Evaluation statistique d'un algorithme bayésien pour la reconstruction de profils moléculaires par spectrométrie de masse

LAURENT GERFAULT[1], GREGORY STRUBEL[1],CAROLINE PAULUS[1],
JEAN FRANCOIS GIOVANNELLI[2], PIERRE GRANGEAT[1]

[1] CEA, LETI, MINATEC, Laboratoire Electronique et Systèmes pour la Santé,
F38054 Grenoble, France
[2]Université de Bordeaux, IMS/LAPS, Equipe Signal et Image,
351 Cours de la Libération, F-33405, Talence cedex, France.

[1] laurent.gerfault@cea.fr, gregory.strubel@cea.fr, caroline.paulus@cea.fr,
[1]pierre.grangeat@cea.fr, [2]Giova@IMS-Bordeaux.fr

Résumé – Ce travail cherche à établir quelle méthodologie utiliser pour comparer des méthodes de quantification absolue de biomarqueurs par spectromètrie de masse quel que soit l'échantillon biologique traité. Le choix du critère de performance est discuté. Les critères retenus sont évalués pour démontrer le gain apporté par une méthode de quantification bayésienne.

Abstract – This work aims to build a methodology to compare absolute quantification methods of biomarkers by mass spectrometry whatever the biological sample is. Performance criteria are discussed and tested to demonstrate the benefices of a bayesian method.

1 Introduction

Le contexte de ce travail 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. Pour assurer la détection et la quantification simultanées de plusieurs protéines avec une haute sensibilité dans des concentrations de l'ordre du nanogramme par millilitre d'échantillon, la chromatographie couplée à la spectrométrie de masse est une technique attrayante. Un problème fondamental est d'améliorer la précision et la reproductibilité de la mesure. Il s'agit ainsi de maîtriser les sources d'erreur et de dispersion induites par les processus physiques et chimiques. Pour cela, nous associons une méthodologie expérimentale d'étalonnage par adjonction de marqueurs étalon dans l'échantillon à mesurer, et une méthodologie algorithmique pour maitriser les impacts des fluctuations expérimentales sur les signaux détectés et sur la mesure qui en découle.

Dès lors se pose la question de l'évaluation de notre algorithme comparativement aux algorithmes classiques. Pour cela, il nous faut définir les critères ainsi que le jeu de données test à utiliser. Dans la littérature, une telle comparaison se réalise sur quelques points de mesures. Cette approche nous semble non satisfaisante. En effet, dans le cas d'une problématique de mesure d'un biomarqueur, la variation de la concentration est inconnue et a priori importante. De plus, l'échantillon typiquement issu d'un prélèvement sanguin varie également de contenu, d'un patient à l'autre, mais également selon l'état du patient.

Le terme complexité utilisé dans le domaine biomédical permet de nommer cette variation de contenu mais ne permet pas de l'appréhender en termes de traitement du signal. Ce terme n'est pas formellement défini, mais il peut signifier le nombre de protéines ou d'autres substances présentes mais non ciblées pour la mesure. La présence de ces éléments a diverses conséquences sur le signal comme décrit en 2. Une première approche est de considérer qu'une complexité croissante provoque une augmentation du bruit de fond. Ainsi, nous nous proposons ici de présenter une première évaluation d'algorithmes de quantification absolue pour une plage de complexité d'échantillons basée sur des simulations. Nous proposons dans cette étude de définir également les critères pertinents pour mesurer l'efficacité de nos choix méthodologiques et algorithmiques. Ces critères viseront à mesurer la robustesse des algorithmes aux sources de fluctuations.

Ainsi, dans une première partie, nous décrivons le contexte de la protéomique: la démarche expérimentale et les sources de fluctuations et d'erreur, et enfin le traitement des données, notamment notre méthode statistique bayésienne. Après la définition des critères, nous comparons notre méthode aux quantifications utilisées classiquement en protéomique (estimations du maximum et du volume du premier pic du massif isotopique des peptides d'intérêt) sur des simulations.

