Xavier Domingo Almenara

AUTOMATED MASS SPECTROMETRY-BASED METABOLOMICS DATA PROCESSING BY BLIND SOURCE SEPARATION METHODS

DOCTORAL THESIS

supervised by Dr. Jesus Brezmes Llecha
and Dr. Alexandre Perera Lluna

Departament d'Enginyeria
Electrònica, Elèctrica i Automàtica
(DEEEA)

UNIVERSITAT
ROVIRA I VIRGILI

Tarragona
2016

UNIVERSITAT
ROVIRA i VIRGILI

Escola Tècnica Superior d'Enginyeria

Departament d'Enginyeria Electrònica, Elèctrica i Automàtica

Av. Paisos Catalans 26

Campus Sescelades

43007 Tarragona

We CERTIFY that the Doctoral Thesis entitled: "AUTOMATED MASS SPECTROMETRY-BASED METABOLOMICS DATA PROCESSING BY BLIND SOURCE SEPARATION METHODS", presented by Xavier Domingo Almenara to obtain the degree of Doctor, has been performed under our supervision in the Departament d'Enginyeria Electrònica, Elèctrica i Automàtica at the Rovira i Virgili University and it meets the requirements for European Mention qualification.

Tarragona, June 2016

Doctoral Thesis Supervisors

Dr. Jesus Brezmes Llecha

Dr. Alexandre Perera Lluna

Who are we? We find that we live on an insignificant planet of a humdrum star lost in a galaxy tucked away in some forgotten corner of a universe in which there are far more galaxies than people.

Standing over humans, gods, and demons, subsuming Caretakers and Tunnel builders, there is an intelligence that antedates the universe.

We live in a society exquisitely dependent on science and technology, in which hardly anyone knows anything about science and technology. This is a prescription for disaster. We might get away with it for a while, but sooner or later this combustible mixture of ignorance and power is going to blow up in our faces.

Imagination will often carry us to worlds that never were. But without it we go nowhere.

Carl Sagan

Automated mass spectrometry-based metabolomics data processing by blind source separation methods

by

Xavier Domingo-Almenara

Submitted to the Department of Electrical and Automation Engineering (DEEEA)
on July 2016, in fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

One of the major bottlenecks in metabolomics is to convert raw data samples into biological interpretable information. Moreover, mass spectrometry-based metabolomics generates large and complex datasets characterized by co-eluting compounds and with experimental artifacts. This thesis main objective is to develop automated strategies based on blind source separation to improve the capabilities of the current methods that tackle the different metabolomics data processing workflow steps limitations. Also, the objective of this thesis is to develop tools capable of performing the entire metabolomics workflow for GC–MS, including pre-processing, spectral deconvolution, alignment and identification. As a result, three automated strategies for spectral deconvolution were developed based on blind source separation methods. These methods were embedded into two computation tools able to automatedly convert raw data into biological interpretable information and thus, allow resolving biological answers and discovering new biological insights. The tools were implemented in a modularized manner and their structure was standardized in a single S4 method known as MetaboSet.

First, an independent component regression (ICR) for GC–MS compound identification as an alternative to multivariate curve resolution (MCR–ALS) was introduced. Whereas the typical approach of the ICA-based methods for GC–MS data processing was based on considering the spectra as the independent component in the chromatogram, in this ICR implementation the concept of independence was twisted: compound profiles were targeted as the independent source of the chromatographic mixture, as opposite to the spectra. Also, an orthogonal signal deconvolution (OSD) approach using principal component analysis as an alternative to the traditional least squares approach was introduced, allowing the extraction of refined spectra when compounds elute under the influence of biological matrices, compound co-elution or other types of noise.

Second, independent component analysis - orthogonal signal deconvolution strategy was proposed for the automated resolution of chromatographic signals in comprehensive gas chromatography – mass spectrometry. The ICA–OSD method was

proposed as an effective method to enhance the spectral deconvolution capability and to increase the processing speed when applied for the automated compound deconvolution in chromatographic signals.

Third, the application of multivariate algorithms, e.g., MCR–ALS or ICA-based approaches, in GC–MS data involve segmenting the chromatogram into regions or windows, which may lead to failure in the detection of compounds. Thus, we proposed the application of ICA–OSD and MCR–ALS through a moving window to avoid the usual practice of chromatographic segmentation into regions or windows.

Fourth, despite the existence of different pieces of free and commercial software for GC–MS data analysis, none of these allow the execution of an integrated workflow that includes spectral deconvolution and alignment, followed by the identification and quantification of metabolites in the same application, and implemented in a modularized and standardized manner. This still leads many researchers to implement separate software for each process, and tedious manual workflows for data processing. In this thesis, eRah was designed to fill this gap. Also, while univariate peak–picking approaches are focused on the ion fragment peak as the analysis entity, multivariate methods such as MCR–ALS or ICA aim at extracting the spectra from GC–MS data by taking advantage of the inherent fragment-redundancy in mass spectrometry. However, multivariate methods performance depend, to a greater degree, on an appropriate estimation of the number of components to build the multivariate model. In eRah, we introduced a multivariate compound detector to detect compounds instead of peaks. We later used OSD to determine the compound spectra. The tandem application of the multivariate compound detector by local covariance (CMLC) with OSD allowed the spectral deconvolution of compounds in GC–MS mixtures without the use of factor analysis techniques. eRah was demonstrated to be capable of robustly conducting the complete metabolomics workflow.

Fifth, a tool called BaitMet was introduced to take advantage of the knowledge provided by metabolomics spectral libraries to process full scan GC–MS chromatograms in a driven manner, and with the possibility of standardize the retention times without the use of internal standards. BaiTMet operates under the assumption that the retention time relation between metabolites naturally found in the samples can be used to predict their respective retention indexes. BaiTMet is an R package for high-throughput quantification of compounds of an entire MS library into GC–MS data. BaiTMet was able to identify compounds by the standardization of retention time without mixing internal standards in the samples. Moreover, BaiTMet is also compatible with the use of internal standards mixed in the samples, and use them to characterize the RI/RT curve in each chromatogram.

Acknowledgments

Crec que un doctorat no només implica una sèrie de resultats i publicacions, sinó aprendre de diferents persones i experiències que m'han influenciat tan professionalment com personalment. A totes aquestes persones que m'han ajudat a arribar fins aquí i que han contribuït a que sigui millor investigador, els hi dedico les següents paraules:

Al Jesus i l'Àlex, els meus directors de tesi. Al Jesus, t'agraeixo el no haver-me tallat les ales, el confiar amb mi, ajudar-me, motivar-me, valorar-me, que m'hagis encoratjat i animat quan no ho veia clar i haver estat sempre allí quan ho he necessitat. Et vull agrair el teu tracte personal, el qual ens ha permès comunicar-nos amb molta confiança, la teva visió oberta i que hagis estat sempre obert a col·laborar i fer més projectes.

A l'Àlex, et vull agrair el que m'hagis ensenyat a ser el que sóc. T'agraeixo la direcció que m'has fet de la tesi, la teva paciència, motivació i interès en tot el procés i que m'hagis parat els peus quan tocava. Has estat la motivació que m'ha fet espavilar per poder estar a la teva alçada. M'has ensenyat com fer les coses bé, i ha estat un plaer aprendre d'una persona que sap treure el millor dels doctorands.

A l'Oscar, t'agraeixo el que hagis confiat amb mi com ho has fet, i tot el teu suport desinteressat que m'has donat durant aquest anys i sobretot durant l'últim tram d'aquest camí. Amb tu he après un altre idioma, el de la metabolòmica i ha estat un honor aprendre de tu. Crec que tens una manera de tractar els doctorands fantàstica, amb respecte i proximitat, i això ha fet fantàstic el dur a terme els projectes que em fet junts.

Al Xavier, et vull agrair el teu suport i la confiança no només amb mi sinó amb els estudiants de doctorat. Amb tu he trobat un líder proper, motivat, amb visió i amb ganes de fer més i més coses.

A la Mariona, t'agraeixo tot el que m'has ajudat, com m'has inspirat i motivat per seguir fent el que faig. Amb tu he trobat una persona que sempre ha estat aquí per discutir científicament, amb algú dels pocs que parlen a la perfecció els dos

llenguatges, els dels metabolomics i els dels computacionals, i que sempre ha valorat la meva feina. Espero poder seguir treballant amb tu amb el futur.

A la Noelia, t'agraeixo el teu suport, ajuda, i el que m'hagis a ensenyat a escriure millor. El teu esperit crític ha tingut un paper clau en la publicació dels meus papers. He trobat una persona amb la qual és un plaer treballar, amb motivació, amb ganes d'aprendre i d'ensenyar. He tingut molta sort d'haver pogut comptar amb tu.

A la Sara, t'agraeixo tot el teu interès i la teva motivació. És tot un plaer poder comptar amb una persona que sap com transmetre les idees quan un no entén tant bé l'idioma dels metabolomics.

Al Gabriel, t'agraeixo el que m'hagis acollit al teu grup i el teu tracte personal. T'agraeixo la teva motivació en el meu projecte i l'ajuda i esforç per millorar-lo.

A la Rosa i a la Sílvia, em falten paraules per agrair el vostre suport desinteressat i l'ajuda que meu donat durant aquests anys. He tingut molta sort de poder treballar amb vosaltres. Mil gracies per la vostra ajuda!

Als meus compis de PhD, el Pepe, Rubén, Dídac, Dani, Pere, Sònia, Míriam, Jordi, Núria, Roger i Rocío. Difícilment hagués trobat companys millors. Amb vosaltres he guanyat uns amics per sempre, i amb el quals he pogut aprendre, compartir i discutir científicament. Gràcies per ser com sou!

A la resta de la gent de la plataforma o del departament, a la Serena, Nico, Miguel, Lorena, Raul i Oriol. Gràcies pel vostre suport i amistat, i molta sort en els vostres projectes!

To all the people in Universiteit van Amsterdam, and specially to Rosa, Andrei, Michael, Martin, Marta, Jana, Petra, Michelle and Hendrick, and to my neighbors Paul and Alex. Thank you all for being such a nice people. It was a pleasure to stay with you and learning from you. The best of luck to you all!

Als meus amics, l'Eli, el Bernardo, la Carmen i el Jan, el Marc i el Jose, el Ricard, i als meus compis de grup, el Kilian, Javi, David, Nano, Víctor, Barrero, i en record al nostre germà Sputch. Gràcies per ser com sou i pel vostre suport durant aquests anys!

Als meus germans petits el Bernat, el Jan i la Karina i als meus germans grans

l'Aleix, la Marina i l'Anna. Sense vosaltres no seria qui sóc.

A la meva família, els meus pares, els meus avis, els meus tiets, i a la meva cosina Marina, que sense el seu suport durant aquests anys no hagués arribat fins aquí.

A la Bet, pel teu suport, per voler compartir la vida amb mi i estar sempre al meu costat, per ser el millor que m'ha passat mai.

De tot cor, gràcies.

Xavi

Contents

List of Figures

List of Tables

Abbreviations

BSS	Blind Source Separation
PCA	Principal Component Analysis
SVD	Singular Value Decomposition
MCR–ALS	Multivariate Curve Resolution – Alternating Least Squares
ICA	Independent Component Analysis
ICR	Independent Component Regression
LS	Least Squares
OSD	Orthogonal Signal Deconvolution
CMLC	Compound Match by Local Covariance
GC–MS	Gas Chromatography – Mass Spectrometry
GC×GC–MS	Comprehensive Gas Chromatography – Mass Spectrometry
LC–MS	Liquid Chromatography – Mass Spectrometry
EI	Electron Impact
MS	Mass Spectra
TOF	Time of Flight
GMD	Golm Metabolome Database
NIST	National Institute of Standards and Technology
HMDB	Human Metabolome Database
RT	Retention Time
RI	Retention Index
IS	Internal Standards
TMS	Trimethylsilyl

Chapter 1

Introduction

1.1 Metabolomics

Metabolomics [1, 2] is the profiling of metabolites in biofluids, cells and tissues, and it is routinely applied as a tool for biomarker discovery [3]. Metabolomics is now widely used to obtain new insights into human, plant and microbial biochemistry, as well as in drug discovery, nutrition research and food control through the study of the organism's metabolome. The metabolome is typically defined as the collection of low weight molecular compounds - in a cell, tissue or organism - that are chemically transformed during metabolism. Metabolites are considered the terminal downstream product of the genome and, as such, they provide a functional readout of cellular state, which allows linking cellular pathways to biological mechanisms [4, 5].

Metabolomics is one of the four most representative -omic sciences: genomics, with epigenomics as one of its important branches, transcriptomics and proteomics (Figure 1-1). Metabolomics is considered the one that comes closest to expressing phenotype, providing a chance to look at genotype-phenotype as well as genotype-envirotype relationships [6]. This is due to the fact that, unlike genes and proteins, the functions of which are subject to epigenetic regulation and post-translational mod-

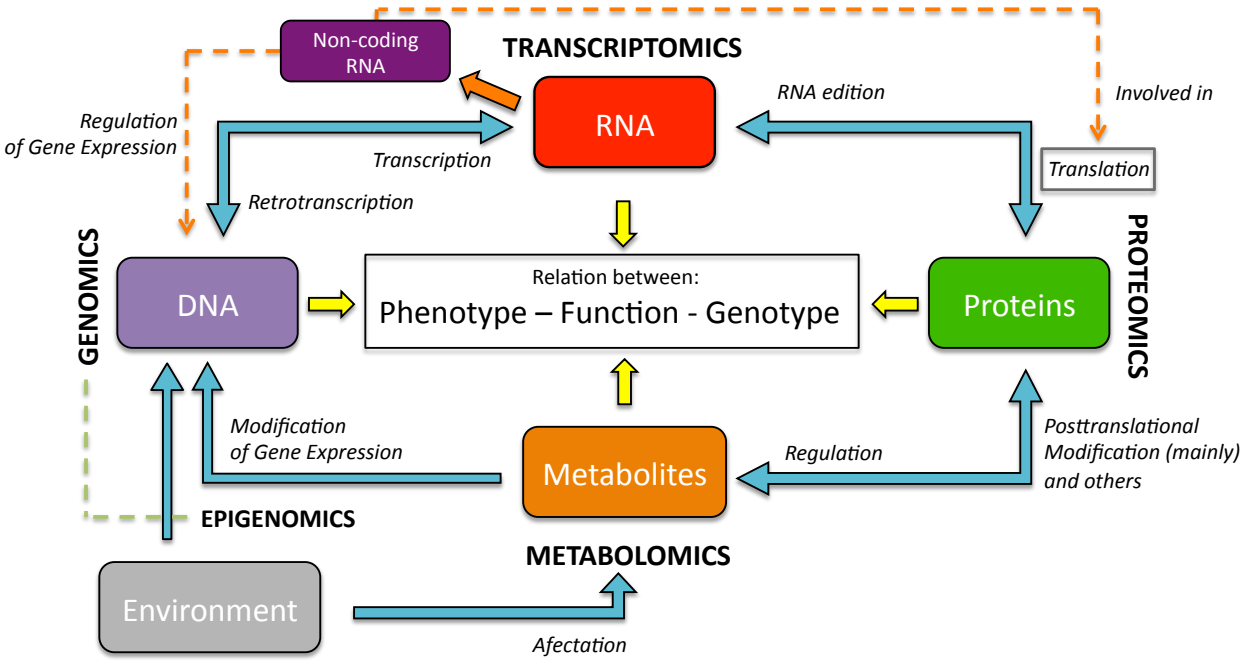

Figure 1-1: Scheme showing the relation between the different -omic sciences and the relation between DNA, RNA, proteins and metabolites. The genes (genomcis) are transcripted into RNA (transcriptomics), which are translated to proteins (proteomics). We have to take into account that not all the genome is transcripted or expressed, and that not all the RNA is translated. Besides, the non-coding RNA is involved in transcription and translation regulation. The non-coding RNA, also known as 'junk RNA', comes from a very large part of the genome that is not expressed and, in the past, it was believed not to have any biological function. Although it seems that some non-coding RNA is the product of spurious transcription, it has recently been shown that it plays an important role in the translation of proteins and the regultation of the gene expression.

ifications respectively, metabolites serve as direct signatures of biochemical activity and are therefore easier to correlate with the phenotype [4]. Metabolites are also more dynamical entities as their concentration may change dramatically in small periods of time. In short, while genomics, transcriptomics and protemics explain what may happen, metabolomics explains what is happening in the organisms [7]. Despite that, and although the metabolic profile can be seen as the ultimate expression of the genome, the metabolome state does not only depend on the complex interactions and processes of the genes, transcripts and proteins, but it is also affected by the environment, including commensal microorganisms, nutritional factors, environmental agents, and drugs or toxic substances [8, 9].

One of the most important differences between metabolomics and the rest of the -

omics sciences is that, whereas the human genome has been sequenced and it is known (estimated in 25 K genes), and also the number of transcripts (150 K) and proteins (1 M) is now well estimated, the exact size of the human metabolome at the present time is still unknown [10]. And although the total number of metabolites registered to the date, for example, at the Human Metabolome Database [11] is 42 K, most of them are (known) unknown metabolites, which are presumed to be breakdown products and molecules transformed by enzyme or microbial activity [12]. Identification of unknown metabolites and data interpretation and translation of analytical platforms results into a biologically interpretable information (data processing) are still some of the most important challenges in metabolomics [13]. These bottlenecks have prevented metabolomics from evolving as fast as the other omic sciences [14, 15, 16] .

1.2 Analytical platforms

Nowadays, different analytical techniques including nuclear magnetic resonance spectroscopy (NMR) or other hyphenated techniques such as liquid chromatography coupled to mass spectrometry (LC-MS), are used for compound profiling in metabolomics. In fact, different analytical techniques are needed as no analytical platform covers the full metabolome, and thus they have to be combined. However, the proof of concept for what we now know as mass spectrometry-based metabolomics was reported in 1966 by Dalgliesh *et al.* [19], which conducted the first GC/MS experiment to separate a wide range of metabolites occurring in urine and tissue extracts. Later in 1971 Horning *et al.* [20] introduced the term metabolic profiles, and along with Pauling and Robinson led to the development of GC–MS methods for monitoring metabolites in biological samples through the 1970s [21, 22]. In this thesis, two analytical platforms were used: GC–MS and GC×GC–MS

1.2.1 GC–MS

Gas chromatography – mass spectrometry has been a long-standing approach used for metabolite profiling of volatile and semi-volatile compounds due to the widespread use of electron impact ionization (EI) mode. EI is a hard ionization technique that has been historically standardized at 70 eV. Unlike soft ionization techniques such as ESI [17] or MALDI [18], EI is a highly reproducible ionization process across many different platforms.

In GC–MS, volatile metabolites can be directly analyzed whereas semi-volatile compounds can only be detected by a previous silylation process of the polar groups (derivatization) [23]. The main objective of derivatization is to block the polar functional groups, leading to an increase in volatility and reduction in polarity. Also, due the robustness and reproducibility of the electron impact (EI) ionization technique, the extensive fragmentation allows a straightforward identification of compounds through spectral libraries. These distinctive advantages have contributed to establish GC–MS as a robust platform for the quantitative analysis of volatile and semi-volatile metabolites.

1.2.2 Comprehensive GC–MS

Gas chromatography separates the metabolites while passing through a single chromatographic column. However, when two or more compounds do not completely separate chromatographically, those compounds are known to be co-eluted, i.e., overlapped with other compounds or matrix components, providing several problems with the identification and quantification of the compounds. To overcome this obstacle, comprehensive gas chromatography - mass spectrometry (GC×GC–MS) [24, 25] emerged over the last two decades as a powerful analytical technique. In the comprehensive GC×GC, the entire first dimension column eluate is further analyzed in the second dimension column. Therefore, the sample pass through two chromato-

graphic columns with orthogonal retention properties, which improves the compound separation space, improving thus the chromatographic resolution and it leads to an increased compound detection capacity as co-elution is diminished.

GC×GC–MS has a higher chromatographic separation power, a broader dynamic range and lower detection limits, and thus it should be the preferred technique for metabolomics analysis [26]. However, the lack of tools that allow retrieving interpretable information (list of compounds and their concentration for each sample) from raw data has limited the application of GC×GC–MS in metabolomics. This lack of tools can be explained due to difficulties of processing the enhanced data complexity of GC×GC–MS respect to the one-dimensional GC.

1.2.3 Mass detectors

Although GC and GC×GC can be used in combination with a wide variety of detectors, the most commonly-used detection technique coupled with GC that is used nowadays is mass spectrometry (MS) detectors. In untargeted GC-based metabolomics, the full scan mode of MS is employed for identifying compound structures, whereas the selective ion monitoring (SIM) is usually employed in targeted metabolomics to achieve higher sensitivity for (selective) quantification. The electron impact is the most widespread mode used in GC–MS and GC×GC–MS, which is a highly reproducible ionization process across many different platforms and that has been historically standardized at 70 eV.

Among the different MS analyzers, the single quadrupole, (quadrupole) time of flight (qTOF/TOF) or triple quadrupole (QqQ) are the most popular in GC. Both TOF and qTOF are the most employed detectors in GC–MS, while the best GC×GC equipment (LECO Corp.) is attached to a TOF detector. While qTOF–MS allows full-scan screenings with high mass resolution and accuracy, TOF detectors allow the full-scan screening only in nominal mass.

1.3 Metabolomics experimental workflow

Metabolomics experiments involve carrying out a series of steps that generally comprises the following workflow: experimental design, sample preparation, analysis of samples, data processing and data analysis [4]. Experimental design aims at preparing the proper specifications of the experiment - analytical platform to use, the type and the number of samples - to answer the biological hypothesis or to discover new biomarkers. After that, sample preparation aims at the extraction of the metabolites of interest for its posterior analysis through the analytical platform chosen. Also, and concretely in GC–MS or GC×GC–MS, metabolites are derivatized to enhance its volatility. Next, the data generated by the analytical platform has to be processed through the so-called metabolomics data processing workflow, which is detailed in the next chapter. Finally, data analysis aims at discovering the up-regulated or down-regulated metabolites between classes or physiological conditions and also the interactions between them through (univariate/multivariate) statistical tests and methods.

Bibliography

[1] Nicholson, J.K., Lindon, J,C., Holmes, E. Metabonomics: understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data *Xenobiotica*, 29 (1999), 1181–1189.

[2] Zhang A., Sun H., Wang X. (2012) Serum metabolomics as a novel diagnostic approach for disease: a systematic review, *Analytical and Bioanalytical Chemistry*, **404**, 1239-1245.

[3] Johnson C.H., Ivanisevic J., Siuzdak G. Metabolomics: Beyond Biomarkers and Towards Mechanisms. *Nature Reviews Molecular Cell Biology*, In press.

[4] G.J. Patti, O. Yanes, G. Siuzdak Innovation: Metabolomics: the apogee of the omics trilogy. *Nature Reviews Molecular Cell Biology*, 13 (4):63–269, 2012.

[5] Roberts LD1, Souza AL, Gerszten RE, Clish CB. Target metabolomics *Curr Protoc Mol Biol*, (2012) Chapter 30:Unit 30.2.1-24

[6] Baraldi E, Carraro S, Giordano G, Reniero F, Perilongo G, Zacchello F. Metabolomics: moving towards personalized medicine. *Italian Journal of Pediatrics*, 35 (2009), 30.

[7] Schmidt C. W. Metabolomics: What's Happening Downstream of DNA. *Environmental Health Perspectives*, 112(7), (2004).

[8] Nicholson JK, Wilson ID. Opinion: understanding 'global' systems biology: metabonomics and the continuum of metabolism. *Nat Rev Drug Discov* 2 (2003) 668-676.

[9] Nicholson JK, Holmes E, Lindon JC, Wilson ID. The challenges of modeling mammalian biocomplexity. *Nat Biotechnol* 22 (2004) 1268–1274.

[10] David S. Wishart, Timothy Jewison, An Chi Guo, Michael Wilson, Craig Knox, Yifeng Liu, Yannick Djoumbou, Rupasri Mandal, Farid Aziat, Edison Dong, Souhaila Bouatra, Igor Sinelnikov, David Arndt, Jianguo Xia, Philip Liu, Faizath Yallou, Trent Bjorndahl, Rolando Perez-Pineiro, Roman Eisner, Felicity Allen, Vanessa Neveu, Russ Greiner, and Augustin Scalbert. Hmdb 3.0the human metabolome database in 2013. *Nucleic Acids Research* 2012.

[11] Wishart DS, Tzur D, Knox C, et al. HMDB: the Human Metabolome Database. *Nucleic Acids Res.* 35 (2007).

[12] Chris Tachibana. What's next in 'omics: The metabolome. *Science*, 345(6203), (2014), 1519–1521.

[13] Aretz I., Meierhofer D. Advantages and Pitfalls of Mass Spectrometry Based Metabolome Profiling in Systems Biology. *Int J Mol Sci.* 17 (2016) 632.

[14] L.W. Sumner, A. Amberg, D. Barrett, M.H. Beale, R. Beger, C.A. Daykin, et al. Proposed minimum reporting standards for chemical analysis. *Metabolomics*, 3 (2007) 211–221.

[15] D.S. Wishart What's next in 'omics: The metabolome. *Bioanalysis*, 3 (2011) 1769–82.

[16] Vinaixa M, Schymanski EL, Neumann S, Navarro M, Salek RM, Yanes O. Mass spectral databases for LC/MS and GC/MS-based metabolomics: state of the field and future prospects. *Trends in Anal. Chem.* 78 (2016) 23–25.

[17] Fenn, J. B. Electrospray wings for molecular elephants (Nobel lecture). *Angew. Chem. Int. Ed. Engl.* **2003**, 42, 3871–3894.

[18] Karas, M.; Ralf, K. Ion formation in MALDI: the cluster ionization mechanism. *Chem. Rev.* **2003**, 103, 427–440.

[19] Dalgliesh, C. E.; Horning, E. C.; Horning, M. G.; Knox, K. L.; Yarger K. A gas-liquid-chromatographic procedure for separating a wide range of metabolites occuring in urine or tissue extracts. *Biochem. J.* **1966**, 101, 792–810.

[20] Horning, E. C.; Horning, M. G. Metabolic profiles: gas-phase methods for analysis of metabolites *Clin. Chem.* **1971**, 17, 802–809.

[21] Teranishi, R.; Mon, T. R.; Robinson, A. B.; Cary, P.; Pauling, L. Gas chromatography of volatiles from breath and urine *Anal. Chem.* **1972**, 44, 18–20.

[22] Matsumoto, K. E.; Partridge, D. H.; Robinson, A. B.; Pauling, L.; Flath, R, A.; Mon, T. R.; Teranishi, R. The identification of volatile compounds in human urine. *J. Chromatogr. A*. **1973**, 85, 31–34.

[23] W.B.Dunn,D.Broadhurst,P.Begley,E.Zelena,S.Francis-McIntyre,N.Anderson, M. Brown, J.D. Knowles, A. Halsall, J.N. Haselden, A.W. Nicholls, I.D. Wilson, D.B. Kell, R. Goodacre, Human Serum Metabolome (HUSERMET) Consortium, Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry, *Nat. Protoc.* 6 (7) (2011) 1060–1083

[24] Luigi Mondello, Peter Quinto Tranchida, Paola Dugo, and Giovanni Dugo. Comprehensive two-dimensional gas chromatography-mass spectrometry: a review. *Mass Spectrometry Reviews*, 27(2):101–124, April 2008.

[25] John V. Seeley and Stacy K. Seeley. Multidimensional Gas Chromatography: Fundamental Advances and New Applications. *Analytical Chemistry*, 85(2):557–578, 2012.

[26] Maud M. Koek, Frans M. van der Kloet, Robert Kleemann, Teake Kooistra, Elwin R. Verheij, Thomas Hankemeier. Semi-automated non-target processing in GCxGC-MS metabolomics analysis: applicability for biomedical studies. *Metabolomics*, 7 (2011), 1–14.

Chapter 2

State of the Art

2.1 Data processing workflow in GC/MS and GC×GC/MS-based metabolomics

In order to identify and extract quantitative information of metabolites across multiple biological samples, workflows for GC–MS and GC×GC–MS data processing are needed. Data processing in gas and comprehensive gas chromatography – mass spectrometry share common steps. These steps include a pre-processing, comprised of noise filtering and baseline removal of chromatograms, peak-picking or deconvolution of compounds and their alignment across samples, and the identification of metabolites by spectral library matching.

2.1.1 Pre-processing

A chromatogram is usually considered to be composed of three additive components: signal, baseline and noise [1, 2]. Hence, all GC–MS and GC×GC–MS chromatograms are usually affected by baseline drift or instrumental noise (Figure 2-1), and the pre-processing of chromatograms by noise filtering and baseline correction may improve the posterior deconvolution and alignment performance. One of the most stablished

method to determine the best baseline model is the asymmetrically weighted least squares regression (ALS) [2], which provided a basis for improvements or variations based on or compared to the original algorithm [3, 1]. Also, different studies have focused on the pre-processing of data as a method for the subsequent extraction of informative features [2, 4, 5, 6].

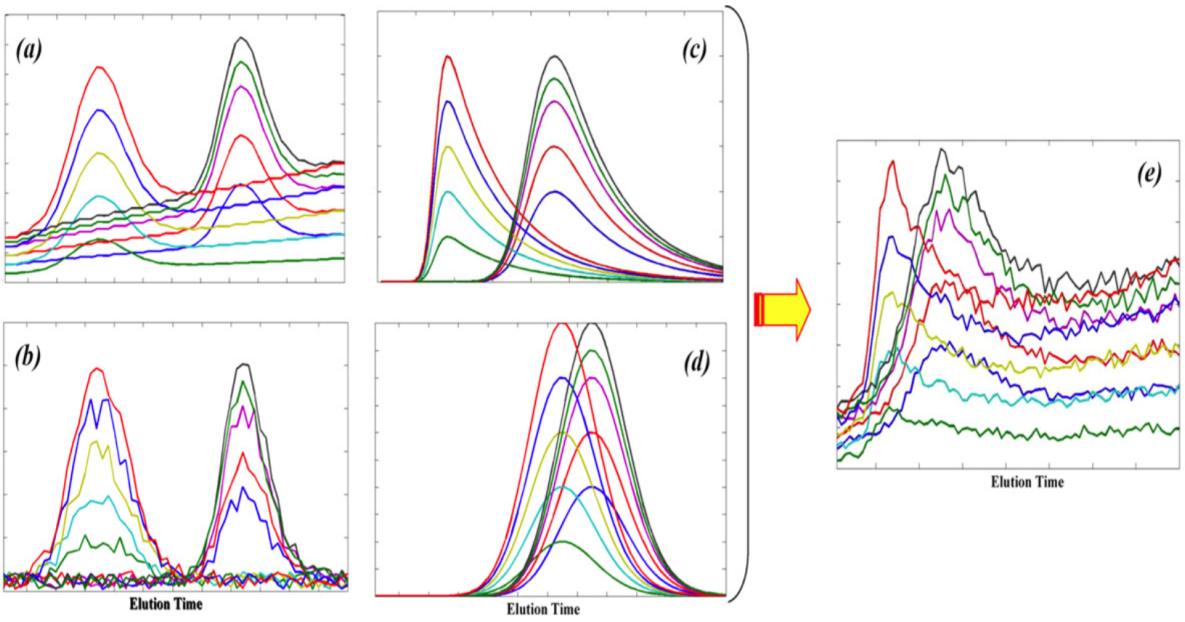

Figure 2-1: Typical GC–MS artifacts or situations: (a) baseline drifts, (b) low S/N ratios due to low concentrated compounds, (c) changes in the peak shape (e.g. peak-tailing or peak-fronting), (d) co-elution of compounds, and (e) combination of all the situations. Picture from Jalali-Heravi et al. [7]

Noise filtering can be conducted by multiple methods imported from the signal processing field [8], by linear and non-liner filters such as the typical finite impulse response (FIR) filters or moving average filters. However, the most widespread filter used in chromatographic data is the Savitzky–Golay filter [9]. Savitzky and Golay were interested in smoothing mass spectral data, and they demonstrated that least squares smoothing reduces noise while maintaining the shape and height of the gaussian-shaped peaks [10]. This is of special importance in chromatography, since some filters are more destructive, and they smooth the data without maintaining the original shape. Savitzky and Golay's paper is one of the most cited papers in

the journal Analytical Chemistry [11]. Later, Eilers claimed that the Whittaker filter [12], based on penalized least squares, was a better and faster alternative to Savitzky–Golay filter.

2.1.2 Peak picking or spectral deconvolution

Data processing in untargeted GC–MS and GC×GC–MS-based metabolomics involve detecting signals that will be ultimately related to metabolies. To do so, methods for data processing can be divided into two main categories: methods based on peak-picking and methods for compound extraction through multivariate algorithms and spectral deconvolution.

Univariate approaches

This first category involves detecting all relevant fragment ion peaks in the spectra to subsequently align them across multiple samples [13, 14] and discover statistical peak variations between experimental groups. The quantitative variables provided by these methods are not based on the compound spectra, but the m/z value, retention time window and area of fragment ion peaks.

The method for what we now know as peak-picking approach for high-throughput data analysis was reported by Smith *et al*, where they introduced XCMS [15], a computational tool for processing LC-MS data. A more sensitive peak-picking called centWave was later reported [16] (Figure 2-2). Smith *et al* popularized the peak-picking approach, and different tools using the same principle were later reported (generally as a part of computational tools and therefore discussed in Section 2.2). Advantages of these methods are that they should be more reliable, as no deconvolution is performed and therefore the true area of the m/z peak is registered. Despite that, no spectra is deconvolved and identification of metabolites is the main bottleneck

of peak-picking approaches. Moreover, peak alignment across samples is conducted without any spectral information, and thus, only the retention time distance between peaks is used. In that sense, some other computational tools attempted to overcome this limitation by grouping the different peaks (based on their shape similarity or peak correlations) into compound spectra. These computational tools are discussed also in Section 2.2.

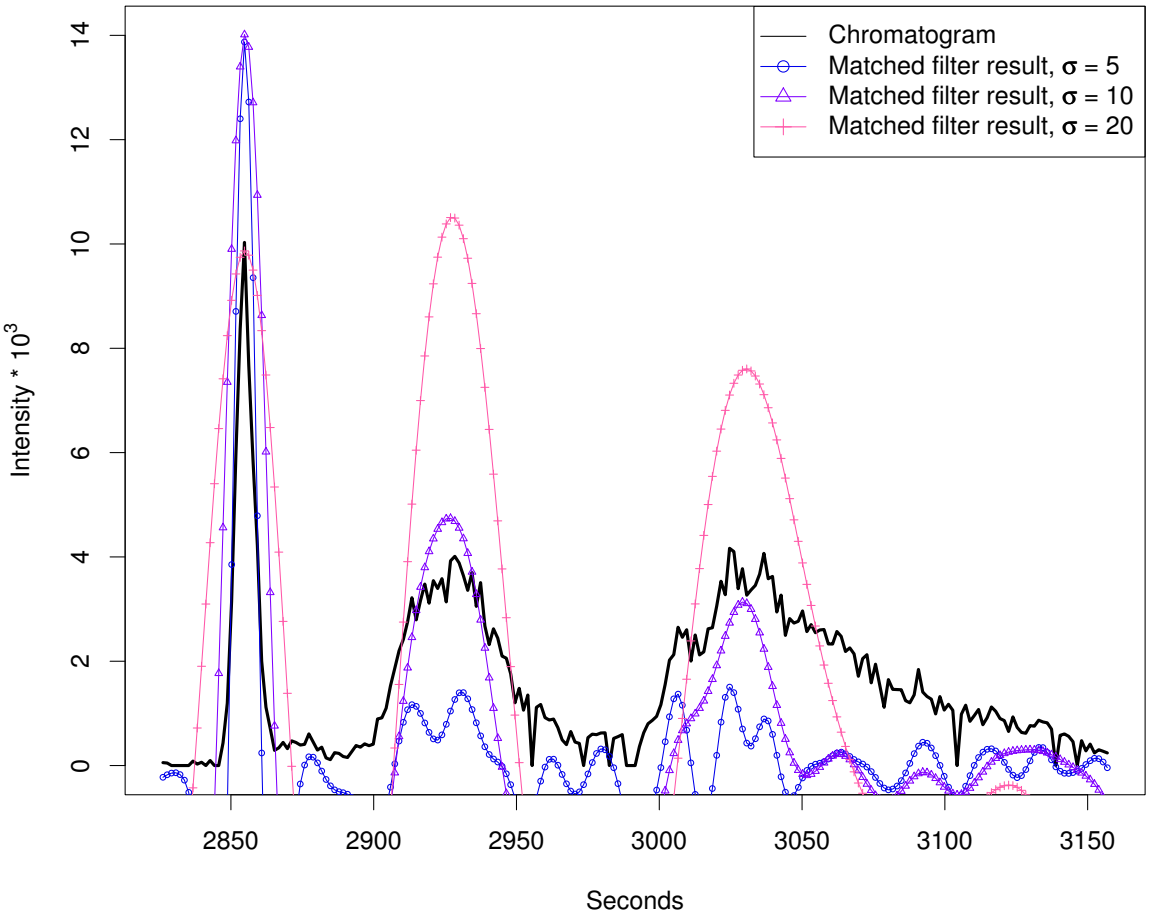

Figure 2-2: This picture illustrates the centWave [16] matched filter effect. Although this is a HPLC/ESI-QTOF-MS of a A. thaliana leaf extract chromatogram, it illustrates the operation of a typical peak-picking approach despite the analytical platform. The match filter filters the data leading to the detection of different peaks. However, each m/z channel is processed separately and thus, the overall spectral information (spectral co-variance) is not taken into account. Picture from Tautenhahn et al. [16]

Multivariate approaches

Multivariate approaches focuses on the compound as the analysis entity, as opposed to the use of individual fragment peaks. Compounds are quantified and identified on the basis of a multivariate deconvolution process [17] that extracts and constructs pure compound spectra from raw data (Figure 2-3). Methods such as multivariate curve resolution – alternating least squares (MCR-ALS), independent component analysis (ICA) or parallel factor analysis (PARAFAC) have been used to process GC–MS and GC×GC–MS data [18].

Advantages of those methods include that the spectra are directly extracted from data by taking advantage of the inherent fragment-redundancy in mass spectrometry. This fragment-redundancy means that for each compound, different fragments or ions elute at the same retention time and with the same elution profile. Then, chromatographic - mass spectrometry data is characterized by a temporal and spectral redundancy which yields to a natural co-variance of fragments that can be used to extract more efficiently the spectra associated to each metabolite. This co-variance redundancy is more emphasized in GC×GC–MS data, where, due to the second retention time dimension, compounds elute in more than one modulation cycle. In these cases, tensor decomposition can be applied.

However, multivariate methods performance depend, to a greater degree, on an appropriate estimation of initialization parameters and/or the correct estimation of the number of components - the number of distinct multivariate sources from which the data can be built. Different methods have been reported to determine the number of components in a mixture, but the most popular methods include the singular value decomposition (SVD) [19], a cross-validation procedure [20], or the evaluation of the residual sum of squares [21].

