Journal of Biomedical Informatics 53 (2015) 291–299

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Comparative analysis of targeted metabolomics: Dominance-based rough set approach versus orthogonal partial least square-discriminant analysis

H. Blasco [a,b,c,*], J. Błaszczyński [d], J.C. Billaut [f], L. Nadal-Desbarats [a,b,g], P.F. Pradat [h], D. Devos [i], C. Moreau [i], C.R. Andres [a,b,c], P. Emond [a,b,g], P. Corcia [a,b,j], R. Słowiński [d,e]

[a] Inserm U930, Tours, France
[b] Université François-Rabelais, Tours, France
[c] Laboratoire de Biochimie et Biologie Moléculaire, CHRU de Tours, Tours, France
[d] Institute of Computing Science, Poznań University of Technology, 60-965 Poznań, Poland
[e] Systems Research Institute, Polish Academy of Sciences, 01-447 Warsaw, Poland
[f] Université François-Rabelais de Tours, CNRS, LI EA 6300, OC ERL CNRS 6305, Tours, France
[g] PPF, Université François-Rabelais, Tours, France
[h] Fédération des Maladies du Système Nerveux, Centre Référent Maladie Rare SLA, Hôpital de la Pitié-Salpétrière, Paris, France
[i] Service de Neurologie, CHRU de Lille, Lille, France
[j] Centre SLA, Service de Neurologie, CHRU Bretonneau, Tours, France

ARTICLE INFO

Article history:
Received 21 May 2014
Accepted 1 December 2014
Available online 11 December 2014

Keywords:
Metabolomics
OPLS-DA
Dominance-based rough set approach
Bayesian confirmation
Diagnosis prediction
Amyotrophic lateral sclerosis

ABSTRACT

Background: Metabolomics is an emerging field that includes ascertaining a metabolic profile from a combination of small molecules, and which has health applications. Metabolomic methods are currently applied to discover diagnostic biomarkers and to identify pathophysiological pathways involved in pathology. However, metabolomic data are complex and are usually analyzed by statistical methods. Although the methods have been widely described, most have not been either standardized or validated. Data analysis is the foundation of a robust methodology, so new mathematical methods need to be developed to assess and complement current methods. We therefore applied, for the first time, the dominance-based rough set approach (DRSA) to metabolomics data; we also assessed the complementarity of this method with standard statistical methods. Some attributes were transformed in a way allowing us to discover global and local monotonic relationships between condition and decision attributes. We used previously published metabolomics data (18 variables) for amyotrophic lateral sclerosis (ALS) and non-ALS patients.
Results: Principal Component Analysis (PCA) and Orthogonal Partial Least Square-Discriminant Analysis (OPLS-DA) allowed satisfactory discrimination (72.7%) between ALS and non-ALS patients. Some discriminant metabolites were identified: acetate, acetone, pyruvate and glutamine. The concentrations of acetate and pyruvate were also identified by univariate analysis as significantly different between ALS and non-ALS patients. DRSA correctly classified 68.7% of the cases and established rules involving some of the metabolites highlighted by OPLS-DA (acetate and acetone). Some rules identified potential biomarkers not revealed by OPLS-DA (beta-hydroxybutyrate). We also found a large number of common discriminating metabolites after Bayesian confirmation measures, particularly acetate, pyruvate, acetone and ascorbate, consistent with the pathophysiological pathways involved in ALS.

Abbreviations: ALS, amyotrophic lateral sclerosis; AHBT, α-hydroxybutyrate; CSF, cerebrospinal fluid; BHBT, β-hydroxybutyrate; Crn, creatine–creatinine; DModX, distance to model plot; DRSA, dominance-based rough set approach; Gln, glutamine; NMR, ^{1}H Nuclear Magnetic Resonance; OPLS-DA, Orthogonal Partial Least Squares Discriminant Analysis OPLS-DA; Par, Pareto scaling; PC, principal component; PCA, Principal Component Analysis; NPV, negative predictive value; PPV, predictive positive value; Q^2, indicates how well a variable can be predicted and estimated by cross validation; R^2, indicates how well the variation of a variable is explained; SD, standard deviation; VIP, variable importance parameters.

* Corresponding author at: Laboratoire de Biochimie et Biologie moléculaire, Hôpital Bretonneau, CHRU de Tours, 2, Bd Tonnellé, 37044 – Tours cedex 1, France. Fax: +33 2 47 47 86 13.
E-mail address: helene.blasco@univ-tours.fr (H. Blasco).

http://dx.doi.org/10.1016/j.jbi.2014.12.001

Conclusion: DRSA provides a complementary method for improving the predictive performance of the multivariate data analysis usually used in metabolomics. This method could help in the identification of metabolites involved in disease pathogenesis. Interestingly, these different strategies mostly identified the same metabolites as being discriminant. The selection of strong decision rules with high value of Bayesian confirmation provides useful information about relevant condition–decision relationships not otherwise revealed in metabolomics data.

