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

Hélène Blasco, Jerzy Blaszczynski, Jean-Charles Billaut, Lydie Nadal-Desbarats, Pierre-Francois Pradat, David Devos, Caroline Moreau, Christian R Andres, Patrick Emond, Philippe Corcia, et al.

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

Blasco H[1,2,3], Błaszczyński J[4], Billaut JC[6], Nadal-Desbarats L[1,2,7], Pradat PF[8], Devos D[9], Moreau C[9], Andres C.R. [1,2,3], Emond P [1,2,7], Corcia P [1,2,10], Słowiński R [4,5]

1- Inserm U930, CNRS 2448, Tours, France
2- Université François-Rabelais, Tours, France
3- Laboratoire de Biochimie et Biologie Moléculaire, CHRU de Tours, Tours, France
4- Institute of Computing Science, Poznań University of Technology, 60-965 Poznań, Poland
5- Systems Research Institute, Polish Academy of Sciences, 01-447 Warsaw, Poland
6- Polytech'Tours, France
7- PPF, Université François-Rabelais, Tours, France
8- Fédération des Maladies du Système Nerveux, Centre Référent Maladie Rare SLA, Hôpital de la Pitié-Salpétrière, Paris, France
9- Service de Neurologie, CHRU de Lille Lille, France
10- Centre SLA, Service de Neurologie, CHRU Bretonneau, Tours, France

Abstract

Background

Metabolomics is an emerging strategy used in health field to ascertaining a metabolic profile from a combination of small molecules. This method is currently applied to find diagnosis biomarkers or to identify pathophysiological ways involved in pathology. However, this approach provides multiple and complex data that are usually analyzed by statistical methods largely described but neither standardized nor validated by rigorous rules. As data preprocessing and analysis are the foundation of a robust methodology, new mathematical methods have to be developed to discuss or complete the current method. Thus, we applied the dominance-based rough set approach (DRSA) to analyze metabolomics data and to compare this method with the standard statistical method. Some of attributes were transformed in a way allowing to discover global and local monotonic relationships between condition and decision attributes. We used previously published metabolomics data (18 variables) of Amyotrophic Lateral Sclerosis (ALS) and non-ALS patients.

Results

Principal Component Analysis (PCA) and Orthogonal Partial Least Square- Discriminant Analyis (OPLS-DA) revealed a correct discrimination (72.73%) between ALS and non-ALS patients. We identified some discriminant compounds such as acetate, acetone, pyruvate and glutamine. Interestingly, the concentrations of some metabolites (acetate, pyruvate) were also significantly different between ALS and non-ALS patients using univariate analysis. The DRSA revealed a correct classification in 68.7% of the cases and established rules involving some of the metabolites previously highlighted in the OPLS-DA such as acetate and acetone. Some rules identified promising biomarkers not revealed by OPLS-DA (beta-hydroxybutyrate). We also found a large number of common discriminating metabolites after

Bayesian confirmation measure, particularly acetate, pyruvate, acetone, ascorbate, that is consistent with the pathophysiological ways previously reported in ALS.

Conclusion

The DRSA provides a complementary mean to improve the predictive performance of the multivariate data analysis usually used in metabolomics. This method could help in the determination of metabolites involved in the pathogenesis of a disease. Interestingly, we noted a high concordance between discriminant metabolites revealed by these different strategies. A selection of strong decision rules with high value of Bayesian confirmation provides useful information about relevant cause-effect relationships not revealed in metabolomics data.

Keywords
Metabolomics, multivariate analysis, OPLS-DA, dominance-based rough set approach, DRSA, Bayesian method, diagnosis prediction, Amyotrophic Lateral Sclerosis, ALS

## Background

Metabolomics refers to a powerful approach to ascertaining metabolic signatures from a combination of small molecules in biological fluid. This emerging “omics” approach is increasingly used in health field to find diagnosis or prognosis biomarkers of diseases or to highlight pathophysiological ways involved in these pathologies. This method selects metabolites across a large spectrum of concentrations, polarity and masses [1-3], based on high-throughput techniques [4]. We usually define “untargeted” metabolomics, based on metabolic profile without systematic identification of metabolites included in this profile and “targeted” methodologies focused on specific metabolites. Whatever the approach, data preprocessing and analysis are the foundation of a robust methodology. Studies in bioinformatics are emerging and try to standardize these steps [5-8]. The greatest risk of high-throughput techniques is mishandling of multiple and complex data, leading to biased results. Multivariate statistical analysis including Principal Component Analysis (PCA :descriptive aim) and Orthogonal Partial least squares Discriminant Analysis (OPLS-DA: predictive aim) remains the most used methodology in “omics” studies but it could be discussed. Standard criteria to validate the models are often unavailable. Efforts have to be done to statistically validate the models or to validate this strategy using another mathematical approach. Internal validation based on cross-validation on the current cohort is the most common strategy; external validation based on experiments performed on another platform using samples from a different origin is rarely applied. Thus, it could be interesting to test different methods of statistical treatment to assess the similarity of highlighted biomarkers and to compare the performance of predictive models.

