ASICS : identifier et quantifier des métabolites à partir d'un spectre RMN ^{1}H

Gaëlle Lefort[1], Laurence Liaubet[2], Hélène Quesnel[3], Cécile Canlet[4,5], Nathalie Vialaneix[1] & Rémi Servien[6]

[1]*MIAT, Université de Toulouse, INRA, Castanet Tolosan, France* {*gaelle.lefort@inra.fr, nathalie.vialaneix@inra.fr*}
[2]*GenPhySE, Université de Toulouse, INRA, ENVT, Castanet Tolosan, France* {*laurence.laubet@inra.fr*}
[3]*PEGASE, INRA, Agrocampus Ouest, 35590, Saint-Gilles, France* {*helene.quesnel@inra.fr*}
[4]*Toxalim, Université de Toulouse, INRA, ENVT, INP-Purpan, UPS, 31027 Toulouse, France,*
[5]*Axiom Platform, MetaToul-MetaboHUB, National Infrastructure for Metabolomics and Fluxomics, 31027 Toulouse, France* {*cecile.canlet@inra.fr*}
[6]*INTHERES, Université de Toulouse, INRA, ENVT, Toulouse, France* {*remi.servien@inra.fr*}

Résumé. La résonance magnétique nucléaire du proton (^{1}H-RMN) est une technologie haut-débit permettant d'obtenir des profils métaboliques, sous forme de spectres, à un coût relativement faible. C'est un outil prometteur pour détecter des biomarqueurs facilement mesurables. Cependant, les métabolites présents dans un mélange complexe ne sont pas identifiables et quantifiables directement, ce qui limite l'interprétabilité de ces approches. Pour faciliter l'utilisation de ces données, nous avons développé une méthode d'analyse automatique, encapsulée dans un nouveau package R/Bioconductor, **ASICS**, qui permet l'identification et la quantification globale et automatique des métabolites dans un spectre RMN. Le package permet d'enchaîner facilement toutes les étapes de l'analyse (pré-traitements, quantification, outils de diagnostic pour juger de la qualité des quantifications, analyses statistiques post-quantification). La méthode de quantification, préexistante (Tardivel et al., 2017), a été testée sur un jeu de données réel (ANR PORCINET). Le but était de juger des performances de la méthode comparativement à celles existantes et de l'améliorer grâce à un paramétrage plus fin.

Mots-clés. Métabolomique, Résonance Magnétique Nucléaire (RMN), Quantification de métabolites, Sélection de variables

Abstract. ^{1}H Nuclear Magnetic Resonance (NMR) is a high-throughput technology that allows to obtain metabolomic profile easily (*e.g.*, from fluids such as blood) and at low cost. It is a promising tool to detect easily measured biomarkers. However, its interpretation can be difficult, because metabolites present in the complex mixture can not be automatically identified and quantified. To ease the use of such data, we developed a new method package, embedded in an R/Bioconductor package **ASICS**, which performs a global and automatic identification and quantification of metabolites in ^{1}H NMR spectra.

The package combines all the steps of the analysis (preprocessing, quantification, diagnosis tools to assess the quality of the quantification, post-quantification statistical analyses). The quantification method, based on Tardivel et al. (2017), have been tested on a real dataset (ANR PORCINET) to assess its performances and provide a fine tuning of all the parameters

Keywords. Metabolomics, Nuclear Magnetic Resonance (NMR), Quantification of metabolites, Variable selection

1 Introduction

La métabolomique est l'étude de l'ensemble des petites molécules impliquées dans les réactions chimiques métaboliques d'un organisme. C'est une approche prometteuse pour la caractérisation des phénotypes et la découverte de biomarqueurs, dans différents domaine comme l'agriculture, la microbiologie, l'environnement ou la santé. Deux approches complémentaires sont utilisées pour obtenir des profils métaboliques : la résonance magnétique nucléaire (RMN) et la spectrométrie de masse. Ces technologies permettent de détecter des centaines de métabolites dans divers types d'échantillons (organes, biofluides...) grâce à la production de spectres (Figure 1). Cependant, à cause de leur complexité et du grand nombre de signaux générés, l'analyse de telles données reste un challenge majeur pour la métabolomique haut-débit.

