Reconstruction de profils moléculaires par inversion d'un modèle paramétrique d'une chaîne d'analyse biologique

Caroline PAULUS, Laurent GERFAULT, Pierre GRANGEAT

CEA, LETI, MINATEC, Laboratoire Electronique et Systèmes pour la Santé
F38054 Grenoble, France
caroline.paulus@cea.fr, laurent.gerfault@cea.fr
pierre.grangeat@cea.fr

Résumé – Cette communication présente une méthode paramétrique adaptative de reconstruction de la concentration de protéines cibles présentes dans un milieu biologique complexe et analysées par un système LC-MS (Chromatographie Liquide et Spectromètre de Masse). L'approche proposée est basée sur l'inversion d'un modèle à espace état décrivant les équations physiques régissant le comportement de la colonne de chromatographie. La nouveauté de cette méthode réside dans la description d'un système complexe en termes de modèle à espace état permettant ainsi une meilleure robustesse de la méthode vis-à-vis du bruit et des fluctuations par rapport aux méthodes classiques. L'efficacité de la méthode est testée sur des données simulées et comparée avec les méthodes de l'état de l'art.

Abstract – This communication presents an adaptive parametric method which enables estimation of targeted proteins' quantity present in complex biological mixtures and analyzed by a LC-MS (Liquid Chromatography and Mass Spectrometry) system. The proposed approach is based on the inversion of a parametric state space model based on physical equations describing the chromatographic column behaviour. The method's novelty comes from the description of a complex system in terms of state space model inducing a robust and adaptive algorithm for quantitative proteomics. Efficiency of the algorithm is tested on synthetic dataset and comparison with state-of-the-art methodologies is performed.

1 Introduction

La problématique de cette communication est la détection précoce du cancer à partir d'une chaîne d'analyse biologique partant de l'échantillon de sang contenant un ensemble de protéines et allant jusqu'au diagnostic médical en combinant des technologies liées à la biologie, aux nanotechnologies et au traitement de l'information. Nous nous intéressons à la partie traitement des données, appelée reconstruction de profils moléculaires (cf. Fig.1). Cette étape doit permettre d'estimer la concentration de certaines protéines cibles (caractéristiques de la maladie) présentes dans l'échantillon sanguin à partir de données 2 dimensions, appelées spectrogrammes, composées d'un ensemble de pics représentatifs des protéines et obtenues en sortie de la chaîne d'analyse. Une dimension de ces données est le temps de rétention, associée à la séparation sur une nano colonne de chromatographie liquide (nano-LC), l'autre dimension correspond au rapport masse sur charge obtenu par un spectromètre de masse (MS). Par la suite, nous utiliserons le terme de peptides pour désigner les fragments de protéines obtenus lors de la préparation de l'échantillon.

La méthodologie proposée se base sur une approche de type problème inverse associé à un modèle dynamique direct reliant les inconnues (les concentrations des protéines cibles) aux mesures (les spectrogrammes).
Une des approches les plus classiques pour identifier et quantifier des protéines est basée sur l'étude de certains pics du spectrogramme. De ces pics sont extraites des caractéristiques de type position, forme, ou maximum qui permettent d'identifier et de quantifier les protéines. Ce type d'approche est mis en défaut dans le cas de pics mal séparés, de pics de faible intensité par rapport au bruit ou encore de pics déformés par saturation. L'approche proposée, basée sur l'inversion d'un modèle dynamique paramétrique d'une partie de la chaîne d'acquisition, est plus globale car elle prend en compte toute l'information sur le signal. De plus, elle doit permettre d'améliorer la sensibilité et la robustesse vis-à-vis du bruit et des perturbations. La méthode est fondée à la fois sur un modèle à espace état [3, 4] décrivant les équations physico-chimiques du comportement de la colonne de chromatographie, sur un ensemble d'étalonnages du système et enfin sur une méthode d'inversion par moindre

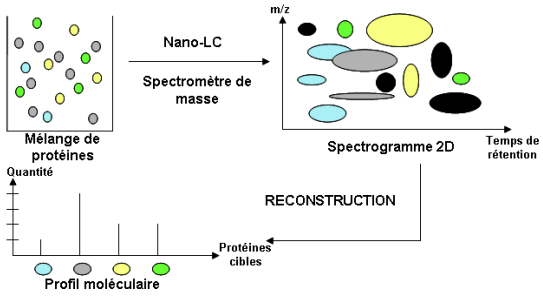

FIG. 1: *Reconstruction de profils moléculaires*

carré pour la quantification.

