A KNIME Pipeline for the Analysis of GC-MS Data in Metabolomics

Sonia Liggi[1], Maria Laura Santoru[1], Cristina Piras[1], Antonio Murgia[2], Pierluigi Caboni[2], Luigi Atzori[1]

[1] Department of Biomedical Sciences, Section of Pathology, University of Cagliari, Italy
[2] Department of Life and Environmental Sciences, High Resolution Mass Spectrometry Laboratory, University of Cagliari, Italy

Introduction

Elucidation of the metabolic changes taking place in pathological conditions can help in the identification of new biomarkers, prediction of response to therapy and better understanding of the pathogenesis [1]. Gas Chromatography coupled with Mass Spectrometry (GC-MS) is one of the leading analytical techniques utilised to deconvolute the metabolic profile of biofluids and tissues. However, the large number of experiments deriving from high-throughput studies along with the complex set of steps required to pre-process and analyse the results obtained from GC-MS measurements (Figure 1) represents a bottleneck. Indeed, several computer programs need to be used to accomplish a number of tasks (namely retention time correction, peak extraction, metabolites deconvolution, blanks removal, normalisation and last but not least statistical analysis), requiring computational competences and resources not always present in an experimental group. In this context, the KNIME Analytics Platform [2] was used to develop a pipeline able to perform the aforementioned pre-processing steps in an automated way even by users unfamiliar with programming. The workflow was utilised to obtain a matrix of all the signals found in the chromatograms of faecal samples deriving from patients affected by Inflammatory Bowel Diseases (IBDs, namely Ulcerative Colitis and Crohn Disease) and a population of healthy subjects. The deriving matrix was then statistically analysed to elucidate differences in the metabolic spectrum of the three classes of samples.

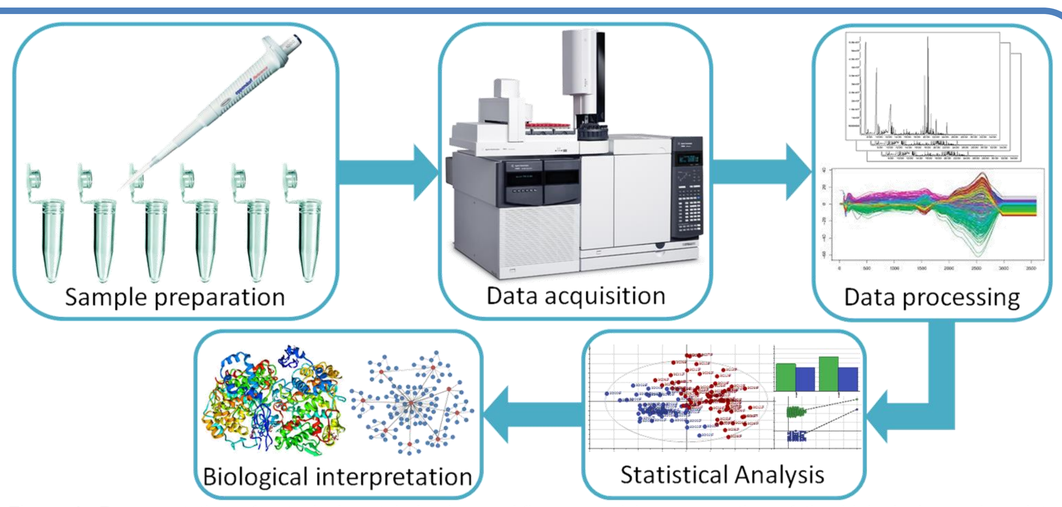

Figure 1: Representation of a typical metabolomics workflow comprising preparation of the biological samples, data acquisition through different techniques (here a GC-MS is represented), processing of the data, statistical analysis and finally biological interpretation with identification of the metabolites implicated in the condition studied as well as the proteins and pathways important in their regulation.

