Projet de thèse : « Intégration de données métabolomiques par quantification de données haut débit et application à l'analyse des causes de la mortalité périnatale chez le porc »

Résumé de la thèse

L'obtention de données métabolomiques est une méthode abordable pour accéder à des informations phénotypiques fines à partir d'échantillons sanguins, de liquide amniotique ou d'urine en utilisant une approche haut-débit, la RMN (Résonnance Magnétique Nucléaire). Cependant, à cause d'un manque de méthodes automatiques permettant de relier l'information fournie par le spectre RMN à la quantification des métabolites, son utilisation est limitée en interprétation et, de ce fait, sous-exploitée. Cette thèse se propose de développer une nouvelle méthode, rapide et efficace, pour la quantification de données métabolomiques à partir des spectres RMN. Les résultats de la méthode seront validés par l'application à des données déjà acquises du projet PORCINET (ANR-09-GENM-005) et les données en voie d'acquérir dans le cadre du projet SuBPig (GISA, 2018-2019) qui s'attaquent au problème de la mortalité périnatale chez le porc. Dans le cadre de ce projet, des quantifications directes de certains métabolites ont été effectuées et seront utilisées pour valider la méthode. Enfin, les résultats obtenus seront intégrés aux autres variables 'omiques disponibles dans le but de proposer des biomarqueurs exploitables pour la prédiction de la mortalité à la naissance chez le porc.

Sujet détaillé

L'objectif de la thèse est d'exploiter une partie des très nombreuses données issues du projet PORCINET, ANR-09-GENM-005 afin de préciser le contrôle génétique et les bases moléculaires de la fin du développement fœtal chez le porc et d'identifier de nouveaux marqueurs de maturité périnatale. Ce projet (coord. L. Liaubet) a concerné 612 fœtus prélevés sur 39 portées à deux stades de fin de gestation (90 et 110 jours, la naissance étant à 114 jours). Le plan d'expérience permet de comparer quatre génotypes : deux génotypes extrêmes pour la survie néonatale, les Meishan (MS), lignée rustique avec peu ou pas de mortalité néonatale, et les Large White (LW), lignée sélectionnée pour plus de productivité ainsi que les croisés réciproques (données décrites dans la partie « Matériel nécessaire » ci-dessous). En particulier, les fluides (urine, plasma et liquide amniotique) ont été prélevés et analysés par Résonance Magnétique Nucléaire (RMN-1H). L'avantage de la RMN est de permettre une analyse haut-débit et, ici, l'analyse des 3 fluides sur tous les fœtus. L'analyse des données discrétisées a rapidement permis d'évaluer la puissance du dispositif expérimental et d'apporter les premières informations biologiques importantes et originales mais non encore publiées. Par contre, si la RMN représente un outil très intéressant pour réaliser un phénotypage fin et haut-débit à coût relativement raisonnable (par rapport au transcriptome), elle est limitée en termes d'interprétabilité. En effet, un des enjeux majeurs avec ce type de données est de pouvoir les transformer en information biologique pour savoir quelles molécules et quelles voies métaboliques sont effectivement mises en jeu dans le contexte expérimental. Ici, il s'agit de comprendre comment la sélection génétique a affecté la mise en place de la régulation du métabolisme énergétique (retard et/ou hétérogénéité de développement identifié chez les LW, [1]).

Actuellement, deux voies principales sont utilisées pour aborder ce problème : la première consiste à travailler sur les spectres bruts ou agglomérés en buckets, en utilisant éventuellement des approches par ondelettes pour le pré-traitement [2]. Les analyses statistiques (fouille de données, tests, méthodes de prédiction…) permettent alors d'extraire des zones du spectre qui sont d'intérêt pour le problème biologique [3]. Cependant, l'identification des métabolites correspondants à ces zones est souvent complexe : les métabolites peuvent avoir plusieurs pics en commun et, de plus, les métabolites identifiés ne sont pas précisément quantifiés. Une approche alternative consiste tout d'abord à utiliser les spectres RMN pour quantifier les métabolites présents dans l'échantillon puis à effectuer les analyses statistiques sur les résultats obtenus lors de cette étape de quantification. La quantification est un une question ouverte en statistique : une première méthode est proposée dans

le package R BATMAN [4,5] qui utilise une approche bayésienne. Cependant, cette approche s'avère très coûteuse en temps de calcul et difficile à mettre en œuvre sur un nombre d'échantillons aussi important que les nôtres. Des résultats préliminaires montrent qu'une approche LASSO, méthodologie adaptée aux données de grande dimension, permet d'obtenir des résultats prometteurs en un temps raisonnable. Cette approche, nommée ASICS, a été récemment développée au sein de l'unité InTheRes, notamment par R. Servien, dans le cadre d'un projet multidisciplinaire IDEX Toulouse. Les bases théoriques de cette méthode ont été étudiées ainsi que ces premières applications à des données métabolomiques. Elle a donné lieu à deux publications, une en statistique, actuellement soumise, et une en métabolomique (à laquelle G. Lefort, l'étudiante pressentie pour cette thèse est associée) [6,7]. L'utilisation de cette méthode permet d'améliorer la qualité de la quantification à un coût computationnel très réduit.

