Méthodes statistiques d'identification et de quantification en métabolomique. Application aux spectres RMN.

Patrick Tardivel, Cécile Canlet, Marie Tremblay-Franco, Laurent Debrauwer, Didier Concordet et Rémi Servien

ENVT-INRA, Université de Toulouse, UMR1331 Toxalim, Research Centre in Food Toxicology, F-31027 Toulouse, France
email : patrick.tardivel@toulouse.inra.fr

Résumé. La métabolomique est une science qui s'intéresse à l'identification et la quantification des métabolites (petites molécules) à partir d'un mélange obtenu dans le sang, l'urine, le plasma, ... Une des techniques les plus employées pour la caractérisation de métabolites est la résonance magnétique nucléaire du proton (RMN). Pour chaque métabolite la RMN produit un spectre spécifique. De même pour un mélange, la RMN génère un spectre qui est une combinaison convexe des spectres des métabolites qui le composent. Cependant, tous ces signaux sont observés bruités (variation de l'amplitude des pics) et déformés (variation de la forme et de la localisation des pics). Ainsi, il est très délicat de calculer exactement les proportions des métabolites du mélange. Nous proposons dans un premier temps d'estimer la déformation associée à chaque métabolite. Dans un deuxième temps, une méthode statistique basée sur une approche par programmation linéaire permet d'obtenir une estimation parcimonieuse des proportions. Enfin, des résultats sur des données réelles montrent l'efficacité de notre méthode.

Mots-clés. Traitement du signal, déformation, programmation linéaire, métabolomique, spectres RMN.

Abstract. Metabolomics is a science concerned with characterization of metabolites (kind of molecules) in a mixture (blood, urine, ...). The most common technique for obtaining such a characterization is proton nuclear magnetic resonance (NMR). For each metabolite, NMR produces a specific spectrum and the mixture has a spectrum that is a convex combination of the different spectra of the metabolites that compose this mixture. However, each signal is observed noisy (variation of magnitude of peaks) and warped (variation of shape and localization of peaks). Thus, it is very tricky to compute the exact value of each proportion. Instead, in the first step we estimate the warping function of each metabolite. An algorithm is developped in this way. In a second step, we propose a statistical method based on a linear programming approach that gives us a sparse estimation of the proportions. Finally, results on real data assess the good performances of our method.

Keywords. Signal processing, warping function, linear programming, metabolomics, NMR spectra.

1 Introduction

La métabolomique est une science qui s'intéresse à la caractérisation et la quantification de métabolites, ces petites molécules que l'on retrouve dans les cellules, les tissus, les fluides biologiques et les organismes. La technique la plus utilisée pour obtenir cette caractérisation est la résonance magnétique nucléaire des protons (RMN). Afin d'identifier ces métabolites, les experts utilisent une bibliothèque personnelle qui contient les spectres des métabolites purs et comparent ces spectres à la main à celui du mélange biologique à analyser. Plus précisément, lorsqu'un expert veut savoir si un métabolite particulier est présent dans un mélange, il vérifie si tous les pics du spectre de ce métabolite se retrouvent dans le spectre du mélange. Cette méthode dépend donc grandement des connaissances de l'expert, notamment du nombre de spectres de métabolites qu'il connaît. Cette identification peut également être rendue délicate par la déformation des spectres (dûe par exemple à une variation de pH) ou le chevauchement de certains des pics des métabolites présents dans le mélange. Etudions en détails un exemple d'identification de métabolite dans un spectre avec la Figure 1.