Data pre-processing – KNIME automation

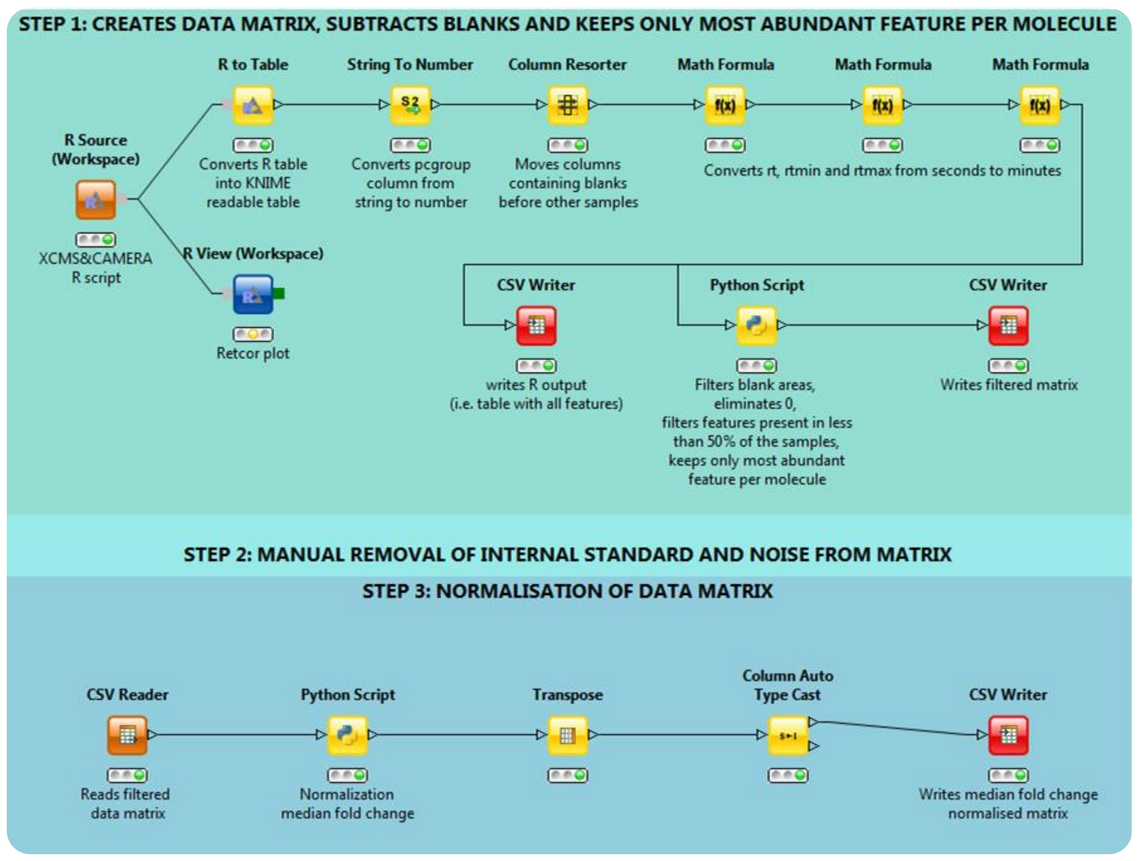

Figure 2: KNIME workflow utilised for data pre-processing by joining KNIME native nodes with in-house python and R scripts.

The chromatograms were processed to obtain a matrix of features present across all samples by using a workflow created with the platform KNIME (Figure 2) joining in-house python and R [3] scripts as well as KNIME native nodes. The workflow comprises two automated steps (step 1 and step 3), and a manual step (step 2). The first step consists in peak detection, retention time correction, peak grouping into pseudospectra according to their retention time, and finally filtering of the matrix. The R library XCMS [4] was utilised for peak detection and retention time correction, whereas grouping of features into pseudospectra was performed using the R library CAMERA [5] (functions and parameters in Table 1). The resulting matrix was processed using a python script to eliminate signals present in the blanks, keep only the most abundant feature per molecule and modify all zeroes present in the matrix by inserting half of the minimum value found for a feature in the matrix. After manual curation of the filtered matrix to eliminate internal standard and any possible remaining noise signal (Figure 2, step 2), median fold change normalisation [6] was performed using a python script in order to compensate for sample dilution biases.

Table 1: The R script included in step 1 of the KNIME workflow utilises the libraries XCMS and CAMERA to perform the tasks listed in the first column by using the functions in the central columns with the arguments specified in the third column.

XCMS		
Peak detection	xcmsSet()	method = "centWave", ppm = 0.1, snthr = 1, peakwidth = c(5, 13), mzdiff = 0.01, prefilter = c(5, 100), integrate = 2, profmethod = "bin"
Retention time correction	retcor()	method = "obiwarp", profStep=0.1, plottype = "deviation"
Grouping	group()	method = "density", bw=3, mzwid=0.25, minfrac=0.5, minsamp=1, max = 100
Missing peaks fill	fillPeaks()	method="chrom"
CAMERA		
Grouping - retention time	groupFWHM()	sigma = 6, perfwhm = 0.6, intval = "into"
Grouping - correlation	groupCorr()	cor_eic_th = 0.75, pval= 0.05

Samples collection, preparation and acquisition

Faecal samples of 78 patients affected by Ulcerative colitis, 48 patients affected from Crohn disease and 49 healthy patients were obtained from the Gastroenterology Unit of University Hospital of Cagliari, Cagliari, Italy. The study was approved by the Institutional Ethics Committee (Azienda Ospedaliero-Universitaria, University of Cagliari) and was performed in accordance with the Declaration of Helsinki. Participants were informed of the purpose and methodology of the study and their written consent was obtained prior to inclusion.
300 mg of faeces were dissolved in 1ml of methanol 80% and left in the solution for 30 minutes to allow extraction of the metabolites. Samples were then centrifuged for 10 minutes at 14000 rpm at a temperature of 4˚C. 300 µL of the surnatant were dried under vacuum over night and derivatised with 50 µL of a solution of methoxamine in pyridine (10mg/ml). After 17h, 100 µL N-Methyl-N-(trimethylsilyl) trifluoroacetamide were added to the samples, followed by addition of 400 µL of solvent and internal standard (undecane in hexane 25ppm) after 1h.
1 µL of each sample was injected in splitless mode into an Agilent 7890A GC coupled with an Agilent 5975C VL MSD, equipped with a 30m X 0.25mm ID fused silica capillary column which was chemically bounded with 0.25 µm HP-5MS stationary phase (Agilent Technologies, Santa Clara, California, USA). The injector temperature was set to 250°C. Carrier gas was helium in constant flow (1mL/min). The column initial temperature was kept at 60°C for 3 minutes, then it was increased from 60°C to 140°C at 7°C per minute, held at 140°C for 4 minutes, increased from 140°C to 300°C at 5°C per minute and held at 300°C for 1 minute. Transfer line temperature was 280°C.

