Analyse différentielle longitudinale des voies métaboliques

Camille Guilmineau [1], Rémi Servien [1], Marie Tremblay-Franco [2,3], Nathalie Vialaneix [4]

[1] *INRAE, Université de Montpellier, LBE, F-11100, Narbonne, France.*
{*camille.guilmineau, remi.servien*}*@inrae.fr*
[2] *INRAE, Université de Toulouse, ENVT, Toxalim, Toulouse F-31027, France.*
[3] *Axiom Platform, MetaToul-MetaboHUB, National Infrastructure for Metabolomics and Fluxomics, Toulouse F-31027, France.*
marie.tremblay-franco@inrae.fr
[4] *Université de Toulouse, INRAE, UR MIAT, Castanet-Tolosan F-31326, France.*
nathalie.vialaneix@inrae.fr

Résumé. La métabolomique permet de décrire le profil métabolique d'un organisme à un instant donné en étudiant les quantités de métabolites, qui sont des molécules de petite taille. Ces métabolites participent au fonctionnement moléculaire des organismes vivants (ou des conglomérats de micro-organismes) au travers de réactions chimiques auxquelles ils participent et les voies métaboliques sont formées par des suites de réactions chimiques impliquant certains métabolites pour une fonction donnée de l'organisme. Ainsi, prendre en compte les voies métaboliques dans les modèles statistiques peut permettre de détecter plus d'effets et de faciliter l'interprétation biologique. Nous nous intéressons ici à l'étude de l'évolution temporelle des métabolites dans un contexte d'analyse différentielle (ou cette évolution est influencée par un facteur d'intérêt) et nous présentons une méthode d'analyse différentielle qui se positionne au niveau de la voie métabolique. Cette méthode comporte deux étapes : la matrice des quantifications des métabolites est d'abord transformée par ACP, puis, un modèle linaire mixte est estimé sur les données transformées. Cette méthode a été appliquée sur des données semi-synthétiques et les résultats ont été comparés à ceux obtenus avec la méthode de référence, l'analyse d'enrichissement. On constate que notre proposition détecte mieux les voies métaboliques différentielles que l'analyse d'enrichissement avec un taux de faux positifs plus faible. La méthode est en cours d'implémentation dans le package R **PHOENICS**.

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

Abstract. Metabolomics describes the metabolic profile of an organism at a given time by studying the quantities of metabolites, which are small molecules. These metabolites are involved in the molecular functioning of living organisms (or conglomerates of micro-organisms) through the chemical reactions in which they participate, and metabolic pathways are formed by sequences of chemical reactions involving certain metabolites for a given function in the organism. Thus, taking metabolic pathways into account in statistical models can help detect more effects and facilitate biological interpretation. We are interested here in studying the temporal evolution of metabolites in a differential analysis context (where this evolution is influenced by a factor of interest), and we present a differential analysis method

that is positioned at the pathway level. This methods involves two steps: first, the matrix of metabolite quantifications is transformed by PCA, then a mixed linear model is estimated on the transformed data. This method was applied on semi-synthetic data and the results were compared with those obtained using reference method, namely enrichment analysis. It was found that our proposal detects differential metabolic pathways better than enrichment analysis with a lower false positive rate. This method is currently being implemented in the R package **PHOENICS**.

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

1 Introduction

La métabolomique consiste à détecter et à quantifier les molécules de petite taille, appelées métabolites, présentes dans des mélanges complexes. Des suites de réactions chimiques se produisent entre les métabolites et sont regroupées dans des voies métaboliques, qui permettent la réalisation d'une fonction utile au système biologique. Des méthodes, présentées dans [Tardivel et al., 2017], permettent de quantifier les métabolites individuellement dans des échantillons biologiques. Ces méthodes haut-débit produisent des données de grande dimension et il est donc nécessaire, pour les analyser, de réduire le nombre de variables largement supérieur au nombre d'individus.

Nous nous intéressons ici au suivi du métabolome, c'est-à-dire à l'évolution de l'ensemble des métabolites d'un système biologique au cours du temps. Dans ce contexte, les données métabolomiques sont acquises à plusieurs dates et sur les mêmes individus, afin d'étudier l'évolution dans le temps du métabolome de ces individus.

Une approche courante pour analyser ce type de données consiste à se baser 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 également d'inclure à la fois des effets fixes et aléatoires. Les effets fixes correspondent à des effets contrôlés et d'intérêt, tels que les conditions expérimentales étudiées ou le temps pour les analyse longitudinales. Au contraire, les effets aléatoires représentent des effets non contrôlés, généralement inhérents à la population étudiée, comme la variabilité entre les individus. Cependant, ces approches ne tiennent pas compte des voies métaboliques.

