SUJET DE POST-DOCTORAT

Découverte de biomarqueur en protéomique :
Stratégie bayésienne, sélection de modèle, échantillonnage stochastique

Ces dernières années, on assiste à l'apparition d'une nouvelle génération de méthodes de recherche clinique, de diagnostic / pronostic, ainsi que de suivi thérapeutique reposant sur des *analyses protéomiques*. Leur objectif est de caractériser certaines pathologies au travers de signatures sur des profils de protéines issus de prélèvements comme une prise de sang ou une biopsie. Sur un plan physiologique, on est confronté à de très faibles concentrations dans des milieux particulièrement complexes en plus d'une forte variabilité entre individus. Dans ce contexte, nous nous intéresserons à la phase préliminaire de *découverte et sélection de biomarqueur*, cruciale mais très délicate.

Pour la réaliser, les instruments d'analyse MALDI (Matrix-Assisted Laser Desorption/Ionisation) sont particulièrement adaptés en raison de leur sensibilité et de leur débit d'analyse. Ils produisent des *spectres* présentant, idéalement, un pic ou un réseau de pics associé à chaque protéine. En pratique, plusieurs effets instrumentaux entrent en jeu : élargissement et déformation des pics conduisant à de possibles recouvrements ou masquages, présence de bruit de mesure et d'une ligne de base,... A ces effets, s'ajoutent des difficultés liées à des incertitudes et variabilités sur certains paramètres instruments (largeurs de pics, niveaux de bruit, gains,...).

Grace à ces instruments, on dispose des spectres relatifs à des prélèvements provenant de deux cohortes dont le statut clinique est identifié (sain et malade). La modélisation sera fondée sur un mélange de lois multidimensionnelles et les travaux seront intimement liés à deux questions classiques : (1) l'*apprentissage* des paramètres des lois pour chaque classe, connaissant le statut clinique et (2) la *classification* des individus connaissant ces paramètres. Cela dit, le cœur des travaux concernera la réflexion sur la découverte et la sélection de biomarqueurs et il s'agit d'un problème beaucoup plus délicat. Des solutions existent mais la question est encore largement ouverte.

La stratégie proposée repose sur des modèles hiérarchiques et des approches statistiques bayésiennes. Une des difficultés est liée à la nécessité d'estimer conjointement les paramètres instrument en même temps que de traiter les questions de la découverte et de la sélection. L'exploration des lois a posteriori sera réalisée grâce à des algorithmes d'échantillonnage (Monte Carlo par Chaînes de Markov, Gibbs, Metropolis-Hastings,...) et / ou des approximations variationnelles. On aura ainsi accès à des moyennes et écarts-types (fournissant des estimées et des marges d'erreurs), à des corrélations (quantifiant des liens entre paramètres) et plus généralement à des marginales (permettant l'évaluation de probabilités d'hypothèses). Au final, nous visons la gestion des variabilités biologiques et technologiques ainsi que leur impact sur les propriétés des tests d'hypothèse (puissance, probabilité d'erreurs,...). Il s'agit d'un point critique pour le développement de ces chaînes d'analyse.

Ces recherches se déroulent dans le cadre du projet BHI-PRO (Bayesian Hierarchical Inversion in Proteomics) soutenu par l'ANR, pour 2011-2013 et qui implique cinq partenaires : le CEA (LETI et LIST), l'IMS, la plateforme de protéomique CLIPP, le LBS et la Société bioMérieux. Les recherches se dérouleront au sein du Groupe Signal – Image de l'IMS à Bordeaux et s'intègreront dans les activités concernant les problèmes d'estimation, de détection, de test d'hypothèse et de sélection de modèle ainsi que les problèmes inverse, dans un contexte myope.

Mots clés : problèmes inverses, modèles hiérarchiques, approches bayésiennes, aspects myopes (semi-aveugle), mélanges de densités, apprentissage et classification, sélection (de modèle, d'ordre, de variables), facteur de Bayes, échantillonnage stochastique.

Applications visées : protéomique, imagerie biologique et médicale, diagnostic / pronostic.

Langage : Matlab ou R sur PC.

Contexte et financement : Projet de recherche BHI-PRO, financement ANR (programme blanc).

Durée – Salaire : 12 mois modulable, environ 2000 Euros nets par mois.

Lieu des travaux : Groupe Signal – Image, Laboratoire de l'Intégration du Matériau au Système, (Université de Bordeaux – CNRS). Déplacements à prévoir (Dijon, Grenoble, Lyon, Saclay).

Contact : Jean – François GIOVANNELLI, Mél : `Giova@IMS-Bordeaux.fr`, Tél : 05 4000 31 76.

Encadrement : Jean – François GIOVANNELLI, Audrey GIREMUS.

Références

[1] A. Gelman, J. C. Carlin, H. S. Stern et D. B. Rubin, *Bayesian Data Analysis*, Chapman & Hall/CRC, Boca Raton, USA, 2nd edition, 2004.

[2] J.-M. Marin et C. Robert, *Bayesian Core. A Practical Approach to Computational Bayesian Statistics*, Texts in statistics. Springer, Paris, France, 2007.

[3] C. P. Robert, *The Bayesian Choice. From decision-theoretic foundations to computational implementation*, Springer Texts in Statistics. Springer Verlag, New York, NY, USA, 2007.

[4] C. P. Robert et G. Casella, *Monte-Carlo Statistical Methods*, Springer Texts in Statistics. Springer, New York, NY, USA, 2000.

[5] K.-A. Do, P. Muller et M. Vannucci, *Bayesian Inference for Gene Expression And Proteomics*, Cambridge University Press, Cambridge, Angleterre, 2006.

[6] G. Strubel, J.-F. Giovannelli, L. Gerfault, P. Szacherski, C. Paulus et P. Grangeat, « Bayesian protein quantification and instrument parameter calibration for LC-MS », *Soumis*, septembre 2011.

[7] P. Szacherski, J.-F. Giovannelli, L. Gerfault et P. Grangeat, « Apprentissage supervisé robuste de caractéristiques de classes. Application en protéomique. », in *Actes du 23e colloque GRETSI*, Bordeaux, France, septembre 2011.

[8] P. Szacherski, J.-F. Giovannelli et P. Grangeat, « Joint Bayesian hierarchical inversion-classification and application in proteomics. », in *Proceedings of the International Conference on Statistical Signal Processing*, Nice, France, juin 2011.