Reconstruction bayésienne de profils moléculaires

G.STRUBEL[1], J.-F. GIOVANNELLI[2], C. PAULUS[1], L. GERFAULT[1], P.GRANGEAT[1]

[1]LETI, MINATEC, Département des microTechnologies pour la Biologie et la Santé, CEA-GRENOBLE, 17 Rue des Martyrs, 38054 GRENOBLE Cedex 9, France, Tel : 0438789049 - Fax :0438785787

[2]LSS, (CNRS-Supélec-UPS) Supélec, Plateau de Moulon, 91192 Gif-sur-Yvette Cedex, France

gregory.strubel@cea.fr giova@lss.supelec.fr

Résumé – L'étude des protéines laisse entrevoir de grands espoirs pour la médecine de demain. Cependant pour répondre à ces promesses, les méthodes actuelles doivent gagner en sensibilité, en spécificité et en robustesse. Dans cette optique, le CEA développe un laboratoire sur puce et des méthodes de traitement numérique dédiées aux analyses protéomiques par LC-MS. Nous présentons dans cet article une approche bayésienne pour la reconstruction des profils de concentrations de protéines. Dans un premier temps nous proposons un modèle pour le dispositif de mesure complet. Puis nous décrivons une méthode d'estimation des concentrations de protéines présentes, les paramètres instrumentaux étant fixés. Enfin nous estimons conjointement les concentrations et les paramètres instrumentaux.

Abstract – There is a need to precisely measure concentration of proteins in biological substance for early diagnosis of disease or knowledge of fundamental biological processes. This paper focuses on data processing of proteomic experiments combining nano liquid chromatography and mass spectrometry techniques. Experimental fluctuations of this process raise an interest for robust methods. Consequently, we propose a model of this acquisition system and a probabilistic Bayesian method to find an estimate of proteins' concentrations.

1. Introduction

L'avènement de la biologie moléculaire pendant la seconde moitié du XX^{ème} siècle a fait comprendre les mécanismes en jeu à l'intérieur de chaque cellule, et leur importance pour le fonctionnement global des organismes. Elle a mis en valeur le patrimoine génétique qui fait l'individualité de chacun et informe des prédispositions à l'égard des maladies. Parallèlement à l'étude de l'ADN, le monde des sciences de la vie développe des méthodes qui permettent de connaître les molécules fabriquées par la cellule à un moment donné. On accède ainsi à une information variant dans le temps permettant de mieux comprendre les mécanismes régissant la vie cellulaire. Parmi ces molécules, les protéines, expression de l'ADN, offrent des débouchés à la médecine de demain, pouvant aider au diagnostic et à la thérapie.

La chromatographie liquide associée à la spectrométrie de masse (LC-MS) est un outil incontournable pour analyser les fluides biologiques complexes, particulièrement dans le domaine de la protéomique [1]. De plus, certaines protéines, les marqueurs cancéreux, donnent de bonnes indications sur l'apparition et l'évolution d'un cancer [2]. Actuellement, la quantification est réalisée grâce à une méthode combinant reconnaissance moléculaire et détection optique, la méthode ELISA [3]. Cependant, les protéines les plus intéressantes, en tant que marqueurs cancéreux, sont en très faible concentration et noyées dans un mélange de protéines très semblables. L'utilisation de techniques combinant chromatographie, spectrométrie de masse et traitement de données s'avère donc nécessaire pour séparer, identifier et quantifier les protéines. A travers le projet européen Loccandia [4], le CEA développe un micro-système d'analyse pour la nanoLC-MS, qui promet de dépasser les performances de sensibilité et de spécificité des techniques actuelles. Un tel dispositif permettra de suivre l'évolution de la concentration de nouveaux marqueurs cancéreux et laisse espérer un meilleur diagnostic. Nous présentons dans cet article une méthode bayésienne adaptée à la LC-MS.