2 Equations physiques décrivant le fonctionnement de la nano-LC

La chromatographie est une technique de séparation, basée sur la rétention sélective des solutés (dans notre cas, les peptides). La séparation se caractérise par un retard différentiel d'arrivée des molécules en sortie de colonne. Trois éléments entre en jeu: la phase mobile qui assure le transport des solutés, la phase stationnaire qui est en fait la surface fonctionnalisée du microréacteur et le soluté qui par ses propriétés d'interaction avec la phase stationnaire est plus ou moins retardé.
La propagation des peptides dans une colonne de chromatographie peut être décrite par une équation différentielle de convection diffusion monodimensionnelle [2]. Une équation de ce type relie, à l'emplacement z et au temps t, la concentration locale de peptide adsorbé q à la concentration locale de peptide dans la phase mobile c:

$$\frac{\partial c(z,t)}{\partial t} + F\frac{\partial q(z,t)}{\partial t} + u_s\frac{\partial c(z,t)}{\partial z} = D_i\frac{\partial^2 c(z,t)}{\partial z^2}. \quad (1)$$

F est le rapport des volumes occupés par la phase mobile et la phase stationnaire (paramètre lié directement à la géométrie de la colonne), u_s la vitesse de propagation du solvant, D_i le facteur de diffusion contribuant à l'étalement des pics de chromatographie.
L'adsorption est un phénomène cinétique et dynamique (réaction dépendant des concentrations des éléments en présence). Nous nous intéressons à sa formulation en régime stationnaire c'est-à-dire à l'équilibre, appelée isotherme. Ce dernier relie la concentration locale de soluté adsorbé q à la concentration locale de soluté dans la phase mobile c. Un exemple d'isotherme linéaire simple est :

$$q(z,t) = kc(z,t) \quad (2)$$

avec k le facteur de rétention.

3 Modélisation de la chaine d'analyse et reconstruction de profils

3.1 Modèle du signal en sortie du système

On note $M_{i,j,k}$ le spectrogramme du peptide i appartenant à la protéine k dans le mélange j et $M^*_{i,j,k}$ celui du peptide i appartenant à la version alourdie de la protéine k. Les protéines alourdies sont des protéines synthétiques ayant les mêmes propriétés physico-chimiques que la protéine simple, seule une petite différence en masse les distingue. Les protéines alourdies sont ajoutées en quantité connue à l'échantillon biologique pour permettre l'étalonnage de certains paramètres du système. $M_{j,k}$ est la somme des spectrogrammes des N_{pep} peptides de la protéine k et $M^*_{j,k}$ la somme des spectrogrammes des N_{pep} peptides de la protéine k alourdie dans le mélange j.
$M_{i,j,k}$ et $M^*_{i,j,k}$ peuvent s'exprimer sous la forme:

$$M_{i,j,k} = c_{j,k}\ \beta_{i,j,k}\ s_{i,k}\ y^T_{i,k} + B_{i,j,k} \quad (3)$$

$$M^*_{i,j,k} = c^*_{j,k}\ \beta_{i,j,k}\ s^*_{i,k}\ y^T_{i,k} + B^*_{i,j,k} \quad (4)$$

où:

- $c_{j,k}$ et $c^*_{j,k}$ sont les concentrations de la protéine k cible et de la protéine k alourdie dans le mélange j,
- $\beta_{i,j,k}$ est le facteur d'étalonnage de la chaîne biologique pour le peptide i de la protéine k. Il est obtenu en utilisant l'information de concentration de la protéine alourdie $c^*_{j,k}$ (voir au 3.3),
- $y_{i,k}$ est la réponse de la partie du système comprenant la colonne de chromatographie pour le peptide i de la protéine k pour les paramètres physiques p (p comprend notamment F, D_i, u_s, k). Elle est modélisée par un modèle à espace état (voir au 3.2).
- $s_{i,k}$ et $s^*_{i,k}$ sont les réponses (supposées connues) de la partie du système comprenant le spectromètre de masse pour le peptide i de la protéine k et de sa version alourdie.
- $B_{i,j,k}$ et $B^*_{i,j,k}$ sont des bruits blancs gaussiens additifs.

Nous avons N mélanges biologiques analysés pour lesquels les $c^*_{j,k}$ sont connus et les $c_{j,k}$ sont à estimer.