2 Description de l'expérimentation

Les sources de fluctuations sont nombreuses dans une chaîne d'analyse protéomique. Pour assurer une quantification de bonne qualité à l'aide d'un spectromètre de masse, l'échantillon va subir plusieurs traitements avant d'être injecté dans celui-ci. Les étapes de préparation consistent en plusieurs phases visant à séparer et concentrer les protéines d'intérêt dans un volume réduit, et à conditionner celles-ci pour un passage dans la colonne de chromatographie. Pour s'adapter à la dynamiques en masse des spectomètres de

masse, les protéines sont découpées en fragments, appelés peptides. On mesure leur quantité sur le spectromètre de masse. La juxtaposition de ces étapes engendre des pertes de matière qu'il est nécessaire d'évaluer pour obtenir une bonne quantification. De plus, les conditions expérimentales vont engendrer des instabilités se traduisant notamment par des fluctuations sur la position temporelle et l'amplitude du signal mesuré.

L'ajout d'étalons internes apporte une réponse à ces fluctuations. Ils se comportent comme leurs «contreparties» naturelles. Ils sont ajoutés en concentration connue, préalablement à tout traitement de l'échantillon. La mesure de leur signal permet ainsi d'étalonner le gain global du système pour chaque peptide.

3 Estimation bayésienne des concentrations

Les signaux obtenus en sortie de la chaîne d'analyse sont des données à 2 dimensions, appelées spectrogrammes, composées d'un ensemble de pics représentatifs des peptides associés aux protéines. Une dimension de ces données est le temps de rétention dans la nano colonne de chromatographie liquide (nano-LC), l'autre dimension correspondant au rapport masse sur charge des peptides délivrés par le spectromètre de masse (MS).

Les approches classiques reposent sur l'estimation du maximum du pic ou le calcul du volume sous le pic. Ces deux méthodes peuvent être interprétées comme du filtrage adapté avec un modèle de signal correspondant respectivement à une impulsion de Dirac ou une fonction créneau entre les bornes d'intégration. Ces deux fonctions ne sont pas adaptées pour décrire la forme des pics. Nous avons donc préféré considérer des formes gaussiennes définies par leur moyenne, leur amplitude et leur inverse variance. Ce choix de formes gaussiennes a été motivé par une étude des processus de la chaîne analytique [1]. La modélisation des signaux obtenus est la suivante:

$$y = F(t).diag(\xi).D.x + F^*(t).diag(\xi).D.x^* + b(\gamma_b)$$

où y est le vecteur des mesures spectrométriques en temps et en masse, x le vecteur profil des concentrations en protéines, x^* le vecteur profil des concentrations en protéines isotopiquement alourdies, b le bruit de mesure, γ_b l'inverse variance du bruit, t le vecteur paramètres des positions temporelles des gaussiennes chromatographiques, $F(t)$ la matrice décrivant les signatures spectrométriques de chaque peptide par des fonctions gaussiennes en temps et en masse, $diag(\xi)$ la matrice des gains du système pour chaque peptide, D la matrice des rendements de digestion entre les concentrations de protéines et les concentrations de peptides.

Notre méthode bayésienne est utilisée pour estimer conjointement ces paramètres [1]. Elle exploite la présence des marqueurs étalons alourdis afin d'estimer conjointement le gain de la chaîne pour chaque protéine et chaque peptide d'intérêt. A partir des lois a priori pour les paramètres et le bruit, la formule de Bayes permet de calculer la loi a posteriori pour l'ensemble des paramètres. Nous choisissons l'estimateur de la moyenne a posteriori. Cet estimateur, activement utilisé ces dernières années en traitement du signal, possède un biais moyen nul et une erreur quadratique moyenne minimale. Nous avons retenu une mise en œuvre fondée sur les outils de l'échantillonnage stochastique de Monte-Carlo par chaîne de Markov. Nous avons conçu un générateur aléatoire simulant la loi a posteriori fondée sur un échantillonneur de Gibbs [2].

