Imputation multiple et prise en compte de l'incertitude pour les données de protéomique quantitative

Marie Chion [1], Frédéric Bertrand [2] & Christine Carapito [3]

[1] *Institut de Recherche Mathématique Avancée, UMR 7501, 7 rue René Descartes, 67084 Strasbourg Cedex et Laboratoire de Spectrométrie de Masse Bio-Organique, IPHC, UMR 7178, 25 rue Becquerel, 67087 Strasbourg Cedex, marie.chion@unistra.fr*

[2] *Institut de Recherche Mathématique Avancée, UMR 7501, 7 rue René Descartes, 67084 Strasbourg Cedex, frederic.bertrand@math.unistra.fr*

[3] *Laboratoire de Spectrométrie de Masse Bio-Organique, IPHC, UMR 7178, 25 rue Becquerel, 67087 Strasbourg Cedex, ccarapito@unistra.fr*

Résumé. L'analyse protéomique consiste à étudier l'ensemble des protéines contenues dans un système biologique donné, à un instant donné et dans des conditions données. Les techniques les plus performantes pour déterminer l'abondance des protéines passent par la mesure des intensités peptidiques. Mais ces données peptidiques comportent des valeurs manquantes. Bien que les techniques statistiques usuelles en protéomique permettent l'imputation de celles-ci, l'incertitude liée à l'imputation n'est pas prise en compte. Nous proposons ici de combiner les techniques d'estimation tempérée de la variance aux méthodes d'imputation multiple, en les rendant utilisables tant au niveau peptidique qu'au niveau protéique.

Mots-clés. Valeurs manquantes, Imputation multiple, tests t-tempérés, Protéomique.

Abstract. Proteomic analysis consists in studying proteins from a given biological system, at a given time and under given conditions. The most efficient techniques for determining protein abundance rely on measuring peptides intensities. These peptide data include missing values. Although usual statistical techniques for proteomics allow missing values imputation, they do not take into account the uncertainty caused by the imputation itself. We propose here to combine moderated t-tests techniques with multiple imputation methods and make them applicable to either peptide or protein data.

Keywords. Missing values, Multiple imputation, Moderated t-tests, Proteomics.

1 Contexte

En analyse protéomique quantitative, il s'agit d'identifier et quantifier l'ensemble des protéines exprimées par une cellule, un tissu, un organe ou un organisme à un moment donné et sous des conditions données. Les progrès technologiques de ces dernières années

permettent aujourd'hui d'identifier et quantifier plusieurs milliers de protéines en quelques heures d'analyse LC-MS (spectromètre de masse couplé à une chromatographie liquide). Cependant, dans les jeux de données de protéomique quantitative, il est fréquent d'avoir des valeurs manquantes. Celles-ci, qu'elles soient décrites par des 0 (Zéro), des NA (Not Available) ou des NULL, peuvent être causées par deux possibilités :

— soit l'intensité mesurée pour la protéine concernée est en-dessous de la limite de quantification ;
— soit la protéine concernée est réellement absente de l'échantillon étudié.

Les valeurs manquantes (Molenberghs et al. 2015) peuvent donc être déterminantes lors d'analyses différentielles, dans lesquelles il s'agit d'identifier les protéines différentiellement exprimées entre deux conditions, par exemple entre des échantillons d'individus sains et des échantillons d'individus malades.

2 Problème

Aujourd'hui, les méthodes d'inférence statistique couramment utilisées en protéomique quantitative sont basées sur la mesure des intensités (extraction de courants d'ions) des peptides. Elles permettent d'en déduire les abondances des protéines à condition de disposer de suffisamment de peptides par protéine. Elles ne prennent pas en compte de manière satisfaisante les peptides ou les protéines dont les intensités sont manquantes dans certaines conditions alors que ceux-ci sont particulièrement intéressants d'un point de vue biologique ou médical, puisque que potentiellement explicatifs d'une différence entre les groupes comparés.

Plusieurs méthodologies, accessibles au travers de logiciels de traitement statistique des données de protéomique, comme celle de ProStaR de Wieczorek et al. 2017, proposent d'imputer ces valeurs manquantes, alors que les autres retirent simplement les protéines pour lesquelles il y a trop de peptides manquants. Lorsque des méthodes d'imputation sont utilisées, notamment d'imputation multiple (Little et Rubin, 2002), le traitement statistique n'est pas complètement satisfaisant. En effet, même si ces outils statistiques sont pertinents dans ce contexte, les jeux de données une fois imputés sont considérés comme ayant toujours été complets dans la suite des analyses : il n'est pas du tout tenu compte de l'incertitude liée à l'imputation.

Ces analyses se terminent généralement par une étude des différences d'abondances des protéines entre les différentes conditions soit à l'aide de test de Student ou de Welch pour les approches les plus rudimentaires, soit à l'aide des techniques de test t-tempérés exposés dans Phipson et al. 2016 qui reposent sur des approches bayésiennes empiriques et qui sont implémentées dans le package Bioconductor `limma`.

3 Amélioration proposée

Des travaux récents ont cherché à améliorer l'inférence des jeux de données protéiques et en particulier l'analyse des abondances différentielles. Plusieurs méthodologies ont été proposées et comparées dans Geominne et al. 2015. Ainsi, des approches *ridge* et robuste associées à des modèles mixtes ont été introduites pour tenir compte de la spécificité des données peptidiques, (Goeminne et al. 2018).

Il ressort de ces recherches que les modèles statistiques se basant sur les données peptidiques sont les plus performants. Ainsi, nous proposons une nouvelle méthodologie qui débute par l'imputation des valeurs manquantes au niveau peptidique, en suivant les recommandations de White et al. 2011, et l'estimation de l'incertitude liée à cette imputation et se prolonge naturellement par l'intégration de cette incertitude aux techniques actuelles d'estimation tempérée de la variance proposées par Phipson et al. 2016 ou utilisées par Goeminne et al. 2018.

4 Résultats

Cette nouvelle méthodologie sera évaluée à l'aide de simulations et à partir d'un jeu de données réelles.

Bibliographie

Geominne, L. J. E., Argentini, A., Martens, L. & Clement, L. (2015). Summarization vs Peptide-Based Models in Label-Free Quantitative Proteomics : Performance, Pitfalls, and Data Analysis Guidelines. *Journal of Proteome Research*, 14, 2457-2465.

Goeminne, L. J. E., Gevaert, K. and Clement, L. (2018). Experimental Design and Data-analysis in Label-free Quantitative LC/MS Proteomics : a Tutorial with MSqRob. *Journal of Proteomics*, 171, 23-36.

Little, R. J. A. & Rubin, D. B. (2002). *Statistical Analysis with missing data*, John Wiley & Sons.

Molenberghs, G., Fitzmauice, G., Kenward, M. G., Tsiatis, A. & Verbeke, G. (2015). *Handbook of Missing Data Methodology*, CRC Press.

Phipson, B., Lee, S., Majewski, I. J., Alexander, W. S. & Smyth, G. K. (2016). Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. *Annals of Applied Statistics*, 10(2), 946-963.

White, I. R., Royston, P. & Wood, A. M. (2011). Multiple imputation using chained equations : Issues and guidance for practice. *Statistics in Medicine*, 30, 377-399.

Wieczorek, S., Combes, F., Lazar, C., Giai-Gianetto, G., Gatto, L., Dorffer, A., Hesse, A.-M., Couté, Y., Ferro, M., Bruley, C., & Burger, T. (2017). DAPAR & ProStaR : software to perform statistical analyses in quantitative discovery proteomics. *Bioinformatics*, 33(1), 135-136.