Universitat Rovira i Virgili

DOLPHIN AND WHALE: DEVELOPMENT, EVALUATION AND APPLICATION OF NOVEL BIOINFORMATICS TOOLS FOR METABOLITE PROFILING IN HIGH THROUGHPUT 1H-NMR ANALYSIS

Josep Gómez Álvarez

Josep Gómez Álvarez

DOLPHIN AND WHALE: DEVELOPMENT, EVALUATION AND APPLICATION OF NOVEL BIOINFORMATICS TOOLS FOR METABOLITE PROFILING IN HIGH THROUGHPUT ^{1}H-NMR ANALYSIS

DOCTORAL THESIS

Supervised by Dr. Nicolau Cañellas

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. Països Catalans 26

Campus Sescelades

43007 Tarragona

I CERTIFY that the Doctoral Thesis entitled: "DOLPHIN AND WHALE: DEVELOPMENT, EVALUATION AND APPLICATION OF NOVEL BIOINFORMATICS TOOLS FOR METABOLITE PROFILING IN HIGH THROUGHPUT ^{1}H-NMR ANALYSIS", presented by Josep Gómez Álvarez to obtain the degree of Doctor, has been performed under my supervision in the Departament d'Enginyeria Electrònica, Elèctrica i Automàtica at the Universitat Rovira i Virgili and it meets the requirements for International Mention qualification.

Tarragona, July 2016-07-24

Doctoral Thesis Supervisor

Dr. Nicolau Cañellas Alberich

“Science never solves a problem without creating ten more”

George Bernard Shaw

“Science, my boy, is made up of mistakes, but they are mistakes which it is useful to make, because they lead little by little to the truth”

Julio Verne

“Touch a scientist and you touch a child”

Ray Bradbury

ACKNOWLEDGMENTS

Per mi, aquest es un dels apartats més importants de la tesi, doncs sense el suport de la gent amb la qui he compartit part del camí aquesta no hagués estat possible...i la veritat es que tinc tants agraïments i per tanta gent que no se per on començar…

Nico, moltes gràcies per tot, per la guia, pel recolzament, per l'exigència, i pels debats i reunions que hem mantingut al llarg de tots aquests anys. No sempre ha estat fàcil i hem tingut les nostres discrepàncies, però després de tot el que hem passat crec que podem dir que tenim una gran relació molt madura, que ens permet discutir qualsevol cosa al despatx i fer-nos una abraçada amb una cervesa a la paella del Xavier o a la calçotada del Jesús...

Xavier i Jesús, moltes gràcies a tots dos per estar sempre a prop pel que fes falta, per fer grans aportacions científiques durant la tesi, per donar un cop de mà amb l'anglès, i per suposat, per les paelles i les calçotades!

Mariona i Miguel, vosaltres vau ser els meus fars al inici de la tesi, doncs recordo que cada cop que tenia algun dubte sobre NMR sempre us emprenyava. Gràcies per la paciència, cada xerrada amb vosaltres és com una *master class* per mi.

Roger i Nuria, només érem nosaltres tres al despatx 2 del bunker als inicis... tu Roger sempre has estat un model a seguir per mi, i sense tu encara estaria barallant-me amb les comandes del Matlab per plotar espectres... i tu Nuria sempre seràs la meva 'compi' de tesi, doncs vam començar junts i això es una cosa que recordo sempre amb molt de carinyo... crec que ens hauríem de veure més!

Dídac, Xavi, Rubén, Daniel, Pere, Sónia i Txell, vosaltres formeu la segona i tercera tongada de bunkis des que vaig entrar jo, doncs tot i que tu Txell vas ser un estel fugaç, també fitxaves dia rere dia allà per fer el projecte final de carrera. Amb vosaltres he tingut grans debats tècnics i grans moments lúdics, inclús he compartit pis (Dídac), runnings (Rubén), concerts, submissions i testimonis de boda (Xavi), comiats de solter (Pere) i hem esgotat les existències d'alcohol dels bars (Sónia). Se que Whale queda en bones mans Daniel.

Sara, Míriam, Jordi, Lorena, Mabel, Rita, Irene i un llarg etc. de gent del COS i els serveis científics-tècnics, gràcies per tots els sopars i ganxets pintxos! Vosaltres si que sou pintxos! (com diria la Mariona).

I would like to thank all the cheminformatics and metabolism group at EBI for their hospitality and specially to Dr Reza Salek, who contribuited greatly in the development of the software packages generated during this thesis.

Mateus, Jose, Carlos, Jeanette, Maria, Yasset, Rosa, Rosela y toda la gente con la que viví gran parte de mis aventuras en Cambridge durante mi estancia allí, muchas gracias por formar parte de ella, sois parte de mi y de esta tesis.

Un agraïment molt gran també als 'Purasangre', aquells amics de tota la vida que passi el que passi sempre estan allà i sé que sempre estaran.

Ja acabant, gràcies als meus pares, germanes, tiets, cosins i avis, perquè si ja els amics de tota la vida sé que sempre estan i estaran allà, la família encara amb més garanties.

Non può mancare all'appello la famiglia napoletana e tutti gli amici che ho conosciuto in questa meravigliosa città.

Per últim els agraïments més importants per a la Floriana Montalto, doncs la nostra relació va començar als pocs mesos de començar la tesi i ha estat sempre al meu costat en els moments més durs. Aquesta tesi i tot el que d'ella es derivarà també es teu, gràcies per tot.

ABSTRACT

Metabolite profiling is the most challenging approach in NMR spectral analysis. It aims to comprehend biological processes occurring in a certain moment through identifying and quantifying metabolites present in complex NMR mixtures. An NMR spectrum is composed by resonances of a huge number of metabolites, and these resonances often overlap between them, shift position depending on the sample pH and can be masked by macromolecules signals. All these drawbacks hinder metabolite identification and quantification, so obtaining a cured metabolite profile of a sample can be a very big issue even for expert users.

One single metabolite may present one, two or several resonances in an NMR spectrum, being these signals what is called the metabolite signature. Some of these signals can be unique according their position or peaks multiplicity, but some others can be easily confused, producing false positives.

To identify a metabolite unequivocally, users should detect all the resonances that belong to the metabolite signature, and to do so, 2-dimensional acquisitions are often needed. To quantify a metabolite accurately, the best option is to calculate the area under the curve of the less congested resonance, provided that the target resonance has enough signal to noise ratio. If this resonance is a totally isolated signal, a bucket integration of the signal region is optimal, being spectral binning the faster approach to use. But if the resonance is partially masked by macromolecules signals or by neighbouring resonances, spectral devoncolution methods are needed in order to properly calculate the area under the curve of the resonance.

In this context, the motivation of this thesis was born with the aim to provide automatisms and user-friendly interactive functions for NMR metabolite profiling, improving the quality of the results and reducing the time span of the analysis. To do so, several algorisms were implemented and embedded into two software packages.

The first package, Dolphin, aims to profile a fixed set of metabolites in biofluids such as aqueous extracts and urine in a fully automated manner. To do so, Dolphin takes profit of the 2-dimensional JRES spectra, where signals are much less superposed. In those spectra, the congestion produced by the multiplicity of signals disappears, since them are projected in an orthogonal dimension. The orthogonal cut at the position of the signal of interest allows the algorithm to compare the multiplicity and j-coupling of the signal placed there with an internal

library, and if all the resonances in the library match with a determined compound, a metabolite is considered identified.

In a subsequent step, Dolphin focalizes the most isolated signal of a targeted metabolite and finds the neighbouring signals that can affect its quantification. In this process, Dolphin annotates the position, multiplicity and j-coupling of the neighbouring signals in a completely automated manner, despite the neighbouring signals identity (known or unknown compounds). Finally, it performs a line-shape fitting of that region, adjusting the intensities of all the signals present and modelling their shapes as lorentzian-gaussian functions.

The second package, Whale, aims to profile a flexible set of metabolites in any kind of biofluid in a semi automated manner, and allows users approaching metabolomics NMR data in different ways depending on the final goal of their studies.

Whale incorporates several functions for exploring data using fingerprinting techniques, which include comparisons between spectra and metadata, finding spectral regions that present significant variability between groups and finding spectral regions that are related between them. Moreover, it offers up to four different quantification modes in order to optimize the time-span of the analysis and gives users the option of manually adjust the quantification parameters to test its performance before running an automated metabolite profiling in the whole dataset.

To facilitate metabolite assignments, it offers metabolite suggestions depending on the spectral region and the biofluid under analysis, and a repository panel where the user can compare target spectra with reference spectra from public databases. Whale outputs allow users to detect where the algorithm failed and re-run the analysis only for those spectra where the algorithm failed.

An evaluation of Dolphin's strategy and the line-shape fitting algorithm for automated quantification included in both tools are presented in this thesis, as well as two full NMR-based metabolomics studies where Whale was applied.

LIST OF ABREVIATIONS

ACQ	acquisition
ART	effective antiretroviral therapy
BMRB	biological magnetic resonance data bank
C	cholesterol
CDV	cardiovascular disease
CI	confidence interval
CIBERDEM	Centro de Investigación Biomédica en Red de Diabetes y Enfermedades Metabólicas Asociadas
COSY	correlation spectroscopy
CPMG	Carr-Purcell-Meiboom-Gill
CSF	cerebrospinal fluid
CSN	total sum normalization
CTLS	constrained total line-shape
DNA	deoxyribonucleic acid
DSS	2,2-dimethyl-2-silapentane-5-sulfonate sodium salt
EDTA	ethylene diamine tetraacetic acid
ERETIC	electronic reference to access in vivo concentrations
FCS	functional class scoring
FID	finite induction decay
FT	Fourier transformation
GUI	graphical user interface
HCV	hepatitis C virus
HDL	high density lipoprotein
HDL-C	HDL cholesterol
HDL-P	HDL particles
HIV	human immunodeficiency virus
HMBC	heteronuclear multiple-bond correlation spectroscopy
HMDB	human metabolome database

HSQC	heteronuclear single-quantum correlation spectroscopy
IISPV	Institut d'Investigació Sanitària Pere Virgili
INR	immunological non-responders
IR	immunological responders
JRES	j-resolved
LDL	low density lipoprotein
LDL-C	LDL cholesterol
LDL-P	LDL particles
LMWM	low molecular weight metabolites
MCMC	Markov chain Monte Carlo
MS	mass spectrometry
NMR	nuclear magnetic resonance
NOESY	nuclear overhauser effect spectroscopy
OPLS-DA	orthogonal partial least squares discriminant analysis
ORA	over-representation analysis
PC	principal component
PCA	principal component analysis
PEPA	positional enrichment by proton analysis
PLS	partial least squares
PLS-DA	partial least squares discriminant analysis
RF	random forest
ROC	receiver operating characteristic
ROI	region of interest
SR	spectral reference
TMS	trimethylsilane
TOCSY	total correlation spectroscopy
TSP	3-trimethylsilylpropionic acid
VLDL	very low density lipoproteins
VLDL-C	VLDL cholesterol
VLDL-P	VLDL particles

LIST OF PUBLICATIONS

· Esther Rodríguez-Gallego*, **Josep Gómez***, Yolanda María Pacheco*, Joaquim Peraire, Consuelo Viladés, Raúl Beltrán-Debón, Roger Mallol, Miguel López-Dupla, Sergi Veloso, Verónica Alba, Julià Blanco, Nicolau Cañellas, Anna Rull, Manuel Leal, Xavier Correig, Pere Domingo and Francesc Vidal. *"Baseline metabolomic signature predicts immunological CD4+ T-cell recovery after 36 months of virologically successful ART in adult HIV-infected patients: a pilot study"*. (Submitted).

· Maria Vinaixa*, Miguel A. Rodríguez, Suvi Aivio, Jordi Capellades, **Josep Gómez**, Nicolau Cañellas, Travis Stracker and Oscar Yanes. *"Positional Enrichment by Proton Analysis (PEPA): a one-dimensional ^{1}H-NMR approach for ^{13}C stable isotope resolved metabolomics"*. (Submitted).

· **Josep Gómez*,** Jesus Brezmes, Roger Mallol, Miguel A. Rogríguez., Maria Vinaixa, Reza M Salek, Xavier Correig and Nicolau Cañellas. *"Dolphin: a tool for automatic targeted metabolite profiling using 1D and 2D ^{1}H-NMR data"*. Analytical and Bioanalytical Chemistry (**2014**). Vol 406, 7967-76. DOI: 10.1007/s00216-014-8225-6.

· **Josep Gómez*,** Maria Vinaixa, Miguel A. Rodríguez, Reza M. Salek, Xavier Correig and Nicolau Cañellas. *"Dolphin 1D: Improving Automation of Targeted Metabolomics in Multi-matrix Datasets of ^{1}H-NMR spectra"*. 9th International Conference on Practical Applications of Computational Biology and Bioinformatics. (2015). Vol 375, 59-67. DOI: 10.1007/978-3-319-19776-0_7.

LIST OF CONGRESSES

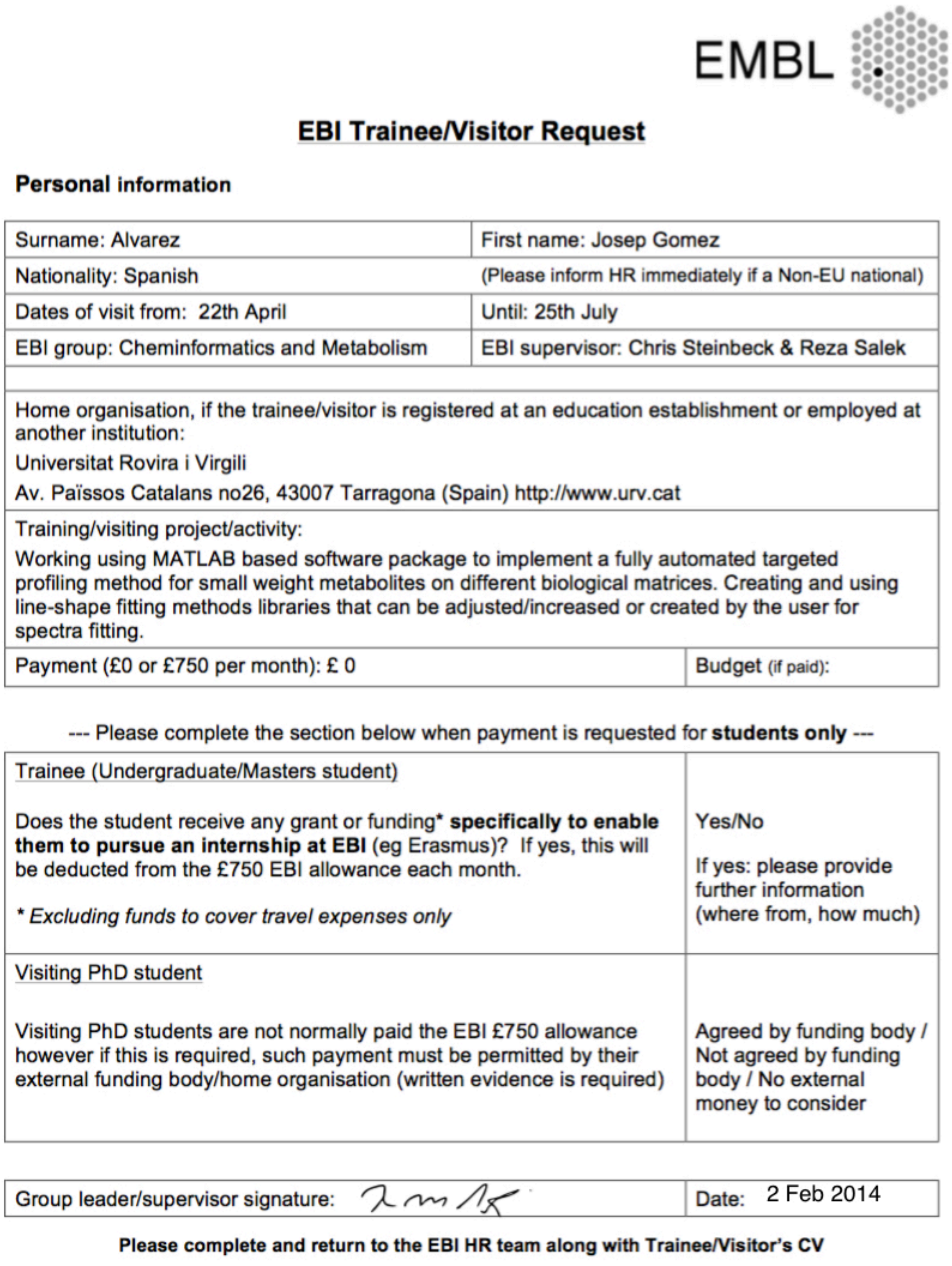
· **Josep Gómez***, Rubén Barrilero, Xavier Domingo, Xavier Correig and Nicolau Cañellas. *"Evaluation of Multivariate Curve Resolution for Macromolecular Baseline Removal in ^{1}H-NMR Spectra".* Small Molecule NMR Conference (SMASH), Baveno, Italy (2015). **Poster.**

· **Josep Gómez***, Maria Vinaixa, Miguel A. Rogríguez, Noelia Ramirez, Reza M. Salek, Xavier Correig and Nicolau Cañellas. *"Whale: a package combining fingerprinting and targeted metabolite profiling to improve the extraction of metabolic information in NMR spectra".* The 11th International conference of the metabolomics society, San Francisco, USA (2015). **Poster.**

· Maria Vinaixa*, Miguel A. Rodríguez, Suvi Aivio, Jordi Capellades, **Josep Gómez**, Nicolau Cañellas, Travis Stracker and Oscar Yanes. *"Positional Enrichment by Proton Analysis (PEPA): a one-dimensional ^{1}H-NMR approach for ^{13}C stable isotope resolved metabolomics".* The 11th International conference of the metabolomics society, San Francisco, USA (2015). **Poster.**

· **Josep Gómez***, Maria Vinaixa, Miguel A. Rodríguez, Reza M. Salek, Xavier Correig and Nicolau Cañellas. *"Dolphin 1D: Improving Automation of Targeted Metabolomics in Multi-matrix Datasets of ^{1}H-NMR spectra".* 9th International Conference on Practical Applications of Computational Biology and Bioinformatics, Salamanca, Spain (2015). **Oral presentation.**

EUROPEAN VISIT



EMBL

EBI Trainee/Visitor Request

Personal information

Surname: Alvarez	First name: Josep Gomez
Nationality: Spanish	(Please inform HR immediately if a Non-EU national)
Dates of visit from: 22th April	Until: 25th July
EBI group: Cheminformatics and Metabolism	EBI supervisor: Chris Steinbeck & Reza Salek
Home organisation, if the trainee/visitor is registered at an education establishment or employed at another institution: Universitat Rovira i Virgili Av. Països Catalans no26, 43007 Tarragona (Spain) http://www.urv.cat	
Training/visiting project/activity: Working using MATLAB based software package to implement a fully automated targeted profiling method for small weight metabolites on different biological matrices. Creating and using line-shape fitting methods libraries that can be adjusted/increased or created by the user for spectra fitting.	
Payment (£0 or £750 per month): £ 0	Budget (if paid):

--- Please complete the section below when payment is requested for **students only** ---

Trainee (Undergraduate/Masters student) Does the student receive any grant or funding* **specifically to enable them to pursue an internship at EBI** (eg Erasmus)? If yes, this will be deducted from the £750 EBI allowance each month. ** Excluding funds to cover travel expenses only*	Yes/No If yes: please provide further information (where from, how much)
Visiting PhD student Visiting PhD students are not normally paid the EBI £750 allowance however if this is required, such payment must be permitted by their external funding body/home organisation (written evidence is required)	Agreed by funding body / Not agreed by funding body / No external money to consider

Group leader/supervisor signature:	Date: 2 Feb 2014

Please complete and return to the EBI HR team along with Trainee/Visitor's CV

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

1.1 METABOLOMICS

Metabolomics is the last of the four most representative “omics” sciences, whereas genomics, transcriptomics and proteomics focuses on the study of genes, transcripts and proteins, metabolomics is focused on the comprehensive characterization of the metabolites present in biological matrices. Metabolites are the intermediates and end products of metabolism, and can be defined as any molecule less than 1kDa in size[1]. However, bigger molecules such as albumin and lipoproteins are also included in metabolomics studies of blood serum and plasma[2]. In human-based metabolomics, metabolites can be divided into two major groups, the endogenous metabolites, which are produced by the host organism, and the exogenous metabolites[3]. Metabolites of foreign substances such as drugs are termed xenometabolites[4].

Metabolites are involved in a wide range of cellular functions, including cell energetics, system defense, signaling, as well as building blocks of structural biopolymers such as proteins and DNA[5]. Endogenous metabolites are often altered in response to environmental factors, disease, nutrition and other aspects in an attempt to maintain cellular homeostasis in the organism. Obtaining the metabolite profile of a sample at a given time may provide a comprehensive view of biochemical reactions and cellular phenotypes. The collection of all the metabolites present in a determined biological tissue, organ or organism is called the metabolome[5]. Metabolomics aims to fully understand the magnitude and the interactions of the metabolome in order to obtain a comprehensive view of the biological behavior of cells, tissues and organisms.

Metabolomics is then a key part of the “omics puzzle”, and its integration with genomics, transcriptomics and proteomics is needed to perform systems biology studies. In this context, the pioneer in applying the scope of systems biology to studies of metabolism was Jeremy Nicholson, creating a new approach called metabonomics[6,7]. Metabonomics is defined as "the quantitative measurement of the dynamic multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification"[6].