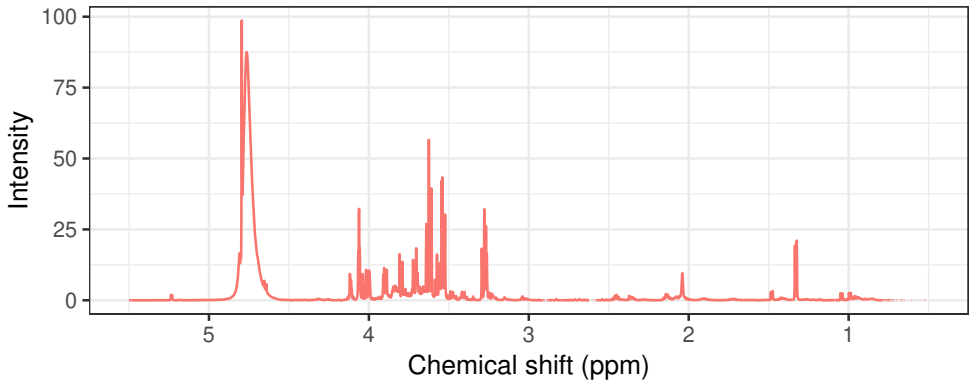

Figure 1: Exemple d'un spectre de RMN d'environ 30 000 points. En fonction de leur structure, les métabolites présents dans le mélange vont avoir un ou plusieurs pics à un déplacement horizontal prédéfini. L'intensité des pics permet de connaître la concentration du métabolite dans le mélange.

Cette communication se focalise sur les spectres issus de la RMN. L'approche usuelle pour traiter ce type de données est dans un premier temps de diviser le spectre en intervalles appelés *buckets*. Ensuite, l'aire sous la courbe est calculée pour chaque *bucket* et

les analyses statistiques sont réalisées sur ces nouvelles variables (environ 700 variables). Cependant, les *buckets* ne sont pas directement liés aux métabolites : un pic du spectre peut correspondre à plusieurs métabolites et un métabolite peut avoir plusieurs pics en fonction de sa structure chimique. Il est donc nécessaire que des experts en RMN identifient manuellement les *buckets* issus de l'analyse pour pouvoir interpréter biologiquement les résultats obtenus. Cette identification est longue, fastidieuse, dépend de l'expert et est très difficilement reproductible. De plus, seuls les *buckets* extraits de l'analyse sont identifiés ce qui entraîne une importante perte d'information (Considine et al., 2018).

Des méthodes ont donc été développées pour identifier et quantifier la concentration des métabolites dans un spectre RMN (Autofit (Weljie et al., 2006), **batman** (Hao et al., 2012), Bayesil (Ravanbakhsh et al., 2015) et **rDolphin** (Cañueto et al., 2018)). Récemment, Tardivel et al. (2017) ont développé une nouvelle méthode statistique pour identifier et quantifier automatiquement les métabolites présents dans un spectre. Cette méthode, basée sur une librairie de spectres purs, semble plus performante que les autres. Néanmoins, elle se focalise principalement sur l'étape de quantification et nécessite d'être couplée à des pré-traitements et post-traitements pour la rendre pleinement utilisable par les biologistes. Le package R, **ASICS** (*Automatic Statistical Identification in Complex Spectra*), a été développé dans cette optique. Les méthodes d'identification et de quantification y sont partiellement basées sur Tardivel et al. (2017) mais ont été testées sur des jeux de données réels et améliorées pour obtenir un paramétrage plus fin.

2 Les étapes de l'analyse de spectres RMN

2.1 Pré-traitements du spectre d'un échantillon (mélange complexe)

Après l'import des spectres depuis les fichiers bruts (FID) grâce au package R **PepsNMR** (Martin et al., 2018) ou en partie déjà prétraités, plusieurs étapes de pré-traitements sont recommandés pour supprimer les biais techniques. Pour cela, nous avons utilisé des méthodes existantes que nous avons combinées et améliorées pour parvenir à un prétraitement efficace. L'algorithme de Wang et al. (2013) avec un paramétrage adapté est utilisé pour corriger la ligne de base. Pour aligner les pics des spectres entre eux, nous avons adapté l'algorithme de Vu et al. (2011). Enfin, nous avons choisi de normaliser les spectres par l'aire sous la courbe (Craig et al., 2006).