Statistical analysis

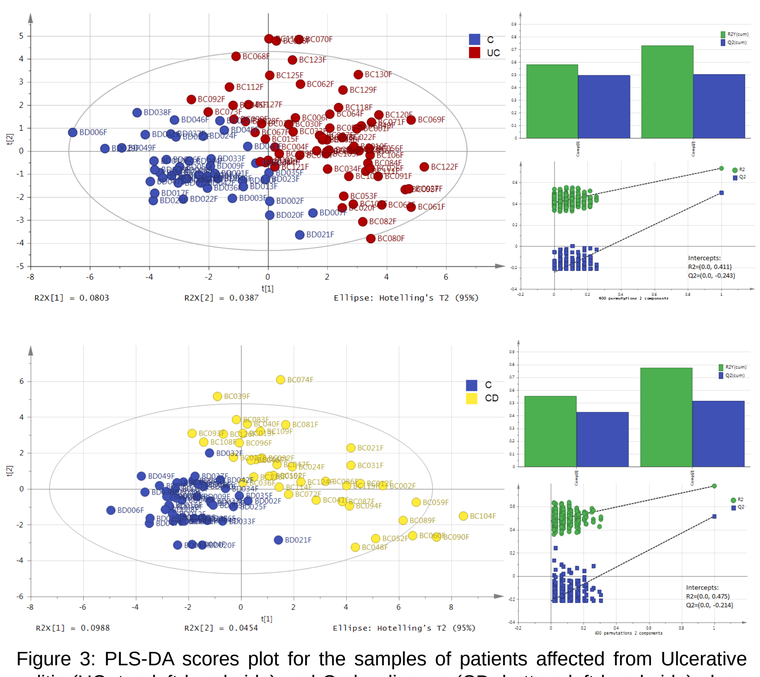

Figure 3: PLS-DA scores plot for the samples of patients affected from Ulcerative colitis (UC, top left hand side) and Crohn disease (CD, bottom left hand side) along with their respective R2 and Q2 values and the results obtained from a permutation test (right hand side).

The normalised matrix was subjected to statistical analyses using the software SIMCA (Umetrics, Umeå, Sweden). Principal components analysis (PCA) was performed on a per-class base to evaluate sample distribution and eliminate samples falling outside the confidence ellipse (Hotelling's T^2 95%). Moreover, partial least squares discriminant analysis (PLS-DA, Figure 3) was utilised to elucidate differences amongst classes, showing an interesting separation between healthy and patients affected by IBDs. Further analyses of these results are currently undergoing with the aim to elucidate inter-class differences and highlight possible biomarkers for an early and not invasive diagnosis of IBDs.

Conclusions and future work

With the increase in computational power and availability of software, analysis of data deriving from high-throughput metabolomics experiments has become easier and more affordable. However, the number of different steps to perform, along with the plethora of parameters to specify, might make this process troublesome and confusing to accomplish for users not familiar with programming. In this context, we developed a workflow able to perform GC-MS data pre-processing and normalisation in a semi-automated way, whose final output can be directly subjected to statistical analysis using external tools. Although the protocol still contains a manual step, and its output needs to be transferred to other platforms to complete the analyses and obtain the results, we are currently working on implementation of these parts to eliminate as much as possible manual curation and external software utilisation.

References

1. R. Madsen *et al.*, Chemometrics in metabolomics—A review in human disease diagnosis, Analytica Chimica Acta, 659 (2010) 23–33
2. M.R. Berthold *et al.*, KNIME: The Konstanz Information Miner, Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007), Springer, ISBN 978-3-540-78239-1, ISSN 1431-8814, 2007
3. R Core Team (2015), R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/
4. C.A. Smith *et al*., XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification Analytical Chemistry, 78 (2006), 779–787
5. C. Kuhl *et al.*, CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets, Analytical Chemistry, 84 (2012), 283–289
6. K.A. Veselkov *et al.*, Optimized preprocessing of ultra-performance liquid chromatography/mass spectrometry urinary metabolic profiles for improved information recovery, Analytical Chemistry, 83 (2011), 5864–5872

Acknowledgements

The authors would like to thank the Gastroenterology and Microbiology Units of the University Hospital of Cagliari for the samples provided as well as for the useful and inspiring discussions. This project was funded by Regione Autonoma della Sardegna, L.R.7/2007, grant number F71J12001180002