A-000

CONTROLLED EXPERIMENT FOR EVALUATION OF STATISTICAL MODELS FOR DATA ANALYSIS IN METABOLOMICS.

T.O. Mendes[1,2], C.O.S. Sorzano[1], Angulo S[1], M.J.V. Bell[2], F.J. Ruperez[1], C. Barbas[1]

[1]CEMBIO (Center for Metabolomics and Bioanalysis) Facultad de Farmacia, Universidad CEU San Pablo, Campus Montepríncipe, Boadilla del Monte, 28668 Madrid, Spain.
[2]Laboratório de Espectroscopia de Materiais, Departamento de Física, Universidade Federal de Juiz de Fora, Juiz de Fora-MG, Brazil.

E-mail corresponding author: cbarbas@ceu.es; Tel.: +34 913724753; Fax: +34 913724712

A typical metabolomics experiment generates large data sets, due to the complexity of biological samples and the high sensitivity of modern analytical techniques such as mass detectors coupled to separation systems, which can generate several hundreds/thousands of variables per sample. The manipulation of these data sets is one of the challenging steps of a study in metabolomics, and it determines which metabolites present in an organism are candidates as potential biomarkers. Therefore, data processing step acts directly on the results generated in a study.

Works in the area of chemometrics have contributed to the interpretation of experiments that generate a lot of data for analysis [1], especially multivariate statistics [2] which includes tools such as principal component analysis (PCA), and discriminate analysis with partial least squares regression (PLS-DA), that are used extensively in studies of metabolomics. Other works point to important observations about the pre-processing of data in multivariate analysis in metabolomics [3, 4]. The present work discusses some changes in the selection of variables as potential biomarkers caused by changes in the stage of data processing.

To evaluate these modifications a controlled experiment was designed. Samples (n=30) contained equal amounts of a pool of urine and 20 standards of metabolites in 5 different concentrations, 13 of them changed randomly keeping the average in the two groups equal, while 7 increased or decreased classifying the samples into two groups . After that, samples were analyzed by capillary electrophoresis with time of flight mass detector (CE-TOF), under conditions previously described [5]. The profiles of the 30 electropherograms were aligned in MassProfiler Professional software (Agilent Technologies), and subjected to different processes of transformation, scaling of data, and treatment of missing values in PCA and PLS-DA models, analyzed in SIMCA-P+ software (Umetrics) and the R platform (open-source).

Data analysis showed significant differences when working with raw data or log transformed data, and scaling type unit variance (UV) or variance type Pareto (Par), modifying the estimates of covariance and errors associated with the variables of the PLSDA models. In addition, the treatment of missing values directly influenced the quality of the PCA models and metabolite concentrations between the groups of individuals prepared in the experimental design.

References

[1] R. Madsen, T. Lundstedt, J. Trygg, Analytica Chimica Acta 659 (2010) 23-33
[2] K.H. Liland, Trends in Analytical Chemistry, 30 (2011) 827-841
[3] M. Katajamaa, M. Oresic, Journal of Chromatography A 1158 (2007) 318-328
[4] M.M.W.B. Hendriks, F.A. Eeuwijk, R.H. Jellema, J.A. Westerhuis, T.H. Reijmers, H.C.J. Hoefsloot, A.K. Smilde, Trends in Analytical Chemistry, 30 (2011) 1685-1698
[5] J.Godzien, D. Gracía-Martínez, P. Martinez-Alcazar, F.J. Ruperez, C. Barbas, Metabolomics (2011)