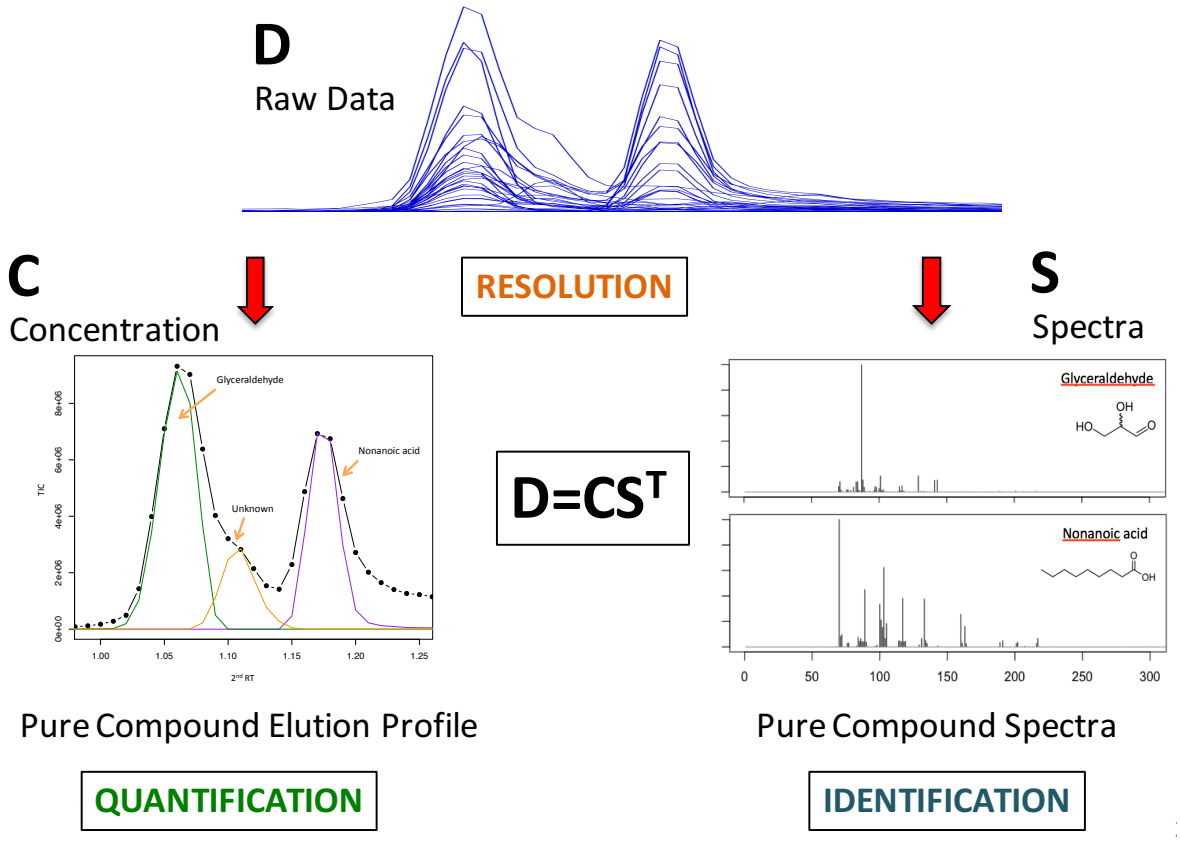

Figure 2-3: Illustration of the bilinear data structure. The bilinear model describes the data (D) as the composition of two matrices (C and S), where C represents the pure chromatographic profiles and S the pure spectra. Generally, multivariate methods aim to decompose the chromatographic mixture into (ideally) quantificable chromatographic profiles and identificable spectra. The top picture represents the raw data (a GC–MS chromatographic segment), and each blue line represents each m/z.

Also, to ensure a correct performance, multivariate two-way methods have to be applied in small regions of the chromatogram [22], and therefore the chromatograms have to be segmented prior to the application of the algorithms. The automation of this segmentation process is a challenging task as it implies separating between informative data and noise from the chromatogram.

Multivariate curve resolution – alternating least squares: Multivariate curve resolution (MCR) is an historical chemometric method, widely used for the resolution of chromatographic mixtures. The first study that inspired MCR was reported in 1971 by Lawton and Sylvestre [23], but the most reported and popular

MCR algorithm, known as multivariate curve resolution – alternating least squares (MCR–ALS), was proposed by Tauler in 1995 [24]. MCR–ALS has been reported for the resolution of GC–MS, GC×GC–MS, and also of LC–MS signals [25]. MCR–ALS is used to decompose a data matrix containing a mixture of compounds into two matrices containing the resolved pure concentration profiles and pure spectra.

Independent component analysis: Independent component analysis (ICA) is a blind source separation (BSS) method developed in the early 1980s, and it is widely used for the processing of signals of different natures (Figure 2-4) including electroencephalographic records [26, 27, 28], or other biomedical signals, image analysis [29] or sonar applications [30] among others. In chemistry, this method is well-stablished in spectroscopy [31, 32, 33, 34, 35], but its use in chromatography was introduced by Shao *et al* [36]. Since then, several algorithms were developed [37], which used ICA for resolution of chromatographic signals, including mean-field ICA (MF–ICA) [38], post-modification based on chemical knowledge (PBCK) [39], window ICA (WICA) [40] and non-negative ICA [41]. Artificial immune system algorithms involving the use of ICA were also proposed [42].

The aforementioned ICA-based approaches for the resolution of GC–MS signals share a common procedure: first, they use ICA to deconvolve the mass spectrum for each compound in the mixture, i.e., they consider the spectra as the independent source in the chromatograms. After that, each above-mentioned algorithm uses different approaches to determine the elution profile of each compound, since the elution profiles determined by ICA tend to be inaccurate or affected by various ICA ambiguities such as negativity or variance (energy) indetermination [43]. However, ICA-based algorithm efficiency has been questioned by Parastar *et al.* [44], where they claimed that up the existing ICA-based algorithms could be considered as an alternative tool for resolving mixed signals in analytical chemistry only in a limited number of cases.

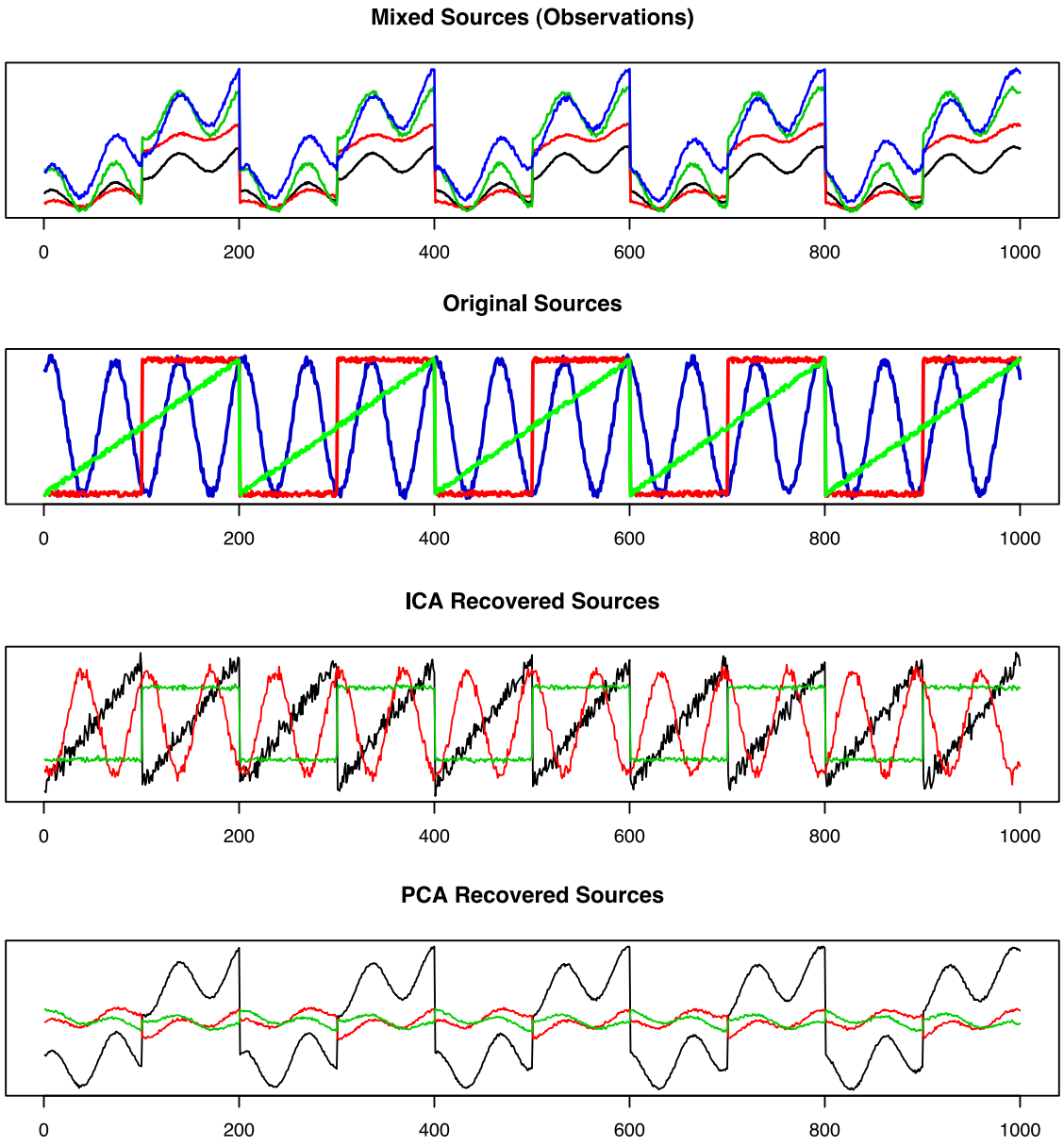

Figure 2-4: Illustration of ICA outcome. In this example, three signals (sine, triangular and square) are mixed. Whereas PCA may find difficulties in recovering the original sources, ICA recovers the qualitative shape of the original signals.

ICA and MCR methods share the same objective, which is the resolution of complex mixtures into pure-components: pure chromatographic profiles and spectra [45]. The most characteristic difference between the methods is that, whereas MCR–ALS resolves a chromatographic mixture by minimizing the residual error between the data and the predicted model, ICA uses another type of measure which is the sta-

tistical independence, and it estimates the original compound sources by maximizing the independence between components.

Tensor decomposition: Generally, GC–MS data is built in a (two-way) matrix of order (Rt × m/z), where Rt and m/z are the retention time and masses respectively. However, when multiple samples are employed, data can be represented in a (three-way) array (or tensor) of order (Rt × m/z × Chr), where Chr is each chromatogram or sample. Contrarily, in GC×GC–MS, each single chromatogram is composed of a three-way array of order (Rt1 × Rt2 × m/z), where each Rt 1 and 2 are the 1st and 2nd retention times (and where the first corresponds to each modulation cycle). These arrays can be seen as a box in which the ways correspond to the vertical, horizontal and depth axis [46].

In those cases where data is composed by a three dimensional array -tensor-, three-way multivariate algorithms can be applied. Although MCR–ALS can be adapted to resolve a three-way structure by data (row or column wise) augmentation or by data unfolding into multiple two-way matrices [47] - since MCR can not be applied directly to three-way a data structure -, the historically known as parallel factor analysis (PARAFAC) [48] - and later known as Canonical Decomposition (CANDECOMP) or Canonical Polyadic Decomposition (CND) -, is the most popular tensor decomposition method which has been shown to be effective for the resolution of chromatographic signals [49], which generally may achieve similar results to MCR-ALS [50].

Three dimensional arrays can be built from multiple GC–MS chromatograms or a single (or also multiple) GC×GC–MS chromatogram. Tensor decomposition does not only take advantage of the temporal and spectral co-variance of metabolites, but also of the variation across chromatograms or modulation cycles. Tensor decomposition is specially powerful in those cases where similar - or co-eluting - compounds may be difficult to separate just from the information provided by single chromatogram. However, those compounds may have different relative concentrations across chro-

matograms, and tensor decomposition aims to use these differences of variation to resolve data even more efficiently.

The main drawbacks of tensor decomposition problems are that generally, the data has to be trilinear, which means that compounds appearing in multiple samples should be well aligned, and this usually does not occur. In that sense, methods such as Tucker3 [51] were proposed to resolve non-trilinear three-way datasets. In the same way that two-way methods, three-way methods depend also on the correct estimation of the number of factors or components, and this parameter critically affects its outcome.

2.1.3 Alignment of metabolites

Alignment of metabolites aim to correct the retention time variation of eluting compounds, facilitating the relative quantification and comparison of compounds across samples. In that sense, metabolites in GC–MS have to be aligned between chromatograms, whereas - and due to the second retention time dimension - metabolites appearing in GC×GC/MS data have to be aligned within and between chromatograms.

Currently, alignment algorithms fall into two distinct branches: first, there are several methods based on the pre-alignment of the chromatogram before the application of any data processing and thus, the application of these methods is usually considered as data pre-processing. The aim of these algorithms is to facilitate the posterior processing of the data, or to directly register metabolite differences among samples since metabolites are expected to be already aligned. Generally, algorithms for pre-alignment of chromatographic signals are based on dynamic time warping [52, 53].

The second category include the alignment of compound or peak lists. These algorithms align the metabolites after their are deconvolved from data by clustering

them on the basis of the retention time distance - and spectral similarity if spectra is obtained - between metabolites. Usually, those algorithms are reported as part of computational tools, but some independent methods focused exclusively in what is known as alignment of peaks (list) or spectra have been reported [54, 55, 56].

Disctinctively in GC×GC–MS, the same metabolites appearing in different modulation cycles have to be also aligned within the chromatogram. Methods that tackle this matter have also been reported [57].

2.1.4 Identification

The extensive fragmentation of the electron impact ionization allows obtaining distinctive and highly reproducible fragmentation patterns - spectra - for each analyzed metabolite. Therefore, and although different levels of identification can be applied [58], identification of metabolites is generally conducted by spectral similarity with reference pattens provided by spectral libraries, and often also by the comparison of a physico-chemical property (e.g., chromatographic retention time). Those two molecular properties are known to be orthogonal, and although stereoisomers are difficult to be distinguished by only these two properties [59], the combination of both allows a reliable metabolite identification in most of the cases.

However, compound retention time is not a reproducible variable, as it depends on the chromatographic method employed in each case. To solve that, retention indices (RI) are used instead [59]. In RI, the retention time is standardized and thus it is given relative to the retention time of known standards - typically fatty acid methyl esters (FAME) or alkanes (ALK) -mixed with the samples. The retention times of those standards vary according to the method, but the RI - the relative chromatographic distance between metabolites and known standards - is a very reproducible feature.

Methods for spectral matching include the widely used cosine dot product/Pearson's correlation [60] or the Stein and Scott's composite score [61], among others [62]. For

retention index (RI) comparison, methods including linear interpolation and known as van den Dool [63] and Kováts [64] methods or spline interpolation are used to standardize the empirical retention time into RI.

Different free and commercial spectral libraries are currently used in metabolomics, which provide information on chemical structures, physico-chemical properties, spectral/fragmentation patterns, biological functions and pathway mapping of metabolites [58]. Among those, libraries with electron impact fragmentation patterns include the Golm Metabolome Database (GMD) [65, 66], the Human Metabolome Database (HMDB) [67], the MassBank repository [68] or the NIST and Wiley libraries. The NIST is the richest library by the number of metabolites recorded. However, NIST EI-MS spectra are recorded using single quadrupole (SQ) mass spectrometer. The advent of other mass detectors such as GC–(q)TOF and GC–Orbitrap with other operational principles than SQ raises the need for EI-based exact mass libraries acquired with these new detectors. One of the most significant variations of spectra from the same compound acquired using SQ and TOF detectors is the difference in relative ion intensities. This directly translates into lower matching scores when querying SQ based spectral libraries with data acquired using TOF detectors. Therefore, GMD is the richest library currently available for GC-TOF detectors. Of note, GMD is however populated using TOF detectors with nominal mass.

2.2 Implementation of the workflow: computational tools for GC–MS data processing

Current computational approaches for GC–MS data processing in untargeted metabolomics fall into two main categories: tools based on peak-picking, and tools for compound extraction through the so-called curve resolution and spectral deconvolution (Table 2.1). The first category involves detecting all relevant molecular ion peaks in the spec-

tra, align them across multiple samples, and discover statistical peak variations between experimental groups. Representative tools from this category include MZmine [69, 70], MetAlign [71, 72], and XCMS [15, 73]. Although these tools were initially intended for liquid chromatography mass spectrometry (LC–MS) data processing, they can also be used for GC–MS data analysis [74, 75]. The quantitative variables provided by these methods are not the compound spectra, but the m/z value, retention time window and area of individual fragment ion peaks. Thus, compound identification is the main bottleneck of peak-picking approaches. In this regard, tools such as metaMS [76], TagFinder [77], MetaboliteDetector [78] and PyMS [79] attempt to overcome this limitation by grouping the different peaks (based on their shape similarity or peak correlations) into compound spectra, allowing the putative identification of compounds by comparing their mass spectra with a reference MS library.

The second category focuses on the compound as the analysis entity, as opposed to the use of individual molecular ion peaks. Compounds are quantified and identified on the basis of a multivariate deconvolution process [17] that extracts and constructs pure compound spectra from raw data. Representative tools falling into this category include TNO-DECO [80] or ADAP-GC [81]. TNO-DECO uses multivariate curve resolution to extract the compound spectra whereas the deconvolution algorithm of ADAP-GC is based on hierarchical clustering of the fragments shape for compound detection. Furthermore, there are other free softwares, such as AMDIS [60] or BinBase [82, 83], that perform parts of the untargeted GC–MS metabolomics workflow. AMDIS is used to identify compounds in samples by using the NIST library, but it does not include spectral alignment. BinBase uses the spectral deconvolution provided by a proprietary algorithm in the commercial software ChromaTOF (LECO Corporation) to align compounds across samples, and provide compound quantification and identification based on self-constructed or downloadable libraries [84]. Finally, TargetSearch [85] is an R package for library-driven compound profiling that relies on

the retention indexes (RI) and the list of selective masses provided by a reference MS library to quantify a high number target compounds with univariate techniques. It requires mixing internal standards with the samples.

Table 2.1: List of representative softwares for GC–MS data processing

Tool	**Language**	**Anal. Plat.**	**Type/Comment**
XCMS	R	LC/MS, GC/MS	Univariate.
mzMine	Java	LC/MS, GC/MS	Univariate.
MetAlign	C++	LC/MS, GC/MS	Univariate. No open-source.
MetaMS	R	LC/MS, GC/MS	Univariate.
TagFinder	Java	GC/MS	Univariate. No open-source.
TargetSearch	R	GC/MS	Univariate.
MetaboliteDetector	C++	GC/MS	Univariate. No open-source.
PyMS	Python	GC/MS	Univariate.
TNO-DECO	Matlab	GC/MS	Multivariate. Open-source
ADAP-GC	Java	GC/MS	Multivariate. No open-source and currently not available.

Bibliography

[1] M. Lopatka, A. Barcaru, M. J. Sjerps, G. Vivo-Truyols. Leveraging probabilistic peak detection to estimate baseline drift in complex chromatographic samples. *Journal of Chromatography A*, 1431 (2016) 122–130.

[2] M. Daszykowski, B. Walczak. Use and abuse of chemometrics in chromatography. *TrAC Trends Anal. Chem.*, 25 (11) (2006) 1081–1096.

[3] J.J. de Rooi, P.H. Eilers. Mixture models for baseline estimation. *Chemomet. Intell. Lab. Sys.* 117 (2012) 56–60.

[4] S. Castillo, P. Gopalacharyulu, L. Yetukuri, M. Oresic. Algorithms and tools for the preprocessing of LC–MC metabolomics data, *Chemomet. Intell. Lab. Syst.* 108 (1) (2011) 23–32.

[5] D.W. Cook, S.C. Rutan, Chemometrics for the analysis of chromatographic data in metabolomics investigations, *J. Chemomet.* 28 (9) (2014) 681–687.

[6] Fu HY, Li HD, Yu YJ, Wang B, Lu P, Cui HP, Liu PP, She YB. Simple automatic strategy for background drift correction in chromatographic data analysis. *J Chromatogr. A* 1449 (2016) 89–99.

[7] Jalali-Heravi M, Parastar H. Recent trends in application of multivariate curve resolution approaches for improving gas chromatography–mass spectrometry analysis of essential oils. *Talanta* 85(2) (2011) 835–49.

[8] S. W. Smith, The scientist and engineer's guide to digital signal processing, *California Technical Publishing San Diego* (1997).

[9] A. Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures *Anal. Chem*, 36 (8) (1964) 1627–1639

[10] R. W. Schafer What Is a Savitzky-Golay Filter, *IEEE Signal Processing Magazine* 28(4), (2011), 111–117.

[11] L. K. Cynthia, J. V. Sweedler, Celebrating the 75th Anniversary of the ACS Division of Analytical Chemistry: A Special Collection of the Most Highly Cited Analytical Chemistry Papers Published between 1938 and 2012 *Anal. Chem* 85 (0) (2013), 4201–4202.

[12] P. H. C. Eilers A Perfect Smoother *Anal. Chem* 75 (2003), 3631–3636.

[13] Koh, Y.; Pasikanti K. K.; Yap, C. W.; Chan, E. C. Comparative evaluation of software for retention time alignment of gas chromatography/time-of-flight mass spectrometry-based metabonomic data. *J. Chromatogr. A.* 2010, 1217(52), 8308–8316.

[14] Niu, W.; Knight, E.; Xia, Q.; McGarvey, B.D. Comparative evaluation of eight software programs for alignment of gas chromatography–mass spectrometry chromatograms in metabolomics experiments. *J. Chromatogr. A.* 2014, 1374, 199–206

[15] Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.* 2006, 78.3, 779–787

[16] Tautenhahn, R.; Bottcher, C.; Neumann, S. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics.* 2008, 9.1, 1–16.

[17] Du, X.; Steven H. Z. Spectral deconvolution for gas chromatography mass spectrometry-based metabolomics: current status and future perspectives. *Comput. Struct. Biotechnol. J.* **2013**, 4, 1–10.

[18] L. W. Hantao, H. G. Aleme, M. P. Pedroso, G. P. Sabin, R. J. Poppi, F. Augusto. Multivariate curve resolution combined with gas chromatography to enhance analytical separation in complex samples: a review. *Anal. Chim. Acta* 2012 (731), 11-23

[19] J. Diewok, A. de Juan, M. Maeder, R. Tauler, B. Lendl. Application of a combination of hard and soft modeling for equilibrium systems to the quantitative analysis of pH - modulated mixture samples. *Anal. Chem* 75 (2003) 641–647.

[20] Marius D'Amboise, Benoit Lagarde. Factor analysis using column cross-validation. *Computers and Chemistry* 13 (1), 1989, 39–44.

[21] D. Jouan-Rimbaud Bouveresse, A. Moya-GonzÃąlez, F. Ammari, D.N. Rutledge. Two novel methods for the determination of the number of components in independent components analysis models. *Chemom. Intell. Lab. Syst.* 112 (2012) 24–32.

[22] Lea G. Johnsen, Jose Manuel Amigo, Thomas Skov, Rasmus Bro Automated resolution of overlapping peaks in chromatographic data *Journal of Chemometrics* Volume 28, Issue 2, pages 71–82, February 2014

[23] W.H. Lawton, E.A. Sylvestre. Self Modeling Curve Resolution. *Technometrics* 13 (1971), 617–633.

[24] R. Tauler. Multivariate curve resolution applied to second order data. *Chemom. Intell. Lab. Syst.* 30 (1995), 133–146.

[25] M. Navarro-Reig, J. Jaumot, A. Garcia-Reiriz, R. Tauler. Evaluation of changes induced in rice metabolome by Cd and Cu exposure using LC-MS with XCMS

and MCR-ALS data analysis strategies. *Anal. and Bioanal. Chem.*, 2015, 407 (29), 8835–8847.

[26] Silvia Comani, Dante Mantini, Paris Pennesi, Antonio Lagatta, and Giovanni Cancellieri. Independent component analysis: fetal signal reconstruction from magnetocardiographic recordings. *Computer Methods and Programs in Biomedicine*, 75(2):163–177, August 2004.

[27] F. J. Martinez-Murcia, J. M. Gorriz, J. Ramirez, C. G. Puntonet, and I. A. Illan. Functional activity maps based on significance measures and Independent Component Analysis. *Computer Methods and Programs in Biomedicine*, 111(1):255–268, July 2013.

[28] S. Spasic, Lj. Nikolic, D. Mutavdzic, and J. Saponjic. Independent complexity patterns in single neuron activity induced by static magnetic field. *Computer Methods and Programs in Biomedicine*, 104(2):212–218, November 2011.

[29] Gonzales, R. and Wintz, P. Digital Image Processing. *Addison-Wesley.* (1987)

[30] N. N. d. Moura, J. M. Seixas, W. S. Filho and A. V. Greco. Independent Component Analysis for Optimal Passive Sonar Signal Detection. *Seventh International Conference on Intelligent Systems Design and Applications (ISDA 2007) Rio de Janeiro* (2007), 671–678.

[31] G. Wang, Q. Ding, Y. Sun, L. He, X. Sun. Estimation of source infrared spectra profiles of acetylspiramycin active components from troches using kernel independent component analysis, *Spectrochim. Acta A: Mol. Biomol. Spectrosc*, 70 (3) (2008) 571–576

[32] M. Toiviainen, F. Corona, J. Paaso, P. Teppola. Blind source separation in diffuse reflectance NIR spectroscopy using independent component analysis, *Journal of Chemometrics* 24 (7-8) (2010) 514–522

[33] I. Schelkanova, V. Toronov. Independent component analysis of broad- band near-infrared spectroscopy data acquired on adult human head, *Biomed. Opt. Express* 3 (1) (2012) 64–74

[34] Y.B. Monakhova, S.S. Kolesnikova, S.P. Mushtakova. Independent component analysis algorithms for spectral decomposition in UV/VIS analysis of metalcontaining mixtures including multimineral food supplements and platinum concentrates, *Anal. Methods* 5 (11) (2013) 2761–2772

[35] I. Toumi, S. Caldarelli, B. Torrsani. A review of blind source separation in NMR spectroscopy, *Prog. Nucl. Magn. Reson. Spectrosc.* 81 (2014) 37–64

[36] Shao X1, Wang G, Wang S, Su Q. Extraction of mass spectra and chromatographic profiles from overlapping GC/MS signal with background. *Anal Chem.* 76(17), (2004), 5143–5148.

[37] G.Wang, Q.Ding, Z.Hou Independent component analysis and its applications in signal processing for analytical chemistry, *TrAC – Trends Anal. Chem* 27 (4) (2008) 368–376

[38] Guoqing Wang, Wensheng Cai, and Xueguang Shao. A primary study on resolution of overlapping GC-MS signal using mean-field approach independent component analysis. *Chemometrics and Intelligent Laboratory Systems*, 82(1-2):137–144, May 2006.

[39] Guoqing Wang, Wensheng Cai, and Xueguang Shao. A post-modification approach to independent component analysis for resolution of overlapping GC/MS signals: from independent components to chemical components *Sci. China Ser. B: Chem.* 50 (4) (2007) 530–537

[40] Zhichao Liu, Wensheng Cai, and Xueguang Shao. Sequential extraction of mass spectra and chromatographic profiles from overlapping gas chromatography-mass spectroscopy signals. *Journal of Chromatography A*, 1190(1-2):358–364, May 2008.

[41] X. Shao, Z. Liu, W. Cai. Extraction of chemical information from complex analytical signals by a non-negative independent component analysis, *Analyst* 134 (10) (2009) 2095–2099,

[42] Xueguang Shao, Zhichao Liu, and Wensheng Cai. Resolving multi-component overlapping GC-MS signals by immune algorithms. *TrAC Trends in Analytical Chemistry*, 28 (11) (2009):1312–1321

[43] J.V. Stone. Independent Component Analysis: A Tutorial Introduction, *A Bradford Book*, Cambridge, MA, 2004

[44] H. Parastar, M. Jalali-Heravi, R. Tauler, Is independent component analysis appropriate for multivariate resolution in analytical chemistry? *Trends Anal. Chem.* 31 (2012) 134–143

[45] C.Ruckebusch, L.Blanchet. Multivariate curve resolution: a review of advanced and tailored applications and challenges *Analytical Chimica Acta*, 765 (2013) 28–36.

[46] Giordani P, Kiers H, Del Ferraro M. Three-Way Component Analysis Using the R Package ThreeWay. *Journal of Statistical Software* 57 (2014) 1–23.

[47] Alejandro Olivieri,Graciela Escandar. Practical Three-Way Calibration, 1st Edition. *Elsevier* (2014)

[48] Nicolaas (Klaas) M. Faber, Rasmus Bro, and Philip K. Hopke. Recent developments in CANDECOMP/PARAFAC algorithms: a critical review. *Chemometrics and Intelligent Laboratory Systems*, 65(1):119–137, January 2003.

[49] Maud M. Koek, Renger H. Jellema, Jan van der Greef, Albert C. Tas, and Thomas Hankemeier. Quantitative metabolomics based on gas chromatography mass spectrometry: status and perspectives. *Metabolomics.* 7 (2011) 307–328.

[50] Bosco, M.V., Larrechi, M.S. PARAFAC and MCR-ALS applied to the quantitative monitoring of the photodegradation process of polycyclic aromatic hydrocarbons using three-dimensional excitation emission fluorescent spectra. Comparative results with HPLC. *Talanta*, 71 (2007) 1703–1709.

[51] Kiers, H. and Smilde, A.K. Constrained three-mode factor analysis as a tool for parameter estimation with second-order instrumental data. *Journal of Chemometrics* 12 (1998) 125–147.

[52] Skov, T., van den Berg, F., Tomasi, G., Bro, R. Automated alignment of chromatographic data. *Journal of Chemometrics* 20 (2006) 484–497.

[53] Bloemberg TG, Gerretzen J, Lunshof A, Wehrens R, Buydens LM. Warping methods for spectroscopic and chromatographic signal alignment: a tutorial. *Anal Chim Acta.* 781 (2013) 14–32.

[54] Perera, V., De Torres Zabala M., Florance, H., Smirnoff, N., Grant, M., Yang, Z. R. Aligning extracted LC-MS peak lists via density maximization. *Metabolomics.* 8 (2012) 175–185.

[55] Wei X, Shi X, Merrick M, Willis P, Alonso D, Zhang X. A method of aligning peak lists generated by gas chromatography high-resolution mass spectrometry. *Analyst.* 138 (2013) 5453-60.

[56] Wehrens R, Bloemberg TG, Eilers PH. Fast parametric time warping of peak lists. *Bioinformatics.* 31 (2015) 3063–3065

[57] Wei X, Shi X, Koo I, Kim S, Schmidt RH, Arteel GE, Watson WH, McClain C, Zhang X. MetPP: a computational platform for comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry-based metabolomics. *Bioinformatics.* 29 (2013) 1786-1792.

[58] Vinaixa M, Schymanski EL, Neumann S, Navarro M, Salek RM, Yanes O. Mass spectral databases for LC/MS and GC/MS-based metabolomics: state of the field and future prospects. *Trends in Anal. Chem.* 78 (2016) 23–25.

[59] Strehmel N, Hummel J, Erban A, Strassburg K, Kopka J. Retention index thresholds for compound matching in GC-MS metabolite profiling. *J Chromatogr B* 871 (2008) 182–190.

[60] Stein, SE. An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. *J. Am. Soc. Mass Spectrom.* 10 (1999), 770–781.

[61] Stein, SE, Scott DR. Optimization and testing of mass spectral library search algorithms for compound identification. *J Am Soc Mass Spectrom.* 5 (1994), 859–866.

[62] Koo I, Kim S, Shi B, Lorkiewicz P, Song M, McClain C, Zhang X. EIder: A compound identification tool for gas chromatography mass spectrometry data. *J Chromatogr A* 1448 (2016) 107–114.

[63] H. van den Dool, P.D. Kratz A generalization of the retention index system including linear temperature programmed gas-liquid partition chromatography *J. Chromatogr.* 11 (1963) 463.

[64] E. Kováts. Gas chromatographische Charakterisierung organischer Verbindungen. *Helv. Chim. Acta*, 41 (1958), 1915.

[65] Hummel, J.; Selbig, J.; Walther, D.; Kopka, J. The Golm Metabolome Database: a database for GC-MS based metabolite profiling. *Metabolomics.* **2007**, 18, 75–95.

[66] Hummel, J.; Strehmel, N.; Selbig, J.; Walther, D.; Kopka, J. Decision tree supported substructure prediction of metabolites from GC-MS profiles. *Metabolomics.* **2010**, 6, 322–333.

[67] Wishart DS, Tzur D, Knox C, et al. HMDB: the Human Metabolome Database. *Nucleic Acids Res.* 35 (2007).

[68] Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K.; Oda, Y.; Kakazu, Y.; Kusano, M.; Tohge, T.; Matsuda, F.; Sawada, Y.; Hirai, MY.; Nakanishi, H.; Ikeda, K.; Akimoto, N.; Maoka, T.; Takahashi, H.; Ara, T.; Sakurai, N.; Suzuki, H.; Shibata, D.; Neumann, S.; Iida, T.; Tanaka, K.; Funatsu, K.; Matsuura, F.; Soga, T.; Taguchi, R.; Saito, K.; Nishioka, T. MassBank: A public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* **2010**, 45, 703–714.

[69] Katajamaa, M.; Jarkko, M.; Matej, O. MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics.* **2006**, 22, 634–636.

[70] Pluskal, T.; Castillo, S.; Villar-Briones, A.; Oresic, M. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics.* **2010**, 11, 395.

[71] Lommen, A. MetAlign: interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing. *Anal. Chem.* **2009**, 81, 3079–3086.

[72] Lommen, A.; Harrie J. K. MetAlign 3.0: performance enhancement by efficient use of advances in computer hardware. *Metabolomics.* **2012**, 8, 719–726.

[73] Tautenhahn, R.; Patti, G. J.; Rinehart, D.; Siuzdak, G. XCMS Online: a web-based platform to process untargeted metabolomic data. *Anal. Chem.* **2012**, 84, 5035–5039.

[74] Aggio, R.; Villas-Boas, S. G.; Ruggiero, K. Metab: an R package for high-throughput analysis of metabolomics data generated by GC-MS. *Bioinformatics.* **2011**, 27, 2316–2318.

[75] Fernandez-Varela, R.; Tomasi, G.; Christensen J. H. An untargeted gas chromatography mass spectrometry metabolomics platform for marine polychaetes. *J. Chromatogr. A.* **2015**, 1384, 133–141.

[76] Wehrens, R.; Georg, W.; Fulvio, M. metaMS: An open-source pipeline for GC–MS-based untargeted metabolomics. *J. Chromatogr. B.* **2014**, 966, 109–116.

[77] Luedemann, A.; Strassburg, K.; Erban, A.; Kopka, J. TagFinder for the quantitative analysis of gas chromatography–mass spectrometry (GC-MS)-based metabolite profiling experiments *Bioinformatics.* **2008**, 24, 732–737.

[78] Hiller, K.; Hangebrauk, J.; Jager, C.; Spura, J.; Schreiber, K.; Schomburg, D. MetaboliteDetector: comprehensive analysis tool for targeted and nontargeted GC-MS based metabolome analysis. *Anal. Chem.* **2009**, 81, 3429–3439.

[79] O'Callaghan, S.; De Souza, D. P.; Isaac, A.; Wang, Q.; Hodkinson, L.; Olshansky, M.; Erwin, T.; Appelbe, B.; Tull, D. L.; Roessner, U.; Bacic, A.; McConville, M. J.; Likic, V. A. PyMS: a Python toolkit for processing of gas chromatography-mass spectrometry (GC-MS) data. Application and comparative study of selected tools. *BMC bioinformatics.* **2012**, 13, 115.

[80] Jellema, R. H.; Krishnan, S.; Hendriks, M. M. W. B.; Muilwijk, B.; Vogels J. T. W. E. Deconvolution using signal segmentation. *Chemometr. Intell. Lab.* **2010**, 104, 132–139.

[81] Ni, Y.; Qiu, Y.; Jiang, W.; Suttlemyre, K.; Su, M.; Zhang, W.; Jia, W.; Du, X. ADAP-GC 2.0: Deconvolution of coeluting metabolites from GC/TOF-MS data for metabolomics studies. *Anal. Chem.* **2012**, 84, 6619–6629.

[82] Fiehn,O., Wohlgemuth,G., Scholz,M. Automatic annotation of metabolomic mass spectra by integrating experimental metadata. *Proc. Lect. Notes Bioinformatics*. **2005**, 3615, 224–239.

[83] Skogerson, K.; Wohlgemuth, G.; Barupal, D. K.; Fiehn, O. The volatile compound BinBase mass spectral database. *BMC Bioinformatics*. **2011**, 12, 321.

[84] Kind, T.; Wohlgemuth, G.; Lee do, Y.; Lu, Y.; Palazoglu, M.; Shahbaz, S.; Fiehn, O. FiehnLib: mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry. *Anal. Chem.* **2009**, 81, 10038–10048.

[85] Cuadros-Inostroza, A, Caldana, C, Redestig, H, Kusano, M, Lisec, J, Pena-Cortes, H, Willmitzer, L, A; Hannah, M. TargetSearch - a Bioconductor package for the efficient preprocessing of GC-MS metabolite profiling data *BMC Bioinformatics*. 2009, 10, 428

Chapter 3

Goals

3.1 Main objective

One of the major bottlenecks in metabolomics is to convert raw data samples into biological interpretable information. Also, mass spectrometry-based metabolomics generates large and complex datasets characterized by co-eluting compounds and with experimental artifacts. This thesis main objectives are to develop new methods based on blind source separation to improve the capabilities of the current strategies that tackle the different metabolomics workflow steps limitations. Also, the objective of this thesis is to develop tools capable of performing the entire metabolomics workflow for GC–MS, including pre-processing, spectral deconvolution, alignment and identification. These tools should be able to convert raw data into biological interpretable information and thus, allow resolving biological answers and discovering new biological insights.

3.2 Goals of the project

- **O1:** Compare and improve the performance of MCR–ALS and ICA applied in GC–MS and GCxGC–MS data. Design an automated strategy for the appli-

cation of multivariate algorithms based on independent component analysis in GC–MS data.

- **O2:** Combine the strengths of univariate (peak-picking) and multivariate (MCR–ALS/ICA) methods for chromatographic data processing. Develop a factor analysis-free method for the multivariate spectral deconvolution of chromatographic signals.

- **O3:** Implement an easy-to-use R library to perform a flexible, robust and automated metabolomics data processing workflow of GC–MS.

- **O4:** Implement an easy-to-use R library to take advantage of the knowledge provided by metabolomics spectral libraries to process GC–MS samples in a driven manner, and with the possibility of standardize the retention times without the use of internal standards.