L'objectif de la thèse est donc double : d'un point de vue méthodologique, il s'agit de comparer les diverses approches d'analyse des données métabolomiques et de valider la quantification en comparaison des dosages effectués directement dans le sang (glucose, lactate, fructose, acides aminés, métabolites qui reflètent l'état des ressources énergétiques). Cette étape doit permettre d'améliorer la méthode ASICS et de l'intégrer à divers outils génériques d'analyse de données métabolomiques (Galaxy, package R) afin d'en faire une méthode de référence pour l'analyse de données métabolomiques. La deuxième partie de la thèse est à visée applicative : l'objectif de la thèse sera donc aussi, en utilisant les méthodologies développées pour la quantification des métabolites, l'identification de biomarqueurs utilisables sur animaux vivants (prélèvement de sang à la naissance). La thèse se rattache à la problématique de l'élevage de précision, dans laquelle le phénotypage fin et haut débit est d'ores et déjà envisagé par les organismes de sélection porcine pour produire des lignées maternelles ayant de meilleures aptitudes pour assurer une plus grande survie des portées.

Références

[1] Voillet V, SanCristobal M, Lippi Y, Martin PG, Iannuccelli N, Lascor C, Vignoles F, Billon Y, Canario L, Liaubet L. Muscle transcriptomic investigation of late fetal development identifies candidate genes for piglet maturity. BMC Genomics. 2014 Sep 17;15:797.

[2] Villa-Vialaneix, N., Hernandez, N., Paris, A., Domange, C., Priymenko, N., & Besse, P. (2016). On combining wavelets expansion and sparse linear models for regression on metabolomic data and biomarker selection. Communications in Statistics - Simulation And Computation, 45(1), 282–298.

[3] Rohart, F., Paris, A., Laurent, B., Canlet, C., Molina, J., Mercat, M.J., Tribout, T., Muller, N., Iannuccelli, N., Villa-Vialaneix, N., Liaubet, L., Milan, D. & San Cristobal, M. (2012) Phenotypic prediction based on metabolomic data for growing pigs from three main European breeds. Journal of Animal Science, 90(13), 4729-4740.

[4] Astle, W., de Iorio, M., Richardson, S., Stephens, D. & Ebbels, T. (2012) A Bayesian model of NMR spectra for deconvolution and quantification of metabolites in complex biological mixtures. Journal of the American Statistical Association, 107(500), 1259-1271.

[5] Hao, J., Liebeke, M., Astle, W., De Iorio, M., Bundy, G. & Ebbels, T.M.D. (2014). Bayesian deconvolution and quantification of metabolites in complex 1D NMR spectra using BATMAN, Nature Protocols, 9, 1416-1427.

[6] Tardivel, P., Servien, R. & Concordet, D. Familywise Error Rate Control With a Lasso Estimator. *Submitted.*

[7] Tardivel, P., Canlet, C., Tremblay-Franco, M., Lefort, G., Debrauwer, L., Concordet, D. & Servien, R. ASICS: an automatic method for identification and quantification of metabolites in NMR 1D 1H spectra. Metabolomics, **13**(109).

Encadrement

- Nathalie Villa-Vialaneix (HDR, directrice de la thèse, MIA)
- Rémi Servien (non HDR, co-directeur de la thèse, InTheRes)
- Laurence Liaubet (non HDR, GenPhysE)
- Hélène Quesnel (HDR, PEGASE)

La thèse se déroulera à l'unité MIAT, INRA Toulouse (site d'Auzeville).
Contact : Nathalie Villa-Vialaneix <nathalie.villa-vialaneix[AT]inra.fr>

Profil souhaité pour le ou la candidat·e

De formation en mathématiques appliquées et statistique, vous avez un goût prononcé pour les applications et, idéalement, une expérience préalable dans l'analyse des données issues des technologies haut débit en biologie.

Pour candidater, envoyer un CV et une lettre de motivation **avant le 30 avril 2018**.