boldsymbol{R}_L^{-1}\boldsymbol{x}$$

Car $\boldsymbol{R}_L$ est symétrique. Détaillons chacun de ces termes $\boldsymbol{\mu}_L^t\boldsymbol{R}_L^{-1}\boldsymbol{\mu}_L$ est une constante ne dépendant pas de $\boldsymbol{x}$. De plus nous avons

$$\begin{aligned} -2\boldsymbol{x}^t\boldsymbol{R}_L^{-1}\boldsymbol{\mu}_L &= -2\gamma_b\boldsymbol{x}^t\left[(\boldsymbol{H}^t\boldsymbol{H})(\boldsymbol{H}^t\boldsymbol{H})^{-1}\right](\boldsymbol{H}^t\boldsymbol{y} - \boldsymbol{H}^t\boldsymbol{H}^*\boldsymbol{x}^*) \\ &= -2\gamma_b\boldsymbol{x}^t(\boldsymbol{H}^t\boldsymbol{y} - \boldsymbol{H}^t\boldsymbol{H}^*\boldsymbol{x}^*) \end{aligned}$$

Et

$$\boldsymbol{x}^t\boldsymbol{R}_L^{-1}\boldsymbol{x} = \gamma_b\boldsymbol{x}^t(\boldsymbol{H}^t\boldsymbol{H})\boldsymbol{x}$$

Nous avons donc $(\boldsymbol{x}-\boldsymbol{\mu}_L)^t\boldsymbol{R}_L^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_L)$ est égale à $\gamma_b\left(-2\boldsymbol{x}^t(\boldsymbol{H}^t\boldsymbol{y} - \boldsymbol{H}^t\boldsymbol{H}^*\boldsymbol{x}^*) + \boldsymbol{x}^t\boldsymbol{H}^t\boldsymbol{H}\boldsymbol{x}\right)$ à une constante additive près.

La vraisemblance peut donc se réécrire :

$$L(\boldsymbol{x}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_L)^t\boldsymbol{R}_L^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_L)\right)$$

A une constante multiplicative près, c'est l'expression d'une gaussienne multivariée centrée en $\boldsymbol{\mu}_L$ et de matrice de covariance $\boldsymbol{R}_L$. Si l'on extrait de cette vraisemblance pour $\boldsymbol{x}$, la vraisemblance par rapport à chacun des x_i on obtient également à une constante multiplicative près, une gaussienne.

7.3.2. Démonstration utilisant la norme matricielle

Nous avons démontré à la section 7.2.1.4 que la norme matricielle peut se mettre sous la forme

$$\chi(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t}) = \boldsymbol{y}^t\boldsymbol{y} - 2\sum_p \left(x_p b_p + x_p^* b_p^*\right) + \sum_p\sum_q x_p x_q a_{pq} + 2\sum_p\sum_q x_p x_q^* a_{pq}^* + \sum_p\sum_q x_p^* x_q^* a_{pq}^{**}$$

$$= \left(\boldsymbol{y}^t\boldsymbol{y} - 2\sum_p x_p^* b_p^* + \sum_p\sum_q x_p^* x_q^* a_{pq}^{**}\right) - 2\sum_p\left(x_p b_p - \sum_q x_p x_q^* a_{pq}^*\right) + \sum_p\sum_q x_p x_q a_{pq}$$

Avec

$$a_{pq} = \sum_{ijk}\sum_{uvw} d_{ip}d_{uq}\xi_i\xi_u\pi_{ij}\pi_{uv}\pi'_{ijk}\pi'_{uvw}\boldsymbol{s}_{ijk}^t\boldsymbol{s}_{uvw}\boldsymbol{c}_i^t\boldsymbol{c}_u$$

$$a_{pq}^* = \sum_{ijk}\sum_{uvw} d_{ip}d_{uq}\xi_i\xi_u\pi_{ij}\pi_{uv}\pi'_{ijk}\pi'^*_{uvw}\boldsymbol{s}_{ijk}^t\boldsymbol{s}_{uvw}^*\boldsymbol{c}_i^t\boldsymbol{c}_u$$

$$a_{pq}^{**} = \sum_{ijk}\sum_{uvw} d_{ip}d_{uq}\xi_i\xi_u\pi_{ij}\pi_{uv}\pi'^*_{ijk}\pi'^*_{uvw}\boldsymbol{s}_{ijk}^{*t}\boldsymbol{s}_{uvw}^*\boldsymbol{c}_i^t\boldsymbol{c}_u$$

$$b_p = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=0}^{K} d_{ip}\xi_i\pi_{ij}\pi'_{ijk}\boldsymbol{s}_{ijk}^t\boldsymbol{Y}\boldsymbol{c}_i$$

$$b_p^* = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=0}^{K} d_{ip}\xi_i\pi_{ij}\pi'^*_{ijk}\boldsymbol{s}_{ijk}^{*t}\,\boldsymbol{Y}\boldsymbol{c}$$