Knowledge discovery from data describing a piece of real or abstract world is a field of computer science that concerns the process of automatically searching the data for patterns that can be considered knowledge about this piece of the world. The patterns are to evidence

by induction some “cause-effect” 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) with some decisions on a dependent variable (called decision attribute). The same representation of patterns is used in multi-attribute classification, thus the data searched for discovery of these patterns can be seen as classification data.

In this paper, we analyzed the metabolomics data using the standard descriptive and predictive methodologies (PCA and OPLS-DA), and we also adopted the classification perspective to apply an original methodology of inducing “cause-effect” relationships from the data and representing them by so-called monotonic decision rules. We compared these approaches to assess their complementarity and the concordance of results, particularly in the identification of relevant variables (condition attributes). To perform this study, we used some data from targeted metabolomics using $^{1}$H Nuclear Magnetic Resonance (NMR) of cerebrospinal fluids (CSF) of patients with amyotrophic lateral sclerosis (ALS) and non-ALS patients. These data were partially previously published [9]. The ultimate aim of this project was to discriminate patients and non-ALS patients and to identify the variables relevant in this discrimination.

This study is the first to apply the dominance-based rough set approach to analyze metabolomics data and to compare this relevant method with other standard statistical methods.

## Methods

### Sample collection and NMR acquisition

The methodology for CSF samples and NMR acquisition is described in our previous study [9]. Briefly, we collected 50 samples from ALS patients and 49 from non-ALS subjects at the time of diagnosis. Information on age at onset and gender were obtained for each subject. The $^{1}H$ NMR spectra were performed on a Bruker DRX-500 spectrometer (Bruker SADIS, Wissembourg, France). Data were processed using XWinNMR version 3.5 software (Bruker Daltonik, Karlsruhe, Germany). Quantification of metabolites peaks was performed 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 anti-oxidant molecule ascorbate, and formate as well as ethanol. Thus, we obtained the following data for each subject: age, gender, concentrations of the CSF metabolites **(additional file 1).**

**Univariate analysis**

A comparison of CSF metabolites concentrations between ALS and non ALS patients was performed using t-tests or Wilcoxon tests in order to highlight the potential disturbances in the metabolic pathway due to ALS. We also compared sex and age between both groups. A correction for multiple tests was applied to adjust the p values by accounting for the 19 parameters evaluated in the analysis. Differences were considered as significant when $p<0.0026$. Statistical analysis was performed with JMP statistical software version 7.0.2 (SAS Institute, Cary, North Carolina).

**Principal Componant analysis**

Multivariate analysis was performed on metabolomics data using Simca-$P^+$-13. We used Pareto scaling (Par), obtained by dividing each variable by the square root of its standard deviation. Moreover, a logarithmic transformation was done to minimise the impact of noise or high variance of the variables [11]. After these transformations, data were used for unsupervised principal components analysis (PCA)[12] to identify similarities or differences between sample profiles. Spectral variation was reduced to a series of principal components (PC), each representing correlated spectral changes, and summarized in a score plot. PCs, new variables that are orthogonal to each other, explained progressively less variance in the data set. The PCs were displayed in a two-dimensional score plot, allowing visualization of the distribution and grouping of the samples in the new variable space. Score plots were visually inspected for grouping, trends and outliers in the data. If outliers were detected in the distance to model plot (DModX), which is based on the residual variance model, they were rejected, and the PCA model was rebuilt. We evaluated the $Q^2/R^2$ overview plot to highlight the cumulative $R^2$ and $Q^2$ values for each variable. The well modelled metabolites have $R^2$ and $Q^2$ values >0.5. The $R^2$ indicates how well the variation of a variable is explained and $Q^2$ how well a variable could be predicted and estimated by cross validation.