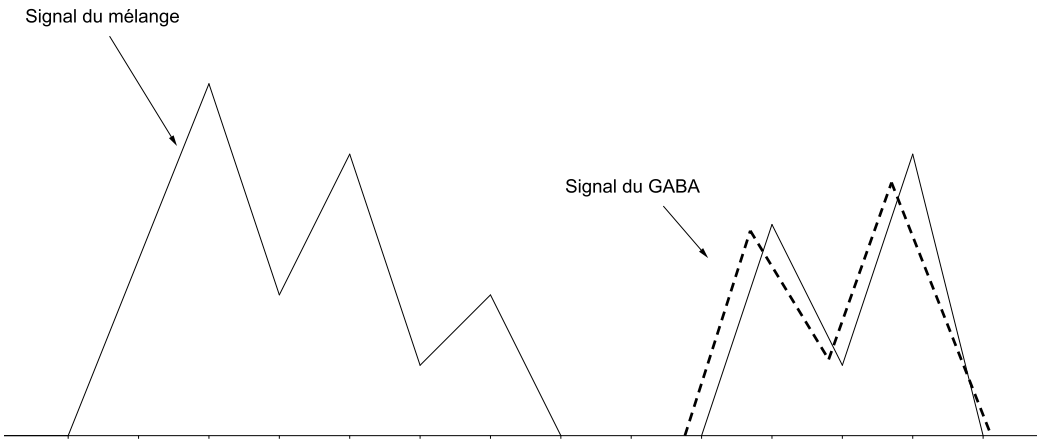

Figure 1: Superposition des spectres du mélange complexe et d'un métabolite particulier (le GABA ici en pointillé).

Dans un mélange complexe, l'expert cherche à identifier la présence d'un métabolite particulier (ici le GABA) en vérifiant si tous les pics du spectre de ce métabolite se retrouve dans le spectre du mélange à une petite déformation près. D'après les spectres de la Figure 1, l'expert conclura que le GABA est présent dans ce mélange.

Certaines méthodes automatiques d'identification et de quantification des métabolites ont été proposées récemment mais elles restent perfectibles. En effet, MetaboHunter [1] est très rapide mais ne permet pas gérer le chevauchement des pics et donc la compétition

entre métabolites. De plus, elle ne fournit qu'un score de présence d'un métabolite lié au nombre de ses pics qui ont été identifiés dans le mélange. D'un autre côté le fort coût computationnel de BATMAN [2,3] ne permet pas de rechercher des dizaines de métabolites. Par conséquent, il n'existe pas encore une méthode de référence dans ce domaine.

2 Modélisation

Afin d'automatiser cette étape d'identification, nous allons tout d'abord modéliser le problème. Pour chaque métabolite, nous possédons un spectre spécifique Z_i dans notre bibliothèque personnelle, Z_i est un processus aléatoire défini par

$$Z_i(t) = f_i(\phi_i(t))(1 + \varepsilon_i(t)),\ t \in T$$

où T est un compact de $\mathbb{R}$, f_i est une fonction positive représentant le spectre pur du i^e métabolite et vérifiant $\int_T f_i(t)dt = 1$; ϕ_i une fonction continue strictement croissante appelée fonction déformante [4] et ε_i un processus stochastique homoscédastique de variance σ^2. Le bruit ε_i agit sur l'amplitude des pics du spectre tandis que la fonction déformante ϕ_i modifie la forme et la localisation des pics. Afin de fixer les idées, la Figure 2 fournit un exemple de fonction déformante.

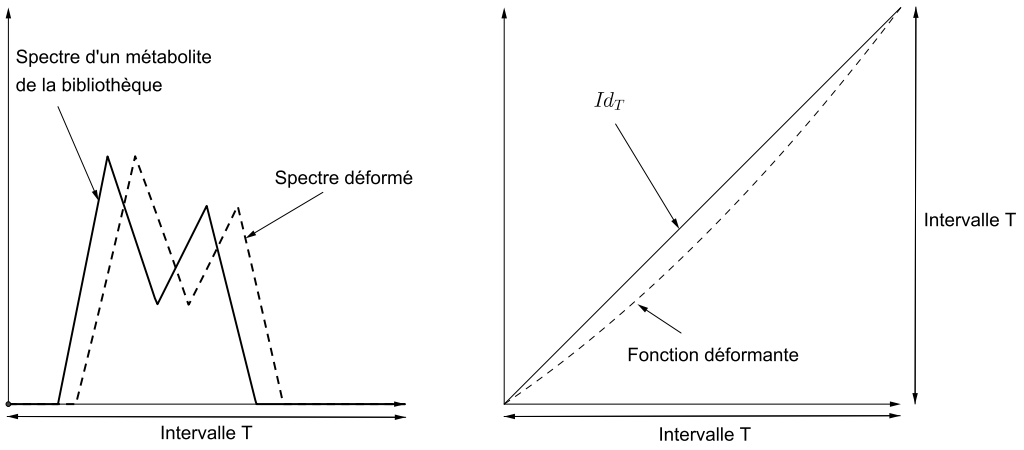

Figure 2: La partie gauche de la figure représente le spectre d'un métabolite de la bibliothèque Z_i et un second spectre en pointillé $Z_i \circ \phi_i$ qui est une déformation du premier spectre. La fonction déformante ϕ_i est représentée en pointillé sur la partie droite de la figure.