La vraisemblance est donc égale à un coefficient multiplicatif près ne dépendant pas de x_p

$$L(x_{p_0}) = \left(2\pi\gamma_b^{-1}\right)^{-N_c N_s/2}\exp\left(-\frac{1}{2}\gamma_b\chi(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t})\right)$$

$$\propto \exp\left(-\frac{1}{2}\gamma_b\left(-2\sum_p x_p\underbrace{\left(b_p - \sum_q x_q^* a_{pq}^*\right)}_{b'_p} + \sum_p\sum_q x_p x_q a_{pq}\right)\right)$$

Mettons en valeur l'inconnue x_{p_0}

$$L(x_{p_0}) \propto \exp\left(-\frac{1}{2}\gamma_b\left(-2\sum_{p=1}^{P} x_p b'_p + x_{p_0}^2 a_{p_0p_0} + \sum_{\substack{q=1\\q\neq p_0}}^{P} x_{p_0}x_q a_{p_0q}\right)\right)$$

$$\times\exp\left(-\frac{1}{2}\gamma_b\left(\sum_{\substack{p=1\\p\neq p_0}}^{P} x_p x_{p_0} a_{pp_0} + \sum_{\substack{p=1\\p\neq p_0}}^{P}\sum_{\substack{q=1\\q\neq p_0}}^{P} x_p x_q a_{pq}\right)\right)$$

La vraisemblance est donc égale à un coefficient multiplicatif près ne dépendant pas de x_{p_0}

$$L(x_{p_0}) \propto \exp\left(-\frac{1}{2}\gamma_b\left(x_{p_0}^2 a_{p_0p_0} + x_{p_0}\left(-2b'_{p_0} + 2\sum_{\substack{q=1\\q\neq p_0}}^{P} x_q a_{p_0q}\right)\right)\right)$$

Mettons le polynôme en x_{p_0} sous forme canonique

$$L(x_{p_0}) \propto \exp\left(-\frac{1}{2}a_{p_0p_0}\gamma_b\left(x_{p_0}^2 - 2\frac{x_{p_0}}{a_{p_0p_0}}\left(b'_{p_0} - \sum_{\substack{q=1\\q\neq p_0}}^{P} x_q a_{p_0q}\right)\right)\right)$$

$$\propto \exp\left(-\frac{1}{2}a_{p_0p_0}\gamma_b\left(x_{p_0} - \frac{1}{a_{p_0p_0}}\left(b'_{p_0} - \sum_{\substack{q=1\\q\neq p_0}}^{P} x_q a_{p_0q}\right)\right)^2\right)$$

La vraisemblance en fonction des concentrations x_{p_0} est une gaussienne centrée en $\frac{1}{a_{p_0p_0}}\left(b'_{p_0} - \sum_{\substack{q=1\\q\neq p_0}}^{P} x_q a_{p_0q}\right)$ et d'inverse variance $a_{p_0p_0}\gamma_b$, à un coefficient multiplicatif près.

7.4. Calcul de la Vraisemblance des positions

Nous avons démontré à la section 7.2.1.4 que la norme matricielle peut se mettre sous la forme

$$\chi(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t}) = \boldsymbol{y}^t\boldsymbol{y} - 2\sum_p\left(x_p b_p + x_p^* b_p^*\right) + \sum_p\sum_q x_p x_q a_{pq} + 2\sum_p\sum_q x_p x_q^* a_{pq}^* + \sum_p\sum_q x_p^* x_q^* a_{pq}^{**}$$

$$= \left(\boldsymbol{y}^t\boldsymbol{y} - 2\sum_p x_p^* b_p^* + \sum_p\sum_q x_p^* x_q^* a_{pq}^{**}\right) - 2\sum_p\left(x_p b_p - \sum_q x_p x_q^* a_{pq}^*\right) + \sum_p\sum_q x_p x_q a_{pq}$$

Avec

$$a_{pq} = \sum_{ijk}\sum_{uvw} d_{ip}d_{uq}\xi_i\xi_u\pi_{ij}\pi_{uv}\pi'_{ijk}\pi'_{uvw}\boldsymbol{s}_{ijk}^t\boldsymbol{s}_{uvw}\boldsymbol{c}_i^t\boldsymbol{c}_u$$

$$a_{pq}^* = \sum_{ijk}\sum_{uvw} d_{ip}d_{uq}\xi_i\xi_u\pi_{ij}\pi_{uv}\pi'_{ijk}\pi'^*_{uvw}\boldsymbol{s}_{ijk}^t\boldsymbol{s}_{uvw}^*\boldsymbol{c}_i^t\boldsymbol{c}_u$$

$$a_{pq}^{**} = \sum_{ijk}\sum_{uvw} d_{ip}d_{uq}\xi_i\xi_u\pi_{ij}\pi_{uv}\pi'^*_{ijk}\pi'^*_{uvw}\boldsymbol{s}_{ijk}^{*t}\boldsymbol{s}_{uvw}^*\boldsymbol{c}_i^t\boldsymbol{c}_u$$

$$b_p = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=0}^{K} d_{ip}\xi_i\pi_{ij}\pi'_{ijk}\boldsymbol{s}_{ijk}^t\boldsymbol{Y}\boldsymbol{c}_i$$

$$b_p^* = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=0}^{K} d_{ip}\xi_i\pi_{ij}\pi'^*_{ijk}\boldsymbol{s}_{ijk}^{*t}\boldsymbol{Y}\boldsymbol{c}$$

