Identification et Quantification de métabolites dans un spectre RMN

Didier Concordet & Rémi Servien

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

Résumé. Nous nous intéressons ici à un problème rencontré en métabolomique. Ce domaine vise à caractériser la composition d'un mélange complexe par ses métabolites i.e. ses petites molécules. Les spectromètres RMN fournissent un spectre de mélange complexe qui est la superposition des spectres des métabolites purs. Chaque métabolite possède un spectre caractéristique, sa signature, qui le rend identifiable. Cependant, la reconnaissance automatique des métabolites dans un mélange complexe est rendu délicate par des problèmes comme la déformation du spectre (translation, dilatation ...) ou la superposition des pics. Nous proposons ici une méthode permettant d'identifier et de quantifier rapidement les métabolites dans un spectre complexe. Nous estimons tout d'abord les déformations à l'aide d'une procédure itérative puis nous calculons les proportions des métabolites de manière simultanée en utilisant un algorithme de programmation linéaire. Cette procédure, testée sur différents mélanges, s'avère performante et rapide.
Mots-clés. Spectre, Métabolomique, Déformation de fonction, Programmation linéaire.

Abstract. We are interested in this paper in metabolomics. All experiments in metabolimcs rely on the identification and the quantification of metabolites in complex biological mixtures. This information is usually obtained using NMR spectroscopy which supplies a spectra. Each metabolite generates a characteristic resonance signature in the NMR spectra with an intensity proportional to its concentration in the mixture. However, the automatic identification of metabolites is complicated by some problems like the distortion of the spectra or the overlapping of the peaks. We propose here an efficient method to identify and to quantify quickly the metabolites in a complex mixture. First, we estimate the warping functions for each metabolite using an iterative procedure. Then we compute simultaneously the proportion of the metabolites using linear programming. This methodology turns out to be efficient and quick on different datasets.
Keywords. Spectra, Metabolomic, Warping function, Linear programming.

1 Introduction

La métabolomique est une discipline récente s'intéressant à la caractérisation des métabolites, ces petites molécules se trouvant dans les cellules, les tissus ou les biofluides. Cette discipline est devenue de plus en plus populaire de par les récentes avancées technologiques. Par conséquent, la recherche dans ce domaine est en pleine expansion avec de

nombreuses applications par exemple en oncologie (Oakman and al., 2011), en génétique (Illig and al., 2010) etc ... La technique la plus commune utilisée en métabolomique est la spectroscopie par résonance magnétique nucléaire (RMN). Cette dernière permet de produire un spectre caractérisant chaque mélange complexe, ce spectre étant la somme des spectres des métabolites présents dans le mélange. Chaque métabolite génère une résonnance spectrale qui lui est propre avec une intensité proportionnelle à sa concentration dans le mélange. Le nombre de pics générés par un métabolite, comme leur localisation sur le spectre, est reproductible et uniquement déterminée : chaque métabolite possède donc sa signature sur le spectre.

Un des challenges majeurs en analyse RMN est par conséquent l'identification et la quantification des métabolites présents dans un mélange complexe. Différentes techniques ont été récemment introduites, tout d'abord basées sur la connaissance des experts (Pontolzeau and al., 2010). Ces approches sont cependant biaisées par l'expertise humaine et sont contraignantes et consommatrices de temps humain. Cependant, la construction de méthodes automatiques se heurte à plusieurs difficultés. Tout d'abord, à cause de plusieurs facteurs comme le pH ou l'intéraction entre ions, les pics générés par un métabolite peuvent être décalés ou déformés. Ensuite, un mélange complexe peut contenir des centaines ou des milliers de métabolites qui peuvent avoir certains pics en commun. Enfin, comme le nombre de métabolites d'intérêt peut rapidement dépasser le millier, on peut avoir à explorer un espace des possibles très important. La combinaison de ces facteurs rend l'identification et la quantification des métabolites assez délicate en pratique.

