Modélisation de données métabolomiques longitudinales par voies métaboliques

Camille Guilmineau, Marie Tremblay-Franco, Nathalie Vialaneix, Rémi Servien

Modélisation de données métabolomiques longitudinales par voies métaboliques

Camille Guilmineau [1], Marie Tremblay-Franco [2], Nathalie Vialaneix [3], Rémi Servien [1]

[1] *INRAE, Université de Montpellier, LBE, 11100, Narbonne, France,*
{camille.guilmineau, remi.servien}@inrae.fr
[2] *INRAE, Université de Toulouse, Toxalim (Research Centre in Food Toxicology), ENVT, INP-Purpan, UPS, MetaboHUB, 31300, Toulouse, France*
marie.tremblay-franco@inrae.fr
[2] *INRAE, Université de Toulouse, UR MIAT, F-31320, Castanet-Tolosan, France*
nathalie.vialaneix@inrae.fr

Résumé. Cette communication présente une méthode d'analyse de données métabolomiques longitudinales prenant en compte les voies métaboliques. La métabolomique permet de caractériser le profil métabolique d'un organisme à un instant donné, via l'étude de petites molécules appelées métabolites. La métabolomique longitudinale permet donc d'étudier l'évolution des métabolites au cours du temps. Or, des suites réactions chimiques se produisent entre les métabolites, formant des voies métaboliques. La prise en compte des voies métaboliques dans les modèles statistiques doit donc permettre de détecter plus d'effets et de faciliter l'interprétation biologique. Pour cela, nous étendons dans un premier temps l'approche ssClustPA proposée par [Wieder et al., 2022] aux données longitudinales, afin de transformer la matrice des quantifications des métabolites en matrice de scores des voies métaboliques. Nous estimons ensuite des modèles mixtes sur la matrice obtenue à l'étape précédente, permettant ainsi de réaliser la modélisation sur les voies métaboliques.

Mots-clés. modèle mixte, données longitudinales, métabolomique, voies métaboliques

Abstract. This proposal presents a method to analyze longitudinal metabolomics data that takes into account metabolic pathways. Metabolomics is used to characterize metabolic profiles of an organism at a given time, via the study of small molecules called metabolites. Longitudinal metabolomics thus allows the study of the evolution of metabolites over time. But, series of chemical reactions occur between the metabolites, forming metabolic pathways. The inclusion of metabolic pathways in statistical models aims at detecting more effects and at facilitating biological interpretation. To do so, we first extend the ssClustPA approach proposed by [Wieder et al., 2022] to longitudinal data, and transform the metabolite quantifications into pathway scores. Then, we fit mixed models based on these scores, allowing to define models on metabolic pathways.

Keywords. mixed model, longitudinal data, metabolomics, metabolic pathways

1 Contexte

Le développement des méthodes haut-débit permet de produire massivement des données omiques et notamment métabolomiques. La métabolomique par résonance magnétique nucléaire (RMN) permet de caractériser des mélanges complexes composés d'un grand nombre de molécules, appelées métabolites. Les voies métaboliques décrivent les suites de réactions chimiques qui se produisent entre ces métabolites.

Nous nous intéressons ici au suivi du métabolome, c'est-à-dire à l'évolution de l'ensemble des métabolites au cours du temps. Il s'agit d'acquérir des données métabolomiques à plusieurs dates et sur les mêmes échantillons afin d'étudier l'évolution de ces échantillons dans le temps.

La méthode classique pour analyser ce type de données est d'utiliser des approches basées sur le modèle linéaire mixte, comme proposé par [Martin and Govaerts, 2020]. Ce type de modèle est bien adapté aux données répétées car il ne nécessite pas d'indépendance entre les mesures. Il permet d'inclure à la fois des effets fixes et des effets aléatoires. Les effets fixes correspondent à des effets contrôlés et d'intérêt, comme les conditions expérimentales que l'on étudie. Dans les analyses longitudinales, le temps est également un effet fixe. Les effets aléatoires représentent au contraire des effets non contrôlés, généralement déjà présents dans la population étudiée, comme la variabilité liée à l'individu observé.

Nous présentons ici une méthode de modélisation de données métabolomiques longitudinales qui prend en compte les voies métaboliques. L'analyse par voies métaboliques doit permettre de détecter plus d'effets que l'analyse métabolite par métabolite [Subramanian et al., 2005] et de faciliter l'interprétation biologique. L'enjeu est donc d'étendre les approches usuelles afin de construire un modèle mixte basé sur une voie métabolique et non sur un métabolite. Le temps est également inclus dans le modèle sous la forme d'un effet fixe.

