

phoenics : UNE NOUVELLE APPROCHE STATISTIQUE POUR L'ANALYSE DIFFÉRENTIELLE LONGITUDINALE DE VOIES MÉTABOLIQUES

Camille Guilmineau¹, Rémi Servien¹, Marie Tremblay-Franco^{2,3}, Nathalie Vialaneix⁴

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 décrit le profil métabolique d'un organisme par les concentrations des métabolites qui le compose à un instant donné. Lorsqu'elles sont étudiées au fil du temps, les concentrations des métabolites peuvent aider à comprendre l'évolution dynamique d'un processus biologique. Cependant, ces métabolites sont impliqués dans des séquences de réactions chimiques, appelées voies métaboliques, liées à une fonction biologique donnée. La prise en compte de ces voies dans les méthodes statistiques d'analyse des données métabolomiques est donc un moyen pertinent d'exprimer directement les résultats en termes de fonctions biologiques et d'augmenter leur interprétabilité. Nous proposons une nouvelle méthode, **phoenics**, qui effectue une analyse différentielle des données métabolomiques longitudinales au niveau des voies. En bref, **phoenics** procède en deux étapes : tout d'abord, la matrice de quantification des métabolites est transformée par une approche de réduction de dimension prenant en compte les informations sur les voies. Ensuite, un modèle linéaire mixte est ajusté sur les données transformées. Cette méthode a été appliquée à des données semi-synthétiques et à un jeu de données réelles évaluant les effets des antibiotiques sur les matières fécales. Les résultats ont montré que **phoenics** contrôle correctement le taux d'erreur de type I et a une meilleure capacité à détecter les voies métaboliques différentielles et à extraire de nouvelles fonctions biologiques impactées que les méthodes alternatives. La méthode est implémentée dans le package **R phoenics** disponible sur CRAN.

Mots-clés. modèles à effets mixtes, données longitudinales, métabolomique, voies métaboliques.

Abstract. Metabolomics describes the metabolic profile of an organism at a given time by the concentrations of its constituent metabolites. When studied over time, metabolite concentrations can help understand the dynamical evolution of a biological process. However, metabolites are involved into sequences of chemical reactions, called metabolic pathways, related to a given biological function. Accounting for these pathways into statistical methods for metabolomic data is thus a relevant way to directly express results in terms of biological functions and to increase their interpretability. We propose a new method, **phoenics**, to

perform differential analysis for longitudinal metabolomic data at the pathway level. In short, **phoenics** proceeds in two steps: First, the matrix of metabolite quantifications is transformed by a dimension reduction approach accounting for pathway information. Then, a mixed linear model is fitted on the transformed data. This method was applied to semi-synthetic data and a real dataset assessing effects of antibiotic on feces. Results showed that **phoenics** properly controls the Type I error rate and has a better ability to detect differential metabolic pathways and to extract new impacted biological functions than alternative methods. The method is implemented in the R package **phoenics** available on CRAN.

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

1 Introduction

Les données métabolomiques correspondent à la quantification de la concentration de petites molécules, appelées métabolites, présentes dans des mélanges complexes à un instant donné. De plus en plus d'expériences ciblent l'évolution du métabolome au cours du temps, dans différentes conditions, ou suite à un événement d'intérêt (Mäkinen and al., 2022). Dans ces cas, les données métabolomiques sont acquises à plusieurs moments sur les mêmes individus. Nous abordons ici la question de l'utilisation de données métabolomiques temporelles acquises dans plusieurs conditions expérimentales pour extraire des biomarqueurs soit entre différentes conditions, soit entre différents temps de mesure.

Les métabolites sont cependant regroupés en séquences de réactions chimiques, appelées voies métaboliques, dans lesquelles ils sont impliqués et qui partagent une fonction biologique commune. La prise en compte des voies dans les méthodes statistiques pour les données métabolomiques est donc un bon moyen d'exprimer directement les résultats en termes de fonctions biologiques et d'augmenter leur interprétabilité. Habituellement, cela se fait par analyse d'enrichissement, une méthode qui consiste à effectuer des tests indépendamment pour chaque métabolite et à combiner les résultats de ces tests pour extraire des voies contenant plus de métabolites différentiels que ce qui aurait été attendu par hasard (Khatri and al., 2012). Un inconvénient est que cette approche est basée sur la puissance d'une analyse primaire où les métabolites sont analysés de manière indépendante est donc incapable de détecter des voies qui correspondent à une accumulation de petits effets.

