ProteoBayes : un cadre bayésien pour l'analyse protéomique différentielle

Marie Chion* Arthur Leroy†

Résumé

Les méthodes statistiques actuelles dans l'analyse protéomique différentielle laissent généralement de côté plusieurs défis, tels que les valeurs manquantes, les corrélations entre les intensités des peptides et la quantification de l'incertitude. En outre, elles fournissent des estimations ponctuelles, telles que l'intensité moyenne pour un peptide ou une protéine donné(e) dans une condition donnée. La décision de considérer ou non un analyte comme différentiel est alors basée sur la comparaison d'une p-valeur avec un seuil de significativité. Nous présentons ici le package R ProteoBayes, disponible sur le CRAN. Il implémente un cadre bayésien pour l'analyse protéomique différentielle, permettant ainsi d'estimer explicitement la taille de l'effet et de quantifier l'incertitude pour une différence de moyennes entre deux conditions biologiques comparées. Une application Shiny a également été mise en place pour les utilisateurs ne codant pas en R.
Mots-clefs : Biostatistique – Statistique Bayésienne – Package – Shiny

Développement

L'analyse protéomique différentielle consiste à mesurer des intensités de peptides (usuellement par spectrométrie de masse), puis à comparer leurs moyennes par condition pour enfin identifier celles qui sont différentiellement exprimées. Celles-ci sont alors considérées comme potentiels biomarqueurs de pathologie. Les méthodes statistiques couramment utilisées ignorent généralement plusieurs problématiques spécifiques à ces données, telles que les valeurs manquantes, les corrélations entre les intensités des peptides et la quantification de l'incertitude. En outre, elles fournissent des estimations ponctuelles, telles que l'intensité moyenne pour une protéine donnée dans une condition donnée. La décision de considérer ou non une protéine comme "différentielle" est alors basée sur la comparaison de la p-valeur avec un seuil de significativité, généralement de 5 %. Dans l'approche du test t-modéré de Smyth [2004], ainsi que dans son extension dans le cadre d'imputation multiple [Chion et al., 2022], un modèle hiérarchique bayésien est utilisé pour déduire la distribution a posteriori de l'estimateur de la variance pour chaque peptide. L'espérance de cette distribution est ensuite utilisée comme une estimation modérée de la variance et est injectée directement dans l'expression de la statistique de test. Cette méthode permet de prendre en compte la structure de variabilité particulière des données de protéomique et plus largement des données d'expression de gènes. Cependant, en considérant des distributions plutôt que des estimateurs ponctuels de position et de dispersion, la statistique bayésienne permet de quantifier l'incertitude. Le travail présenté dans Chion and Leroy [2023] suit une idée similaire en tirant parti des résultats standard de l'inférence bayésienne avec des distributions *a priori* conjuguées dans les modèles hiérarchiques pour développer une méthodologie adaptée au traitement des contextes d'imputation multiple.

Formellement, si l'on cherche à comparer l'intensité moyenne d'un peptide $p = 1, \dots, P$ entre différents groupes $k = 1, \dots, K$, pour lesquels nous avons observé $n = 1, \dots, N_k$ échantillons, l'inférence porte alors sur la quantité $\mu_k^p - \mu_{k'}^p$ (la différence des moyennes de groupes). Si l'on note $y_{k,n}^p$ l'observation

*MRC Biostatistics Unit, University of Cambridge, marie.chion@mrc-bsu.cam.ac.uk
†Department of Computer Science, The University of Manchester, arthur.leroy.pro@gmail.com