Aussi, en général, les analyses métabolomiques (comme les tests dans les modèles linéaires mixtes décrits ci-dessus), réalisées pour chaque métabolite individuellement, sont post-traitées avec une approche d'analyse d'enrichissement. Elle permet d'étudier si une voie métabolique est enrichie, c'est-à-dire si elle contient significativement plus de métabolites identifiés par l'analyse primaire qu'au hasard.

Nous présentons ici une méthode de modélisation de données métabolomiques longitudinales par voie métabolique. L'analyse par voies métaboliques doit permettre de détecter plus d'effets que l'analyse métabolite par métabolite, car les métabolites d'une voie sont analysés ensemble, permettant de détecter des effets plus faibles qui ne seraient pas identifiés par

l'analyse individuelle des métabolites. Cela doit également faciliter l'interprétation biologique des résultats. 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.

Dans la suite, nous présenterons la méthode proposée dans la section 2. Les données utilisées seront présentées dans la section 3 et la procédure d'évalutation de la méthode sera décrite dans la section 4. Les résultats seront présentés dans la section 5. Enfin, dans la conclusion, nous aborderons des pistes de réflexion et les développements à venir.

2 Description de la méthode proposée

La matrice de quantification des métabolites, notée X, est de dimension $(n \times T) \times m$, où le nombre total d'observations est égal à $n \times T$, avec n est le nombre d'individus et T le nombre de dates auxquelles ont été mesurées les données, et où m est le nombre de métabolites. On note p le nombre de voies métaboliques contenant ces métabolites. Chaque métabolite appartient à au moins une voie, mais peut aussi être impliqué dans plusieurs voies.

Afin de permettre l'analyse des voies métaboliques, une approche de transformation de la matrice des quantifications des métabolites en une matrice au niveau des voies métaboliques, contenant des scores des voies métaboliques pour chaque individu, a été proposée par [Wieder et al., 2022]. Cependant ce type d'approche ne permet pas la prise en compte des mesures longitudinales.

La méthode proposée est découpée en deux étapes :

1. La première étape consiste à transformer la matrice X des quantifications des métabolites. Pour cela, pour chaque voie métabolique $\mathcal{M}_l$, une ACP est réalisée sur la matrice $Z_l = (X_{ij})_{i=1,...,n,\, j \in \mathcal{M}_l}$, la matrice des quantifications des métabolites de la voie métabolique $\mathcal{M}_l$. Les m_l^* premières composantes principales sont sélectionnées par un critère défini préalablement. Nous choisissons ici de sélectionner autant de composantes principales que de facteurs d'intérêts dans le design expérimental. La voie métabolique est alors représentée par les coordonnées des individus sur ces m_l^* composantes principales stockées dans la matrice A_l avec $(n \times T)$ lignes et m_l^* colonnes. Dans la suite, on notera, de manière générique, a, une des colonnes d'une des matrices A_l, qui correspond donc aux coordonnées sur une des composantes principales.

2. La deuxième étape consiste à estimer un modèle mixte à partir de cette nouvelle matrice, pour décomposer les effets du temps et des conditions expérimentales pour chaque voie métaboliques. Pour cela, on estime le modèle suivant :

$$a = U\beta + D\alpha + \epsilon$$

avec

- U la matrice des F effets fixes, $U = (1|U_1|U_2|...|U_F)$, le temps étant défini comme l'un des effets fixes ;

- β 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 le vecteur des paramètres des effets aléatoires, $\alpha \sim \mathcal{N}(0, \sigma_r^2 \mathbf{I}_{q_r})$ pour un effet aléatoire r ayant q_r niveaux ;
- $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2 \mathbf{I}_{n \times T})$ les résidus du modèle, supposés i.i.d.

La significativité des effets fixes est testée par ANOVA en comparant le modèle complet à un modèle restreint, de la forme

$$a = U_f \beta_f + D\alpha + \epsilon$$

pour un effet f, où U_f est une sous-matrice de la matrice des effets fixes U ne contenant pas l'effet f.

Pour une voie métabolique donnée, $\mathcal{M}_l$, cette estimation est répétée pour chacune des colonnes de A_l, ce qui conduit à l'obtention de m_l^* p-valeurs. La procédure de Simes [Simes, 1986] est utilisée pour agréger, par voie métabolique, ces m_l^* p-valeurs : cette procédure contrôle l'erreur de type I de l'hypothèse nulle $\mathrm{H}_0 = \cap_{j=1}^{m_l^*} \mathrm{H}_{0j}$ où H_{0j} est la nullité de l'effet f pour la j-ème colonne de A_l. Ainsi, une unique p-valeur par voie métabolique est obtenue.