D'autres méthodes, dites de notation de classe fonctionnelle (FCS), ont également été utilisées pour effectuer des analyses différentielles. La plupart de ces méthodes ont été développées dans le domaine de la transcriptomique mais elles peuvent être facilement étendues à l'analyse des voies métabolomiques (Maleki and al., 2020). Dans le domaine de la métabolomique, Wieder and al. (2022, 2024) développent une de ces méthodes, où les données au niveau du métabolite sont transformées en données au niveau de la voie avant qu'une analyse différentielle soit réalisée sur cette transformation. Aussi, Ozier-Lafontaine and al. (2024) proposent un test non linéaire pour les ensembles de gènes basé sur une approche par noyau. Dans son implémentation actuellement disponible, le test est cependant limité aux différences de test entre deux conditions sans covariables.

Nous présentons ici une extension de la méthode de Wieder and al. (2022) pour per-

mettre l'analyse de mesures longitudinales répétées. Notre méthode, **phoenics**, procède en deux étapes : une transformation des données similaire à Wieder and al. (2022) et un modèle mixte. La méthode est implémentée dans le package R **phoenics** disponible sur le CRAN¹.

2 Méthode

Soit X une matrice ($nT \times m$) de quantifications, où n est le nombre d'individus, T le nombre de points dans le temps et m le nombre de métabolites. X est organisée de telle sorte que ses n premières lignes correspondent aux mesures de quantification des n individus au premier point dans le temps. Chaque métabolite appartient à une ou plusieurs voies métaboliques $\mathcal{M}_l$. La méthode proposée est composée de deux étapes :

1. **Représentation par analyse factorielle pour extraire les signatures des voies.** Pour chaque voie métabolique $\mathcal{M}_l$, et de manière similaire à Wieder and al. (2022), une ACP est réalisée sur la matrice de quantification des métabolites de cette voie, $Z_l = (X_{ij})_{i=1,\dots,nT, j \in \mathcal{M}_l}$. Les scores des m_l^* premières composantes principales sont extraits dans la $(nT \times m_l^*)$ -matrice A_l . Dans ce qui suit, pour simplifier les notations, nous ferons référence génériquement à a comme l'une des colonnes d'une des matrices A_l , qui correspond à l'une des composantes principales pour une voie donnée.
2. **Modèle mixte pour tester les voies.** Pour chaque composante principale a , le modèle mixte suivant est alors estimé :

$$a = \beta_t + \theta_d + u_i + \epsilon_{tdi} \quad (1)$$

avec

- β_t l'effet fixe du temps t ;
- θ_d l'effet fixe de la condition d ;
- $u_i \sim \mathcal{N}(0, G)$ l'effet aléatoire de l'individu i , avec G la matrice de covariance ;
- $\epsilon_{tdi} \sim \mathcal{N}(0, \sigma_\epsilon^2)$ les résidus du modèle, qui sont supposés être i.i.d.

La signification d'un effet fixe donné f est ensuite testée à l'aide du test du rapport de vraisemblance comparant le modèle complet de l'équation (1) à un modèle restreint qui ne contient pas l'effet f .

Pour une voie métabolique donnée $\mathcal{M}_l$, cette estimation est répétée pour chaque colonne de A_l , ce qui donne m_l^* p -valeurs. La procédure de Simes (1986) est utilisée pour agréger les m_l^* p -valeurs : cette procédure contrôle l'erreur de type I de l'hypothèse nulle $H_0 = \cap_{j=1}^{m_l^*} H_{0j}$, où H_{0j} est l'égalité des moyennes dans les différents niveaux de f pour la j ème colonne de A_l . En conséquence, cette approche nous conduit à obtenir une seule valeur p par voie métabolique pour chaque effet testé f . Enfin, comme plusieurs valeurs p sont obtenues (une pour chaque voie), une correction supplémentaire par test multiple est nécessaire. Nous avons choisi de contrôler le FDR en utilisant la procédure de Benjamini-Hochberg (BH).

¹<https://CRAN.R-project.org/package=phoenics>

En pratique, la méthode nécessite seulement de choisir le paramètre m_l^* , correspondant au nombre de composantes principales sur lesquelles les modèles mixtes sont estimés pour chaque voie. Pour permettre à la méthode de conserver suffisamment d’informations dans les données pour estimer chaque effet possible, nous proposons de fixer m_l^* égal au nombre d’effets fixes, F .