2.2 Pré-traitements de la librarie de référence

Une librairie de spectres de métabolites purs est utilisée comme référence pour identifier et quantifier les métabolites dans le mélange complexe. Une telle librairie, composée de 190 spectres, est disponible dans le package. Comme pour le mélange complexe, des pré-

traitements sont nécessaires. Cependant, aucune méthode n'était disponible pour réaliser une pré-sélection des spectres nous permettant ensuite de réaliser une régression sur un nombre réduit de spectre. Elle ont donc dues être développées.

Suppression du bruit Tous les spectres RMN contiennent du bruit mais alors qu'il est délicat de le supprimer dans un mélange complexe cela est utile sur un spectre pur grâce à un seuillage. Cela va permettre de déterminer plus facilement la position des pics lors des prochaines étapes.

Première étape de sélection Un métabolite ne peut pas appartenir au mélange complexe si tous ses pics ne sont pas présents. Des biais techniques peuvent également faire varier légèrement les déplacements chimiques des spectres. Partant de ces deux constatations, un spectre de la librairie de référence est conservé si tous ses pics sont présents dans le mélange complexe avec un décalage horizontal maximal de M entre ces deux spectres.

Translation et déformation Pour réaliser la quantification, il est nécessaire d'aligner les spectres de la librairie avec le mélange complexe. Pour ce faire, une procédure en deux étapes est utilisée. Dans un premier temps, les spectres purs sont alignés en maximisant la corrélation croisée de la transformé discrète de Fourier avec un décalage maximal M (Wong et al., 2005). Dans un second temps, chaque pics est aligné individuellement, sur un intervalle plus petit, $m = \frac{M}{5}$, en minimisant les résidus de la régression linéaire entre le mélange complexe et le spectre pur.

2.3 Quantification relative des concentrations des métabolites

En utilisant le mélange complexe et la librairie pré-traités, la quantification est réalisée comme décrite dans Tardivel et al. (2017). Le mélange complexe est défini comme une combinaison linéaire des spectres de la librairie de référence :

$$g(t) = \sum_{i=1}^{p} \beta_i f_i(\Phi_i(t)) + \epsilon(t) \qquad \text{with } \beta_i \geq 0 \tag{1}$$

où g correspond au mélange complexe, $f_i \circ \Phi_i$ aux spectre de la librairie, $\beta = (\beta_1, \ldots, \beta_p)$ aux coefficients associés à ces spectres et ϵ au bruit. Les coefficients β sont estimés par moindres carrés sous contraintes grâce à l'algorithme de Goldfarb and Idnani (1983). Une procédure de sélection de variables est, ensuite, implémentée pour obtenir un β parcimonieux en contrôlant le Family Wise Error Rate (FWER) avec un risque α. Une fois les métabolites sélectionnés, les quantifications $(\beta_i)_i$ pour ces métabolites sont ré-estimés en restreignant l'équation (1) à ce sous-ensemble et des quantifications relatives, dépendant de propriétés chimiques de chaque métabolite, sont finalement calculées. Cette méthode permet de limiter le biais d'estimation des procédures parcimonieuses.

3 Validation des quantifications

Pour tester les différentes méthodes de quantification, les corrélations entre les quantifications et des dosages biochimiques de trois métabolites ont été réalisés sur 32 spectres de plasma. Ces spectres correspondent à un sous-ensemble représentatif des données obtenues dans le cadre du projet ANR PORCINET. Ce projet a pour but de comprendre les mécanismes de maturation en fin de gestation chez le porcelet.

Table 1: Corrélations entre les dosages biochimiques de trois métabolites et les quantifications relatives obtenues grâce à trois méthodes concurrentes et les *buckets* connus correspondants aux métabolites cibles. *Bucket* du lactate: 1.335; *bucket* du fructose: 3.995; *bucket* du glucose: 5.235. Le temps de calcul est donné pour un spectre.