3.2 Modèle à espace état discret

Nous proposons d'utiliser un modèle de type espace état [3, 4] pour décrire le comportement de la colonne de chromatographie. Dans le cas stationnaire, discret et linéaire, un tel système peut s'écrire sous la forme:

$$\begin{cases} x(n+1) = A(p)x(n) + B(p)u(n) \\ y(n) = C(p)x(n) + D(p)u(n) \\ x(0) = x_0(p) \end{cases} \quad (5)$$

avec x le vecteur d'état, p l'ensemble des paramètres physiques du système, $u(n)$ et $y(n)$ respectivement les signaux d'entrée et de sortie du système, x_0 les conditions initiales du vecteur d'état ; $A(p)$, $B(p)$, $C(p)$, $D(p)$ respectivement les matrices d'état, d'entrée, de sortie et de boucle retour du système d'état.
Afin d'obtenir une solution numérique unique, nous spécifions les conditions initiales et aux limites de notre système qui sont choisies de type Dirichlet.
Nous utilisons la méthode des différences finies pour approximer les opérateurs différentiels de l'EDP (eq.1) [1]. Les index i et n sont utilisés pour désigner les dimensions espace et temps. Les pas d'échantillonnage en temps et espace sont Δ_t et Δ_z.
Pour des raisons de stabilité du schéma numérique, nous proposons de faire les approximations suivantes:

- une différence finie décentrée en amont pour approximer la dérivée partielle d'ordre 1 en espace (terme convectif),

– une différence finie centrée pour approximer la dérivée partielle d'ordre 2 en espace (terme diffusif),
– une différence finie décentrée en aval pour approximer la dérivée partielle d'ordre 1 en temps.

Ainsi, nous avons les approximations suivantes:

$$\frac{\partial c(z,t)}{\partial z}\Big|_{i,n} = \frac{1}{\Delta_z}\big[c(i,n) - c(i-1,n)\big] + O(\Delta z) \quad (6)$$

$$\frac{\partial^2 c(z,t)}{\partial z^2}\Big|_{i,n} = \frac{1}{{\Delta_z}^2}\big[c(i+1,n) - 2c(i,n) + c(i-1,n)\big] + O(\Delta_z^2) \quad (7)$$

$$\frac{\partial c(z,t)}{\partial t}\Big|_{i,n} = \frac{1}{\Delta_t}\big[c(i,n+1) - c(i,n)\big] + O(\Delta_t) \quad (8)$$

En remplaçant les dérivées partielles de l'équation (1) par leurs approximations, nous obtenons la relation suivante:

$$\begin{aligned} c(i,n+1) = \quad & \Big(\tfrac{u_s\Delta_z\Delta_t + D_i\Delta_t}{{\Delta_z}^2(1+Fk)}\Big)c(i-1,n) \\ & + \Big(1 - \tfrac{u_s\Delta_z\Delta_t + 2D_i\Delta_t}{{\Delta_z}^2(1+Fk)}\Big)c(i,n) \\ & + \Big(\tfrac{D_i\Delta_t}{{\Delta_z}^2(1+Fk)}\Big)c(i+1,n) \end{aligned} \quad (9)$$

Soit:

$$c(i,n+1) = K(p)c(i-1,n) + J(p)c(i,n) + I(p)c(i+1,n) \quad (10)$$

L'équation précédente peut être mise sous la forme d'un modèle à espace état discret du type de l'équation 5 où les états du vecteur x correspondent aux $c(i,n)$ pour i allant de 1 à nz (le nombre d'états utilisés pour décrire la colonne) avec $nz = L/\Delta_z$ (L, la longueur de la colonne). La matrice $A(p)$ est une matrice tri-diagonale de taille (nz, nz):

$$A(p) = \begin{pmatrix} J(p) & I(p) & 0 & & 0 \\ \ddots & \ddots & \ddots & & \\ 0 & K(p) & J(p) & I(p) & 0 \\ & & \ddots & \ddots & \ddots \\ 0 & & 0 & K(p) & J(p) \end{pmatrix} \quad (11)$$

Le vecteur $B(p)$ de taille $(nz, 1)$ est:

$$B(p) = \begin{pmatrix} 1 & 0 & \dots & 0 \end{pmatrix}^T \quad (12)$$

et le vecteur $C(p)$ de taille $(1, nz)$ est:

$$C(p) = \begin{pmatrix} 0 & \dots & 0 & 1 \end{pmatrix}. \quad (13)$$

De sorte que nous avons $y(n) = c(L/\Delta_z, n)$. La matrice D est un scalaire nul.