De nombreuses méthodes de traitement de données associées aux données LC-MS ont été développées ces dernières années se basant sur les méthodes de traitement des années 90. On peut citer les travaux de Muller [5] et plus récemment de Gambin [6]. La caractéristique de ces méthodes est le découpage du traitement en plusieurs étapes séquentielles (estimation du bruit, détection des pics les plus grands, regroupement des pics, quantification). Le découpage des traitements induit des erreurs supplémentaires dans la quantification, et une difficulté de réglage des hyperparamètres et de la manière optimale d'ordonnancer les étapes [7]. Ainsi Listgarten propose une méthode bayésienne globale inspirée de la reconnaissance vocale et basée sur la comparaison de plusieurs expériences[8], mais qui n'essaie pas d'injecter de

connaissance a priori sur les molécules observées. Nous nous concentrerons dans cet article sur l'estimation des concentrations de protéines spécifiques ce qui nous permet d'injecter de l'information spécifique.

2. Modèle instrument

Le système est formé de deux dispositifs, une colonne de chromatographie et un spectromètre de masse. Ainsi le mélange de protéines est analysé suivant deux dimensions correspondant à deux paramètres physico-chimiques.

La colonne de chromatographie est un système permettant de retarder différemment chacune des protéines. L'entrée et la sortie de ce système sont des fonctions représentant une quantité de protéines en fonction du temps. La propagation des protéines dans une colonne de chromatographie peut être décrite par une équation différentielle de convection-diffusion monodimensionelle [9]. Elle peut être également approchée par un modèle convolutif avec une réponse impulsionelle gaussienne, dont la moyenne T_i est appelée temps de rétention de la protéine i. Un système appelé boucle d'injection chromatographique génère une entrée impulsionnelle. Le signal de sortie d'une protéine $c_i(t)$ peut donc être décrit par l'équation suivante :

$$c_i(t) = \left(2\pi\gamma_c^{-1}\right)^{-1/2} \exp\left(-0.5\gamma_c (t-T_i)^2\right),$$

où γ_c est l'inverse variance de la gaussienne.

À chaque instant, la sortie de la colonne est analysée par le spectromètre de masse. Un spectromètre de masse est sensible au rapport entre la masse d'une molécule et sa charge. Il prend en entrée une certaine quantité de matière et fournit en sortie une fonction donnant la quantité de molécules en fonction de leur rapport masse sur charge.

Une protéine peut regrouper plusieurs molécules de même formule peptidique mais portant un nombre de charges différent ou n'ayant pas la même masse. En effet, chaque élément chimique existe sous la forme de plusieurs isotopes. Classiquement, une protéine pourra avoir jusqu'à 3 neutrons et jusqu'à 3 charges supplémentaires. Le spectre de masse théorique d'une protéine est donc formé de plusieurs impulsions de Dirac. Le spectromètre de masse pourra être modélisé par un modèle convolutif gaussien [10]. Finalement, nous obtenons le modèle suivant pour le signal d'une protéine :

$$s_i(m) = \sum_{j=1}^{3}\sum_{k=0}^{3} \Pi_{ijk} \exp\left(-0.5\gamma_s\left(m - \frac{M_i + kM_n}{j}\right)^2\right)$$

avec $\Pi_{ijk} = \dfrac{\gamma_s^{1/2}\pi_{ij}\pi'_{ik}}{(2\pi)^{1/2}}$, $\sum_j \pi_{ij} = 1$ et $\sum_k \pi'_{ik} = 1$.

Les paramètres de l'équation sont :

- j, le nombre de charges portées par la protéine,
- k, le nombre de neutrons supplémentaires,
- π_{ij}, proportions de la protéine i ayant j charges,
- π'_{ik}, proportions de la protéine i ayant k neutrons supplémentaires,
- γ_s, inverse variance des pics spectrométriques,
- M_i, masse de la protéine i sans neutrons supplémentaires,
- M_n, masse d'un neutron.

Le signal global donne une image dont l'axe des abscisses représente le temps, l'axe des ordonnées la masse et l'intensité des pixels représente la quantité de protéines détectée.