De récentes approches ont tenté de surmonter ces difficultés en comparant le spectre du mélange complexe à ceux, déjà obtenus, d'une bibliothèque de métabolites. Par exemple, MetaboHunter (Tulpan and al., 2011) consiste à comparer chaque métabolite de la bibliothèque avec la liste des pics du mélange complexe, en se basant exclusivement sur la position des pics dans le spectre. Une fonction de score est ensuite calculée pour chaque métabolite et donne sa probabilité de présence dans le spectre. Cette méthode est très simple et très rapide mais souffre de plusieurs inconvénients. Sa manière de gérer la superposition des pics est par exemple un peu simpliste et pas forcément adaptée. D'autres méthodes, basées sur la modélisation des pics par des courbes de Lorentz puis une estimation par algorithme MCMC, ont également été développées (Astle and al., 2012). Elles donnent de bons résultats mais sont très coûteuses en temps de calcul et donc difficilement applicables sur une grosse bibliothèque de métabolites.

Dans cet article, nous développerons une méthode originale pour l'identification et la quantification automatique des métabolites dans un mélange complexe. Cette méthode devra traiter les différents problèmes évoqués ci-dessus tout restant très rapide, le but final étant de l'appliquer à une bibliothèque de plusieurs dizaines de milliers de métabolites.

2 Pré-traitement des spectres

Les spectres bruts sortis directement du spectromètre RMN ne sont pas utilisables directement. Le but de cette étape est de "nettoyer" ces spectres de manière à les simplifier tout en conservant les spécificités de chaque métabolite. Il faut obtenir un compromis entre la simplification des spectres pour rendre les futurs calculs plus rapides et la nécessité de conserver suffisament d'information pour pouvoir identifier chaque métabolite.

En utilisant la connaissance des experts et en calibrant la variance du bruit sur des zones sans pics, on est capable de nettoyer de manière efficace les différents spectres. Cette étape permet de diminuer de 99.85% les zones du spectres informatives (i.e. avec une intensité supérieure à 0). Ensuite, les spectres sont normalisés afin de pouvoir être comparés. Ce même pré-traitement est ensuite appliqué au spectre du mélange complexe.

Finalement, le problème général est de déterminer les proportions $(\alpha_i)_{1\leq i\leq K}$ des K métabolites f_i dans le mélange Y tels que $\sum_{i=1}^{K}\alpha_i$ soit maximale dans

$$Y(t) = \sum_{i=1}^{K} \alpha_i f_i(\varphi_{\theta_i}(t)) + R(t) + \varepsilon(t)$$

où $\varphi_{\theta_i}(t)$ est la fonction de déformation de paramètres θ_i au poinr t, $R(t)$ est une fonction positive représentant la part de $Y(t)$ ne pouvant pas être expliquée à l'aide de notre bibliothèque de métabolites et $\varepsilon(t)$ un bruit inconnu. La fonction $R(t)$ est positive et ne peut pas être estimée de par sa nature. Le bruit $\varepsilon(t)$ est estimé de manière classique et on montre qu'il est multiplicatif, c'est à dire plus grand avec de grandes valeurs pour $Y(t)$. Donc nous pouvons borner $\varepsilon(t)$ par une fonction positive $M(Y(t))$ telle que

$$G(t) = Y(t) + M(Y(t)) \geq \sum_{i=1}^{K} \alpha_i f_i(\varphi_{\theta_i}(t)).$$

La fonction $M(Y(t))$ représente une tolérance accordée à la somme des métabolites et lui permettant de dépasser légèrement le mélange complexe Y. Il nous faut donc estimer les déformations (dûes au pH, à l'expérimentateur, au calibrage de la machine etc ...) et surtout les proportions des métabolites qui représentent les résultats d'intérêt pour les biologistes.

3 Estimation

Avant toute chose il est nécessaire de faire une hypothèse. S'il n'y a pas identifiabilité entre les métabolites (i.e. si le spectre d'un métabolite peut s'écrire comme la somme d'autres spectres), nous ne serons pas capables d'estimer les α_i. Cette hypothèse n'est pas

propre à notre méthode.

D'après le modèle ci-dessus, les estimations des déformations et des proportions doivent être simultanées, chacune étant fortement liée à l'autre. Cependant cette estimation simultanée est délicate et très longue, rendant la procédure inexploitable pour notre application. Une autre approche est donc préférée. Nous allons tout d'abord estimer les fonctions de déformation à travers leurs paramètres $\theta^* = (\theta_1^*, \ldots, \theta_K^*)$ où $\theta_i^* = \arg\sup_{\theta_i} U_i(\theta_i)$ et $U_i(\theta_i) = \arg\sup_{z_i>0} L(z_i(\theta_i))$ avec $L(\alpha(\theta)) = \inf_t G(t) - \sum_i \alpha_i f_i(\varphi_{\theta_i}(t))$. En d'autres termes, nous estimons pour chaque métabolite le paramètre de déformation θ_i^* qui nous donnerait le plus grand α_i si le métabolite était seul dans la bibliothèque. L'optimisation des paramètres θ_i^* est donc faite séparément, métabolite par métabolite. Puis nous estimons les différentes proportions en simultané, i.e. avec compétition entre les différents métabolites.