**Orthogonal Partial Least Square – Discriminant analysis**

Orthogonal partial least-squares discriminant analysis (OPLS-DA) evaluated variations in frames areas between groups: variation in the measured data was partitioned into 2 blocks by the program, one containing variations that correlated with the class identifier and the other

containing variations that were orthogonal to the first block and thus did not contribute to discrimination between groups [13]. Next, we created a score plot to visualize the OPLS-DA model and characterized the contribution of variables to the separation of classes using the loading plot and the contribution plot. OPLS-DA was cross validated by withholding one-seventh of the samples in seven successive simulations such that each sample was omitted once in order to guard against over fitting [14]. This approach meant that the OPLS-DA was built from one "predictive" component and two or more orthogonal components. $Q^2$ and $R^2$ assessed the robustness of the model. $R^2$ is defined as a fraction of the variance explained by a component. Cross validation of $R^2$ gives $Q^2$, which represents the proportion of total variation predicted by a component. The quality of the models was described by the cumulative modeled variation in the X matrix (metabolites) $R^2$X(cum), the cumulative modeled variation in the Y matrix (CSF samples) $R^2$Y(cum), and the cross validated predictive ability $Q^2$(cum) values. Models were rejected if there was complete overlap of $Q^2$ distributions ($Q^2$(cum) < 0) or low classification rates ($Q^2$(cum)< 0.05 and eigenvalues > 2). We considered a model robust if $Q^2$> 40% and $R^2$>50%, but these cut off values need to be confirmed under biological conditions. The set of multiple models resulting from the cross validation was used to calculate jack-knife uncertainty measures. The predicted data are then compared with the original data and the sum of squared errors calculated for the whole dataset. We fixed the maximum number of iterations at 200 to ensure convergence of the OPLS algorithm [15]. A misclassification table was generated to show the proportion of correctly classified observations in the dataset (ALS *vs* non-ALS patients). We evaluated the performance of the predictive model using sensitivity, specificity, predictive positive value (PPV) and predictive negative value (PNV). We built the models from the 17 metabolites and age. Variable importance parameters (VIP) ranked the compounds according to their contribution to the

model, and the variable selection led to the involvement of the most relevant VIP in the final model.

**Dominance-based rough set approach**

Data from targeted metabolomics of cerebrospinal fluids of ALS and non-ALS can be seen as classification data, where concentrations of particular metabolites are condition attributes (independent variables), and ALS or non-ALS are class labels assigned to patients by a decision attribute (dependent variable). To explain the class assignment in terms of condition attributes, we used the rough set concept [16], and its particular extension called Dominance-based Rough Set Approach (DRSA)[17-20]. DRSA proved to be an effective tool in analysis of classification data which are partially inconsistent [21-22]. In our context, inconsistency means that two patients have similar concentrations of metabolites, while one is in the ALS class, and another is in non-ALS class. The rough sets representing ALS and non-ALS classes discern between consistent and inconsistent patients and prepare the ground for induction of decision rules. DRSA assumes that the value sets of condition attributes are ordered and monotonically dependent on the order of decision classes. In consequence, the rules induced from data structured using the concept of dominance-based rough sets are monotonic, which means that they have the following syntax:

"if $at_i(patient) \geq val_i$ and $at_j(patient) \geq val_j$ and … and $at_p(patient) \geq val_p$, then patient is ALS",

"if $at_k(patient) \leq val_k$ and $at_l(patient) \leq val_l$ and … and $at_s(patient) \leq val_s$, then patient is non-ALS",

where $at_h$ is an $h$-th condition attribute and $val_h$ is a threshold value of this attribute which makes an elementary condition $at_h(patient) \geq val_h$ or $at_h(patient) \leq val_h$ entering the condition part of a rule indicating the assignment of a patient to either class ALS or non-ALS, respectively. In the above syntax of the rules, it is assumed that value sets of all condition attributes are numerical

and ordered such that the greater the value, the more likely is that the patient is ALS; analogously, it is assumed that the smaller the value, the more likely is that the patient is non-ALS. Attributes ordered in this way are called gain-type. Cost-type attributes have value sets ordered in the opposite way, so elementary conditions on these attributes have opposite relation signs. In case of metabolomics data, it is of course impossible to assume *a priori* if concentrations of metabolites are gain or cost attributes, thus we proceed as described in [23], i.e., we are considering each original attribute in two copies, and for the first copy we assume it is gain-type, while for the second copy we assume it is cost-type. The applied transformation of data is non-invasive, i.e., it does not bias the matter of discovered relationships between condition attributes and the decision attribute. Then, the induction algorithm constructs decision rules involving elementary conditions on one or both copies of particular attributes. For example, in a rule indicating the assignment of a patient to class ALS there may appear the following elementary conditions concerning attribute $at_i$:

- $\uparrow at_i(patient) \geq val_{i1}$,
- $\downarrow at_i(patient) \leq val_{i2}$,
- $\uparrow at_i(patient) \geq val_{i1}$ and $\downarrow at_i(patient) \leq val_{i2}$, which boils down to $at_i(patient) \in [val_{i1}, val_{i2},]$ if $val_{i1} \leq val_{i2}$,

where $\uparrow at_i$ and $\downarrow at_i$ are gain-type and cost-type copies of attribute $at_i$, respectively. Remark that the above transformation of attributes permits discovering global and local monotonic relationships between concentration of metabolites and class assignment.

Sets of decision rules, which are essential for the analysis, were induced from metabolomics data transformed in the way described above and structured using the concept of dominance-based rough sets. The induction algorithm is called VC-DomLEM [24]; it has been implemented as software package called jMAF (http://idss.cs.put.poznan.pl/site/139.html), based on java Rough Set (jRS) library. The induced sets of rules were used to construct

component classifiers in variable consistency bagging [25-26]. Variable consistency bagging (VC-bagging) was applied to increase the accuracy of results produced by VC-DomLEM. Estimation of attribute relevance in rules has been performed through measuring Bayesian confirmation, as described in [27]. In this process, decision rules are induced repetitively on bootstrap samples and tested on patients who are not included in the samples. In this way, attributes present in the premise of a rule that assigns patients correctly, or attributes absent in the condition part of a rule that assigns patients incorrectly, are getting more relevant. Bayesian confirmation measure quantifies the degree to which the presence of an attribute in the premise of a rule provides evidence for or against the conclusion of the rule.

## Results and discussion

### Univariate analysis

We found no difference of mean age between both groups (59.7+/-10.6 years in ALS vs 63.9+/-14.5 years in non-ALS patients, p=0.09). The percentage of men was similar in both groups (64% in ALS vs 55.10% in non-ALS patients, p=0.41). We found lower concentrations of acetate (mean +/- SD : 174+/-159 µmol/L vs 84+/-101 µmol/L ($p<0.0001$)) in ALS patients compared to non ALS patients. According to bonferroni adjustment, acetone concentration had a trend to be higher in ALS patients (13.4 +/-12.2 µmol/L vs 8.7+/-8.2 µmol/L, p=0.02). Glutamine concentration had a trend to be lower in ALS patients (652+/-225 vs 790+/-348, p=0.03). Moreover, we found higher concentrations of pyruvate (58+/-37 µmol/L vs 39+/-48, p=0.0026) in ALS patients than in non-ALS patients. The colour map, based on p-values evaluating correlation between all variables revealed that variables were highly correlated (**Additional file 2**).

**PCA**

The resulting PCA score plot showed that the two first PCs explained 64.1 % of the variation in the selected metabolites (**Figure 1**A). The PCA scores plot for the quantitative data showed quite separated clusters for ALS and non-ALS patients. To sum the observation, each spectrum can be viewed as an observation in PCA space where the proximity of observations represents the similarity of the metabolic profiles of CSF samples. The contribution score plot (**figure 1B**) identified the relative importance of each variable in differentiating ALS from non-ALS patients. The Q2/R2 overview plot showed that 76.4% of the metabolites were well modeled (**figure s2**).

**OPLS-DA**

The model including the 17 metabolites and age showed quite good results with a correct prediction in 72.73% of cases (sensitivity: 78%, specificity; 67.3%, PPV: 70.9% PNV: 67.3%, **additional file 3**). The metabolomics data treatment is based on the optimisation of the predictive model based on the selection of variables. This approach provides the best performance criteria of the models. According to the high number of variables compared to the size of the cohort, the maximal reduction of the number of selected metabolites leads to the most rigorous results.

From the predictive variation between X (metabolites, age) and Y (CSF samples) given by $R^2$X(cum), the best model used 4 components, and interpreted approximately 99.9% of the total variation in X. Pareto scaling of the models created by OPLS-DA (**figure 2A**), explained approximately 24.5% of the variations in the various samples ($R^2Y$(cum)). We found low

