Présentation d'un algorithme d'inversion hiérarchique bayésien pour la quantification et la classification de données protéomiques

Laurent Gerfault[1], Pascal Szacherski[1,2], Jean François Giovannelli[2], Jean Philippe Charrier[3], Pierre Mahé[4], Bruno Lacroix[3] and Pierre Grangeat[1]
[1]CEA-LETI, MINATEC CAMPUS, DTBS, 17 rue des martyrs, 38054, Grenoble cedex, France
[2] Université Bordeaux, IMS, UMR 5218, Talence, France; [3] bioMérieux, 69280 Marcy l'Etoile, France, [4]bioMérieux, Parc Polytec, 5 rue des Berges, 38000 Grenoble, France

laurent.gerfault@cea.fr, pascal.szacherski@cea.fr, Jean-Francois.Giovannelli@ims-bordeaux.fr, jean-philippe.charrier@eu.biomerieux.com, pierre.mahe@eu.biomerieux.com, bruno.lacroix@eu.biomerieux.com,pierre.grangeat@cea.fr

Abstract

Maîtriser la variabilité technologique sur les chaînes d'analyse protéomique par spectrométrie de masse est un point critique. Nous proposons de contrôler la variabilité technologique grâce à une modélisation paramétrique de la chaîne d'acquisition et à un algorithme d'inversion Bayésien adaptatif innovant pour estimer la concentration des protéines et le statut clinique de l'échantillon. Nous présentons le principe de cette approche et les résultats de quantification et de classification que nous avons obtenus sur une cohorte de patients dans le cadre d'une étude liée au cancer colorectal.

1 Introduction

Maîtriser la variabilité technologique sur les chaînes d'analyse protéomique par spectrométrie de masse est un point critique. Dans le cas des spectromètres de masse SRM, les performances en sensibilité atteintes permettent d'envisager des applications cliniques. Cependant, le CPTAC a démontré que les variabilités d'estimation étaient une contrainte importante pour envisager le déploiement de cette technique et une utilisation en routine. Pour minimiser ces variabilités, des réponses sont apportées par la standardisation des protocoles de préparation des échantillons. Cependant, ces efforts n'ont pas d'effet sur les variabilités instrumentales antérieures à l'injection de l'échantillon dans la chaîne analytique. Ces variabilités sont liées à chaque étape de la chaîne d'analyse à savoir le passage de protéine à peptide, peptide à ion précurseur et d'ion précurseur à fragment. Par ailleurs, la présence de contaminants est envisageable malgré la sélectivité améliorée de ce type d'instrument.

Par ailleurs, une utilisation en clinique nécessite l'usage d'algorithme automatique pour gérer le grand nombre de données générées, améliorer la reproductibilité de l'estimation à partir des données, et aussi délivrer des informations concernant l'incertitude des mesures. Dans le cadre du projet « BHI-PRO », nous proposons d'estimer la concentration des protéines et le statut clinique de l'échantillon, et de contrôler la variabilité technologique grâce d'une part à une modélisation hiérarchique

paramétrique de la chaîne d'acquisition et d'autre part à un algorithme d'inversion hiérarchique Bayésien adaptatif innovant.

2 Méthode

Dans cette approche, nous décrivons la chaîne d'acquisition analytique selon un modèle hiérarchique qui suit la cascade d'étapes appliquées aux protéines et à leurs sous-produits et nous introduisons les paramètres technologiques afférents aux molécules et au système d'acquisition. Typiquement, pour une protéine, 3 peptides sont ciblés, et 3 transitions par peptide sont mesurées. Ainsi, notre algorithme doit permettre de délivrer une estimation des protéines ciblées à partir de 9 transitions.

Par inversion de ce modèle, en utilisant des techniques d'estimation bayésienne de type moyenne de la distribution a posteriori, nous réalisons l'estimation conjointe des concentrations d'intérêt et des paramètres technologiques en combinant toutes les informations issues des protéines d'intérêt. Cet algorithme permet d'intégrer la variabilité biologique sur les concentrations de protéine et la variabilité instrumentale sur les paramètres technologiques présents dans le modèle. D'autre part, l'analyse bayésienne permet de fournir, en plus de l'estimation des paramètres, l'incertitude de cette estimation. Contrairement aux autres algorithmes de type N-PLS (N-way Partial Least Square) qui procèdent par apprentissage pour estimer la signature d'une protéine, l'estimation de la signature n'utilise que l'acquisition en cours.

Par ailleurs, dans le cadre de la thèse de P. Szacherski, cette approche a été étendue à l'estimation conjointe des concentrations des protéines et de la classe d'appartenance des échantillons. Ceci ouvre notamment le champ d'application aux études différentielles en sciences de la vie et à l'aide au diagnostic en médecine.

3 Résultats

Dans cette communication, nous considérons des analyses sur des échantillons sanguins avec standard AQUA sur une cohorte de 237 patients. Cependant, cette technique peut aussi être appliquée sur des mesures utilisant des standard PSAQ pour des données Full MS et SRM.

Les tests ELISA font partie des tests de référence utilisés en diagnostic clinique pour quantifier la concentration d'une protéine. Les tests SRM permettraient de réaliser ces études plus rapidement, à moindre coût, sans nécessité de développer des anticorps. Nous montrons que les performances de notre estimation de la concentration de la protéine LFABP en termes de corrélation avec un test ELISA sont similaires à celles existant entre 2 tests ELISA commerciaux. De plus, avec l'approche BHI-PRO, nous disposons d'une procédure automatique pour traiter les données SRM et d'une estimation de l'incertitude sur la valeur quantifiée.

La diminution de la variabilité instrumentale dans l'incertitude de l'estimation permet aussi de réduire les risques d'erreur à l'étape suivante de classification des patients. Cette question est au centre du projet ANR BHI-PRO. Nous avons notamment proposé de fusionner ces deux approches de quantification et de classification en une seule opération. Nous montrons que cette approche conjointe permet de minimiser le risque moyen de mauvaise classification.