Une limitation de cette approche est que l’ACP ne tient pas compte de la structure spécifique des données où certaines observations d’un individu donné peuvent être plus corrélées que d’autres. Ainsi, nous proposons également de remplacer l’ACP globale par des analyses ACP indépendantes, dans le même esprit que l’analyse multi-tables appelée Analyse Factorielle Multiple (MFA) (Escofier and Pagès, 1994). La matrice de quantification des métabolites de la voie testée est alors vue comme un ensemble de matrices, chacune correspondant à un instant donné, t : $Z_l^{(t)} = (X_{ij})_{i=n(t-1)+1,\dots,nt, j \in \mathcal{M}_l}$ pour $t = 1, \dots, T$. La MFA effectue ensuite une ACP individuelle (dite partielle) de chaque matrice (une ACP globale est aussi disponible, qui correspond à une moyenne pondérée des ACP individuelles). Ici, nous nous en tenons à la première étape de la MFA et récupérons les scores des m_l^* premières composantes principales des analyses partielles pour les empiler dans une $(nT \times m_l^*)$ -matrice A_l . La deuxième étape (correspondant au modèle mixte) est ensuite réalisée de la même manière que l’approche précédente. Cependant, si cette approche paraît plus appropriée pour tester l’effet condition, elle n’est pas pertinente pour l’effet temps, car ce dernier devrait être effacé en effectuant les analyses partielles indépendantes.

3 Résultats

Données : Les performances de la méthode ont été évaluées à l’aide de données expérimentales de Choo and al. (2017), disponibles dans l’entrepôt de données MetaboLights avec l’identifiant MTBLS422. Cette étude examine les changements du métabolisme fécal induits par les antibiotiques. Les données contiennent des spectres de résonance magnétique nucléaire (RMN) obtenus à partir de 2×8 fluides de souris soumises ou non à un traitement antibiotique. Dans la suite, le statut de traitement de la souris est appelé la « condition » et est un effet fixe d’intérêt. De plus, dans chaque condition, des spectres ont été acquis à partir d’échantillons collectés à trois moments et l’effet fixe temporel est donc un deuxième effet fixe d’intérêt.

Les quantifications des métabolites ont été estimées à partir des spectres RMN avec le package **R ASICS** (Lefort and al., 2019). Cela a permis de quantifier 176 métabolites au total. Les voies métaboliques ont été récupérées à l’aide du package **R KEGGREST** qui interroge la base de données KEGG pour les voies spécifiques à un organisme donné (ici, *Mus musculus*), comme implémenté dans notre package **phoenics**. Un nombre total de 98 voies ont été obtenues, contenant au moins un des métabolites quantifiés pour un total de 128 métabolites trouvés dans les voies.

Ce jeu de données a ensuite été utilisé pour générer des jeux de données semi-synthétiques permettant de tester un effet temps, un effet condition et les deux effets simultanément. Ceci avec différentes intensités de l’effet, d’un effet fort (Scenario 1) à un effet faible (Scenario 4).

Les jeux de données semi-synthétiques ont permis de tester 3 caractéristiques des méthodes : le contrôle de l'erreur de type I, la puissance statistique et une analyse de la puissance selon le nombre de métabolites différentiels dans voie testée.

Méthodes : Nous avons comparé **phoenics** à la méthode de référence pour l'analyse des voies métaboliques, qui est l'analyse d'enrichissement. Pour cette approche, nous avons utilisé le même modèle mixte linéaire que dans **phoenics** mais par métabolite et non par voie et avons également corrigé les p -valeurs issues du test de ce modèle par la méthode BH. **phoenics** a aussi été comparé à une méthode FCS multivariée, **ktest**. Comme sa mise en œuvre est actuellement limitée à deux conditions, des comparaisons par paires des trois dates ont été réalisées pour le test de l'effet du temps, en utilisant les deux premières dimensions de l'espace image associé au noyau. Les deux premières dimensions ont également été utilisées pour le test de l'effet de condition. De manière similaire à ce qui est fait dans **phoenics**, la procédure de Simes a été utilisée pour agréger les six ($2 \text{ dimensions} \times 3 \text{ paires de dates}$) ou les deux ($2 \text{ dimensions} \times 1 \text{ paire de conditions}$) valeurs p résultantes. La multiplicité des tests a été prise en compte par la méthode BH. Dans la suite, **phoenicsPCA** fera référence à l'approche basée sur l'ACP alors que l'approche basée sur la MFA sera notée **phoenicsMFA**.