Cette méthode nous permet d'obtenir en un temps raisonnable une bonne approximation de $(\alpha_i(\theta_i))_{i=1,\ldots,K}$ car, triviallement, $\sup \sum_{i=1}^K \alpha_i(\theta_i) \geq \sup \sum_{i=1}^K \alpha_i(\theta_i^*)$.

3.1 Estimation des déformations

La première étape est donc d'obtenir le paramètre des fonctions de déformation θ^*. Comme expliqué ci-dessus, chaque θ_i^* est estimé séparément des autres. La procédure est la suivante : nous avons un métabolite de spectre f_i et un mélange complexe modifié G. Nous voulons estimer θ_i par θ_i^*, son estimateur qui maximise α_i dans $G(t) \geq \alpha_i f_i(\varphi_{\theta_i^*}(t))$. Donc, $\varphi_{\theta_i^*}$ est la fonction de déformation qui maximise la proportion du métabolite i dans le mélange complexe si ce métabolite est seul dans la librairie. Il convient de remarquer qu'il exite au moins un point t_0 tel que $G(t_0) = \alpha_i f_i(\varphi_{\theta_i^*}(t_0))$.

L'estimation de fonctions de déformation est un domaine en plein essor, en particulier en analyse d'images (Bigot et al., 2009). Cependant, pour différentes raisons (simplicité, temps de calcul ...) nous avons choisi de développer notre propre procédure. Notre méthode consiste en 3 étapes à répéter autant que nécessaire. Nous notons $\theta_i^{*,j}$ l'estimateur θ_i^* à l'itération j et $\alpha_{i,j}$ la proportion liée. Tout d'abord, nous prenons la valeur t_0 définie précédemment ainsi que sa composante connexe i.e le plus grand intervalle $[t_0 - a, t_0 + b]$ avec $f(x) > 0$ pour tout $x \in [t_0 - a, t_0 + b]$. Ensuite, nous déformons cette composante en lui appliquant la fonction $t \to t + \gamma t(1-t)$ ce qui nous fournit un nouveau $\alpha_i^{\gamma,j}$. Nous faisons ces calculs pour $\gamma \in]-1, 1[$ avec un pas de 0.05. Le plus grand $\alpha_i^{\gamma,j}$ nous fournit la valeur de γ choisie et, par conséquence, celle de $\theta_i^{*,j}$ (voir Figure 1). Quand $\sup_\gamma \alpha_i^{\gamma,j} = \alpha_{i,j-1}$ la procédure prend fin. Cette procédure permet d'estimer une large variétés de déformations contenant les translations, les dilatations ... Il est important de noter que ces déformations sont encadrées par des bornes données par les experts, un pic ne pouvant pas être déplacé de plus de 0.02ppm.