- **O5:** Implement all the R libraries with a standardized S4 method, and with a modular structure, so each familiarized programmer can attach their own modules (deconvolution, alignment, identification) to the main package.

3.3 Expected contributions

Expected contributions include the development of new strategies based on blind source separation for GC–MS data processing contextualized in metabolomics. Those strategies are expected to be applied as a part of a R package capable of performing the entire metabolomics workflow. The developed methods should provide a new and original strategy to resolve the typical problems of the univariate (peak-picking) and multivariate approaches for GC–MS data resolution, in order to advance in the automated interpretation of mass spectrometry data.

Chapter 4

Compound identification in gas chromatography/mass spectrometry–based metabolomics by blind source separation

Published as: Domingo-Almenara X, Perera A, Ramírez N, Cañellas N, Correig X, Brezmes J. Journal of Chromatography A. Vol. 1409 (2015) 226-233. DOI: 10.1016/j.chroma.2015.07.044.

Abstract

Metabolomics GC–MS samples involve high complexity data that must be effectively resolved to produce chemically meaningful results. Multivariate curve resolution–alternating least squares (MCR–ALS) is the most frequently reported technique for that purpose. More recently, independent component analysis (ICA) has been reported as an alternative to MCR. Those algorithms attempt to infer a model describing the observed data and, therefore, the least squares regression used in MCR assumes that the data is a linear combination of that model. However, due to the high complexity of real data, the construction of a model to describe optimally the observed data is a critical step and these algorithms should prevent the influence from outlier data. This study proves independent component regression (ICR) as an alternative for GC–MS compound identification. Both ICR and MCR though require least squares regression to correctly resolve the mixtures. In this paper, a novel orthogonal signal deconvolution (OSD) approach is introduced, which uses principal component analysis to determine the compound spectra. The study includes a compound identification comparison between the results by ICA–OSD, MCR–OSD, ICR and MCR–ALS using pure standards and human serum samples. Results shows that ICR may be used as an alternative to multivariate curve methods, as ICR efficiency is comparable to MCR–ALS. Also, the study demonstrates that the proposed OSD approach achieves greater spectral resolution accuracy than the traditional least squares approach when compounds elute under undue interference of biological matrices.

4.1 Introduction

The analysis of samples from a metabolomics perspective allows the phenotyping of organisms at a molecular level [1]. At the same time, metabolomics provides a means of detecting early biochemical changes in organisms before the appearance of a disease and thus, a means of finding predictive biomarkers [2]. Among the analytical techniques used in metabolomics, gas chromatography-mass spectrometry (GC–MS) is a well stablished platform due to its robustness and its applicability to a wide range of matrices and metabolites through silylation of the polar groups.

Because of the high complexity of biological fluids, the complete chromatographic resolution of all the metabolites in a sample cannot be easily achieved as the co-elution of two or more of them usually occurs. The correct identification of co-eluted compounds depends mostly on the degree of the chromatographic separation and their spectral dissimilarity. Likewise, the metabolites in the samples usually occur at low concentrations and the background signal, inherent in the instrument and the sample biological matrix, interferes in their correct identification and quantification. The use of resolution algorithms, which can help extract the purest compound elution profile and spectra, is mandatory for GC–MS data processing.

One of the best–established algorithms for application to chromatographic data to resolve co-eluted compounds is multivariate curve resolution–alternating least squares (MCR–ALS) [3, 4]. MCR–ALS can resolve a mixture of compounds into a pure concentration profile matrix and a pure spectra matrix [5]. In recent years, a blind source separation (BSS) technique known as independent component analysis (ICA) [6], already widely applied for the resolution of spectroscopic mixtures [7, 8, 9, 10, 11], has also been applied for the resolution of GC–MS samples [12]. In a GC–MS chromatogram, the compounds elution profiles appear mixed with their respective spectra. In these cases, ICA-based approaches are able to recover the different independent sources contained in data and, eventually, resolve GC–MS data. MCR–ALS ap-

proaches this problem by minimizing the residual error between the data and the predicted model, whereas ICA focuses on estimating the original sources - or components - by maximizing their statistical independence. Actual ICA-based methods to resolve chromatographic data include mean-field ICA (MF–ICA) [13], post-modification based on chemical knowledge (PBCK) [14], window ICA (WICA) [15] and non-negative ICA [16]. Artificial immune system algorithms involving the use of ICA have also been proposed [17]. The first step of the resolution procedure in these methods is the use of ICA to resolve the mass spectrum for each compound in the mixture. The above-mentioned algorithms use different approaches to determine the elution profile of each compound, since the elution profiles determined by ICA tend to be inaccurate or affected by various ICA ambiguities such as negativity or variance (energy) indetermination [18]. Recently, these ICA-based methods were compared with MCR for the resolution of GC–MS data by Parastar and coworkers [19] who showed that the ICA-based resolutions methods show the same performance than MCR. A natural extension of ICA to recover co-eluted profiles might be independent component regression (ICR), which was first used to resolve mixtures in near infrared (NIR) spectra by Shao et al. [20], but whose efficiency on GC–MS data treatment has not yet been studied.

The use of least squares (LS) regression, common to most algorithms in GC–MS data resolution, has a major drawback, induced by the inherent correlation between ions related to the same compound. This correlation yields an ion-redundancy which means that, for each compound, different ions, also called fragments or m/z, elute at the same retention time and with the same elution profile. When fitting the elution profiles to data, no correlation information between the ions is taken into account, so the LS regression does not distinguish between noise and the compound ions that are being regressed; this may introduce a bias into the LS regressors. This effect includes instrumental or experimental noise as baseline, peak-tailing, or compound

co-elution. The performance of the resolution of mixtures with least squares may, therefore, depend on the correct estimation of the underlying model from the data.

This study proposes the use of ICR for GC–MS compound identification. In this approach, we integrate ICA and MCR with a novel orthogonal spectra deconvolution (OSD) as an alternative to least squares regression with a view to improve the determination of the compound spectra when compounds elute under the interference of a biological matrix.

4.2 Materials and methods

This section describes MCR–ALS, ICR and their variants integrated with the OSD algorithm (ICA–OSD and MCR–OSD). The proposed methods were evaluated by comparing the resolution of the spectra of 38 compounds in a pure standards sample and 25 compounds in a human serum sample. A match score between the resolved and the reference spectra was determined for each compound and method. The samples were processed by MCR, the proposed ICR, both ICA and MCR using the OSD approach (ICA-OSD and MCR-OSD). The goal was to use the different methods compared in this study to extract the most pure spectra for each compound. The spectra extracted were matched against a reference MS spectra database. For this study, the Golm Metabolome Database (GMD) [21] was used as a reference database.

4.2.1 Materials

A set of four pure standards samples - four sample repetitions - and a total of eight biological samples - four sample repetitions of a human serum sample, and two repetitions of two human urine samples from healthy volunteers - were used for evaluation. The standard mixture was composed of 26 metabolites (see Table A.1 of the Supplementary Material) previously found in the human serum and urine metabolome

[22]. First, all samples were characterized by a curated identification of the reference compounds (standards). The pure standards samples were taken as a reference to later identify the same compounds in the human serum and urine samples. Two compounds identified in the biological samples that are not included in the pure standards set were validated also analyzing their corresponding standard references.

The metabolites of the human serum and urine samples were extracted and derivatized following a standard protocol [23] with slight modifications to optimize the process. Extracts were analyzed using a 7890 gas chromatograph from Agilent (Palo Alto, CA. USA) coupled to a Pegasus IV TOF/MS from Leco (St. Joseph, MI, USA) using a DB5–MS capillary column (30 m $\times$ 0.25 mm $\times$ 0.25 μm, 5% diphenyl, 95% dimethylpolysiloxane) from Agilent. Analyses were performed by injecting 1 μL of the extracts into a split/splitless inlet at 250°C with a split flow of 5 mL min^{-1} and a helium constant flow of 1 mL min^{-1} (99.999%, Abelló Linde, Barcelona). The oven temperature of the GC was initially held at 50°C for 1 min, then raised to 285°C at a rate of 20°C min^{-1} and held at that temperature for 5 min. The GC–TOF/MS interface was set at 280°C and the ion source at 250°C. The mass spectrometer acquired m/z ratios between 35 and 600 amu at 10 Hz and an electron impact energy of 70 eV.

4.2.2 Data pre-processing and analysis

In order to analyze an entire dataset using the MCR or ICA-based approaches, each chromatogram was divided in chromatographic peak features (CPFs) using the same criteria as in [24]. The different CPFs contained several compounds, so the algorithm had to deconvolve them in case of co-elution. The number of factors or components used to initialize both MCR and ICA was determined by cross-validation (described in Section 2.6). A unimodality constraint [25] was applied to the resolved profiles and the same non-negative least squares algorithm was applied for both MCR and ICR. The simple mean spectra determined either by ICA–OSD, MCR–OSD, ICR or MCR

in the different samples for each compound were compared using the dot product [26] against the GMD MS spectra database.

The masses 73, 74, 75, 147, 148, and 149 m/z were excluded before processing the sample, since they are ubiquitous mass fragments typically generated from compounds carrying a trimethylsilyl moiety [21]. They were also excluded in the identification. Only the fragments from m/z 70 to 600 were taken into account when comparing reference and empirical spectra, since this is the m/z range included in the downloadable GOLM database. Also, the human serum and urine samples signal was filtered using a Savitzky–Golay filter [27] and the baseline was removed using a semi-supervised spline interpolation to reduce the interaction of the biological matrix (described in Section 3.2). The ICA algorithm used was the joint approximate diagonalization of eigenvalues (JADE) [30].

4.2.3 Resolution of GC/MS mixtures by multivariate curve resolution–alternating least squares (MCR–ALS)

The purpose of multivariate curve resolution – alternating least squares (MCR–ALS) is to decompose a data matrix containing a mixture of compounds into two matrices containing the resolved pure concentration profiles and pure spectra. MCR can mathematically be expressed as:

$$D = CS^T + E \tag{4.1}$$

where D ($N \times M$) is the raw data matrix containing the mixture of compounds, C ($N \times k$) is the resolved concentration profile matrix, S ($M \times k$) is the resolved spectra matrix and E ($N \times M$) is the error matrix. In this notation, N is the number of chromatographic scans (retention time), M is the range of acquisition of the mass-charge ratio (m/z), and k is the number of components or compounds in the model.

MCR–ALS uses an iterative least squares algorithm (ALS) to determine both C and S matrices by minimizing the error matrix E. A detailed explanation of MCR–ALS, together with pseudocode, is given elsewhere [29]. To optimize execution speed, we used our own implementation of the MCR–ALS algorithm. This was based on the R package *NNLS*, which uses the Lawson–Hanson non-negative least squares (NNLS) implementation. The package uses C routines to increase the computational speed.

4.2.4 Resolution of GC/MS mixtures by independent component regression (ICR)

The proposed independent component regression (ICR) method consists of applying an independent component analysis (ICA), followed by a least squares regression (LS) using the ICA output as a regressor. In this manner, ICA is used to determine the elution profile of the different compounds in the mixture. Then, a least squares regression is used to determine the spectra of each compound by fitting the extracted elution profiles to the data. This implementation is the opposite of the extraction of the compound spectra to later determine the elution profile, used in the above mention ICA-based implementations. Our ICA model can be expressed as:

$$D^T = AZ^T \tag{4.2}$$

Analogously to (Eq. 4.1), D ($N \times M$) is the original chromatographic raw data matrix, A ($M \times K$) is the mixing-matrix and Z ($N \times K$) is the independent components matrix. The Z matrix holds the elution response of the underlying components, but it presents two main ambiguities: (i) we cannot determine the energy or intensity of the resolved components and therefore they are not ordered by explained variance, and (ii) recovered sources do not fulfill non-negativity. Due to the first ambiguity, the recovered sources in Z are arbitrarily scaled and consequently they cannot be

used for quantifying the concentration of compounds. Due to the second ambiguity, the extracted components can be negative or contain negative values - known to be caused by source signal overlapping, as explained in [16]-. According to [7, 12], the estimated sources in Z may appear negatively correlated with the data, i.e., the estimated elution profile may be a negative mirror image of the real one. Thus, the Z matrix contains only the qualitative shape—the elution profile model in the retention time dimension—of the underlying compounds. A natural strategy for avoiding such negativity ambiguity is the use of non-negative ICA (nnICA). This, however, adds a significant computational cost and does not solve the first ambiguity, which still has to be resolved by a least squares regression. Therefore, the following strategy is proposed to overcome both ICA ambiguities: all the profiles in Z that express more negative variance than positive variance are negatively rotated. After this step, a non-negative least squares regression (NNLS) is applied to resolve the variance ambiguity and to retrieve the spectrum for each compound. This is to determine a non-negative spectra matrix S that minimizes the error matrix E:

$$D = \hat{Z}S^T + E \tag{4.3}$$

where D ($N \times M$) is the raw data matrix, Z ($N \times k$) is the elution matrix and S ($M \times k$) the spectra matrix. The hat in $\hat{Z}$ denotes a normalized matrix, since real energies are not known a priori. The determined profiles are fitted in the different columns of the data matrix containing the different m/z values. For ICR, Z is the matrix analogous to the C matrix in MCR. The *JADE* R package is used for the implementation of the ICA-based algorithms.

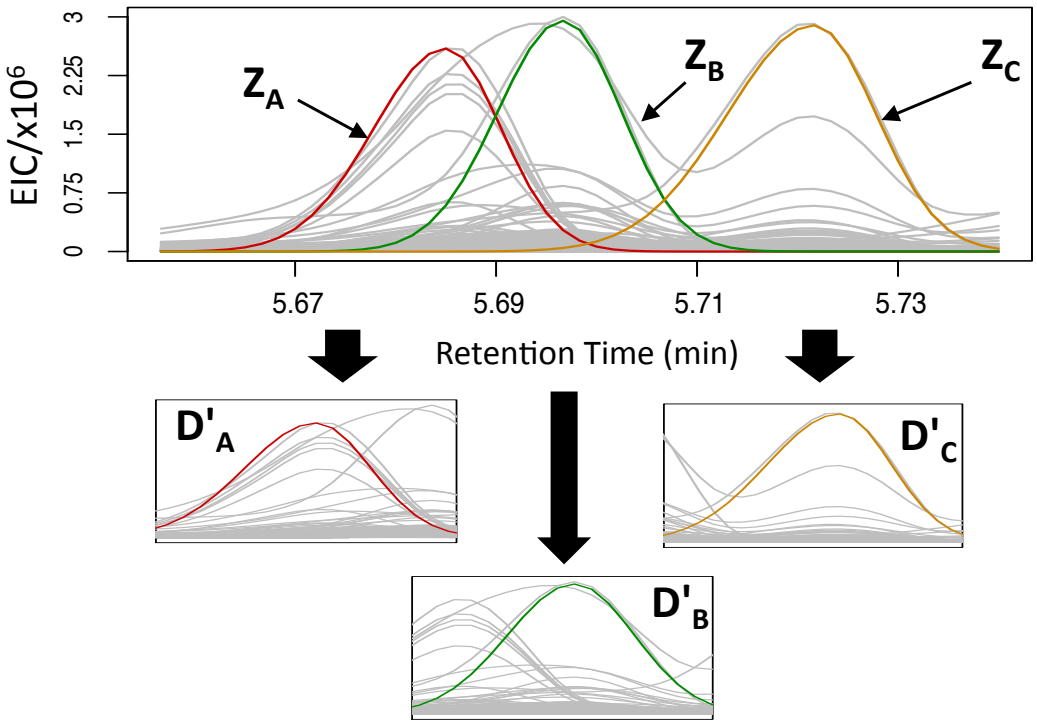

Figure 4-1: Determination of D'_j for a given data matrix D, where three compounds appear co-eluted. The extracted ion chromatogram (EIC) of the original D matrix is shown (top). The grey lines represent the different m/z masses whereas the coloured lines represent the three resolved compounds for the case given. Each sub-data matrix D'_j is determined comprising the data for which each compound profile D_j is eluting. A cut-off of 5% is applied to all the profiles, so the D'_j sub-data matrix comprises the data in D for which the profile Z_j is non-zero.

4.2.5 Spectra extraction by orthogonal signal deconvolution (OSD)

Orthogonal signal deconvolution (OSD) is a method to extract and deconvolve the spectra given only the compounds elution profile. In multivariate curve resolution or independent component regression, the spectra is determined by means of non-negative least squares, instead, in OSD principal component analysis (PCA) is used to determine the spectra of each compound as opposite to the use of least squares. For this study, a pre-process to determine the elution profiles is conducted by independent component analysis (ICA) or multivariate curve resolution (MCR), and are referred as ICA-OSD and MCR-OSD, respectively. In OSD, PCA is used to decorrelate the

sub-data matrix and to determine which ions co-vary along the retention time, thus detecting the different ion-redundancies or ion-correlations related to each compound. However, PCA cannot be used directly to resolve an entire chromatographic mixture, since it is constrained to fulfill maximum variance and orthogonality [10]. PCA can be used to deconvolve spectra though if we force PCA to fulfill maximum variance and orthogonality just in the eluting space of the compound whose spectrum is to be extracted. Then, for each extracted compound profile j in Z (2), a D_j sub-data matrix is determined comprised only of the data of the retention time in which the compound Z_j is eluting (Figure 4-1). After that, a PCA is applied for each given window, i.e., each compound profile. Following the same notation, PCA can be mathematically described as:

$$D'_j = YW^T \tag{4.4}$$

where D'$_j$ ($N \times M$) is the sub-data matrix to decompose, Y ($N \times M$) is the score matrix and W ($M \times M$) the loading or eigenvectors matrix. Matrix Y holds the retention time response of the different decomposed components and matrix W holds the spectra associated with each component, which includes the spectrum of the compound of interest and other unknown noise interferences. In both decomposed matrices, each component may have negative or positive variance. The component of interest associated with the compound whose spectrum is to be extracted is determined by comparing the different covariance responses in matrix Y with the reference profile in Z. This is to determine which component has the highest absolute correlation with the elution profile of the compound of interest. The spectra associated with the selected components are rotated according to the sign of the correlation coefficient with the compounds profile models. OSD algorithm can be summarized in the following steps:

1. Given a Z_j compound elution profile, determine a D_j sub-data matrix comprised only of the data of the retention time in which the compound is eluting.

2. Apply a PCA over D_j. The result is a score matrix Y and loading matrix W.

3. Determine the correlation coefficient between Z_j and each component in Y and select the component h with the highest absolute correlation value.

4. Select the component h in W, rotate W_h according to the sign of the previous determined correlation coefficient, and clip to zero all the negative values. W_h is now considered to be the spectrum of Z_j.

OSD uses a PCA-based approach in order to avoid the use of an LS regressor, which finds difficulties in discriminating noise and the compound ions that are being regressed, which itself may introduce a bias to the LS regressors. This effect results in the extraction of the spectra with fragments that may not belong to the true compound spectrum or its intensity is over or underestimated. In OSD, principal component analysis is proposed to improve this limitation and to take advantage of the multivariate nature of GC/MS data. The difference in the application of PCA instead of NNLS resides in the fact that the PCA model takes into account the inherent noise always present in real data and which may have not been included in the ICA or MCR model.

4.2.6 Determination of number of components

Both MCR and ICA/ICR require a fixed number of components, also known as factors, to define their respective models. This parameter clearly affects the ICA or MCR outcome, as a correct estimation of components in the mixture leads to the construction of a model which better fits in data. In this study, a cross-validation approach was used to assure an appropriate determination of the number of components, which was implemented by the following steps: (i) Similarly to [31], divide

the D matrix into D_{even} and D_{odd}. Each matrix contains every second row (scans) in D, and thus, all the columns (m/z channels) are preserved. (ii) Compute PCA over D_{even} and determine a L1 matrix containing the PCA loadings. (iii) For each column j in L_1 matrix, determine a matrix $T = [l_1, l_2, ..., l_j]$ containing all the L_1 columns from 1 to j, (iv) determine a rotation matrix T_p and compute the S_2 scores over D_{odd} (5), (v) project the variance explained by S_2 scores into D_{odd} by constructing the M_2 matrix (6). For each iteration j determine the residual sum of squares (RSS) error between D_{odd} and M_2.

$$T_p = (T^T T)^{-1} T \qquad \Rightarrow \qquad S_2 = D_{odd} T_p^T \tag{4.5}$$

$$M_2 = S_2 T^T \tag{4.6}$$

This method yields a decreasing RSS curve. The proper number of factors is determined when the addition of more components does not significantly decrease the explained variance, i.e., when the RSS error reaches a minimum.

4.3 Results and discussion

4.3.1 Pure standards dataset processing

The synthetic sample was processed by all the alternative strategies described. All the approaches led to the correct identification of all 26 metabolites in the original mixture design. Some of the compounds appeared in different trimethylsilyl (TMS) derivatives and therefore a total of 38 compounds was identified. Table A.2 (see Supplementary Material) shows the complete list of metabolites identified, along with their match score by the different methods for a quantitative comparison reference. The identification match score is determined by the following steps: first, the nor-

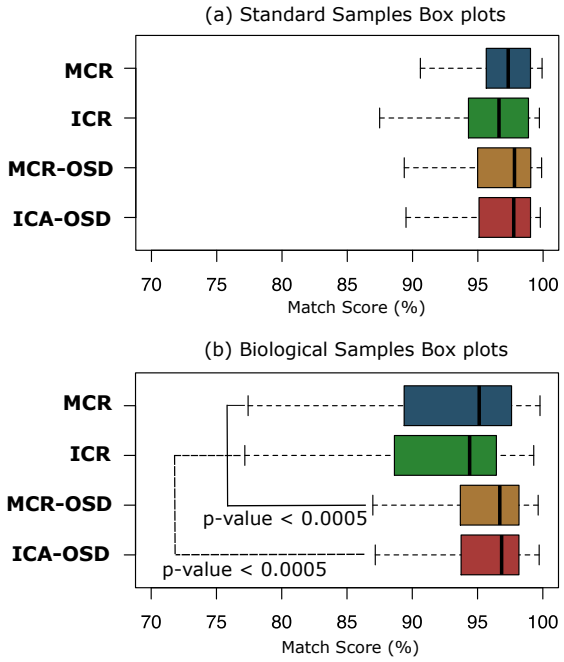

Figure 4-2: Match score box plots. (a) The match score boxplot for the case of pure standards dataset and (b) for the case of biological samples dataset. Outliers in the boxplot are not shown. The ρ-values were determined with a paired wilcoxon test, with an alternative hypothesis that the OSD method performs better than LS. The sample size N was of N=152 for (a) and of N=80 for (b).

malized spectra for each compound in the four samples is averaged by a simple mean - the total sum of the spectra in each sample -. Then, the match score is determined by the dot product between the average resolved - extracted or empirical - and the reference spectra. The closer the score to one hundred, the more exact and pure the spectra extracted.

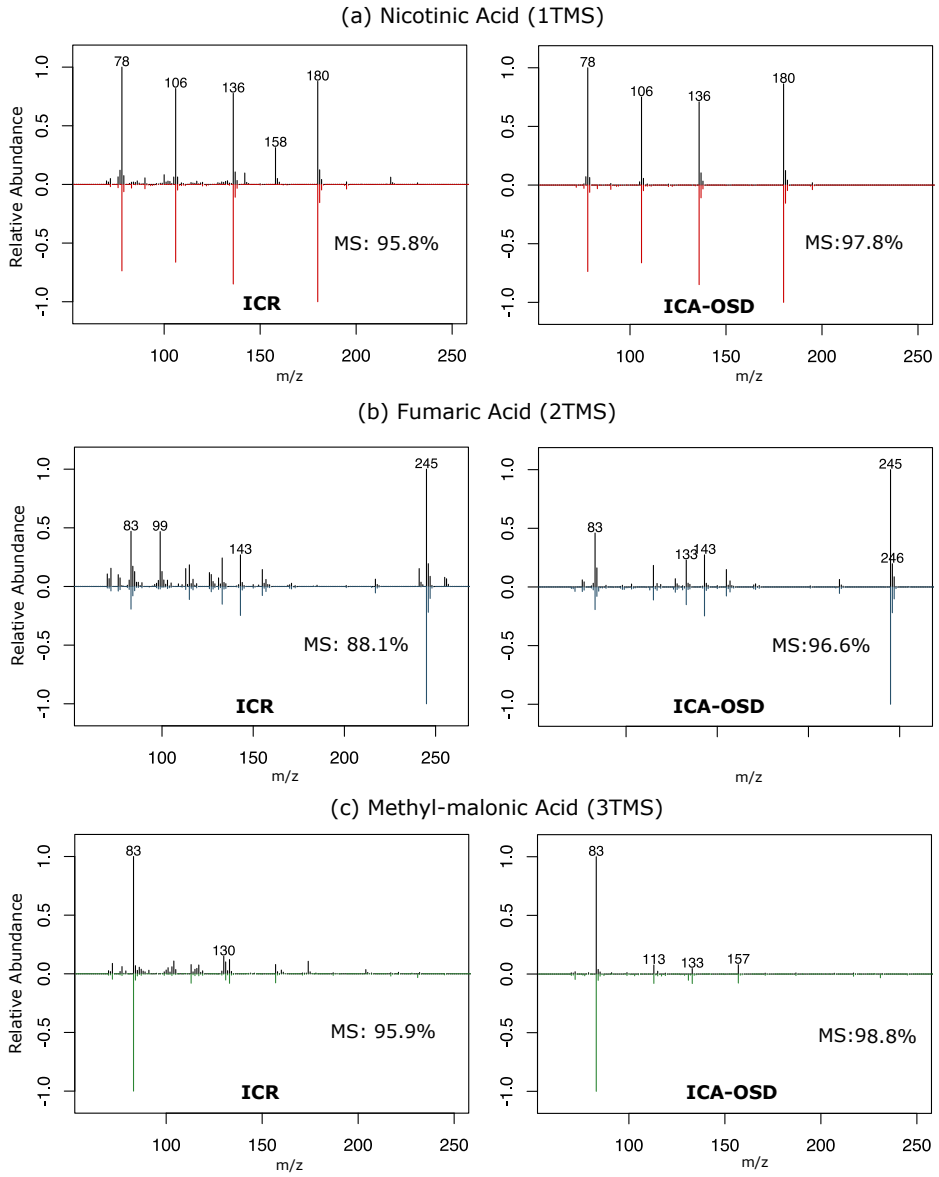

Figure 4-3: Comparison of the standards dataset extracted spectra (black and positively displayed) and the reference GMD spectra (color and negatively displayed). Qualitative spectra differences can be seen between least squares (ICR) and OSD (ICA–OSD) approaches. The extracted spectra by ICR and ICA–OSD are shown in black for (a) nicotinic acid, (b) fumaric acid and (c) methyl-malonic acid. The reference spectra (color) are shown in the same axis, negatively rotated, for better visual appreciation. The match score (MS) is noted in each plot.

Overall identification performance for the studied methods is shown in the box plots of Figure 4-2 (a). To increase the statistical power, these box plots were constructed by the match score for each metabolite and sample separately — each spec-

trum of the same compound in each sample was matched independently against the reference spectra —. It is clear that in the capability of the proposed ICR and OSD methods is comparable to the best extended MCR–ALS. Still, some qualitative differences between ICR and ICA-OSD extracted spectra can be appreciated when visually comparing the empirical (resolved) and reference spectra of certain compounds, specially in co-elution situations. Compounds showing important qualitative spectral differences between methods include nicotinic acid, fumaric acid and methyl-malonic acid (Figure 4-3), for which the OSD approach performs a better isolation of the compound-related ions from the other ions or fragments product of the co-elution with neighboring compounds. In Figure 4-3 (a), the OSD approach is able to discard the ion m/z number 158 for nicotinic acid as it is an interference due to the co-elution with isoleucine. The same observation can be seen in Figure 4-3 (b) where the ion number 99 for fumaric acid is detected as an outlier by the OSD approach and discarded from its resolved spectrum; this interference occurs as fumaric acid in co-elution with uracil (See Table A.2 of Supplementary Material). Also, Figure 4-3 (c) shows the case of methyl-malonic acid, for which OSD extracted purer spectra, specially at low ion intensity levels.

4.3.2 Biological samples processing

In this case, the methods under study were tested in biological samples, where compounds appear in very low concentrations and with the interference of a biological matrix. Processing of the human serum and urine samples by the different methods led to the extraction of a total of an average of 230 compounds or components per sample by the OSD approaches, ICR and MCR. From all of them, 15 metabolites from the original pure standards experiment, and two that were not included in the standards dataset, were identified in different TMS derivatives, so a total of 25 compounds were found (Table 4.1) - 21 in human serum and 4 of them both in serum and

urine -.

Table 4.1: Identification score results for the human serum and urine samples.

Name	ICA OSD	MCR OSD	ICR	MCR
Serum				
Leucine (1TMS)	99.61	99.26	86.55	89.26
Proline (1TMS)	99.34	99.49	95.42	95.60
Urea (2TMS)	98.45	96.78	95.77	95.24
Isoleucine (2TMS)	98.83	97.20	94.63	96.07
Proline (2TMS)	98.01	98.13	95.08	95.62
Glycine (3TMS)	99.25	99.26	98.56	98.60
Serine (3TMS)	98.18	98.24	96.77	96.81
Allo-threonine (3TMS)	97.35	94.67	87.52	96.21
Methionine (2TMS)	92.24	96.22	85.22	84.85
Aspartic acid (3TMS)	96.51	94.61	87.03	88.91
Phenylalanine (1TMS)	98.65	98.42	99.06	98.99
Cysteine (3TMS)	93.84	93.68	72.12	72.88
2-oxo-glutaric acid (2TMS)	87.48	87.43	73.48	74.40
Proline [+CO2] (2TMS)	98.78	98.74	98.28	98.53
Phenylalanine (2TMS)	97.35	97.01	95.11	95.09
Ornithine (3TMS)	98.22	98.19	97.47	97.91
Ornithine (4TMS)	98.21	98.19	98.92	98.99
Citric acid (4TMS)	96.81	96.78	95.30	95.27
Tyrosine (2TMS)	96.73	96.76	95.54	95.04
Myo-inositol (6TMS)	97.92	97.98	96.19	98.03
Cholesterol (1TMS)	92.58	92.23	92.84	92.23
Urine				
Urea (2TMS)	94.26	97.20	91.19	91.17
2-oxo-glutaric acid (2TMS)	80.26	77.99	73.12	73.26
Citric acid (4TMS)	94.41	90.67	88.26	88.83
Myo-inositol (6TMS)	90.74	91.10	91.97	95.35

Raw data pre-processing included signal filtering using a Savitzky–Golay filter of third order with a 1.1 seconds window length, i.e., half the average peak width. Baseline was removed using a three-step spline interpolation. For each m/z channel, first, (i) a running minimum filter was used with window length 10 times the average peak width (k_{filter}) and from the resulting signal Υ_{min} the baseline standard deviation was determined (σ_b). After that, (ii) a same window length running medians filter was applied, and the resulting signal was Υ_{base}. The running medians filtered signal outcomes a good approximation of the underlying baseline, but to refine it and to

avoid outliers each point in Υ_{base} was constrained not to have intensity above Υ_{min} plus σ_b. Finally, (iii) a spline interpolation was applied - with $k_{filter}/2$ degrees of freedom - to smooth Υ_{base}. The smoothed Υ_{base} was subtracted from the original raw data.

An overall identification capability for the studied methods is shown in the box plots of Figure 4-2 (b). In the biological dataset, the OSD implementations display a more accurate identification of the metabolites in terms of match score and major qualitative differences between the regression (ICR/MCR–ALS) and OSD approaches can be observed. Compounds showing an important match score enhancement between least squares and OSD methods include proline (1TMS) and (2TMS), serine, methionine, aspartic acid, 2-oxo-glutaric acid, cysteine, phenylalanine or urea.

Compounds showing important qualitative and quantitative differences between the least squares and OSD approaches include isoleucine, urea, aspartic acid and cysteine (Figure 4-4). Figure 4-4 shows that isoleucine low intense interfering ions are removed in the OSD approach. In the case of urea, the spectra is structurally the same between methods but in the OSD approach, the intensities of their ions are closer to the pure spectrum values, and this enhances the match score for the OSD case. Figure 4-4 also shows that m/z signals 128 and 176 for aspartic acid and 91 and 120 for cysteine are clearly interfering with the underlying pure spectrum of the compound, as the least squares approach is not able to diminish the signal disturbance. On the contrary, the OSD approach is able to deconvolve or discard those signals, and to correct their intensity so that they are closer to the pure spectrum value. This also reveals that OSD is not only an m/z classifier but also has a distinct multivariate deconvolution property. OSD is not only able to discard those m/z signals unrelated to the compound of interest, but is also able to correct, and therefore deconvolve, the intensity of the m/z response. This deconvolution property, a product of the benefits of the application of multivariate over univariate methods, is specially observable in

the case of urea or cysteine (Figure 4-4) but also occurs in the remainder of the cases.

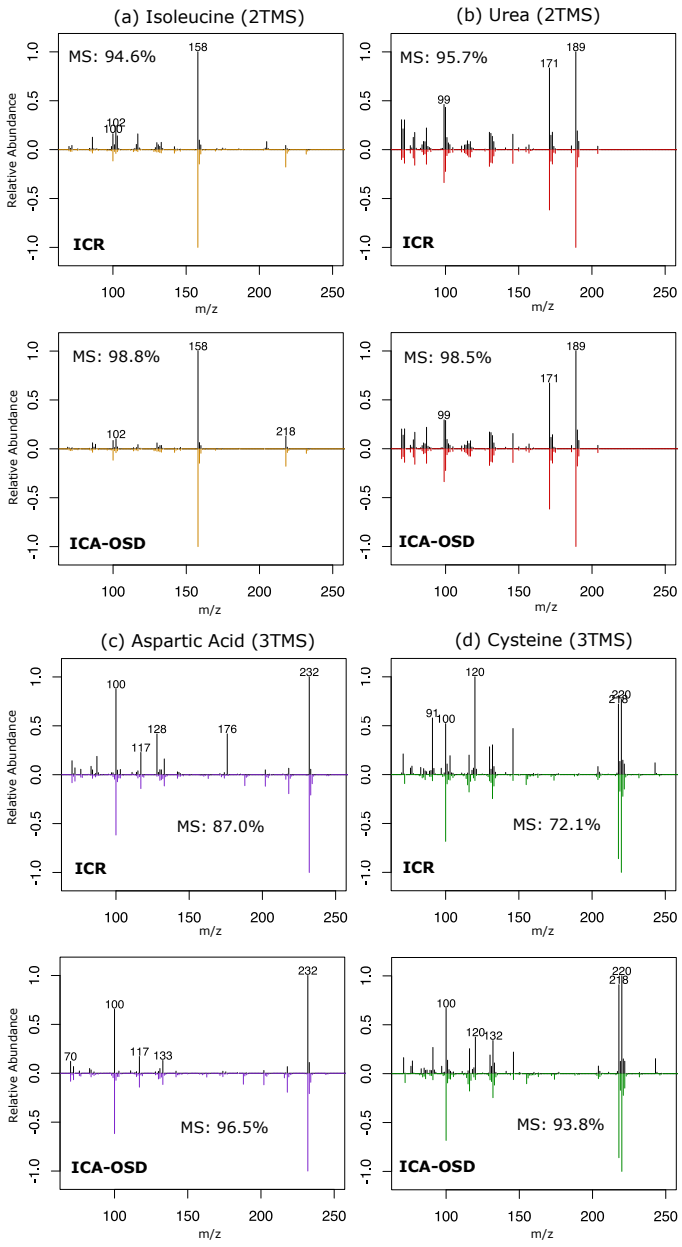

Figure 4-4: Comparison of the extracted spectra (black) and the reference GMD spectra (color) in the biological samples. Significant qualitative and quantitative differences can be appreciated between least squares (ICR) and OSD (ICA–OSD) approaches. The extracted spectra by ICR (top row) and ICA–OSD (bottom row) are shown in black for (a) isoleucine, (b) urea, (c) aspartic acid and (d) cysteine. The reference spectra are shown in the same axis for a better visual appreciation. The match score (MS) is noted in each plot.

To compare the multivariate deconvolution capacity of the different approaches, the euclidean error distance was computed for all the normalized spectra and methods

(Figure 4-5). For each compound, the euclidean distance was computed between the m/z of each reference spectra and the m/z of the different empirical spectra by each method, as described in the Supplementary Information. The figure shows that both ICA–OSD and MCR–OSD methods appear closer to the original spectra, since their distances are generally smaller. These results confirm that OSD acts as a multivariate method for spectra deconvolution.

In some cases, the use of OSD led to a decrease of the match score in comparison with LS, this occurs in the case of phenylalanine (1TMS) or myo-inositol (urine). This can be explained as the LS approaches are more conservative, since they do not make any presumption whether a certain fragment belongs or not to the compound spectrum being extracted. Therefore, the OSD approach may fail in detecting covariability between ions which may lead to an incorrect association of the true fragments of the compound. Both ICA-OSD and MCR-OSD exhibit similar performance as can be observed from Table A.2 (Supplementary Materials) and Table 4.1, but there exist some differences in the match score for certain compounds between both OSD methods. This can be explained as the only input for OSD is the elution profile, previously determined in this study by MCR or ICA. Consequently, the elution profiles determined modify the amount of variance captured by PCA, including the amount of variance related to the spectra to be extracted, and the amount of outlier variance from neighboring compounds or noise, and this clearly determines the purity of the eventual extracted spectra. This thought, is the main advantage of OSD, as it is able to deconvolve the spectrum for a certain compound only with the shape of its elution profile, and therefore independent of the quality of the elution profiles extracted for the rest of the compounds.

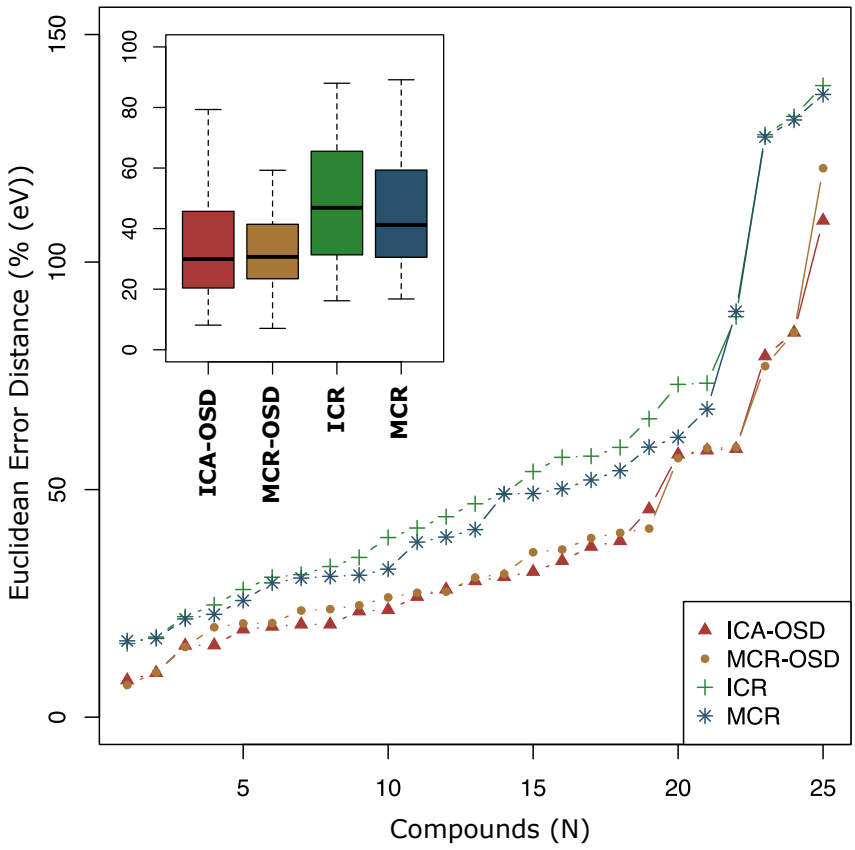

Figure 4-5: Euclidean error distance curves. This shows how close each compound is to the original spectrum in terms of relative error. This graphic assists the evaluation of the deconvolution capability between the methods compared. Outliers in the boxplot are not shown. The ρ-values for the euclidean error distances between LS and OSD approaches show that those differences are statistically significative (ρ-value < 0.0005).