Cette méthode est en cours d'implémentation dans un package R nommé **PHOENICS** que nous présenterons lors des Journées de Statistique.

3 Données

Dans le but de tester les capacités de détection de la méthode développée, nous souhaitons utiliser un jeu de données maîtrisé. Pour cela, nous avons créé un jeu de données métabolomiques semi-synthétiques en utilisant la procédure présentée par [Wieder et al., 2022], adaptée aux données longitudinales, et en nous basant sur des données réelles. Ces données proviennent de l'article de [Choo et al., 2017] et sont accessibles dans le dépôt de données métabolomiques MetaboLights (identifiant MTBLS422). Elles contiennent des données de métabolomiques obtenues par résonance magnétique nucléaire (RMN) générées à partir d'échantillons issus d'une étude sur l'effet d'antibiotiques sur des souris. Deux traitements antibiotiques (utilisant de la ciprofloxacin ou du vancomycin-imipenem) ont été comparés à une condition contrôle mais nous nous limiterons à l'utilisation des traitements vancomycin-imipenem et contrôle. Pour chacune des conditions, des mesures ont été réalisées à 3 dates sur 8 souris. Le jeu de données contient ainsi 68 échantillons (4 échantillons sont manquants).

Nous avons traité ces données avec le package R **ASICS** [Lefort et al., 2019] pour obtenir les quantifications des métabolites dans ces échantillons. Les voies métaboliques ont été retrouvées avec le package R **MetaboAnalystR** [Chong and Xia, 2018], qui s'appuie sur la base de données KEGG [Kanehisa and Goto, 2000] spécifique pour l'organisme *Mus musculus* (souris).

Pour créer ces données semi-synthétiques il faut tout d'abord supprimer le signal dans les données réelles. Dans ce but nous avons tout d'abord permuté les échantillons pour les affecter aléatoirement à un groupe (vancomycin-imipenem ou contrôle) et à une date (5.5, 7.5 et 9). Nous avons ensuite choisi $k = 3$ voies métaboliques aléatoirement dans lesquelles nous introduisons une différence entre dates (mais pas entre groupes). Pour cela, les quantifications des métabolites de ces k voies $\{\mathcal{M}_1 \cup \ldots \cup \mathcal{M}_k\}$ sont modifiées en les multipliant par une constante α_t, selon la date t :

$$Y_{tij} = X_{tij} \times \alpha_t$$

avec $j \in \{\mathcal{M}_1 \cup \ldots \cup \mathcal{M}_k\}$, X_{tij} la matrice des quantifications et Y_{tij} la matrice des données semi-synthétiques.

Plusieurs scénarios avec différents α_t ont été testés :

Scénario 1 : $\begin{cases} \alpha_t = 1 & \text{si } t = 5.5, \\ \alpha_t = 5 & \text{si } t = 7.5, \\ \alpha_t = 10 & \text{si } t = 9, \end{cases}$

Scénario 2 : $\begin{cases} \alpha_t = 1 & \text{si } t = 5.5, \\ \alpha_t = 2 & \text{si } t = 7.5, \\ \alpha_t = 3 & \text{si } t = 9, \end{cases}$

Scénario 3 : $\begin{cases} \alpha_t = 1 & \text{si } t = 5.5, \\ \alpha_t = 1.5 & \text{si } t = 7.5, \\ \alpha_t = 2 & \text{si } t = 9. \end{cases}$

Afin de permettre l'évaluation de la qualité de la méthode, le tirage aléatoire des 3 voies métaboliques qui sont simulées différentielles a été répété 100 fois.

4 Évaluation de la méthode

4.1 Comparaison avec les méthodes existantes

Afin d'évaluer notre approche, nous l'avons comparée avec la référence pour l'analyse de voies métaboliques : l'analyse d'enrichissement, qui est basée sur un test exact de Fisher [Wieder et al., 2021]. Une analyse individuelle des métabolites a d'abord été faite en estimant, pour chaque métabolite, un modèle mixte à partir de la matrice des quantifications des métabolites (avec comme effet fixe la date et le traitement et comme effet aléatoire l'individu) puis en testant la significativité des effets fixes. Les métabolites significatifs constituent les métabolites d'intérêt. Pour l'analyse d'enrichissement, un ensemble de métabolites de référence doit également être défini. Il contient généralement l'ensemble des métabolites qui peuvent être détectés dans l'expérimentation. La définition de l'ensemble de référence a un impact important sur les résultats, comme cela a été mis en évidence dans [Wieder et al., 2021], car utiliser un ensemble de référence non spécifique peut mener à un grand nombre de faux positifs. Nous avons testé ici deux ensembles de référence : le premier est constitué de l'ensemble des métabolites de la base de données KEGG, le deuxième est constitué de l'ensemble des métabolites identifiables, ce qui correspond aux métabolites de la base de données du pa-

ckage **ASICS**, utilisé pour identifier les métabolites (soit 180 métabolites). Les deux analyses d'enrichissement ont été réalisées avec le package R **MetaboAnalystR**.