Dans la suite, nous présentons la méthode basée sur les modèles mixtes développée par [Martin and Govaerts, 2020] ainsi que l'approche proposée par [Wieder et al., 2022] pour l'analyse des voies métaboliques dans la section 2. Puis, dans la section 3, nous présenterons la méthode que nous proposons pour la modélisation des données métabolomiques longitudinales par voie métabolique. Enfin, dans la conclusion, nous aborderons plusieurs pistes de réflexions et les développements à venir.

2 État de l'art

2.1 LiMM-PCA

La méthode ANOVA–simultaneous components analysis (ASCA) [Smilde et al., 2005] a été développée comme une généralisation de l'ANOVA pour des données multivariées complexes, en particulier pour la métabolomique. ASCA combine l'ANOVA et l'ACP et permet d'interpréter la variance induite par les différents facteurs expérimentaux. Cependant,

cette méthode ne permet pas d'analyser des données déséquilibrées ni d'inclure des effets aléatoires. [Martin and Govaerts, 2020] ont développé LiMM-PCA pour étendre ASCA aux modèles mixtes et lever ces limitations.

Dans la suite, nous noterons X la matrice des quantifications des métabolites de dimensions $n \times m$, avec n le nombre d'individus et m le nombre de métabolites. x_{ij} désigne la quantification du métabolite j pour l'individu i.

Dans LiMM-PCA, la matrice de réponse normalisée X est d'abord transformée par ACP en une matrice X^* de dimension $n \times m^*$ telle que

$$X = X^* Q^{*'} + H^*$$

où m^* est le nombre de composantes principales conservées de l'ACP sur X et Q^* est la matrice des facteurs principaux. Ensuite, un modèle mixte est construit sur chaque composante principale (colonne), x_j^*, de la nouvelle matrice X^* :

$$\forall\, j = 1, ..., m^*, \qquad x_j^* = U\beta_j + D\alpha_j + \epsilon_j$$

où les matrices U et D sont organisées en blocs correspondant aux effets fixes et aux effets aléatoires du modèle, respectivement, β_j et α_j sont les paramètres fixes et aléatoires du modèle et ϵ_j les résidus i.i.d. du modèle. $\alpha_j \sim \mathcal{N}(0, \sigma_r^2 \mathbf{I}_{q_r})$ pour un effet aléatoire r ayant q_r niveaux et $\epsilon_j \sim \mathcal{N}(0, \sigma_{\epsilon_j}^2 \mathbf{I}_n)$. Dans un modèle longitudinal, le temps est l'un des effets fixes et les effets aléatoires sont classiquement liés à l'individu observé. L'estimation de ces m^* modèles résulte en un modèle multivarié estimé, de la forme

$$X^* = U\widehat{\Theta} + D\widehat{\Gamma} + \widehat{E}$$

La matrice de réponse peut ainsi être décomposée en une somme de matrice d'effets fixes et aléatoires

$$X^* = \widehat{M_0} + \sum_{f=1}^{F} \widehat{M_f} + \sum_{r=1}^{R} \widehat{M_r} + \widehat{E}$$

où $\forall\, f = 1, \ldots, F$, $\widehat{M_f} = U_f \widehat{\Theta}_f$ et $\forall\, r = 1, \ldots, R$, $\widehat{M_r} = D_r \widehat{\Gamma}_r$. Les matrices d'effets sont ensuite analysées par ACP. Enfin, la significativité d'un effet fixe ou aléatoire peut être testée en utilisant une extension multivariée du « log likelihood ratio test » (LLR).

2.2 ssClustPA

Alternativement, d'autres auteurs ont proposé des analyses de données métabolomiques non longitudinales avec des approches qui se fondent sur l'analyse des voies métaboliques plutôt que sur les quantifications individuelles des métabolites. Ces approches sont, par contre, combinées avec des tests d'analyse différentielle qui ne permettent pas la prise en compte des mesures répétées des études longitudinales.