1. Background

Metabolomics is defined as the study of all metabolites in a system, and includes the identification of metabolic signatures from a combination of small molecules in biological fluid. This emerging "omics" approach is increasingly used in medicine to find diagnostic and prognostic biomarkers of diseases and to identify pathophysiological pathways involved in these pathologies. The methods involve studying metabolites across a large spectrum of concentrations, polarity and masses [1–3], based on high-throughput techniques [4]. "Untargeted" metabolomics is based on metabolic profiles without systematic identification of metabolites included in this profile, and "targeted" methodologies focus on specific metabolites. Whatever the approach, data preprocessing and analysis are the foundation of a robust methodology. There has been increasing work in bioinformatics to try to standardize these steps [5–8]. The greatest risk associated with high-throughput techniques is mishandling multiple and complex data, leading to biased results. Multivariate statistical analysis including Principal Component Analysis (PCA: the aims of which include description) and Orthogonal Partial Least Squares Discriminant Analysis (OPLS-DA: for predictions) are the most widely used statistical methodologies in "omics" studies. However, standard criteria to validate the models are often unavailable. Efforts have to be made to validate the models statistically or to validate this strategy using another mathematical approach. Internal validation based on cross-validation with the current cohort is the most common strategy; external validation based on experiments performed on another platform using samples from a different origin is rare. Consequently, it would be valuable to test different methods of statistical treatment to assess the similarity of the biomarkers identified and the performance of predictive models. Also, it would be helpful to provide the proof of concept that different mathematical tools could be used to analyze metabolomics data.

Knowledge discovery from data describing a piece of the real or an abstract world is a field of computer science: it involves the process of searching the data automatically for patterns that can be considered knowledge about this piece of the world. The patterns are evidenced by induction some "condition–decision" relationships hidden in the data. The most natural representation of these relationships is by "*if...,then...*" decision rules relating some conditions on independent variables (called condition attributes) to some decisions on a dependent variable (called the decision attribute). The same representation of patterns is used in multi-attribute classification, and therefore the data searched for discovery of these patterns can be seen as classification data.

In this paper, we analyze metabolomics data using the standard descriptive and predictive methodologies (PCA and OPLS-DA). We also adopt the classification perspective to apply an original methodology based on dominance-based rough set approach (DRSA) to induction of "condition-decision" relationships from the data and representing them by so-called monotonic decision rules. The DRSA was never applied to metabolomics. Consequently, we assessed the complementarity and the usefulness of the novel concept of analyzing metabolomics data by DRSA; this included testing for the potential concordance of results between DRSA and standard statistical methods of metabolomics data treatment, particularly in the identification of relevant variables (condition attributes). We used some data from targeted metabolomics investigations: ^{1}H Nuclear Magnetic Resonance (NMR) analysis of cerebrospinal fluids (CSF) from patients with amyotrophic lateral sclerosis (ALS) and non-ALS patients. Some of these data were published previously [9]. The ultimate aims of this project were to discriminate ALS patients and non-ALS subjects, and to identify the variables relevant for this discrimination.

This study is the first to apply the DRSA to the analysis of metabolomics data and to assess the "added value" of this method over standard statistical methods.

2. Methods

2.1. Sample collection and NMR acquisition

The methodology for collection, handling and analysis of CSF samples and for NMR acquisition has been described elsewhere [9]. Briefly, we collected 50 samples from ALS patients at the time of diagnosis and 49 from non-ALS subjects. Information on age at onset and gender were obtained for each subject. The ^{1}H NMR spectra were acquired with a Bruker DRX-500 spectrometer (Bruker SADIS, Wissembourg, France). Data were processed using XWinNMR version 3.5 software (Bruker Daltonik, Karlsruhe, Germany). Metabolite peaks were determined with the ERETIC peak as a quantitative reference.

We quantified (by XWin NMR software) 17 CSF metabolites in ALS and non ALS patients, defined as follows: amino-acids (alanine, glutamine, tyrosine), organic acids (citrate, acetate, α-hydroxybutyrate (AHBT)), ketone bodies (β-hydroxybutyrate (BHBT), acetone, acetoacetate), glucose, fructose, metabolites involved in glucose metabolism (pyruvate, lactate), creatinine and creatine, recently identified as markers of mitochondrial dysfunction [10], the antioxidant molecule ascorbate, formate, and ethanol. Thus, we obtained the following data for each subject: age, gender, concentrations of the CSF metabolites (additional file 1).

2.2. Univariate analysis

Wilcoxon tests and *t*-tests were used to compare CSF metabolite concentrations between ALS and non ALS patients to identify disturbances in metabolic pathways associated with ALS. We also compared sex and age between the two groups. A correction for multiple tests (Bonferroni adjustment) was applied to adjust the *p* values by accounting for the 19 parameters evaluated in the analysis. Differences were considered as significant when $p < 0.0026$. JMP statistical software version 7.0.2 (SAS Institute, Cary, North Carolina) was used for all statistical analyses.