D'après les experts, la distance en norme infinie entre la fonction déformante ϕ et l'identité Id_T est plus petite qu'une valeur $\epsilon > 0$. Cette condition nous indique que la localisation d'un même pic d'un métabolite ne peut varier que faiblement d'un mélange à l'autre.

De façon similaire, le spectre d'un mélange peut être modélisé par

$$Y(t) = \left(\sum_i \alpha_0(i) f_i(t) \right) (1 + \varepsilon(t)),$$

où $\alpha_0(i) \in [0, 1]$ représente la proportion du métabolite i dans le mélange et ε est un processus stochastique homoscédastique de variance σ^2. Notre objectif principal est de proposer une méthode automatique permettant de caractériser la composition du mélange Y en déterminant les proportions α_i des métabolites.

Les paramètres de notre modèle sont donc les proportions α_i et les fonctions déformantes ϕ_i. L'estimation simultanée de tous ces paramètres est difficile, voire impossible en un temps raisonnable. Néanmoins une bonne approximation peut être obtenue en estimant dans un premier temps les fonctions déformantes ϕ_i.

Nous avons développé une méthode itérative basée sur une composition de fonctions affines par morceaux pour estimer ϕ_i. Nous savons que l'ensemble des fonctions déformantes affines par morceaux est dense dans l'ensemble des fonctions déformantes. De plus, nous avons prouvé que toutes fonctions déformantes affines par morceaux s'écrit comme une composition de fonctions déformantes affines par morceaux élémentaires de la forme

$$\phi(a) = a, \phi(x_1) = x_1, \phi(x_2) = y_2, \phi(x_3) = x_3 \text{ et } \phi(b) = b.$$

Avec, $T = [a, b], a \leqslant x_1 < x_2 < x_3 \leqslant b$ et $x_1 \leqslant y_2 \leqslant x_3$. À l'itération $n + 1$, la fonction déformante sera de la forme

$$\hat{\phi}_{n+1} = \hat{\phi}_n \circ \phi,$$

avec ϕ une fonction déformante élémentaire. Cette méthode fournit des estimations $\hat{\phi}_i$. Les proportions sont ensuite estimées par programmation linéaire [5] en utilisant $\hat{\phi}_i$ au lieu de ϕ_i. Ensuite, en maximisant $\sum_i \alpha_i$ sous les contraintes

$$\sum_i \alpha_i \hat{Z}_i \leq Y, \quad \forall i, \alpha_i \in [0, 1], \text{ et } \hat{Z}_i = Z_i \circ \hat{\phi}_i,$$

nous obtenons des estimateurs $\hat{\alpha}_i$.

Cette méthode a l'avantage d'être parcimonieuse et permet de gérer la compétition entre métabolites (chevauchement de pics issus de spectres différents) en calculant simultanément les proportions pour chaque métabolite.

3 Données réelles

Nous avons testé notre méthode sur deux mélanges de composition connue. Nous allons détailler les résultats obtenus pour le premier, les deux résultats étant sensiblement similaires. La bibliothèque contient 56 spectres de métabolites dont les 4 spectres de métabolites présents dans le mélange, ainsi que 52 autres spectres de métabolites non présents dans le mélange (Choline, Créatinine,...). La première étape de déformation permet de conserver 13 métabolites potentiellement présents dans le mélange (Acide Acetic, Alanine, Acide Aspartique, Betaine, Cadaverine, Carnosine, Créatinine, Maltose, Dimethylamine, Glucose, Glycine, Lysine et Trimethylamine) et, par conséquent, d'en éliminer 43. Ces 43 métabolites sont ceux dont les spectres nécessitent une déformation trop importante ($\|\phi - Id_T\|_\infty > \epsilon$) pour être insérés dans le spectre du mélange.

La seconde étape d'optimisation des proportions permet d'exclure la présence de 9 métabolites et identifie correctement les 4 métabolites présents avec les proportions indiquées dans le Tableau 1. Il est également important de noter que ces deux étapes ne prennent pas plus d'une quinzaine de secondes.