There has been some disagreement over the exact differences between 'metabolomics' and 'metabonomics'. While there is still no absolute agreement, there is a growing consensus that 'metabolomics' places a greater emphasis on metabolic profiling at a cellular or organ level and is primarily concerned with normal endogenous metabolism. 'Metabonomics' extends metabolic profiling to include information about perturbations of metabolism caused by environmental factors (including diet and toxins), disease processes, and the involvement of extragenomic influences, such as gut microflora. This is not a trivial difference; metabolomic studies should, by definition, exclude metabolic contributions from extragenomic sources, because these are external to the system being studied. However, in practice, within the field of human disease research there is still a large degree of overlap in the way both terms are used, and they are often in effect synonymous[8].

Obtaining the metabolic profile of an entire metabolome using only one detection method and one protocol is not possible due to the huge number of metabolites present and their chemical differences. The most widely used techniques for metabolite detection in metabolomics are mass spectrometry (MS) and nuclear magnetic resonance spectroscopy (NMR). While MS presents higher sensitivity, NMR is highly quantitative and reproducible. Ultimately, both techniques are complementary and allow researchers to obtain more comprehensive metabolic profiling[9]. Moreover, depending on the final goal, different protocols of extraction, ionization, isotope labeling, etc… can be performed to obtain a snap-shot covering a certain part of the global metabolism.

The amount of data generated in the metabolomics field has increased during the last years. Innovations in instrumentation, data mining and bioinformatic tools are constantly emerging with the aim of improving the comprehension of the metabolism of biological organisms[10].

The metabolomics community usually organizes events annually via the metabolomics society, dedicated to promoting the growth, use and understanding of metabolomics in the life sciences. The Metabolomics Standards Initiative (MSI) helps to coordinate the work and to ensure that approaches and data relationships shared by the various working groups are maintained[11,12]. Part of this effort is currently carried out by the coordination of standards in metabolomics (COSMOS) initiative. The MetaboLights team is coordinating this consortium of 14 European partners, playing a role as repository of metabolomics experiments[13,14]. A key aspect of this effort aims to develop efficient policies ensuring that metabolomics data is encoded in open standards, tagged with a community-agreed and complete set of metadata[15].

1.2 NUCLEAR MAGNETIC RESONANCE (NMR) SPECTROSCOPY IN METABOLOMICS

Nuclear magnetic resonance (NMR) spectroscopy is a quantitative and non-destructive technique. It is also a robust and reliable analytical method with paramount reproducibility and repeatability[16].

Its physical principle relies on the fact that spins of each molecule within a sample are excited by means of a magnetic field followed by a radio frequency pulse that aligns the spins in a manner that molecules can be measured. Once the radio frequency is switched off, spins recover their initial state at a rate that depends on the molecular weight of the molecule. This recovering step is known as relaxation, and the signal obtained between the alignment and the total recovery of the spins is called finite impulse decay (FID). The FID shows a decaying sinusoid containing all the signals from all the molecules within a sample. The NMR spectrum of a given sample is the result of applying the Fourier transformation (FT) to this FID, and it provides information of both the environment of the molecule moieties (structure elucidation) and the abundance of a given molecule (quantification).

The main biofluids analyzed in NMR-based metabolomics studies are blood plasma or serum, urine, cerebrospinal fluid (CSF) and cell or tissue extracts. Most of these biofluids can be obtained quite easily with minimal invasion. Moreover, a high sampling frequency can be achieved[17].

The detection limit of metabolite concentration using NMR spectroscopy is of the order of micromolar (μM) and the number of observable metabolites in biofluids largely depends on the magnetic field strength of NMR spectrometer. Therefore, working at the highest available magnetic field is recommended. Generally, 500 or 600 MHz NMR instruments are used in metabolomics studies, because these fields are easily accessed. However, the use of 800 or 900 MHz has already been reported[18,19]. From the first real application of NMR to the analysis of biofluids into the early 80s up to now, the increase in field strength has tremendously improved the technique resolution.

A very important benefit of NMR spectroscopy for metabolic profiling is that it is quantitative and does not require time-consuming sample preparation steps, like separation or derivatization. Moreover, it does not require a prior knowledge about compounds present in a sample and is thus ideally suited for non-targeted profiling[20].

Figure C1.1 shows the typical processes in NMR-metabolomics workflow, and each of them will be detailed in the following subsections.

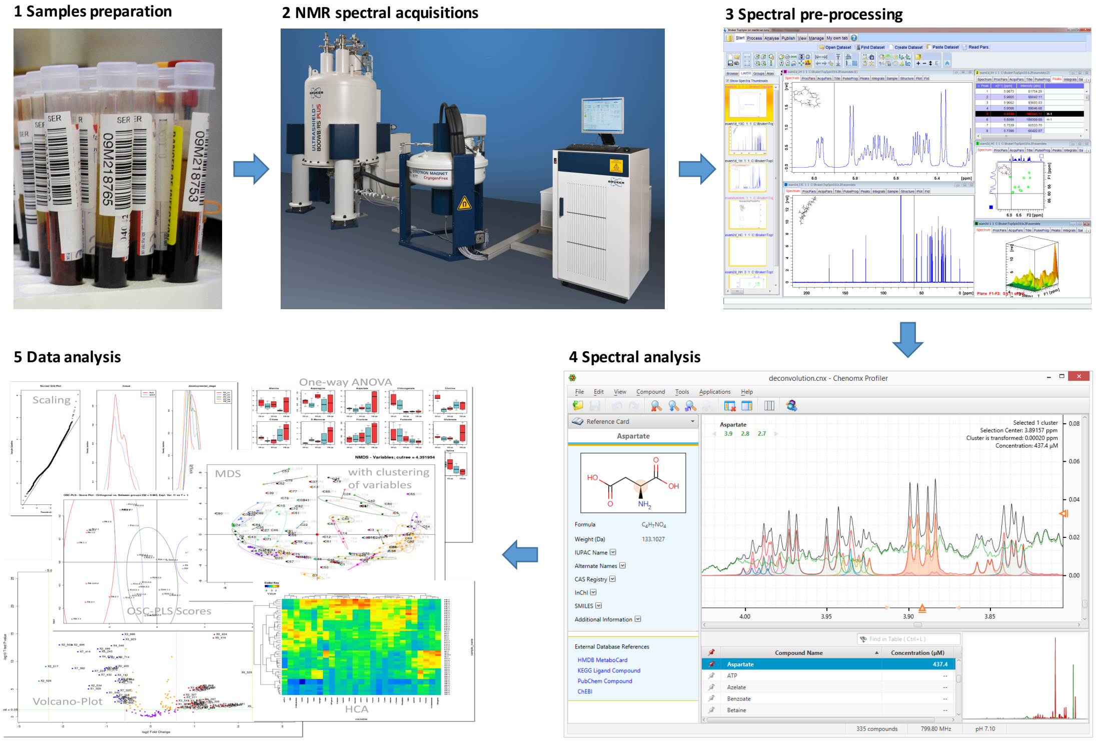


Figure C1.1. NMR-metabolomics workflow.

1.2.1 Samples and its preparation

The biological sample must be collected under strict conditions regardless of its type. For instance, blood is usually collected by venipuncture into standard vials containing either ethylene diamine tetra acetate (EDTA) or lithium heparin as anti-coagulant. When using EDTA, the NMR spectrum will show extra resonances. These resonances correspond to complexes formed by the EDTA with Ca2+ and Mg2+ ions that are present in plasma[21].

Urine samples need addition of sodium azide to control bacterial growth, while plasma and serum can be measured directly with minimal sample preparation. Dilution of plasma or serum is recommended, since it reduces the sample's viscosity and releases plasma-protein bound metabolites. Both ultra-filtration and solvent precipitation are used for protein removal in serum and plasma samples, but recent comparisons between these two methods have demonstrated that ultra-filtration is better for metabolic NMR measurement[22,23]. Filters need to be centrifuged with water before usage in order to avoid contaminations due to the presence of molecules such as glycerol, which has been found in many commercially available filters.

As the pH of samples has significant influence on the chemical shifts of the resonances along the NMR spectrum, its monitoring is essential in order to expedite metabolite profiling, especially in urine samples. Additions of NaOH and HCl may be used to manually adjust the pH when analyzing urine[24,25], but the most common method to maintain a controlled pH range between any kind of samples is by adding a phosphate buffer stock solution with D_2O at pH 7.4[26].

Finally, additions of internal standards at known concentrations are typically used for referencing purposes in terms of absolute quantification and chemical shifting. The most commonly used are the sodium salt of 3-trimethylsilylpropionic acid (TSP), the 2,2-dimethyl-2-silapentane-5-sulfonate sodium salt (DSS) and the trimethylsilane (TMS) for organic solvents. Another strategy is to introduce a synthetic electronic reference signal called Electronic Reference to access In vivo Concentrations (ERETIC)[27], which substitutes the addition of chemical compounds to obtain absolute concentrations in the metabolic profiling. In this case, other signals present in the samples have to be used for referencing chemical shifting, such as glucose[28].

Detailed procedures to collect, store and measure the mostly studied biofluids have been provided in the literature[17,26].

1.2.2 NMR Spectral acquisitions

The NMR approach allows users to obtain different types of spectral acquisitions by applying different pulse sequences. Moreover, a wide variety of spin ½ nuclei such as ^{1}H, ^{13}C, ^{31}P, ^{15}N

or ^{19}F can be measured.

^{1}H-NMR approach is the most used in NMR-based metabolomics due to the natural abundance of hydrogen in biological molecules. Most NMR metabolomics experiments, and especially those performed on biofluids, depend on effective suppression of the water resonance. The one-dimensional nuclear Overhauser effect spectroscopy with spoil gradient (1D NOESY) pulse is the most popular method for solvent suppression, because of its robustness and ease of implementation. All the necessary concepts of parameter optimization and the mechanism of 1D NOESY for water suppression have already been reviewed[29].

In some cases, as in serum or plasma, obtaining a reliable metabolic profiling of the low-molecular-weight metabolites (LMWMs) in a 1D NOESY spectrum o is severely compromised by the resonances of macromolecules such as lipids and proteins. The most popular method for avoiding this macromolecular baseline is to apply the Carr-Purcell-Meiboom-Gill (CPMG) filter[30], which is able to remove these broad resonances exploiting their shorter transverse relaxation rate[31]. The main problem of working with 1D CPMG spectra is that all the resonances, including those of the LMWMs, have been reduced by its own spin-spin relaxation time (T_2), and obtaining the absolute concentration of the metabolites present in a sample is not trivial. While in 1D NOESY spectra the absolute quantification of a metabolite can be obtained directly as the ratio between the area of one of its resonances and the area of a reference compound[32], in 1D CPMG spectra each resonance needs to be corrected by its own T_2 first. Few studies show T_2 factors of some resonances for a subset of metabolites under determined conditions, but the reality is that several quantifications of each resonance are needed in order to accurately calibrate its T_2 decay and obtain its absolute concentration value[33].

Depending on the complexity of the biofluid, 1D spectra are not enough to obtain a large and reliable metabolic profiling of the samples due to the high signal overlap. This signal overlap masks the resonance structures with all their specific attributes (position, multiplicity and J-coupling), making very difficult the assignment and quantification of metabolites. There exist a lot of two-dimensional (2D) experiments with the aim of contribute with extra information about the resonances and their relations, which is very helpful to achieve reliable assignments of those resonances to metabolites.

Homonuclear 2D J-resolved (2D JRES) is very often used in metabolomics studies of biofluids[34]. This approach separates the effects of chemical shift and J-coupling into independent dimensions[35,36]. Another popular homonuclear 2D NMR experiment is the correlation spectroscopy (COSY) sequence, which is used to identify spins which are coupled to each other[37]. Total correlation spectroscopy (TOCSY) creates correlations between all protons within a given spin system, not just between geminal or vicinal protons as in COSY. Correlations are seen between distant protons as long as there are couplings between every intervening proton[38].

Heteronuclear single-quantum correlation spectroscopy (HSQC) detects correlations between nuclei of two different types that are separated by one bond. This method gives one peak per pair of coupled nuclei, whose two coordinates are the chemical shifts of the two coupled atoms[39]. Heteronuclear multiple-bond correlation spectroscopy (HMBC) detects heteronuclear correlations between two nuclei separated by ranges of about 2-4 bounds[40]. Even if both methods can combine different nuclei, ^{1}H and ^{13}C are the most commonly used in NMR-based metabolomics studies.

1.2.3 Spectral pre-processing

Transformation of the time-domain FID into the frequency domain spectrum by FT is the initial step of all NMR analysis in metabolomics. Before FT apodization, a zero filling of FID can be applied in order to obtain better spectral resolution. After the FT apodization, the phase is corrected to obtain absorption line shape. The software packages provided by the spectrometer vendors as well as the freely available NMRPipe[41] or commercial such as Chenomx NMR Suite[42] and MestreNova[43] contain numerous spectral processing tools for spectral pre-treatment.

A key factor for spectral comparison between samples is area normalization. Signal intensities in NMR spectra can be distorted by a great number of factors such as the spectrometer, the probe, the NMR pulse sequence, the temperature and the sample itself (relaxation times, J-couplings, etc.)[44]. One of the most commonly used normalization method is integral normalization, where the spectra are normalized by dividing each signal of a spectrum by the total peak areas. This method is called total sum normalization (CSN)[45]. It assumes that the total peak area of a spectrum remains constant across the samples, which is not always the case of biological samples. This method can easily fail in metabolomic studies of blood serum and plasma, where molecules such as lipids and proteins are greatly affecting the total area of a spectrum. To avoid this kind of deviations, other normalization techniques have been developed such as normalization to the creatine concentration[46] (commonly used in urine samples) or to an internal quantitative reference[47] (e.g. TSP or DSS) by which the intensity of the spectrum is scaled.

Before starting the spectral analysis, some other processes are needed in order to accurately compare the samples under study. Depending on the quality of the spectra, a baseline correction may be necessary to remove certain artifacts or perform spectral enhancement. The most popular methods of baseline correction involve fitting a polynomial[48–50], fitting a regression curve to a spectrum using a penalized least square approach[51], using B-splines[52] or applying mixture models[53]. Removal such artifacts is critical to yield accurate results by any subsequent method of quantitative analysis[54].

The last issue to overcome as a pre-processing step is spectral shifting. Correcting shift deviations along large datasets is key in order to facilitate posterior comparisons between data, especially in large-scale studies where hundreds or even thousands of spectra may be analyzed in high-resolution by pattern recognition analysis. Algorisms to locate and calibrate large datasets of complex biofluids such as blood serum and plasma have been already reported[28].

1.2.4 Spectral analysis

Once all the necessary pre-processing filters have been applied and the spectra are ready, the user can start with the spectral analysis. There are two basic different approaches for spectral analysis: fingerprinting and profiling. Fingerprinting is the most commonly used method in NMR metabolomics and it is based upon the multivariate analysis of a dataset consisting on a large amount of sample NMR spectra, where each spectrum can be considered as a fingerprint of unassigned signals arising from low molecular weight analytes. Profiling is more challenging but ultimately more meaningful approach for analyzing NMR spectra. It is based on the analysis of an array of metabolites known to be involved in a given biochemical pathway. It does not allow for fast and high-throughput automated measurements since considerable human intervention is needed to guide the process of identifying and quantifying metabolites in NMR spectral data, especially when dealing with complex mixtures such as biofluids[55]. Within the profiling approach there exist two different approaches, targeted profiling and untargeted profiling.

1.2.4.1 Fingerprinting

Metabolite fingerprinting provides a powerful method for discriminating between biological samples on the basis of differences in metabolism without entering in details about individual metabolites. In this application, statistical analysis of metabolic datasets is used to compare the overall metabolic composition of related samples[56]. This approach is very useful for sorting and classifying data into subgroups, enabling conclusions to be drawn about the classification of the samples. This kind of analysis can be performed as exploratory analysis in a first step, allowing users to focus a posterior metabolic analysis on the variables that are most important in achieving these discriminations in a second step.

The statistical methods used to analyze sets of metabolite fingerprints fall into two main categories, unsupervised and supervised methods[57]. Unsupervised methods classify the spectra without the knowledge of the class of biological specimens (such as disease or control) by using the NMR frequencies of each sample as input. The most commonly applied unsupervised method in multivariate analysis is principal component analysis (PCA)[58]. In this procedure the variation between spectra is reduced to a series of principal components (PCs) and generates two classes of descriptors known as scores and loadings. Scores are linear combinations of the original spectral variables, and in a score plot each point represents the specified principal component of a single fingerprint (spectrum). Loadings describe the weightings that are applied to each of the original spectral variables as they are combined to generate each PC and thus identify the regions of the spectra responsible for defining the distribution of the samples. Other unsupervised approaches, such as nonlinear mapping[59] or hierarchical cluster analysis[60], may be used to explore groupings within the data and to aid their visualization.

The second main strategy for comparing fingerprints is to use supervised methods that exploit information about the grouping of the samples. Supervised techniques can be appropriate to force classification (such as in determining which metabolites distinguish between groups) or to regress a pattern against a trend (such as correlating a temporal progression with metabolic changes)[61]. The most commonly applied approaches are those based on partial least squares projection to latent structures (PLS)[62]. PLS derives latent variables, analogous to principal components of PCA, which describe the maximum proportion of the covariance between the measured data and the response variable, where the latter is the information that is to be explained or predicted by the model. This basic approach may be extended in several ways. For instance, PLS discriminant analysis (PLS-DA) applies the logic of PLS to discriminate between groups of samples that are defined as separate response variables[63]. Orthogonal PLS-DA (OPLS-DA) provides a further refinement of this approach in which variation in the measured data is partitioned into two blocks: one containing variation that correlates with the class identifier (defined by the response variable) and the other containing variation that is orthogonal to the first block and, thus, does not contribute to discrimination between the defined groupings[64]. The advantage of this approach is that, by separating between group and within-group variation, it permits the contribution of different regions of the spectra to the discrimination between classes to be identified, and the source of the independent variability between samples to be investigated[65].

1.2.4.2 Profiling

Metabolite profiling is the most challenging approach in NMR spectral analysis. It aims to comprehend biological processes occurring in a certain moment through identifying and quantifying metabolites present in complex NMR mixtures. An NMR spectrum is composed by resonances of a huge number of metabolites, and these resonances often overlap between them, shift position depending on the sample pH and can be masked by macromolecules signals. All these drawbacks hinder metabolite identification and quantification, so obtaining a cured metabolite profile of a sample can be a very big issue even for expert users.

Within the profiling, there exist two main approaches, which are targeted profiling and untargeted profiling. Untargeted metabolomics studies are global in scope and have the aim of simultaneously measuring as many metabolites as possible from biological samples without bias. This strategy, known as top-down strategy, avoids the need for a prior specific hypothesis on a particular set of metabolites and, instead, analyses the global metabolomic profile[66]. Conversely, targeted metabolomics studies are hypothesis driven experiments and are characterized by the measurement of predefined sets of metabolites, typically focusing on one or more related pathways of interest[67].

One single metabolite may present one, two or several resonances in an NMR spectrum, being these signals what is called the metabolite signature. Some of these signals can be unique according their position or peaks multiplicity, but some others can be easily confused, producing false positives. To identify a metabolite unequivocally, the user should detect all the resonances that belong to the metabolite signature, and to do so, 2D acquisitions are often needed. All these 2D acquisitions decongest the resonances allowing the user to discriminate signals in crowded regions.

There exist public metabolite libraries such as the Human Metabolome Database (HMDB)[68], the Biological Magnetic Resonance Data Bank (BMRB)[69] or the Birmingham Metabolite Library (BML)[70] where any user can check the metabolite signature of a determined compound both in 1D and 2D acquisitions. Moreover, the user can find other relevant information, such as the biofluids where metabolite can be found and its estimated average concentration in normal conditions.

To quantify a metabolite accurately, the best option is to calculate the area under the curve of the less congested resonance, provided that the target resonance has enough signal to noise ratio. If this resonance is a totally isolated signal, a bucket integration of the signal region is optimal, being spectral binning the faster approach to use[71–74]. If the resonance is partially masked by macromolecules signals or by neighbouring resonances, spectral deconvolution methods are needed in order to properly calculate the area under the curve of the resonance. In this sense, two main approaches have demonstrated to provide efficient deconvolutions in highly overlapped spectral regions, which are Bayesian decompositions[75,76] and constrained total line shape fitting (CTLS)[77–80].

Within the CTLS approach, one of the most novel advances is the proposed by Mika Tainen, where CTLS is combined with a quantum mechanical (QM) theory. This combination is called quantitative Quantum Mechanical Total-Line-Shape (qQMTLS) and has been tested in human serum mimics[33]. The current version demands the support of PERCH NMR Software (http://www.perchsolutions.com).

Within the Bayesian decompositions approach, there exist two promising bioinformatics tools for ^{1}H-NMR spectra profiling. One of them is the Bayesian AuTomated Metabolite Analyser for NMR spectra (BATMAN)[81,82]. It is a free available R package created to quantify a set of metabolites in 1D ^{1}H-NMR spectra. It combines Bayesian models with a Markov Chain Monte Carlo (MCMC) algorithm to estimate metabolite concentrations. It incorporates an internal library of signals associated with metabolites, and users can select the signals they want to quantify and some other specific parameters through the R console before starting the analysis. The main limitations are that it does not offer any assistance in signal identification and that the user interactivity is very improvable. Moreover, the final concentration estimates of each compound depend a lot on the parameters set-up, and due to its deconvolution algorithm strategy (where the final shape is always 100% filled by the sum of the signals plus the wavelet) it can easily produce over-fitting if those parameters are not ideally optimized.