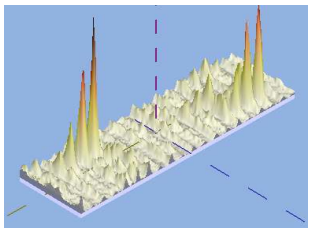

Fig 1 : Exemple de signal visualisé avec le logiciel MSight (SIB).

En première approximation, le système est séparable et linéaire :

$$Y(t,m) = \sum_{i=1}^{N} x_i\, c_i(t)\, s_i(m) + B(t,m) \qquad (1)$$

Nous nous intéressons à N protéines d'intérêt. x_i est la concentration de la protéine i dans l'échantillon initial. $Y(t,m)$ sont les données observées.

L'équation (1) peut se discrétiser sous la forme :

$$\boldsymbol{Y} = \sum_{i=1}^{N} x_i \boldsymbol{s}_i \boldsymbol{c}_i^t + \boldsymbol{B}$$

La relation est linéaire donc il existe une matrice $\boldsymbol{H}$ qui relie $\boldsymbol{y}$, le vecteur des données, à $\boldsymbol{x}$ le vecteur des concentrations.

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

$\boldsymbol{y}$ est la forme vectorisé de la matrice $\boldsymbol{Y}$ obtenue par concaténation des colonnes de la matrice. $\boldsymbol{y}$ et $\boldsymbol{b}$ sont des vecteurs colonnes de dimension M ($M \gg N$).

3. Inversion

3.1 Estimation des variables d'intérêt x

Dans un premier temps, nous allons construire une méthode d'inversion bayésienne utilisant des distributions *a priori* gaussiennes pour le bruit et pour les paramètres d'intérêt.

$$p\left(\boldsymbol{y}|\boldsymbol{x},\gamma_b,\gamma_c\right) = \left(2\pi\gamma_b^{-1}\right)^{-M/2} \exp\left(-0.5\gamma_b \left\|\boldsymbol{y} - \boldsymbol{H}_{\gamma_c}\boldsymbol{x}\right\|^2\right)$$

$$p\left(\boldsymbol{x}|\gamma_x\right) = \left(2\pi\gamma_x^{-1}\right)^{-N/2} \exp\left(-0.5\gamma_x \left\|\boldsymbol{x} - \boldsymbol{x_0}\right\|^2\right)$$

γ_b et γ_x sont les inverse variances de ces distributions. Leurs valeurs permettent de régler l'information dont nous disposons sur $\boldsymbol{x}$ et $\boldsymbol{b}$.

Ces distributions nous permettent de calculer la distribution *a posteriori* gaussienne suivante :

$$p\left(\boldsymbol{x}|\boldsymbol{y},\gamma_b,\gamma_c,\gamma_x\right) \propto \exp\left(-\frac{\gamma_b}{2}\left\|\boldsymbol{y} - \boldsymbol{H}_{\gamma_c}\boldsymbol{x}\right\|^2 - \frac{\gamma_x}{2}\left\|\boldsymbol{x} - \boldsymbol{x_0}\right\|^2\right)$$

La distribution étant gaussienne, l'estimateur de la moyenne, de la médiane et du maximum sont égaux. Nous choisissons l'estimateur de la moyenne a posteriori.

$$\hat{\boldsymbol{x}} = \left(\boldsymbol{H}^t \boldsymbol{H} + \frac{\gamma_x}{\gamma_b} \boldsymbol{I} \right)^{-1} \left(\boldsymbol{H}^t \boldsymbol{y} + \frac{\gamma_x}{\gamma_b} \boldsymbol{x}_0 \right)$$

Nous pouvons faire trois remarques à propos de ce premier estimateur. Premièrement, il a l'avantage de se calculer facilement grâce à sa structure linéaire. De plus même si la matrice $\boldsymbol{H}$ est de grande taille (MxN), $\boldsymbol{H}^t\boldsymbol{H}$ n'est que de taille NxN ce qui permet une inversion rapide. Enfin, si nous n'injectons pas de connaissance a priori sur les concentrations des protéines, γ_x tend vers 0 et nous retrouvons la solution des moindres carrés. Par la suite nous considérerons γ_x comme étant nul.