Métabolites	Glycine	Glucose	Acide Aspartique	Lysine
Vraie proportion	75%	15%	8%	2%
Proportion estimée	26%	12%	3%	4%

Table 1: Comparaison des vraies proportions et des proportions estimées sur un mélange connu.

On remarque que l'estimation des proportions par programmation linéaire semble globalement biaisée. Ceci vient du fait que la combinaison linéaire des estimations par programmation linéaire $\hat{\alpha}$ avec les spectres déformés $\hat{Z}$ est plus petite que le spectre du mélange Y. En d'autre termes, on a

$$\hat{\alpha}_{Gly}\hat{Z}_{Gly} + \hat{\alpha}_{Glu}\hat{Z}_{Glu} + \hat{\alpha}_{Asp}\hat{Z}_{Asp} + \hat{\alpha}_{Lys}\hat{Z}_{Lys} \leqslant Y.$$

La Figure 3 ci-dessous illustre cette inégalité.

Afin de débiaiser nos estimations, nous nous intéressons actuellement à la convergence en loi de la variable aléatoire $(\hat{\alpha}_i - \alpha_0(i))/\sigma$ lorsque σ tend vers 0. Le calcul de l'espérance et des quantiles de la loi limite, nous permettrait de réduire le biais de notre estimateur et d'obtenir des intervalles de confiance asymptotiques.

Ce projet bénéficie du soutien financier du Ministère de l'Écologie, du Développement Durable et de l'Énergie dans le cadre du programme national de recherche Risk'OGM ainsi que de l'IDEX Toulouse "Transversalité 2014".

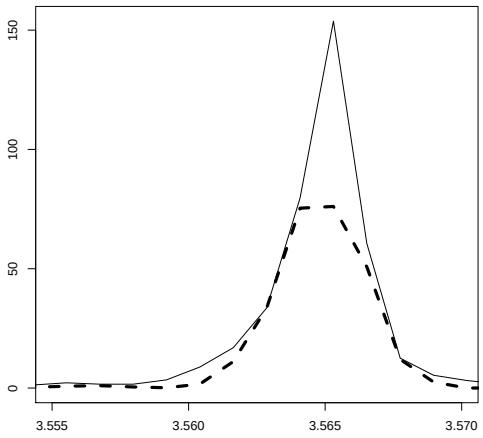

Figure 3: La fonction $\hat{\alpha}_{Gly}\hat{Z}_{Gly} + \hat{\alpha}_{Glu}\hat{Z}_{Glu} + \hat{\alpha}_{Asp}\hat{Z}_{Asp} + \hat{\alpha}_{Lys}\hat{Z}_{Lys}$ est en pointillé, le mélange Y est en trait plein. On observe graphiquement que, pour satisfaire la condition $\hat{\alpha}_{Gly}\hat{Z}_{Gly}+\hat{\alpha}_{Glu}\hat{Z}_{Glu}+\hat{\alpha}_{Asp}\hat{Z}_{Asp}+\hat{\alpha}_{Lys}\hat{Z}_{Lys} \leqslant Y$, la combinaison linéaires des 4 métabolites ne peut investir tout le pic. Les proportions $\hat{\alpha}$ sont alors "bloquées", laissant apparaître le biais inhérent à l'utilisation de la programmation linéaire.

Bibliographie

[1] Tulpan, D., Léger, S., Belliveau, L., Culf, A. et Čuperlović-Culf, M. (2011). MetaboHunter: an automatic approach for identification of metabolites from 1H-NMR spectra of complex mixtures. BMC bioinformatics, 12(1), 400.
[2] Astle, W., De Iorio, M., Richardson, S., Stephens, D. et Ebbels, T. (2012). A Bayesian model of NMR spectra for the deconvolution and quantification of metabolites in complex biological mixtures. Journal of the American Statistical Association, 107(500), 1259-1271.
[3] Hao, J., Astle, W., De Iorio, M. et Ebbels, T. M. (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.
[4] Wierzbicki, M. R., Guo, L. B., Du, Q. T. et Guo, W. (2014). Sparse Semiparametric Nonlinear Model With Application to Chromatographic Fingerprints. Journal of the American Statistical Association, 109(508), 1339-1349.
[5] Boyd, S. et Vandenberghe, L. (2004). Convex optimization. Cambridge university press.