Nous pouvons réécrire cette norme de façon à mettre en valeur les seules variables dépendant des positions : les vecteurs chromatographiques $\boldsymbol{c}_i$.

$$\chi(\boldsymbol{x},\boldsymbol{x}^*,\boldsymbol{\xi},\boldsymbol{t}) = \boldsymbol{y}^t\boldsymbol{y} - 2\sum_i \boldsymbol{\beta}_i''\boldsymbol{c}_i + \sum_i\sum_u \alpha'_{iu}\boldsymbol{c}_i^t\boldsymbol{c}_u$$

Avec

$$\alpha'_{iu} = \sum_{pjk}\sum_{qvw} d_{ip}d_{uq}\xi_i\xi_u\pi_{ij}\pi_{uv}\left(x_p x_q\pi'_{ijk}\pi'_{uvw}\boldsymbol{s}_{ijk}^t\boldsymbol{s}_{uvw} + 2x_p x_q^*\pi'_{ijk}\pi'^*_{uvw}\boldsymbol{s}_{ijk}^t\boldsymbol{s}_{uvw}^* + x_p^* x_q^*\pi'^*_{ijk}\pi'^*_{uvw}\boldsymbol{s}_{ijk}^{*t}\boldsymbol{s}_{uvw}^*\right)$$

$$\boldsymbol{\beta}_i'^t = \sum_{pjk} d_{ip}\xi_i\pi_{ij}\left(x_p\pi'_{ijk}\boldsymbol{s}_{ijk}^t + x_p^*\pi'^*_{ijk}\boldsymbol{s}_{ijk}^{*t}\right)\boldsymbol{Y}$$

La vraisemblance des positions se réécrit donc

$$L(t_{i_0}) = (2\pi\gamma_b^{-1})^{-N_c N_s/2} \exp\left\{-\frac{1}{2}\gamma_b \chi(\boldsymbol{x}, \boldsymbol{x}^*, \boldsymbol{\xi}, \boldsymbol{t})\right\}$$

$$\propto \exp\left\{-\frac{1}{2}\gamma_b\left(-2\sum_i \boldsymbol{\beta}_i'' \boldsymbol{c}_i + \sum_i \sum_u \alpha'_{iu} \boldsymbol{c}_i^t \boldsymbol{c}_u\right)\right\}$$

Mettons maintenant en valeurs les variables t_i. Pour cela, nous avons montré à la section 7.2.4, que le produit scalaire de deux vecteurs $\boldsymbol{c}_i$ pouvait se formuler à l'aide d'une fonction gaussienne. De même nous pouvons expliciter le produit $\boldsymbol{\beta}_i'' \boldsymbol{c}_i$

$$L(t_{i_0}) \propto \exp\left\{-\frac{1}{2}\gamma_b\left(-2\sum_{i=1}^{I}\sum_{n=1}^{N_c}\beta'_{in} N(nT_e^c; t_i, \gamma_c) + \sum_{i=1}^{I}\sum_{u=1}^{I}\alpha'_{iu}\frac{1}{T_e^c} N\left(t_i; t_u, \frac{\gamma_c}{2}\right)\right)\right\}$$

Avec β'_{in} les éléments du vecteur $\boldsymbol{\beta}_i'$. Concentrons-nous sur la variable t_{i_0} en faisant sortir de la fonction exponentielle les éléments qui n'en dépendent pas

$$L(t_{i_0}) \propto \exp\left\{-\frac{1}{2}\gamma_b\left(-2\sum_{n=1}^{N_c}\beta'_{i_0 n} N(nT_e^c; t_{i_0}, \gamma_c) + \alpha'_{i_0 i_0}\frac{1}{T_e^c} N\left(t_{i_0}; t_{i_0}, \frac{\gamma_c}{2}\right) + 2\sum_{\substack{u=1 \\ u\neq i_0}}^{I}\alpha'_{i_0 u}\frac{1}{T_e^c} N\left(t_{i_0}; t_u, \frac{\gamma_c}{2}\right)\right)\right\}$$

De plus, notons que

$$N\left(t_{i_0}; t_{i_0}, \frac{\gamma_c}{2}\right) = (2\pi)^{-1/2}\left(\frac{\gamma_c}{2}\right)^{1/2} \exp\left(-\frac{1}{2}\frac{\gamma_c}{2}(t_{i_0} - t_{i_0})^2\right)$$

$$= (2\pi)^{-1/2}\left(\frac{\gamma_c}{2}\right)^{1/2}$$

ce terme ne dépend pas donc pas de t_{i_0} et peut être sorti de la fonction exponentielle

$$L(t_{i_0}) \propto \exp\left\{-\frac{1}{2}\gamma_b\left(-2\sum_{n=1}^{N_c}\beta'_{i_0 n} N(nT_e^c; t_{i_0}, \gamma_c) + 2\sum_{\substack{u=1 \\ u\neq i_0}}^{I}\frac{\alpha'_{i_0 u}}{T_e^c} N\left(t_{i_0}; t_u, \frac{\gamma_c}{2}\right)\right)\right\}$$