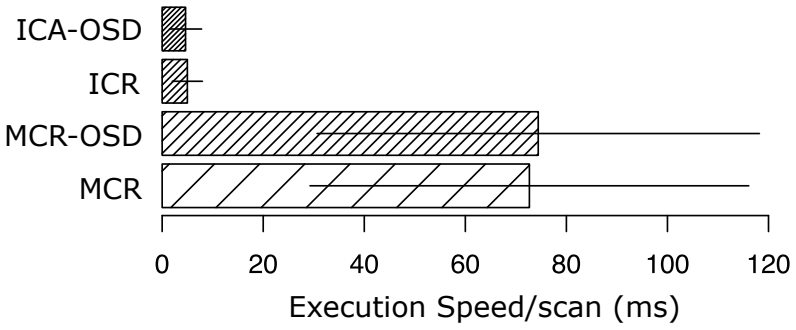

Figure 4-6: Time comparison between methods. The barplot shows the mean and standard deviation speed of execution, in miliseconds, necessary to proces one scan of data by each method.

4.3.3 Execution time comparison

Finally, the execution time differences between ICA–OSD, MCR–OSD, ICR and MCR are shown in Figure 4-6. Each method was tested by processing 2000 scans of raw data (3.3 min of sample) with a range of 566 m/z fragments. These bar plots show the mean speed of execution per scan. From this picture, it can be appreciated that both ICR and ICA–OSD offer the most rapid processing of the chromatogram. The total time difference between the ICA- and MCR-based methods becomes more important as the number of samples to process increases. Data were processed using a 2.4 GHz Intel Core 2 Duo processor with 4 GB of 1067 MHz DDR3 RAM.

4.4 Conclusion

This paper demonstrates the capability and suitability of independent component regression (ICR) for GC–MS compound identification as an alternative to multivariate curve resolution. The results given by ICR are comparable to the results given by MCR, but ICR is is superior in terms of execution time. This is of special interest in metabolomics due to the high amount of data that GC–MS currently generates and the quantity of samples that are analyzed in metabolomics experiments. Also, a novel OSD approach using principal component analysis as an alternative to the traditional least squares approach is introduced, allowing the extraction of refined spectra when compounds elute under the influence of biological matrices, compound co-elution or other types of noise.

Bibliography

[1] G.J. Patti, O. Yanes, G. Siuzdak Innovation: Metabolomics: the apogee of the omics trilogy. *Nature Reviews Molecular Cell Biology*, 13 (2012) 63–269.

[2] Aihua Zhang, Hui Sun, and Xijun Wang. Serum metabolomics as a novel diagnostic approach for disease: a systematic review. *Analytical and Bioanalytical Chemistry*, 404 (2012) 1239–1245.

[3] C.Ruckebusch, L.Blanchet. Multivariate curve resolution: a review of advanced and tailored applications and challenges. *Analytical Chimica Acta*, 765 (2013) 28–36.

[4] A. de Juan, J. Jaumot, R. Tauler. Multivariate Curve Resolution (MCR). Solving the mixture analysis problem. *Anal. Methods*, 6 (2014) 4964.

[5] P. Gemperline Practical Guide to Chemometrics, second ed., *CRC Press*, 2012.

[6] J.-F. Cardoso and A. Souloumiac. Blind beamforming for non-Gaussian signals. *Radar and Signal Processing, IEE Proceedings F*, 140 (1993) 362–370.

[7] G. Wang, Q. Ding, Y. Sun, L. He, X. Sun. Estimation of source infrared spectra profiles of acetylspiramycin active components from troches using kernel independent component analysis, *Spectrochim. Acta A: Mol. Biomol. Spectrosc*, 70 (2008) 571–576

[8] M. Toiviainen, F. Corona, J. Paaso, P. Teppola. Blind source separation in diffuse reflectance NIR spectroscopy using independent component analysis, *Journal of Chemometrics* 24 (2010) 514–522.

[9] I. Schelkanova, V. Toronov. Independent component analysis of broad- band near-infrared spectroscopy data acquired on adult human head, *Biomed. Opt. Express* 3 (2012) 64–74

[10] Y.B. Monakhova, S.S. Kolesnikova, S.P. Mushtakova. Independent component analysis algorithms for spectral decomposition in UV/VIS analysis of metalcontaining mixtures including multimineral food supplements and platinum concentrates, *Anal. Methods* 5 (2013) 2761–2772

[11] I. Toumi, S. Caldarelli, B. Torrsani. A review of blind source separation in NMR spectroscopy. *Prog. Nucl. Magn. Reson. Spectrosc.* 81 (2014) 37–64

[12] G.Wang, Q.Ding, Z.Hou Independent component analysis and its applications in signal processing for analytical chemistry. *TrAC – Trends Anal. Chem* 27 (2008) 368–376.

[13] Guoqing Wang, Wensheng Cai, and Xueguang Shao. A primary study on resolution of overlapping GC-MS signal using mean-field approach independent component analysis. *Chemometrics and Intelligent Laboratory Systems*, 82 (2006) 137–144.

[14] Guoqing Wang, Wensheng Cai, and Xueguang Shao. A post-modification approach to independent component analysis for resolution of overlapping GC/MS signals: from independent components to chemical components. *Sci. China Ser. B: Chem.* 50 (2007) 530–537.

[15] Zhichao Liu, Wensheng Cai, and Xueguang Shao. Sequential extraction of mass spectra and chromatographic profiles from overlapping gas chromatography-mass spectroscopy signals. *Journal of Chromatography A* 1190 (2008) 358–364.

[16] X. Shao, Z. Liu, W. Cai. Extraction of chemical information from complex analytical signals by a non-negative independent component analysis. *Analyst* 134 (2009) 2095–2099.

[17] Xueguang Shao, Zhichao Liu, and Wensheng Cai. Resolving multi-component overlapping GC-MS signals by immune algorithms. *TrAC Trends in Analytical Chemistry*, 28 (2009) 1312–1321.

[18] J.V. Stone. Independent Component Analysis: A Tutorial Introduction, *A Bradford Book*, Cambridge, MA, 2004

[19] H. Parastar, M. Jalali-Heravi, R. Tauler, Is independent component analysis appropriate for multivariate resolution in analytical chemistry? *Trends Anal. Chem.* 31 (2012) 134–143.

[20] X. Shao, W. Wang, Z. Hou, W. Cai A new regression method based on independent component analysis. *Talanta* 69 (3) (2006) 676–680

[21] Jan Hummel, Nadine Strehmel, Joachim Selbig, Dirk Walther, and Joachim Kopka. Decision tree supported substructure prediction of metabolites from GC-MS profiles. *Metabolomics*, 6 (2010) 322–333.

[22] N.Psychogios,D.D.Hau,J.Peng,A.C.Guo,R.Mandal,S.Bouatra,I.Sinelnikov,R. Krishnamurthy, R. Eisner, B. Gautam, N. Young, J. Xia, C. Knox, E. Dong, P. Huang, Z. Hollander, T.L. Pedersen, S.R. Smith, F. Bamforth, R. Greiner, B. McManus, J.W. Newman, T. Goodfriend, D.S. Wishart, The human serum metabolome, *PLoS ONE* 6 (2011) e16957

[23] W.B.Dunn,D.Broadhurst,P.Begley,E.Zelena,S.Francis-McIntyre,N.Anderson, M. Brown, J.D. Knowles, A. Halsall, J.N. Haselden, A.W. Nicholls, I.D. Wilson, D.B. Kell, R. Goodacre, Human Serum Metabolome (HUSERMET) Consortium, Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry, *Nat. Protoc.* 6 (2011) 1060–1083.

[24] Y. Ni, Y. Qiu, W. Jiang, K. Suttlemyre, M. Su, W. Zhang, W. Jia, X. Du, ADAP- GC 2.0: deconvolution of coeluting metabolites from GC/TOF-MS data for metabolomics studies, *Anal. Chem*, 84 (2012) 6619–6629.

[25] A de Juan, Y Vander Heyden, R Tauler, and D. L Massart. Assessment of new constraints applied to the alternating least squares method. *Analytica Chimica Acta*, 346 (1997) 307–318.

[26] Katty X. Wan, Ilan Vidavsky, and Michael L. Gross. Comparing similar spectra: from similarity index to spectral contrast angle. *Journal of the American Society for Mass Spectrometry*, 13 (2002) 85–88.

[27] A. Savitzky, M.J.E. Golay. Smoothing and differentiation of data by simplified least squares procedures *Anal. Chem*, 36 (1964) 1627–1639.

[28] D.N.Rutledge, D. Jouan-Rimbaud Bouveresse. Independent components analysis with the JADE algorithm, *TrAC – Trends Anal. Chem.* 50 (2013) 22–32.

[29] Ivo H. M. van Stokkum, Katharine M. Mullen, and Velitchka V. Mihaleva. Global analysis of multiple gas chromatography-mass spectrometry (GC/MS) data sets: A method for resolution of co-eluting components with comparison to MCR-ALS. *Chemometrics and Intelligent Laboratory Systems*, 95 (2009) 150–163.

[30] Y.B. Monakhova, A.M. Tsikin, T. Kuballa, D.W. Lachenmeier, S.P. Mushtakova. Independent component analysis (ICA) algorithms for improved spectral deconvo-

lution of overlapped signals in 1H NMR analysis: application to foods and related products. *Magn. Reson. Chem.* 52 (2014) 231–240.

[31] S. Peters, H.-G. Janssen, G. Vivo-Truyols. A new method for the automated selection of the number of components for deconvolving overlapping chromatographic peaks. *Anal. Chim. Acta* 799 (2013) 29–35.

Chapter 5

Automated resolution of chromatographic signals by independent component analysis - orthogonal signal deconvolution in comprehensive gas chromatography/mass spectrometry-based metabolomics

Published as: Domingo-Almenara X, Perera A, Ramírez N, Brezmes J. Computer Methods and Programs in Biomedicine. Vol. 130 (2016) 135–141. DOI: 10.1016/j.cmpb.2016.03.007.

Abstract

Comprehensive gas chromatography - mass spectrometry (GC×GC–MS) provides a different perspective in metabolomics profiling of samples. However, algorithms for GC×GC–MS data processing are needed in order to automatically process the data and extract the purest information about the compounds appearing in complex biological samples. This study shows the capability of independent component analysis - orthogonal signal deconvolution (ICA–OSD), an algorithm based on blind source separation and distributed in an R package called *osd*, to extract the spectra of the compounds appearing in GC×GC–MS chromatograms in an automated manner. We studied the performance of ICA–OSD by the quantification of 38 metabolites through a set of 20 Jurkat cell samples analyzed by GC×GC–MS. The quantification by ICA–OSD was compared with a supervised quantification by selective ions, and most of the R^2 coefficients of determination were in good agreement (R^2>0.90) while up to 24 cases exhibited an excellent linear relation (R^2>0.95). We concluded that ICA–OSD can be used to resolve co-eluted compounds in GC×GC–MS.

5.1 Introduction

Metabolomics is the study of low molecular weight compounds in biological systems [1]. Particularly, metabolomics focuses on comparing healthy versus metabolomic disease organisms and, therefore, it attempts to discover predictive biomarkers by detecting early biochemical changes before the appearance of the disease [2]. For that purpose, metabolomics experimental designs include non-targeted analysis of the samples as there is no prior knowledge of the metabolites that may be involved not only in fully developed metabolomic diseases, but also in pre-symptomatic stages.

Analytical techniques to identify and quantify metabolites include the best-established gas chromatography-mass spectrometry (GC–MS). Gas chromatography separates the compounds contained in a sample while passing through a chromatographic column. However, when two or more compounds do not completely separate chromatographically, those compounds are known to be co-eluted, and this clearly affects the correct quantification and identification of the metabolites. In that sense, comprehensive gas chromatography - mass spectrometry (GC×GC–MS) [3, 4] was devised to minimize co-elution. In GC×GC–MS, the sample pass through two chromatographic columns with orthogonal polarity properties, which improves the compound separation and it leads to an increased compound detection capacity as co-elution is diminished.

However, compounds in the samples usually appear at trace levels and different sources of noise derived from the instrument and the sample biological matrix may interfere with the correct identification of the compounds. In the same way, GC×GC–MS generates large quantity of data and its interpretation can not be conducted manually. In that sense, GC×GC–MS data processing algorithms are needed to turn the chromatographic signals into interpretable biological information. Besides, GC×GC–MS samples are composed by a large amount of data in comparison with GC–MS samples, and algorithms for GC×GC–MS data processing should be

optimized for a fast data processing.

As reviewed in [5], some of the existing data processing algorithms that can be applied to resolve mixtures in comprehensive gas chromatography include PARAFAC [6] and multivariate curve resolution - alternating least squares (MCR–ALS) [7]. Contrarily to MCR, PARAFAC can be only applicable to a three-way data set.

In the past years, independent component analysis (ICA) [8] has been introduced as an alternative to the traditional MCR for GC–MS data analysis [9, 10, 11]. ICA is a blind source separation (BSS) technique used to separate linearly mixed sources, i.e., it is capable of separating and retrieve the original compound sources - elution profile or spectra - from a mass spectra chromatogram. Whereas MCR–ALS resolves a chromatographic mixture by minimizing the residual error between the data and the predicted model, ICA uses another type of measure which is the statistical independence, and it estimates the original compound sources by maximizing the independence between components. ICA is widely applied in biomedical sciences, including data processing in electroencephalography recordings [12, 13, 14], and it is also one of the most reported algorithms for resolution of spectroscopy mixtures. More recently, we have developed a new method known as independent component analysis - orthogonal signal deconvolution (ICA–OSD) [15], embedded in an R package, that uses a combination of ICA and principal component analysis (PCA) to identify co-eluted compounds in GC–MS. In ICA–OSD, PCA is proposed as an alternative to the typical use of least squares (LS) in MCR–ALS. The application of LS for spectra extraction has different drawbacks, detailed in [15], which can be summarized in the fact that no correlation or covariance information is taken into account when applying LS, and therefore LS may find difficulties in distinguishing noise and the different compound fragments. This may lead to introducing a bias into the LS regressors specially in situations of co-elution or under undue biological matrix interference. Besides, whereas the current ICA-based methods consider the spectra as the

independent source in the chromatograms, in ICA–OSD we implemented a different approach where we assumed that the elution profile was the independent source, as opposite to the spectra. In that sense, we used ICA to extract the elution profiles and then determine the spectra by means of OSD. Finally, ICA–OSD shown itself as a computationally faster alternative to MCR–ALS. Up to the date, the capability of independent component analysis - orthogonal signal deconvolution for compound quantification in chromatographic signals has not been studied.

In this paper we propose an automated method to deconvolve compounds appearing in GC×GC–MS samples by independent component analysis - orthogonal signal deconvolution.

5.2 Materials and methods

5.2.1 Materials

The performance of ICA–OSD was evaluated through a set of 38 metabolites appearing in 20 Jurkat cell samples extracted from human acute T cell lymphoblastic leukemia cell line Jurkat. The samples of this experiment were previously used to report the intersection of phosphoethanolamine with menaquinone-triggered apoptosis by Styczynski *et al.* [16]. More details on the dataset, sample preparation and methods can be found in the original study.

5.2.2 Data analysis and pre-processing

ICA–OSD was used to automatically extract and deconvolve the compounds concentration profiles and spectra. The GC×GC–MS chromatograms were processed by analyzing each modulation cycle separately. Each modulation cycle was first divided in chromatographic peak features (CPFs) using the same criteria as in [17]. The different CPFs contained several compounds, so the algorithm had to deconvolve them

in case of co-elution. The number of factors or components for ICA was determined by evaluation of residual sum of squares (described in Section 3.2).

The chromatograms were automatically processed by ICA–OSD. From the ICA–OSD output we only took into account those metabolites appearing in at least 15 of the 20 samples, so a total of 38 compounds with KEGG number (Kyoto Encyclopedia of Genes and Genomes) were identified. Metabolite identities were curated by spectral similarity with the reference spectra and retention index error by retention time standardization using fatty acid methyl esters (FAME) standards. However, the identity was not confirmed with the analysis of reference standards and therefore, the list of identified metabolites is putative, and a name is assigned to facilitate the interpretation of the results. For this sub-set of 38 compounds, reference relative compound concentration - relative across samples - was determined by the area of a selective ion. The most selective ion was manually determined for each compound.

The spectra determined by ICA–OSD were compared using the dot product [18] against the Golm Metabolome Database (GMD) [19] MS spectra library. The masses 73, 74, 75, 147, 148, and 149 m/z were excluded before processing the sample, since they are ubiquitous mass fragments typically generated from compounds carrying a trimethylsilyl moiety [19]. They were also excluded in the identification. Only the fragments from m/z 70 to 600 were taken into account when comparing reference and empirical spectra, since this is the m/z range included in the downloadable GOLM database. Also, chromatographic signals were filtered using a Savitzky–Golay filter [20]. The ICA algorithm used was the joint approximate diagonalization of eigenvalues (JADE) [21].

5.3 Computational methods and theory

This section describes the ICA–OSD algorithm together with the methodology to determine the number of compounds.

5.3.1 Resolution of GC×GC–MS mixtures by independent component analysis – orthogonal signal deconvolution

Orthogonal signal deconvolution (OSD) is a multivariate method which purpose is to extract and deconvolve the spectrum of a given compound only with the information relative to the compound elution profile. OSD is based on principal component analysis, avoiding thus, the use of least squares used in multivariate curve resolution - alternating least squares (MCR–ALS). Here, the elution profiles are determined by ICA to later determine the spectra using OSD, and in this manner we will refer the complete approach as ICA–OSD.

ICA is mathematically expressed as:

$$X = AZ^T \tag{5.1}$$

where X (N×M) is the matrix containing the mixture of compounds, A (N×k) is the mixing matrix and Z^T(k×M) is the source matrix. N and M are the number of rows and columns of the data matrix X, and k denotes the number of components or compounds in the model. Each row in X holds a m/z channel whereas each column holds the retention time scans. ICA decomposes the data matrix by finding the independent sources contained in X.

As mentioned above, generally ICA-based approaches are based on extracting first the spectra using ICA - the spectra are considered the independent sources - to later estimate the elution profile using different approaches. In our ICA–OSD

implementation, the elution profiles of the compounds are considered the independent sources and thus Z^T holds the elution profile for each compound. Since the elution profiles determined by ICA may be affected by the ICA ambiguity of negativity, the sources in Z^T that express more negative variance than positive are negatively rotated. Moreover, all the components in Z^T are submitted to unimodality constraint to force one local maxima per source. ICA has a second ambiguity related to variance (energy) indetermination, which means that the energy of the recovered compound profiles do not correspond to the real energy of that component. To overcome that, a least squares regression is performed with the estimated sources hold in Z^T against the base ion chromatogram of the matrix X. The base ion chromatogram or BIC is determined by representing the maximum m/z value for each point in the chromatogram.

Once the elution profiles are determined, OSD is applied to extract each corresponding spectra. In OSD, an X'_j sub-data matrix is determined for each compound j in Z^T. This sub-data matrix comprises only the data from X in which the compound profile in Z^T_j is non-zero - the elution profile in Z^T is used as a mask to suppress the surrounding data non-related to the compound -. A PCA is performed over the sub-data matrix to determine the spectra associated to each compound. PCA can be mathematically expressed as:

$$X'_j = YW^T \tag{5.2}$$

where $X'(N\times M)$ is the sub-data matrix to decompose, $Y(N\times M)$ is the score matrix and $W(M\times M)$ is the loading or eigenvectors matrix. For each compound profile, the PCA decorrelates the information of the sub-data matrix and decomposes it into a matrix W^T (Eq. 5.2) which is a set of orthogonal spectra and a matrix Y which is associated to the retention time covariance response for each spectrum in W^T. The matrix W^T holds the spectra of the compound of interest together with the spectra of the different sources of noise - such as co-eluted substances or biological matrix interference -. To determine which spectrum is related to the compound of interest

we compute the correlation between the profile of the compound in $Z^T{}_j$ and the information of the covariance responses determined by the PCA in Y. The component with the highest absolute correlation is the candidate spectra for the compound of interest.

OSD can be summarized in the following steps:

1. Given a Z_j^T compound elution profile, determine a X_j sub-data matrix comprised only of the data of the retention time in which the compound is eluting.

2. Apply a PCA over X_j. The result is a score matrix Y and loading matrix W.

3. Determine the correlation coefficient between Z_j^T and each component in Y and select the component h with the highest absolute correlation value.

4. Select the component h in W, rotate W_h according to the sign of the previous determined correlation coefficient, and clip to zero all the negative values. W_h is now considered to be the spectrum of Z_j^T.

After the spectra are determined, the elution profiles are refined by the application of a NNLS regression of all the spectra against the data matrix X.

5.3.2 Determination of number of components

To define the ICA model, it requires a fixed number of components. The number of components is closely related to the number of compounds present in the mixture, as usually the model to define the data is not only constructed by pure compounds but also by baseline, noise, or other interferences. An iterative residual sum of squares (RSS) approach was used to automatically determine the number of components for the ICA model. The RSS can be expressed as:

$$RSS(k) = \sum_{i=1}^{N} (X - X^*(k))^2 \tag{5.3}$$

where, X is the original mixture matrix, $X^*(k)$ is the resolved matrix by ICA–OSD using k components, and N is the total length of the unfolded X matrix. For each k in $k = 1, 2, ..., N$, ICA–OSD resolves the X data with k components and it determines the RSS. This method yields a decreasing RSS curve that tends to a minimum. The proper number of factors is determined when the addition of more components does not significantly decrease the explained variance, i.e., when the RSS error reaches a certain threshold.

5.4 Results and discussion

The chromatographic data was automatically processed with our proposed method ICA–OSD. Metabolites eluting in more than one modulation cycle were associated based on their identity and quantified together (sum of concentrations). The metabolites across samples were aligned also based on their identity. Table 5.1 shows the list of the identified compounds along with their 1st and 2nd retention times and the identification match factor (MF). The identification match factor is determined by dot product between the averaged compound spectra across samples and the reference spectra (GMD). The closer the score to one hundred, the more exact and pure the spectra extracted. The table also shows the linear regression coefficient of determination (R^2) between our empirical method ICA–OSD and the selective ion area (reference model). In order to demonstrate the ICA–OSD quantification capability along a wide dynamic range of metabolite concentration, we determined the relative compound concentration (Rel. C.) which is the quotient between the mean concentration of each compound and the mean concentration of all the compounds listed in the table.

In this study, we use the coefficient of determination R^2 as a metric to describe the relative deviation between our proposed method for quantification (ICA–OSD)

and our reference model (selective ion). From the given results, most of the R2 coefficients are in good agreement (R^2>0.90) while up to 24 cases exhibit an excellent linear relation (R^2>0.95). Overall, ICA–OSD conducted a reliable quantification of compounds even when those occurred at low concentration or appeared co-eluted.

The efficiency of ICA–OSD is directly conditioned by the degree of noise and co-elution with other compounds. To illustrate this, and the operation of ICA–OSD for compound deconvolution we shown two different examples of co-elution situations in GC×GC–MS. Figure 5-1 shows the total ion chromatogram (BIC) in dotted grey line, and the resolved compound elution profiles by ICA–OSD in color lines, of two selected retention time windows from different modulation cycles.

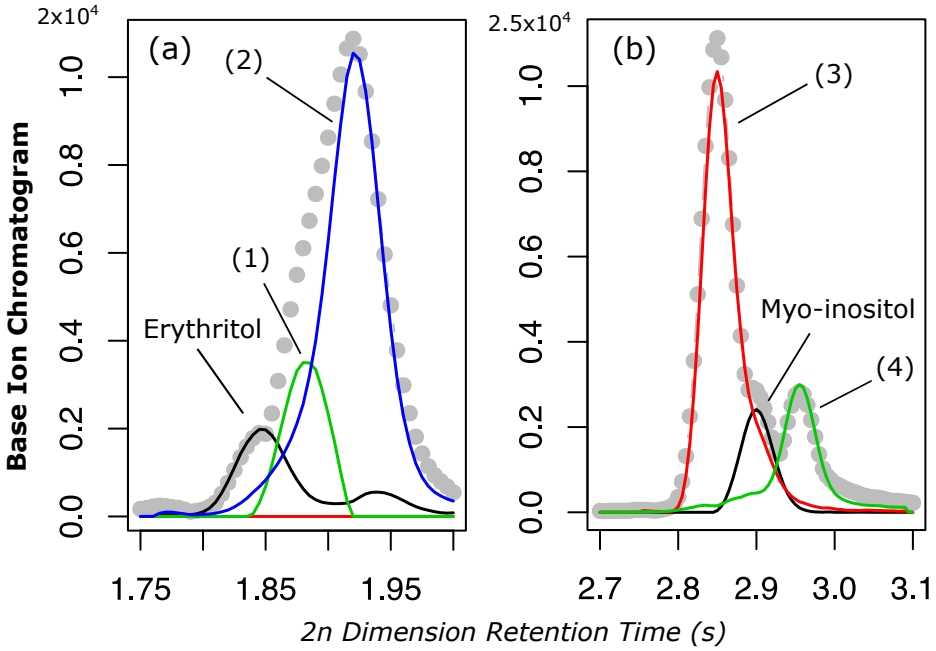

Figure 5-1: Two cases of co-elution resolved by ICA–OSD. The dotted grey line represents the BIC whereas the resolved profiles are shown in the solid-colored line. In (a), erythritol appear in co-elution with other unkown compounds (1, 2). In (b), myo-inositol appear also in co-eluted with an unkown compound (3, 4).

In Figure 5-1 (a), three compounds appear under the same chromatographic peak, those three compounds were resolved by ICA–OSD and one of them was identified as erythritol (4TMS). Similarly, in Figure 5-1 (b) three compounds appear co-eluted

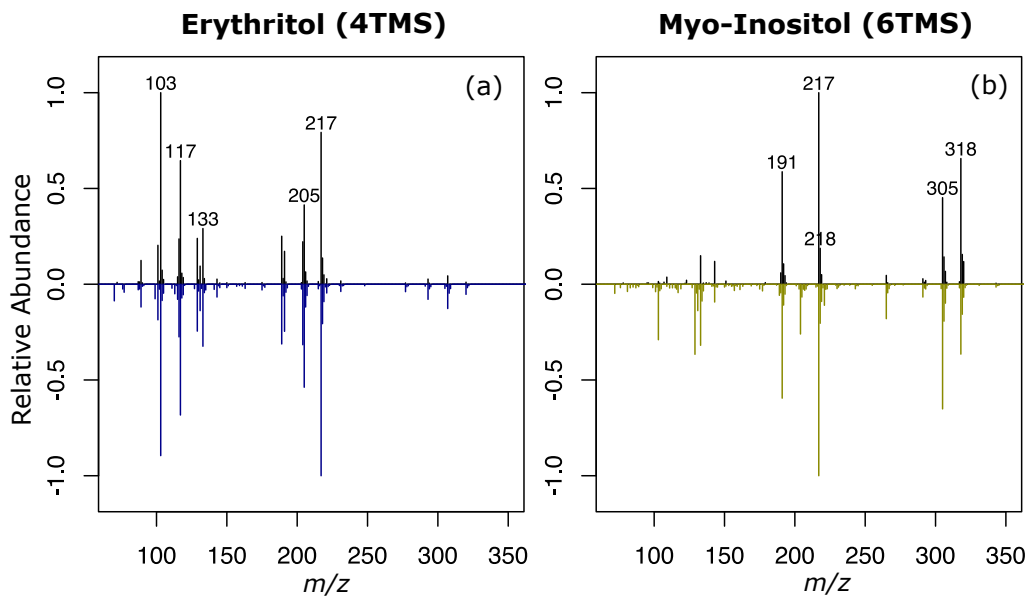

Figure 5-2: Representation of the extracted spectra (black) by ICA–OSD and the reference GMD spectra (color), for the cases shown in Figure 1, erythritol and myo-inositol. Reference spectra are shown negatively rotated in the same axis for a better visual appreciation.

but resolved by ICA–OSD; on of them was identified as myo-inositol (6TMS). The resolved spectra for erythritol and myo-inositol are shown in Figure 5-2 where we can visually compare the empirical (black and positive) and the reference (color and negative) spectra. In both cases ICA–OSD successfully extracted the spectra needed to properly identify both compounds. In the Figure 5-1 (a) case, erythritol appears low concentrated and in co-elution with a more intense compound. Despite that, ICA–OSD is capable of extracting a sufficient pure spectrum to allow a correct identification, with a match score of 98 % - for the given sample case -. In Figure 5-1 (b), myo-inositol appears strongly interfered by another more concentrated compound. As a result, ICA–OSD fails in correctly associate the fragments between m/z 100 and 150 (Figure 5-2 (b)), which appears in the reference spectrum but they do not appear in the empirical spectrum. Also, the ions m/z 305 and 318 appears to be interfered, and their relative intensities differ from the reference pattern. Consequently, the match score of myo-inositol in this given case is 87 %. This is a clear example of the problems for the correct identification of metabolites that co-elution brings. The identification

performance can be assessed also in an example of a set of spectra extracted by ICA–OSD shown in Figure 5-3, where we can visually compare the empirical (black and positive) and the reference (color and negative) spectra for each compound. The figure shows the spectra extracted for lactic acid (2TMS), phosphoric acid (3TMS), fumaric acid (2TMS) and glycerol (3TMS), and this exemplifies the capability of ICA–OSD to successfully extract spectra from chromatographic mixtures.

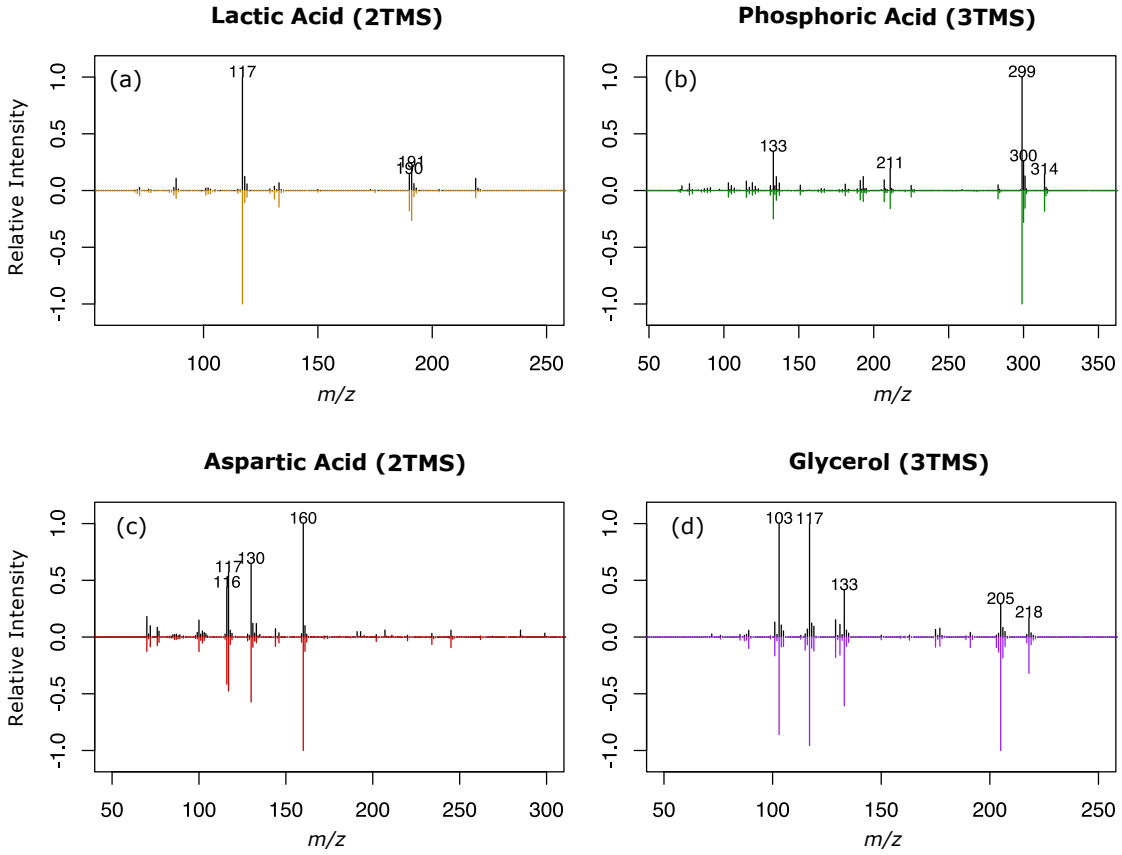

Figure 5-3: Representation of a set of extracted average - across samples - spectra (black) by ICA–OSD and the reference GMD spectra (color). Reference spectra are shown negatively rotated in the same axis for a better visual appreciation.

As mentioned before, one of the most important factors that difficulties the identification is co-elution. In those cases, the spectrum of each compound has to be correctly separated - resolved or deconvolved - from co-eluted compounds or other noise interferences. Despite that one of the differential characteristics of GC×GC–MS with

respect to GC–MS is the reduction of the co-elution problem, we still find co-eluted peaks across the second retention time dimension. Here we show how ICA–OSD is also an effective method for the resolution of chromatographic signals including those generated by GC×GC–MS. Due to noise and other interferences, OSD may fail in correctly classify the m/z when deconvolving spectra. This means that OSD would fail in associating a certain m/z to a compound where other methods based on least squares, such as MCR–ALS would probably not, as OSD is a more conservative approach . On the contrary, OSD brings more accuracy generally in co-eluted situations as attempts to differentiate which ions correspond to the compound of interest [15].

Here we applied ICA–OSD in each modulation cycle separately. We later grouped the compounds appearing in different modulation cycles according to their identity. This may also affect the quantification of compounds as the same compound can be identified with a different name between or within samples. Automatic alignment or grouping of compounds between and within samples after deconvolution is still an important problem that has to be tackled.

5.5 Conclusions

We previously shown that ICA–OSD was able to successfully extract the spectra from co-eluted compounds in GC–MS [15], but the capability of ICA–OSD to quantify metabolites was not evaluated. In this study we evaluated a method to automatically resolve chromatographic data in GC×GC–MS samples with ICA–OSD. Besides, ICA–OSD is an efficient method in terms of speed of execution as previously shown in [15], which is an important advantage for GC×GC–MS data processing due to the large amount of data that metabolomics experiments generate with this analytical platform. This study concludes that ICA–OSD can be used to resolve co-eluted compounds in GCxGC/MS-based metabolomics samples. The package *osd*

is available from the Comprehensive R Archive Network (CRAN) at http://cran.r-project.org/package=osd and it comes under the GNU General Public Licence (GPL) 2.0 or higher licence.

Table 5.1: List of identified compounds in Jurkat cell samples. MF is the match factor, R^2 is the linear regression coefficient, and Rel. C is the relative concentration.

No.	Rt1	Rt2	Name	MF	R^2	Rel. C (%)
1	4.5	1.8	Boric acid (3TMS)	92	0.96	137.28
2	4.58	2.9	Alanine (2TMS)	95	0.82	24.87
3	5.33	1.99	Valine (1TMS)	98	0.92	27.06
4	5.58	2.06	Lactic acid (2TMS)	99	0.99	939.18
5	5.75	2.12	Glycolic acid (2TMS)	98	0.90	61.17
6	5.92	1.94	Ethanolamine (3TMS)	87	0.84	16.26
7	6.5	1.9	Isovaleric acid, 2-oxo- (1MEOX) (1TMS) MP	89	0.98	101.7
8	6.67	2.38	Furan-2-carboxylic acid (1TMS)	98	1.00	25.02
9	7.5	2.78	Phosphoric acid (3TMS)	98	0.97	12.84
10	7.6	1.86	Glycerol (3TMS)	90	0.98	1294.11
11	8.1	2.38	Succinic acid (2TMS)	98	0.85	49.99
12	8.6	2.12	Nonanoic acid (1TMS)	91	0.98	105.88
13	9.1	2.04	Threonine, allo- (3TMS)	98	0.90	12.23
14	9.5	2.48	Aspartic acid (2TMS)	95	0.85	20.99
15	9.6	2.06	Malic acid (3TMS)	72	0.99	13.83
16	9.8	2.11	Decanoic acid (1TMS)	96	1.00	11.63
17	10.6	1.86	Erythritol (4TMS)	97	0.99	56.44
18	11.4	2.48	Proline [+CO2] (2TMS)	99	0.98	7.99
19	11.6	2.54	Hypotaurine (3TMS)	97	0.98	74.63
20	11.8	2.26	Glutamic acid (3TMS)	98	0.99	93.16
21	12.23	3.79	Pyroglutamic acid (2TMS)	99	0.89	112.16
22	12.23	3.05	Proline, 4-hydroxy-, cis- (3TMS)	98	0.82	14.45
23	12.82	4.28	Glutamic acid (2TMS)	97	0.97	17.37
24	13.23	3.26	Glutamic acid (3TMS)	98	0.99	80.1
25	13.48	3.01	Dodecanoic acid (1TMS)	98	0.94	25.84
26	13.9	3.65	Pyrophosphate (4TMS)	96	0.99	5.25
27	14.23	3.94	Glucose, 2-amino-2-deoxy- (4TMS) MP	91	0.99	8.38
28	14.57	2.89	Xylitol (5TMS)	98	0.92	24.63
29	14.98	3.41	Glycerol-3-phosphate (4TMS)	98	0.92	93.59
30	15.4	3	Ornithine (4TMS)	97	1.00	3.95
31	15.57	3.02	Tetradecanoic acid (1TMS)	98	0.97	154.25
32	16.07	3.25	Tyrosine (2TMS)	99	0.84	3.84
33	16.15	2.85	Psicose (1MEOX) (5TMS) BP	99	0.96	270.47
34	16.4	2.85	Glucose (1MEOX) (5TMS) MP	97	1.00	149.68
35	16.48	2.83	Mannose (1MEOX) (5TMS) MP	98	1.00	66.15
36	17.65	2.9	Inositol, allo- (6TMS)	94	0.95	19.81
37	18.98	2.98	Octadecenoic acid, 9-(Z)- (1TMS)	91	0.89	30.99
38	22.9	2.8	Sucrose (8TMS)	94	1.00	3.43

Bibliography

[1] G.J. Patti, O. Yanes, G. Siuzdak Innovation: Metabolomics: the apogee of the omics trilogy. *Nature Reviews Molecular Cell Biology*, 13 (2012) 63–269.