Résultats : Le contrôle de l'erreur de Type I a toujours été assuré pour toutes les méthodes. Pour le jeu de données incluant seulement un effet condition, **phoenicsMFA** s'est avéré plus puissant que **phoenicsPCA** et que les autres méthodes, les différences de performances se creusant quand l'amplitude de l'effet diminue. Cependant, comme supposé, lorsqu'il y a seulement un effet temps **phoenicsMFA** (tout comme **ktest**) ne détecte pas cet effet. **phoenicsPCA** est alors plus performant que l'analyse d'enrichissement et la différence de performances est également plus forte quand l'effet est faible.

Pour le jeu de données comprenant les 2 effets, le test de l'effet temps a amené aux mêmes conclusions que précédemment. Le test de l'effet condition est plus intéressant, **phoenicsPCA** a une sensibilité plus élevée que **phoenicsMFA** et que l'analyse d'enrichissement dans tous les scénarios (Figure 1). De plus, **phoenicsPCA** présente un nombre plus faible de faux positifs par rapport à **phoenicsMFA** et a donc une Valeur Prédictive Positive (VPP) plus élevée. Aussi, **ktest** présente une sensibilité et une VPP élevées dans le scénario 1 mais ne détecte plus de positif lorsque la taille de l'effet diminue dans les autres scénarios.

L'application de **phoenics** sur le jeu de données réelles a ensuite permis de détecter les voies pertinentes pour l'effet des antibiotiques sur la métabolomique des matières fécales.

4 Conclusion

Nous présentons ici une approche permettant de réaliser une analyse différentielle longitudinale des voies métaboliques, disponible dans le package **R phoenics**. La méthode est basée sur l'analyse factorielle et des modèles linéaires mixtes pour réaliser l'analyse au niveau des voies et ne nécessite pas d'analyse préalable des métabolites. L'application sur des données a montré que les méthodes **phoenics** détectent les voies différentielles avec une puissance supérieure aux méthodes existantes tout en contrôlant le taux d'erreur de type I.

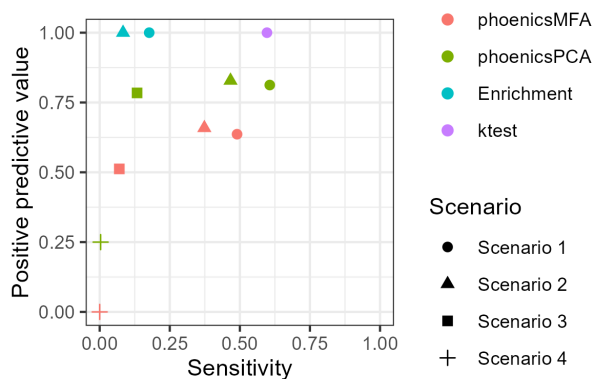


FIGURE 1 : VPP et sensibilité pour tester l'effet condition. L'analyse d'enrichissement pour les Scenarios 3 et 4 et **ktest** pour les Scenarios 2, 3 et 4 n'apparaissent pas car ils ne font pas de détections positives.

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

References

- Choo, J. M. and al. (2017). Divergent relationships between fecal microbiota and metabolome following distinct antibiotic-induced disruptions. *mSphere*, 2(1).
- Escofier, B. and Pagès, J. (1994). Multiple factor analysis (AFMULT package). *Computational Statistics and Data Analysis*, 18(1):121–140.
- Khatri, P. and al. (2012). Ten years of pathway analysis: Current approaches and outstanding challenges. *PLoS Computational Biology*, 8(2):e1002375.
- Lefort, G. and al. (2019). ASICS: an R package for a whole analysis workflow of 1D 1H NMR spectra. *Bioinformatics*, 35(21):4356–4363.
- Maleki, F. and al. (2020). Gene set analysis: challenges, opportunities and future research. *Frontiers in Genetics*, 11:654.
- Mäkinen, V.-P. and al. (2022). Longitudinal profiling of metabolic ageing trends in two population cohorts of young adults. *International Journal of Epidemiology*, 51:1970–1983.
- Ozier-Lafontaine, A. and al. (2024). Kernel-based testing for single-cell differential analysis. *Genome Biology*, 25:114.
- Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754.
- Wieder, C. and al. (2022). Single sample pathway analysis in metabolomics: performance evaluation and application. *BMC Bioinformatics*, 23:481.
- Wieder, C. and al. (2024). PathIntegrate: multivariate modelling approaches for pathway-based multi-omics data integration. *PLoS Computational Biology*, 20(3):e1011814.