4 Evaluation des performances d'algorithmes de quantification: critères et jeu de données

4.1 jeu de données

Nous cherchons à évaluer la performance des algorithmes de quantification absolue et leur robustesse aux sources de fluctuations. Pour cela, il nous faut définir les critères ainsi que le jeu de données à utiliser. Les paramètres fondamentaux à estimer pour la validation bio-analytique sont: la précision, la sélectivité, la sensibilité, la reproductibilité et la stabilité [3]. Le test expérimental habituel de la qualité de la quantification absolue est de réaliser un test de linéarité. Des peptides sont ajoutés en quantité croissante et connue dans un échantillon de base que l'on appellera matrice. Sa complexité, à savoir la quantité de signaux parasites venant de protéines présentes dans l'échantillon mais dont on ne souhaite pas mesurer la concentration, est stable. Ensuite, pour des concentrations croissantes de peptides ciblés, on crée des réplicats d'échantillons, puis également des réplicats de mesures. Les concentrations connues sont alors comparées aux concentrations estimées. Ce jeu de données est donc pertinent pour estimer la performance de la mesure d'une concentration inconnue pour un échantillon ayant un contenu moléculaire donné.

En modifiant le contenu moléculaire de la matrice, nous pouvons évaluer la robustesse de la quantification par rapport à la variabilité inter-patient ou au changement de la provenance de l'échantillon c'est à dire issu par exemple d'un prélèvement urinaire ou sanguin.

4.2 critères

Nous cherchons à trouver un critère global permettant d'estimer la performance des algorithmes de quantification sur ce groupe d'expériences. Dans un cas réel, la concentration recherchée étant inconnue, chacun des critères sera défini pour une gamme de concentration et pour un niveau de bruit donné.

Les critères classiques sont l'erreur absolue moyenne ou le coefficient de variation. On étendra ce dernier critère à une gamme de concentration en réalisant la moyenne des valeurs obtenues sur la gamme.

Ainsi, l'erreur absolue moyenne et le coefficient de variation s'écrivent respectivement :

$$EAM = \sum_{i=1}^{Nc} \sum_{j=1}^{Nr} \left| \hat{y}_{ij} - y_i \right| / y_i$$

$$CV = \frac{1}{Nc} \sum_{i=1}^{Nc} \frac{Nr}{Nr-1} \sum_{j=1}^{Nr} \frac{\left(\hat{y}_{ij} - \bar{y}_i\right)^2}{\bar{y}_i} .$$

avec Nr le nombre de réplicats pour une concentration, Nc le nombre de concentrations par gamme de mesure, y_i exprime la valeur théorique, $\hat{y}_i$ la valeur observée, $f(y_i)$ la valeur régressée et $\bar{y} = \sum_{j=1}^{Nr} \hat{y}_{ij}$ la moyenne des observations.

On réalise également une analyse par régression linéaire de la courbe concentration estimée par rapport à la concentration vraie. On considère alors le coefficient de détermination R^2. Il représente la proportion de la variabilité expliquée par le modèle de régression par rapport à la variabilité totale.

Il est calculé comme R^2 = 1- SCR/SCT où:

$SCT = \sum_{i=1}^{Nc} \sum_{j=1}^{Nr} \left(\hat{y}_{ij} - \bar{y}\right)^2$ est la somme des carrés des écarts des observations à leur moyenne, $SCR = \sum_{i=1}^{Nc} \sum_{j=1}^{Nr} \left(\hat{y}_{ij} - f(y_i)\right)^2$ la somme des carrés des résidus [4].

5 Comparaison des méthodes

5.1 réalisation du jeu de données

Nous étudions la robustesse des méthodes par rapport à un bruit additif. Des signaux sont simulés, à partir d'un modèle de pic gaussien dans les 2 dimensions, pour représenter le signal d'un peptide. Le signal de la solution tampon est simulé par l'ajout d'un bruit de moyenne nulle avec 6 niveaux de bruit différents (écart type de 0 à 0,3) mimant l'augmentation de la complexité de la solution. Pour un échantillon tampon (un niveau de bruit), 5 concentrations de peptide non marqué (0,6; 1,2; 2,3; 4,6; 9,2) et une concentration de peptide marqué (5) sont considérées. Pour une concentration et un niveau de bruit, nous simulons 9 mesures différentes correspondant à une répétition de l'expérience.

Par la simulation, nous assurons que la concentration dans l'échantillon est exacte. Ainsi, il n'existe pas d'incertitude sur la valeur vraie, ce que l'on ne peut assurer sur un échantillon réel. Par conséquent, toutes les sources de variabilités des estimations proviennent du signal. Nous éliminons ainsi la variabilité liée à la préparation des échantillons.