predictive value of the models ($Q^2$(cum) = 0.167). The cross-validation performance was confirmed by analysis of variance CV-ANOVA (median *p*-value = 0.021). The contribution score plot (**figure 2B**) identified the relative importance of each compound in differentiating ALS from non-ALS patients and showed that high levels of acetone and pyruvate, and low levels of glutamine and acetate were discriminating in this model. The loading scatter plot (**figure 3**) showing which variables expressed differentially between ALS and non-ALS patients confirmed these results. The VIP panel including these 4 metabolites predicted the diagnostic group with mean probability of 76.77%, with 82% specificity, 71.4% sensitivity, 74.5 PPV, 79% PNV. Although the performance criteria of the model are not excellent, the predictive ability of the model is correct. However, we have to take into account the high optimist results of such strategy: fitting a PLS model using VIP identified in a previous step corresponds to a statistical variable selection coupled with a PLS predictor. This strategy provides artificially optimistic cross-validation results, as the $Q^2$ value is only derived for the second PLS model alone, not for the combined two-step "variable selection + PLS predictor" model.

Most importantly, we found that discriminative metabolites were the same as those highlighted in univariate analysis, including those being statistically different and those having a trend to be different between both groups.

**Dominance-based rough set approach**

The model constructed by VC-bagging with VC-DomLEM component classifiers showed good classification performance in 3-fold stratified cross validation, which was repeated 100 times for a better repeatability. On average, 68.7% of cases were correctly classified (sensitivity: 70%, and specificity: 67.3%). The values of a Bayesian confirmation measure

calculated for all variables (condition attributes) give more insight into the constructed classification model (**figure 4**). The variables having the highest values of confirmation are the ones, which are the most relevant from the viewpoint of correct prediction by the DRSA rule model. What is the most important, we found that the same discriminative metabolites, as discovered by the univariate analysis and OPLS-DA, are distinguished as the most relevant by the Bayesian confirmation.

The rules induced from the information table structured using DRSA represent relevant patterns of cause-effect relationships, which are free of inessential and redundant information. Some of the most relevant rules among all rules that constitute DRSA model are presented below.

1: if Acetate $\geq$ 0.125 and Ascorbate $\leq$ 0.01964669 then patient is not ALS;

{strength: 0.21, confirmation: 0.57}

2: if Gln $\geq$ 0.804629579 and Acetone $\leq$ 0.007626821 then patient is not ALS;

{strength: 0.11, confirmation: 0.37}

3: if Pyruvate $\leq$ 0.029593053 and Crn $\geq$ 0.192354842 then patient is not ALS;

{strength: 0.09, confirmation: 0.2}

4: if Acetate $\leq$ 0.055457394 and Ascorbate $\leq$ 0.0786 then patient is ALS;

{strength: 0.25, confirmation: 0.55}

5: if Acetone $\in$ [0.0087;0.009534583] then patient is ALS;

{strength: 0.04, confirmation: 0.5}

6: Alanine $\leq$ 0.039024531 and Pyruvate $\geq$ 0.0326 then patient is ALS;

{strength: 0.08, confirmation: 0.55}

**Comparison of results obtained using different approach**es

Interestingly, the different approaches led to identification of the same relevant metabolites. Even if the criteria of internal validation are different, including the number of iterations, the number of samples in the test set…, the performance and the relevance of the findings are quite the same.

First, from the descriptive methods of the dominance-based rough set approach, the most relevant metabolites to explain ALS are low concentration of acetate (<55 µmol/L in the rules, and the median is 88 µmol/L), and high concentration of pyruvate (>32.6 µmol/L in the rules and the median is 34.9 µmol/L. Although, alanine was not relevant in the PCA and was not statistically different between ALS and non-ALS patients using univariate analysis, this metabolite appears important when associated with high levels of pyruvate. To explain the non-ALS groups, we also found the same relevant metabolites with an inverse tendency compared to the previous rules. The dominance-based rough set approach confirmed the relevance of high acetate and low ascorbate concentration in the non-ALS group. High concentrations of glutamine associated with low concentrations of acetone remain relevant to explain the non-ALS patients. Although, creatine-creatinine was not relevant in the PCA and was not statistically different between ALS and non-ALS patients using univariate analysis, this metabolite appears important when associated with low levels of pyruvate, thus highlighting the potential crucial role of pyruvate in the pathogenesis of ALS. These findings are consistent with the hypothesis of disturbance in glucose metabolism largely reported in ALS [9, 28-29].