3.2 Estimation conjointe

Dans le cadre où les paramètres instruments sont moins connus, nous pouvons mettre en œuvre une démarche similaire qui conduit à estimer conjointement paramètres d'intérêt et paramètres instruments. Nous verrons qu'il s'agit d'une extension de la méthode développée à la section précédente. Nous étudierons dans cette section l'estimation conjointe des paramètres $\boldsymbol{x}$, γ_b et γ_c.

3.2.1 Distribution *a posteriori*

Nous devons tout d'abord définir les distributions *a priori* de chacun des trois paramètres. Ainsi $\boldsymbol{x}$ aura la même distribution *a priori* que dans la section précédente. Un choix classique pour la distribution *a priori* de l'inverse variance d'un bruit gaussien est la distribution de Jeffrey. Enfin nous choisissons une distribution uniforme entre $\gamma_{c,min}$ et $\gamma_{c\,max}$ pour γ_c.

$$\gamma_b \sim J \text{ et } \gamma_c \sim \mathcal{U}_{[\gamma_{c\min};\gamma_{c\max}]}$$

Ce qui nous donne la distribution *a posteriori* suivante :

$$p(\boldsymbol{x}, \gamma_b, \gamma_c | \boldsymbol{y}) \propto p(\boldsymbol{x}) p(\gamma_b) p(\gamma_c) p(\boldsymbol{y} | \boldsymbol{x}, \gamma_b, \gamma_c) \quad (2)$$

3.2.2 Echantillonnage stochastique

La distribution *a posteriori* (2) n'étant plus gaussienne, nous calculons l'estimateur de la moyenne par une méthode d'échantillonnage stochastique. En effet si nous disposons d'un générateur de vecteurs aléatoires simulant la loi (2), l'estimateur de la moyenne s'approche en calculant la moyenne des échantillons produits par ce générateur :

$$\begin{bmatrix} \hat{x} & \hat{\gamma}_b & \hat{\gamma}_c \end{bmatrix}^t = \frac{1}{K} \sum_{k=1}^{K} \begin{bmatrix} x^{(k)} & \gamma_b^{(k)} & \gamma_c^{(k)} \end{bmatrix}$$

La construction d'un tel générateur peut se faire grâce à une structure de Gibbs [11], qui permet de transformer l'échantillonnage d'une loi multivariée en un problème d'échantillonnage de lois plus simples, soit monovariées, soit gaussiennes.

En effet, échantillonner une distribution *a posteriori* (2) est équivalent à échantillonner successivement les conditionnelles *a posteriori*. L'algorithme du générateur s'écrit donc :

- Initialisation $\boldsymbol{x}^{(0)}$, $\gamma_b^{(0)}$, $\gamma_c^{(0)}$.
- Pour k=1 jusqu'à K
  - Générer $\boldsymbol{x}^{(k+1)} \sim p\left(\boldsymbol{x} \middle| \boldsymbol{y}, \gamma_b^{(k)} \gamma_c^{(k)}\right)$,
  - Générer $\gamma_b^{(k+1)} \sim p\left(\gamma_b \middle| \boldsymbol{y}, \boldsymbol{x}^{(k+1)}, \gamma_c^{(k)}\right)$,
  - Générer $\gamma_c^{(k+1)} \sim p\left(\gamma_c \middle| \boldsymbol{y}, \boldsymbol{x}^{(k+1)}, \gamma_b^{(k+1)}\right)$.
- Fin pour.

3.2.2.1 *Distribution conditionnelle a posteriori de x*

La variable $\boldsymbol{x}$ est distribuée selon la même gaussienne multivariée que celle présentée à la section 3.1 :

$$p\left(\boldsymbol{x} \middle| \boldsymbol{y}, \gamma_b, \gamma_c\right) = (2\pi)^{-\frac{N}{2}} |\boldsymbol{R}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^t \boldsymbol{R}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right)$$

avec $\boldsymbol{\mu} = \left(\boldsymbol{H}_{\gamma_c}^t \boldsymbol{H}_{\gamma_c} \right)^{-1} \boldsymbol{H}_{\gamma_c}^t \boldsymbol{y}$, $\boldsymbol{R} = \gamma_b^{-1} \left(\boldsymbol{H}_{\gamma_c}^t \boldsymbol{H}_{\gamma_c} \right)^{-1}$.