[2] Aihua Zhang, Hui Sun, and Xijun Wang. Serum metabolomics as a novel diagnostic approach for disease: a systematic review. *Analytical and Bioanalytical Chemistry*, 404 (2012) 1239–1245.

[3] Luigi Mondello, Peter Quinto Tranchida, Paola Dugo, and Giovanni Dugo. Comprehensive two-dimensional gas chromatography-mass spectrometry: a review. *Mass Spectrometry Reviews*, 27 (2008) 101–124.

[4] John V. Seeley and Stacy K. Seeley. Multidimensional Gas Chromatography: Fundamental Advances and New Applications. *Analytical Chemistry*, 85 (2012) 557–578.

[5] J. T. V. Matos, Regina M. B. O. Duarte, and Armando C. Duarte. Trends in data processing of comprehensive two-dimensional chromatography: State of the art. *Journal of Chromatography B*, 910 (2012) 31–45.

[6] Nicolaas (Klaas) M. Faber, Rasmus Bro, and Philip K. Hopke. Recent developments in CANDECOMP/PARAFAC algorithms: a critical review. *Chemometrics and Intelligent Laboratory Systems*, 65 (2003) 119–137.

[7] A. de Juan, J. Jaumot, R. Tauler. Multivariate Curve Resolution (MCR). Solving the mixture analysis problem. *Anal. Methods*, 6 (2014) 4964.

[8] Stephen Roberts and Richard Everson. *Independent Component Analysis: Principles and Practice*. Cambridge University Press, March 2001.

[9] Guoqing Wang, Wensheng Cai, and Xueguang Shao. A primary study on resolution of overlapping GC-MS signal using mean-field approach independent component analysis. *Chemometrics and Intelligent Laboratory Systems*, 82 (2006) 137–144.

[10] Zhichao Liu, Wensheng Cai, and Xueguang Shao. Sequential extraction of mass spectra and chromatographic profiles from overlapping gas chromatography-mass spectroscopy signals. *Journal of Chromatography A*, 1190 (2008) 358–364.

[11] Xueguang Shao, Zhichao Liu, and Wensheng Cai. Resolving multi-component overlapping GC-MS signals by immune algorithms. *TrAC Trends in Analytical Chemistry*, 28 (2009) 1312–1321.

[12] Silvia Comani, Dante Mantini, Paris Pennesi, Antonio Lagatta, and Giovanni Cancellieri. Independent component analysis: fetal signal reconstruction from magnetocardiographic recordings. *Computer Methods and Programs in Biomedicine*, 75 (2004) 163–177.

[13] F. J. Martinez-Murcia, J. M. Gorriz, J. Ramirez, C. G. Puntonet, and I. A. Illan. Functional activity maps based on significance measures and Independent Component Analysis. *Computer Methods and Programs in Biomedicine*, 111 (2013) 255–268.

[14] S. Spasic, Lj. Nikolic, D. Mutavdzic, and J. Saponjic. Independent complexity patterns in single neuron activity induced by static magnetic field. *Computer Methods and Programs in Biomedicine*, 104 (2011) 212–218.

[15] X. Domingo-Almenara, A. Perera, N. Ramirez, N. Canellas, X. Correig, J. Brezmes. Compound identification in gas chromatography/mass spectrometry-based metabolomics by blind source separation. *Journal of Chromatography A*, 1409 (2015) 226–233.

[16] S. Dhakshinamoorthy, N. Dinh, J. Skolnick, M.P. Styczynski. Metabolomics identifies the intersection of phosphoethanolamine with menaquinone-triggered apoptosis in an in vitro model of leukemia. *Molecular Biosystems*, 11 (2015) 2406–2416.

[17] Y. Ni, Y. Qiu, W. Jiang, K. Suttlemyre, M. Su, W. Zhang, W. Jia, X. Du, ADAP- GC 2.0: deconvolution of coeluting metabolites from GC/TOF-MS data for metabolomics studies, *Anal. Chem*, 84 (2012) 6619–6629.

[18] Katty X. Wan, Ilan Vidavsky, and Michael L. Gross. Comparing similar spectra: from similarity index to spectral contrast angle. *Journal of the American Society for Mass Spectrometry*, 13 (2002) 85–88.

[19] Jan Hummel, Nadine Strehmel, Joachim Selbig, Dirk Walther, and Joachim Kopka. Decision tree supported substructure prediction of metabolites from GC-MS profiles. *Metabolomics*, 6 (2010) 322–333.

[20] A. Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures *Anal. Chem*, 36 (1964) 1627–1639.

[21] J.-F. Cardoso and A. Souloumiac. Blind beamforming for non-Gaussian signals. *Radar and Signal Processing, IEE Proceedings F*, 140 (1993) 362–370.

Chapter 6

Avoiding hard chromatographic segmentation: a moving window approach for the resolution of GC–MS signals in metabolomics by multivariate methods.

Abstract

Gas chromatography – mass spectrometry (GC–MS) produces large and complex datasets characterized by co-eluted compounds and at trace levels, and with a distinct ion-redundancy as a result of the high fragmentation produced by electron impact ionization. Compounds in GC–MS can be resolved by taking advantage of the multivariate nature of GC–MS data by applying multivariate resolution methods. However, to ensure a correct performance, multivariate methods have to be applied in small regions of the chromatogram, and therefore the chromatogram is segmented prior to the application of the algorithms. The automation of this segmentation process is a challenging task as it implies separating between informative data and noise from the chromatogram. This study demonstrates the capabilities of independent component analysis – orthogonal signal deconvolution (ICA–OSD) and multivariate curve resolution – alternating least squares (MCR–ALS) with a moving window implementation. We evaluated the proposed methods through a quantitative analysis of GC-qTOF MS data from 25 serum samples. The quantitative performance of both ICA–OSD and MCR–ALS moving window-based implementations was compared with the quantification of 33 compounds by the XCMS package. Results shown that most of the R^2 coefficients of determination exhibited a high correlation (R^2>0.90). This demonstrates the capability of both ICA–OSD and MCR–ALS moving window-based to resolve and quantify compounds appearing in GC–MS samples.

6.1 Introduction

Gas chromatography – mass spectrometry (GC–MS) has been extensively applied for compound profiling in metabolomics experiments due to the highly reproducible electron impact ionization process. Electron impact (EI) is a high fragmentation ionization method which leads to an extensive fragmentation. Therefore, the richness of GC–MS data relies on an inherent correlation – or ion-redundancy – between fragments or ions from the same compound, i.e., different peak fragments elute at the same retention time and with the same elution profile [1]. However, compounds in GC–MS may appear co-eluted - chromatographycally not completely separated or resolved - and/or at trace levels. Due to the multivariate nature of GC–MS data, some approaches for its processing have been focused on the implementation of multivariate methods.

The most reported multivariate methods applied for the resolution of GC–MS signals are those based on multivariate curve resolution - alternating least squares (MCR–ALS) [2, 3], or parallel factor analysis (PARAFAC) [4, 5]. Agorithms based on independent component analysis (ICA) have also been applied for GC–MS signal resolution [6, 7, 8]. More recently, an alternative application of ICA, called independent component analysis – orthogonal signal deconvolution (ICA–OSD) [1, 9], for the resolution of GC–MS chromatograms has been introduced, where the concept of independence was twisted: whereas the aforementioned ICA-based methods consider the spectra as the independent source in the chromatograms, ICA–OSD considers the elution profile as the independent source, as opposite to the spectra [9]. Contrarily to the spectra, chromatography aims to separate the compounds along the chromatogram, so compound chromatographic profiles are naturally independent between them and their degree of independence depends on their degree of co-elution. In that sense, in ICA–OSD, ICA is employed to extract the elution profiles and then determine the spectra by means of OSD. Orthogonal signal deconvolution (OSD) is a method that

uses principal component analysis (PCA) as an alternative to the typical use of least squares (LS) used for example in MCR–ALS. When applying LS, no correlation or covariance information is taken into account, and this may introduce a bias into the LS regressors specially in situations of co-elution or under undue biological matrix interference [1, 9]. OSD allows the extraction of more pure spectra in comparison with least squares-based algorithms.

Despite the availability of multivariate methods for GC–MS signal resolution, the correct answering to biological hypothesis or the discovering of new biological insights is the current untargeted GC–MS-based metabolomics challenge. In that sense, all the implementations of multivariate methods should be fully automated, and this automatization should not be limited to the deconvolution process but also the posterior alignment of the resolved metabolites. There is a need for high-throughput application of these multivariate methods. Several automated methods based on the aforementioned algorithms have been reported [10, 11, 12, 13, 14]. However, as curve resolution techniques work in small and regional intervals [13], the application of multivariate methods in high-throughput GC–MS resolution is usually conducted by a hard chromatographic segmentation, i.e., windowing or dividing the chromatogram by selection those regions with putative information – compounds – to be resolved. The automation of this segmentation process is a challenging task as it implies separating what is useful data and what is noise from the chromatogram and thus, selecting regions of the chromatogram without splitting compounds on window borders or loosing useful information, i.e., considering compounds at trace levels as noise.

Moving windows have been used in GC–MS for factor analysis [15, 16, 17, 18]. In these studies, factor analysis techniques are applied trough a moving window with the aim of detecting components or spectral features. Those spectral features can be later resolved for a posterior resolution and comparison between samples. More recently, the concept of sliding window multivariate curve resolution (SW-MCR) [19]

was introduced for the resolution of ion-mobility gas chromatography data. When using a moving - or sliding - window to resolve the chromatogram, the consecutive windows have to be overlapped to ensure that one compound that could be split on window borders is fully covered by the next window. In SW-MCR, they tackle this issue by grouping compounds through consecutive windows based on the similarity of their spectra. Grouping compounds across windows based on spectral similarity is a challenging task, as due to noise, the spectra of the same compounds deconvolved from two consecutive windows may change, which difficults the selection of the best spectra. To our knowledge, the performance and suitability of moving window MCR–ALS and ICA–OSD-based approaches for the automated resolution of GC–MS metabolomics samples has not yet been studied.

In this study we propose an automated application of moving window-based ICA–OSD and MCR–ALS approaches for the resolution of GC–MS signals in biological samples. This approach avoids hard segmentation or windowing of the chromatogram. We propose a duplicity filter based on the minimization of the residual error to filter duplicated compounds resolved across windows, and thus selecting the best models. Also, to increase the automated reproducibility of the results, we use an existing automated method for aligning compounds across samples. We evaluated the proposed methods through a quantitative analysis of GC-qTOF MS data from serum samples of adolescents with hyperinsulinaemic androgen excess and healthy controls and the quantitative results were compared with centWave [20], the peak-picking algorithm implemented in the widely used XCMS package [21, 22].

6.2 Materials and methods

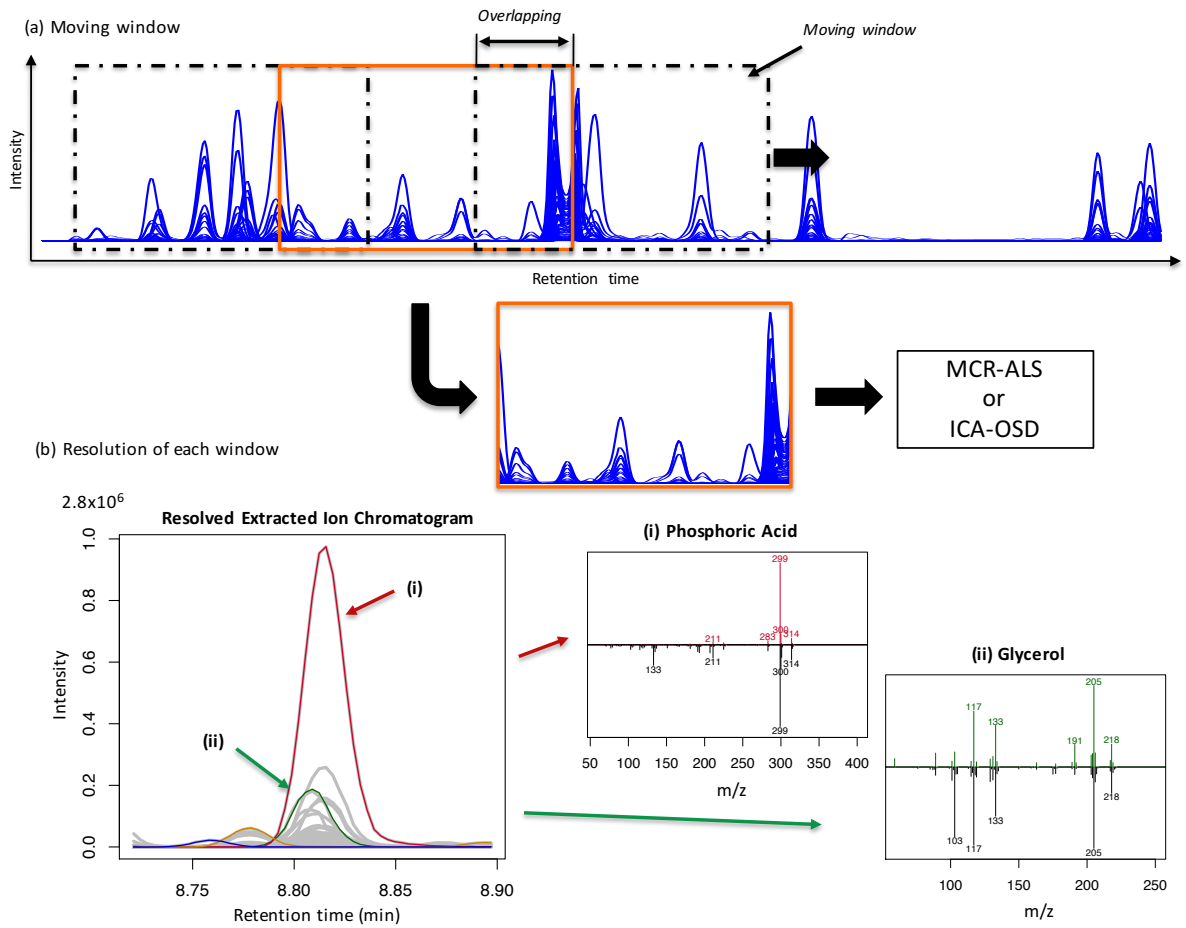

Figure 6-1: In (a), illustration of the moving window approach. A fixed-length window is displaced with a certain overlapp along the chromatogram. The blue lines represent each m/z (extracted ion chromatogram). Each chromatographic window (b) is resolved by ICA–OSD or MCR–ALS into pure chromatographic profiles and spectra. This case shows the resolution of (i) glycerol and (ii) phosphoric acid, which appear strongly co-eluted. For this case, the extracted ion chromatogram is shown in grey, whereas colored solid lines represent the resolved chromatographic profiles. Compound resolved spectra are shown in color red and green along with each reference spectrum negatively rotated in the same axis and shown in black. In this example, the resolved spectra of both phosphoric acid and glycerol - by comparing it with the reference - seems to be affected by the strong co-elution in which they appear.

6.2.1 Materials

The methods were compared by the quantification of 33 metabolites across 25 serum samples (from 11 young, non-obese adolescents with HIAE and 14 age-, weight- and ethnicity-matched healthy controls) [23], analyzed through GC-qTOF MS. This work-

bench was previously used to demonstrate the capabilities of the eRah R package [24]. More details on the dataset, sample preparation and methods can be found in the original study. Briefly, analysis was carried out on a qTOF MS 7200 (Agilent, Santa Clara, CA, USA) coupled to an Agilent 7890A gas chromatography (GC). Derivatized samples (1 μL each) were injected in the gas chromatograph system with a split inlet equipped with a J&W Scientific DB5–MS+DG stationary phase column (30 mm $\times$ 0.25 mm i.d., 0.1 μm film, Agilent Technologies). Helium was used as a carrier gas at a flow rate of 1 mL/min in constant flow mode. The injector split ratio was adjusted to 1:5 and oven temperature was programmed at 70 °C for 1 min and increased at 10 °C/min to 325 °C. The MS was operated in the electron impact ionization mode at 70 eV. Mass spectral data were acquired in full scan mode from m/z 35 to 700 with an acquisition rate of 5 spectra per second.

6.2.2 Data analysis and pre-processing

GC–MS chromatograms were processed using XCMS in order to detect and align features. A feature is defined as an ion entity with a unique m/z and a specific retention time (mzRT). The parameters used in the XCMS workflow were: `xcmsSet (method = 'centWave', ppm = 15, peakwidth = c(1,5)); retcor (method = 'peakgroups', extra = 1, missing=1)` and `group (mzwid = 0.0025, minfrac = 0.5, bw = 5)`. XCMS analysis provided an xcmsSet object containing the retention time, m/z value, and peak intensity (or area) of each feature for every serum sample. For each compound, we selected the feature to be used as a selective ion for quantification reference. Raw GC–MS files are available at MetaboLights with accession number MTBLS321.

Both moving window-based ICA–OSD and MCR–ALS implementations were used to automatically extract and deconvolve the compounds concentration profiles and spectra. The methods were compared using different lengths of window, concretely,

we used 50, 75 and 100 scans length corresponding to 10, 15 and 20 seconds respectively. We used an overlapping of 50 % for all the implementations. The number of factors or components for both ICA and MCR was determined by a singular value decomposition (SVD), as described in [25]. MCR–ALS was initialized by means of a principal component analysis (PCA).

Reference spectra were obtained from the Golm Metabolme Database (GMD) [26, 27]. The fragments at m/z 73, 74, 75, 147, 148, and 149 were excluded before processing the sample, since they are widespread mass fragments typically generated from compounds carrying a trimethylsilyl-moiety [27]. Also, chromatographic signals were filtered by noise and baseline removal as described in [1, 24].The ICA algorithm used was the joint approximate diagonalization of eigenvalues (JADE) [28] implemented in the R package *JADE* [29]. Both MCR–ALS and ICA–OSD algorithms employed were those included in the R package *osd*, freely available as an R package on the Comprehensive R Archive Network (CRAN) at http://CRAN.R-project.org/package=osd. Once resolved, compounds were aligned across samples with *eRah* [24] alignment algorithm (http://CRAN.R-project.org/package=erah).

6.2.3 Moving window resolution of chromatographic signals

The aim of the method is to achieve the resolution of an entire chromatogram. Then, a moving window is proposed where, in each iteration, the window is displaced with a determined overlapping along the retention time (Figure 6-1 (a)). Each chromatographic window is resolved into pure chromatographic profiles and spectra (Figure 6-1 (b)).

We employed two methods for the resolution of mixtures, one is the widely used multivariate curve resolution – alternating least squares (MCR–ALS), and the other is independent component analysis - orthogonal signal deconvolution (ICA–OSD). Both algorithms share the same objective based on the assumption of the Lambert-Beer's

law, which can be mathematically described as follows:

$$D = CS^T + E \tag{6.1}$$

where D ($N \times M$) is the chromatographic window to be resolved, C ($N \times k$) is the resolved concentration profile matrix, S ($M \times k$) is the resolved spectra matrix and E ($N \times M$) is the error matrix. In this notation, N is the number of chromatographic scans (retention time), M is the range of acquisition of the mass-charge ratio (m/z), and k is the number of components or compounds in the model. MCR–ALS uses an iterative least squares algorithm (ALS) to determine both C and S matrices by minimizing the error matrix E. A detailed explanation of MCR–ALS, together with pseudocode, is given elsewhere [30]. In ICA–OSD, independent component analysis is used to extract the chromatographic profile matrix C by considering those the independent source of the chromatogram. After that, orthogonal signal deconvolution (OSD) is applied to determine S. OSD purpose is to extract and deconvolve the spectrum of a given compound only with the information relative to the compound elution profile - which are previously determined by ICA -. A detailed explanation of ICA–OSD is given elsewhere [1, 9].

Both methods obtain a local C and S matrices, corresponding to the resolution of each window into pure chromatographic profiles and spectra, respectively. Each local C and S matrices are appended to a general C_g and S_g matrices containing the resolution of all the chromatogram. As mentioned before, when using a moving window to resolve the chromatogram, the consecutive windows have to be overlapped to ensure that one compound that could be split on window borders is fully covered by the next window. Then, compounds - or often a part of it when it is split by the window border - are expected to be resolved in more than one window. This leads to multiple duplicates that difficulties the selection of the quantitative - correctly resolved - compound. To ensure only one chromatographic profile and spectrum per

compound, a duplicity filter is proposed. First, a correlation matrix for C_g is determined, and those groups of chromatographic profiles that correlate in more than a certain threshold - typically 75 % - are considered that may be duplicated. These groups may be composed of two or more chromatographic profiles. After that, all the possible combinations are considered. As an illustrative example, let us consider that three (N=3) chromatographic profiles C_1, C_2 and C_3 correlate between them. Then, 8 ($2^{N=3}$) possible scenarios are considered. For each scenario, first, a chromatographic matrix D is determined comprising the retention time of the all the considered chromatographic profiles, and after that, a putative D^* matrix is determined by:

$$D^*(k) = C_j S_j^T \tag{6.2}$$

where $D^*(k)$ is the reconstructed matrix and the subindex j denotes the compounds considered in each k=1,2,...,N case. Then, a residual sum of squares for each scenario is determined as follows:

$$RSS(k) = \sum_{i=1}^{N} (D - D^*(k))^2 \tag{6.3}$$

The scenario with the least RSS errors is considered to be the combination that best describes the data, and the chromatographic profiles that are not included in this combination, are removed from C_g and S_g.

6.3 Results

The moving window-based ICA–OSD and MCR–ALS implementations were used to automatically extract and deconvolve the compounds concentration profiles and spectra from all the 25 serum samples. Three different window lengths were employed - 50, 75 and 100 scans length corresponding to 10, 15 and 20 seconds respectively -, all

Table 6.1: Retention time (Rt), quantitative fragment ion (m/z) (XCMS), relative concentration (Rel. C) of 33 compounds. Coefficients of determination (R^2) of the regression between the area and intensity of the resolved chromatographic profile (ICA–OSD and MCR–ALS) and the quantitative ion peak (XCMS) is shown. WL and NF stands for *window length* and *not found* respectively. The number of trimethylsilyl (TMS) derivatives groups are not shown, with the exception of those compounds that appear duplicated. For those cases, the number of trimethylsilyl (TMS) groups is shown in brackets.

					R^2											
					ICA–OSD						MCR–ALS					
Cp.	Rt	m/z	Rel.C	Name	WL=10		WL=15		WL=20		WL=10		WL=15		WL=20	
	(min)		(%)		Area	Int	Area	Int	Area	Int	Area	Int	Area	Int	Area	Int
1	5.73	117	1055	Lactic acid	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2	5.85	173	9	Hexanoic acid	0.41	0.88	0.71	1.00	NF	NF	0.87	1.00	0.93	1.00	NF	NF
3	6.11	72	65	Valine (1)	0.93	1.00	0.98	1.00	0.98	1.00	0.85	1.00	0.95	1.00	0.95	1.00
4	6.53	113	5	Hydroxylamine	0.79	0.74	0.75	0.74	0.74	0.74	0.79	0.70	0.82	0.71	0.80	0.71
5	6.69	131	18	2-Hydroxybutyrate	0.80	1.00	0.96	1.00	1.00	1.00	0.97	1.00	0.99	1.00	0.98	1.00
6	7.08	86	33	Leucine (1)	0.97	1.00	0.90	1.00	NF	NF	0.93	1.00	0.95	1.00	0.85	1.00
7	7.15	191	19	3-Hydroxybutyrate	0.99	1.00	0.99	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
8	7.97	145	4	Valine (2)	0.32	0.95	0.97	0.97	0.95	0.97	0.97	0.97	0.99	0.97	0.98	0.97
9	8.19	130	206	Urea	0.90	0.91	0.96	0.91	0.92	0.94	0.96	0.89	0.95	0.93	0.95	0.94
10	8.36	179	133	Benzoic acid	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	0.99	1.00	0.99	1.00
11	8.55	116	24	Serine	0.90	0.98	0.98	1.00	0.99	1.00	0.98	1.00	0.98	1.00	0.99	1.00
12	8.75	158	18	Leucine (2)	NF	NF	NF	NF	NF	NF	0.98	1.00	NF	NF	NF	NF
13	8.80	205	209	Glycerol	0.99	1.00	1.00	1.00	0.99	1.00	NF	NF	0.31	0.31	0.85	0.90
14	8.81	299	649	Phosphoric acid	1.00	1.00	1.00	1.00	1.00	1.00	0.74	0.75	0.61	0.73	0.8	0.82
15	9.06	218	2	Isoleucine	0.82	0.9	0.79	0.91	NF	NF	0.97	0.94	0.92	0.93	NF	NF
16	9.10	142	9	Proline	0.99	1.00	1.00	1.00	NF	NF	0.99	1.00	0.99	1.00	NF	NF
17	9.58	189	5	Glyceric acid	0.97	1.00	0.98	1.00	0.99	0.99	0.98	0.99	0.98	0.99	0.98	0.99
18	9.85	215	14	Nonanoic acid	0.96	0.99	0.99	0.99	0.98	0.99	0.99	0.98	0.98	0.99	0.99	0.99
19	10.34	291	7	Threonine	0.92	0.96	0.89	0.92	0.98	0.93	0.94	0.97	0.98	0.97	0.98	0.93
20	11.82	230	3	3-hydroxy-trans-proline	0.87	0.99	0.82	1.00	0.95	1.00	0.81	0.99	0.95	0.99	0.95	1.00
21	12.02	156	216	Glumatic Acid (2)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
22	12.73	142	14	Proline [+CO2]	0.92	0.99	0.97	1.00	0.99	1.00	0.99	1.00	0.99	1.00	0.99	1.00
23	13.17	246	6	Glutamic acid (3)	0.99	1.00	0.99	1.00	0.99	1.00	0.99	1.00	1.00	1.00	0.99	1.00
24	13.27	218	24	Phenylalanine	0.99	1.00	0.98	1.00	0.99	1.00	0.99	1.00	0.99	1.00	0.99	1.00
25	13.42	117	101	Dodecanoic acid	0.98	1.00	0.97	1.00	0.98	1.00	0.99	0.99	0.99	0.99	0.99	0.99
26	15.39	142	16	Ornithine	1.00	1.00	0.99	1.00	0.99	1.00	0.99	1.00	0.99	1.00	1.00	1.00
27	15.46	273	15	Citric acid	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	0.99	1.00	0.99	1.00
28	15.54	285	22	Tetradecanoic acid	0.96	0.92	0.88	0.85	0.95	0.92	0.97	0.93	0.96	0.87	0.98	0.94
29	16.44	229	<1	Lysine	0.49	0.87	0.5	0.87	0.25	0.79	0.10	0.78	0.11	0.78	0.24	0.78
30	17.48	129	324	Hexadecanoic acid	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
31	18.23	305	30	Myo-inositol	0.88	0.99	0.91	1.00	0.93	1.00	0.94	0.99	0.87	0.97	0.7	0.98
32	18.24	441	11	Uric acid	0.99	0.98	0.99	0.99	1.00	0.98	0.91	0.9	0.88	0.91	0.82	0.86
33	19.26	356	16	Octadecanoic acid	1.00	0.99	1.00	0.99	1.00	0.99	1.00	0.99	1.00	0.99	1.00	0.99

with an overlap of 50 %. After being resolved, compounds were automatically aligned across samples by the eRah's alignment algorithm. The area and intensity of a set of 33 compounds by ICA–OSD and MCR–ALS were determined by the area and intensity of the resolved compounds elution profile, whereas for the case of XCMS, we manually selected a quantitative mzRT feature corresponding to a selective/quantitative ion. Table 1 shows the list of the 33 metabolites with their retention time, the quantification m/z and a linear regression coefficient of determination (R^2) between our the proposed methods (ICA–OSD and MCR–ALS) and the selective ion area (reference model). In order to demonstrate the ICA-OSD/MCR–ALS quantification capability along a wide dynamic range of metabolite concentration, we determined the relative compound concentration (Rel. C.) which is the quotient between the mean concentration of each compound and the mean concentration of all the compounds listed in the table. Overall, results shown excellent linear correlations (R^2>0.90) for most compounds and methods. Those compounds noted with NF (*Not Found*) have not been found by the algorithm: we considered one compound as NF when it was detected in less than 9 samples. We define the term detected as a compound that has been correctly resolved in a sample and also correctly aligned with the same compounds appearing in the rest of the samples.

From the table, differences between windows size were observed. For example, Valine (1) eluting at 6.11 min, was not found for the 20 seconds window size. This can be attributed at the fact that its low concentration (9 %) affected its detection when more compounds were included in a window. Contrarily, in small windows, the compound had relatively more importance - variance - respect the whole window, which benefited its correct detection in the 10 and 15 s windows cases. The same occurred with isoleucine and proline at 9.06 and 9.10 min, with relative concentrations of 2 and 9 % respectively.

Also, some coefficients of determination varied when comparing area and intensity.

Intensity is expected to be a more robust variable, as the fragments shape may be easily affected by noise or co-elution, whereas in those cases the peak intensity remains more stable. This is because the peak apex is relatively more difficult to be found co-eluted.

From the results, both ICA–OSD and MCR–ALS shown a comparable quantitative performance. Both methods led to similar R^2 values. Differences were found, for example, for the case of Leucine (2) at 8.75 min, where ICA–OSD failed in detecting it whereas MCR–ALS did not, and ICA–OSD successfully quantified glycerol at 8.8 min, whereas MCR–ALS shown a poor performance. Also, the number of samples for where each compound was automatically detected varied between methods (Supplementary Table B.1). For example, hexanoic acid eluting at 5.85 min was detected by ICA–OSD 13 and 18 times for the 10 and 15 s windows cases respectively, whereas MCR–ALS successfully detected it in all the 25 samples. This also explains the fact that glycerol was not correctly detected by MCR–ALS, as it was detected only in 4, 17 and 9 samples for the 10, 15 and 20 seconds windows cases whereas it was detected in almost all the samples by ICA–OSD.

After being resolved, compounds were automatically aligned by the eRah's alignment algorithm. This multivariate algorithm groups the compounds across samples by taking into account the retention time distance and spectral similarity. Thus, a good resolution is important to achieve a good alignment, as automation should not only be limited to deconvolution, but it should also contemplate the automated alignment in order to register the concentrations changes among samples to obtain new biological insights.

Finally, the most significant difference between ICA–OSD and MCR–ALS is their speed of execution. ICA–OSD is a fast resolution method, whereas the implementation based on MCR–ALS resolved an entire sample in approximately 3 min for the 20 s window size, ICA–OSD only took approximately 1.4 min. A fast speed of execution

is and advantageous feature due to the large amount of data that metabolomics experiments generate and also because when a moving window approach is employed, the same data is analyzed twice due to the overlapping, leading to a lower speed of execution in comparison with the traditional hard segmentation approaches. Processing was conducted with a 2.4 GHz Intel Core i7 with 16 GB of DDR3 memory at 1333 MHz.

6.4 Conclusions

Different multivariate methods have been reported in literature for the processing of GC–MS data. However, its application in GC–MS data involve segmenting the chromatogram into regions or windows, which may lead to failure in the detection of compounds. In this study, we proposed the application of moving window-based independent component analysis – orthogonal signal deconvolution (ICA–OSD) and multivariate curve resolution – alternating least squares (MCR–ALS) approaches. We evaluated the proposed methods through their quantification capabilities in comparison with the XCMS package. Results shown that the proposed methodology was able to correctly quantify compounds appearing in biological matrices with the advantage that the automation of the method was not limited only to the resolution, but also the alignment of compounds across samples. Altogether, our results strengthen the suitability of the challenged independent component analysis (ICA) technique for multivariate resolution in analytical chemistry [31], and they demonstrate the robustness of ICA–OSD as a complementary method to MCR–ALS for the automated resolution of GC–MS mixtures in metabolomics experiments.

Bibliography

[1] X. Domingo-Almenara, A. Perera, N. Ramirez, N. Canellas, X. Correig, J. Brezmes. Compound identification in gas chromatography/mass spectrometry-based metabolomics by blind source separation. *J. Chromatogr. A.* 1409 (2015) 226–33.

[2] C. Ruckebusch, L. Blanchet. Multivariate curve resolution: a review of advanced and tailored applications and challenges *Analytical Chimica Acta*, 765 (2013) 28–36,

[3] A. de Juan, J. Jaumot, R. Tauler. Multivariate Curve Resolution (MCR). Solving the mixture analysis problem. *Anal. Methods*, 6:4964, 2014.

[4] Nicolaas (Klaas) M. Faber, Rasmus Bro, and Philip K. Hopke. Recent developments in CANDECOMP/PARAFAC algorithms: a critical review. *Chemometrics and Intelligent Laboratory Systems*, 65(1):119–137, January 2003.

[5] J. M. Amigo, T. Skov, R. Bro, J. Coello, S. Maspoch Solving GC–MS problems with PARAFAC2 *Trends in Anal. Chem.* 27(8) (2008) 714–725.

[6] Guoqing Wang, Wensheng Cai, and Xueguang Shao. A primary study on resolution of overlapping GC-MS signal using mean-field approach independent component analysis. *Chemometrics and Intelligent Laboratory Systems*, 82(1-2):137–144, May 2006.

[7] Zhichao Liu, Wensheng Cai, and Xueguang Shao. Sequential extraction of mass spectra and chromatographic profiles from overlapping gas chromatography-mass spectroscopy signals. *Journal of Chromatography A*, 1190(1-2):358–364, May 2008.

[8] Xueguang Shao, Zhichao Liu, and Wensheng Cai. Resolving multi-component overlapping GC-MS signals by immune algorithms. *TrAC Trends in Analytical Chemistry*, 28(11):1312–1321, December 2009.

[9] X. Domingo-Almenara, A. Perera, N. Ramirez, J. Brezmes. Automated resolution of chromatographic signals by independent component analysis - orthogonal signal deconvolution in comprehensive gas chromatography/mass spectrometry-based metabolomics. *Comput. Methods Programs Biomed* 130 (2016) 135–141.

[10] Jellema, R. H.; Krishnan, S.; Hendriks, M. M. W. B.; Muilwijk, B.; Vogels J. T. W. E. Deconvolution using signal segmentation. *Chemometr. Intell. Lab.* **2010**, 104, 132–139.

[11] Amigo JM, Popielarz MJ, Callejon RM, Morales ML, Troncoso AM, Petersen MA, Toldam-Andersen TB. Comprehensive analysis of chromatographic data by using PARAFAC2 and principal components analysis. *J. Chromatogr. A.* 1217(26) (2010) 4422–4429.

[12] Furbo S, Christensen JH Automated peak extraction and quantification in chromatography with multichannel detectors. *Anal. Chem.* 84(5) (2012) 2211–2218

[13] Lea G. Johnsen, Jose Manuel Amigo, Thomas Skov, Rasmus Bro Automated resolution of overlapping peaks in chromatographic data *Journal of Chemometrics* Volume 28, Issue 2, pages 71–82, February 2014

[14] Jochen Vestner, Gilles de Revel, Sibylle Krieger-Weber, Doris Rauhut, Maret du Toit, Andre de Villiers. Toward automated chromatographic fingerprinting: A

non-alignment approach to gas chromatography mass spectrometry data. *Analytica Chimica Acta* Volume 911, 10 March 2016, Pages 42–58.

[15] Zhong-Da Zeng, Cheng-Jian Xu, Yi-Zeng Liang, Bo-Yan Li. Sectional moving window factor analysis for diagnosing elution chromatographic patterns. *Chemometrics and Intelligent Laboratory Systems* Volume 69, Issues 1–2, 28 November 2003, Pages 89–101

[16] Zhong-Da Zeng, Yi-Zeng Liang, Ya-Li Wang, Xiao-Ru Li, Lu-Ming Liang, Qing-Song Xu, Chen-Xi Zhao, Bo-Yan Li, Foo-Tim Chau. Alternative moving window factor analysis for comparison analysis between complex chromatographic data. *Journal of Chromatography A* 1107 (2006) 273–285.

[17] Da-Lin Yuan, Lun-Zhao Yi, Zhong-Da Zeng and Yi-Zeng Liang. Alternative moving window factor analysis (AMWFA) for resolution of embedded peaks in complex GC–MS dataset of metabonomics/metabolomics study. *Anal. Methods* 2 (2010) 1125–1133.

[18] Sergio Lopez-Urena , Miriam Beneito-Cambra , Rosa M. Donat-Beneito, Guillermo Ramis-Ramos. Overlapped moving windows followed by principal component analysis to extract information from chromatograms and application to classification analysis. *Anal. Methods* 7 (2015) 3080–3088.

[19] S. Oller-Moreno, G. Singla-Buxarrais, J.M. Jimenez-Soto, A. Pardo, R. Garrido-Delgado, L. Arcec, S. Marco. Sliding window multi-curve resolution: Application to gas chromatography–ion mobility spectrometry. *Sensors and Actuators B: Chemical* 217 (2015) 13–21.

[20] Tautenhahn, R.; Bottcher, C.; Neumann, S. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics*. **2008**, 9, 1–16.

[21] Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.* **2006**, 78, 779–787.

[22] Tautenhahn, R.; Patti, G. J.; Rinehart, D.; Siuzdak, G. XCMS Online: a web-based platform to process untargeted metabolomic data. *Anal. Chem.* **2012**, 84, 5035–5039.

[23] Samino, S.; Vinaixa, M.; Díaz, M.; Beltran, A.; Rodríguez, M. A.; Mallol, R.; Heras, M.; Cabre, A.; Garcia, L.; Canela, N.; Zegher, F.; Correig, X.; Ibàñez, L.; Yanes, O. Metabolomics reveals impaired maturation of HDL particles in adolescents with hyperinsulinaemic androgen excess. *Sci. Rep.* **2015**, 5, 11496.

[24] X. Domingo-Almenara, J. Brezmes, M. Vinaixa, S. Samino, M. Diaz, L. Ibanez, X. Correig, A. Perera, O. Yanes. eRah: a computational tool integrating spectral deconvolution and alignment with quantification and identification of metabolites in GC–MS-based metabolomics. *Anal. Chem.* Submitted.

[25] J. Diewok, A. de Juan, M. Maeder, R. Tauler, B. Lendl. Application of a combination of hard and soft modeling for equilibrium systems to the quantitative analysis of pH - modulated mixture samples. *Anal. Chem* 75 (2003) 641–647.

[26] Hummel, J.; Selbig, J.; Walther, D.; Kopka, J. The Golm Metabolome Database: a database for GC–MS based metabolite profiling. *Metabolomics*. **2007**, 18, 75–95.