$$= \exp\left\{\sum_{n=1}^{N_c}\beta'_{i_0 n}\gamma_b N(nT_e^c; t_{i_0}, \gamma_c) - \sum_{\substack{u=1 \\ u\neq i_0}}^{I}\frac{\alpha'_{i_0 u}\gamma_b}{T_e^c} N\left(t_{i_0}; t_u, \frac{\gamma_c}{2}\right)\right\}$$

$$= \exp\left\{\sum_{n=1}^{N_c}\beta'_{i_0 n}\gamma_b N(nT_e^c; t_{i_0}, \gamma_c) - \sum_{\substack{u=1 \\ u\neq i_0}}^{I}\frac{\alpha'_{i_0 u}\gamma_b}{T_e^c} N\left(t_u; t_{i_0}, \frac{\gamma_c}{2}\right)\right\}$$

Avec

$$\alpha'_{iu} = \sum_{pjk}\sum_{qvw} d_{ip} d_{uq} \xi_i \xi_u \pi_{ij} \pi_{uv} \left(x_p x_q \pi'_{ijk} \pi'_{uvw} \boldsymbol{s}_{ijk}^t \boldsymbol{s}_{uvw} + 2x_p x_q^* \pi'_{ijk} \pi'^*_{uvw} \boldsymbol{s}_{ijk}^t \boldsymbol{s}_{uvw}^* + x_p^* x_q^* \pi'^*_{ijk} \pi'^*_{uvw} \boldsymbol{s}_{ijk}^{*t} \boldsymbol{s}_{uvw}^*\right)$$

$$\boldsymbol{\beta}_i'^t = \sum_{pjk} d_{ip} \xi_i \pi_{ij} \left(x_p \pi'_{ijk} \boldsymbol{s}_{ijk}^t + x_p^* \pi'^*_{ijk} \boldsymbol{s}_{ijk}^{*t}\right)\boldsymbol{Y} = \begin{bmatrix}\beta'_{i1} & \dots & \beta'_{iN_c}\end{bmatrix}$$

La log-vraisemlance des positions est donc constitué d'une somme de fonctions gaussiennes.

8. Bibliographie personnelle

8.1. Traitement de données protéomique :

- Brevet
  - G. Strubel, J.-F. Giovannelli, and P. Grangeat, "Procédé d'estimation de concentrations de molécules dans un relevé d'échantillon et appareillage," Brevet Français n°07 57 131 du 22.08.2007.
- Communication à un congrès international
  - G. Strubel, J.-F. Giovannelli, C. Paulus, L. Gerfault, and P. Grangeat, "Bayesian estimation for molecular profile reconstruction in proteomics based on liquid chromatography and mass spectrometry," in *IEEE EMBS 2007.* pp. 5979-5982.
  - C. Paulus, S. Bonnet, L. Gerfault, E. Mery, G. Strubel, F. Ricoul, and P. Grangeat, "Chromatographic alignment combined with chemometrics profile reconstruction approaches applied to LC-MS data," in *IEEE EMBS 2007*, pp. 5983-5986.
  - C. Paulus, V. Brun, L. Gerfault, G. Strubel, J. Garin, and P. Grangeat, "Quantitative proteomics based on spectrum signal model," in *Human Proteome Organisation*, Seoul, 2007.
  - P. Grangeat, N. Sarrut, L. Gerfault, C. Paulus, G. Strubel, J.-F. Giovannelli, E. Mery, A. Fonverne, C. Demesmay, R. Ossig, J. Schnekenburger, V. Brun, A. Dupuis, J. Garin, M. Kalaitzakis, V. Kritsotakis, M. Tsiknakis, D. Kafetzopoulos, M. Perez, C. Reina, and B. Jordan, "Micro-nano technologies and information processing applied to proteomic analysis with high sensitivity " in *8th IEEE EMBS international summer school on biomedical imaging*, Berder Island, France 2008.
  - G. Strubel, C. Paulus, E. Mery, L. Gerfault, F. Ricoul, J.-F. Giovannelli, and P. Grangeat, "Robust protein quantification in mass spectrometry," in *Human Proteome Organization*, Amsterdam, Pays-Bas, 2008.
- Communication à un congrès national
  - G. Strubel, J.-F. Giovannelli, C. Paulus, L. Gerfault, and P. Grangeat, "Reconstruction bayésienne de profils moléculaires," in *Colloque GRETSI* Troyes, 2007.
  - C. Paulus, G. Strubel, L. Gerfault, and P. Grangeat, "Robustesse des approches chimiométriques pour la reconstruction de profils moléculaires," in *Colloque Gretsi* Troyes, 2007.
  - P. Grangeat, C. Paulus, G. Strubel, J.-F. Giovannelli, L. Gerfault, E. Mery, F. Mittler, N. Sarrut, V. Brun, A. Dupuis, and J. Garin, "Reconstruction de profils moléculaires pour la protéomique haute sensibilité à composants intégrés," in *première journée du groupe AQS (Analyse d'images, quantification et statistique)*, Paris, 2008.