The other tool is called BAYESIL[83]. It is a web system based application that automatedly identifies and quantifies metabolites in 1D ^{1}H-NMR spectra of ultra-filtered plasma, serum or CSF. It uses Bayesian models to match the metabolite signatures of more than 50 metabolites of its internal library with the NMR spectrum. The main limitations are that users have to collect the spectra in a standardized fashion for Bayesil to perform optimally and that it only works for NMR spectra of three ultra-filtered biofluids.

Actually, there exist some free-available software tools which aim is to automate or assist metabolite identification and quantification. Some of them are focused in metabolite identification, where tools such as MetaboHunter[84] and MetaboID[85] use 1D data only and CCPN[86], COLMAR[87] and MetaboMiner[88] use 2D acquisitions for this purpose. Some others, such as BATMAN and MetaboQuant[89], are focused in metabolite quantification, and few try to cover the two processes in a unique tool such as Focus[90], rNMR[91] and BAYESIL.

Despite the number of free-available tools that have emerged to guide, assist or automate NMR data analysis, none of them has been well established yet, and commercial packages such as Bioref AMIX (Bruker, GmbH, Silberstreifen, Rheinstetten, Germany) and Chenomx NMR Suite (Chenomx Inc., Edmonton, Alberta, Canada) are the most popular tools used in the NMR metabolomics field.

In terms of profiling, AMIX allows to visualize spectra in both 1D and 2D acquisitions and manually assign metabolites to signals. One can easily process 1D spectra with ‘bucketing’ selection, and not than easily with signal deconvolution. The final output contains the area of each signal in arbitrary units in a matrix format. Chenomx is quite more user-friendly to use, and the final output contains the absolute concentrations of each metabolite, which is calculated through a reference standard included in the samples (such as TSP). Its strategy consists in manually fit reference spectra of pure metabolites stored in its library with the sample spectra. However, the vast number of metabolites included in its library can generate confusion in metabolite assignments to non-expert users. Recent versions of Chenomx offer automatic fitting, but by now they recommend users to manually check its performance due to the poor reliability of the fully automatic approach.

Even combining commercial and free available packages, the challenges for high throughput metabolite identification and quantification remain an open topic for research.

1.2.5 Data analysis

The analysis of NMR data can be carried out by different approaches, depending on the final goal. The unsupervised and supervised methods described in the fingerprinting section are the most used methods in profiling as well, but using metabolite concentrations as input instead of spectral features. Therefore, PCA and PLS are usually applied to analyze NMR data obtained from metabolite profiling.

Besides PCA and PLS, there are two important methods to analyze NMR data obtained from the profiling approach, pathway analysis and time course analysis. Pathway analysis allows users to detect the biological mechanisms in which identified metabolites are involved. Methods such as over-representation analysis (ORA) and functional class scoring (FCS) are the most important within pathway analysis, as well as pathway simulation methods. All these methods are core methods currently in use on metabolomics data analysis platforms, such as MetaboAnalyst[92,93]. Detailed explanations of how to apply these three methods have been recently reviewed[94].

Finally, basic statistical testing is commonly applied to each individual metabolite in order to find significant differences between groups. Normality test such as saphiro-test can provide information about the distribution of the target metabolite in the samples, and according to it, the user can apply a parametric test, such as a t-test, or a nonparametric test, such as the wilcoxon-mann-whitney test, to obtain its p-value. Besides a p-value list, fold change values, heat-maps and boxplots are usually applied to individual metabolites to analyze its importance in NMR-based metabolomics studies.

1.2.6 Applications

NMR-based metabolomics has a huge number of applications. The high reproducibility of the technique gives this method a number of advantages over other analytical techniques in large-scale and long-term metabolomics studies.

One of the most important applications is medical research[95], which includes human disease diagnosis[96,97], biomarker discovery[98,99], pharmaceutical research[100,101] and personalized treatments[102,103].Moreover, this approach has been used as a potential diagnostic tool for a wide variety of human diseases such as cancer[104,105], neurological disorders[106,107], cardiovascular disease[108,109] and diabetes[110,111].

Important applications are placed in the food and nutrition field[112], which includes wine quality analysis[113], growth monitoring of animals[114] and analysis of storage and processing of food[115]. Moreover, NMR-based is applied in toxicity tests[116,117] and forensic science[118].

1.3 THESIS MOTIVATION AND OBJECTIVES

The doctoral thesis presented in this document is the result of the research conducted in the Department of Electronic, Electrical and Automation Engineering at the Rovira i Virgili University (URV). The Metabolomics Platform, a joint research facility created by the URV and the Spanish Biomedical Research Center in Diabetes and Metabolic Disorders (CIBERDEM), has also been involved in this research. The Metabolomics Platform is part of the Pere Virgili Health Research Institute (IISPV), a major medical research organization that undertakes numerous research initiatives in the country.

As a metabolomics platform, we have several collaborations with different research groups, which means a huge number of NMR datasets to analyze. Metabolic profiles of different biofluids including boold plasma and serum, urine, tissue extracts and cell cultures are continuously demanded in order to obtain a better understanding of the behavior of the metabolome under determined conditions. None of the existing tools was satisfactory enough providing users good agreements between interactivity and automation when performing untargeted and targeted metabolic profiling in NMR datasets of different matrices.

In this context, the motivation of this thesis was born with the aim to provide automatisms and user-friendly interactive functions for NMR metabolic profiling, improving the quality of the results and reducing the time span of the analysis. To do so, we have implemented several algorisms that have ended up becoming bioinformatic packages. More concretely, the main objectives of this thesis are the following:

1 Develop and evaluate a set of algorithms able to find and quantify a significant number of target signals in a completely automated manner combining 1D (NOESY and CPMG) spectra with its own 2D JRES complementary spectra.

2 Develop and evaluate a collection of algorithms able to allow the user to interact with the spectra and to use different automated quantification modes under a user-friendly an intuitive GUI, avoiding the necessity of 2D JRES complementary spectra to perform the profiling.

3 Develop and evaluate a group of functions able to combine fingerprinting approaches and to import reference spectra from public databases to perform the most reliable NMR profiling minimizing user-subjectivities but avoiding black-box processes at the same time in a useful and versatile tool.

Apart of these three principal objectives, some secondary objectives were proposed in order to enlarge the knowledge acquired in the NMR-based metabolomics field at the end of this thesis, such as:

1 Apply the bioinformatics tools developed to achieve the main objectives of this thesis to obtain reliable and biologically contrastable results.

2 Develop a set of functions able to handle all the main statistical approaches applied in NMR-based metabolomics to convert metabolic profiles into significant and reliable results.

3 Participate in the design and development of NMR-based metabolomics studies from the initial hypothesis to the discussion and conclusion of the final results.

1.4 ORGANIZATION OF THE DOCUMENT

This chapter has provided the objectives of this thesis, a generic introduction to the metabolomics field and a more detailed explanation of all processes involved in NMR-based metabolomics studies, emphasizing spectral analysis. A brief state of the art of fingerprinting and profiling approaches, including an overview of the current available packages to perform this kind of analyses, has been exposed. The motivation of this thesis comes from the necessity of optimizing spectral analysis in NMR datasets, since none of the current tools is infallible and several processes can be improved. In this sense, novel automatic profiling methods and strategies for approaching NMR spectral data have been developed during this thesis.

Chapter 2 presents Dolphin, a tool for automated targeted metabolite profiling using 1D and 2D ^{1}H-NMR data. The contents of this chapter have been published as a research article in the journal *Analytical and Bioanalytical Chemistry*. It describes a new methodology for automated metabolite identification in 2D-JRES spectra and its posterior automated quantification in 1D ^{1}H-NMR spectra. Briefly, its strategy consists in finding metabolite specific resonances stored in an internal library based on their position, multiplicity and j-coupling, and annotate these three characteristics of all the signals surrounding the target in order to quantify the target resonance using its line-shape fitting algorithm. Dolphin's performance was evaluated using a pull of standards at known concentrations and a spike-in experiment in human urine samples. Moreover, its performance in biological samples was compared with two of the most used methods for metabolite quantification, manual reference deconvolution and bucket integration, represented by two of the most established packages, Chenomx and AMIX respectively.

Chapter 3 presents Whale, a tool that optimizes metabolite profiling in ^{1}H-NMR datasets combining fingerprinting functions and automated quantification methods with user interactivity under a user-friendly and intuitive GUI. This chapter contains a description of all Whale's functions and two full studies where the tool has been applied. The first of them presents a new one-dimensional ^{1}H-NMR approach for ^{13}C stable isotope resolved metabolomics called Positional Enrichment by Proton Analysis (PEPA) and has been submitted to *Angewandte Chemie* journal. The second study presents a pilot study where NMR-based metabolomics has been applied for finding baseline biomarkers able to predict immunological CD4^{+} T-cell recovery after 36 months of virologically successful ART in adult HIV-infected patients, and has been submitted to *The Lancet Infectious Diseases* jornal.

Chapter 4 and chapter 5 contain a general discussion and the general conclusions of this thesis, and the Annexes section at the end of this document contains the user manual of Whale.

1.5 REFERENCES

1. Samuelsson, L. M. & Larsson, D. G. J. Contributions from metabolomics to fish research. *Mol. Biosyst.* **4,** 974–9 (2008).

2. Nicholson, J. K., Foxall, P. J. D., Spraul, M., Farrant, R. D. & Lindon, J. C. 750 MHz 1H and 1H-13C NMR Spectroscopy of Human Blood Plasma. *Anal. Chem.* **67,** 793–811 (1995).

3. Nordström, A., O'Maille, G., Qin, C. & Siuzdak, G. Nonlinear data alignment for UPLC-MS and HPLC-MS based metabolomics: quantitative analysis of endogenous and exogenous metabolites in human serum. *Anal. Chem.* **78,** 3289–95 (2006).

4. Crockford, D. J. *et al.* 1 H NMR and UPLC-MS E Statistical Heterospectroscopy: Characterization of Drug Metabolites (Xenometabolome) in Epidemiological Studies. *Anal. Chem.* **80,** 6835–6844 (2008).

5. Monteiro, M. S., Carvalho, M., Bastos, M. L. & De Pinho, P. G. Metabolomics analysis for biomarker discovery: Advances and challenges. *Curr. Med. Chem.* **20,** 257–271 (2013).

6. 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,** 1181–1189 (1999).

7. Nicholson, J. K., Connelly, J., Lindon, J. C. & Holmes, E. Metabonomics: a platform for studying drug toxicity and gene function. *Nat. Rev. Drug Discov.* **1,** 153–61 (2002).

8. Robertson, D. G. Metabonomics in toxicology: a review. *Toxicol. Sci.* **85,** 809–22 (2005).

9. Pan, Z. & Raftery, D. Comparing and combining NMR spectroscopy and mass spectrometry in metabolomics. *Anal. Bioanal. Chem.* **387,** 525–527 (2007).

10. Patti, G. J., Yanes, O. & Siuzdak, G. Metabolomics Science Rev.

11. Goodacre, R. *et al.* Proposed minimum reporting standards for data analysis in

metabolomics. *Metabolomics* **3,** 231–241 (2007).

12. Sumner, L. W. *et al.* Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics* **3,** 211–221 (2007).

13. Haug, K. *et al.* MetaboLights--an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res.* **41,** D781-6 (2013).

14. Steinbeck, C. *et al.* MetaboLights: towards a new COSMOS of metabolomics data management. *Metabolomics* **8,** 757–760 (2012).

15. Emwas, A.-H. M., Salek, R. M., Griffin, J. L. & Merzaban, J. NMR-based metabolomics in human disease diagnosis: applications, limitations, and recommendations. *Metabolomics* **9,** 1048–1072 (2013).

16. Dumas, M.-E. *et al.* Assessment of Analytical Reproducibility of 1 H NMR Spectroscopy Based Metabonomics for Large-Scale Epidemiological Research: the INTERMAP Study. *Anal. Chem.* **78,** 2199–2208 (2006).

17. Keun, H. C. & Athersuch, T. J. *Metabolic Profiling. Methods in molecular biology (Clifton, N.J.)* **708,** (Humana Press, 2011).

18. Nagana Gowda, G. A., Gowda, Y. N. & Raftery, D. Massive glutamine cyclization to pyroglutamic acid in human serum discovered using NMR spectroscopy. *Anal. Chem.* **87,** 3800–5 (2015).

19. Liu, J. *et al.* 1H nuclear magnetic resonance brain metabolomics in neonatal mice after hypoxia-ischemia distinguished normothermic recovery from mild hypothermia recoveries. *Pediatr. Res.* **74,** 170–9 (2013).

20. Smolinska, A., Blanchet, L., Buydens, L. M. C. & Wijmenga, S. S. NMR and pattern recognition methods in metabolomics: from data acquisition to biomarker discovery: a review. *Anal. Chim. Acta* **750,** 82–97 (2012).

21. Nicholson, J. K., Buckingham, M. J. & Sadler, P. J. High resolution 1H n.m.r. studies of vertebrate blood and plasma. *Biochemical Journal* **211,** 605–615 (1983).

22. Tiziani, S. *et al.* Optimized metabolite extraction from blood serum for 1H nuclear magnetic resonance spectroscopy. *Anal. Biochem.* **377,** 16–23 (2008).

23. Daykin, C. A., Foxall, P. J. D., Connor, S. C., Lindon, J. C. & Nicholson, J. K. The comparison of plasma deproteinization methods for the detection of low-molecular-weight metabolites by (1)H nuclear magnetic resonance spectroscopy. *Anal. Biochem.* **304,** 220–30 (2002).

24. Slupsky, C. M. *et al.* Investigations of the effects of gender, diurnal variation, and age in human urinary metabolomic profiles. *Anal. Chem.* **79,** 6995–7004 (2007).

25. Viant, M. R., Ludwig, C., Rhodes, S., Günther, U. L. & Allaway, D. Validation of a urine metabolome fingerprint in dog for phenotypic classification. *Metabolomics* **3,** 453–463 (2007).

26. Beckonert, O. *et al.* Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nat. Protoc.* **2,** 2692–2703 (2007).

27. Akoka, S., Barantin, L. & Trierweiler, M. Concentration Measurement by Proton NMR Using the ERETIC Method. *Anal. Chem.* **71,** 2554–7 (1999).

28. Pearce, J. T. M. *et al.* Robust algorithms for automated chemical shift calibration of 1D 1H NMR spectra of blood serum. *Anal. Chem.* **80,** 7158–62 (2008).

29. Mckay, R. T. How the 1D-NOESY suppresses solvent signal in metabonomics NMR spectroscopy: An examination of the pulse sequence components and evolution. *Concepts Magn. Reson. Part A* **38A,** 197–220 (2011).

30. Meiboom, S. & Gill, D. Modified Spin-Echo Method for Measuring Nuclear Relaxation Times. *Rev. Sci. Instrum.* **29,** 688 (1958).

31. Bharti, S. K. *et al.* Improved quantification from 1H-NMR spectra using reduced repetition times. *Metabolomics* **4,** 367–376 (2008).

32. Serkova, N., Fuller, T. F., Klawitter, J., Freise, C. E. & Niemann, C. U. H-NMR-based metabolic signatures of mild and severe ischemia/reperfusion injury in rat kidney transplants. *Kidney Int.* **67,** 1142–51 (2005).

33. Tiainen, M., Soininen, P. & Laatikainen, R. Quantitative Quantum Mechanical Spectral Analysis (qQMSA) of (1)H NMR spectra of complex mixtures and biofluids. *J. Magn. Reson.* **242,** 67–78 (2014).

34. Aue, W. P., Karhan, J. & Ernst, R. R. Homonuclear broad band decoupling and two-

dimensional J-resolved NMR spectroscopy. *J. Chem. Phys.* **64,** 4226–4227 (1976).

35. Ludwig, C. & Viant, M. R. Two-dimensional J-resolved NMR spectroscopy: review of a key methodology in the metabolomics toolbox. *Phytochem. Anal.* **21,** 22–32 (2010).

36. Huang, Y., Cai, S., Zhang, Z. & Chen, Z. High-resolution two-dimensional J-resolved NMR spectroscopy for biological systems. *Biophys. J.* **106,** 2061–70 (2014).

37. Aue, W. ., Bachmann, P., Wokaun, A. & Ernst, R. . Sensitivity of two-dimensional NMR spectroscopy. *J. Magn. Reson.* **29,** 523–533 (1978).

38. Braunschweiler, L. & Ernst, R. . Coherence transfer by isotropic mixing: Application to proton correlation spectroscopy. *J. Magn. Reson.* **53,** 521–528 (1983).

39. Bodenhausen, G. & Ruben, D. J. Natural abundance nitrogen-15 NMR by enhanced heteronuclear spectroscopy. *Chem. Phys. Lett.* **69,** 185–189 (1980).

40. Bax, A. & Summers, M. F. 1H and 13C assignments from sensitivity-enhanced detection of heteronuclear multiple-bond connectivity by 2D multiple quantum NMR. *J. Am. Chem. Soc.* **108,** 2093–2094 (1986).

41. Delaglio, F. *et al.* NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6,** 277–93 (1995).

42. Weljie, A. M., Newton, J., Mercier, P., Carlson, E. & Slupsky, C. M. Targeted profiling: quantitative analysis of 1H NMR metabolomics data. *Anal. Chem.* **78,** 4430–4442 (2006).

43. Cobas, J. C. & Sardina, F. J. Nuclear magnetic resonance data processing. MestRe-C: A software package for desktop computers. *Concepts Magn. Reson.* **19A,** 80–96 (2003).

44. Bampos, N. Book Review:A Handbook of Nuclear Magnetic Resonance. Ray Freeman. Addison Wesley Longman, Harlow, 2nd Edition, 1997. ISBN 0 582 25184 2. *Magn. Reson. Chem.* **36,** 311–311 (1998).

45. Bollard, M. E., Stanley, E. G., Lindon, J. C., Nicholson, J. K. & Holmes, E. NMR-based metabonomic approaches for evaluating physiological influences on biofluid composition. *NMR Biomed.* **18,** 143–62 (2005).

46. Jatlow, P. Correction of Urine Cotinine Concentrations for Creatinine Excretion: Is It

Useful? *Clin. Chem.* **49,** 1932–1934 (2003).

47. Xu, Q., Sachs, J. R., Wang, T.-C. & Schaefer, W. H. Quantification and identification of components in solution mixtures from 1D proton NMR spectra using singular value decomposition. *Anal. Chem.* **78,** 7175–85 (2006).

48. Gan, F., Ruan, G. & Mo, J. Baseline correction by improved iterative polynomial fitting with automatic threshold. *Chemom. Intell. Lab. Syst.* **82,** 59–65 (2006).

49. Chang, D., Banack, C. D. & Shah, S. L. Robust baseline correction algorithm for signal dense NMR spectra. *J. Magn. Reson.* **187,** 288–92 (2007).

50. Xi, Y. & Rocke, D. M. Baseline correction for NMR spectroscopic metabolomics data analysis. *BMC Bioinformatics* **9,** 324 (2008).

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

52. Eilers, P. H. C. & Marx, B. D. Flexible smoothing with B-splines and penalties. *Stat. Sci.* **11,** 89–121 (1996).

53. de Rooi, J. & Eilers, P. Deconvolution of pulse trains with the L0 penalty. *Anal. Chim. Acta* **705,** 218–26 (2011).

54. Malz, F. & Jancke, H. Validation of quantitative NMR. *J. Pharm. Biomed. Anal.* **38,** 813–23 (2005).

55. Griffiths, W. J. *Metabolomics, Metabonomics and Metabolite Profiling*. **4,** (Royal Society of Chemistry, 2008).

56. Kruger, N. J., Troncoso-Ponce, M. A. & Ratcliffe, R. G. 1H NMR metabolite fingerprinting and metabolomic analysis of perchloric acid extracts from plant tissues. *Nat. Protoc.* **3,** 1001–12 (2008).

57. Trygg, J., Holmes, E. & Lundstedt, T. Chemometrics in metabonomics. *J. Proteome Res.* **6,** 469–79 (2007).

58. Kemsley, E. K. *et al.* Multivariate techniques and their application in nutrition: a metabolomics case study. *Br. J. Nutr.* **98,** 1–14 (2007).

59. Gartland, K. P. *et al.* Pattern recognition analysis of high resolution 1H NMR spectra of urine. A nonlinear mapping approach to the classification of toxicological data. *NMR Biomed.* **3,** 166–172 (1990).

60. Tikunov, Y. *et al.* A novel approach for nontargeted data analysis for metabolomics. Large-scale profiling of tomato fruit volatiles. *Plant Physiol.* **139,** 1125–37 (2005).

61. Gowda, G. A. N. *et al.* Metabolomics-based methods for early disease diagnostics. *Expert Rev. Mol. Diagn.* **8,** 617–33 (2008).

62. Eriksson, L. *Multi- and megavariate data analysis*. (Umetrics AB, 2006).

63. Jonsson, P. *et al.* Extraction, interpretation and validation of information for comparing samples in metabolic LC/MS data sets. *Analyst* **130,** 701–7 (2005).