[27] Jan Hummel, Nadine Strehmel, Joachim Selbig, Dirk Walther, and Joachim Kopka. Decision tree supported substructure prediction of metabolites from GC-MS profiles. *Metabolomics*, 6(2):322–333, June 2010.

[28] J.-F. Cardoso and A. Souloumiac. Blind beamforming for non-Gaussian signals. *Radar and Signal Processing, IEE Proceedings F*, 140(6):362–370, December 1993.

[29] Klaus Nordhausen, Jean-Francois Cardoso, Jari Miettinen, Hannu Oja, Esa Ollila and Sara Taskinen. JADE: Blind Source Separation Methods Based on Joint Diagonalization and Some BSS Performance Criteria *R package* version 1.9-93 (2015). https://CRAN.R-project.org/package=JADE.

[30] Ivo H. M. van Stokkum, Katharine M. Mullen, and Velitchka V. Mihaleva. Global analysis of multiple gas chromatography-mass spectrometry (GC/MS) data sets: A method for resolution of co-eluting components with comparison to MCR-ALS. *Chemometrics and Intelligent Laboratory Systems*, 95(2):150–163, February 2009.

[31] H. Parastar, M. Jalali-Heravi, R. Tauler, Is independent component analysis appropriate for multivariate resolution in analytical chemistry? *Trends Anal. Chem.* 31 (2012) 134–143

Chapter 7

eRah: a computational tool integrating spectral deconvolution and alignment with quantification and identification of metabolites in GC–MS-based metabolomics

Abstract

Gas chromatography coupled to mass spectrometry (GC–MS) has been a long-standing approach used for identifying small molecules due to the highly reproducible ionization process of electron impact ionization (EI). However, the use of GC–EI MS in untargeted metabolomics produces large and complex datasets characterized by co-eluting compounds and extensive fragmentation of molecular ions caused by the hard electron ionization. In order to identify and extract quantitative information of metabolites across multiple biological samples, integrated computational workflows for data processing are needed. Here we introduce eRah, a free computational tool written in the open language R composed of five core functions: (i) noise filtering and baseline removal of GC–MS chromatograms, (ii) an innovative compound deconvolution process using multivariate analysis techniques based on compound match by local covariance (CMLC) and orthogonal signal deconvolution (OSD), (iii) alignment of mass spectra across samples, (iv) missing compound recovery, and (v) identification of metabolites by spectral library matching using publicly available mass spectra. eRah outputs a table with compound names, matching scores and the integrated area of compounds for each sample. The automated capabilities of eRah are demonstrated by the analysis of GC–qTOF MS data from plasma samples of adolescents with hyperinsulinaemic androgen excess and healthy controls. The quantitative results of eRah are compared to centWave, the peak-picking algorithm implemented in the widely used XCMS package, and further validated using pure standards and targeted analysis by GC–QqQ MS, LC–QqQ and NMR. eRah is freely available at http://CRAN.R-project.org/package=erah.

7.1 Introduction

Metabolomics is widely used to obtain new insights into human, plant and microbial biochemistry, as well as in drug discovery, nutrition research and food control. Although different technologies are nowadays used to achieve these objectives [1], the proof of concept for what we now know as mass spectrometry-based metabolomics was reported in 1966 by Dalgliesh et al. [2], which conducted the first GC–MS experiment to separate a wide range of metabolites occurring in urine and tissue extracts. Later in 1971, Horning et al.[3] introduced the term "metabolic profiles", and along with Pauling and Robinson led to the development of GC–MS methods for monitoring metabolites in biological samples through the 1970s [4, 5].

GC–MS has been a long-standing approach used for metabolite profiling of volatile and semi-volatile compounds due to the widespread use of electron impact ionization (EI). EI is a hard ionization technique that has been historically standardized at 70 eV. Unlike soft ionization techniques such as ESI[6] or MALDI[7], EI is a highly reproducible ionization process across many different platforms. However, co-elution of compounds from complex biological samples in GC along with extensive fragmentation of molecular ions by EI ionization, result in large and complex datasets. Reconstructing GC–MS profile data into identified and quantified metabolites across multiple samples remains a challenging task due to the lack of integrated computational tools in GC–MS-based untargeted metabolomics.

Current computational approaches for GC–MS data processing fall into two main categories: tools based on peak-picking, and tools for compound extraction through the so-called curve resolution or spectral deconvolution. The first category involves detecting all relevant fragment ion peaks in the spectra, and align them across multiple samples [8, 9] to subsequently discover statistical peak variations between experimental groups. Representative tools from this category include MZmine [10, 11], MetAlign [12, 13], and XCMS [14, 15]. Although these tools were initially intended

for liquid chromatography mass spectrometry (LC-MS) data processing, they can also be used for GC–MS data analysis [16, 17]. The quantitative variables provided by these methods are not based on the compound spectra, but the m/z value, retention time window and area of fragment ion peaks. Thus, compound identification is the main bottleneck of peak-picking approaches. In this regard, tools such as metaMS [18], TagFinder [19], MetaboliteDetector [20] and PyMS [21] attempt to overcome this limitation by grouping the different peaks (based on their shape similarity or peak correlations) into compound spectra, allowing the putative identification of compounds by comparing their mass spectra with a reference MS library.

The second category focuses on the compound as the analysis entity, as opposed to the use of individual fragment ion peaks. Compounds are quantified and identified on the basis of a multivariate deconvolution process [22] that extracts and constructs pure compound spectra from raw data. Representative tools falling into this category include TNO–DECO [23] or ADAP–GC [24]. TNO–DECO uses multivariate curve resolution to extract the compound spectra, whereas the deconvolution algorithm of ADAP–GC is based on an hierarchical clustering of fragment ions. Other free software, such as AMDIS [25] or BinBase [26, 27] perform parts of the GC–MS metabolomics workflow. AMDIS is used to identify compounds by using the NIST library, but it processes samples independently and it does not include spectral alignment. BinBase uses the spectral deconvolution provided by a proprietary algorithm in the commercial software ChromaTOF (LECO Corporation) in order to align compounds across samples, and it provides compound quantification and identification based on self-constructed libraries [28].

Despite these efforts, there is a need for a free and open source software that integrates all the necessary steps for data processing in GC–MS-based untargeted metabolomics. Here we introduce eRah, an R package with an integrated design that incorporates a novel spectral deconvolution method using multivariate techniques

based on blind source separation (BSS), alignment of spectra across samples, quantification, and automated identification of metabolites by spectral library matching. We demonstrate the functionality of eRah through a comparative analysis of serum samples from adolescents with hyperinsulinaemic androgen excess (HIAE) and healthy controls.

7.2 Experimental Section

7.2.1 Materials

A dataset of 25 serum samples (from 11 young, non-obese adolescents with HIAE and 14 age-, weight- and ethnicity-matched healthy controls) [29] were analyzed by GC-EI-qTOF-MS (Agilent Technologies). Pure standards nicotinic acid, leucine, proline, methionine, aspartic acid, myo-inositol, ornithine, urea and lactic acid were purchased from Sigma Aldrich (Steinheim, Germany). Analytical grade methanol was purchased from SDS (Peypin, France). Water was produced in an in-house Milli-Q purification system (Millipore, Molsheim, France). N-methyl-N-trimethylsilyltrifluoroacetamide, methoxamine hydrochloride and pyridine were purchased from Sigma-Aldrich (Steinheim, Germany). Myristic-d27 acid and succinic acid-2,2,3,3-d4 were from Isotec Stable Isotopes (Miamisburg, USA).

7.2.2 Metabolite extraction method

Serum aliquots (25 μL) were thawed at 4 °C. Samples were briefly vortex-mixed and each aliquot was supplemented with 20 μL of 1 μg/μL succinic-d4 acid (internal standard). Proteins were then precipitated by the addition of 475 μL cold methanol/water (8:1 vol/vol) followed by 3 min of ultrasonication and 10 s of vortex-mixing. Aliquots were subsequently maintained on ice for 10 min. After centrifugation for 10 min (19.000 g, 4 °C), 100 μL of supernatant were transferred to a

GC autosampler vial and lyophilized. We incubated the lyophilized serum residues with 50 μL methoxyamine in pyridine (40 μg/μL) for 30 min at 60 °C. To increase the volatility of the compounds, we silylated the samples using 30 μL N-methyl-N-trimethylsilyltrifluoroacetamide with 1% trimethylchlorosilane (Thermo Fisher Scientific) for 30 min at 60 °C.

7.2.3 GC-qTOF MS analysis

Analysis was carried out on a qTOF MS 7200 (Agilent, Santa Clara, CA, USA) coupled to an Agilent 7890A gas chromatography (GC). Derivatized samples (1 μL each) were injected in the gas chromatograph system with a split inlet equipped with a J&W Scientific DB5–MS+DG stationary phase column (30 mm $\times$ 0.25 mm i.d., 0.1 μm film, Agilent Technologies). Helium was used as a carrier gas at a flow rate of 1 mL/min in constant flow mode. The injector split ratio was adjusted to 1:5 and oven temperature was programmed at 70 °C for 1 min and increased at 10 °C/min to 325 °C. The MS was operated in the electron impact ionization mode at 70 eV. Mass spectral data were acquired in full scan mode from m/z 35 to 700 with an acquisition rate of 5 spectra per second.

7.2.4 GC-QqQ MS analysis

Myo-inositol, ornithine, urea and lactic acid were analyzed using an Agilent 7890A GC coupled to a triple quadrupole (QqQ) MS (7000 Agilent Technologies, Santa Clara, CA, USA) operating in single ion monitoring (SIM) mode and electron impact ionization of 70 eV and a emission intensity of 35 μA. We acquired quantitative and qualitative ions for myo-inositol (318 m/z, 305 m/z and 265 m/z at RT 18.23 min), ornithine (200 m/z, 174 m/z and 142 m/z at RT 15.40 min), urea (189 m/z, 130 m/z and 100 m/z at RT 8.19 min) and lactic acid (191 m/z, 133 m/z and 117 m/z at RT 5.73 min).

7.2.5 Data processing methods

With the aim of comparing the quantitative results of GC–MS serum samples, the data set was processed using eRah and XCMS [14, 15]. GC–MS data files were converted to .mzXML format using Proteowizard software [30]. Converted files were processed using XCMS in order to detect and align features. A feature is defined as an ion entity with a unique m/z and a specific retention time (mzRT). The parameters used in the XCMS workflow were: `xcmsSet (method = 'centWave', ppm = 15, peakwidth = c(1,5)); retcor (method = 'peakgroups', extra = 1, missing=1)` and `group (mzwid = 0.0025, minfrac = 0.5, bw = 5)`. XCMS analysis provided an xcmsSet object containing the retention time, m/z value, and peak intensity (or area) of each feature for every serum sample. Converted files were also processed using eRah through a fast script in R, which includes (i) data preprocessing, (ii) spectral deconvolution, (iii) spectral alignment, (iv) missing compound recovery, and (v) compound identification (see details below). The samples raw-data are classified in folders, where each folder is a class. Signals at m/z 73, 74, 75, 147, 148, and 149 were excluded for data processing, since these are ubiquitous mass fragments typically generated from compounds carrying a trimethylsilyl moiety. These m/z were thus not used for mass spectral matching and metabolite identification. We used the mass range 70-600 m/z (except for the six excluded m/z) for comparison between deconvoluted and reference spectra. The selected/excluded masses can be modified according to the user criterion. The eRah parameters for the deconvolution were: `setDecPar(min.peak.width=1, min.peak.height=2000, noise.threshold=500, avoid.processing.mz= c(73:75,147:149))`, and for the alignment: `setAlPar(min.spectra.cor=0.90, max.time.dist=3, mz.range= 1:600)`. The minimum number of samples was set to 8 for the missing compound recovery step. The complete analysis of 25 samples was performed in less than 30 minutes (in a 2.4 GHz Intel Core i7 computer). The eRah package includes a tutorial

and the description of each function and parameter through the R help.

7.3 Results and discussion

7.3.1 Computational workflow

This section describes the five steps of the eRah workflow (Figure 7-1): (i) data pre-processing, (ii) spectral deconvolution, (iii) spectral alignment, (iv) missing compound recovery, and (v) compound identification. A detailed explanation of eRah methods can be found in the Supplementary Information.

(i) Pre-processing.

GC–MS chromatograms are usually affected by baseline drift and instrumental noise. Smoothing the data by noise filtering and baseline removal improves the eRah's deconvolution and alignment algorithms. Both baseline and noise are filtered according to a minimum compound peak width σ_{MIN}, a value (in seconds) selected by the user. eRah then approximates the baseline drift by a moving-minimum filter [31] to correct the chromatogram, and removes noise using Savitzky-Golay filter [32].

(ii) Deconvolution.

eRah performs a two-step compound deconvolution. First, a multivariate matched filter called compound match by local covariance (CMLC) is applied. The CMLC filter is based on the covariance match filter [33] applied using local covariance matrices [34]. This multivariate approach benefits from the inherent correlation of fragment ions of each compound in EI-MS. CMLC uses covariance matrices to detect patterns of ion redundancy that characterize each compound within the chromatogram. The patterns of ion redundancy approximate to a gaussian peak shape. This matched filter outputs a signal with local minima on spots of ion redundancy in the chromatogram, which are

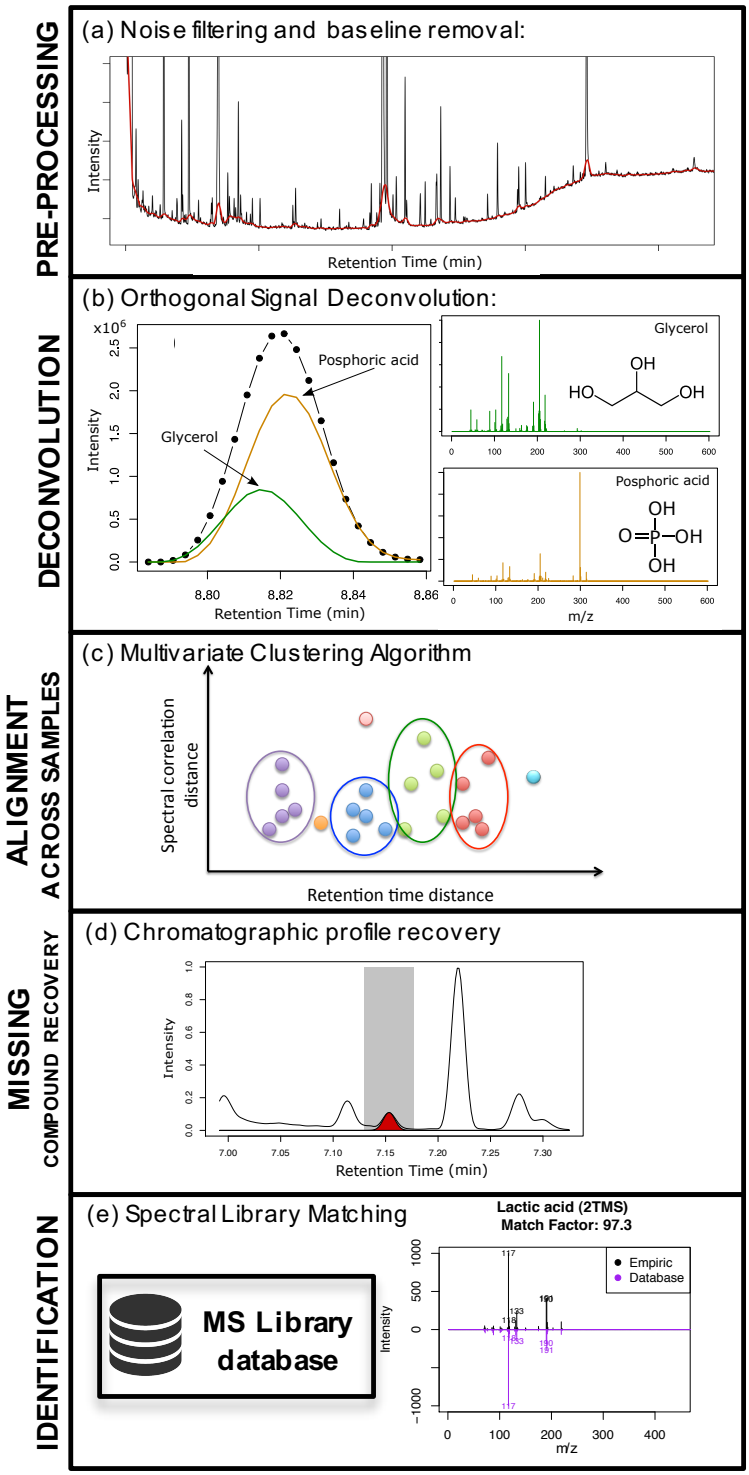

Figure 7-1: eRah's workflow. First, a pre-processing step (a) is applied to remove the noise and the baseline (red) from the chromatogram (black). Second, the deconvolution stage (b) extracts the chromatographic compound profiles and spectra from each sample. Third, compound spectra are aligned (c) across all samples and a missing compounds recovery step (d) retrieves those compounds that were not found in certain samples. Finally, extracted spectra are matched against an MS library (e), providing a list of metabolites and their intensity (or area) in each sample.

determined by compounds with a peak width equal or greater than the selected σ_{MIN} (Figure 7-2 (a)). Upon compound detection by CMLC, the pure compound spectrum is determined using a blind source separation-based algorithm known as orthogonal signal deconvolution (OSD) [35, 36]. OSD is a method able to retrieve a compound spectrum given a compound elution profile. We approximate the elution profile for OSD with the same gaussian model used in CMLC. After the spectrum is determined, we obtain the quantitative compound profile with a least absolute deviation (LAD) regression [37] between the spectrum found by OSD and the chromatogram.

This two-step deconvolution in eRah is depicted in Figure 7-2 using a mixture of five standard compounds, where two different co-elution scenarios are shown. Figures 7-2(b) and 7-2(c) show the eRah's resolved chromatograms for nicotinic acid (I), isoleucine (II) and proline (III) (minutes 5.65–5.74), and methionine (IV) and aspartic acid (V) (minutes 7.13–7.19), respectively. The five compounds were detected using CMLC and their corresponding spectra successfully deconvolved by OSD.

(iii) Alignment.

This step aims to correct the retention time variation of the eluting compounds, facilitating the relative quantification and comparison of compounds across samples. Firstly, the user selects the maximum retention time drift (in seconds) and the minimum spectral similarity (from 0 to 1, being 0 no similarity and 1 the highest similarity) that will be allowed for alignment. This means that two or more compounds with a retention time distance above a maximum retention time drift are not aligned because they are seen as different compounds. The same occurs with the minimum spectral similarity. The alignment is then performed by clustering compounds within these boundaries of retention time distance and spectral similarity (Figure 7-3(a) and Supplementary Information). To determine these clusters eRah computes the Euclidean distance between retention time distance and spectral similarity for all compounds in

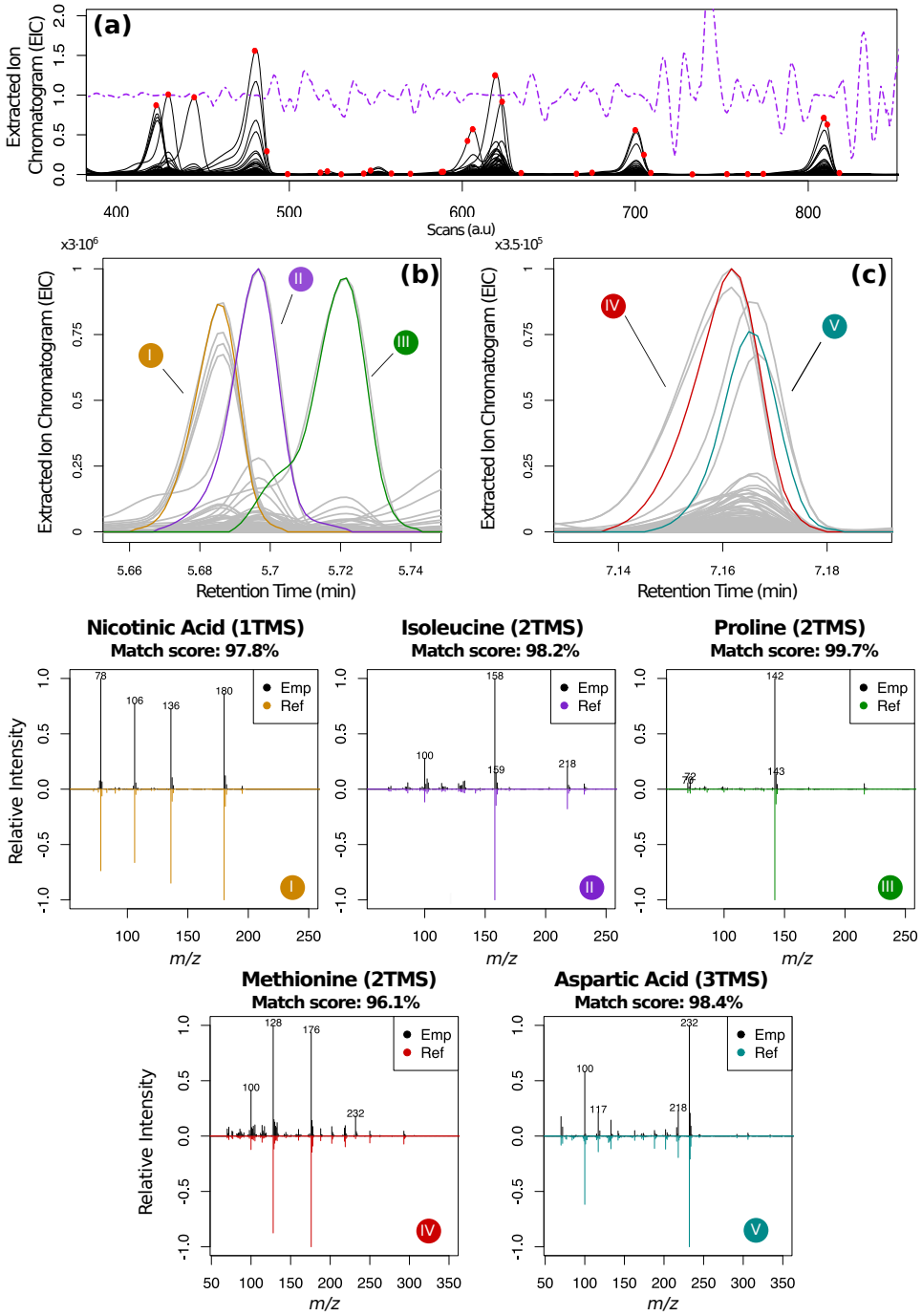

Figure 7-2: Top image (A), shows the operation of the CMLC filter: the black lines depict extracted ion chromatograms (EIC) in the sample, the purple line is the filter output characterized by local minima (marked with red dots in the EIC). Figures B and C show two co-elution situations. The extracted ion chromatograms are shown, where each gray line corresponds to a different m/z peak. Colored solid lines are the deconvolved profiles of the compounds. The deconvolved spectra for each compound are shown in black in figures I-V along with each reference spectrum negatively rotated in the same axis. The match factor is also noted (see details below).

the chromatograms, resulting in compounds appearing across the maximum number of samples and with the least retention time and spectral distance. As an indicative example, Figure 7-3(b) shows the profile of urea before and after the alignment step.

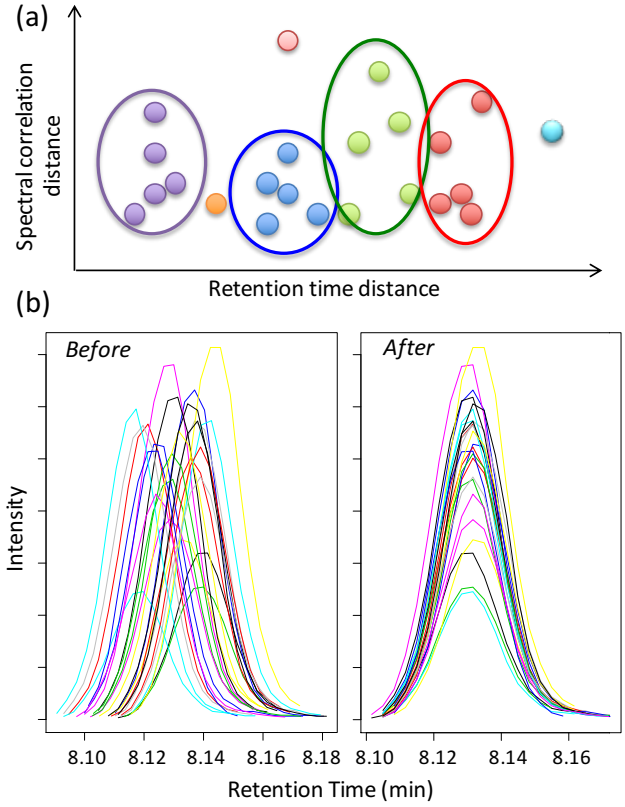

Figure 7-3: (a) Representation of the alignment algorithm. The spheres represent four resolved compounds after deconvolution by eRah. Each compound (purple, blue, green and red spheres) appears in five different samples. We have included three additional compounds as an interference (orange, pink and light blue sphere). The compounds are projected into a two-dimensional space for illustration purposes where their proximity is determined by the spectral similarity and retention time distance. The algorithm aims to cluster the same compound in one group on the basis of proximity in spectra similarity and retention time. (b) Elution profile of urea across samples before and after alignment.

(iv) Missing Compounds Recovery.

Alignment in eRah is a blind step where the algorithm is not forced to find compounds throughout the samples. This means that, in certain circumstances where a strong variation in a compound spectrum occurs, for instance, due to low concentration in a sample (leading noise to disturb the compound spectrum), the alignment step may fail to group the same compound in all samples. To resolve this situation

we have implemented a missing compound recovery step. Similarly to XCMS, the user may impose that compounds appearing in at least e.g., 80% of the samples in an experimental class, may also be found in all other samples. To do this, eRah determines a target spectrum from the mean spectra of each aligned compound. As in the deconvolution step, eRah retrieves the compound chromatographic profile by a LAD regression between the target spectrum and a chromatographic window around the expected elution time in the samples where the compound is missing.

(v) Identification.

Aligned compounds are identified by comparing the mean spectra to reference spectra in a MS library [38]. The current eRah package integrates the free and downloadable version of the MassBank [39] repository containing a set of ∼500 EI GC-TOF mass spectra. However, users may import other libraries such as the Golm Metabolome Database (GMD) [40, 41], Fiehn [28], Human Metabolome Database (HMDB) [42] or an internal database, as long as the library is available in an interpretable format. To use the NIST library, users can export compound spectra from eRah to .MSP format to be read by the MS Search software (NIST) for spectral matching and identification. By comparing the empirical spectra with a reference MS library, eRah generates a list of candidate metabolites along with a similarity match factor, determined using the cosine product [43] (see Supplementary Information for details).

7.3.2 Comparative analysis of serum samples from adolescents with hyperinsulinaemic androgen excess and healthy controls

To illustrate the integrative workflow of eRah, we carried out a comparative metabolomic analysis using 11 serum samples from girls with hyperinsulinemic androgen excess

(HIAE), and 14 age-, weight- and ethnicity-matched healthy controls [29]. HIAE in post-menarcheal adolescent girls is recognized as the phenotypic core of a broader pathological entity traditionally known as polycystic ovary syndrome (PCOS) [44], which affects 8–21% of women of reproductive age worldwide [45, 46]. HIAE usually precedes a broader pathological phenotype in adulthood that is associated with anovulatory infertility, metabolic syndrome, type 2 diabetes [47] and possibly cardiovascular disease. [48]. Therefore, unveiling metabolic derangements in early stages can bring a better understanding of these long-term health risks.

Samples were analyzed using GC-EI-qTOF MS (see Methods section for further details). Raw GC–MS files are available at MetaboLights with accession number MTBLS321. With the aim of comparing the quantitative results of the deconvolved compounds by eRah, mass spectra were also processed using XCMS [14, 15] (Supplementary File 1 and 2). The latter uses centWave, a highly sensitive peak detection algorithm [49]. The output of eRah contained 169 resolved and aligned compounds (Supplementary File 3). We focused, however, on 33 compounds to assess the quantitative accuracy of eRah by comparison with XCMS (Table 7.1). These compounds showed a high similarity match factor (>80.0) to reference MS spectra in the GMD and MassBank. We manually selected a quantitative mzRT feature from the xcmsSet object for each of the 33 compounds. Given the multivariate nature of the spectral deconvolution in eRah, compound quantification is based on the area of the deconvolved compound elution profile and not just a fragment ion peak. Table 7.1 shows the list of 33 compounds with their retention time (RT), identification match factor (MF), and quantitative ion from XCMS. To demonstrate that eRah performs well in a wide dynamic range of metabolite concentrations, we determined the relative compound concentrations (Rel. C.) defined as the quotient between the mean concentration of each compound (i.e., mean area of each deconvolved compound profile) and the mean concentration of all the compounds listed in the table. In addition, the table shows

Table 7.1: Retention time (RT), quantitative fragment ion (m/z) (XCMS), relative concentration (Rel. C) and identification match factor (MF) (eRah) of 33 compounds. The coefficient of determination (R^2) of the regression between the area and intensity of the deconvolved compound elution profile (eRah) and the quantitative ion peak (XCMS) is shown. Percentage of variation between HIAE and control groups is also indicated for both compound (eRah) and peak (XCMS) intensity and area. The percentage was calculated as 100*(mean(HIAE)-mean(CTR)/mean(CTR)).

Cp.	Rt	m/z	Rel.C	Name	MF	R^2		Percentage of variation (%)			
								Area		Intensity	
No.#	(min)		(%)		(%)	Area	Int	eRah	XCMS	eRah	XCMS
1	5.73	117	635	Lactic acid (2TMS)	96.5	1.00	1.00	35	38	37	35
2	5.85	173	5	Hexanoic acid (1TMS)	92.3	0.93	1.00	26	26	25	25
3	6.11	72	38	Valine (1TMS)	96.8	0.97	1.00	18	18	21	21
4	6.53	113	63	Hydroxylamine (3TMS)	96.4	0.86	0.74	-3	-7	-6	-6
5	6.69	131	11	2-hydroxy-butanoic acid (2TMS)	98.2	1.00	1.00	2	4	1	3
6	7.08	86	20	Leucine (1TMS)	99.1	0.98	1.00	23	23	25	25
7	7.15	191	11	3-hydroxy-butanoic acid (2TMS)	90.8	1.00	1.00	-13	-13	-13	-12
8	7.97	145	20	Valine (2TMS)	98.0	0.99	0.97	65	65	66	55
9	8.19	130	948	Urea (2TMS)	91.9	0.93	0.92	13	19	15	17
10	8.36	179	76	Benzoic acid (1TMS)	90.6	0.99	1.00	20	19	20	20
11	8.55	116	14	Serine (2TMS)	95.0	1.00	1.00	48	46	49	48
12	8.75	158	11	Leucine (2TMS)	98.6	1.00	1.00	85	73	77	70
13	8.80	205	122	Glycerol (3TMS)	90.1	1.00	1.00	-8	-10	-7	-8
14	8.81	299	354	Phosphoric acid (3TMS)	95.0	0.99	1.00	25	31	31	31
15	9.06	218	7	Isoleucine (2TMS)	98.5	0.98	0.97	61	64	64	60
16	9.10	142	4	Proline (2TMS)	97.8	1.00	1.00	56	57	66	63
17	9.58	189	3	Glyceric acid (3TMS)	80.5	0.97	0.99	23	21	23	20
18	9.85	215	8	Nonanoic acid (1TMS)	93.2	0.98	1.00	11	11	11	12
19	10.34	291	10	Threonine (3TMS)	85.8	0.99	0.98	28	29	30	30
20	11.82	230	2	3-hydroxy-trans-proline (3TMS)	95.6	0.98	1.00	42	39	47	43
21	12.02	156	125	5-oxoproline 2TMS	99.5	1.00	1.00	35	36	35	34
22	12.73	142	7	Proline [+CO2] (2TMS)	98.4	1.00	1.00	28	29	27	28
23	13.17	246	3	Glutamic acid (3TMS)	92.0	0.99	1.00	118	105	107	103
24	13.27	218	14	Phenylalanine (2TMS)	95.4	1.00	1.00	30	28	27	27
25	13.42	117	59	Dodecanoic acid (1TMS)	92.4	0.91	1.00	5	6	3	3
26	15.40	142	9	Ornithine (4TMS)	98.2	1.00	1.00	64	63	65	64
27	15.46	273	8	Citric acid (4TMS)	96.7	1.00	1.00	25	24	26	25
28	15.54	285	20	Tetradecanoic acid (1TMS)	91.7	0.99	0.94	4	6	2	5
29	16.43	229	15	Lysine (4TMS)	94.3	0.60	0.59	45	30	28	29
30	17.48	129	414	Hexadecanoic acid (1TMS)	91.1	1.00	1.00	7	6	8	6
31	18.23	305	17	Myo-inositol (6TMS)	80.6	0.98	1.00	18	22	16	21
32	18.24	441	6	Uric acid (4TMS)	92.0	0.99	1.00	2	6	4	6
33	19.26	356	224	Octadecanoic acid (1TMS)	89.9	1.00	0.99	6	6	7	9

the coefficient of determination (R^2) of the regression between the mean area - and intensity - of the deconvolved compound profile (eRah) and the quantitative mzRT feature (XCMS). Finally, the percentage of variation between disease (HIAE) and control was also calculated for each compound.

Table 7.2: Percentages of variation for lactate, urea, ornithine and myo-inositol using peak intensity (int) and area determined using eRah, XCMS, MassHunter (MH) and GC-QqQ MS analysis. The percentage was calculated as 100*(mean(HIAE)-mean(CTR)/mean(CTR)).

			QqQ	MH	eRah		XCMS	
Rt	Name	m/z	area	area	area	int	area	int
5.73	Lactic acid	117	32	39	35	38	38	35
8.19	Urea	130	1	13	15	13	19	17
15.40	Ornithine	142	44	64	64	65	63	64
18.23	Myo-inositol	305	11	23	18	16	22	21

The analysis indicated an excellent linear correlation (R^2>0.90) for most compounds. Even for coeluted (e.g., glycerol and phosphoric acid) and low concentration compounds (e.g., myo-inositol and uric acid), the correlation between the area - and intensity - of deconvolved compounds and selective mzRT features was high. Only the area of hydroxylamine and lysine showed R^2<0.90, however these two compounds exhibited similar percentages of variation between HIAE and control groups when compared to XCMS. We also noted that for some compounds the coefficients of determination varied when comparing area and intensity. We attribute these differences to the fact that eRah and XCMS use distinct methodology for quantifying compounds and peaks, respectively, which may lead to some disagreements when comparing areas linearly. Moreover, although XCMS is a very reliable reference, its results should not be taken as ground truth. For this reason, we validated eRah's results using two additional analytical platforms (Table 7.2). Due to availability of pure standards in our laboratory, we focused on lactate, myo-inositol, urea and ornithine for validation experiments. Manual integration of peaks using MassHunter (Agilent Technologies) revealed similar differences, and targeted analysis using GC-triple quadrupole (QqQ) MS (see Methods section for details) reproduced similar variations between HIAE

and control. Altogether, Table 7.2 consistently shows similar quantitative differences using eRah, XCMS, MassHunter and QqQ MS analysis, which further supports the strength of eRah for GC–MS-based untargeted metabolomics studies.

Table 7.3: Percentage of variation and p-values (Wilcoxon–Mann–Whitney test) of statistically significant metabolites. The positive variations indicate higher levels in girls with HIAE relative to healthy controls.

Rt	Name	p-value	%Var
5.73	Lactic acid (2TMS)	0.0090	35
7.08	Leucine (1TMS)	0.0014	23
8.55	Serine (2TMS)	0.0022	48
8.75	Leucine (2TMS)	0.0034	85
11.82	5-oxoproline (2TMS)	0.0042	35
13.17	Glutamic acid (3TMS)	0.0028	118
15.40	Ornithine (4TMS)	0.0002	64
16.43	Lysine (4TMS)	0.0034	45

Consequently, we determined statistical significant differences between HIAE and control using eRah. Lactic acid, leucine, serine, 5-oxoproline (pyroglutamic acid), glutamic acid, ornithine and lysine showed higher levels (p-value$<$0.01) in HIAE relative to control serum samples (Table 7.3). Next, we focused on changes in 5-oxoproline, glutamic acid, lactic acid and leucine to be validated by complementary analyses on the same serum samples using nuclear magnetic resonance (NMR) or liquid chromatography (LC–QqQ MS) as previously described [29] (Figure 7-4). Metabolites in NMR spectra were quantified using Dolphin [50, 51]. Interestingly, analytical platforms such as NMR, which analyze serum non-destructively, and LC-MS, which produces intact molecular ions due to soft ionization, revealed very similar percentages of variation and p-values. Moreover, Zhao et al. [52] observed a positive association of lactate and leucine concentrations with insulin resistance independently of obesity in adult (28-29 years old) PCOS patients. In this previous study, serine levels were increased in PCOS plasma samples as compared with the normal controls independently of obesity or insulin resistance [52]. Our results also revealed elevated levels of lactate, leucine and serine, although in non-obese adolescents with hyperinsulinaemic

androgen excess. Finally, elevated level of ornithine suggests the imbalance of urea cycle in adolescents with HIAE.

Altogether, our results strengthen the feasibility of the recently challenged [53] GC–MS technique for metabolomic applications, and demonstrate the robustness of eRah for data processing.