Le principe général consiste, à partir de la matrice X de dimensions $n \times m$ qui contient les quantifications des métabolites pour chaque individu, à obtenir une matrice A de dimensions

$n \times p$ contenant des scores des voies métaboliques pour chaque individu, où p est le nombre de voies métaboliques. Les méthodes de « single sample pathway analysis » (ssPA) permettent de réaliser cette transformation. [Wieder et al., 2022] proposent la méthode ssClustPA, basée sur un clustering par la méthode des k-means avec $k = 2$. Cette méthode est adaptée au cas où l'on cherche à comprendre les différences entre deux groupes d'individus, avec par exemple un groupe traitement et un groupe contrôle.

Pour chaque voie métabolique l ($l = 1, \ldots, p$), vu comme un sous-ensemble $\mathcal{M}_l$ de l'ensemble des métabolites $\{1, \ldots, m\}$, on définit

$$Z_l = (X_{ij})_{i=1,\ldots,n,\, j\in\mathcal{M}_l}$$

la matrice des quantifications des métabolites dans la voie l (avec $|\mathcal{M}_l| = m_l$). L'algorithme k-means, avec $k = 2$, est ensuite appliqué sur les lignes de Z pour définir deux groupes d'individus dont on note c_1^l et c_2^l les centroïdes (qui sont donc des vecteurs de $\mathbb{R}^{m_l}$). L'information contenue dans Z_l est alors résumée par

$$a_l = Z_l u_l \qquad \text{où } u_l = c_1^l - c_2^l.$$

Ainsi, $a_l \in \mathbb{R}^n$ correspond aux scores, pour chaque individu, de la voie métabolique l. La matrice générale des scores est alors définie comme $A = [a_1, \ldots, a_p]$. L'idée générale est que la projection de la matrice des quantifications des métabolites sur le vecteur de la différence des centroïdes permet de capter la variation liée à la différence entre les groupes, si celle-ci est importante pour la voie métabolique concernée.

3 Approche proposée

Dans le cas de données longitudinales, le nombre total d'observations est égal à $n \times T$, où T est le nombre de dates auxquelles ont été mesurées les données. La matrice X des quantifications des métabolites est donc de dimension $(n \times T) \times m$.

3.1 Description de la méthode

Nous présentons ici une méthode permettant de modéliser les données métabolomiques longitudinales en tenant compte des voies métaboliques. Pour cela, nous proposons une adaptation de ssClustPA aux données longitudinales en appliquant la méthode par pas de temps. Ainsi, on applique ssClustPA à la matrice $Z_{lt} = (X_{ijt})_{i=1,\ldots,n,\, j\in\mathcal{M}_l,\, t=1,\ldots,T}$ la matrice des quantifications des métabolites dans la voie l pour les observations du pas de temps t. On obtient les scores $a_{lt} \in \mathbb{R}^n$, pour chaque observation, pour la voie métabolique l et le pas de temps t. On note alors a_l, le vecteur des scores pour la voie métabolique l, défini par $a_l = [a_{11}^\top, \ldots a_{1T}^\top]^\top \in \mathbb{R}^{nT}$.

Un modèle mixte est alors estimé pour décomposer les effets du temps et des conditions expérimentales pour chaque voie métabolique. De manière plus précise, on estime :

$$\forall\, l = 1, \ldots, p, \qquad a_l = U\beta_l + D\alpha_l + \epsilon_l$$

avec

- U la matrice des F effets fixes, $U = (1|U_1|U_2|...|U_F)$. Le temps est défini comme l'un des effets fixes ;
- β_l le vecteur des paramètres des effets fixes ;
- D la matrice des R effets aléatoires, $D = (D_1|D_2|...|D_R)$. Un effet aléatoire pourra être l'individu sur lequel est réalisée l'observation ;
- α_l le vecteur des paramètres des effets aléatoires, $\alpha_l \sim \mathcal{N}(0, \sigma_r^2 \mathbf{I}_{q_r})$ pour un effet aléatoire r ayant q_r niveaux ;
- $\epsilon_l \sim \mathcal{N}(0, \sigma_{\epsilon_l}^2 \mathbf{I}_{n \times T})$ les résidus du modèle, supposés i.i.d.

3.2 Mise en œuvre

L'approche que nous proposons nécessite d'appliquer des pré-traitements aux données avant de pouvoir être mise en œuvre.