Among the rules obtained from dominance-based rough set approach (**additional file 5**), other metabolites are frequently involved in a high number of rules such as beta-hydroxybutyrate. Surprisingly, beta-hydroxybutyrate was not identified by other methods. However, this observation is greatly consistent with the involvement of acetate and acetone in the predictive models. All these metabolites are ketone bodies, previously highlighted by

Kumar et al.[30], in a metabolomics study performed in serum of ALS patients. Interestingly, some variables are involved in some rules but the range of values are opposite, dependant on the other variables associated in this rule (i.e., to explain the pathology in predictive rules : high values of beta-hydroxybutyrate when associated with high values of acetone and low values of beta-hydroxybutyrate when associated with high values of ascorbate) Such observations seem the most informative for the interpretation of biochemical ways, because they reveal the dynamism and the kinetic of the overall metabolism.

Second, we also found a large number of common discriminating metabolites after Bayesian confirmation measure computed for each of 18 variables: acetate, pyruvate, acetone, ascorbate. Although ascorbate was not highly discriminative in the final OPLS-DA model, the concentration of ascorbate has a trend to be higher in ALS patients and it is the 3$^{rd}$ metabolites the most important in the OPLS-DA model included all variables. The role of ascorbate, considered as an anti-oxidant compound is consistent with the oxidative stress involved in the pathogenesis of ALS [31-32].

The dominance-based rough set approach provides a complementary mean to improve the predictability of the models usually used in metabolomics field. Above all, this approach could help in the determination of metabolites involved in the pathogenesis of a disease. The ability to suggest some mechanistic explanations about pathophysiological ways is promising. Moreover, a selection of strong decision rules with high value of Bayesian confirmation provides useful information about relevant cause-effect relationships hidden in metabolomics data.

**Availability of supporting data**

The data set supporting the results of this article is included within the additional files of the article (table S1)

**List of abbreviations**

ALS :amyotrophic lateral sclerosis
AHBT: α-hydroxybutyrate CSF: cerebrospinal fluids
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 components
PCA: Principal Component Analysis
PNV predictive negative value
PPV predictive positive value
$Q^2$ :indicates how well a variable could 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

**Competing interests**

The authors declare that they have no competing interests.

**Authors' contributions**

HB participated in the design of the study, in the acquisition of the data; performed the multivariate analysis (PCA, OPLS-DA), contributed to the interpretation of the data, and wrote the first part of the manuscript. JB performed the analysis using the dominance-based rough set approach, contributed to the interpretation of the data and wrote the second part of the manuscript. JCB contributed to the design of this study and its coordination and revised the manuscript. LND participated in the acquisition and analysis of the data, was involved in

the interpretation of the data and revised the manuscript. PFP, DD and CM, participated to the collection of the data, to the interpretation of the data and revised the manuscript. PE participated to the design and the coordination of this study, was involved in the interpretation of the data, and revised the manuscript. PC participated in the design and the coordination of this study, was involved in the collection of data, contributed to the interpretation of the data and revised the manuscript. RM participated to the design of this study, performed the analysis using the dominance-based rough set approach, contributed to the interpretation of the data and wrote the second part of the manuscript. All authors have given final approval of the version to be published agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

**Acknowledgements**

J.Błaszczyński and R.Słowiński wish to acknowledge financial support from the Polish National Science Center, grant no. DEC-2011/01/B/ST6/07318. We acknowledge Catherine Antar for her technical help in this study.

**References**

1. Patti GJ, Yanes O, Siuzdak G: Innovation: Metabolomics: the apogee of the omics trilogy. *Nat Rev Mol Cell Biol* 2012, 13(4):263-269.
2. Wang JH, Byun J, Pennathur S: Analytical approaches to metabolomics and applications to systems biology. *Semin Nephrol* 2010, 30(5):500-511.
3. Koek MM, Jellema RH, van der Greef J, Tas AC, Hankemeier T: Quantitative metabolomics based on gas chromatography mass spectrometry: status and perspectives. *Metabolomics* 2011, 7(3):307-328.
4. Quinones MP, Kaddurah-Daouk R: Metabolomics tools for identifying biomarkers for neuropsychiatric diseases. *Neurobiol Dis* 2009, 35(2):165-176.
5. Courant F, Royer AL, Chereau S, Morvan ML, Monteau F, Antignac JP, Le Bizec B: Implementation of a semi-automated strategy for the annotation of metabolomic fingerprints generated by liquid chromatography-high resolution mass spectrometry from biological samples. *Analyst* 2012, 137(21):4958-4967.