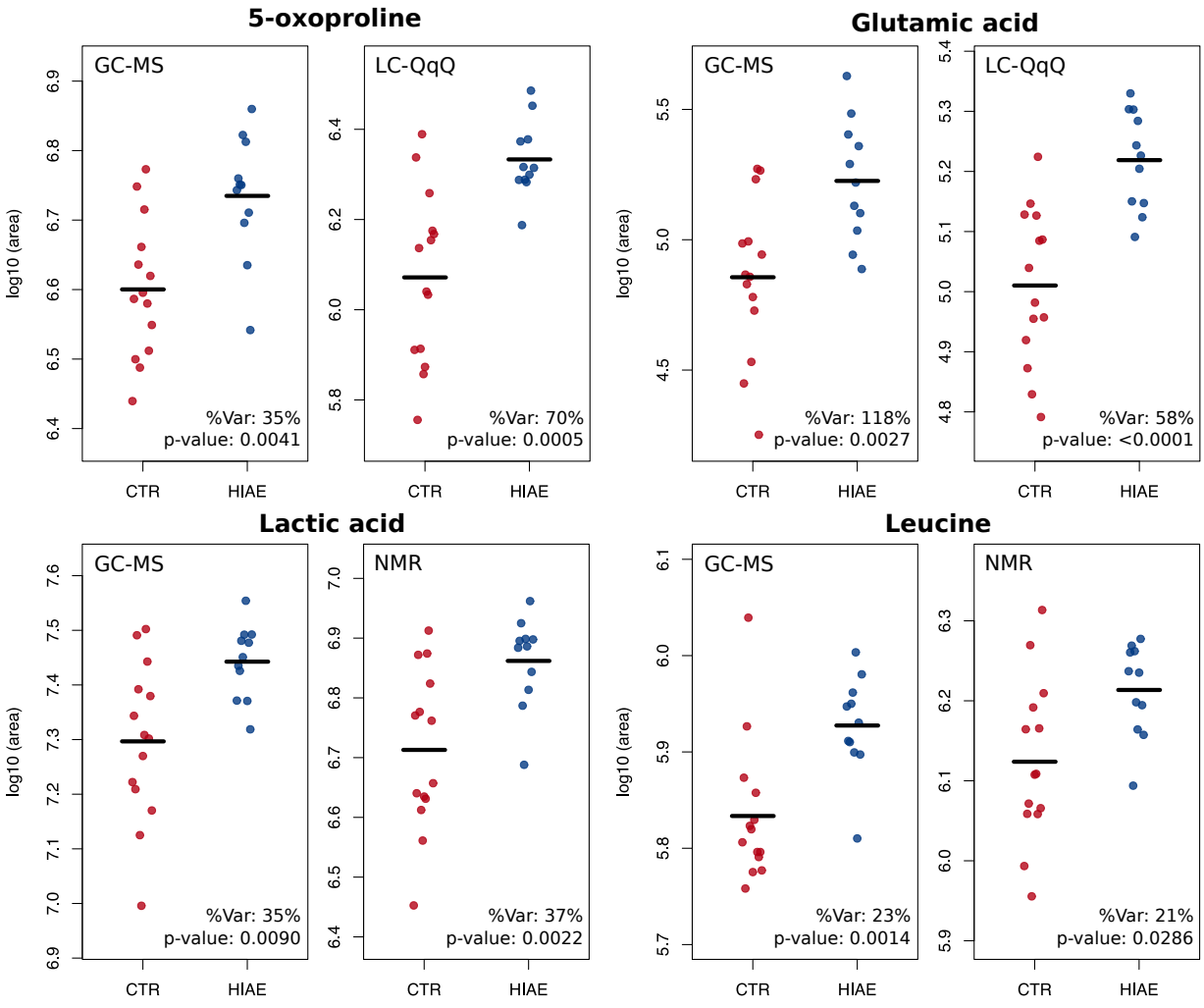

Figure 7-4: Scatter plots of metabolites identified and quantified by GC–MS (eRah), LC–QqQ-MS targeted analysis and NMR. The scatter plots show the abundance of 5-oxoproline, glutamic acid, lactic acid and leucine in controls and HIAE serum samples and trimmed mean (controls are depicted in red and HIAE in blue). Percentage variation (%Var) and p-values (Wilcoxon–Mann–Whitney test) are also shown.

7.4 Conclusions

Despite the existence of different pieces of free and commercial software for GC–MS data analysis, none of these allow the execution of an integrated workflow that includes spectral deconvolution and alignment, followed by the identification and quantification of metabolites in the same application. This still leads many researchers to implement separate software for each process, and tedious manual workflows for data processing. We have developed eRah to fill this gap. eRah is a free computational tool (http://CRAN.R-project.org/package=erah) written in the open language R that enables users to execute a complete automated workflow for data analysis in GC–MS untargeted metabolomics. Moreover, eRah incorporates an innovative deconvolution process based on multivariate compound detection and blind source separation that differs from existing tools. The comparative analysis of serum samples of adolescents with hyperinsulinaemic androgen excess and healthy controls by GC-qTOF MS provided test data demonstrating an excellent correlation with alternative quantitative approaches and analytical platforms such as XCMS, GC-QqQ MS, LC-QqQ MS and NMR, with the manifest advantage that eRah provides a complete automated data analysis workflow. Collectively, we anticipate that eRah will help to expedite and facilitate the analysis of GC–MS data resulting in a greater implementation of this technique in untargeted metabolomic studies.

Bibliography

[1] Patti, G. J.; Yanes, O.; Siuzdak, G. Innovation: Metabolomics: the apogee of the omics trilogy. *Nat. Rev. Mol. Cell Biol.* **2012**, 13, 63–269.

[2] Dalgliesh, C. E.; Horning, E. C.; Horning, M. G.; Knox, K. L.; Yarger K. A gas-liquid-chromatographic procedure for separating a wide range of metabolites occuring in urine or tissue extracts. *Biochem. J.* **1966**, 101, 792–810.

[3] Horning, E. C.; Horning, M. G. Metabolic profiles: gas-phase methods for analysis of metabolites *Clin. Chem.* **1971**, 17, 802–809.

[4] Teranishi, R.; Mon, T. R.; Robinson, A. B.; Cary, P.; Pauling, L. Gas chromatography of volatiles from breath and urine *Anal. Chem.* **1972**, 44, 18–20.

[5] Matsumoto, K. E.; Partridge, D. H.; Robinson, A. B.; Pauling, L.; Flath, R, A.; Mon, T. R.; Teranishi, R. The identification of volatile compounds in human urine. *J. Chromatogr. A.* **1973**, 85, 31–34.

[6] Fenn, J. B. Electrospray wings for molecular elephants (Nobel lecture). *Angew. Chem. Int. Ed. Engl.* **2003**, 42, 3871–3894.

[7] Karas, M.; Ralf, K. Ion formation in MALDI: the cluster ionization mechanism. *Chem. Rev.* **2003**, 103, 427–440.

[8] Koh, Y.; Pasikanti K. K.; Yap, C. W.; Chan, E. C. Comparative evaluation of software for retention time alignment of gas chromatography/time-of-flight mass spectrometry-based metabonomic data. *J. Chromatogr. A.* **2010**, 1217, 8308–8316.

[9] Niu, W.; Knight, E.; Xia, Q.; McGarvey, B.D. Comparative evaluation of eight software programs for alignment of gas chromatography–mass spectrometry chromatograms in metabolomics experiments. *J. Chromatogr. A.* **2014**, 1374, 199–206.

[10] Katajamaa, M.; Jarkko, M.; Matej, O. MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics.* **2006**, 22, 634–636.

[11] Pluskal, T.; Castillo, S.; Villar-Briones, A.; Oresic, M. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics.* **2010**, 11, 395.

[12] Lommen, A. MetAlign: interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing. *Anal. Chem.* **2009**, 81, 3079–3086.

[13] Lommen, A.; Harrie J. K. MetAlign 3.0: performance enhancement by efficient use of advances in computer hardware. *Metabolomics.* **2012**, 8, 719–726.

[14] Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.* **2006**, 78, 779–787.

[15] Tautenhahn, R.; Patti, G. J.; Rinehart, D.; Siuzdak, G. XCMS Online: a web-based platform to process untargeted metabolomic data. *Anal. Chem.* **2012**, 84, 5035–5039.

[16] Aggio, R.; Villas-Boas, S. G.; Ruggiero, K. Metab: an R package for high-throughput analysis of metabolomics data generated by GC–MS. *Bioinformatics.* **2011**, 27, 2316–2318.

[17] Fernandez-Varela, R.; Tomasi, G.; Christensen J. H. An untargeted gas chromatography mass spectrometry metabolomics platform for marine polychaetes. *J. Chromatogr. A.* **2015**, 1384, 133–141.

[18] Wehrens, R.; Georg, W.; Fulvio, M. metaMS: An open-source pipeline for GC–MS-based untargeted metabolomics. *J. Chromatogr. B.* **2014**, 966, 109–116.

[19] Luedemann, A.; Strassburg, K.; Erban, A.; Kopka, J. TagFinder for the quantitative analysis of gas chromatography–mass spectrometry (GC–MS)-based metabolite profiling experiments *Bioinformatics.* **2008**, 24, 732–737.

[20] Hiller, K.; Hangebrauk, J.; Jager, C.; Spura, J.; Schreiber, K.; Schomburg, D. MetaboliteDetector: comprehensive analysis tool for targeted and nontargeted GC–MS based metabolome analysis. *Anal. Chem.* **2009**, 81, 3429–3439.

[21] O'Callaghan, S.; De Souza, D. P.; Isaac, A.; Wang, Q.; Hodkinson, L.; Olshansky, M.; Erwin, T.; Appelbe, B.; Tull, D. L.; Roessner, U.; Bacic, A.; McConville, M. J.; Likic, V. A. PyMS: a Python toolkit for processing of gas chromatography-mass spectrometry (GC–MS) data. Application and comparative study of selected tools. *BMC bioinformatics.* **2012**, 13, 115.

[22] Du, X.; Steven H. Z. Spectral deconvolution for gas chromatography mass spectrometry-based metabolomics: current status and future perspectives. *Comput. Struct. Biotechnol. J.* **2013**, 4, 1–10.

[23] Jellema, R. H.; Krishnan, S.; Hendriks, M. M. W. B.; Muilwijk, B.; Vogels J. T. W. E. Deconvolution using signal segmentation. *Chemometr. Intell. Lab.* **2010**, 104, 132–139.

[24] Ni, Y.; Qiu, Y.; Jiang, W.; Suttlemyre, K.; Su, M.; Zhang, W.; Jia, W.; Du, X. ADAP-GC 2.0: Deconvolution of coeluting metabolites from GC/TOF-MS data for metabolomics studies. *Anal. Chem.* **2012**, 84, 6619–6629.

[25] Stein, S. An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. *J. Am. Soc. Mass Spectrom.* **1999**, 10, 770–781.

[26] Fiehn,O., Wohlgemuth,G., Scholz,M. Automatic annotation of metabolomic mass spectra by integrating experimental metadata. *Proc. Lect. Notes Bioinformatics.* **2005**, 3615, 224–239.

[27] Skogerson, K.; Wohlgemuth, G.; Barupal, D. K.; Fiehn, O. The volatile compound BinBase mass spectral database. *BMC Bioinformatics.* **2011**, 12, 321.

[28] Kind, T.; Wohlgemuth, G.; Lee do, Y.; Lu, Y.; Palazoglu, M.; Shahbaz, S.; Fiehn, O. FiehnLib: mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry. *Anal. Chem.* **2009**, 81, 10038–10048.

[29] Samino, S.; Vinaixa, M.; Díaz, M.; Beltran, A.; Rodríguez, M. A.; Mallol, R.; Heras, M.; Cabre, A.; Garcia, L.; Canela, N.; Zegher, F.; Correig, X.; Ibàñez, L.; Yanes, O. Metabolomics reveals impaired maturation of HDL particles in adolescents with hyperinsulinaemic androgen excess. *Sci. Rep.* **2015**, 5, 11496.

[30] Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics.* **2008**, 24, 2534–2536.

[31] Pitas, J. Fast Algorithms for Running Ordering and Max/Min Calculation. *IEEE Trans. Circuits Syst.* **1989**, 36, 795–804.

[32] Savitzky, A.; Golay, M. J. E. Smoothing and differentiation of data by simplified least squares procedures *Anal. Chem.* **1964**, 36, 1627–1639.

[33] Chang C. I. Hyperspectral Imaging *Springer Science+Bussiness Media New York*. **2003**.

[34] Caefer C. E.; Rotman S. R. Local covariance matrices for improved target detection performance. *First Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*. **2009**, 1–4.

[35] Domingo-Almenara, X.; Perera, A.; Ramirez, N.; Canellas, N.; Correig, X.; Brezmes, J. Compound identification in gas chromatography/mass spectrometry-based metabolomics by blind source separation. *J. Chromatogr. A.* **2015**, 1409, 226–33.

[36] Domingo-Almenara, X.; Perera, A.; Ramirez, N.; Brezmes, J. Automated resolution of chromatographic signals by independent component analysis - orthogonal signal deconvolution in comprehensive gas chromatography/mass spectrometry-based metabolomics. *Comput. Methods Programs Biomed.* **2016**, 130, 135–141.

[37] Li, Y.; Gonzalo, A. R. A maximum likelihood approach to least absolute deviation regression. *EURASIP J. Adv. Signal Process.* **2004**, 12, 1762-1769.

[38] Vinaixa, M.; Schymanski, E .L.; Neumann, S.; Navarro, M.; Salek, R. M.; Yanes, O. Mass spectral databases for LC/MS and GC/MS-based metabolomics: state of the field and future prospects. *Trends Analyt. Chem.* **2016**, 78, 23–35.

[39] Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K.; Oda, Y.; Kakazu, Y.; Kusano, M.; Tohge, T.; Matsuda, F.; Sawada, Y.; Hirai, MY.; Nakanishi, H.; Ikeda, K.; Akimoto, N.; Maoka, T.; Takahashi, H.; Ara, T.; Sakurai, N.; Suzuki, H.; Shibata, D.; Neumann,

S.; Iida, T.; Tanaka, K.; Funatsu, K.; Matsuura, F.; Soga, T.; Taguchi, R.; Saito, K.; Nishioka, T. MassBank: A public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* **2010**, 45, 703–714.

[40] Hummel, J.; Strehmel, N.; Selbig, J.; Walther, D.; Kopka, J. Decision tree supported substructure prediction of metabolites from GC-MS profiles. *Metabolomics.* **2010**, 6, 322–333.

[41] Hummel, J.; Selbig, J.; Walther, D.; Kopka, J. The Golm Metabolome Database: a database for GC–MS based metabolite profiling. *Metabolomics.* **2007**, 18, 75–95.

[42] Wishart, D. S.; Jewison, T.; Guo, A. C.; Wilson, M.; Knox, C.; Liu, Y.; Djoumbou, Y.; Mandal, R.; Aziat, F.; Dong, E.; Bouatra, S.; Sinelnikov, I.; Arndt, D.; Xia, J.; Liu, P.; Yallou, F.; Bjorndahl, T.; Perez-Pineiro, R.; Eisner, R.; Allen, F.; Neveu, V.; Greiner, R.; Scalbert, A. HMDB 3.0-The Human Metabolome Database in 2013 *Nucleic Acids Res.* **2013**, 41, 801–807.

[43] Stein, S. E.; Donald R. S. Optimization and testing of mass spectral library search algorithms for compound identification. *J. Am. Soc. Mass Spectrom.* **1994**, 5, 859–866.

[44] Ibáñez, L.; Ong, K. K.; López-Bermejo, A.; Dunger, D. B.; de Zegher, F. Hyperinsulinaemic androgen excess in adolescent girls. *Nat. Rev. Endocrinol.* **2014**, 10, 499–508.

[45] Franks, S. Polycystic ovary syndrome. *N. Engl. J. Med.* **1995**, 333, 853–861.

[46] March, W. A.; Moore, V. M.; Willson, K. J.; Phillips, D. I.; Norman, R. J.; Davies, M. J. The prevalence of polycystic ovary syndrome in a community sample assessed under contrasting diagnostic criteria. *Hum. Reprod.* **2010**, 25, 544–551.

[47] Gambineri, A.; Patton, L.; Altieri, P.; Pagotto, U.; Pizzi, C.; Manzoli, L.; Pasquali, R. Polycystic Ovary Syndrome Is a Risk Factor for Type 2 Diabetes Results From a Long-Term Prospective Study. *Diabetes*. **2012**, 61, 2369–2374.

[48] Talbott, E. O.; Guzick, D. S.; Sutton-Tyrrell, K.; McHugh-Pemu, K. P.; Zborowski, J. V.; Remsberg, K. E.; Kuller, L. H. Evidence for association between polycystic ovary syndrome and premature carotid atherosclerosis in middle-aged women. *Arterioscler. Thromb. Vasc. Biol.* **2011**, 20, 2414–2421.

[49] Tautenhahn, R.; Bottcher, C.; Neumann, S. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics*. **2008**, 9, 1–16.

[50] Gómez, J.; Brezmes, J.; Rodriguez, M. A; Vinaixa, M.; Salek, R. M.; Correig, X.; Canellas, N. Dolphin: a tool for automatic targeted metabolite profiling using 1D and 2D (1)H-NMR data. *Anal. Bioanal. Chem.* **2014**, 406, 7967–7976.

[51] Gómez, J.; Vinaixa, M.; Rodriguez , M. A; Salek, R. M.; Correig, X.; Canellas, N. Dolphin 1D: Improving automation of targeted metabolomics in multi-matrix datasets of 1H-NMR spectra. *Advances in Intelligent Systems and Computing. 9th International Conference on Practical Applications of Computational Biology and Bioinformatics*. **2015**, 375, 59–67.

[52] Zhao Y.; Li Fu, L.; Li R.;, Wang L.; Yang Y.; Liu N.; Zhang C.; Wang Y.; Liu P.; Tu B.; Zhang X.; Qiao J. Metabolic profiles characterizing different phenotypes of polycystic ovary syndrome: plasma metabolomics analysis. *BMC Medicine*. **2012**, 10, 1–12.

[53] Fang, M.; Ivanisevic, J.; Benton, H. P.; Johnson, C. H.; Patti, G. J.; Hoang, L. T.; Uritboonthai, W.; Kurczy, M. E.; Siuzdak, G. Thermal Degradation of Small Molecules: A Global Metabolomic Investigation. *Anal. Chem.* **2015**, 87, 10935–10941.

Chapter 8

Targeting the untargeted: BaitMet, an R package for GC–MS library-driven compound profiling in metabolomics

Abstract

Targeted metabolomics aims to answer biological hypotheses by the quantification of known (target) metabolites. Target analysis is generally achieved by limiting the mass detector to analyze (monitor) only either those ions or transitions which are sufficiently selective (unique) for a given set of compounds at a certain retention time. This process clearly limits the coverage of the target analysis. An alternative to this is to combine the strengths of the targeted and the untargeted types of analysis: targeting the untargeted. This means using a previous knowledge about the metabolome (MS libraries) and target this knowledge in chromatograms acquired in full scan mode (untargeted). This study introduces BaiTMet, and R package for high-throughput quantification of compounds of an entire MS library into GC–MS data using blind source separation methods. In GC–MS, internal standards are mixed with the samples to standardize the retention times into retention indexes (RI), and thus increase the identification confidence. Mixing internal standards also increases the complexity to the already complex GC–MS samples and increases the wet laboratory sample preparation time. BaitMet was applied for the automated profiling of serum samples of patients with chronic kidney disease. BaitMet was able to identify target compounds in chromatograms acquired in full scan mode given a reference library and to automatically standardize the retention time without the use of internal standards.

8.1 Introduction

Metabolomics, which is the profiling of metabolites in biofluids, cells and tissues, is routinely applied as a tool for biomarker discovery [1]. In this context, gas chromatography – mass spectrometry (GC–MS) has been a long–standing analytical platform for the quantification of volatile and semivolatile metabolites due to the reproducibility of electron impact ionization and retention time robustness of capillary compounds.

Metabolomics experiments are branched into untargeted - measuring as many metabolites as possible - and targeted - measuring a set of known metabolites -. While a great attention has been laid on untargeted analysis, the quantification of a set of known (target) metabolites - typically focusing on certain related pathways of interest -, also has merits in addressing hypothesis-driven biological questions [2]. Depending on the configuration of GC–MS instrument, targeted analysis can be conducted using SIM (selective ion monitoring) in the case of GC–MS (single quadrupole) and GC–TOF instruments; either MRM (multiple reaction monitoring) or SRM (single reaction monitoring) for GC–QqQ and GC–Orbitrap MS/MS instruments and full spectrum of Product Ions (SIM–TOF) in the case of GC–qTOF. In all cases limiting the mass detector limits the metabolite coverage of the target analysis as it is constricted to analyze (monitor) only either those ions or transitions which are sufficiently selective (unique) for a given set of compounds at a certain retention time. Higher sensitivity and usually wider dynamic ranges are achieved through this approach as compared to full scan data acquisition modes commonly employed in untargeted analysis.

An alternative to this is to combine the strengths of the targeted and the untargeted types of analysis: targeting the untargeted. This means using a previous knowledge about the metabolome and target this knowledge in chromatograms acquired in full scan mode (untargeted). Due to the rising interest in metabolomics, specific MS libraries have been developed [3], including NIST, Golm Metabolome Database (GMD)

[6], Human Metabolme Database (HMD) [4] or MassBank [5]. Those libraries provide the knowledge needed for a library-driven compound profiling: metabolite MS spectra and retention time; the last can be predicted using retention indexes by internal standards.

Some tools have already combined the strengths of the targeted and the untargeted types of analysis, including TargetSearch [7] and targeted mass spectral ratio analysis (TMSRA) [8]. Those tools are able to quantify known metabolites, but in chromatograms acquired in full scan mode. TMSRA requires to input the expected retention time of the target compounds, and their mass spectra. Then, it attempts to quantify the 'target' compounds by automatically determine the selective masses of each target analyte. TargetSearch is an R package for library-driven compound profiling that relies on the retention indexes (RI) and the list of selective masses provided by a reference MS library to quantify a high number target compounds with univariate techniques. It requires mixing internal standards with the samples. The use of internal standards - typically n-alkanes (ALK) and n-alkyl fatty acid methyl esters (FAME) - may chromatographically mask other compounds, besides of increasing the wet laboratory sample preparation time.

This paper introduces BaiTMet, an R package that allows a high-throughput search of an entire MS library into full-scan GC–MS data, using the library as a bait (Bait), to quantify metabolites (Met) and thus performing a library-driven compound profiling. BaitMet can quantify compounds with (i) multivariate methods and without any prior information about the selective masses or (ii) by the integration of a previously defined selective mass for each compound. Also, internal standards (IS) for RT standardization are not needed in each sample, instead, only a single separated analysis of the IS is needed. BaitMet automatically determines which is the instrumental retention time variation of each samples with respect to a fixed retention index/retention time curve. BaiTMet outputs a table with compounds name,

spectral matching score, RI error, and the peak intensity or integrated area - relative concentration - of the compound in each sample.

8.2 Methods

The automated capabilities of BaitMet were demonstrated by the analysis of GC–MS data from a total of 182 human serum samples from age and weight matched volunteers with chronic kidney disease. Concretely, we used 60 different biological samples in 3 replicates plus 2 serum as quality control. Each sample were mixed with 9 FAMEs (C10-25).

For this set of samples, BaitMet was used to (i) automatically find the internal standards (FAME) mixed in the samples. Also, (ii) BaitMet was evaluated by its capability to infer the chromatographic RI/RT curve variation without any information about the internal standards (as if the IS were not mixed with the samples). Additionally, the quantitative results of BaiTMet were compared with a reference concentration by a selective (quantitative) molecular ion for a set of 33 identified metabolites.

8.2.1 Metabolite extraction method

Serum samples were thawed on ice at 4 °C for 30-60 min. Each aliquot of serum (70 μL) was spiked with 300 μL of acetonitrile isopropanol water (3:3:2) (deproteinization), added 5 μL of internal standard solution (myristic acid D27 - 3 mg/mL). After vortex for 15 sec, the mixture was centrifuged for 15 min at 15,800xg at 4 °C. Supernatant (320 μL) was transferred to a new microcentrifuge tube, followed by lyophilization in a speedvac concentrator. Subsequently, 5 μL of FAME (C10-25) was added in the residue and was resuspended in 50 μL of methoxyamine in pyridine (Sigma-Aldrich) solution (40 mg/mL) and vortexed for 3 min. This methoximation

reaction was performed at room temperature for 16h, followed by trimethylsilylation for 1 h adding 100 μL MSTFA (N-methyl-Ntrimethylsilyltrifluoroacetamine) with 1% TMCS (trimethylchlorosilane) (Sigma-Aldrich). After derivatization, 1 μL of this derivative was used for Gas Chromatography Mass Spectrometry (GC/MS) analysis.

8.2.2 GC–MS analysis

Extracts were analyzed by a 7890B gas chromatograph from Agilent (Palo Alto, CA. USA) coupled to an Agilent 5977A mass selective detector, using a DB5–MS Duragard capillary column (10 m) from Agilent (Agilent 122-5532G). Analyses were performed by injecting 1 μL of the extracts into a splitless inlet at 250 °C and a helium constant flow of 1.1 mL min^{-1}. The oven temperature of the GC was initially held at 60 °C for 1 min, then raised to 310 °C at a rate of 10 °C min^{-1}. The mass spectrometer operated in the electron impact ionization mode (70 eV) and mass spectra were recorded after a solvent delay of 6.5 min with 3 scans per second in full scan mode (from 50 to 500 Da). The MS quadrupole temperature was set at 180 °C and the ion source temperature was set at 280 °C.

8.2.3 Data analysis

BaiTMet was provided with a reference library and a fixed RI/RT curve determined by the mean RT of each FAME. The library used was a subset of the downloadable version of the Golm Metabolome Database (GMD) [6] (Version at 2011-11-21) containing only those compounds with KEGG number (a total of 1152). FAMEs were not included in the library. The RI error was set to 0.5% according to thresholds proposed by [11] and the FWHM$_{min}$ was of 2 seconds. Thus, BaitMet parameters as in code were as follows:

```
> setBaitPar(ri.error=0.05, min.peak.width=2, avoid.processing.mz=
c(35:69,73:75,147:149), min.peak.height=1000, noise.threshold=100)
```

8.2.4 Computational workflow

Compound deconvolution

First, the compound chromatographic location where each compound may be eluting is determined. BaiMet compound location is based upon retention time indexes (RI) tabulated in the library and a series of reference standards (either FAME or ALK, commonly employed in GC–MS). These reference standards should be measured once in a single sample using the same chromatographic method that will be further used to measure samples. A RT/RI curve is built by a linear interpolation between - and linear extrapolation outside - the reference standards retention times. Then, an Expected Elution Window (EEW) for each compound in the library can be estimated as $RI_j - RI_j * ri_e$ to $RI_j + RI_j * ri_e$, where the subindex j denotes each metabolite RI, and ri_e is the expected error for RT variation respective to reference RI (Figure 8-1 (a)). The spectra of each compound in the library is correlated against each spectrum recorded within EEW in the real sample. The RT that maximizes this correlation is retained as the specific compound retention time in the real sample. A more accurate compound elution window ROI (Region of Interest) (Figure 8-1 (b)) is defined at this specific retention with boundaries calculated as two times the minimum compound full width at half maximum (FWHM_{min}), an user-value in seconds, before and after the compound retention time.

Once the ROI is determined, BaitMet aims to recover the compound chromatographic profile. An intuitive way to recover a compound chromatographic profile given its spectrum is by a least squares regression of the spectrum against the chromatogram. This process though, can be substituted by a L_1 estimation, also known as least absolute deviation (LAD). Whereas the least squares regression minimizes the squared residuals between the data and the regressor, the L_1 estimation minimizes

the sums of absolute residuals. L_1 estimation is mathematically defined as:

$$min_\alpha \ f(\alpha) = |\alpha \mathbf{s_r} - \mathbf{D_j}| \ \ ; \quad j = 1, 2, ..., N \tag{8.1}$$

where D $(N \times M)$ is the chromatographic ROI and s_r $(M \times 1)$ is the reference spectrum from the MS library. In this notation, N is the number of chromatographic scans (retention time) and M is the range of acquisition of the mass-charge ratio (m/z). The L_1 estimation is more robust to outliers than least squares as it weights all the observations equally. Physically, this means that all the ions of the spectrum that are regressed against the chromatogram have the same importance. This is a natural way to give the same importance also to the selective m/z - which best describe the pure chromatographic profile - and that tend to be less abundant.

Next, although we initially have the reference spectrum of the target compound, we extract the empirical spectrum for each chromatographic profile deconvolved. The empirical spectra will be subsequently compared with the reference spectra by spectral matching allowing users to evaluate wether or not the compound is in the sample. BaitMet uses orthogonal signal deconvolution (OSD) [9, 10] for spectral deconvolution. OSD is a method, based on blind source separation, to extract the spectra from data given the compound profile, in an accurate and fast manner.

RT/RI elastic curve correction

The hypothesis behind BaiTMet is that, when using internal standards (IS) to characterize the relative retention indexes (RI) along the chromatographic retention time (RT), the curve RI/RT is the same for all the samples under the same chromatographic method. This RI/RT though may suffer of an elastic variation due to instrumental error, i.e., the RI/RT curve for each sample can be approximated given a fixed RI/RT curve (representative of the chromatographic method) plus an elastic variation of the

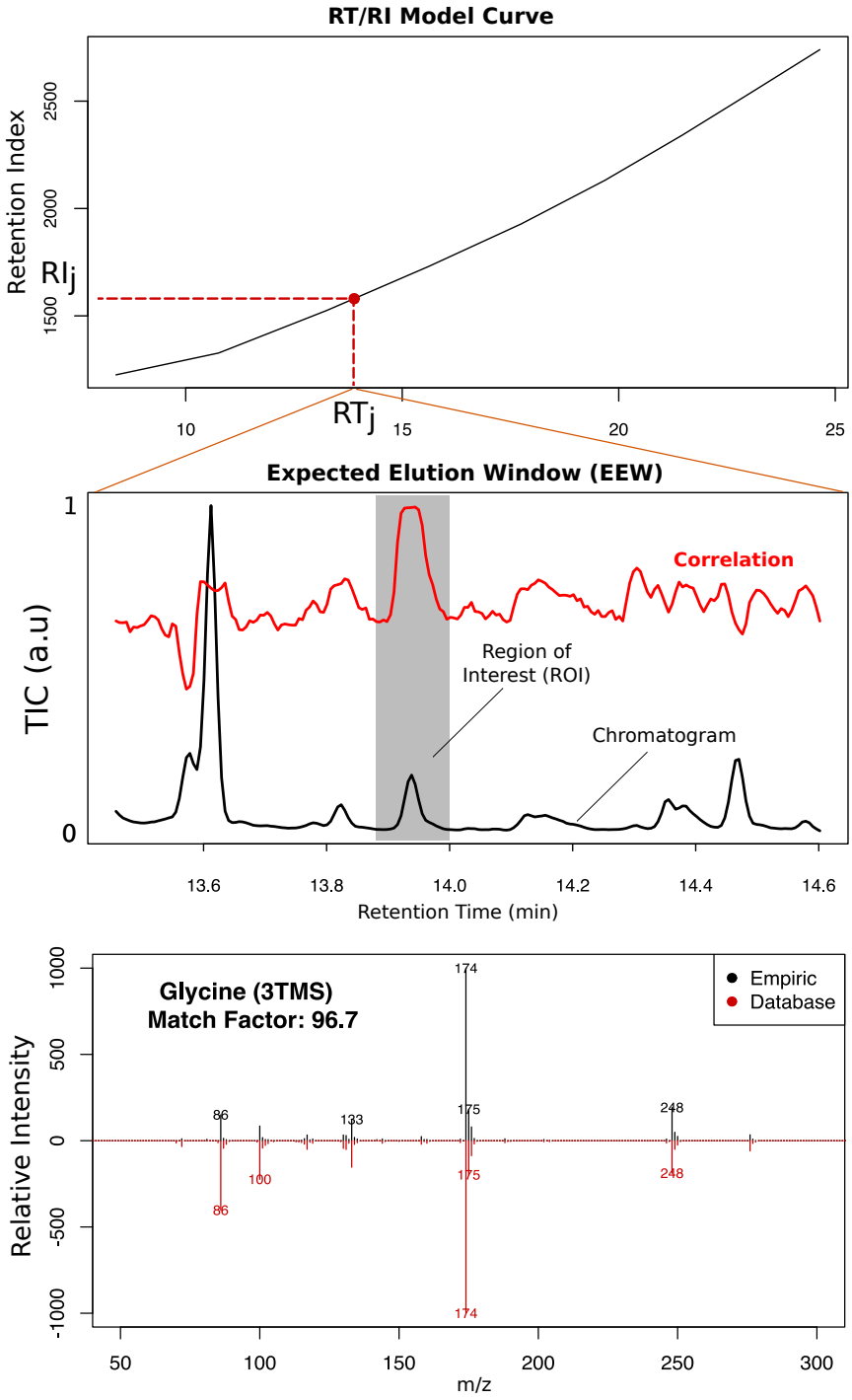

Figure 8-1: Illustration of BaitMet deconvolution stage: first (a), the approximated retention time of the compound j is approximated by projecting its retention index into the fixed RT/RI curve. Next, (b) the target spectrum is correlated against a wide expected elution window (EEW), where a region of interest (ROI) is later determined around the RT that maximizes this correlation. Finally, (c) the compound empirical spectrum is extracted for its further comparison with the reference.

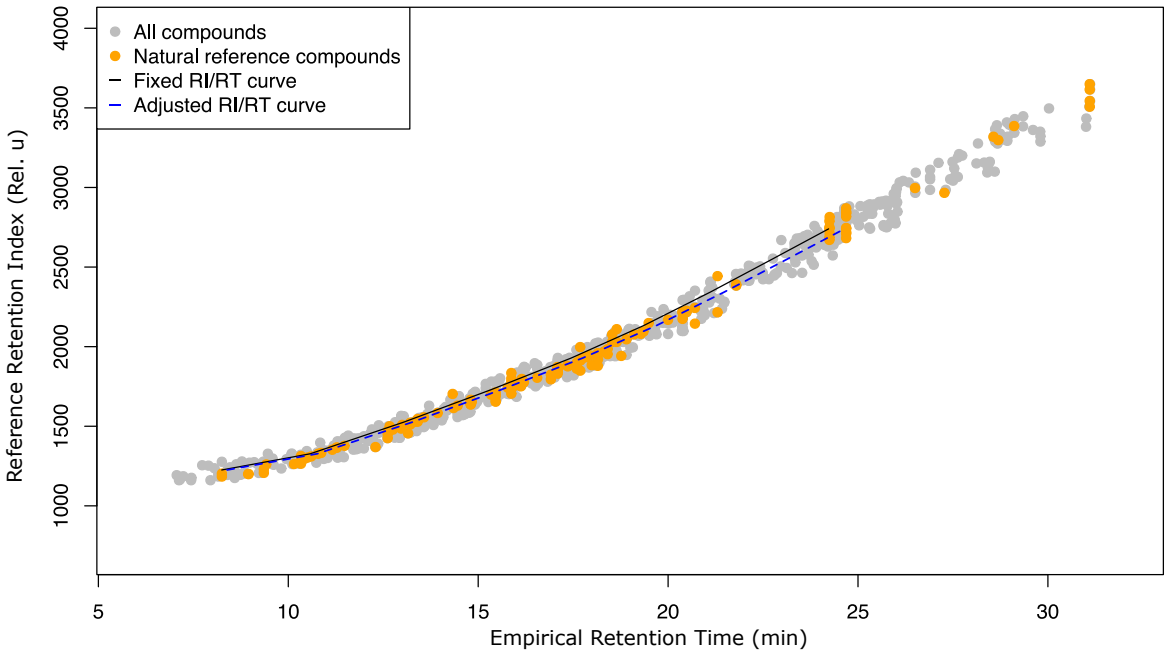

Figure 8-2: This figure shows how the original RI/RT curve (black) is elastically modified into the new curve (blue). The grey points are all the compounds detected by BaiTMet, in orange, are all the compounds that are taken into account to infer the elastic variation.

form:

$$x_n = x\beta + y0 \tag{8.2}$$

where x_n is the vector that defines the particular RI curve for each retention time and for each sample n, x is the fixed RI/RT curve, and (β) and $y0$ are the scalars - to be inferred - that define the elastic variation and the offset value respectively. BaiTMet automatically determines which is the variation of the curve in each chromatogram by the following steps: those compounds with a match factor above a user-defined threshold of 90 % are considered tentatively correctly identified. Then, an optimization procedure determines which elastic variation minimizes the median absolute error between the reference RI - provided by the library - and the predicted RI by a modified RI/RT curve (Figure 8-2). The compounds empirical RI is deter-

mined following this new inferred curve.

8.3 Results

We analyzed a total of 182 human serum samples from age and weight matched subjects with chronic kidney disease. Each sample was mixed with 9 FAMEs (C10-25). BaiTMet detected a total of 95 compounds with less than 0.1 % of error and with a spectra similarity score above 90 %. Of all the compounds detected, we focused on a subset of 33 compounds. The list of the metabolites along with the the spectral similarity match factor is shown in Table 8.1. The empirical RI for each compound found was automatically determined by the BaiTMet algorithm without FAMEs. For evaluation purposes, we additionally determined the RI with the van den Dool and Kratz algorithm [12] using the FAMEs retention time. The area of a selective (quantitative) molecular ion from each metabolite was used as quantitative reference. Table 8.1 shows the list of the 33 compounds with their respective match factor and RI error provided by the BaitMet algorithm and by the use of FAME. Table 8.1 also shows the coefficient of determination between the quantification by BaiTMet and the reference concentration. From the table, almost all the compounds exhibited an excellent linear relation (R^2>0.95).

Figure 8-3 shows the overall RI error barplot of the 33 compounds by both BaiTMet and FAMEs. Statistical significative differences using a paired t-test were observed - FAME mean RI error was less than BatiMet - (p-value <0.0001). The absolute mean difference between both methods is of 0.01 %, which is significantly less than the typical identification RI error (0.5 - 1 %). This proves the capability of BaiTMet to automatically standardize the RT into RI. Both methods yield to the identification of the studied compounds with mean relative retention index errors below 0.5 %.

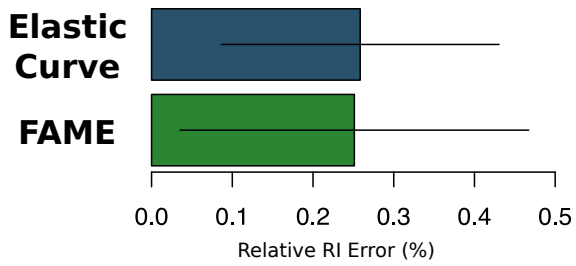

Figure 8-3: Overall RI error barplot of the 34 identified compounds by both BaiTMet (elastic curve) and FAMEs. Outliers (those with values above 2σ) were removed. The sample size was of N=6188 (34 compounds in 182 samples).

8.4 Conclusion

Currently, target analysis is achieved by limiting the mass detector to analyze those ions or transitions which are sufficiently selective for a given set of compounds at a certain retention time. This process clearly limits the coverage of the target analysis. This study introduces BaitMet, an R package for high-throughput quantification of compounds of an entire MS library into full scan GC–MS chromatograms. BaitMet combines thus the strengths of the targeted and the untargeted types of analysis. BaiTMet is also able to identify compounds by the standardization of retention time without mixing internal standards in the samples. Alternatively, BaiTMet is also compatible with the use of internal standards mixed in the samples, and using them to characterize the RI/RT curve in each chromatogram.

Table 8.1: Adjusted coefficients of determination (R^2) for the regression between the quantification by BaiTMet and the reference concentration by the selective/quantitative ion for each metabolite. The table also lists the quantitative m/z, the spectral similarity match factor (MF), the retention index error (RI_e) by the elastic curve modification inferred by BaitMet (BM) and by using internal standards (IS), and the relative concentration (RC).