64. Cloarec, O. *et al.* Evaluation of the orthogonal projection on latent structure model limitations caused by chemical shift variability and improved visualization of biomarker changes in 1H NMR spectroscopic metabonomic studies. *Anal. Chem.* **77,** 517–26 (2005).

65. Bylesjö, M. *et al.* OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. *Journal of Chemometrics* **20,** 341–351 (2006).

66. Alonso, A., Marsal, S. & JuliÃ , A. Analytical Methods in Untargeted Metabolomics: State of the Art in 2015. *Front. Bioeng. Biotechnol.* **3,** 1–20 (2015).

67. Patti, G. J., Yanes, O. & Siuzdak, G. Innovation: Metabolomics: the apogee of the omics trilogy. *Nat. Rev. Mol. Cell Biol.* **13,** 263–9 (2012).

68. Wishart, D. S. *et al.* HMDB: the Human Metabolome Database. *Nucleic Acids Res.* **35,** D521-6 (2007).

69. Doreleijers, J. F. *et al.* BioMagResBank database with sets of experimental NMR constraints corresponding to the structures of over 1400 biomolecules deposited in the Protein Data Bank. *J. Biomol. NMR* **26,** 139–46 (2003).

70. Ludwig, C. *et al.* Birmingham Metabolite Library: a publicly accessible database of 1-D 1H and 2-D 1H J-resolved NMR spectra of authentic metabolite standards (BML-NMR). *Metabolomics* **8,** 8–18 (2011).

71. De Meyer, T. *et al.* NMR-based characterization of metabolic alterations in hypertension using an adaptive, intelligent binning algorithm. *Anal. Chem.* **80,** 3783–90 (2008).

72. Alves, A. C., Rantalainen, M., Holmes, E., Nicholson, J. K. & Ebbels, T. M. D.

Analytic properties of statistical total correlation spectroscopy based information recovery in 1H NMR metabolic data sets. *Anal. Chem.* **81,** 2075–84 (2009).

73. Anderson, P. E., Reo, N. V., DelRaso, N. J., Doom, T. E. & Raymer, M. L. Gaussian binning: a new kernel-based method for processing NMR spectroscopic data for metabolomics. *Metabolomics* **4,** 261–272 (2008).

74. Jacob, D., Deborde, C. & Moing, A. An efficient spectra processing method for metabolite identification from 1H-NMR metabolomics data. *Anal. Bioanal. Chem.* **405,** 5049–61 (2013).

75. Ochs, M. F., Stoyanova, R. S., Arias-Mendoza, F. & Brown, T. R. A new method for spectral decomposition using a bilinear Bayesian approach. *J. Magn. Reson.* **137,** 161–76 (1999).

76. Stoyanova, R., Nicholson, J. K., Lindon, J. C. & Brown, T. R. Sample classification based on Bayesian spectral decomposition of metabonomic NMR data sets. *Anal. Chem.* **76,** 3666–74 (2004).

77. Laatikainen, R., Niemitz, M., Malaisse, W. J., Biesemans, M. & Willem, R. A computational strategy for the deconvolution of NMR spectra with multiplet structures and constraints: analysis of overlapping 13C-2H multiplets of 13C enriched metabolites from cell suspensions incubated in deuterated media. *Magn. Reson. Med.* **36,** 359–65 (1996).

78. Soininen, P., Haarala, J., Vepsäläinen, J., Niemitz, M. & Laatikainen, R. Strategies for organic impurity quantification by 1H NMR spectroscopy: Constrained total-line-shape fitting. *Anal. Chim. Acta* **542,** 178–185 (2005).

79. Jukarainen, N. M. *et al.* Quantification of 1H NMR spectra of human cerebrospinal fluid: a protocol based on constrained total-line-shape analysis. *Metabolomics* **4,** 150–160 (2008).

80. Mihaleva, V. V *et al.* Automated quantum mechanical total line shape fitting model for quantitative NMR-based profiling of human serum metabolites. *Anal. Bioanal. Chem.* **406,** 3091–102 (2014).

81. Hao, J. *et al.* Bayesian deconvolution and quantification of metabolites in complex 1D NMR spectra using BATMAN. *Nat. Protoc.* **9,** 1416–27 (2014).

82. Hao, J., Astle, W., De Iorio, M. & Ebbels, T. M. D. BATMAN--an R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a Bayesian model. *Bioinformatics* **28,** 2088–90 (2012).

83. Ravanbakhsh, S. *et al.* Accurate, Fully-Automated NMR Spectral Profiling for Metabolomics. *PLoS One* **10,** e0124219 (2015).

84. Tulpan, D., Léger, S., Belliveau, L., Culf, A. & Čuperlović-Culf, M. MetaboHunter: an automatic approach for identification of metabolites from 1H-NMR spectra of complex mixtures. *BMC Bioinformatics* **12,** 400 (2011).

85. MacKinnon, N. *et al.* MetaboID: a graphical user interface package for assignment of 1H NMR spectra of bodyfluids and tissues. *J. Magn. Reson.* **226,** 93–9 (2013).

86. Chignola, F. *et al.* The CCPN Metabolomics Project: a fast protocol for metabolite identification by 2D-NMR. *Bioinformatics* **27,** 885–6 (2011).

87. Zhang, F., Robinette, S. L., Bruschweiler-Li, L. & Brüschweiler, R. Web server suite for complex mixture analysis by covariance NMR. *Magn. Reson. Chem.* **47 Suppl 1,** S118-22 (2009).

88. Xia, J., Bjorndahl, T. C., Tang, P. & Wishart, D. S. MetaboMiner--semi-automated identification of metabolites from 2D NMR spectra of complex biofluids. *BMC Bioinformatics* **9,** 507 (2008).

89. Klein, M. S., Oefner, P. J. & Gronwald, W. MetaboQuant: a tool combining individual peak calibration and outlier detection for accurate metabolite quantification in 1D (1)H and (1)H-(13)C HSQC NMR spectra. *Biotechniques* **54,** 251–6 (2013).

90. Alonso, A. *et al.* Focus: a robust workflow for one-dimensional NMR spectral analysis. *Anal. Chem.* **86,** 1160–9 (2014).

91. Lewis, I. A., Schommer, S. C. & Markley, J. L. rNMR: open source software for identifying and quantifying metabolites in NMR spectra. *Magn. Reson. Chem.* **47,** S123–S126 (2009).

92. Xia, J., Mandal, R., Sinelnikov, I. V, Broadhurst, D. & Wishart, D. S. MetaboAnalyst 2.0--a comprehensive server for metabolomic data analysis. *Nucleic Acids Res.* **40,** W127-33 (2012).

93. Xia, J. & Wishart, D. S. Web-based inference of biological patterns, functions and

pathways from metabolomic data using MetaboAnalyst. *Nat. Protoc.* **6,** 743–60 (2011).

94. Ren, S., Hinzman, A. A., Kang, E. L., Szczesniak, R. D. & Lu, L. J. Computational and statistical analysis of metabolomics data. *Metabolomics* **11,** 1492–1513 (2015).

95. Nordström, A. & Lewensohn, R. Metabolomics: moving to the clinic. *J. Neuroimmune Pharmacol.* **5,** 4–17 (2010).

96. Wen, H. *et al.* A new NMR-based metabolomics approach for the diagnosis of biliary tract cancer. *J. Hepatol.* **52,** 228–33 (2010).

97. Sinclair, A. J. *et al.* NMR-based metabolomic analysis of cerebrospinal fluid and serum in neurological diseases--a diagnostic tool? *NMR Biomed.* **23,** 123–32 (2010).

98. Blasco, H. *et al.* Untargeted 1H-NMR metabolomics in CSF: toward a diagnostic biomarker for motor neuron disease. *Neurology* **82,** 1167–74 (2014).

99. Gebregiworgis, T. & Powers, R. Application of NMR Metabolomics to Search for Human Disease Biomarkers. *Comb. Chem. High Throughput Screen.* **15,** 595–610 (2012).

100. Fan, T. W.-M. *et al.* Stable isotope-resolved metabolomics and applications for drug development. *Pharmacol. Ther.* **133,** 366–91 (2012).

101. Beyoğlu, D. & Idle, J. R. Metabolomics and its potential in drug development. *Biochem. Pharmacol.* **85,** 12–20 (2013).

102. Coen, M. *et al.* Pharmacometabonomic investigation of dynamic metabolic phenotypes associated with variability in response to galactosamine hepatotoxicity. *J. Proteome Res.* **11,** 2427–40 (2012).

103. Nicholson, J. K., Wilson, I. D. & Lindon, J. C. Pharmacometabonomics as an effector for personalized medicine. *Pharmacogenomics* **12,** 103–11 (2011).

104. Maxwell, R. J. *et al.* Pattern recognition analysis of 1H NMR spectra from perchloric acid extracts of human brain tumor biopsies. *Magn. Reson. Med.* **39,** 869–77 (1998).

105. Carrola, J. *et al.* Metabolic signatures of lung cancer in biofluids: NMR-based metabonomics of urine. *J. Proteome Res.* **10,** 221–30 (2011).

106. Kork, F. *et al.* A possible new diagnostic biomarker in early diagnosis of Alzheimer's

disease. *Curr. Alzheimer Res.* **6,** 519–24 (2009).

107. Blasco, H. *et al.* 1H-NMR-based metabolomic profiling of CSF in early amyotrophic lateral sclerosis. *PLoS One* **5,** e13223 (2010).

108. Mäkinen, V.-P. *et al.* 1H NMR metabonomics approach to the disease continuum of diabetic complications and premature death. *Mol. Syst. Biol.* **4,** 167 (2008).

109. Mallol, R. *et al.* Liposcale: a novel advanced lipoprotein test based on 2D diffusion-ordered 1H NMR spectroscopy. *J. Lipid Res.* **56,** 737–46 (2015).

110. Nicholson, J. K. *et al.* Proton-nuclear-magnetic-resonance studies of serum, plasma and urine from fasting normal and diabetic subjects. *Biochem. J.* **217,** 365–75 (1984).

111. Mallol, R., Rodriguez, M. A., Brezmes, J., Masana, L. & Correig, X. Human serum/plasma lipoprotein analysis by NMR: application to the study of diabetic dyslipidemia. *Prog. Nucl. Magn. Reson. Spectrosc.* **70,** 1–24 (2013).

112. Laghi, L., Picone, G. & Capozzi, F. Nuclear magnetic resonance for foodomics beyond food analysis. *TrAC Trends Anal. Chem.* **59,** 93–102 (2014).

113. Alañón, M. E., Pérez-Coello, M. S. & Marina, M. L. Wine science in the metabolomics era. *TrAC Trends Anal. Chem.* **74,** 1–20 (2015).

114. Schock, T. B. *et al.* Evaluation of Pacific white shrimp (Litopenaeus vannamei) health during a superintensive aquaculture growout using NMR-based metabolomics. *PLoS One* **8,** e59521 (2013).

115. Piras, C., Scano, P., Locci, E., Sanna, R. & Marincola, F. C. Analysing the effects of frozen storage and processing on the metabolite profile of raw mullet roes using ^{1}H NMR spectroscopy. *Food Chem.* **159,** 71–9 (2014).

116. Xu, H.-D. *et al.* (1)H NMR based metabolomics approach to study the toxic effects of herbicide butachlor on goldfish (Carassius auratus). *Aquat. Toxicol.* **159,** 69–80 (2015).

117. Jordan, J., Zare, A., Jackson, L. J., Habibi, H. R. & Weljie, A. M. Environmental contaminant mixtures at ambient concentrations invoke a metabolic stress response in goldfish not predicted from exposure to individual compounds alone. *J. Proteome Res.* **11,** 1133–43 (2012).

118. Kwon, B. *et al.* 1H NMR spectroscopic identification of a glue sniffing biomarker. *Forensic Sci. Int.* **209,** 120–5 (2011).

CHAPTER II

DEVELOPMENT AND EVALUATION OF DOLPHIN: A TOOL FOR AUTOMATED TARGETED METABOLITE PROFILING USING 1D AND 2D ^{1}H-NMR DATA

2.1 ABSTRACT

One of the main challenges in nuclear magnetic resonance (NMR) metabolomics is to obtain valuable metabolic information from large datasets of raw NMR spectra in a high throughput, automated, and reproducible way.

Although new approaches try to identify and quantify in an automated or semi-automated way, they are still a long way from being truly automated. Most of the existing methods can identify and fit metabolites contained in their libraries, but they do not properly process the possible superposition of "unknown" signals not found in the libraries; these unmatched peaks represent a source of noise and often lead to inaccurate fitting results.

This chapter presents Dolphin, a software package born to achieve the first objective of this thesis, which is to develop and evaluate a set of algorithms able to find and quantify a set of target signals in a completely automated manner combining 1D (NOESY and CPMG) spectra with its own 2D JRES complementary spectra. Dolphin takes advantage of the 2D J-resolved NMR spectroscopy signal dispersion to avoid inconsistencies in signal position detection, enhancing the reliability and confidence in metabolite matching. Furthermore, in order to improve accuracy in quantification, Dolphin uses 2D NMR spectra to obtain additional information on all neighboring signals surrounding the target metabolite, considering not only those signals present in its library, but also those unknown signals visualized in the 2D spectrum. In this way, Dolphin carries out the automated quantification of a fixed target set of metabolites in a collection of experimental samples obtaining high throughput target metabolite profiling with minimum intervention from the operator.

In order to evaluate our approach, we used three different datasets to test Dolphin and compared them with the results from existing packages. The first dataset contained a pool mixture of standard compounds at various concentrations, the second dataset is formed by a collection of 24 rat liver aqueous extracts and the third dataset consisted in an spike-in experiment using human urine samples. We then compared Dolphin's performance against other NMR approaches, namely an integration based quantification method and the commercial package Chenomx NMR Suite 7.0, which is a line-fitting and library based approach.

2.2 DOLPHIN'S APPROACH AND METHODOLOGY

Dolphin utilizes a 1D line-shape fitting approach supported by a 2D complementary source of spectral data. There is a wide range of 2D NMR pulse sequences that address component separation in biological samples, such as COSY, TOCSY and HSQC[1], each one with its particular benefits when performing metabolite profiling. The common drawback of most of these 2D pulse sequences is that they require a lot of machine time to be acquired, a situation that can severely limit the throughput analysis of huge data sets. Two-dimensional J-resolved (2D JRES) NMR spectroscopy[2] can acquire a second-dimension spectrum of a metabolite mixture with relatively little overlapping of signals. This approach separates the effects of chemical shift and J-coupling into independent dimensions[3,4]. The drawback is that J-coupling can only be quantified for narrow signals, and so the use of Dolphin is limited to the automated profiling of low molecular weight metabolites in samples free from broad resonances. This is less problematic while working with aqueous extracts of biofluids and tissues, since these samples are free of broad resonances arising from high molecular weight macromolecules. Otherwise, if it is not possible to work with extracts, the dominating broad resonances and background signals can be reduced using different NMR pulse sequences and editing techniques based on NMR relaxation times[5].

Dolphin's flowchart is represented in Figure C2.1. The workflow starts by importing the NMR experiments and selecting the library to be used (see the next section). The line shape-fitting algorithm works optimally when we have a precise referencing between 1D and 2D JRES spectra. By default, this referencing uses the alpha-glucose doublet[6], but it can be changed to the TSP signal or any other peak of interest (in this work we used the default referencing option). In addition, during the data import process, the user can choose the data normalization method (selected peak, total area, or none). This step is especially important if quantification results have to be reported.

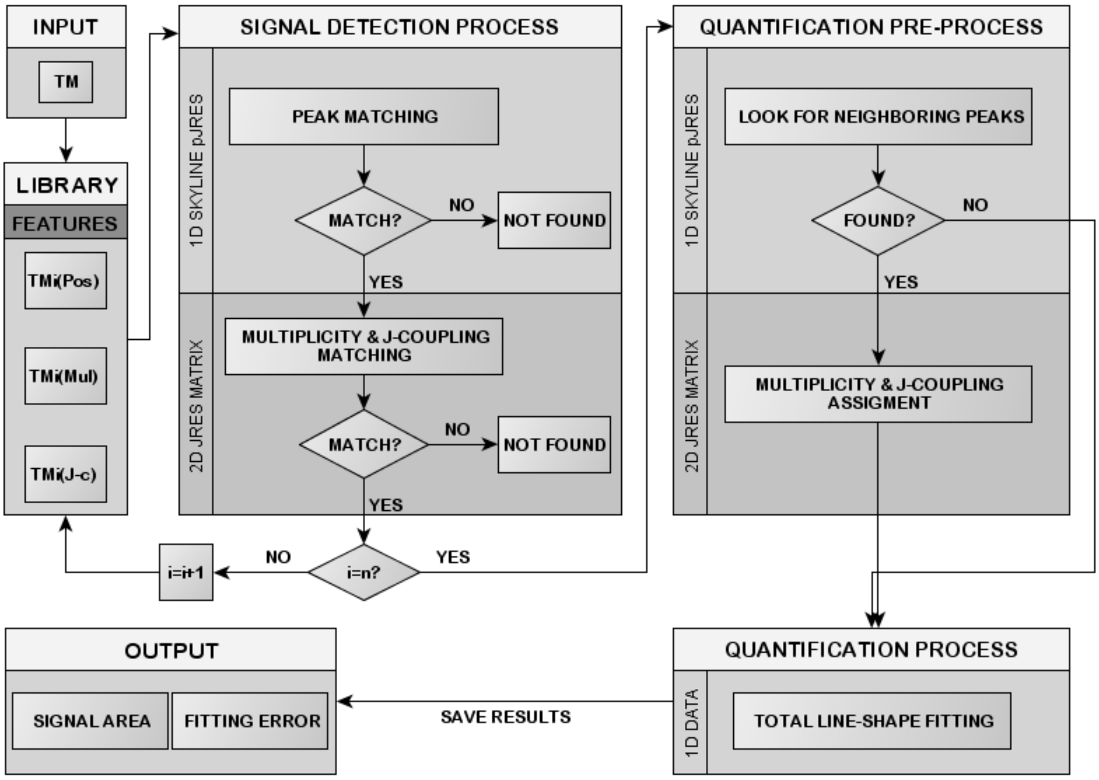


Figure C2.1. Dolphin's flowchart. Here we show the process of matching and quantification of a metabolite *TM* with *n* signals represented as *Mi*, where *i* value goes from *1* to *n*. It is important to note that quantification is only performed after a validated match.

After these initial preprocessing steps, the user can run the automated quantification of those metabolites selected in the user library. The process works in a two-stage approach detailed in the following steps:

1st: In a first step, Dolphin takes all signals that belong to that metabolite from the library; in some cases it could be one (as fumarate or formate) or more than one (as valine or leucine). Then, starting with one of them, the software goes to the signal detection step. (Note that the number of signals goes from i to n).

2nd: To detect a signal, the first step is to find a peak in the 1D JRES skyline projection (1D SKYLINE pJRES) at the ppm position annotated in the library for our target signal (TMi(pos)). If a peak is found, the software goes to the 2D matrix in order to check the multiplicity and j-coupling of that peak.

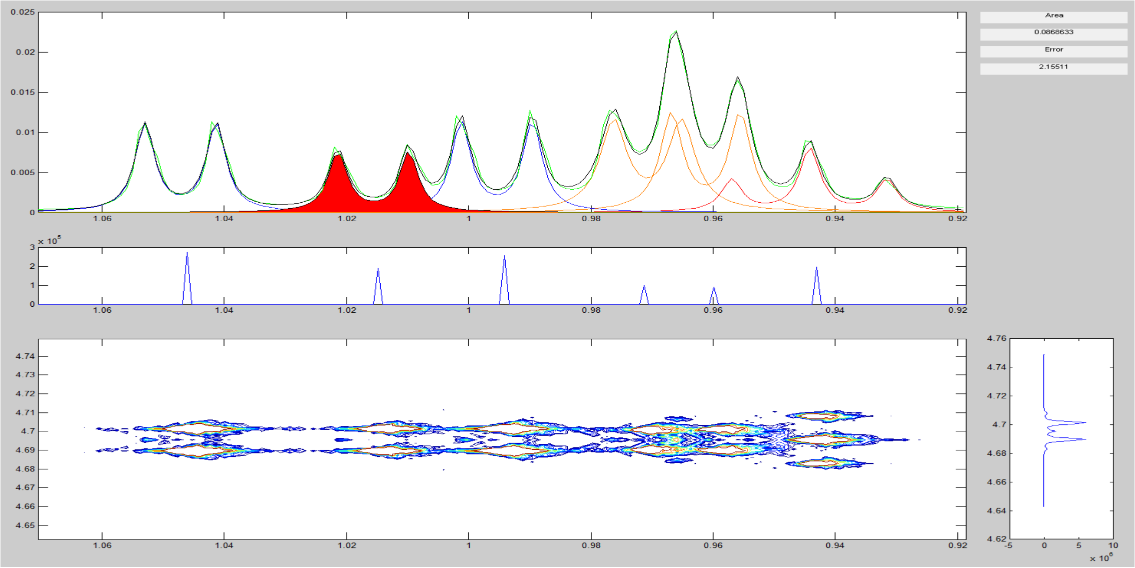
3rd: Each peak in the skyline projection has its multiplicity and j-coupling dispersed in the 2D matrix. To check both features, the software makes a vertical cut in the 2D matrix at the ppm position of our candidate peak. If the number of peaks in that cut (multiplicity) and the distance between them (j.-coupling) coincide with the multiplicity and j-coupling annotated in our library for our target signal (TMi(mul) and TMi(j-c)) the signal is considered as found.