8.2. Etalonnage géométrique de scanner X

- Communication à un congrès international
  - G. Strubel, R. Clackdoyle, C. Mennessier, and F. Noo, "Analytic calibration of cone-beam scanners," in *Nuclear Science Symposium Conference Record, 2005 IEEE*, 2005, pp. 2731-2735.
- Rapport de Master recherche
  - G. Strubel, "Etalonnage géométrique en tomographie conique," Lyon, Master Recherche SIDS-ISSI, 2005.

9. Bibliographie

[1] Alberts, Bray, Johnson, Lewis, Raff, Roberts, and Walter, *L'essentiel de la biologie cellulaire*, 2ème édition ed.: Flammarion Medecine Sciences, 2005.

[2] L. Stryer, J. M. Berg, and J. L. Tymoczko, *Biochimie*, 5e édition ed.: Flammarion Medecine-Sciences, 2003.

[3] C. Moussard, *Biologie moléculaire. Biochimie des communications cellulaires*: De Boeck, 2005.

[4] J. D. Wulfkuhle, L. A. Liotta, and E. F. Petricoin, "Proteomic applications for the early detection of cancer," *Nat Rev Cancer,* vol. 3, pp. 267-75, Apr 2003.

[5] E. F. Petricoin, K. C. Zoon, E. C. Kohn, J. C. Barrett, and L. A. Liotta, "Clinical proteomics: translating benchside promise into bedside reality," *Nat Rev Drug Discov,* vol. 1, pp. 683-95, Sep 2002.

[6] H. Mischak, R. Apweiler, R. E. Banks, M. Conaway, J. Coon, A. Dominiczak, J. H. H. Ehrich, D. Fliser, M. Girolami, H. Hermjakob, D. Hochstrasser, J. Jankowski, B. A. Julian, W. Kolch, Z. A. Massy, C. Neusuess, J. Novak, K. Peter, K. Rossing, J. Schanstra, O. J. Semmes, D. Theodorescu, V. Thongboonkerd, E. M. Weissinger, J. E. Van Eyk, and T. Yamamoto, "Clinical proteomics: A need to define the field and to begin to set adequate standards." vol. 1, 2007, pp. 148-156.

[7] E. P. Diamandis, "Mass Spectrometry as a Diagnostic and a Cancer Biomarker Discovery Tool: Opportunities and Potential Limitations." vol. 3, 2004, pp. 367-378.

[8] N. L. Anderson and N. G. Anderson, "The Human Plasma Proteome: History, Character, and Diagnostic Prospects," *Molecular & Cellular Proteomics,* vol. 1, pp. 845-867, November 1, 2002 2002.

[9] S. F. Kingsmore, "Multiplexed protein measurement: technologies and applications of protein and antibody arrays," *Nat Rev Drug Discov,* vol. 5, pp. 310-20, Apr 2006.

[10] R. Aebersold and M. Mann, "Mass spectrometry-based proteomics," *Nature,* vol. 422, pp. 198-207, Mar 13 2003.

[11] E. F. Petricoin, A. M. Ardekani, B. A. Hitt, P. J. Levine, V. A. Fusaro, S. M. Steinberg, G. B. Mills, C. Simone, D. A. Fishman, E. C. Kohn, and L. A. Liotta, "Use of proteomic patterns in serum to identify ovarian cancer," *Lancet,* vol. 359, pp. 572-7, Feb 16 2002.

[12] B. Domon and R. Aebersold, "Mass spectrometry and protein analysis," *Science,* vol. 312, pp. 212-7, Apr 14 2006.

[13] L. S. Ettre, "Nomenclature for chromatography," *Pure &Appl Chem.,* vol. 65, pp. 819-872, 1993.

[14] G. Guiochon, A. Felinger, D. G. Shirazi, and A. M. Katti, *Fundamentals of Preparative and Nonlinear Chromatography* Elsevier, 2006.

[15] A. Felinger, A. Cavazzini, and F. Dondi, "Equivalence of the microscopic and macroscopic models of chromatography: stochastic-dispersive versus lumped kinetic model," *Journal of Chromatography A,* vol. 1043, pp. 149-157, 2004.

[16] A. M. Kuznetsov and H. H. Girault, "Reformulating the Kinetic Approach of Column Chromatography for a Single Component," *Helvetica Chimica Acta,* vol. 80, p. 1176, 1997 1997.