						RI_e		
No.	m/z	RT	Name	MF	R^2	BM	IS	RC
1	247	10.57	Succinic acid (2TMS)	91.0	0.93	0.44	0.45	5
2	205	11.14	Serine (3TMS)	96.1	0.94	0.09	0.05	40
3	129	11.19	Nonanoic acid (1TMS)	92.5	0.74	0.45	0.31	7
4	101	11.46	Threonine (3TMS)	98.3	0.99	0.25	0.09	48
5	261	11.75	Glutaric acid (2TMS)	90.4	0.81	0.40	0.36	3
6	158	12.61	Proline, 4-hydroxy-, trans- (2TMS)	97.8	1.00	0.38	0.20	8
7	233	12.75	Malic acid (3TMS)	97.8	0.99	0.15	0.05	4
8	157	13.01	Pyroglutamic acid (1TMS)	97.5	0.92	0.58	0.36	15
9	232	13.15	Aspartic acid (3TMS)	95.1	0.96	0.03	0.25	10
10	156	13.21	Pyroglutamic acid (2TMS)	99.9	1.00	0.36	0.15	194
11	84	13.31	Glutamic acid (2TMS)	99.0	1.00	0.21	0.01	222
12	205	13.44	Erythronic acid (4TMS)	94.8	0.99	0.52	0.73	36
13	120	13.57	Phenylalanine (1TMS)	98.5	0.96	0.54	0.33	34
14	142	13.93	Proline [+CO2] (2TMS)	99.8	1.00	0.28	0.06	33
15	247	14.34	Glutamic acid (3TMS)	94.1	1.00	0.03	0.25	125
16	219	14.46	Phenylalanine (2TMS)	98.0	0.99	0.29	0.08	22
17	217	15.46	Ribitol (5TMS)	94.9	0.99	0.00	0.22	14
18	175	15.87	Ornithine (3TMS)	96.6	0.98	0.10	0.12	31
19	357	15.93	Glycerol-3-phosphate (4TMS)	96.7	1.00	0.22	0.45	144
20	133	16.55	Citric acid (4TMS)	90.6	0.99	0.33	0.56	292
21	156	17.01	Lysine (3TMS)	96.7	0.94	0.20	0.44	66
22	181	17.32	Tyrosine (2TMS)	95.4	1.00	0.29	0.01	29
23	156	17.65	Lysine (4TMS)	86.5	0.91	0.02	0.09	177
24	218	17.83	Tyrosine (3TMS)	97.1	1.00	0.13	0.13	56
25	333	18.38	Gluconic acid (6TMS)	94.0	0.96	0.09	0.35	58
26	174	18.83	Lysine, 5-hydroxy- (4TMS)	94.4	0.40	0.19	0.48	3
27	129	18.90	Hexadecanoic acid (1TMS)	97.4	1.00	0.27	0.01	526
28	191	19.28	Inositol, myo- (6TMS)	96.3	1.00	0.05	0.33	302
29	441	19.35	Uric acid (4TMS)	94.3	1.00	0.25	0.02	51
30	200	20.38	Tryptophan (2TMS)	98.4	0.73	0.47	0.17	72
31	117	20.48	Octadecenoic acid, 9-(E)- (1TMS)	95.4	0.97	0.38	0.08	315
32	117	20.71	Octadecanoic acid (1TMS)	97.5	1.00	0.01	0.29	291
33	204	24.57	Maltose (1MEOX) (8TMS) MP	91.2	0.99	0.08	0.47	66

Bibliography

[1] Johnson C.H., Ivanisevic J., Siuzdak G. Metabolomics: Beyond Biomarkers and Towards Mechanisms. *Nature Reviews Molecular Cell Biology*, In press.

[2] Dudley, E., Yousef, M., Wang, Y. Griffiths, W. J. Targeted metabolomics and mass spectrometry. *Adv. Protein Chem. Struct. Biol.*, (2010) 80, 45–83

[3] Vinaixa M, Schymanski EL, Neumann S, Navarro M, Salek RM, Yanes O. Mass spectral databases for LC/MS and GC/MS-based metabolomics: state of the field and future prospects. *Trends in Analytical Chemistry*, (78) 23–35, 2016

[4] Wishart DS, Tzur D, Knox C, et al. HMDB: the Human Metabolome Database. *Nucleic Acids Res.* 35 (2007).

[5] Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K.; Oda, Y.; Kakazu, Y.; Kusano, M.; Tohge, T.; Matsuda, F.; Sawada, Y.; Hirai, MY.; Nakanishi, H.; Ikeda, K.; Akimoto, N.; Maoka, T.; Takahashi, H.; Ara, T.; Sakurai, N.; Suzuki, H.; Shibata, D.; Neumann, S.; Iida, T.; Tanaka, K.; Funatsu, K.; Matsuura, F.; Soga, T.; Taguchi, R.; Saito, K.; Nishioka, T. MassBank: A public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* **2010**, 45, 703–714.

[6] Hummel, J.; Selbig, J.; Walther, D.; Kopka, J. The Golm Metabolome Database: a database for GC–MS based metabolite profiling. *Metabolomics*. **2007**, 18, 75–95.

[7] Cuadros-Inostroza, A, Caldana, C, Redestig, H, Kusano, M, Lisec, J, Pena-Cortes, H, Willmitzer, L, A; Hannah, M. TargetSearch - a Bioconductor package for the efficient preprocessing of GC–MS metabolite profiling data *BMC Bioinformatics*. 2009, 10, 428

[8] Benjamin Kehimkar, Jamin C. Hoggard, Jeremy S. Nadeau, Robert E. Synovec. Targeted mass spectral ratio analysis: A new tool for gas chromatography–mass spectrometry. *Talanta* 103 (2013) 267–275.

[9] Domingo-Almenara X, Perera A, Ramirez N, Canellas N, Correig X, Brezmes J. Compound identification in gas chromatography/mass spectrometry-based metabolomics by blind source separation. *Journal of Chromatography A*. 2015, 1409C, 226–33.

[10] X. Domingo-Almenara, A. Perera, N. Ramirez, J. Brezmes. Automated resolution of chromatographic signals by independent component analysis - orthogonal signal deconvolution in comprehensive gas chromatography/mass spectrometry-based metabolomics. *Comput. Methods Programs Biomed.* 2016, 130C, 135–141.

[11] Strehmel N, Hummel J, Erban A, Strassburg K, Kopka J. Retention index thresholds for compound matching in GC–MS metabolite profiling *Journal of Chromatography B*, (2008) 182–190

[12] H. van den Dool, P.D. Kratz A generalization of the retention index system including linear temperature programmed gas-liquid partition chromatography *J. Chromatogr.* 11 (1963) 463.

Chapter 9

Results and Conclusions

9.1 Summary of the results

Independent component regression (ICR), multivariate curve resolution (MCR–ALS) and the orthogonal signal deconvolution-based (OSD) tandem strategies ICA–OSD and MCR–OSD were tested in both pure standard and biological samples. A total of 38 and a total of 25 compounds were identified in the standard and the biological samples respectively. The identification performance of the methods was evaluated through the spectral similarity match score in comparison with the reference spectra provided by the Golm Metabolome Database (GMD). For the case of the standards samples, no significative differences in match score were appreciated between methods. For the case of the biological matrices - where compounds appear in very low concentrations and with the interference of a biological matrix -, statistical significative differences were appreciated between methods, where the OSD implementations displayed a more accurate identification of the metabolites in terms of match score and major qualitative (structure of the spectra) differences could be observed between the regression (ICR/MCR–ALS) and OSD approaches (ICA–OSD and MCR–OSD). Finally, the ICA-based approaches were faster in terms of execution time.

The capability of ICA–OSD to quantify metabolites in chromatographic samples was later evaluated by the automated resolution of GC×GC–MS chromatograms. We studied the performance of ICA–OSD by the quantification of 38 metabolites through a set of 20 Jurkat cell samples analyzed by GC×GC–MS. Results shown that ICA–OSD could be used to resolve co-eluted compounds in chromatographic signals, as the automated ICA–OSD approach was able to correctly quantify a set of compounds across different samples.

The application of multivariate algorithms in GC–MS data, e.g., MCR–ALS or ICA-based approaches, involve segmenting the chromatogram into regions or windows, which may lead to failure in the detection of compounds. Thus, we proposed the application of ICA–OSD and MCR–ALS through moving window to avoid the usual practice of chromatographic segmentation into regions or windows. We evaluated this strategy through its quantification capability in comparison with the XCMS package. Results shown that the proposed methodology was able to correctly quantify compounds appearing in biological matrices.

To illustrate the integrative workflow of eRah, we carried out a comparative metabolomic analysis using 11 serum samples from girls with hyperinsulinemic androgen excess (HIAE), and 14 age-, weight- and ethnicity-matched healthy controls. With the aim of comparing the quantitative results of the deconvolved compounds by eRah, mass spectra were also processed using XCMS. The output of eRah contained 169 resolved and aligned compounds. We focused, however, on 33 compounds to assess the quantitative accuracy of eRah in comparison with XCMS. These compounds showed a high similarity match factor (>80.0) to reference MS spectra in the GMD and MassBank. The analysis indicated an excellent linear correlation (R^2>0.90) between eRah and XCMS for most compounds. Even when appearing coeluted and/or low concentration compounds the correlation between the area - and intensity - of deconvolved compounds and selective mzRT features was high. Moreover, although

XCMS is a very reliable reference, its results should not be taken as ground truth. For this reason, we validated eRah's quantitative results of four compounds by an integration using MassHunter on the raw data and an additional GC-triple quadrupole (QqQ) MS targeted analysis, which consistently shown similar percentages of variation between HIAE and control groups in comparison with eRah. Finally, we focused on changes in four additional metabolites to be validated by complementary analyses on the same serum samples using nuclear magnetic resonance (NMR) or liquid chromatography (LC–QqQ MS). Interestingly, analytical platforms such as NMR, which analyze serum non-destructively, and LC–MS, which produces intact molecular ions due to soft ionization, revealed very similar percentages of variation and p-values

BaitMet was evaluated by the analysis of 182 human serum samples subjects with chronic kidney disease. Each sample was mixed with 9 FAMEs (C10-25). BaiTMet detected a total of 95 compounds with less than 0.1 % of error and with a spectra similarity score above 90 %. Of all the compounds detected, we focused on a subset of 33 compounds. The empirical RI for each compound found was automatically determined by the BaiTMet algorithm without FAMEs. For evaluation purposes, we additionally determined the RI with the van den Dool and Kratz algorithm [12] using the FAMEs retention time. Although statistical significative differences using a paired t-test were observed (p-value <0.0001) - FAME mean RI error was less than BaitMet -, the absolute mean difference between both methods was of 0.01 %, which is significantly less than the typical identification RI error (0.5 - 1 %). Also, the quantitative performance of BaiTMet was evaluated by comparing its quantification with a selective (quantitative) ion for each metabolite. Almost all the compounds exhibited an excellent linear relation ($R^2>0.95$).

Overall, the results of this thesis aim at improving and integrating all the steps of the metabolomics workflow, including a new factor analysis-free spectral deconvolution strategy which attempts to evolve from peak-picking to compound picking. These

methods rely on orthogonal signal deconvolution. OSD was also used in combination with independent component analysis and multivariate curve resolution, which highlighted the necessity of new methodologies for the application of multivariate methods for high- throughput compound deconvolution. In that sense, GC–MS chromatograms can be processed using eRah for untargeted purposes, which provides with a list of compounds found and putative names, and their relative concentration among samples. Also, to obtain confirmatory results or focus on certain metabolites of interest, BaitMet can be used for compound profiling of MS library available metabolites. The methods were embedded into R libraries publicly available.

9.2 Discussion of the results and further work

Metabolomics, along with the rest of the -omics sciences that use analytical chemistry platforms, have boosted the application of computational methods to deal with the large and complex datasets that those platforms generate. If we focus in metabolomics though, we can state that the the current computational methods that tackle the data processing problems lag behind the analytical platforms capabilities. In that sense, many strategies can still been designed to improve the different steps of the metabolomics workflow, and more importantly, embed them into computational tools. Those computational tools should not only integrate the complete metabolomics workflow, allowing to obtain biological interpretable information from raw data, but they should also be implemented in a highly modularized and standardized manner. This modularization should serve two purposes: in one hand, it allows the scientific community to easily implement their own specific algorithms and improvements of each step (module) of the workflow. On the other hand, it allows a straightforward comparison between different and new methods (e.g., new or improved methods for deconvolution, alignment), since the rest of the modular frame remains the same.

From the literature, we can also view two different schools that focus on the processing of chromatographic data. The first is the chemometrician school, which use multivariate methods of mathematical complexity to resolve the data generated by, in this case, GC–MS or GC×GC–MS. Generally, this school focuses on the design of strategies that, although through the use of multivariate techniques achieve a greater performance, those clearly lack of automation capabilities or tools integrating those methods. Besides, this school is focused on solving chromatographic mixtures, independently of the field of research (metabolomics). On the other hand, the bioinformatics school has put special interest in strategies that are fully automated and included into dedicated metabolomics tools (software), but this school has mainly contributed with univariate techniques.

In that sense, we designed independent component analysis – orthogonal signal deconvolution (ICA–OSD), which although it could be seen as a chemometric method, it has given a boost to the suitability of independent component analysis/blind source separation in GC–MS-based metabolomics, it has also been embedded in an R package (osd), and finally, it has been shown that its application can be fully automated, including an original moving window strategy, where ICA–OSD was compared with MCR–ALS and with univariate approaches in real metabolomics applications.

Concretely, ICA–OSD included two main improvements. First, in this ICA approach, the concept of independence was twisted: compound profiles were targeted as the independent source of the chromatographic mixture, as opposite to the spectra, which was up to the date the typical approach of the ICA-based methods for GC–MS data processing. We believe that this approach is a more natural implementation of ICA in GC–MS data. Second, orthogonal signal deconvolution included an alternative extraction of spectra, based on principal component analysis as opposite to the typical use of least squares in MCR–ALS or ICR. We have seen in OSD, a powerful and reliable tool for multivariate spectral deconvolution.

Despite the existence of different pieces of free and commercial software for GC-MS data analysis, none of these allow the execution of an integrated workflow that includes spectral deconvolution and alignment, followed by the identification and quantification of metabolites in the same application, and implemented in a modularized and standardized manner. This still leads many researchers to implement separate software for each process, and tedious manual workflows for data processing. This lack of tools integrating the complete metabolomics workflow with multivariate methods, inspired the design of eRah. When using multivariate resolution methods (e.g., ICA–OSD or MCR–ALS) the number of components or factors has to be estimated (automatedly), which is a key parameter that clearly affects the outcome of the algorithms, and for which any optimal solution has still not been found. This has clearly limited the application of multivariate algorithms for high-throughput GC-MS data processing. In eRah, we attempted to overcome this limitation by evolving the traditional peak-picking approach into an innovative multivariate compound detector, where compounds are detected as opposite to peaks. Spectra are later determined by OSD, the previously designed multivariate method for spectral deconvolution based on principal component analysis. We demonstrated the capabilities of eRah with a comparative analysis of plasma samples of adolescents with hyperinsulinaemic androgen excess, where eRah allowed to turn raw data into biological interpretable information.

eRah shown itself as a robust and efficient tool for processing GC–MS-based untargeted metabolomics data. However, tackling the data processing from different angles (ideally orthogonal, e.g., peak picking with multivariate deconvolution, or completely different approaches) could provide not only confirmatory analysis but also complementary information. This, together with the idea of integrating physicochemical knowledge into computational methods to improve the performance of those, lead to the design of BaitMet. In BaitMet, the concept of targeting the untargeted

was achieved by targeting those compounds from a mass spectral library into chromatograms acquired in full scan mode. In that sense, BaitMet was able to perform a library-driven compound profiling.

Finally, future work includes the design of new and improved compound detector match filters, and strategies that outperform the current methods for spectra deconvolution, but that avoid the use of factor analysis. Also, the capabilities of eRah's methodology could be tested in other analytical platforms including comprehensive gas chromatography or liquid chromatography – mass spectrometry.

9.3 Conclusions

This section summarizes the conclusions of the doctoral thesis.

- First, an independent component regression (ICR) for GC–MS compound identification as an alternative to multivariate curve resolution (MCR–ALS) was introduced. However, the typical approach of the ICA-based methods for GC–MS data processing was based on considering the spectra as the independent component in the chromatogram. In this thesis, the concept of independence was twisted: compound profiles were targeted as the independent source of the chromatographic mixture, as opposite to the spectra. Also, the results given by ICR were comparable to the results given by MCR–ALS, but ICR was superior in terms of execution time. This is of special interest in metabolomics due to the high amount of data that GC–MS currently generates and the quantity of samples that are analyzed in metabolomics experiments. Also, a novel orthogonal signal deconvolution (OSD) approach using principal component analysis as an alternative to the traditional least squares approach was introduced, allowing the extraction of refined spectra when compounds elute under the influence of biological matrices, compound co-elution or other types of noise. Also, we

concluded that ICA–OSD could also be used to robustly quantify compounds in chromatographic signals. This accomplished the objective O1.

- The application of multivariate algorithms in GC–MS data, e.g., MCR–ALS or ICA-based approaches, involve segmenting the chromatogram into regions or windows, which may lead to failure in the detection of compounds. Thus, we proposed the application of ICA–OSD and MCR–ALS through moving window to avoid the usual practice of chromatographic segmentation into regions or windows. We evaluated this strategy through its quantification capability in comparison with the XCMS package. Results shown that the proposed methodology was able to correctly quantify compounds appearing in biological matrices. This accomplished the objective O1.

- While univariate peak–picking approaches are focused on the ion fragment peak as the analysis entity, multivariate methods such as MCR–ALS or ICA aim at extracting the spectra from GC–MS data by taking advantage of the inherent fragment-redundancy in mass spectrometry. However, multivariate methods performance depend, to a greater degree, on an appropriate estimation of the number of components to build the multivariate model. This bottleneck has limited the use of multivariate methods in high-throughput GC–MS data processing tools. In this thesis we introduced a multivariate compound detector to detect compounds instead of peaks. We later used OSD to determine the compound spectra. The tandem application of the multivariate compound detector by local covariance (CMLC), with OSD allowed the spectral deconvolution of compounds in GC–MS mixtures without the use of factor analysis techniques. This accomplished the objective O2.

- Despite the existence of different pieces of free and commercial software for GC–MS data analysis, none of these allow the execution of an integrated workflow

that includes spectral deconvolution and alignment, followed by the identification and quantification of metabolites in the same application, and implemented in a modularized and standardized manner. This still leads many researchers to implement separate software for each process, and tedious manual workflows for data processing. In this thesis, eRah was designed to fill this gap. eRah was demonstrated to be capable of conducting the complete metabolomics workflow with robust identification and quantification methods. This accomplished the objective O3.

- BaiTMet was designed to take advantage of the knowledge provided by metabolomics spectral libraries to process full scan GC–MS chromatograms in a driven manner, and with the possibility of standardize the retention times without the use of internal standards. BaiTMet operates under the assumption that the retention time relation between metabolites naturally found in the samples can be used to predict their respective retention indexes. BaiTMet is an R package for high-throughput quantification of compounds of an entire MS library into GC-MS data. BaiTMet was able to identify compounds by the standardization of retention time without mixing internal standards in the samples. Moreover, BaiTMet is also compatible with the use of internal standards mixed in the samples, and use them to characterize the RI/RT curve in each chromatogram. This accomplished the objective O4.

- eRah and BaitMet libraries were implemented in a modularized manner, and also their structure was standardized in a single S4 method known as MetaboSet. This class is inspired in the expressionSet class widely used in genomics. This class contains all the objects to hold, not only the information of the results provided by eRah, but it may also hold results from other softwares and platforms(e.g., it could store the XCMS results after processing LC–MS data if

XCMS were adapted for that purpose). The main functions of the eRah or BaitMet packages have a MetaboSet object as an input and give another MetaboSet object as an output containing the new results. Overall, this allows a familiarized programmer to attach their own modules (e.g., deconvolution, alignment, identification) to the main package, and easily understand the internal main operation core. Then, this allows the user to easily access to the object created as a result of the processing of the data, and therefore, access to the internal structure of the results and create modifications or functions to customize their results. This accomplished the objective O5.

Chapter 10

Publications

During the development of this thesis there were produced a set of publications as scientific papers in indexed journals and in international conferences as oral or poster communications. As a result of the methods developed in this thesis, a set of R packages were also made publicly available.

10.1 Indexed Journal Papers

- Domingo-Almenara X, Perera A, Ramirez N, Canellas N, Correig X, Brezmes J. **Compound identification in gas chromatography/mass spectrometry-based metabolomics by blind source separation.** *Journal of Chromatography A* (2015). Vol. 1409: 226-233.

- Domingo-Almenara X, Perera A, Ramirez N, Brezmes J. **Automated resolution of chromatographic signals by independent component analysis - orthogonal signal deconvolution in comprehensive gas chromatography/mass spectrometry-based metabolomics.** *Computer Methods and Programs in Biomedicine* (**2016**). Vol. 130, 135-141.

- Domingo-Almenara X, Brezmes J, Vinaixa M, Samino S, Ramirez N, Ramon-

Krauel M, Lerin C, Diaz M, Ibanez L, Correig X, Perera-Lluna A, Yanes O. **eRah: a computational tool integrating spectral deconvolution and alignment with quantification and identification of metabolites in GC-MS-based metabolomics.** *Analytical Chemistry* (**2016**). Accepted.

- Domingo-Almenara X, Perera A, Venturini G, Vivo-Truyols G, Vinaixa M, Brezmes J. **Targeting the untargeted: BaiTMet, an R package for library-driven GC-MS compound profiling in metabolomics.** (*Submitted*).

- Domingo-Almenara X, Perera A, Brezmes J. **Avoiding hard chromatographic segmentation: a moving window approach for the resolution of GC-MS signals in metabolomics by multivariate methods.** (*Submitted*).

10.2 Conference Proceedings

- Domingo-Almenara X, Perera A, Ramirez N, Brezmes J. **Compound Identification in Comprehensive Gas Chromatography – Mass Spectrometry-Based Metabolomics by Blind Source Separation.** 9th International Conference on Practical Applications of Computational Biology and Bioinformatics. (2015). DOI: 10.1007/978-3-319-19776

10.3 Oral or Poster Communications

- Navarro M, Senan O, Domingo-Almenara X, Capellades J, Aguilar-Mogas A, Brezmes J, Sales-Pardo M, Guimera R, Yanes O. From 'peakomics' to metabolomics in LC-MS global profiling of human plasma. 12th Annual Conference of the Metabolomics Society (June 2016), Dublin, Ireland. *(Poster)*.

- Domingo-Almenara X, Brezmes J, Vinaixa M, Samino S, Diaz M, Ibanez L, Correig X, Perera A, Yanes O. eRah: a computational tool integrating spectral deconvolution and alignment with quantification and identification of metabolites in GC-MS-based metabolomics. Chemometrics in Analytical Chemistry (CAC), (June 2016), Barcelona, Spain. *(Poster)*.

- Domingo-Almenara X, Perera A, Vivo-Truyols G, Brezmes J. R2D2: an R package for the automated profiling of GCxGC-MS samples in untargeted metabolomics. 15th GCxGC Symposium (June 2016), Riva del Garda, Italy. *(Poster)*.

- Gomez J, Barrilero R, Domingo-Almenara X, Correig X, Brezmes J, Canellas N. Evaluation of Multivariate Curve Resolution for Macromolecular Baseline Removal in 1H-NMR Spectra, Small Molecule NMR Conference (SMASH) (September 2015), Baveno, Italy. *(Poster)*.

- Domingo-Almenara X, Brezmes J, Vinaixa M, Samino S, Ramirez N, Correig X, Perera A, Yanes O. eRah: an R package for automatic spectra deconvolution, alignment, and library matching of metabolites from GC/TOFMS untargeted metabolomics. The 11th International conference of the metabolomics society (July 2015), San Francisco, California, US. *(Poster)*.

- Domingo-Almenara X, Perera A, Ramirez N, Brezmes J. Compound Identification in Comprehensive Gas Chromatography – Mass Spectrometry-Based Metabolomics by Blind Source Separation. 9th International Conference on Practical Applications of Computational Biology and Bioinformatics. (June 2015), Salamanca, Spain. *(Oral)*.

- Domingo-Almenara X, Fernandez F, Canellas N, Perera A, Correig X, Brezmes J. Automated compound deconvolution and alignment in comprehensive double gas chromatography-mass spectrometry by blind source separation. 13th

GCxGC Symposium (June 2014), Riva del Garda, Italy. *(Poster)*.

10.4 Computational Tools and packages developed

- **osd:** *Orthogonal Signal Deconvolution for Spectra Deconvolution in GC-MS and GCxGC-MS Data*. Compound deconvolution for chromatographic data, including gas chromatography - mass spectrometry (GC-MS) and comprehensive gas chromatography - mass spectrometry (GCxGC-MS). The package includes functions to perform independent component analysis - orthogonal signal deconvolution (ICA-OSD), independent component regression (ICR), multivariate curve resolution (MCR-ALS) and orthogonal signal deconvolution (OSD) alone. URL: http://CRAN. R-project.org/package=osd.

- **erah:** *Automated Spectral Deconvolution, Alignment, and Metabolite Identification in GC/MS-Based Untargeted Metabolomics.* Automated compound deconvolution, alignment across samples, and identification of metabolites by spectral library matching in Gas Chromatography - Mass spectrometry (GC-MS) untargeted metabolomics. eRah outputs a table with compound names, matching scores and the integrated area of the compound for each sample. URL: http://CRAN. R-project.org/package=erah.

- **BaitMet** *Library driven compound profiling in full scan GC–MS.* Automated quantification of metabolites by targeting MS library into the chromatograms. BaiTMet outputs a table with compounds name, spectral matching score, general across-samples RI error, and the area of the compound for each sample. BaitMet automatically determines which is the compounds retention index without mixing internal standards. (To be uploaded).

Appendix A

Supporting Information: Compound identification in gas chromatography/mass spectrometry–based metabolomics by blind source separation

A.1 Determination of the euclidean error distance

For each method and compound, the euclidean error distance ρ is determined by the sum of euclidean differences between each m/z for a given reference spectrum S_r and an empirical spectrum S_e (Eq 1). First, both spectra have been normalized to unity.

$$\rho = \sqrt{\sum_{i=1}^{n} (S_r(i) - S_e(i))^2} \tag{A.1}$$

A.2 List of standards used in the pure standards mixture.

Table A.1: List of standards used in the pure standards mixture.

Name	
2-oxo-glutaric acid	Isoleucine
Allo-threonine	Leucine
Asparagine	Malonic acid
Aspartic acid	Methionine
Benzoic acid	Methylmalonic acid
Citric acid	Myo-inositol
Cholesterol	Nicotinic acid
Cysteine	Phenylalanine
Dodecanoic acid	Proline
Fumaric acid	Serine
Glycerol	Tryptophan
Glycine	Tyrosine
Heptadecanoic acid	Uracil

A.3 Pure standards sample identification scores

Table A.2: Identification score results for the pure standards sample.

No.#	RT	Name	ICA OSD	MCR OSD	ICR	MCR
1	4.40	Glycine (2TMS)	99.55	99.88	99.19	99.53
2	4.66	Leucine (1TMS)	99.77	99.82	99.63	99.73
3	4.81	Proline (1TMS)	99.79	99.82	99.65	99.90
4	4.82	Isoleucine (1TMS)	99.78	99.70	99.42	99.59
5	5.05	Malonic acid (2TMS)	89.90	89.97	89.38	90.69
6	5.13	Malonic acid, methyl- (2TMS)	90.65	90.71	90.49	91.27
7	5.30	Benzoic acid, (1TMS)	97.64	98.35	98.71	98.88
8	5.40	Serine (2TMS)	98.08	97.47	97.41	97.34
9	5.60	Glycerol (3TMS)	94.37	94.31	94.60	96.42
10	5.68	Nicotinic acid (1TMS)	97.76	97.90	95.82	97.81
11	5.69	Isoleucine (2TMS)	99.01	98.88	96.85	99.02
12	5.72	Proline (2TMS)	99.77	99.84	99.54	99.78
13	5.78	Glycine (3TMS)	98.07	97.59	96.43	97.72
14	6.01	Uracil (2TMS)	90.65	90.51	86.16	93.98
15	6.02	Fumaric acid (2TMS)	96.61	96.28	88.13	97.02
16	6.14	Serine (3TMS)	94.96	94.98	95.16	95.22
17	6.33	Threonine, allo- (3TMS)	99.05	99.06	99.05	99.04
18	6.46	Methionine (1TMS)	98.77	98.87	98.09	98.26
19	6.49	Malonic acid, methyl- (3TMS)	98.82	98.82	95.86	97.20
20	7.15	Methionine (2TMS)	98.70	98.78	98.13	99.06
21	7.17	Aspartic acid (3TMS)	95.03	95.08	93.61	96.02
22	7.33	Phenylalanine (1TMS)	96.47	96.72	92.17	93.53
23	7.37	Cysteine (3TMS)	97.36	97.38	96.33	96.93
24	7.48	Serine (4TMS)	97.44	97.44	95.85	95.90
25	7.53	Glutaric acid, 2-oxo- (2TMS) ‡	96.40	96.34	96.23	96.61
26	7.55	Proline [+CO2] (2TMS)	97.99	97.98	97.31	97.61
27	7.68	Asparagine (4TMS) MP	88.13	87.34	82.96	93.96
28	7.80	Phenylalanine (2TMS)	94.98	94.99	94.79	96.37
29	7.87	Dodecanoic acid (1TMS)	98.46	98.46	98.47	98.47
30	8.88	Citric acid (4TMS)	93.42	93.42	93.16	93.17
31	9.20	Tyrosine (2TMS)	99.19	99.16	98.83	98.90
32	9.47	Tyrosine (3TMS)	99.49	99.50	99.50	99.51
33	10.22	Inositol, myo- (6TMS)	96.59	96.63	96.64	96.92
34	10.34	Heptadecanoic acid (1TMS)	99.24	99.27	99.22	99.25
35	10.75	Tryptophan (2TMS)	98.53	98.52	98.31	98.70
36	10.83	Tryptophan (3TMS)	99.68	99.73	99.62	99.70
37	11.12	Cystine (4TMS)	98.39	98.40	98.27	98.42
38	14.81	Cholesterol (1TMS)	93.27	93.28	93.21	92.68

‡ Glutaric acid, 2-oxo- (1MEOX) (2TMS) MP

Appendix B

Supporting Information: Avoiding hard chromatographic segmentation: a moving window approach for the resolution of GC-MS signals in metabolomics by multivariate methods.

B.1 Supplementary Tables

Table B.1: Number of samples for where each compound was automatically detected between methods (ICA–OSD and MCR–ALS) and window length (WL) of 10, 15 and 20 seconds.

Cp.	Name	ICA–OSD			MCR–ALS		
		WL 10	WL 15	WL 20	WL 10	WL 15	WL 20
1	Lactic acid (2TMS)	25	25	25	25	25	25
2	Hexanoic acid (1TMS)	13	18	3	25	25	2
3	Valine (1TMS)	25	25	25	25	25	25
4	Hydroxylamine (3TMS)	25	25	25	25	25	25
5	Butanoic acid, 2-hydroxy- (2TMS)	20	10	25	25	25	25
6	Leucine (1TMS)	24	25	24	24	24	24
7	Butanoic acid, 3-hydroxy- (2TMS)	20	22	24	24	23	24
8	Valine (2TMS)	22	24	23	25	25	25
9	Urea (2TMS)	25	25	25	25	25	25
10	Benzoic acid, (1TMS)	24	25	25	25	25	25
11	Serine (2TMS)	24	25	25	25	25	25
12	Norleucine (2TMS)	6	5	3	10	5	5
13	Glycerol (3TMS)	24	22	22	4	17	9
14	Phosphoric acid (3TMS)	25	24	24	22	23	23
15	Isoleucine (2TMS)	17	19	24	18	17	25
16	Proline (2TMS)	17	18	24	18	17	25
17	Glyceric acid (3TMS)	18	20	18	25	25	25
18	Nonanoic acid (1TMS)	24	25	25	25	25	25
19	Threonine (3TMS)	17	21	19	25	25	24
20	Norleucine (3TMS)	10	13	16	18	20	20
21	Glumatic Acid (2TMS)	25	25	25	25	25	25
22	Proline [+CO2] (2TMS)	18	24	23	25	25	24
23	Glutamic acid (3TMS)	18	21	20	21	23	22
24	Phenylalanine (2TMS)	23	25	25	25	25	25
25	Dodecanoic acid (1TMS)	25	25	24	25	25	25
26	Ornithine (4TMS)	17	24	25	25	25	24
27	Citric acid (4TMS)	25	25	23	24	25	25
28	Tetradecanoic acid (1TMS)	25	25	25	25	25	25
29	Lysine (4TMS)	24	23	23	21	21	22
30	Hexadecanoic acid (1TMS)	25	25	25	25	25	25
31	Inositol, myo- (6TMS)	19	20	21	25	25	24
32	Uric acid (4TMS)	18	17	18	20	21	19
33	Octadecanoic acid (1TMS)	25	25	25	25	25	25

Appendix C

Supporting Information: eRah: a computational tool integrating spectral deconvolution and alignment with quantification and identification of metabolites in GC–MS-based metabolomics.

C.1 Supplementary Theory

The eRah package includes a tutorial and the description of each function and parameter through the R help. Also, a user forum is available at http://erah.lefora.com/.

This section details the methods described in the original paper.

Multivariate compound detector: compound match by local covariance (CMLC). The multivariate match filter is adapted from the versions of match filter by local co-variance, by applying the constraints of multiple channels (m/z), leading to the following equation:

$$MF(D, x) = \sum_{i}^{N} |x^T C^{-1} D_i| \tag{C.1}$$

where C (N $\times$ N) is the covariance matrix of D (N $\times$ M), where D is the data matrix comprising a sub-window (local) of the chromatogram, and x (N $\times$ 1) is the known pattern (gaussian peak shape). In this notation, N is the number of chromatographic scans (retention time), M is the range of acquisition of the mass-to-charge ratio (m/z). This filter detects the presence of a known pattern (a gaussian peak shape with a standard deviation of at least σ_{MIN}). In each scan of the chromatogram, a Region of Interest (ROI) is determined. First, the gaussian peak shape is centered on each scan. Then, each ROI is a sub-window (local) which comprises the data where the gaussian peak shape is non-zero. The local covariance matrix is determined from this ROI sub-window.

Compound spectra deconvolution: orthogonal signal deconvolution. Orthogonal signal deconvolution (OSD) is used to retrieve each compound spectrum.

To determine an spectrum, OSD needs an approximation of the compound elution profile whose spectrum is to be determined. ERah assumes that in each spot where the CMLC filter detects a compound, there is a compound with a peak width σ_{MIN} and intensity equal to the maximum value in the data point, i.e., eRah approximates the elution profile for OSD with the same gaussian model used in CMLC.

Compound chromatographic profile deconvolution: least absolute deviation. The chromatographic profile deconvolution aims at determining the quantitative compound profile. In this step, we know the spectrum whose compound profile is to be determined. An intuitive way to recover a compound chromatographic profile given its spectrum is by a least squares regression of the spectrum against the chromatogram. This process though, can be substituted by a L_1 estimation, also known as least absolute deviation (LAD). Whereas the least squares regression minimizes the squared residuals between the data and the regressor, the L_1 estimation minimizes the sums of absolute residuals. L_1 estimation is mathematically defined as:

$$min_{\alpha}\ f(\alpha) = |\alpha \mathbf{s_r} - \mathbf{D_j}| \quad ; \quad j = 1, 2, ..., N \tag{C.2}$$

where D $(N \times M)$ is the chromatographic data and s_r $(M \times 1)$ is the spectrum already determined by OSD. In this notation, N is the number of chromatographic scans (retention time) and M is the range of acquisition of the mass-charge ratio (m/z). The L_1 estimation is more robust to outliers than least squares as it weights all the observations equally. Physically, this means that all the ions of the spectrum that are to be regressed against the chromatogram have the same importance. This is a natural way to give the same importance also to the selective m/z - which best describes the pure chromatographic profile - and that tend to be less abundant. Moreover, we are fitting a single component into the chromatogram, which, due to noise and other co-eluted compounds, would not always be composed only by the compo-

nent we are regressing. We can reduce the influence from outlier data by applying a L_1 estimation. The L_1 estimation is computationally faster than the typical robust regression through M estimators.

However, sometimes the L_1 estimation introduces an energy ambiguity. This means that although the deconvolved profile represents the true chromatographic profile, its intensity does not correspond to the true one. To correct this, the intensity of the deconvolved chromatographic profile can be now adjusted by means of a least squares regression, of the chromatographic profile against the chromatogram.

Alignment. A custom clustering algorithm aligns the compounds across the different samples. First, two distance matrices are determined, one containing the retention time distance and the other containing the spectral correlation distance between the different compounds of all the samples in the experiment. The upper boundaries for those distances are required, i.e., the maximum distances up to which two or more compounds are allowed to be grouped. The distances in the matrix that are outside these boundaries are omitted. The two matrices are transformed to a single matrix containing the Euclidean distance from the retention time and correlation distances.

The algorithm is shown in **Algorithm 1**. Distances between compounds belonging to the same sample are set as null in the input euclidean distance matrix **H** for this algorithm.

Notation: k: number of components; n: number of samples; **H**: euclidean distance matrix; s: vector containing the sample for each compound; **G**: list containing grouped indexes; g: vector containing the distances for each compound versus its neighbors compounds; q vector containing local grouped indexes.

Algorithm 1 Sample-constrained clustering

Input: H $\in \mathrm{R}^{k \times k}$, s $\in \mathrm{R}^{k}$
Output: G

G $\leftarrow \emptyset$; $p \leftarrow 0$
repeat
  $g \leftarrow \emptyset$; $q \leftarrow \emptyset$
  for $j = 1$ to k **do**
    $m \leftarrow \emptyset$
    for $i = 1$ to n **do**
      Set in p the indexes for which the condition s$==i$ is True
      $m_i \leftarrow \arg\min(\mathrm{H}_{j,p})$
    end for
    $g_j \leftarrow \sqrt{\sum_{i=1}^{n} m^2}/\|m\|$
  end for
  Set in q the indexes for which **arg min(g)** is true.
  Append q to $\mathbf{G}_j$.
  $\mathbf{H}_{q,q} \leftarrow \emptyset$.
until All elements of **H** are $\emptyset$, or no alignment is feasible

The euclidean distance matrix is submitted to the clustering algorithm along with a vector containing the sample identifier-tag to which each compound belongs. The aim of the algorithm is to group the different compounds under one constraint: the compounds must belong to different samples i.e., two compounds belonging to the same sample can not be align. The algorithm determines, for each compound, the mean euclidean distance over all the other compound in the different samples. In each iteration, compounds present in at least two different samples and with the minimum mean distance are grouped together. The algorithm iterates until all the compound have been aligned or no alignment is feasible.

Identification

The identification match factor is determined by the following steps: first, the normalized spectra for each compound in all the samples is averaged by a simple mean. Then, the match score is determined by the dot product between the average

empirical spectra and the reference MS spectra. Each compound is matched against the entire library, and eRah provides with a list of putative hits (metabolite names) ordered by match score.

The cosine dot product is determined by the following equation:

$$cos(\alpha) = \frac{x \cdot y^T}{\sqrt{(x \cdot y)^T \cdot (x \cdot y)}} \tag{3}$$

where x and y are the vectors comprising the two spectra to be compared.