Pour chaque série d'expériences, c'est-à-dire pour l'ensemble des réplicats sur toute la gamme de concentration, et pour chaque niveau de bruit, nous estimons les critères définis précédemment. Position du maximum du pic et zone d'intégration sont définies à partir de la position et de la dimension exactes des pics non bruités. A contrario, l'algorithme bayésien n'est pas initialisé à partir des valeurs vraies des paramètres. Pour introduire la variabilité de l'estimation due aux différentes initialisations possibles de l'algorithme bayésien, nous réalisons 2 estimations par échantillon.

Les courbes suivantes sont représentées en fonction du rapport bruit sur signal. Ce rapport bruit sur signal est celui considéré par les bio-analystes, à savoir le rapport entre l'amplitude du bruit de fond par l'amplitude maximale du signal du peptide. Nous prendrons pour valeur de l'amplitude du bruit de fond, l'écart type du bruit de fond simulé. L'amplitude du peptide considéré est celle du peptide étalon.

Chaque point de mesure sur les courbes suivantes représente une gamme de concentration pour un niveau de bruit de fond. Ainsi, on retiendra que l'estimation est réalisée sur 5 concentrations pour lesquelles le rapport bruit sur signal varie entre 8 fois et 0,5 fois la valeur reportée sur l'axe des abscisses.

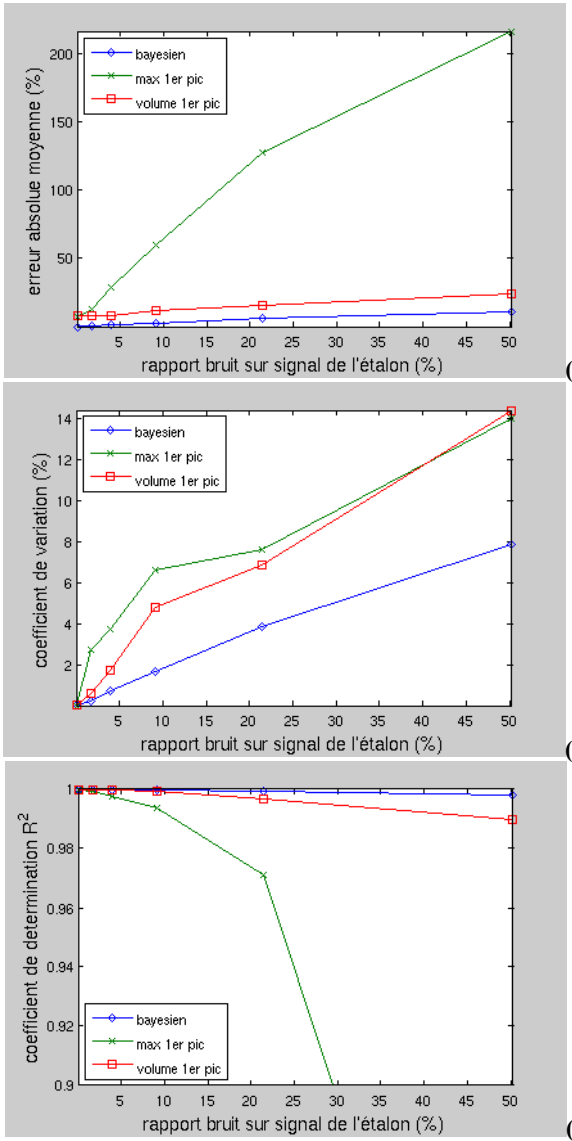

Figure 1: comparaison des différents critères pour les méthodes de mesure utilisées
(a) erreur absolue moyenne, (b) coefficient de variation moyen, (c) coefficient de détermination

5.2 Discussion sur les critères

Le bruit augmentant, la dispersion entre les mesures augmente ce que l'on constate avec tous les critères évalués.
On observe une convergence du coefficient de variation pour le maximum et le volume du pic. Pourtant, au regard des autres critères, il n'y a pas convergence des performances de ces deux méthodes. De ce fait, ce critère ne peut être retenu. Cette convergence s'explique par la prédominance du bruit sur le signal. Par l'utilisation d'une règle de proportionnalité pour calculer la quantité absolue à partir du signal étalon, la quantification en présence de fort bruit tend vers la valeur de la concentration de l'étalon.