[17] M. J. E. Golay, "Theory of chromatography in open and coated tubular columns with round and rectangular cross sections," in *Gas Chromatography*: Academic Press, New York, 1958, pp. 36-68.
[18] V. B. Di Marco and G. G. Bombi, "Mathematical functions for the representation of chromatographic peaks," *Journal of Chromatography A,* vol. 931, pp. 1-30, 2001.
[19] L. R. Snyder and J. W. Dolan, "The linear solvent strength model for gradient elution," *Adv. Chromatogr.,* vol. 38, pp. 115-188, 1998.
[20] P. Jandera, "Can the theory of gradient liquid chromatography be useful in solving practical problems?," *Journal of Chromatography A,* vol. 1126, pp. 195-218, 2006.
[21] J. Listgarten and A. Emili, "Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry," *Mol Cell Proteomics,* vol. 4, pp. 419-34, Apr 2005.
[22] U. D. Neue, *HPLC Troubleshooting Guide*: Waters, 2001.
[23] P. J. Arpino, *Trends in Analytical Chemistry,* vol. 1, pp. 154-158, 1982.
[24] F. Bökman, "Analytical aspects of atmospheric pressure ionisation in mass spectrometry," Uppsala University, Suède, 2002.
[25] R. B. Cole, "Some tenets pertaining to electrospray ionization mass spectrometry," *Journal of Mass Spectrometry,* vol. 35, pp. 763-772, 2000.
[26] T. C. Rohner, N. Lion, and H. H. Girault, "Electrochemical and theoretical aspects of electrospray ionisation," *Physical Chemistry Chemical Physics,* vol. 6, p. 3056, 2004.
[27] W. Gilbert, *De Magnete*. Londres, 1600.
[28] T. C. Rohner, N. Lion, and H. H. Girault, "Electrochemical and theoretical aspects of electrospray ionisation " *Physical Chemistry Chemical Physics,* vol. 6, pp. 3056-3068, 2004.
[29] A. Jaworek and A. Krupa, "Classification of the modes of EHD spraying," *J. Aerosol Sci.,* vol. 30, pp. 873-893 1999.
[30] M. E. Bier and J. E. P. Syka, "US Patent number 5,420,425: Ion trap mass spectrometer system and method ", 1995.
[31] J. C. Schwartz, M. W. Senko, and J. E. P. Syka, "A two-dimensional quadrupole ion trap mass spectrometer " *Journal of the American Society for Mass Spectrometry,* vol. 13, pp. 659-669, 2002.
[32] J. C. Schwartz and M. W. Senko, "US patent 6797950," 2004.
[33] D. J. Douglas, A. J. Frank, and D. Mao, "Linear ion traps in mass spectrometry," *Mass Spectrom Rev,* vol. 24, pp. 1-29, 2005.
[34] K. R. Jonscher and J. R. Yates, 3rd, "The quadrupole ion trap mass spectrometer - a small solution to a big challenge," *Anal Biochem,* vol. 244, pp. 1-15, Jan 1 1997.
[35] R. E. March, "An Introduction to Quadrupole Ion Trap Mass Spectrometry," *Journal of Mass Spectrometry,* vol. 32, pp. 351-369, 1997.
[36] R. E. March, R. J. Hughes, and J. F. J. Todd, *Quadrupole Storage Mass Spectrometry*: Wiley-Interscience 1989.
[37] D. E. Goeringer, W. B. Whitten, J. M. Ramsey, S. A. McLuckey, and G. L. Glish, "Theory of high-resolution mass spectrometry achieved via resonance ejection in the quadrupole ion trap," *Anal. Chem.,* vol. 64, pp. 1434-1439, 1992.
[38] G. Dieter, "Inhomogeneous RF Fields: A Versatile Tool for the Study of Processes with Slow Ions," in *Advances in Chemical Physics*, C.-Y. Ng, M. Baer, I. Prigogine, and S. A. Rice, Eds., 1992, pp. 1-176.
[39] F. W. Aston, *Phil. mag.,* vol. 38, p. 709, 1919.
[40] J. J. Thomson, *Rays of positive electricity*: Longmans, Green & Co, 1913.
[41] A. J. Dempster, *Phys. Rev.,* vol. 11, p. 316, 1918.
[42] W. Paul and H. Steinwedel, "US Patent 2939952," 1960.
[43] E. Mathieu, "Mémoire sur Le Mouvement Vibratoire d'une Membrane de forme Elliptique," *Journal des Mathématiques Pures et Appliquées,* vol. 13, pp. 137-203, 1868.
[44] M. Abramowitz and I. A. Stegun, *Handbook of mathematical functions - with formulas, graphs, and mathematical tables*: Dover publications, 1970.
[45] G. Bolbach, "Les détecteurs en spectrométrie de masse," in *École de Printemps - CJSM*, Saint-Pierre de Chartreuse, 2004.