4th: If all signals belonging to our target metabolite (which means that i=n) are found, the software takes one of them and moves to the quantification pre-process. The goal of this pre-process is the automated annotation of all the signals needed to perform a correct fitting.

5th: By creating a window around our target signal position, the software looks for neighbouring peaks. In case of no neighbouring peaks, the software goes directly to perform the line-shape fitting of our isolated signal in our 1D data using the information annotated in the library. If neighbouring peaks are found, the software makes a vertical cut in the 2D matrix and takes the multiplicity and j-coupling information for each neighbouring peak. All that information is then passed automatedly to the line-shape fitting algorithm in order to perform a realistic fitting, taking into account all surrounding signals present in the fitting region to distribute correctly the 1D data area between them.

When all the steps are finished, the program displays the measured area, the chemical shift of the signal used for quantification, and the fitting error for the region containing the signal of interest in the selected sample. This fitting error will give us information on whether the modeling of known and unknown metabolites has been successful or not, giving an indicative value of the confidence in measurement. Finally, the results for the complete dataset can be exported as an Excel file for further analysis.

As shown in Figure C2.2, the user can verify the fitting performance, both graphically and visually, . Dolphin's user interface plots four figures to show the relevant information for the selected sample. The upper graph contains the 1D spectrum plot of the fitting region; the middle segment contains the pJRES spectrum, where all signals are unified and displayed as a unique singlet, reducing congestions and overlapping; finally, the bottom figure contains the 2D JRES matrix contour plot, and the lower-right figure contains the sum projection of the vertical cut of the 2D JRES spectrum.



Area

0.0868633

Error

2.15511

Figure C2.2. Image of Dolphin's isoleucine quantification in STN1. We can see here how Dolphin uses its automated signal annotation in 2D data to deconvolve our target signal (second doublet from the left of the image, near 1.02 ppm) in the 1D spectra successfully. Dolphin detects 5 more signals that may affect our final goal, without previous knowledge of what those signals are, just taken them as unknowns.

All the functions are programmed under the matrix calculation platform MATLAB (ver. 7.5.0; The Mathworks, Inc., Natick, MA, USA). Dolphin is available by request.

2.3 AUTOMATED MATCHING OF TARGET METABOLITES USING 2D-JRES

In order to check for the presence of each target metabolite in a sample, Dolphin looks for a subset of its spectral pattern in the 2D-JRES NMR spectra. This characteristic subset of spectral resonances is stored in a proprietary library edited specifically for the kind of sample under study. The metabolite information about peak positions, signal multiplicity, and J-coupling values is included in the library, and is based on public domain databases such as the Human Metabolome Database (HMDB)[7], the Birmingham Metabolite Library[8], and the BioMagResBank (BMRB)[9], or commercial packages such as Chenomx NMR Suite (Chenomx Inc., Edmonton, Alberta, Canada) and Bioref AMIX (Bruker, GmbH, Silberstreifen, Rheinstetten, Germany).

The spectral pattern of a metabolite can be very complex and, in some cases, we can observe more than 10 different types of signals for a metabolite (e.g., glucose); however, most metabolites have a fixed pattern of signals that can be used to identify a metabolite in a mixture using just one, two, or three of these features since they are unique for that metabolite and have a fixed location in the spectrum.

The information on peak positions (or metabolite chemical shifts) is used in the first step of the metabolite match process. In this step, Dolphin matches the chemical shift of a signal annotated in the library with a peak in the processed 1D SEM-skyline pJRES spectrum (the 2D-JRES data was processed using a combined sine-bell-exponential (SEM) function, and a combination of SEM-skyline was applied for the 1D projection of the JRES data (pJRES)[10]). If a matching peak is present in the 2D JRES datasets, the algorithm tries to match the multiplicity and J-coupling of this signal annotated in the library with that of the vertical cut in the same peak position in the 2D JRES matrix. This is one of the main advantages of Dolphin, since all other approaches use identifications based only on 1D libraries that can lead to misidentifications.

2.4 AUTOMATED QUANTIFICATION OF PEAKS USING LINE-SHAPE FITTING

Dolphin continues with the quantification process only if location, multiplicity, and J coupling values match those annotated in the library for a given metabolite. The quantification step is carried out by measuring the area under one of the signals of the metabolite in question. The library defines which signal is going to be used in the peak quantification step, and a line-shape fitting algorithm based on the sum of Lorentzian functions modulated with different Gaussian proportions (Voigt profile with up to 10 % of Gaussian shape) is applied to discriminate the area of our target signal from other interferences sharing the same spectral region.

To achieve an optimal fitting, it is important to take into account the proportions between Lorentzians conforming the multiplet of each signal. Often, the presence of unknown signals surrounding or overlapping the signal of interest makes it difficult to accurately quantify a metabolite, especially in congested regions. In Dolphin, the fitting process is improved by using more detailed information from the position and the multiplicity of all the resonances surrounding the signal of interest as well. The processed 1D SEM-skyline pJRES spectrum gives the position of all the resonances surrounding our signal of interest, whereas the vertical cuts at these resonances indicate the multiplicity and J-coupling of all the signals surrounding the signal. In this way, the software is aware not only of the resonances included in the metabolite library but also of the presence of potential unknown interfering signals that are overlapping with the signal in question; taking this information into account results in a more accurate fit.

The implemented line-shape fitting approach is constrained using well-known intensity relationships within the singlets, doublets, and triplets. Multiplets, however, are fitted according to a sum of singlets. When the position, multiplicity, and Jcoupling values are calculated for all signals present in the region to fit, Dolphin is ready to proceed with the quantification step. Finally, both fitting errors and intensity values for all targeted metabolite quantifications are returned in an excel format file.

2.5 COMPARISON OF METHODS

We compared the results obtained with Dolphin's automated process, with two other approaches: one, a widely used integration-based quantification method (in our case we use that mode in AMIX) and second, using the commercial package Chenomx NMR Suite 7.0. Chenomx is a well-known commercial program that incorporates a large list of compounds in its library with the specific resonance signatures for each stored metabolite along the NMR spectra. Although there is a semi-automated fitting mode available, the software is mainly manually operated and requires expertise to adjust the metabolite peaks correctly. Therefore, the process can be considered as "user supervised."

Area integration is a standard quantification method, easy to implement with homemade tools and often included in commercial packages. It allows the quantification of metabolites by means of measuring the area in the spectrum associated with known metabolites. However, quantification of crowded regions using this approach might be challenging because of the potential unknown contribution from overlapping peaks and unidentified signals in the same segment.

2.5.1 Samples preparation and NMR data acquisition

As a proof of concept, we built a library of 15 hydrophilic metabolites that frequently appear in liver aqueous extracts (isoleucine, leucine, valine, β-hydroxybutyrate, alanine, acetate, succinate, creatine, glucose, fumarate, tyrosine, phenylalanine, uridine, formate, and adenosine). We tested Dolphin quantifying these target metabolites in two different real sample sets. The first test set comprised eight standard mixtures (labeled from STN1 to STN8) of pure metabolites, at different concentration ranges. The second dataset contained the NMR measurements of a set of 24 rat liver aqueous extracts (EXP1 to EXP24).

Three additional reference mixtures (REF1 to REF3) were also prepared for calibration purposes. Each one of the mixtures contained five of the library compounds, chosen to ensure a spectrum free of overlapped signals so that all three methods could obtain accurate quantification using their individual units of measure. In this manner, we were able to normalize each approach for comparison purposes.

Stock solutions for every single compound were prepared separately in 1 mL tubes with D_2O solution containing 0.6753 mM trimethylsilylpropionic acid (TSP), and the same D_2O solution was added to the final mixtures to obtain 700 μL for the NMR measurements. The liver extraction protocol was performed as described previously in this other study[11].

Additionally, we performed a spike-in experiment in human urine to test Dolphin's behavior in a more complex matrix. For that purpose, we prepared 5 aliquots of 300μL of human urine (SPK0 to SPK4) and added standard compounds to 4 of them (SPK1 to SPK4) at different concentration ranges. All the aliquots were filled to a final volume of 600μL with the buffer solution. We used a set of 5 compounds present in human urine: lactate, alanine, citrate, taurine and glycine. Finally, a tube with standard additions of the 5 compounds at known concentrations raised to 600μL with buffer solution was prepared for calibration purposes (SPKRef).

The 1D nuclear Overhauser effect spectroscopy with a spoil gradient (NOESY) was used to record 1D ^{1}H NMR spectra using a 600.2 MHz frequency Avance III-600 Bruker spectrometer

(Bruker, Germany) equipped with an inverse TCI 5 mm cryoprobe. A total of 256 transients were collected across a 12 kHz spectral width at 300 K into 64 k data points, and exponential line broadening of 0.3 Hz was applied before Fourier transformation. A recycling delay time of 8 s was applied between scans to ensure correct quantification. Two-dimensional J-RES spectra were acquired using the standard pulse sequence with pre-saturation and spoil gradients (jresgpprqf). Spectral widths of 7.5 kHz in the F2 dimension and 64 Hz in the F1 dimension were used; eight scans per increment and 32 increments were accumulated into 8 k data points. Zero filled into 16 k points and linear prediction to 128 increments was applied prior to Fourier transformation (FT) followed by tilting and symmetrisation. The acquired NMR spectra were phased and baseline-corrected using TopSpin software (ver. 2.1; Bruker BioSpin GmbH, Silberstreifen, Rheinstetten, Germany).

2.5.2 Equations

Chenomx gives quantification results in terms of the absolute concentration of the metabolite, but AMIX and Dolphin report their results in terms of non-normalized area units. Therefore, in order to compare the different values, we have to convert each method's own quantification units into absolute concentrations of the metabolite. Assuming zero error when quantifying clean and isolated signals in the reference mixtures, for each quantification signal of each metabolite we calculate a normalization factor as described in Eq. C2.1,

(Eq.C2.1) $$FM_m = \frac{RC_m}{QM_m}$$

where FM_m is the normalization factor of the method M for the metabolite m, QM_m are the quantification units of the method M quantifying the metabolite m, and RC_m is the real concentration of the metabolite m.

For the STN set, as we know the real concentration of each metabolite in the sample, we can evaluate the accuracy of each one of the methods (M) on every sample *(i)* for each one of the metabolites *(m)*. Eq. C2.2 gives the relative error of quantification of a metabolite m in a sample i (EM_{im}), (Eq.C2.2) $$EM_{im}(\%) = \frac{|(QM_{im} - C_{im})|}{C_{im}} \cdot 100$$

where *QM_{im}* is the measured concentration of the metabolite *m* in the sample *i* of the method *M*, and *C_{im}* is the real absolute concentration of the metabolite *m* in the STN sample *i*.

For the liver aqueous extracts set, the real concentration values in biological samples were not known. Therefore, we evaluated the differences of the three methods by comparing their results against each other in pairs, using the following relative difference estimation:

(Eq.C2.3) $$DM_{im}^{ab}(\%) = \frac{|(QM_{im}^{a} - QM_{im}^{b})|}{QM_{im}^{b}} \cdot 100$$

where *DM_{im} ab* is the relative difference quantifying a metabolite *m* in the sample *i* between method *a* and method *b*, *QM_{im} a* is the concentration of this metabolite *m* estimated by the method *a* in the sample *i*, and *QM_{im} b* is the concentration of the metabolite *m* estimated by the method *b* in the sample *i*.

For the urine spike-ins set, we measured the accuracy of each method subtracting the metabolite concentration quantified in the tube without additions (SPK0) from the values quantified in each spike-in tube (SPK1 to SPK4) and comparing the result to the values spiked-in. The quantification error, expressed in %, is given by (Eq.4):

(Eq.C2.4) $$EM_{im}(\%) = \frac{|(QM_{im} - QM_{0m} - C_{im})|}{C_{im}} \cdot 100$$

where *QM_{im}* is the measured concentration of the metabolite *m* in the sample *i* of the method *M, QM_{0m}* is the measured concentration of the metabolite *m* in the sample SPK0 and *C_{im}* is the real absolute concentration of the metabolite *m* added in the SPK sample *i*.

Final concentrations of each metabolite in the biological datasets were obtained converting each method values using the normalization factor described in Eq.C2.1.

2.6 RESULTS

2.6.1 Targeted profiling in standard pooled samples

All target metabolites were matched and quantified in the standard samples by all three approaches. In most cases, the quantifications were in good agreement with the known metabolite concentrations used to prepare the samples. When evaluating the results, it is important to distinguish those metabolites that can be quantified by an isolated signal from those that can be quantified by a signal that is present in overlapping region. For example, the region from 1.07 to 0.91 ppm, concentrates 11 peaks, corresponding to six signals (five doublets and 1 triplet) from three different metabolites (valine, isoleucine, and leucine). The fitting errors bymethods used here in this region are very different and depend on the metabolite. The results from Dolphin were in very good agreement with actual concentrations in the standard pooled samples, showing very little quantification error for all above metabolites, even for the case of overlapped signals.

The results are summarized in Table C2.1, showing a maximum mean relative error of quantification less than 10 % for all those metabolites that are quantified by an isolated signal. In this case, none of the methods is better than the other two in precision terms, but Dolphin has the advantage of full automation. Focusing on the quantification of areas within the overlapped signals, both Dolphin and Chenomx performed similarly with less than 10 % of relative error. However, the integration approach performed worse in the overlapped signals region, with mean relative errors between 10%and 20 %, with isoleucine as high as 46 %. See radar plot in Fig.C2.3 to graphically evaluate relative errors.

Method	**Chenomx**		**Integration**		**Dolphin**		**All**	
Statistic	**ME**	**SD**	**ME**	**SD**	**ME**	**SD**	**ME**	**SD**
Isoleucine	5	5	46	18	3	2	18	24
Leucine	9	5	12	7	8	4	10	2

Valine	5	2	5	7	4	2	5	1
DL-β-Hydroxybutirate	5	3	4	3	4	3	4	0
Alanine	6	4	14	5	5	3	8	5
Acetate	3	2	6	4	4	4	4	1
Succinate	7	4	20	8	6	6	11	7
Creatine	5	4	13	5	7	1	8	4
Glucose	8	7	7	7	4	3	6	2
Fumarate	3	1	9	7	3	3	5	3
Tyrosine	6	3	4	5	2	2	4	2
Phenylalanine	6	5	5	5	5	4	5	0
Uridine	6	6	5	4	2	1	5	2
Adenosine	3	3	7	6	5	4	5	2
Formate	10	6	6	7	4	2	6	3
All	6	4	11	7	4	3	7	4

Table C2.1. Quantification errors (in %) of the three methods in the STN mixtures. The table contains both the mean relative error (ME in %) and the standard deviation (SD in %) for every metabolite in the STN set.

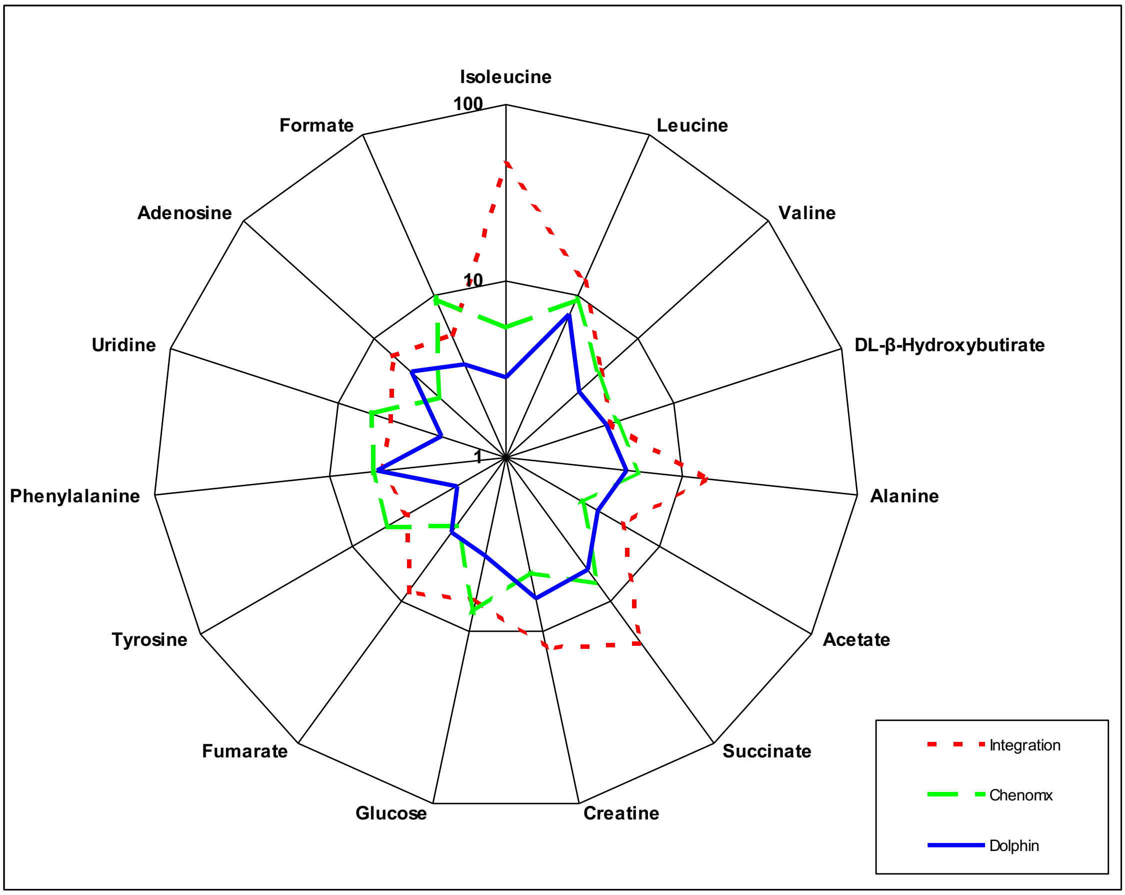


Figure C2.3. Radar plot of the quantification errors (in %) in the STN set. Here we can check graphically the performance of the three approaches quantifying metabolites in the STN samples. The graph shows the mean quantification error values (in %) of Chenomx, Dolphin and the standard integration approach on a logarithmic scale.

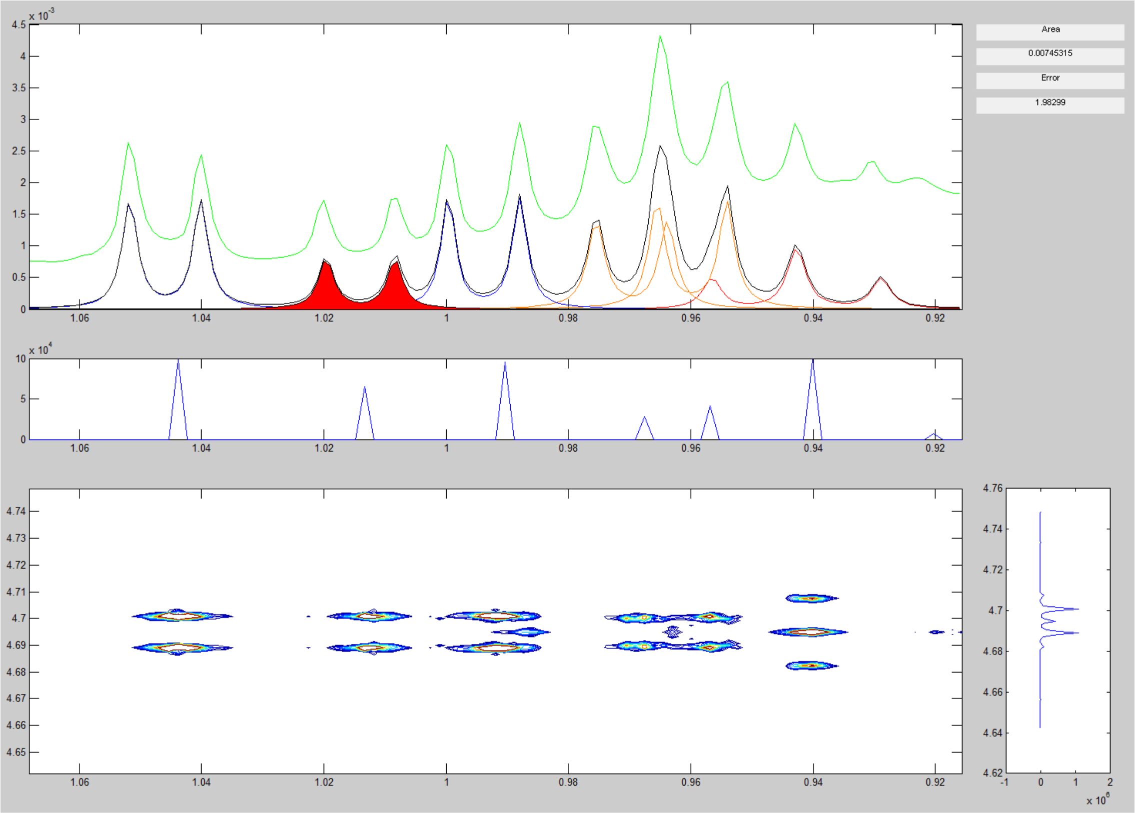
2.6.2 Targeted profiling in biological samples

The second dataset was composed by 24 rat liver aqueous extracts, (EXP1 to EXP24). In this case, we had no previous information about their metabolite content and, therefore, we did not know the correct identifications and quantification values. Absolute quantification errors could not be computed in this case and so we used Eq.C2.3 to perform comparisons by pairing methods. The main goal of this second study was to evaluate Dolphin's performance in biological samples, where overlapping and base lining problems are more pronounced.