On échantillonne facilement les distributions gaussiennes multivariées de la façon suivante :

- Calculer la décomposition de Cholesky de $\boldsymbol{R} = \boldsymbol{\Lambda}^t \boldsymbol{\Lambda}$.
- Générer $\boldsymbol{g}$ un vecteur de N variables gaussiennes indépendantes.
- Calculer $\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{\Lambda}^t \boldsymbol{g}$.

3.2.2.2 *Distribution conditionnelle a posteriori de γ_b*

La variable γ_b est distribuée selon une distribution gamma :

$$p\left(\gamma_b \middle| \boldsymbol{y}, \boldsymbol{x}, \gamma_c\right) = \frac{\gamma_b^{\alpha-1}}{\beta^\alpha \Gamma(\alpha)} \exp\left(-\frac{\gamma_b}{\beta} \right)$$

avec $\alpha = M/2$, $\beta = \left\| \boldsymbol{y} - \boldsymbol{H}_{\gamma_c} \boldsymbol{x} \right\|^2 / 2$, Γ est la fonction gamma.

On a utilisé Matlab qui fournit des routines générant des variables sous des distributions gamma.

3.2.2.3 *Distribution conditionnelle a posteriori de γ_c*

La variable γ_c n'est pas distribuée selon une loi de probabilité classique à cause de la dépendance en γ_c de la matrice $\boldsymbol{H}$.

$$p\left(\gamma_c \middle| \boldsymbol{y}, \boldsymbol{x}, \gamma_b\right) \propto \exp\left(-\frac{1}{2} \gamma_b \left\| \boldsymbol{y} - \boldsymbol{H}_{\gamma_c} \boldsymbol{x} \right\|^2 \right) \mathcal{U}_{[\gamma_{c\,\min};\gamma_{c\max}]}(\gamma_c) \quad (3)$$

Nous utilisons pour l'échantillonner l'algorithme de Metropolis-Hasting [11], qui permet de générer des échantillons de n'importe quelle distribution à l'aide d'une distribution instrumentale $Q(\gamma_c)$ pour laquelle nous disposons d'un générateur aléatoire. Nous choisissons ici pour cette loi instrumentale l'a priori de γ_c. La distribution (3) sera notée par la suite $P(\gamma_c)$. L'algorithme se déroule ainsi :

- Générer $\gamma'_c \sim \mathcal{U}_{[\gamma_{c\,\min};\gamma_{c\max}]}$ et $u \sim \mathcal{U}_{[0;1]}$
- Calculer $\delta = \dfrac{P(\gamma'_c)\, Q\left(\gamma_c^{(k)}; \gamma'_c\right)}{P\left(\gamma_c^{(k)}\right) Q\left(\gamma'_c; \gamma_c^{(k)}\right)}$ qui se simplifie ici sous la forme :

$$\delta = \exp\left(-\frac{1}{2} \gamma_b^{(k+1)} \left(\left\| \boldsymbol{y} - \boldsymbol{H}_{\gamma'_c} \boldsymbol{x}^{(k+1)} \right\|^2 - \left\| \boldsymbol{y} - \boldsymbol{H}_{\gamma_c^{(k)}} \boldsymbol{x}^{(k+1)} \right\|^2 \right) \right)$$

- Si $\delta > \log(u)$ Alors $\gamma_c^{(k+1)} \leftarrow \gamma'_c$ Sinon $\gamma_c^{(k+1)} \leftarrow \gamma_c^{(k)}$

