Predictive analysis of metabolomics data in chronic kidney disease using a subgroup discovery algorithm

Margaux Luck¹, Gildas Bertho³, Eva-Eric Thervet², Philippe Beaune², François d’Ormesson¹, Mathilde Bateson¹, Anastasia Yartseva¹, Cécilia Damon¹, Nicolas Pallet²,4,6,7

¹ Hypercube Institute, Paris, France; ² Paris Descartes University, Paris, France; ³ CNRS UMR 8601, Paris, France; ⁴ Magnetic Nuclear Resonance Platform “Metabo Paris-Santé”, CICB-Paris, Paris, France; ⁵ Renal Division, Georges Pompiou European Hospital, Assistance Publique Hôpitaux de Paris, Paris, France; ⁶ INSERM U1147, Paris, France; ⁷ Clinical Chemistry Department, Georges Pompiou European Hospital, Paris, France

¹H Nuclear Magnetic Resonance (NMR) spectroscopy is widely used for the identification of metabolites in biofluids and enables the understanding of the underlying biological mechanisms of a disease. Within the study of chronic kidney disease (CKD), the analysis of metabolomic variations, acquired with ¹H NMR in urine, may be useful for the diagnosis of the different CKD severity stages.

Due to the high dimension of NMR spectra datasets and the complex mixture of metabolites in biological samples including the overlaps of some spectra peaks, the identification and quantification of discriminant biomarkers in the biological context of a disease is a key challenge. Among the widely used chemometric methods in NMR metabolomics, we can mention unsupervised data mining tools such as Principal Component Analysis (PCA) or clustering, and supervised classification methods such as Partial Least Squares (PLS) or Support Vector Machine (SVM). More recently, a local data mining approach, Random Forests, has been applied for NMR prediction [1].

To our knowledge, none of these methods performs local exhaustive exploration of the data. Indeed, they look at the global information of each feature rather than searching for discriminant local phenomena. Random Forest only explores random parts of the feature space and have problems of interpretability. Here, we used for the metabolomics data analysis, a descriptive and easily understandable approach with the HyperCube® software [2]. First, it selects the most discriminant features from the dataset based on both the normalized mutual information between them and the CKD severity variable, and the chi-2 test. Then, it studies the local distribution of the patient subgroups with identical degree of CKD severity on each feature, using the proprietary algorithm, HyperCube. It extracts range or modalities of values of the discriminant explanatory features with respect to each CKD severity stage. Further, logistic regression on these discriminant features was used to build a predictive model of the CKD severity stage.

We studied a database composed of urinary ¹H NMR spectra of a cohort of 110 individuals referred to the Nephrology Department for the evaluation of CKD [3]. For each patient, we have their clinical, demographical, biochemical and histological data and we assessed the CKD disease severity based on the glomerular filtration rate, eGFR>60 ml/min/1.73m² for low to mild CKD and eGFR<60 ml/min/1.73m² for moderate CKD to established kidney failure.

We first showed that the HyperCube algorithm combined with logistic regression supports the discriminant metabolities obtained with standard Orthogonal Projection to Latent Structure Discriminant Analysis (O-PLS-DA) model. Unlike the O-PLS-DA model, HyperCube algorithm provided clues into the distribution of the CKD severity subgroups with respect to spectral data. Then, we took into account other sources of patient information (clinical, demographical, biochemical and histological) that may have an influence on the spectra data. The constructed predictive model included urinary metabolites for patients with advanced-stage CKD and low fibrosis levels for the patients with early-stage CKD.

Bibliographical references


Keywords

Human, Urine, CKD, ¹H NMR, HyperCube, Subgroup Discovery