All 15 target metabolites (those included in this Dolphin library) were matched and quantified by all three approaches in all the 24 biological samples. As expected, the quantification differences between methods (summarized in Table C2.2) were higher than in the case of the standard mixtures. There was a big difference comparing integration vs. the two fitting approaches, but the differences were not so important when comparing between the two fitting approaches. Although the mean difference in quantification of the whole set of metabolites between Chenomx and Dolphin is around 7 %, the same difference rises to 18 % and 23 % when Integration is compared with Dolphin and Chenomx, respectively. While only adenosine presents the same difference value between methods (4 %), 12 metabolites present lower values when the two fitting approaches are compared. In fact, only in the case of glucose, the difference value between the Integration approach and Dolphin (10 %) is lower than the difference between Dolphin and Chenomx (23 %). When taking Chenomx as reference against Integration, only eight of the whole set present quantification differences under 20 %, highlighting a difference of almost 80 % in the case of creatine. The comparisons between Dolphin and the Integration method followed a similar pattern, but with lower difference values, with the exception of phenylalanine. Figures C2.4 and C2.5 show Dolphin performance in biological samples, quantifying signals in crowded regions and baseline produced by macromolecule residuals.

Method	Chenomx vs Dolphin		Integration vs Dolphin		Chenomx vs Integration	
Statistic	**MD**	**SD**	**MD**	**SD**	**MD**	**SD**
Isoleucine	5	5	18	13	25	21
Leucine	8	8	29	15	35	27
Valine	5	3	9	5	14	9
DL-β-Hydroxybutirate	3	4	12	9	13	13
Alanine	3	5	16	5	16	3
Acetate	5	4	19	10	22	12
Succinate	6	5	6	6	11	8
Creatine	4	3	45	12	79	32
Glucose	23	5	10	5	36	7
Fumarate	7	6	13	11	14	7
Tyrosine	8	5	20	13	21	8
Phenylalanine	10	5	39	15	35	7
Uridine	4	3	7	6	8	8
Adenosine	4	2	4	3	4	3
Formate	8	5	19	9	19	4
All	7	5	18	9	23	11

Table C2.2. Quantification differences (in %) of each method vs. the others by pairs in the EXP dataset. The table here presents the mean relative difference (MD in %) and the standard deviation (SD in %) for every metabolite in the EXP set.



Area

0.00745315

Error

1.98299

Figure C2.4. Image of Dolphin's isoleucine quantification in EXP1. In this regions the metabolite pattern is quite similar to the generated in the STN mixtures, but with an important background signal coming from high weight molecular residuals. The results presented in the main article prove the accuracy of Dolphin's deconvoloution algorithm even in such conflict situations.

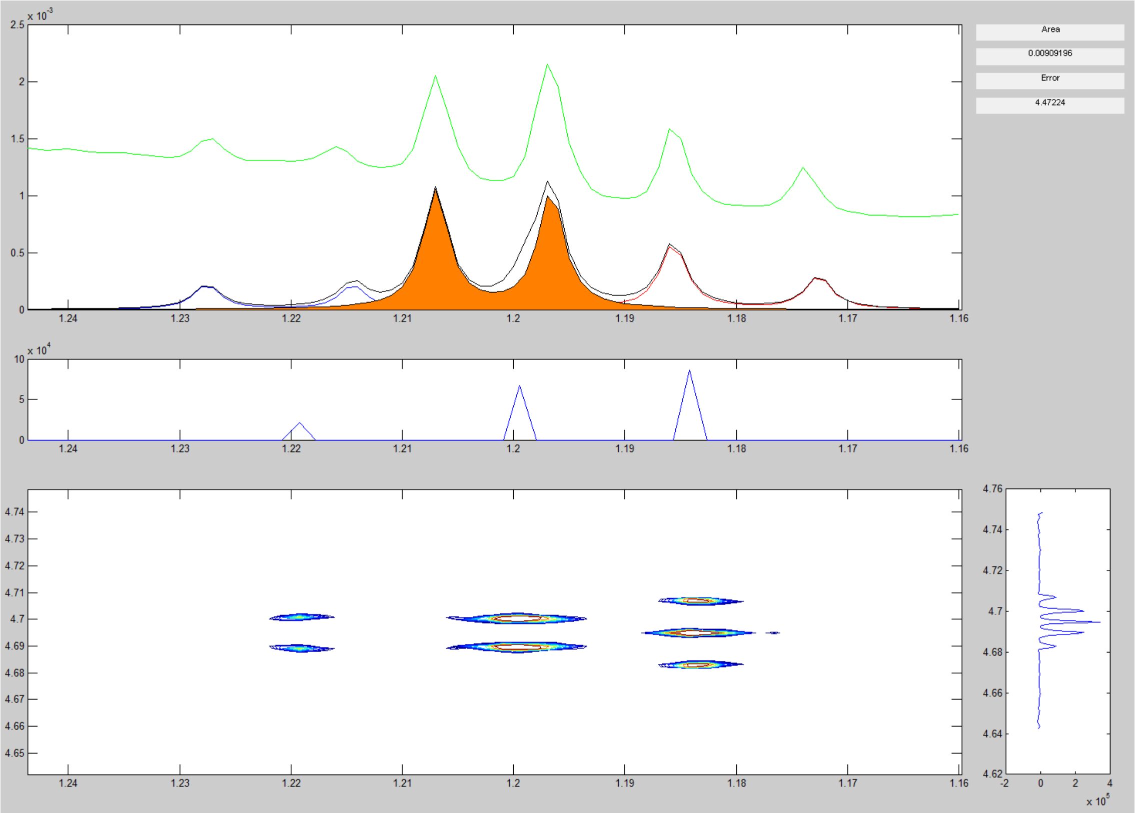


Figure C2.5. Image of Dolphin's β-hydroxybutyrate quantification in EXP1. This region of real biological sample contains unknown signals (not included in the library and not used in the STN mixtures) surrounding our target which are taken into account in order to perform a correct the line-shape fitting.

2.6.3 Targeted profiling in human urine spike-ins

When analyzing the results of this spike-in experiment, is important to take into account that all tubes contained human urine of the same stock sample, which produces minimal biological variation between samples. In such case, the integration approach is always affected by almost the same background signal, so the area difference between samples when integration is used is only attributable to the produced by the increment of the compound added, which explains why integration is the best approach in most of cases (Table C2.3 and Figure C2.6). However, if we focus on the absolute concentrations calculated in SPK0 (Table C2.4), we see that integration is always giving higher values than Chenomx and Dolphin, which means that the error of quantification of the integration approach is being produced since the biological sample without additions and is being masked by this analysis.

Focusing now in Dolphin's behavior, we can see how the results are in very good agreement with those present by Chenomx manual fitting. Chenomx gives better results only when quantifying Glycine, and with a minimal error difference of 1%. We want to highlight Dolphin's accuracy when catching and quantifying citrate through its shifting doublets with roof effect. Figure C2.7 shows Dolphin performance in urine samples.

Method	**Chenomx**		**Integration**		**Dolphin**		**All**	
Statistic	**ME**	**SD**	**ME**	**SD**	**ME**	**SD**	**ME**	**SD**
Alanine	10	5	6	3	4	2	7	3
Citrate	7	6	2	2	6	5	5	3
Glycine	10	4	6	2	11	3	9	3
Lactate	12	6	4	2	11	8	9	4
Taurine	6	6	4	4	6	5	5	1
All	9	5	4	3	8	5	7	2

Table C2.3. This table shows the relative errors of quantification (in %) of the three methods in the SPK additions. The table contains both the mean relative error (ME in %) and the standard deviation (SD in %) for every metabolite in the SPK set.

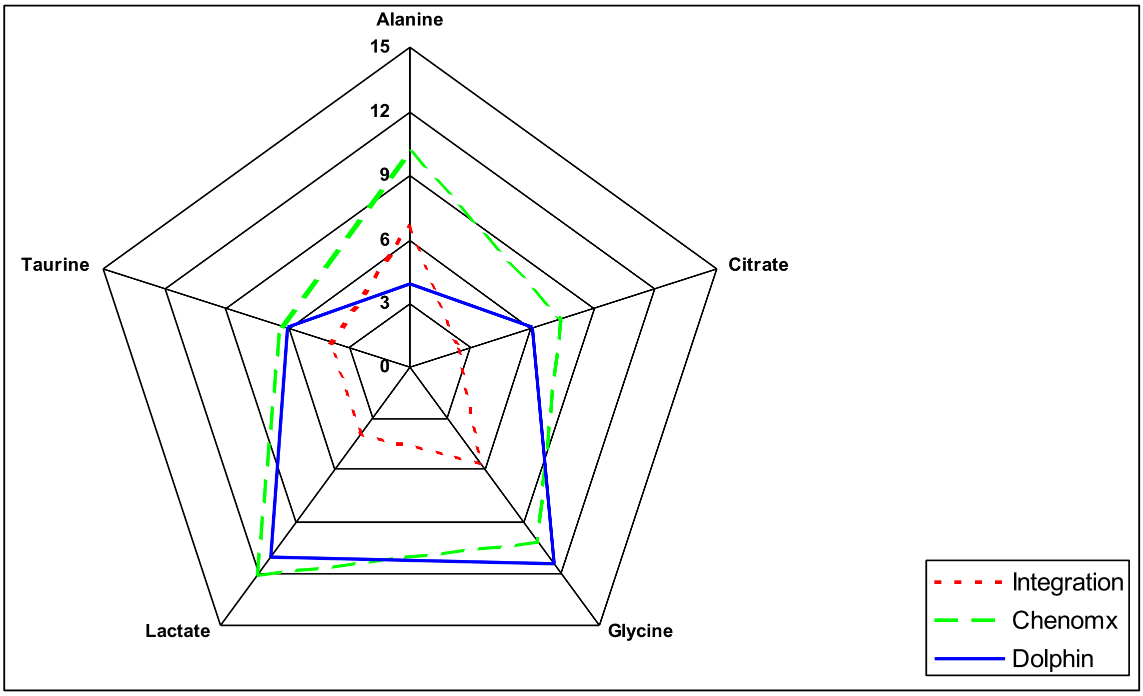


Figure C2.6. Radar plot of the quantification errors (in %) in the SPK set. Here we can check graphically the performance of the three approaches quantifying metabolites in the SPK samples. The graph shows the mean quantification error values (in %) of Chenomx, Dolphin and the standard integration approach on a lineal scale.

Tube	SPK0				SPK1			
Method	**Spiked**	**Chenomx**	**Integration**	**Dolphin**	**Spiked**	**Chenomx**	**Integration**	**Dolphin**
Alanine	0	0.043	0.093	0.047	0.051	0.100	0.150	0.099
Citrate	0	0.450	0.671	0.423	0.163	0.586	0.838	0.561
Glycine	0	0.168	0.482	0.166	0.225	0.380	0.736	0.379
Lactate	0	0.025	0.081	0.024	0.050	0.078	0.132	0.074
Taurine	0	0.147	0.368	0.145	0.213	0.360	0.593	0.353

Tube	SPK2				SPK3			
Method	**Spiked**	**Chenomx**	**Integration**	**Dolphin**	**Spiked**	**Chenomx**	**Integration**	**Dolphin**
Alanine	0.062	0.111	0.166	0.110	0.072	0.113	0.168	0.115
Citrate	0.195	0.642	0.903	0.636	0.228	0.603	0.896	0.595
Glycine	0.270	0.426	0.793	0.419	0.315	0.454	0.821	0.454
Lactate	0.060	0.084	0.147	0.083	0.070	0.091	0.149	0.089
Taurine	0.255	0.426	0.658	0.402	0.298	0.444	0.675	0.436

Tube	SPK4			
Method	**Spiked**	**Chenomx**	**Integration**	**Dolphin**
Alanine	0.082	0.121	0.178	0.127
Citrate	0.260	0.681	0.911	0.687
Glycine	0.360	0.515	0.857	0.508
Lactate	0.080	0.101	0.158	0.101
Taurine	0.340	0.501	0.708	0.506

Table C2.4. This table shows the concentrations of all compounds in each tube, both the spiked-in and the calculated by each method. The arbitrary units of Integration and Dolphin were converted to absolute concentrations units according to Eq.1 in section 2.4 Comparison of methods using the reference signals quantified in SPKRef.

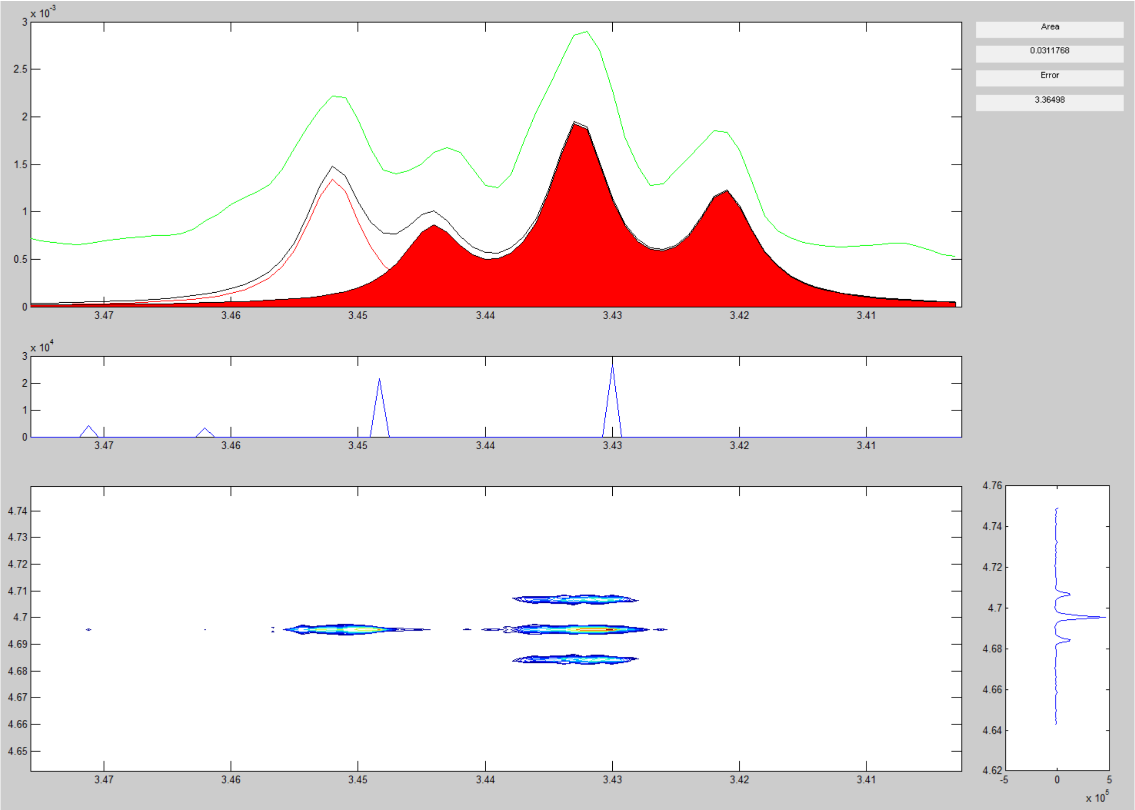


Figure C2.7. Image of Dolphin’s performance in taurine’s quantification in SPK0 sample. Once again we can see the efficiency of Dolphin’strategy, which is able to automatedly add unknown signals surrounding our target in order to improve the fitting.

2.7 DISCUSSION

The differences in quantification accuracies for the three options are directly related to the approaches used and in the particular signal characteristics in the regions of interest.

Area integration is a good approach when quantifying isolated peaks with intensities clearly higher than background noise; on the other hand, this approach is less accurate when used in congested areas or when weak signals need to be measured. These considerations explain the method accuracy in the case of metabolites that can be quantified using isolated signals. In fact, valine and phenylalanine present the lowest errors in quantification for the STN samples using the metabolite integration approach (see Table C2.1).

Otherwise, and due to the shape constraints of its spectral library, Chenomx is the tool that usually presents higher quantifying errors for these metabolites. Chenomx uses reference deconvolutions to adjust the shim for the spectra in its library for all metabolites, but it is only effective in removing line-shape distortions that affect all signals in the spectrum equally; in addition, there are other factors affecting the shape of he signals[12,13], especially if they are unusually intense and, consequently, the Chenomx library signals rarely fit accurately such peaks present in the real spectrum. This lack of shape flexibility makes it difficult to achieve a precise quantification and increases error.

Dolphin, on the other hand, quantifies using a line-shape fitting algorithm based on the addition of Lorentzian signals modulated with a different Gaussian proportion (to correct such problems associated with magnetic field inhomogeneity while acquiring data). The results show that this option works better in metabolite quantification; in fact, Dolphin presents even lower error values compared with the two software tools tested for most of the cases applied (see Table C2.1).

The problem is completely different in the case of areas of the spectra with moderate to high signal congestion. It is in these regions where both Dolphin and Chenomx software present better results than the integration-based approach. Chenomx, through its manually adjusted fit, allows a trained user to solve the fitting puzzle really well, even though its library signals rarely fit in an accurate manner for those peaks present in the real spectrum. However, this manual approach is time-consuming and prone to subjective quantification. For example, in the branched chain amino acids area (from 1.07 to 0.91 ppm) the program presents a list of 21 possible compounds that may be adjusted for each sample. On the other hand, Dolphin's line-shape fitting approach does not require manual supervision since all the necessary information to deconvolve the signal of interest is obtained from the 2D JRES analysis. Figure C2.1 presents the Dolphin analysis of the branched chain amino acids region of one of the STN mixtures (Figures C2.4, C2.5 and C2.7 show Dolphin's performance in the EXP and the SPK samples). In this example, we fit the doublet near 1.00 ppm coming from isoleucine to quantify this metabolite concentration, and the area that corresponds to the signal is adjusted taking into account all neighboring signals as unknowns (five doublets and one triplet, which were adjusted automatedly with our line-shape fitting algorithm). Therefore, Dolphin was more accurate in these situations, even without human intervention. This enables the user to calculate the concentrations of numerous metabolites in an automated, high-throughput, and accurate manner.

The quantifications carried out by the peak integration method produced several inaccuracies. The main problem of this approach was associated with the integration technique, which measures the area under the curve assigned to a metabolite, ignoring the interferences generated by signals from adjacent overlapped metabolites. The quantification of isoleucine in the STN set (see Table C2.1) is a representative example of such a problem; the measured value is clearly affected by the amount of valine, which has a signal overlapped with our target signal (Figure C2.8).

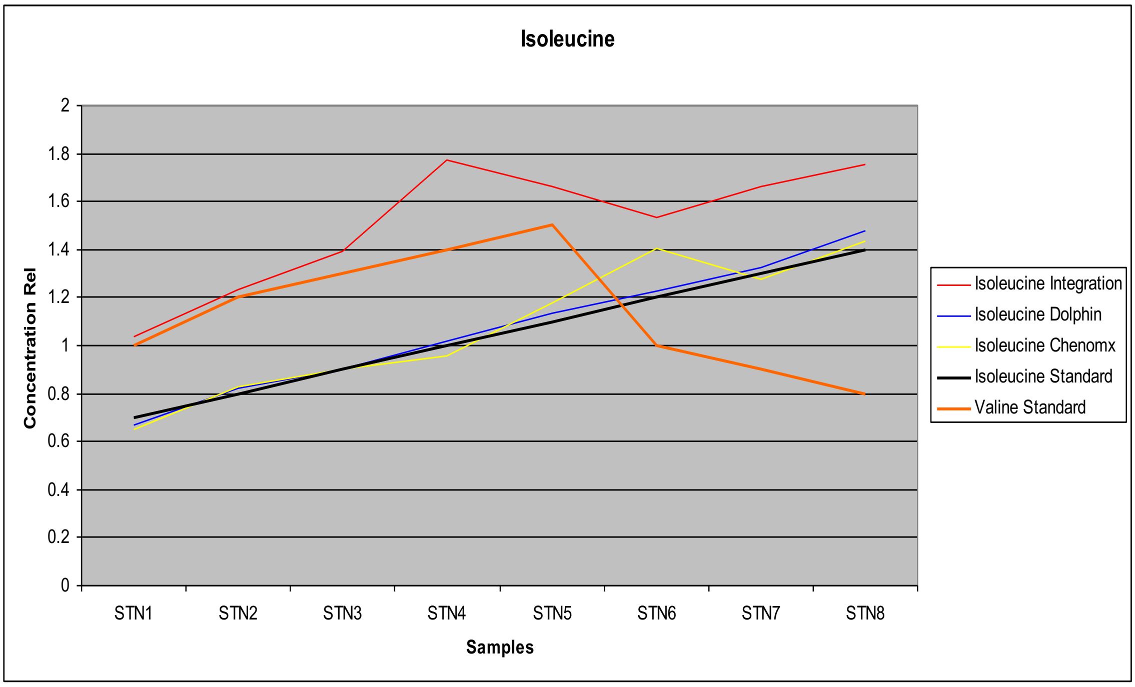


Figure C2.8. Isoleucine quantification behaviour of the three approaches in the STN set. Here we can see how Integration of isoleucine signal, which is moderately overlapped with a valine signal, follows up an over estimated relative concentration against the REF value. While this overestimated trend seems to follow a good correlation with the real Standard, we can observe how the high concentration of valine in STN4 and STN5 are distorting this trend. Finally, when the concentration of valine decreases, isoleucine recovers the original trend, always with an over estimation of its relative concentration.