3.2.1 Quantification des métabolites

La métabolomique par résonnance magnétique nucléaire (RMN) produit des signaux, appelés spectres, dans lesquels l'axe des abscisses représente le déplacement chimique de la molécule et l'axe des ordonnées indique son intensité. La courbe forme des pics dont l'aire sous la courbe est proportionnelle à la quantité de métabolites dans l'échantillon. La RMN produit un spectre par échantillon, ainsi tous les métabolites contenus dans cet échantillon se trouvent sur le même spectre. Il est donc nécessaire d'utiliser une méthode de déconvolution, qui identifie et quantifie les métabolites présents dans ces spectres. Cela permet d'obtenir les résultats de RMN sous forme de matrice des quantifications, sur laquelle il est possible d'appliquer la méthode présentée précédemment.

Pour cela, nous proposons d'utiliser la méthode ASICS proposée par [Tardivel et al., 2017]. Cette méthode effectue l'identification et la quantification automatique des métabolites dans un spectre à partir d'une librairie de spectres purs, c'est-à-dire de spectres correspondant à un seul métabolite. Elle est implémentée dans le package R **ASICS** [Lefort et al., 2019], qui inclut également les pré-traitements nécessaires à la transformation des signaux bruts de RMN en spectres.

3.2.2 Voies métaboliques

L'approche que nous proposons nécessite également de connaître les voies métaboliques auxquelles appartiennent les métabolites identifiés précédemment. Pour cette étape, il est possible d'utiliser le package R **MetaboAnalystR** [Chong and Xia, 2018] qui permet de requêter une base de voies métaboliques spécifiques d'un organisme donné.

4 Conclusion

La méthode présentée ici est actuellement en cours d'implémentation et sera mise en œuvre pour étudier la formation des photogranules. Les photogranules sont des aggrégats de micro-organismes variés qui présentent des propriétés intéressantes pour le traitement des eaux usées. Nous souhaitons étudier leur développement au cours du temps, avec des données métabolomiques longitudinales. Ces données seront mesurées dans différentes conditions expérimentales. L'objectif est d'identifier les voies métaboliques impliquées dans le développement des photogranules et les périodes temporelles importantes.

Par ailleurs, plusieurs aspects de la méthode nécessitent d'être approfondis. Par exemple, il est courant que des voies métaboliques se chevauchent, c'est-à-dire qu'elles soient composés de métabolites communs. Les voies métaboliques, retrouvées à partir des métabolites identifiées dans les données, sont également de tailles très variables. Il sera donc nécessaire d'étudier l'importance de ces caractéristiques sur la qualité des résultats. De plus, afin d'améliorer la modélisation du temps, nous envisageons d'introduire une structure temporelle dans le modèle mixte (càd de contraindre la structure de dépendance des résidus qui ne seraient alors plus i.i.d.).

Bibliographie

[Chong and Xia, 2018] Chong, J. and Xia, J. (2018). MetaboAnalystR : an R package for flexible and reproducible analysis of metabolomics data. *Bioinformatics*, 34(24) :4313–4314.

[Lefort et al., 2019] Lefort, G., Liaubet, L., Canlet, C., Tardivel, P., Père, M.-C., Quesnel, H., Paris, A., Iannuccelli, N., Vialaneix, N., and Servien, R. (2019). ASICS : an R package for a whole analysis workflow of 1D 1H NMR spectra. *Bioinformatics*, 35(21) :4356–4363.

[Martin and Govaerts, 2020] Martin, M. and Govaerts, B. (2020). LiMM-PCA : combining ASCA+ and linear mixed models to analyse high-dimensional designed data. *Journal of Chemometrics*, 34(6) :e3232.

[Smilde et al., 2005] Smilde, A. K., Jansen, Jeroen, j., Hoefsloot, H. C., Lamers, R.-J. A., van der Greef, J., and Timmerman, M. E. (2005). ANOVA-simultaneous component analysis (ASCA) : a new tool for analyzing designed metabolomics data. *Bioinformatics*, 21(13) :3043–3048.

[Subramanian et al., 2005] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis : a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43) :15545–15550.

[Tardivel et al., 2017] Tardivel, P. J., Canlet, C., Lefort, G., Tremblay-Franco, M., Debrauwer, L., Concordet, D., and Servien, R. (2017). ASICS : an automatic method for identification and quantification of metabolites in complex 1D 1H NMR spectra. *Metabolomics*, 13(10) :109.

[Wieder et al., 2022] Wieder, C., Lai, R. P. J., and Ebbels, T. M. D. (2022). Single sample pathway analysis in metabolomics : performance evaluation and application. *BMC Bioinformatics*, 23(1) :481.