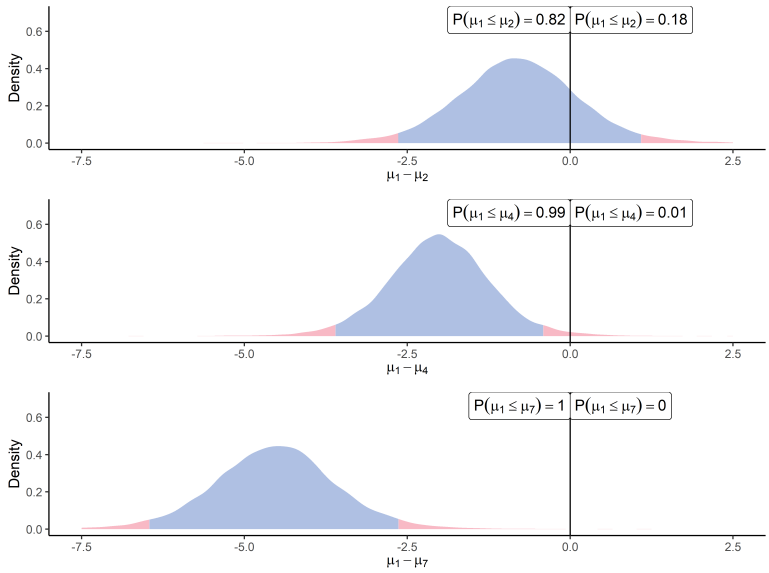

FIGURE 1 – Illustration de la loi a posteriori des moyennes de groupes pour un même peptide à travers 3 comparaisons distinctes (groupe 1 vs 2, 1 vs 4, 1 vs 7) présentant un écart croissant. La probabilité que la moyenne d'un groupe soit plus grande que l'autre est indiquée de part et d'autre de la droite d'abscisse 0. L'intervalle de crédibilité à 95% est représenté par l'aire en bleue sur la densité.

n, du groupe k, pour le peptide p, le modèle génératif est défini comme :

$$y_{k,n}^p = \mu_k^p + \varepsilon_n, \quad \forall p = 1, \dots, P, \ \forall k = 1, \dots, K, \ \forall n = 1, \dots, N_k,$$

avec la vraisemblance et les lois à priori suivantes : $\varepsilon \sim \mathcal{N}(0, \sigma_k^{p2})$, $\mu_k^p \mid \sigma_k^{p2} \sim \mathcal{N}\left(\mu_0, \frac{1}{\lambda_0}\sigma_k^{p2}\right)$, $\sigma_k^{p2} \sim \Gamma^{-1}(\alpha_0, \beta_0)$, où $\{\mu_0, \lambda_0, \alpha_0, \beta_0\}$ sont les hyper-paramètres associés. Ces hypothèses décrivent une loi a priori Gaussienne-inverse-gamma pour les paramètres de moyenne μ_k^p et de variance σ_k^{p2}, qui est conjuguée à la vraisemblance Gaussienne. Puisque l'objet d'intérêt pour l'inférence est le paramètre de moyenne μ_k^p pour chaque groupe, la loi a posteriori marginale $p(\mu_k^p \mid \mathbf{y}_k^p)$ résulte en une t-distribution généralisée dont l'expression analytique est connue. En échantillonnant pour chaque groupe à partir de ces distributions, il est ainsi possible de calculer la loi a posteriori de notre quantité d'intérêt $\mu_k^p - \mu_{k'}^p$, qui peut être visualisée comme sur la Figure 1, où 3 exemples de différence (plus ou moins importante) entre groupes sont représentés. Cette loi offre un bien plus large panel d'informations qu'un résultat de t-test pour conduire l'inférence. En effet, nous estimons ici explicitement la taille d'effet, ainsi que l'incertitude associée, de la différence entre groupes. Il est alors trivial de définir une procédure de décision probabiliste, se basant sur la probabilité que cette différence soit suffisamment distincte de 0.

Cette approche globale a été implémentée et rendue librement disponible à travers un package R, ProteoBayes, disponible sur le CRAN, et une application web offrant une interface graphique pour les praticiens ne codant pas en R.

Références

Marie Chion and Arthur Leroy. A Bayesian Framework for Multivariate Differential Analysis accounting for Missing Data, July 2023.

Marie Chion, Christine Carapito, and Frédéric Bertrand. Accounting for multiple imputation-induced variability for differential analysis in mass spectrometry-based label-free quantitative proteomics. *PLOS Computational Biology*, 18(8) :e1010420, August 2022. ISSN 1553-7358. doi : 10.1371/journal.pcbi.1010420.

Gordon K Smyth. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1) :1–25, January 2004. ISSN 1544-6115. doi : 10.2202/1544-6115.1027.