6. Hoffmann N, Keck M, Neuweger H, Wilhelm M, Hogy P, Niehaus K, Stoye J: Combining peak- and chromatogram-based retention time alignment algorithms for multiple chromatography-mass spectrometry datasets. *BMC Bioinformatics* 2012, 13:214.
7. Eliasson M, Rannar S, Madsen R, Donten MA, Marsden-Edwards E, Moritz T, Shockcor JP, Johansson E, Trygg J: Strategy for optimizing LC-MS data processing in metabolomics: a design of experiments approach. *Anal Chem* 2012, 84(15):6869-6876.
8. Kirwan GM, Johansson E, Kleemann R, Verheij ER, Wheelock AM, Goto S, Trygg J, Wheelock CE: Building multivariate systems biology models. *Anal Chem* 2012, 84(16):7064-7071.
9. Blasco H, Corcia P, Moreau C, Veau S, Fournier C, Vourc'h P, Emond P, Gordon P, Pradat PF, Praline J *et al*: 1H-NMR-based metabolomic profiling of CSF in early amyotrophic lateral sclerosis. *PLoS One* 2010, 5(10):e13223.
10. Shaham O, Slate NG, Goldberger O, Xu Q, Ramanathan A, Souza AL, Clish CB, Sims KB, Mootha VK: A plasma signature of human mitochondrial disease revealed through metabolic profiling of spent media from cultured muscle cells. *Proc Natl Acad Sci U S A* 2010, 107(4):1571-1575.
11. van den Berg RA, Hoefsloot HC, Westerhuis JA, Smilde AK, van der Werf MJ: Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics* 2006, 7:142.
12. Kemsley EK, Le Gall G, Dainty JR, Watson AD, Harvey LJ, Tapp HS, Colquhoun IJ: Multivariate techniques and their application in nutrition: a metabolomics case study. *Br J Nutr* 2007, 98(1):1-14.
13. Madsen R, Lundstedt T, Trygg J: Chemometrics in metabolomics--a review in human disease diagnosis. *Anal Chim Acta* 2010, 659(1-2):23-33.
14. Cloarec O, Dumas ME, Trygg J, Craig A, Barton RH, Lindon JC, Nicholson JK, Holmes E: Evaluation of the orthogonal projection on latent structure model limitations caused by chemical shift variability and improved visualization of biomarker changes in 1H NMR spectroscopic metabonomic studies. *Anal Chem* 2005, 77(2):517-526.
15. Westerhuis JA, van Velzen EJ, Hoefsloot HC, Smilde AK: Multivariate paired data analysis: multilevel PLSDA versus OPLSDA. *Metabolomics* 2010, 6(1):119-128.
16. Pawlak Z: Rough sets. Theoretical aspects of reasoning about data. Dordrecht: Kluwer; 1991.
17. Błaszczyński J, Greco S, Słowiński R, Szelag M: Monotonic Variable Consistency Rough Set Approaches. *International Journal of Approximate Reasoning* 2009, 50:979-999.
18. Greco S, Matarazzo B, Słowiński R: Rough sets theory for multicriteria decision analysis. *European J of Operational Research* 2001, 129:1-47.
19. Slowinski R, Greco S, Matarazzo B: Rough Sets in Decision Making. In: *Encyclopedia of Complexity and Systems Science.* Edited by R.A.Meyers. New York: Springer; 2009: 7753-7786.
20. Słowiński R, Greco S, Matarazzo B: Rough-Set-Based Decision Support. In: *Search Methodologies: Introductory Tutorials in Optimization and Decision Support Techniques.* Edited by Kendall EKBaG, 2nd edition edn. New York: Springer; 2014: 557-609.
21. Greco S, Matarazzo B, Słowiński R: Multicriteria classification. In: *Handbook of Data Mining and Knowledge Discovery* Edited by J.Żytkow WKa. New York: Oxford University Press; 2002: 318-328.