4. Mise en œuvre sur des données expériementales

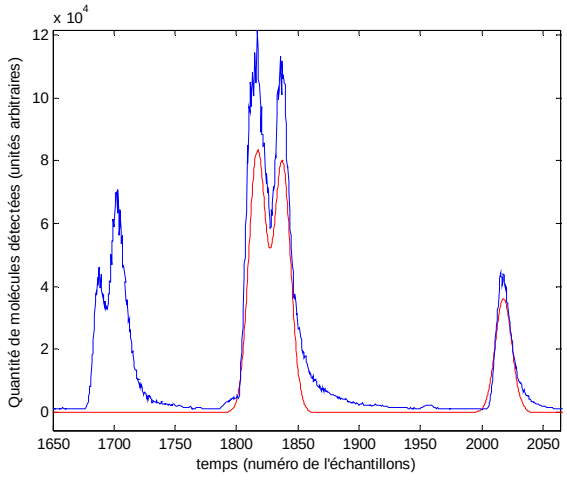

FIG 2 : Données réelles et reconstruites par l'algorithme. Les trois pics reconstruits correspondent aux protéines cibles, les autres à des contaminants inconnus. L'abscisse représente le temps, l'ordonnée correspond à l'intensité maximum du spectre de masse : $\max_m Y(t,m)$.

Nous testons l'algorithme sur un mélange de protéines issues de la digestion du cytochrome C, molécule intervenant dans la respiration. Et plus particulièrement des protéines ayant comme code en acide aminés TGPNLHGLFGR, MIFAGIK, EDLIAYLK. Ces trois protéines sont dans des concentrations identiques dans chaque expérience. L'analyse chromatographique en mode gradient dure 60 min, et les T_i et les M_i correspondant à chaque protéine ont été obtenus grâce à une méthode non décrite dans cet article. Pour cette expérience nous choisissons $\gamma_{c\,min} = 2$, $\gamma_{c\,max} = 3$.

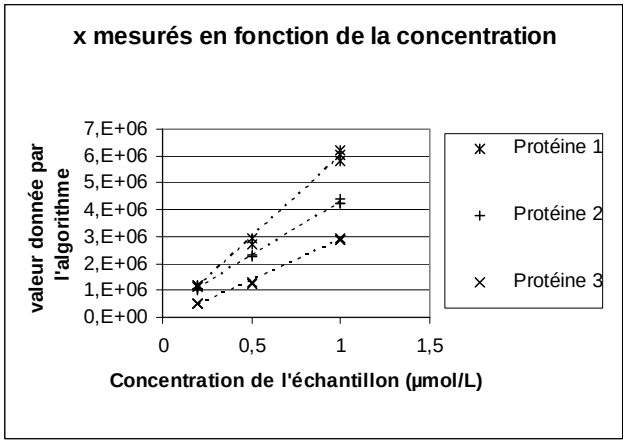

FIG 3 : Estimation des x_i en fonction de la concentration des protéines

La figure 3 nous montre les résultats de l'estimation des x_i sur les trois protéines pour plusieurs concentrations des protéines en entrée du système. Les mesures suivent linéairement les concentrations en entrée du système mais on n'obtient pas la fonction identité souhaitée. Le système comporte donc un gain spécifique à chaque protéine non prévu dans la modélisation (1), dont il faudra tenir compte pour améliorer la quantification.

5. Conclusion

Dans cet article nous proposons une méthode de traitement des données LC-MS prenant en compte un modèle physique du dispositif de mesure. La méthode permet également de s'ajuster aux variations des paramètres de ce modèle. Toutefois la mise en œuvre de l'algorithme proposé a mis en évidence la présence d'un gain à prendre en compte dans la modélisation.

Références

[1] R. Aebersold and M. Mann, "Mass spectrometry-based proteomics," *Nature,* vol. 422, pp. 198-207, Mar 13 2003.

[2] 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.

[3] 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.

[4] "Site web du projet européen Loccandia. www.loccandia.eu."

[5] M. J. Müller, "Molecular Scanner Data Analysis," Université de Genève, 2003.

[6] 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.

[7] 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.

[8] J. Listgarten, "Analysis of sibling time series data: alignment and difference detection," University of Toronto, 2007.

[9] J. C. Giddings, *Dynamics of Chromatography: Principles and Theory Pt. 1* 1965.

[10] 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.

[11] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, Second Edition ed.: Springer, 2004.