4.2 Évaluation de la qualité de la méthode

La méthode a été évaluée à partir des résultats obtenus sur les données semi-synthétiques, avec la méthode que nous proposons et avec les deux tests d'enrichissement. Pour cela, nous avons classé les voies métaboliques en catégories, en fonction de si elles sont significatives et si elles ont été simulées différentielles, comme présenté dans la Table 1. Cependant, les métabolites des voies simulées différentielles peuvent également appartenir à d'autres voies, à cause du chevauchement entre les voies. Les voies chevauchant les voies simulées différentielles ont donc une partie de leurs métabolites qui ont été simulés comme différentiels. Il est donc difficile de conclure pour ces voies car elles ne peuvent pas être considérées complètement comme des voies non différentielles. Elles ont donc été classées dans une catégorie spécifique.

	Significative	Non significative
Differentielle	Vrai positifs	Faux négatifs
Non differentielle	Faux positifs	Vrai negatifs
Non differentielle (chevauchement)	« Faux positifs » (chevauchement)	« Faux négatifs » (chevauchement)

TABLE 1 : Catégories des voies métaboliques.

Le nombre de voies métaboliques dans chaque catégorie a ensuite été compté sur l'ensemble des répétitions de simulation, pour chacune des méthodes comparées.

5 Résultats

Dans le cas des scénarios de simulation 1 et 2, la méthode que nous proposons détecte un plus grand nombre de vrai positifs que les deux tests d'enrichissement. C'est-à-dire que notre méthode retrouve mieux que l'enrichissement les voies métaboliques qui sont effectivement différentielles. Le nombre de faux positifs détectés par notre méthode est également plus faible qu'avec les tests d'enrichissement. Pour les voies qui ont un chevauchement avec les voies différentielles, notre méthode les détecte plus fréquemment significatives que les tests d'enrichissement.

Dans le cas du scénario 3, où les différences sont plus faibles, très peu de voies métaboliques différentielles sont retrouvées par les trois méthodes. Le nombre de faux positifs est également très faible.

6 Conclusion

Plusieurs aspects de la méthode nécessiteraient d'être approfondis. Le chevauchement entre les voies métaboliques fait qu'il est difficile de conclure pour ces voies. Il serait intéressant de mieux les étudier et de les caractériser pour comprendre pourquoi certaines voies sont significatives et d'autres non. La calibration du nombre de composantes principales retenues après l'ACP est également un point à approfondir. Ce critère doit permettre de conserver la variabilité dans les données tout en limitant le bruit. Enfin, à plus long terme, nous souhaitons étendre la méthode à l'intégration de données multi-omiques.

D'un point de vue applicatif, la méthode présentée ici sera utilisée pour étudier la formation des photogranules. Les photogranules sont des aggrégats de divers micro-organismes qui présentent des propriétés intéressantes pour le traitement des eaux usées. L'objectif est d'utiliser des données métabolomiques longitudinales afin d'étudier leur développement au cours du temps, ainsi que dans différentes conditions expérimentales. Cela doit permettre d'identifier les voies métaboliques impliquées dans le développement des photogranules et les périodes temporelles importantes.

7 Remerciements

Cette recherche a été financée par l'Agence Nationale de la Recherche (ANR) au titre du projet ANR-21-CE45-0036-01.

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.

[Choo et al., 2017] Choo, J. M., Kanno, T., Zain, N. M. M., Leong, L. E. X., Abell, G. C. J., Keeble, J. E., Bruce, K. D., Mason, A. J., and Rogers, G. B. (2017). Divergent relationships between fecal microbiota and metabolome following distinct antibiotic-induced disruptions. *mSphere*, 2(1):10.1128/msphere.00005–17.

[Kanehisa and Goto, 2000] Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30.

[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.

[Simes, 1986] Simes, R. J. (1986). An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754.

[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., 2021] Wieder, C., Frainay, C., Poupin, N., Rodríguez-Mier, P., Vinson, F., Cooke, J., Lai, R. P., Bundy, J. G., Jourdan, F., and Ebbels, T. (2021). Pathway analysis in metabolomics: Recommendations for the use of over-representation analysis. *PLOS Computational Biology*, 17(9):e1009105.

[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.