22. Greco S, Matarazzo B, Slowinski R: Rough sets methodology for sorting problems in presence of multiple attributes and criteria. *European J of Operational Research* 2002, 138(2):247-259.
23. Błaszczyński J, Greco S, Słowiński R: Inductive discovery of laws using monotonic rules. *Engineering Applications of Artificial Intelligence* 2012, 25(2):284–294.
24. Błaszczyński J, Słowiński R, Szeląg M: Sequential covering rule induction algorithm for variable consistency rough set approaches. *Information Sciences* 2011, 181:987-1002.
25. Błaszczyński J, Słowiński R, Stefanowski J: Proc. of the Workshop From Local to Global Models : Feature Set-based Consistency Sampling in Bagging Ensembles. In: *European Conference on Machine Learning & Principles of Knowledge Discovery in Databases (ECML/PKDD 2009): September 7-11 2009; Bled, Slovenia*. 19-35.
26. Błaszczyński J, Słowiński R, Stefanowski J: Variable consistency bagging ensembles. . In: *Transactions on Rough Sets XI (LNCS 5946): 2010; Springer, Berlin*. 40-52.
27. Błaszczyński J, Słowiński R, Susmaga R: Rule-based estimation of attribute relevance. In: *RSKT 2011, LNCS 6954: 2011; Berlin*. Springer: 36-44.
28. Pradat PF, Bruneteau G, Gordon PH, Dupuis L, Bonnefont-Rousselot D, Simon D, Salachas F, Corcia P, Frochot V, Lacorte JM *et al*: Impaired glucose tolerance in patients with amyotrophic lateral sclerosis. *Amyotroph Lateral Scler* 2009:1-6.
29. Blasco H, Corcia P, Pradat PF, Bocca C, Gordon PH, Veyrat-Durebex C, Mavel S, Nadal-Desbarats L, Moreau C, Devos D *et al*: Metabolomics in Cerebrospinal Fluid of Patients with Amyotrophic Lateral Sclerosis: An Untargeted Approach via High-Resolution Mass Spectrometry. *J Proteome Res* 2013.
30. Kumar A, Bala L, Kalita J, Misra UK, Singh RL, Khetrapal CL, Babu GN: Metabolomic analysis of serum by (1) H NMR spectroscopy in amyotrophic lateral sclerosis. *Clin Chim Acta* 2010.
31. Parakh S, Spencer DM, Halloran MA, Soo KY, Atkin JD: Redox regulation in amyotrophic lateral sclerosis. *Oxid Med Cell Longev* 2013, 2013:408681.
32. Wallis N, Zagami CJ, Beart PM, O'Shea RD: Combined excitotoxic-oxidative stress and the concept of non-cell autonomous pathology of ALS: Insights into motoneuron axonopathy and astrogliosis. *Neurochem Int* 2012.

**Illustrations and figures**

**Figure 1** : **Principal Component Analysis obtained from 18 variables for ALS (A) and non-ALS patients (B)**

Principal Component Analysis obtained from 18 variables (17 metabolites and age), A: score plot for ALS (n=50), with yellow dots and non-ALS patients (n=49), with green dots; B : Contribution plot from model included the 18 variables

**Figure 2: OPLS-DA model obtained from 4 VIP including Scores scatter plot (A) and contribution plot (B).**

OPLS-DA model obtained from 4 VIP (acetone, pyruvate, acetate, glutamine), A:Scores scatter plot of the first principal component (yellow dot :ALS patients, n=50 and green dots : non-ALS patients; n=49), $R^2$X(cum) = 0.999, $R^2$Y(cum) = 0.245, and $Q^2$(cum)=0.167, B: Contribution plot from model included these 4 VIP, peaks having positive contribution scores correspond to metabolites with higher levels in ALS (pyruvate, acetone), and those having negative contribution score correspond to metabolites with higher levels in non-ALS patients (acetate, glutamine).

**Figure 3: Loading score plot from the model based on 4 VIP**

Loading score plot from the model based on 4 VIP. Scatter plot of the X- and Y-loadings. This plot shows how the responses (Y's) varied in relation to each other, i.e. which provided similar information, and their relationship to the terms of the model.

**Figure 4: Bayesian confirmation measure computed for each of 18 variables**

Bayesian confirmation measure computed for each of 18 variables (condition attributes). This plot shows how each of the variables used in DRSA model confirms correct classification of ALS patients.

**additional files**

**Additional file 1 : Data set supporting the results of this study**

**Additional file 2 : Color Map on p-values showing the significance of the correlations between the 18 variables**

Color Map on p-values showing the significance of the correlations between the 18 variables (17 metabolites and age) on a scale from p = 0 (red) to p = 1(blue)

**Additional file 3: $Q^2/R^2$ overview plot**

$Q^2/R^2$ overview plot highlighting the cumulative $R^2$ and $Q^2$ values for each variable (17 metabolites and age). The well modelled metabolites have $R^2$ and $Q^2$ values >0.5.

**Additional file 4: OPLS-DA model obtained from all variables including scores scatter plot (A) and contribution plot (B)**

OPLS-DA model obtained from all variables (17 metabolites and age), A:Scores scatter plot of the first principal component (yellow dot :ALS patients, n=50 and green dots : non-ALS patients; n=49), B: Contribution plot from model included all variables, peaks having positive contribution scores correspond to metabolites with higher levels in ALS and those having negative contribution score correspond to metabolites with higher levels in non-ALS patients.