Les contraintes de stabilité [1] induisent:

$$\left.\begin{array}{l} I(p) \geq 0 \\ J(p) \geq 0 \\ K(p) \geq 0 \end{array}\right\}. \quad (14)$$

Ainsi, pour des paramètres physiques fixés, les contraintes sur les pas d'échantillonnage en espace et en temps sont:

$$\Delta_t \leq \frac{\Delta_z^2(1+Fk)}{u_s\Delta_z + 2D_i} \quad (15)$$

Après résolution du système, nous obtenons un modèle pour la sortie de la colonne de chromatographie y, notée par la suite $y_{i,k}$ pour faire référence au peptide i de la protéine k.

3.3 Identification des paramètres et inversion du modèle

Avant l'étape de quantification permettant l'estimation des $c_{j,k}$ dans les différents mélanges, il est nécessaire d'estimer les facteurs d'étalonnage du système, $\beta_{i,j,k}$, ainsi qu'un certain nombre de paramètres physiques p du système (notamment D_i et k qui ne peuvent être connus physiquement). Nous avons à disposition N mélanges pour lesquels les $c^*_{j,k}$ sont connus et les $c_{j,k}$ sont à estimer. L'identification par optimisation consiste à estimer les paramètres inconnus en prenant l'optimum d'un critère quadratique portant notamment sur l'erreur. Il est alors nécessaire de construire une fonction coût dont il faut trouver le minimum. Par ses performances éprouvées, l'algorithme de Levenberg-Marquardt est une référence en optimisation non-linéaire car il offre un bon compromis entre robustesse aux conditions initiales et vitesse de convergence. Par conséquent, nous avons choisi d'utiliser cet algorithme.

Etape 1: estimation des facteurs d'étalonnage $\beta_{i,j,k}$ et des paramètres physiques p sur les peptides alourdis

$$\min_{\beta_{i,j,k},p} \left\| M^*_{i,j,k} - c^*_{j,k}\,\beta_{i,j,k}\,s^*_{i,k}\,y^T_{i,k} \right\|^2 \quad \text{pour } i = 1 \text{ à } N_{pep} \text{ et } j = 1 \text{ à } N \quad (16)$$

Etape 2: estimation de la concentration des protéines cibles $c_{j,k}$ dans les N mélanges

$$\min_{c_{j,k}} \left\| M_{j,k} - \sum_{i=1}^{N_{pep}} c_{j,k}\,\beta_{i,j,k}\,s_{i,k}\,y^T_{i,k} \right\|^2 \text{ pour } j = 1 \text{ à } N \quad (17)$$

Cette étape d'optimisation aboutit à l'obtention des concentrations des protéines cibles (profils moléculaires).

4 Application

4.1 Données simulées

Nous simulons un ensemble de spectrogrammes obtenus pour différentes concentrations théoriques de la protéine cible et pour différents rapport signal à bruit (RSB). Ainsi, pour des RSB faibles, les données seront à l'image de données réelles pouvant être obtenues dans des milieux complexes (urine, sang). Nous comparons ensuite les concentrations estimées aux concentrations théoriques. Trois méthodes d'estimation sont testées: la méthode basée sur le maximum du pic, celle basée sur le volume sous le pic et enfin la méthode présentée dans ce papier. La méthode basée sur le volume réalise une estimation du volume sous le pic du peptide cible et du volume sous le pic du peptide alourdi. La concentration du peptide alourdi étant connue, la concentration du peptide cible est obtenue par une règle de proportionnalité. Le même principe est utilisé pour la méthode basée sur le maximum à la différence que l'estimation se fait à partir des maximums des pics des peptides cibles et alourdis.