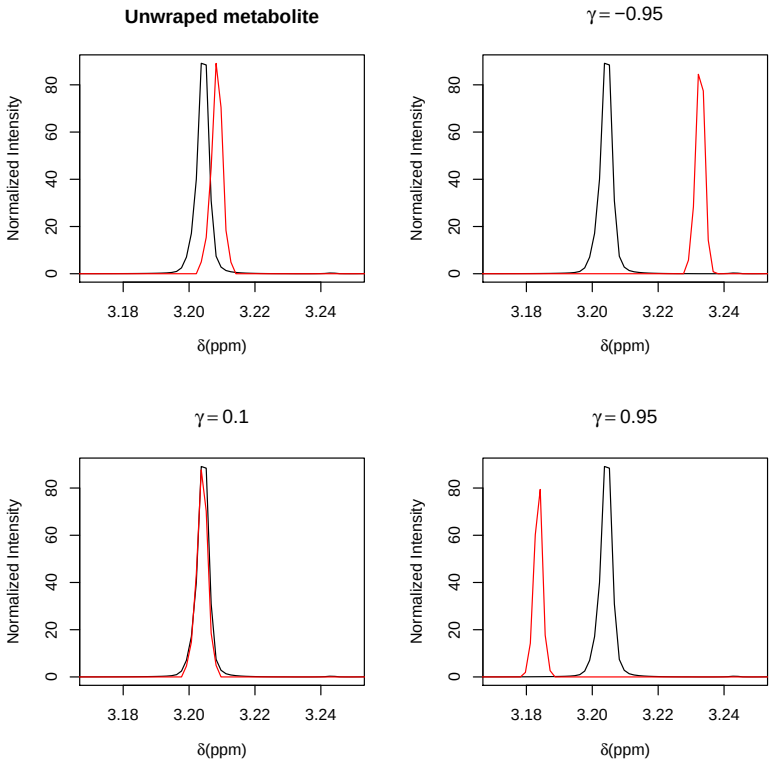

FIGURE 1 – Exemple de la seconde étape, le métabolite est en rouge et le mélange en noir. Le premier graphique est l'état des lieux avant le début de la procédure, ici, clairement, $\alpha_i = 0$ car il y a des points t avec $f_i(t) > 0$ et $G(t) = 0$. Les trois autres graphiques sont des versions déformées du métabolite avec différentes valeurs pour γ. La valeur finalement choisie $\gamma = 0.1$ nous donne un $\alpha_i > 0$.

3.2 Estimation des proportions

Une fois les déformations estimées il nous faut estimer les proportions des métabolites dans le mélange de manière simultanée, donc avec compétition entre les différents métabolites. Il est possible de ramener le problème posé plus haut à un problème de programmation linéaire (Boyd and Vandenberghe, 2004) pour lequel des algorithmes performants existent. Ces algorithmes permettent de déterminer l'ensemble des α_i tels que $\sum_i \alpha_i$ soit maximale sous les différentes contraintes de notre problème et en utilisant les déformations estimées à l'étape précédente. Ceci nous permet donc de répondre aux attentes des biologistes en réalisant l'identification et la quantification en même temps, un métabolite i pour lequel $\alpha_i = 0$ étant absent du mélange.

4 Résultats

Pour cette partie nous disposons d'une bibliothèque de 36 métabolites. Notre méthodologie a été testée sur 3 types de mélange complexe : des mélanges simulés à partir des spectres de la bibliothèque puis déformés, deux mélanges connus c'est-à-dire réalisés à partir de proportions connues des 36 métabolites puis, enfin, un mélange complexe de composition inconnue.

Comme espéré, les résultats sur les mélanges simulés donnent d'excellents résultats, les erreurs étant très rares voire anecdotiques. Pour les mélanges connus, les résultats sont également très bons, l'identification des métabolites étant à chaque fois parfaitement exacte. Leur quantification est elle-aussi très bonne bien qu'encore perfectible. En ce qui concerne le temps de calcul, en programmant en parallèle avec le logiciel R et un serveur possédant 24 coeurs, la détermination de la composition d'un mélange complexe dure une quinzaine de secondes. Il est à noter que certaines étapes comme le pré-traitement des spectres n'ont pas à être réitérées pour chaque mélange. Enfin, sur le mélange complexe, les experts s'accordent à dire que nos résultats sont fortement plausibles et ils sont par conséquent enthousiastes sur la future utilisation de cette méthodolgie.

Références

Astle, W. and al. (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.

Bigot, J., Gadat, S., and Loubes, J.-M. (2009). Statistical M-estimation and consistency in large deformable models for image warping. *Journal of Mathematical Imaging and Vision*, 34(3) :270–290.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, New York, NY, USA.

Illig, T. and al. (2010). A genome-wide perspective of genetic variation in human metabolism. *Nature Genetics*, 42(2) :137–141.

Oakman, C. and al. (2011). Uncovering the metabolomic fingerprint of breast cancer. *The International Journal of Biochemistry & Cell Biology*, 43(7) :1010–1020.

Pontolzeau, C. and al. (2010). Targeted projecion NMR spectroscopy for unambiguous metabolic profiling of complex mixtures. *BMC Bioinformatics*, 9 :507.

Tulpan, D. and al. (2011). Metabohunter : an automatic approach for identification of metabolites from ^{1}H-NMR spectra of complex mixtures. *BMC Bioinformatics*, 12(1) :400.