Le critère basé sur l'erreur est plus discriminant pour distinguer les algorithme bayésien et volume du pic que le coefficient de détermination. Cependant, pour le coefficient de détermination, la différence avec la méthode du volume est réduite de par la nature du bruit et du fait que cette estimation est réalisée de manière optimale. En effet, la zone d'étude utilisée est déterminée à partir des positions et écart type exacts de la simulation. Une détermination manuelle de cette zone aurait engendré des fluctuations bien plus importantes, en particulier pour les plus fortes valeurs du bruit. De plus, compte tenu de l'utilisation d'un jeu de données issu d'un test de linéarité, le coefficient de détermination permet de tester conjointement la linéarité et les dispersions. Enfin, l'usage de données simulées permet d'assurer qu'il n'existe aucune incertitude sur les valeurs de l'axe des abscisses pour la courbe estimation en fonction de la concentration vraie. Ainsi, par construction, la pente de la droite de régression est proche de l'unité, et les résidus de la régression mesurent uniquement les variabilités introduites par l'estimateur sur toute la gamme de mesure.

5.3 résultat de la comparaison des méthodes de quantification

Comme attendu, l'ajout d'un bruit gaussien de moyenne nulle affecte beaucoup le maximum de la courbe et peu les autres méthodes. Cependant, aux fortes valeurs de bruit, on observe une meilleure performance de la méthode bayésienne qui est celle qui utilise le maximum d'information et réalise son calcul sur la totalité du signal peptidique.

Pour tous les critères calculés, l'algorithme bayésien permet d'obtenir les meilleurs résultats. Notre méthode résiste le mieux aux cas où le rapport bruit sur signal augmente, c'est-à-dire en limite de détection. De plus, nous constatons que les performances obtenues sont bonnes au-delà de la limite de quantification par défaut utilisée par les bio-analystes qui est fixée à 10 fois le niveau de bruit. .

6 Conclusion

Les variabilités biologique et technologique étant importantes , les signaux étudiés en protéomique sont très divers. Comment assurer dés lors que, quel que soit l'échantillon biologique, notre algorithme de quantification bayésien est le plus performant ? Pour répondre à cette question de la comparaison de méthodes de reconstruction de profils moléculaires, nous avons discuté comment mesurer la performance des algorithmes de quantification en évaluant plusieurs critères et présenté quel jeu de données utilisé.

Par la simulation de tests de linéarité, nous pouvons évaluer dans toutes les conditions réelles la performance des algorithmes de quantification. Un critère de performance adapté à ce jeu de données a été proposé: le coefficient de détermination R^2. Il permet de mesurer globalement la capacité des méthodes à maîtriser les conséquences des sources de fluctuations sur les signaux. Dans le cas présenté, nous nous sommes intéressés à la robustesse vis-à-vis d'un bruit additif de loi normale. Nous avons démontré que notre méthode d'estimation bayésienne permet d'obtenir les meilleurs résultats.

Cette étude est une première étape pour une évaluation exhaustive de l'impact des autres sources d'erreur sur les méthodes de quantification comme la présence de contaminants ajoutant des pics parasites au signal.

Références

[1] G. Strubel, "Reconstruction de profils moléculaires: modélisation et inversion d'une chaîne de mesure protéomique," Institut polytechnique de Grenoble, 2008.

[2] G. Strubel, J.F. Giovannelli, C. Paulus, L. Gerfault, et P. Grangeat, *Reconstruction bayésienne de profils moléculaires*, Troyes France: 2007.

[3] I. Taverniers, M. De Loose, et E. Van Bockstaele, "Trends in quality in the analytical laboratory. II. Analytical method validation and quality assurance," *TrAC Trends in Analytical Chemistry*, vol. 23, Sep. 2004, pp. 535-552.

[4] BERTRAND, Dominique , "régression linéaire simple," *Étalonnage multidimensionnel : application aux données spectrales*, E.T.I, 2005.