[46] "PIR Molecular weight calculator http://pir.georgetown.edu/pirwww/search/comp_mw.shtml."
[47] G. Demoment, *Probabilités, modélisation des incertitudes, inférence logique, traitement de données expérimentales. Tome premier, bases de la théorie.*: Université Paris-Sud XI, Faculté des sciences d'Orsay, 2005.
[48] R. T. Cox, *The Algebra of Probable Inference*: Johns Hopkins University Press, Baltimore, 1961.
[49] E. W. Weisstein, "Frobenius Norm. From MathWorld--A Wolfram Web Resource. http://mathworld.wolfram.com/FrobeniusNorm.html ".
[50] S. Wold, "Chemometrics; what do we mean with it, and what do we want from it?," *Chemometrics and Intelligent Laboratory Systems,* vol. 30, pp. 109-115, 1995.
[51] R. G. Brereton, *Chemometrics: Data Analysis for the Laboratory and Chemical Plant* Wiley, 2003.
[52] B. Domon and R. Aebersold, "Challenges and opportunities in proteomics data analysis," *Mol Cell Proteomics,* vol. 5, pp. 1921-6, Oct 2006.
[53] M. J. Müller, "Molecular Scanner Data Analysis," Thèse de doctorat de l'université de Genève, 2003.
[54] A. Gambin, J. Dutkowski, J. Karczmarski, B. Kluge, K. Kowalczyk, J. Ostrowski, J. Poznanski, J. Tiuryn, M. Bakun, and M. Dadlez, "Automated reduction and interpretation of multidimensional mass spectra for analysis of complex peptide mixtures," *International Journal of Mass Spectrometry,* vol. 260, pp. 20-30, 2007.
[55] P. M. Palagi, P. Hernandez, D. Walther, and R. D. Appel, "Proteome informatics I: Bioinformatics tools for processing experimental data," *Proteomics,* vol. 6, pp. 5435-5444, 2006.
[56] H. Tang, R. J. Arnold, P. Alves, Z. Xun, D. E. Clemmer, M. V. Novotny, J. P. Reilly, and P. Radivojac, "A computational approach toward label-free protein quantification using predicted peptide detectability," *Bioinformatics,* vol. 22, pp. e481-8, Jul 15 2006.
[57] G. Wang, W. W. Wu, W. Zeng, C. L. Chou, and R. F. Shen, "Label-Free Protein Quantification Using LC-Coupled Ion Trap or FT Mass Spectrometry: Reproducibility, Linearity, and Application with Complex Proteomes," *Journal of Proteome Research,* vol. 5, pp. 1214-1223, 2006.
[58] R. D. Smith, Y. Shen, and K. Tang, "Ultrasensitive and quantitative analyses from combined separations-mass spectrometry for the characterization of proteomes," *Acc Chem Res,* vol. 37, pp. 269-78, Apr 2004.
[59] S. Sechi and Y. Oda, "Quantitative proteomics using mass spectrometry," *Current Opinion in Chemical Biology,* vol. 7, pp. 70-77, 2003.
[60] S. Julka and F. Regnier, "Quantification in proteomics through stable isotope coding: a review," *J Proteome Res,* vol. 3, pp. 350-63, May-Jun 2004.
[61] W. Yan and S. S. Chen, "Mass spectrometry-based quantitative proteomic profiling," *Brief Funct Genomic Proteomic,* vol. 4, pp. 27-38, May 2005.
[62] T. Nakamura and Y. Oda, "Mass spectrometry-based quantitative proteomics," *Biotechnol Genet Eng Rev,* vol. 24, pp. 147-63, 2007.
[63] V. Brun, A. Dupuis, A. Adrait, M. Marcellin, D. Thomas, M. Court, F. Vandenesch, and J. Garin, "Isotope-labeled protein standards: toward absolute quantitative proteomics," *Mol Cell Proteomics,* vol. 6, pp. 2139-49, Dec 2007.
[64] A. Dupuis, "PSAQ, un mètre étalon pour la protéomique," *Lettre Scientifique de l'iRTSV,* vol. 11, 2008.
[65] A. Felinger, *Data Analysis and Signal Processing in Chromatography*: Elsevier, 1998.
[66] M. Hilario, A. Kalousis, C. Pellegrini, and M. Muller, "Processing and classification of protein mass spectra," *Mass Spectrom Rev,* vol. 25, pp. 409-49, May-Jun 2006.
[67] J. Listgarten, "Analysis of sibling time series data: alignment and difference detection," PhD thesis, University of Toronto, 2007.
[68] L. De Lathauwer, "Signal Processing based on Multilinear Algebra," PhD thesis, Faculty of Engineering, K.U.Leuven (Leuven, Belgium), 1997.
[69] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems,* vol. 2, pp. 37-52, 1987.