It is important to note that Dolphin performed well while quantifying metabolites in the STN dataset compared with the other two methods. This fact highlights the combined accuracy of the Chenomx line-fitting approach for congested signals regions and the efficient area integration for isolated peaks that are combined in the Dolphin software package.

Dolphin's quantification accuracy is directly related to the quality of the spectra. In complex biological matrices, especially in the presence of inhomogeneities in the samples, some fitting errors arise, mostly produced by spectral shapes that are quite far from being Lorentzian (the addition of 10 % of Gaussian shape does not resolve this issue totally).

2.8 CONCLUSIONS

This work presents a new NMR data processing tool (Dolphin) that assists with identification and quantification of target metabolites in an automated manner once datasets are imported with the import parameters manually entered by the user. The tool has been compared against two popular NMR metabolite quantification approaches. Dolphin quantifications were more accurate compared with those obtained with bin integration approaches, and in good agreement with the manual line-shape fitting solutions. Our software is based on both the 1D and 2D spectra for the same sample, matching data to a library for metabolite identification, and automatedly fitting peaks for quantification purposes. One advantage of such workflow is the increase in identification confidence. Moreover, this approach takes into account all neighboring signals, even if they are unknown, which leads to a more accurate quantification. This is especially relevant and useful in very congested regions, where current automated tools and methods mostly fail. The maximum number of automatedly identifiable and quantifiable metabolites is strongly related to the 2D-JRES spectra resolution and the type of biofluid matrix under study. Future versions of Dolphin will incorporate better filters and spectral processing routines in order to enlarge the target metabolite list without increasing the processing time requirements.

2.9 ACKNOWLEDGEMENTS

We acknowledge CIBER de Diabetes y Enfermedades Metabólicas, an initiative of ISCIII (Ministerio de Ciencia e Innovación) for partially funding this work.

2.10 REFERENCES

1. Bingol, K. & Brüschweiler, R. Multidimensional approaches to NMR-based metabolomics. *Anal. Chem.* **86,** 47–57 (2014).

2. Aue, W. P., Karhan, J. & Ernst, R. R. Homonuclear broad band decoupling and two-dimensional J-resolved NMR spectroscopy. *J. Chem. Phys.* **64,** 4226–4227 (1976).

3. Ludwig, C. & Viant, M. R. Two-dimensional J-resolved NMR spectroscopy: review of a key methodology in the metabolomics toolbox. *Phytochem. Anal.* **21,** 22–32 (2010).

4. Huang, Y., Cai, S., Zhang, Z. & Chen, Z. High-resolution two-dimensional J-resolved NMR spectroscopy for biological systems. *Biophys. J.* **106,** 2061–70 (2014).

5. Tang, H., Wang, Y., Nicholson, J. K. & Lindon, J. C. Use of relaxation-edited one-dimensional and two dimensional nuclear magnetic resonance spectroscopy to improve detection of small metabolites in blood plasma. *Anal. Biochem.* **325,** 260–72 (2004).

6. Pearce, J. T. M. *et al.* Robust algorithms for automated chemical shift calibration of 1D 1H NMR spectra of blood serum. *Anal. Chem.* **80,** 7158–62 (2008).

7. Wishart, D. S. *et al.* HMDB 3.0--The Human Metabolome Database in 2013. *Nucleic Acids Res.* **41,** D801–7 (2013).

8. Ludwig, C. *et al.* Birmingham Metabolite Library: a publicly accessible database of 1-D 1H and 2-D 1H J-resolved NMR spectra of authentic metabolite standards (BML-NMR). *Metabolomics* **8,** 8–18 (2011).

9. Doreleijers, J. F. *et al.* BioMagResBank database with sets of experimental NMR constraints corresponding to the structures of over 1400 biomolecules deposited in the

Protein Data Bank. *J. Biomol. NMR* **26,** 139–46 (2003).

10. Tiziani, S., Lodi, A., Ludwig, C., Parsons, H. M. & Viant, M. R. Effects of the application of different window functions and projection methods on processing of 1H J-resolved nuclear magnetic resonance spectra for metabolomics. *Anal. Chim. Acta* **610,** 80–8 (2008).

11. Vinaixa, M. *et al.* Metabolomic assessment of the effect of dietary cholesterol in the progressive development of fatty liver disease. *J. Proteome Res.* **9,** 2527–38 (2010).

12. Alam, T. M. *et al.* Investigation of chemometric instrumental transfer methods for high-resolution NMR. *Anal. Chem.* **81,** 4433–43 (2009).

13. Saude, E. J., Slupsky, C. M. & Sykes, B. D. Optimization of NMR analysis of biological fluids for quantitative accuracy. *Metabolomics* **2,** 113–123 (2006).

CHAPTER III

DEVELOPMENT AND APPLICATIONS OF WHALE: OPTIMIZING METABOLITE PROFILING IN 1D ^{1}H-NMR SPECTRA AND APPLYING IT TO TECHNICAL AND CLINICAL STUDIES

3.1 WHALE: A PACKAGE COMBINING FINGERPRINTING AND TARGETED METABOLITE PROFILING APPROACHES TO IMPROVE THE EXTRACTION OF METABOLIC INFORMATION IN ^{1}H-NMR SPECTRA.

3.1.1 Abstract

The complexity of NMR spectra varies a lot depending on the biofluid matrix, the sample treatment, the acquisition parameters, and several other factors. In fact, fitting the spectral processing parameters to the dataset complexity is an optional strategy to save time in the spectral analysis. Moreover, depending on the final goal of the study, different approaches can be applied to analyse samples in a dataset. An untargeted analysis of the sample fingerprint can be enough to discriminate groups into a study, or very useful for matching spectra with metadata values and/or finding 'hot spots' in an exploratory analysis. Otherwise, a targeted analysis of the metabolite profile can be performed in order to find molecular patterns of the samples under study.

With the aim to cover the widest range of NMR spectral analysis, we developed and evaluated a set of algorithms that have become new functions in a new package called Whale. Actually, these new functions cover the second and third objectives of this thesis, which consist in adding fingerprinting approaches, importing reference spectra from public databases and allowing users to interact with the spectra and to use different automated quantification modes under a user-friendly an intuitive GUI. All these improvements allow users to perform the most reliable NMR profiling minimizing user-subjectivities but avoiding black-box processes at the same time in a useful and versatile tool.

3.1.2 Package overview

Whale has born as an upgrade of Dolphin with the aim to cover a wide range of challenges that users may face when analyzing NMR datasets in metabolomics studies. Mainly, it incorporates some fingerprint options that allow users to explore data before starting more complex analysis and a new strategy for metabolite profiling based on user interaction for signal identification with intelligent assistance and different modes of automated quantification along samples. Moreover, it gives the user the option to import reference spectra from public databases and to graphically check them with the spectra of the dataset under study. The aim of the tool is to provide different ways to obtain reliable results depending on their goal, their experience and their necessities. The package has been re-designed to work with NMR datasets of any kind of biofluid using Bruker data as input without needing 2D acquisitions to perform the profiling. As in the previous version, all functions of the package have been programmed using MATLAB (ref matlab) and compiled to work as standalone application.

3.1.3 Inputs and outputs

Inputs and outputs are key elements in NMR spectral analysis tools for metabolomics studies. Inputs are very important in order to establish data import parameters that will define the future results, and outputs are as important as inputs because they need to be informative, easy to interpret and well structured, since in most cases they will be used as inputs for posterior statistical analysis.

Whale uses one input parameters file with some values that need to be filled. The file format is detailed in the appendix section 1, and includes some paths, information about importing parameters such as spectral referencing and normalization and two quality thresholds that will be very important for the analysis. These two thresholds correspond to the minimum ratio of 'area of the signal of interest / area of the total spectrum in that location' and the maximum fitting error calculated as the difference between the real spectrum and the generated by the model. They are directly linked to the output files because all the quantifications that do not pass these thresholds will generate a figure of the quantification performance in a folder called ‘Plots2Check’. It will allow users to rapidly check those quantifications that could not be reliable, detect the mistake and re-run the analysis only in those cases.

The package generates different outputs: the already explained ‘Plots2Check’, a file with all the quantification values, the exact ppm position where signals have been quantified, the fitting error and the ratio 'area of the signal of interest / area of the total spectrum in that location' of each case will also been generated. The output format allows easy interpretation of the reliability of the results and is easily adaptable to be used as input for posterior statistical analysis in Matlab, R, SPSS, and more.

3.1.4 Fingerprint options

One of the most common ways to approach NMR metabolomics data is to apply fingerprinting analysis of the samples in a dataset. Fingerprinting allows users to obtain valuable information of a dataset rapidly before starting more complex analysis. We have incorporated a fingerprint block into the package with three different functions to give users the possibility to approach data in an explorative way.

3.1.4.1 Correlate sample to region

This function allows users to select a region of the spectrum and run a correlation test of that region against the rest of spectral points. This correlation test calculates a correlation factor from 0 to 1 (R2) and a p-value of the as a result, the fingerprinting panel of the package will show a spectrum with the mean intensities of the dataset for each ppm, where spectral points will be colored depending on their correlation factor from blue (R^2=0) to red (R^2=1) and those with correlation factors higher than 0.75 and p-values lower than 0.05 will be highlighted with grey circles. This kind of correlation is very useful for finding metabolite signatures, since under normal conditions, the intensity variation of a resonance of a determined compound along samples is in totally agreement with the intensity variation of the rest of resonances that belong to the same compound. Moreover, users can find correlations with other spectral regions, which can indicate covariation of compounds. Is important to note that signal overlap, which is inherent in NMR spectra of complex biofluids, can mask some correlations producing false negatives. Figure C3.1 shows an example of this function applied to the creatine singlet at 3.04 ppm, where the singlet at 3.93 ppm is highly correlated.

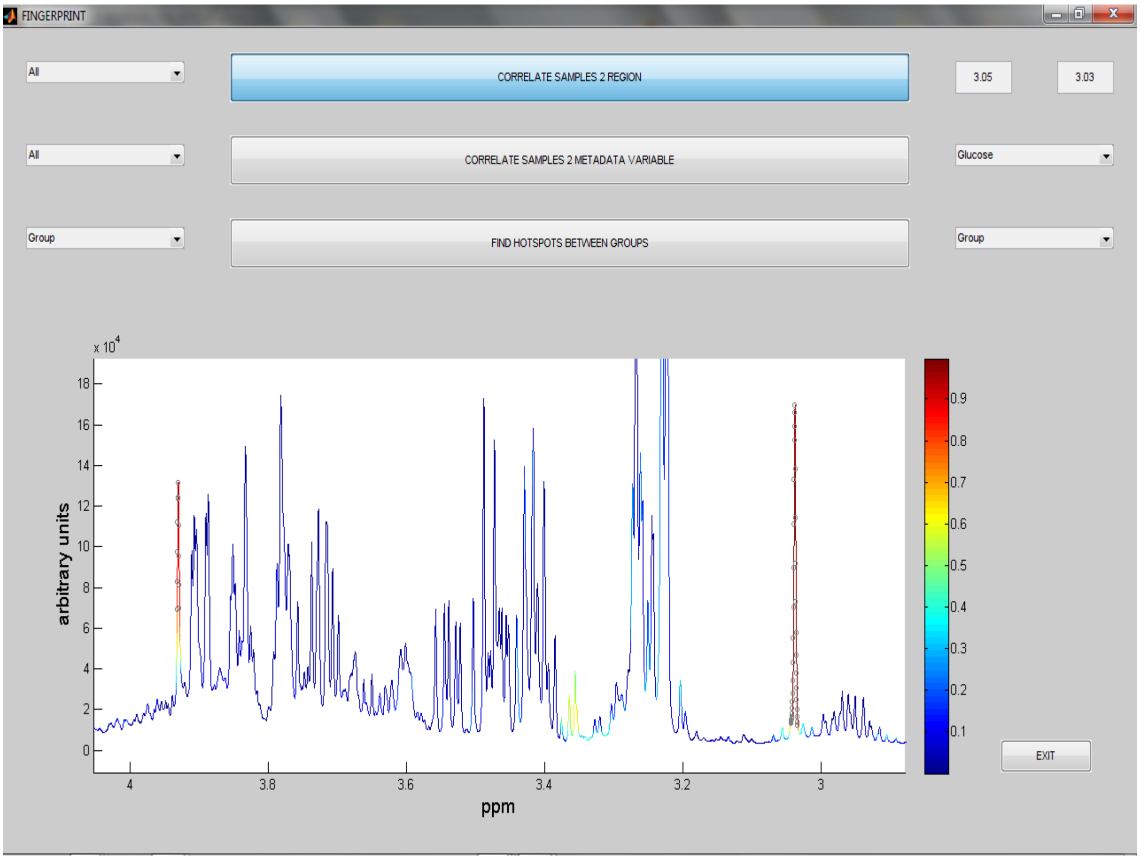


Figure C3.1. An example of the function 'correlate samples to region'. In this example we can see how the region selected by the user from 3.03 to 3.05 ppm presents different correlation coefficients with all the spectral points in the dataset. In this case, the region corresponds to the peak of creatine, which is highly and significantly correlated with itself and another peak at 3.93 ppm. This function of the package is suggesting that these two peaks could be resonances of the same compound, and indeed they are.

3.1.4.2 Correlate sample to metadata variable

This function is very similar to the previously explained, but in this case each spectral point will be correlated with the values of a metadata variable instead of a selected region. Finding a correlation between a metadata variable and spectral regions may signify that one or some metabolites present in those regions could be related to that variable, which could be a starting point to focus a posterior targeted metabolite profile analysis. This function can also be applied as quality control between the spectra and the metadata, since in some studies the metadata contain metabolite concentrations determined by biochemical parameters of compounds that can be easily profiled in NMR spectra such as glucose. Good correlations between glucose concentrations in metadata and the doublet at 5.22 ppm (and several peaks between 3-4 ppm) is an indicator of good agreement between the NMR spectra and the metadata of a dataset (Figure C3.2).

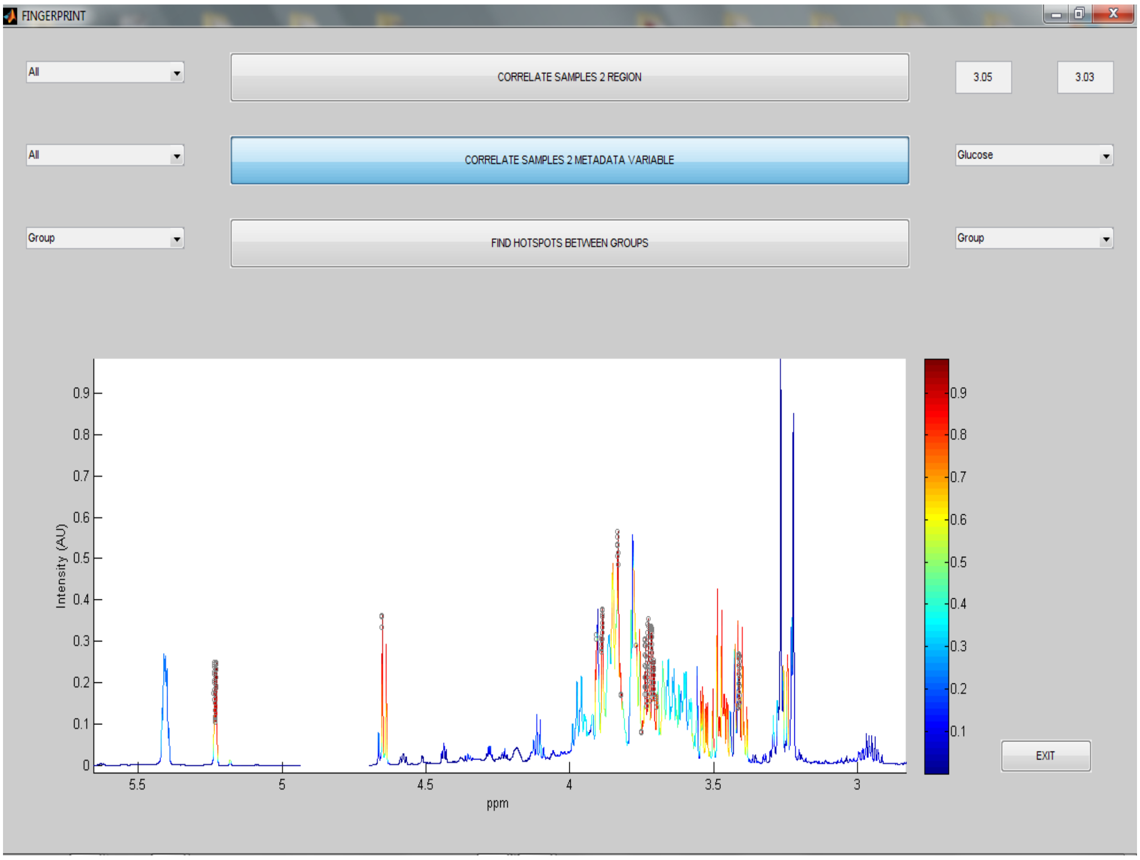


Figure C3.2. An example of the function 'correlate samples to metadata variable'. In this case we looked for spectral correlations with glucose concentrations determined by biochemical parameters. High and significant correlations can be found in a lot of points between 3 and 4 ppm, where most of glucose resonances are placed. As expected, the higher and most significant correlations are focused in the doublet at 5.22 ppm, due to be an isolated signal. This result suggests that there is a very good agreement between the samples labeling and their corresponding NMR spectra.

3.1.4.3 Find hotspots between groups

Finding differences in metabolite concentrations between two groups under study is usually the main objective in NMR metabolomics studies. Even if for obtaining reliable differences in spectra of complex biofluids spectral deconvolution solutions are needed to avoid signal overlap distortions, a fast analysis of intensity differences with all the spectral points can give valuable information in some cases. This kind of screening analysis helps users to focus the posterior deconvolution analysis in regions with more variability in signal intensity. In this case, the algorithm will perform a Wilcoxon rank sum test and a fold change test with all the spectral points to obtain the significance of the differences found between the two groups. As a result, the panel will show a spectrum where the intensity of each point will be the mean intensity of the whole dataset multiplied by the fold change factor obtained when comparing the two groups and colored by its significance from blue (p=1) to red (p=0). According to that, reddish regions will be considered as 'hot spots' to focus further analysis.

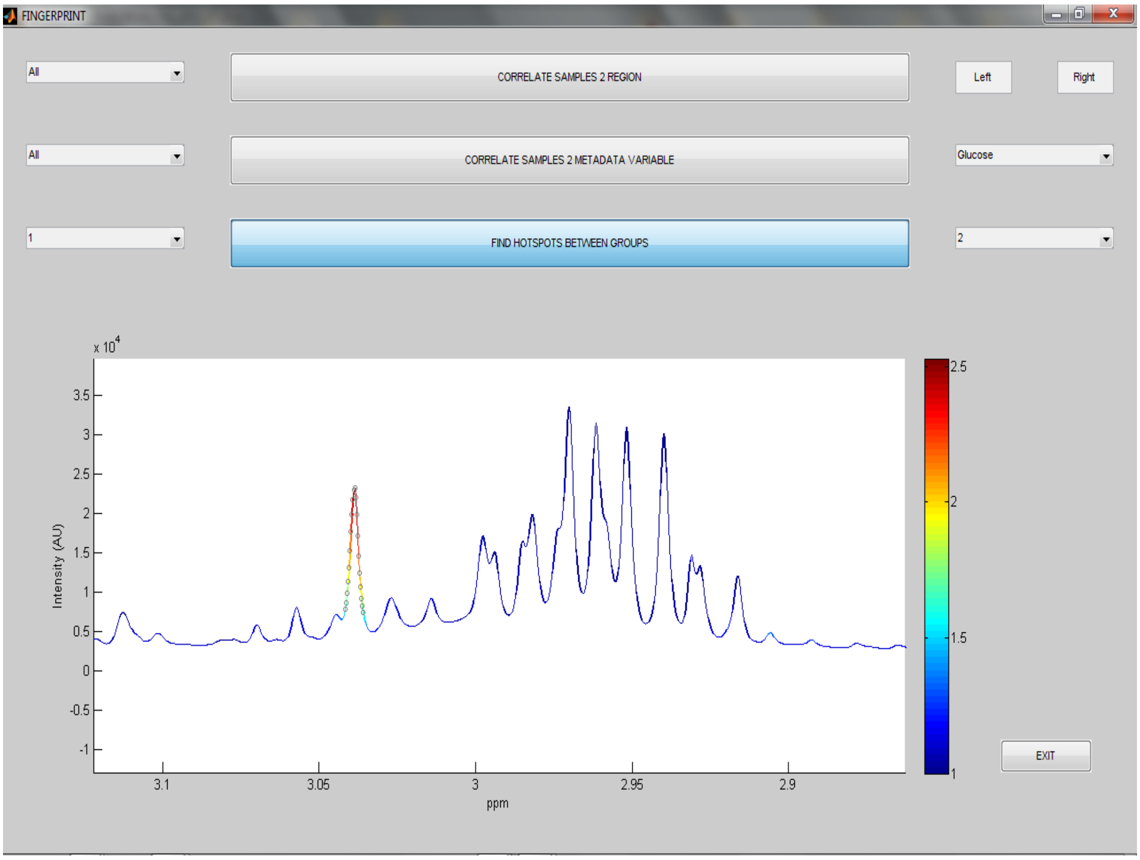


Figure C3.3. An example of the function ‘find hotspots between groups’. In this case we can clearly see the difference between peaks that presented a high fold change ratio and a statistically significant p-value in comparison with those that did not. Gathering the results of this test with those of the showed above in Figure C3.1 users can foretell that differences in creatinine levels between the two groups will be a possible result of the study.

