# Combinaison de méthodes numériques et symboliques pour l'analyse de données métabolomiques

Dhouha-Grissa*,** Blandine Comte*
Estelle Pujos-Guillot*,*** Amedeo Napoli**

*INRA, UMR1019, UNH-MAPPING, F-63000 Clermont-Ferrand, France
blandine.comte@clermont.inra.fr, estelle.pujos@clermont.inra.fr
**LORIA, B.P. 239, F-54506 Vandoeuvre-lès-Nancy, France
amedeo.napoli@loria.fr, dhouha.grissa@loria.fr
***INRA, Plateforme d'Exploration du Métabolisme, F-63000 Clermont-Ferrand, France

La métabolomique est un outil de phénotypage puissant permettant de mieux comprendre les mécanismes biologiques impliqués dans les processus pathologiques et d'en identifier des marqueurs. Cependant, cette approche génère des données massives et complexes (Sugimoto et al., 2012) qui nécessitent des outils de traitement adaptés afin d'extraire les informations biologiquement pertinentes et enrichir les connaissances et compréhension des systèmes biologiques. Le traitement des données métabolomiques présente un enjeu de taille du fait de leur nature (issues d'un signal instrumental, bruitées, volumineuses avec de fortes colinéarités...) et de leur structure (le nombre de variables est très élevé par rapport au nombre d'individus). Malgré l'existence de nombreux outils chimiométriques (Eliasson et al., 2011), il a été reporté des difficultés de classification dus à la grande dimensionnalité des données. Il y a aujourd'hui un besoin de méthodes et workflows robustes et permettant l'obtention de résultats justes et fiables.

Dans ce travail, notre objectif a consisté en l'élaboration d'un workflow d'analyse de données métabolomiques, combinant plusieurs techniques de fouille de données numériques et symboliques, supervisées et non supervisées, dans le but de proposer une solution avancée pour la découverte de biomarqueurs. Notre démarche s'est basée sur l'extraction, au moyen de méthodes numériques, de classes d'attributs qui ont été par la suite organisées par la technique non supervisée d'Analyse Formelle de Concepts (FCA, (Ganter et Wille, 1999)) pour la visualisation et l'interprétation.

Le travail a été divisé en trois étapes : dans la 1ère étape ou prétraitement des données, il s'agissait de choisir les méthodes les plus adéquates à la fois de transformation de données (normalisation, scaling) en fonction de la méthode de sélection de motifs utilisée ; et de réechantillonnage (bootstrap, validation croisée,...) en les évaluant selon des critères de précision, spécificité et sensibilité. La 2ème étape s'est intéressée à la sélection des variables, ou "Feature selection" (Guyon et Elisseeff, 2003). Cette approche, l'une des plus répandues dans le domaine bioinformatique (Zhou et Dickerson, 2014), permet d'identifier des variables significatives (élimination des redondances) qui présentent la meilleure capacité discriminante et prédictive pour la construction du modèle. Dans ce travail, nous avons combiné diverses méthodes de Feature selection. Des méthodes de filtre (utilisant différents scores : information mutuelle et coefficient de corrélation) ont tout d'abord été considérées pour éliminer les variables

redondantes/dépendantes et réduire la dimension de la matrice de données. Puis, des méthodes d'apprentissage supervisée, SVM (Support Vector Machine, ou Machines à vecteurs de support (Corinna et Vladimir, 1995)), RF (Random forest, forêts d'arbres décisionnels (Breiman, 2001)), SVM-RFE (SVM-Recursive Feature Elimination) ont été utilisées afin d'ordonner les variables et sélectionner les plus discriminantes et prédictives. Cette sélection de variables a été basée sur les mesures de précision, les indices "gini" et "kappa", ainsi que le poids "w". De plus, des tests statistiques univariés ont été réalisés. Une étude comparative des $k$ meilleures variables issues de la combinaison de ces différentes approches (au total 10 combinaisons) a permis d'identifier leurs degrés de stabilité (1 à 10). Une matrice binaire a ainsi été construite de la forme (*N variables* $\times$ *10 techniques-d'analyse*) par la méthode de présence/absence des variables. Cette matrice a été le point de départ pour l'application de la FCA et la construction du treillis de concepts. Selon la logique du treillis, les variables partagées par toutes les techniques (stabilité maximale égale à 10) seront des candidats biomarqueurs que nous retiendrons pour l'étape de prédiction. La 3ème étape ou post-traitement de données, s'est concentrée sur la visualisation et l'interprétation des données issues du treillis.

Les résultats obtenus ont pu faciliter l'identification des variables les plus intéressantes et servir de base à l'interprétation par les spécialistes du domaine. Un workflow enchainant la meilleure combinaison de méthodes numériques et symboliques pourra être proposé à la communauté du domaine.

# Références

Breiman, L. (2001). Random forests. In *Machine Learning*, pp. 5–32.

Corinna, C. et V. Vladimir (1995). Support-vector networks. *Mach. Learn. 20*(3), 273–297.

Eliasson, M., S. Rannar, et J. Trygg (2011). *Current Pharmaceutical Biotechnology 12*(7).

Ganter, B. et R. Wille (1999). *Formal concept analysis - mathematical foundations*. Springer.

Guyon, I. et A. Elisseeff (2003). *J. Mach. Learn. Res. 3*, 1157–1182.

Sugimoto, M., M. Kawakami, M. Robert, T. Soga, et M. Tomita (2012). *Curr. Bioinform 7*(1), 96–108.

Zhou, W. et J. Dickerson (2014). *Comput. Biol. Med. 47*, 66–75.

# Summary

Our work consists in developing a workflow using Knowledge Discovery methodologies to propose advanced predictive biomarkers discovery solutions from metabolomic data. We propose to use machine learning algorithms for feature selection and FCA for visualization.