[70] R. Bro, "Multi-way analysis in the food industry: Models, algorithms and applications," Phd Thesis. University of Amsterdam, 1998.
[71] B. Rasmus, "Multiway calibration. Multilinear PLS," *Journal of Chemometrics,* vol. 10, pp. 47-61, 1996.
[72] R. Bro, "PARAFAC. Tutorial and applications," *Chemometrics and Intelligent Laboratory Systems,* vol. 38, pp. 149-171, 1997.
[73] R. A. Harshman, "Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis," *UCLA Working Papers in Phonetics,* vol. 16, pp. 1-84, 1970.
[74] J. V. Candy, *Model-Based Signal Processing* IEEE Computer Society Press 2005.
[75] J. Idier, *Approche bayésienne pour les problèmes inverses*: Traité IC2, Série traitement du signal et de l'image, Hermès, 2001.
[76] E. T. Jaynes, *Probability Theory : The Logic of Science*: Cambridge University Press, 2003.
[77] H. Jeffreys, *Theory of Probability*: Clarendon Press, Oxford, 1939.
[78] D. Mackay, *Information Theory, Inference & Learning Algorithms*: Cambridge University Press, 2002.
[79] I. Hacking, *L'Emergence de la Probabilité* Seuil, 2002.
[80] A. Mohammad-Djafari, J.-F. Giovannelli, G. Demoment, and J. Idier, "Regularization, maximum entropy and probabilistic methods in mass spectrometry data processing problems," *International Journal of Mass Spectrometry,* vol. 215, pp. 175-193, 2002.
[81] S. Moussaoui, D. Brie, A. Mohammad-Djafari, and C. Carteret, "Separation of Non-Negative Mixture of Non-Negative Sources Using a Bayesian Approach and MCMC Sampling," *Signal Processing, IEEE Transactions on,* vol. 54, pp. 4133-4145, 2006.
[82] S. Moussaoui, "Séparation de sources non-négatives. Application au traitement des signaux de spectroscopie.," Thèse de doctorat de l'Université Henri Poincaré. Nancy, 2005.
[83] T. Schwarz-Selinger, R. Preuss, V. Dose, and W. von der Linden, "Analysis of multicomponent mass spectra applying Bayesian probability theory," *Journal of Mass Spectrometry,* vol. 36, pp. 866-874, 2001.
[84] V. Mazet, "Développement de méthodes de traitement de signaux spectroscopiques : estimation de la ligne de base et du spectre de raies," Thèse de doctorat de l'Université Henri Poincaré. Nancy, 2005.
[85] "UniProt the Knowledgebase of The European Bioinformatics Institute and Swiss Institute of Bioinformatics http://www.expasy.org/sprot/."
[86] H. Raiffa and R. Schlaifer, *Applied Statistical Decision Theory* Harvard Business, 1961.
[87] E. W. Weisstein, "Gamma Distribution. From MathWorld, a Wolfram Web Resource. http://mathworld.wolfram.com/GammaDistribution.html ".
[88] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, Second Edition ed.: Springer, 2004.
[89] P. G. A. Pedrioli, J. K. Eng, R. Hubley, M. Vogelzang, E. W. Deutsch, B. Raught, B. Pratt, E. Nilsson, R. H. Angeletti, R. Apweiler, K. Cheung, C. E. Costello, H. Hermjakob, S. Huang, R. K. Julian, E. Kapp, M. E. McComb, S. G. Oliver, G. Omenn, N. W. Paton, R. Simpson, R. Smith, C. F. Taylor, W. Zhu, and R. Aebersold, "A common open representation of mass spectrometry data and its application to proteomics research," *Nat Biotech,* vol. 22, pp. 1459-1466, 2004.
[90] M. M. Dinges, P. M. Orwin, and P. M. Schlievert, "Exotoxins of Staphylococcus aureus," *Clinical Microbiology Reviews,* vol. 13, pp. 16-34, January 1, 2000 2000.
[91] J. Kast, M. Gentzel, M. Wilm, and K. Richardson, "Noise filtering techniques for electrospray quadrupole time of flight mass spectra," *Journal of the American Society for Mass Spectrometry,* vol. 14, pp. 766-776, 2003.
[92] S.-Q. Zhang, X. Zhou, H. Wang, A. Suffredini, D. Gonzales, W.-K. Ching, M. K. Ng, and S. Wong, "Peak Detection with Chemical Noise Removal Using Short-Time FFT for a Kind of MALDI Data," in *The First International Symposium on Optimization and Systems Biology (OSB'07)*, Beijing, China, 2007, pp. 222–231.

10.Liste des figures

Résumé

Des systèmes basés sur la chromatographie et la spectrométrie de masse sont utilisés pour analyser les échantillons biologiques comme l'urine ou le sang. Cette thèse propose une méthode, qui à partir des données, mesure la concentration de biomarqueurs. Dans la première partie du travail, nous élaborons un modèle décrivant chaque module de la chaîne d'analyse. Cependant, pour s'abstraire des fluctuations expérimentales, notre méthode doit évaluer certains paramètres instrument en plus des concentrations. La seconde partie consiste à traiter ce problème d'estimation non linéaire dans le cadre des approches statistiques bayésiennes. Cette démarche nous permet d'introduire de l'information supplémentaire, sous la forme de lois de probabilité, afin de régulariser le problème. La méthode est structurée autour d'un estimateur de la moyenne *a posteriori*. Sa mise en œuvre algorithmique utilise une boucle de Gibbs incluant un échantillonneur de Metropolis-Hastings.

Mots clefs : problème inverse, approche bayésienne, méthodes de Monte Carlo par chaîne de Markov, mesure de concentration, protéomique clinique, protéine, peptide, spectrométrie de masse, chromatographie, electrospray**.**

Abstract

Systems based on chromatography and mass spectrometry are used to analyse biologic samples like urine or blood. This thesis proposes a method which measures the concentration of biomarkers in data. In the first part we elaborate a model which describes each module of the analytic chain. In order to manage experimental fluctuation, our method estimates some instrumental parameters in addition to concentrations. The second part deals with this non linear estimation problem in a Bayesian statistical framework. This approach gives us the possibility to include additional information by using probability laws to regularize the problem. The method is built on a posterior mean estimator. Its implementation uses a Gibbs algorithm including a Metropolis-Hastings sampler.

Key words: inverse problem, Bayesian approach, Monte Carlo Markov Chain, concentration measurements, clinical proteomics, protein, peptide, mass spectrometry, chromatography, electrospray.

Laboratoire Electronique et Systèmes pour la Santé
Service Technologies pour la Détection
Département micro Technologies pour la Biologie et la Santé
Commissariat à l'Energie Atomique de Grenoble
17 rue des Martyrs
38054 Grenoble Cedex 9, France