4.2 Caractérisation des performances

Pour caractériser les performances de chacune des méthodes, nous nous proposons de les comparer en estimant différents paramètres statistiques.
Le premier paramètre est le coefficient de détermination R^2. En effet, pour juger de la robustesse des méthodes, nous effectuons une régression entre les concentrations estimées et les concentrations théoriques. La droite obtenue devrait idéalement avoir une pente de 1 et un coefficient de détermination R^2 le plus proche de 1, ce coefficient indiquant la dispersion des mesures.
Le deuxième paramètre est l'erreur relative absolue moyenne entre les concentrations estimées et les concentrations théoriques. Ce paramètre qui inclut toutes les sources d'erreur possible doit être le plus faible possible.
Le troisième paramètre correspond à la largeur de la bande de régression aussi appelé intervalle de confiance. Nous choisissons un intervalle de confiance de 95% correspondant au fait que l'on a une probabilité de 0.95 de contenir la valeur du paramètre que l'on chercher à estimer. Plus l'intervalle de confiance est petit et plus l'estimateur est robuste.

4.3 Résultats

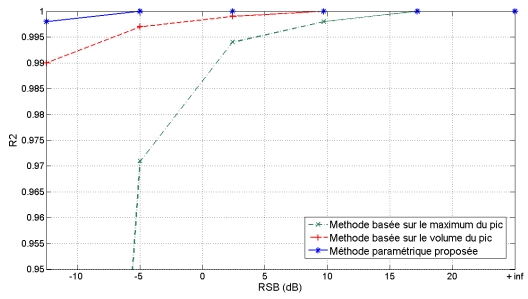

FIG. 2: *Coefficient de détermination R^2 en fonction du RSB*

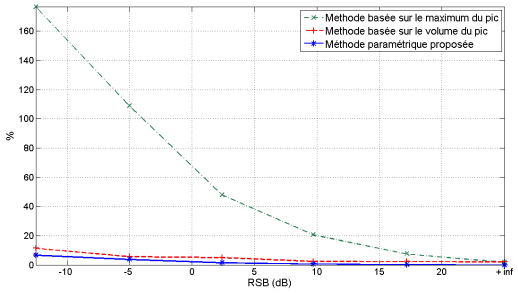

FIG. 3: *Erreur absolue relative moyenne en fonction du RSB*

Les figures 2, 3, 4 montrent la précision des méthodes à travers les trois paramètres statistiques décrits précédemment et obtenus pour différents RSB. Quand il n'y a pas de bruit, le coefficient de détermination vaut 1, l'erreur et la largeur de la bande de régression sont nulles. Dans ce cas, toutes les méthodes renvoient la bonne valeur sans aucune dispersion des résultats. Plus le bruit augmente, plus la dispersion entre les mesures augmente. On voit que la méthode estimant les concentrations à l'aide du maximum des pics est la plus impactée par le bruit. La méthode du volume sous le pic qui moyenne les valeurs est plus robuste que la méthode du maximum des pics. Enfin, notre méthode basée sur un modèle paramétrique adaptatif résiste mieux à l'augmentation du niveau de bruit, illustrant ainsi le gain en robustesse.

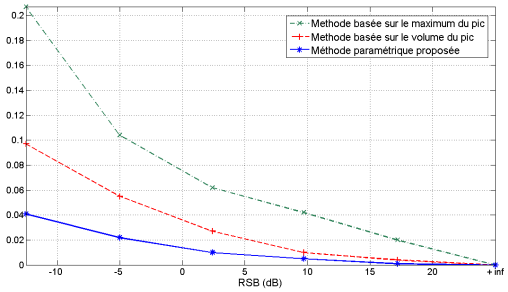

FIG. 4: *Largeur de bande de régression en fonction du RSB*

5 Conclusions

Cette communication présente une méthode de reconstruction de la concentration de protéines cibles présentes dans un milieu complexe et analysées par un système LC-MS. L'approche proposée est basée sur l'inversion d'un modèle dynamique paramétrique décrivant le comportement de la colonne de chromatographie. Cette méthode apporte une meilleure robustesse et adaptativité vis-à-vis du bruit, des fluctuations et des déformations des signaux par rapport aux méthodes classiques. Une application sur des données simulées a montré l'efficacité de l'approche présentée.

Remerciements

Nous remercions les projets européens LOCCANDIA et Technosanté CAPSI pour le financement de cette étude.

Références

[1] G. Allaire. *Numerical Analysis and Optimization: An Introduction to Mathematical Modelling and Numerical Simulation*. Oxford University Press, 2007.

[2] G. Guiochon. Preparative liquid chromatography. *Journal of Chromatography A*, 965(1-2):129–161, 2002.

[3] L. Ljung. State of the art in linear system identification: Time and frequency domain methods. In *American Control Conference*, Boston, Massachussetts, 2004.

[4] L. Ljung. *System Identification*. Prentice Hall, 2006.