	Lactate	Fructose	Glucose	Temps de calcul	Structure de calcul parallèle
ASICS	0.93	0.95	0.90	$\sim$ 1'30 min	Oui
Autofit	0.52	0.74	0.75	$<$ 1min	Non
batman (avec 160 métab.)	0.46	0.56	0.22	$\sim$ 2 jours	Oui
rDolphin	0.82	NA	0.77	$\sim$ 1'30 min	Non
Buckets	0.93	0.95	0.90	2 s	Oui

Les corrélations (Tableau 1) montrent que le package **ASICS** est plus performant que les méthodes Autofit, **batman** et **rDolphin** pour ces trois métabolites. De plus, les corrélations sont identiques à celles obtenus entre les *buckets* et les dosages. En terme de temps de calcul, les pré-traitements et la quantification pour un spectre prennent environ 1'30min et peuvent être lancés en parallèle pour diminuer le temps global.

4 Conclusion

ASICS permet de réaliser toutes les étapes de l'analyse de spectres RMN. Il intègre une méthode automatique d'identification et de quantification des métabolites basée sur une librairie de spectres purs. Sur ce point, **ASICS** montre de meilleurs résultats que les méthodes existantes et permet de réaliser une étude complète d'un jeu de données de plusieurs centaines de spectres en seulement quelques heures. Son utilisation sur un jeu de données réelles produit des résultats similaires à l'analyse standard sur les *buckets* suivie d'une identification par un expert. Mais elle permet également d'apporter de nouvelles informations. Cependant, comme c'est le cas avec les autres données omiques, il peut être nécessaire de valider les métabolites détectés avec d'autres techniques comme de la RMN en 2 dimensions ou des dosages spécifiques.

Comme les autres méthodes automatiques, **ASICS** a toutefois quelques limitations : l'algorithme a des difficultés à identifier des métabolites en faibles concentrations ou dont

tous les pics sont localisés dans une région dense en pics. Ce sera donc le point naturel de focalisation des futurs travaux, par exemple en utilisant l'ensemble des informations extraites sur tous les spectres d'un jeu de données afin d'améliorer les quantifications individuelles.

Remerciements

Les données utilisées dans cet article ont été produites dans le cadre d'un projet soutenu par l'ANR (PORCINET grant ANR-09-GENM005). La thèse de Gaëlle Lefort est financée par l'Institut de Convergence #DigitAg (Agriculture Digitale, `http://www.hdigitag.fr/`, grant ANR-16-CONV-0004), et par les départements Mathématiques et Informatique Appliquées, Génétique Animale et Santé Animale de l'INRA.

Bibliographie

Cañueto, D. *et al.* (2018). rDolphin: a GUI R package for proficient automatic profiling of 1D 1H-NMR spectra of study datasets. *Metabolomics*, 14(3):24.

Considine, E. *et al.* (2018). Critical review of reporting of the data analysis step in metabolomics. *Metabolomics*, 14(1):7.

Craig, A. *et al.* (2006). Scaling and normalization effects in NMR spectroscopic metabonomic data sets. *Analytical Chemistry*, 78(7):2262–2267.

Goldfarb, D. and Idnani, A. (1983). A numerically stable dual method for solving strictly convex quadratic programs. *Mathematical Programming*, 27(1):1–33.

Hao, J. *et al.* (2012). BATMAN – an R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a Bayesian model. *Bioinformatics*, 28(15):2088–2090.

Martin, M. *et al.* (2018). PepsNMR for 1H NMR metabolomic data pre-processing. *Analytica Chimica Acta*, 1019:1–13.

Ravanbakhsh, S. *et al.* (2015). Accurate, fully-automated NMR spectral profiling for metabolomics. *PLOS ONE*, 10(5):e0124219.

Tardivel, P. *et al.* (2017). ASICS: an automatic method for identification and quantification of metabolites in complex 1D 1H NMR spectra. *Metabolomics*, 13(10):109.

Vu, T. *et al.* (2011). An integrated workflow for robust alignment and simplified quantitative analysis of NMR spectrometry data. *BMC Bioinformatics*, 12(1):405.

Wang, K. *et al.* (2013). Distribution-based classification method for baseline correction of metabolomic 1D proton nuclear magnetic resonance spectra. *Analytical Chemistry*, 85(2):1231–1239.

Weljie, A. *et al.* (2006). Targeted profiling: quantitative analysis of 1 H NMR metabolomics data. *Analytical Chemistry*, 78(13):4430–4442.

Wong, J. *et al.* (2005). Application of fast Fourier transform cross-correlation for the alignment of large chromatographic and spectral datasets. *Analytical Chemistry*, 77(17):5655–5661.