3.1.5 Profiling options

Metabolite profiling techniques are more much complex than fingerprinting techniques, since metabolite identification and quantification is severely compromised by the huge amount of signals present in NMR spectra of complex biofluids. Moreover, depending on the biofluid matrix and its treatment, the processes of identification and quantification will follow different strategies in order to optimize the metabolite profiling. For example, in total serum and plasma samples there is very low variation in the number of metabolites that can be profiled, and the useful resonances where a metabolite can be identified and quantified are in most cases the same and do not present shift problems. So in these cases, metabolite profiling can be easily automated with a good deconvolution algorithm able to discriminate between the signals of interest and neighboring and background signals coming from other metabolites and macromolecule residuals. However, when we talk about urine samples, the problems are totally different. In this case, there exist a high variation of the number of metabolites that can be profiled depending on the cases under study, where factors such as life style and diet will probably influence the metabolic profile of the sample. Moreover, due to the sample pH, ionic strength and density, signals may shift a lot between samples. In this case, an agreement between automation and user interaction will be the optimal strategy to use. This agreement between user interactivity and automation will be also the best way to perform the most extended metabolite profile in NMR spectra of tissues and cell extracts. Even if a big subset of metabolites and resonances can be easily profiled in these cases, some extra resonances can be added to the analysis if the quality of the sample allows it.

Taking into account all the possible ways to optimize metabolite profiling in any kind of biofluid we have equipped Whale with 3 different blocks. One of this blocks is a supporting tool for identification purposes. It is called 'Repository', and it launches a panel where users can plot reference spectra of pure compounds that can be obtained in public databases such as the BMRB and the Birmingham database and compare them with their samples. The most important block is called 'ROIs testing'. It launches a panel where users can edit signal parameters in order to graphically check the performance of the four automated quantification modes that the program offers. The third block is called 'Auto run', and it performs a totally automated metabolite profiling along all samples of the dataset.

3.1.5.1 Repository

The repository block allows the user to compare each spectrum of the dataset with reference spectra of pure metabolites. It offers the possibility of importing different types of NMR acquisitions (1D and 2D) and generating artificial spectra by combining reference compounds. Figure C3.4 shows the repository panel in a comparison between a sample spectrum and a reference spectrum produced by the combination of three aminoacids. This block is very useful to discriminate resonances with the aim to unequivocally identify a metabolite.

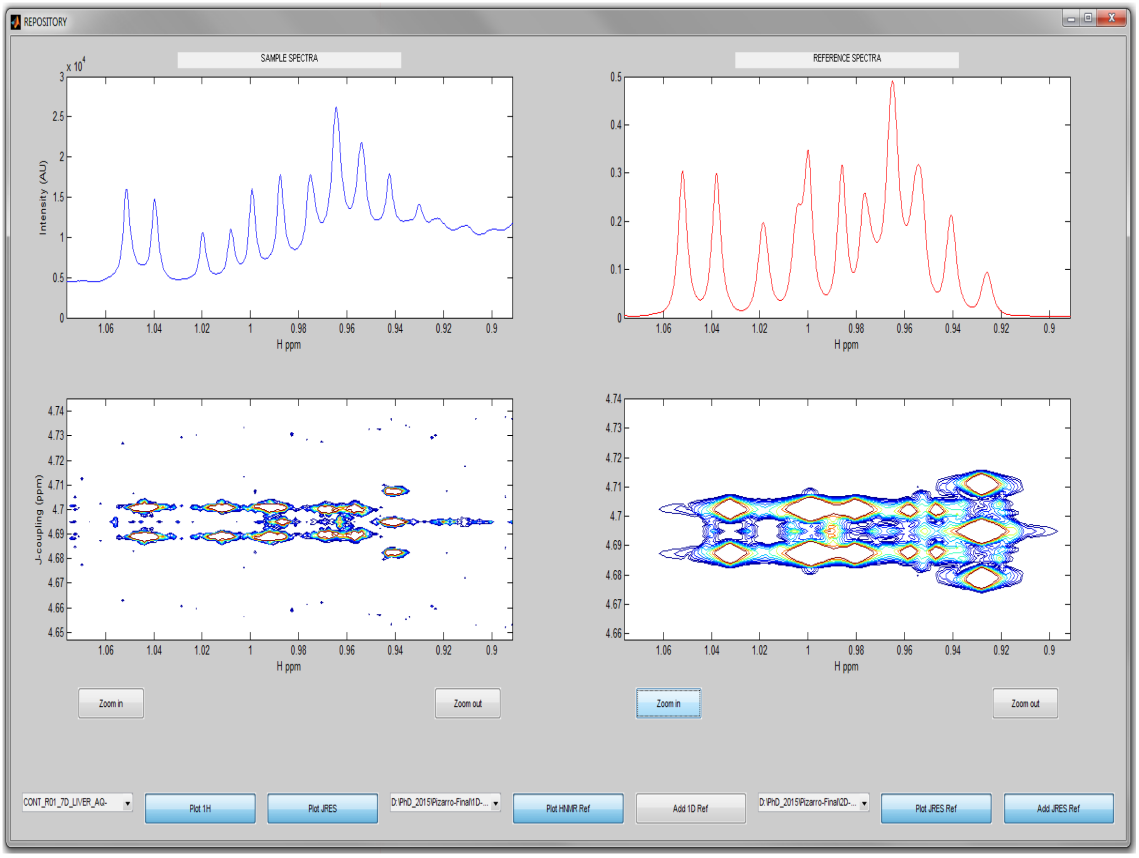


Figure C3.4. Here in this figure we show the Branched-Chain Amino Acids (BCAAs) region in our sample spectra in 1D NOESY (upper-left corner) and 2D JRES (lower-left corner); and the superposition of those BCAAs in 1D NOESY (upper-right corner) and 2D JRES (lower-right corner) using reference spectra stored in the repository.

3.1.5.2 ROIs testing

The ROIs testing block is the core of the package. Its strategy consists in analysing spectra through Regions Of Interest (ROIs). We consider a ROI as a region of the spectra that contains one or more resonances that are susceptible to be quantified. This block launches a panel where users can try and test which quantification mode is the most appropriate in each case before applying it along the whole dataset samples.

The ‘Clean Sum’ mode is very useful for those regions that contain isolated and pure (without any baseline) signals because the computation time is severely reduced while the quantification remains accurate.

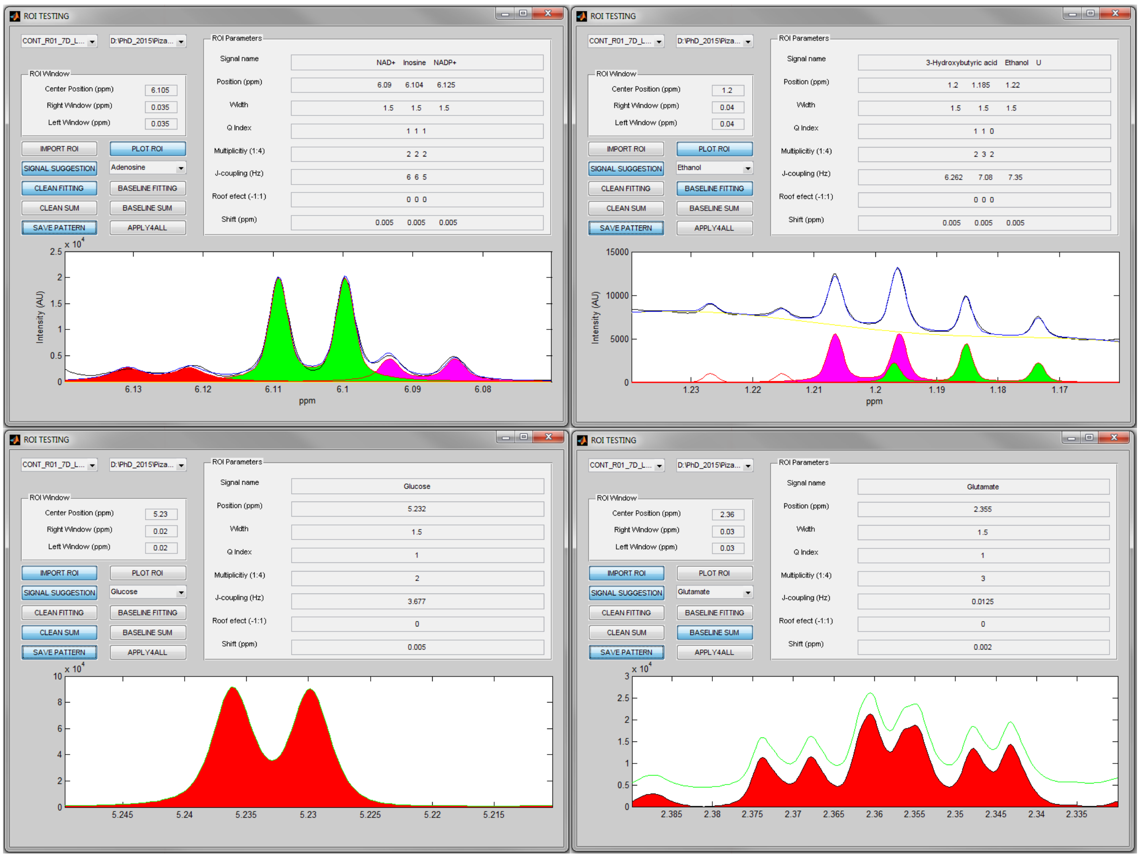
A very similar approach is used by the mode ‘Baseline Sum’, but in this case the algorithm removes what it considers is not part of the signal. It is almost faster as the ‘Clean Sum’ but reduces quantification error in regions with isolated but not pure (contain baseline) signals.

The ‘Clean Fitting’ mode is able to quantify accurately overlapping signals in regions where neither baseline nor broad signals are affecting.

Finally, the ‘Baseline Fitting’ mode allows to deconvolve targeted signals in regions where baseline or broad signals are affecting the final shape of the region. It takes more computation time but is the optimal solution in those cases.

The fitting algorithm is very similar to the previously tested and published in Dolphin (ref25) but with some improvements. The panel supports 2D data, which helps users to be sure about the signals present in a ROI and perform the most accurate deconvoloution. Moreover, there is a function that facilitates metabolite assignments through signal suggestions in each ROI. These signal suggestions come from an internal curated database built from public databases such as the Human Metabolome Database (HMDB), the Birmingham Metabolite Library (BML) and the BioMagResBank (BMRB). The curation is based on the selection of the most useful signals of the most relevant metabolites that frequently appear in ^{1}H-NMR spectra of biofluids and tissues according to recent literature.

The user can expand and reduce the library and the number of ROI pattern files, and adjust the features within the ROI patterns to the experiment requirements. In the ROIs Testing Panel, the user has the option to plot a single spectrum, all spectra or the average spectrum of a ROI, which is very useful to check graphically what could be the performance behavior between samples. Figure C3.5 shows an example of a ROI testing using the Manual Panel.



CONT_R01_7D_L... D:\PhD_2015\Piza...

ROI Window
Center Position (ppm) 6.105
Right Window (ppm) 0.035
Left Window (ppm) 0.035

Adenosine

ROI Parameters
Signal name NAD+ Inosine NADP+
Position (ppm) 6.09 6.104 6.125
Width 1.5 1.5 1.5
Q Index 1 1 1
Multiplicitiy (1:4) 2 2 2
J-coupling (Hz) 6 6 5
Roof efect (-1:1) 0 0 0
Shift (ppm) 0.005 0.005 0.005

CONT_R01_7D_L... D:\PhD_2015\Piza...

ROI Window
Center Position (ppm) 1.2
Right Window (ppm) 0.04
Left Window (ppm) 0.04

Ethanol

ROI Parameters
Signal name 3-Hydroxybutyric acid Ethanol U
Position (ppm) 1.2 1.185 1.22
Width 1.5 1.5 1.5
Q Index 1 1 0
Multiplicitiy (1:4) 2 3 2
J-coupling (Hz) 6.262 7.08 7.35
Roof efect (-1:1) 0 0 0
Shift (ppm) 0.005 0.005 0.005

CONT_R01_7D_L... D:\PhD_2015\Piza...

ROI Window
Center Position (ppm) 5.23
Right Window (ppm) 0.02
Left Window (ppm) 0.02

Glucose

ROI Parameters
Signal name Glucose
Position (ppm) 5.232
Width 1.5
Q Index 1
Multiplicitiy (1:4) 2
J-coupling (Hz) 3.677
Roof efect (-1:1) 0
Shift (ppm) 0.005

CONT_R01_7D_L... D:\PhD_2015\Piza...

ROI Window
Center Position (ppm) 2.36
Right Window (ppm) 0.03
Left Window (ppm) 0.03

Glutamate

ROI Parameters
Signal name Glutamate
Position (ppm) 2.355
Width 1.5
Q Index 1
Multiplicitiy (1:4) 3
J-coupling (Hz) 0.0125
Roof efect (-1:1) 0
Shift (ppm) 0.002

Figure C3.5. This figure shows the performance of the four modes of automated quantification that the package offers. Choosing the most appropriate quantification approach for each ROI allows users optimizing the time span in high-throughput analysis without compromising the reliability of the results.

3.1.5.3 Auto Run

Once all ROI patterns have been tested and considered satisfactory users can take profit of the potential of Whale as a high throughput tool. The auto run block processes all the ROI files stored in a folder into all spectra of the dataset, quantifying all the target signals automatedly without the user supervision. All the quantifications that have not passed the two quality thresholds (previously detailed in section 3.3) will be saved in the folder 'Plots2Check', where the user can graphically check the result and re-run those which are not satisfactory enough.

3.1.6 Applications

Several datasets of different biological matrices have been analyzed using Whale. Actually, the package has been developed in parallel to the necessities of expert users during the analysis of those datasets. More specifically, the package has been tested using human plasma, human urine, rat serum, rat urine, rat liver extracts, rat brain extracts and cell extracts. Depending on the final goal of each study, one or several signals per metabolite have been quantified, and the libraries extension has been increased with unknown signals. By now, a total number of more than 100 resonances belonging to more than 70 metabolites have been successfully profiled along these seven matrices. Moreover, some unknown signals, albumin regions and non-metabolite related signals such as EDTA-Mg and EDTA-Ca have been quantified as well. All the functions of the package have resulted useful to explore data and perform a reliable targeted metabolite profile analysis.

Next, two full studies will be exposed in order to illustrate the proficiency of the tool. With this two studies, the three secondary objectives of this thesis were reached, since the tool was applied to obtain the final results of each study, several algorithms able to handle the metabolomics data were developed and I could actively participate in each part of each study from the beginning to the end. In this two studies, the performance of the tool is referenced as Dolphin due to Whale's manuscript is still under development. Both studies are near to be published in prestigious journals and the package is a key element for the obtainment of the final results.

3.2 POSITIONAL ENRICHMENT BY PROTON ANALYSIS (PEPA): A ONE-DIMENSIONAL ^{1}H-NMR APPROACH FOR ^{13}C STABLE ISOTOPE RESOLVED METABOLOMICS

3.2.1 Abstract

A novel approach for NMR-based stable isotope tracer studies called PEPA is presented, and its performance validated using human cancer cells. PEPA detects the position of carbon label in isotopically enriched metabolites and quantifies their fractional enrichment by indirect determination of ^{13}C-satellite peaks using 1D-^{1}H-NMR spectra. In comparison with ^{13}C-NMR, TOCSY and HSQC, PEPA improves sensitivity, accelerates the elucidation of ^{13}C positions in labelled metabolites and the quantification of the percentage of stable isotope enrichment. Altogether, PEPA provides a novel framework for extending the high-throughput of ^{1}H-NMR metabolic profiling to stable isotope resolved metabolomics, calling to facilitate and complement the information derived from 2D-NMR experiments and expanding the range of isotopically enriched metabolites detected in cellular extracts.

3.2.2 Introduction

Nuclear magnetic resonance (NMR) is, together with mass spectrometry, the primary analytical tool to profile metabolite levels in biological samples[1]. However, the steady-state concentration of metabolites is not sufficient to study the regulation of cell metabolism. Elucidating the flow of chemical moieties through the complex set of metabolic reactions that happen in the cell is essential for understanding the regulation of metabolic pathways[2]. For that, stable isotope tracer studies are needed, which usually require model cellular lines that are fed a stable isotopically labeled substrate. Due to the unique ability of NMR to characterize isotopomers of metabolites by detecting labeling patterns in individual atoms, ^{13}C-NMR in combination with selective ^{13}C-stable isotope tracers have traditionally been used to analyze ^{13}C enrichment of extracted metabolites[3], enabling models of metabolic flux to be generated[4]. The advent of comprehensive metabolic profiling technologies has broadened the coverage of these studies allowing for an unbiased mapping of fluxes through multiple metabolic pathways[5]. The so-called Stable Isotope Resolved Metabolomics (SIRM)[6,7] has emerged as a powerful untargeted approach to classical ^{13}C-NMR-based stable isotope tracing studies. NMR-based SIRM primarily uses homonuclear 2D ^{1}H-^{1}H TOCSY (Total Correlation SpectroscopY) and heteronuclear 2D ^{1}H-^{13}C HSQC (Heteronuclear Single-Quantum Correlation) for analyzing isotopomers in crude cell extracts[8]. However, NMR-acquisition times necessary to obtain suitable TOCSY or HSQC spectra for SIRM applications are two to three orders of magnitude higher than 1D-^{1}H-NMR. Even so, sensitivity issues with respect to enriched metabolites in complex biological extracts and technical complexity in data acquisition and interpretation have prevented many researchers from routinely using 2D-NMR for isotope tracer studies. Thus, NMR-based SIRM is rather often used to obtain qualitative inputs on one or few samples, whereas quantitative information on larger sample sets is obtained via complementary mass spectrometry isotopologue analysis[5,9].

3.2.3 Methods

3.2.3.1 PEPA approach

To make NMR-based SIRM more accessible to comprehensive metabolic analyses, we have developed a new methodology called PEPA (Positional Enrichment by Proton Analysis). PEPA detects the position of carbon label in isotopically enriched metabolites and quantifies their fractional enrichment in 1D-^{1}H-NMR spectra. In principle, the ability of 1D-^{1}H-NMR to observe ^{13}C-satellite peaks makes it suitable for tracer studies. However, direct quantification of ^{13}C-satellites in 1D-^{1}H-NMR spectra of cell extracts is generally only possible for a few metabolites characterized by well-resolved resonances with recognizable ^{13}C-satellite peaks such as lactate and alanine. The vast majority of ^{13}C-satellites remain overlapped by the large body of redundant resonances that populate 1D-^{1}H-NMR spectra of biological cell extracts. This has prevented the use of 1D-^{1}H-NMR for comprehensive isotope tracer studies. PEPA circumvents this limitation by indirectly determining ^{13}C-satellites from the decayed central peak area observed in 1D-^{1}H-NMR spectra of biological equivalent replicates of labeled experiments as compared to the spectra acquired from unlabeled controls (Scheme 1).

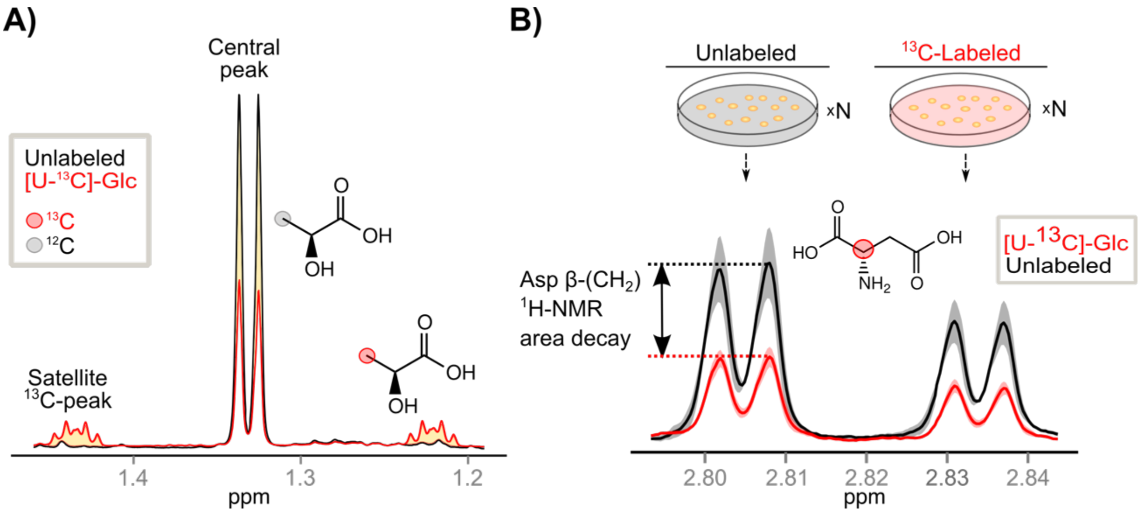
According to PEPA, the area of ^{13}C-satellite peaks in 1D-^{1}H-NMR spectra can be indirectly determined as detailed in Eq. (1):

$$I_{\text{satellites}}^{[\text{U-}^{13}\text{C}]} \approx \bar{I}_{\text{central peak}}^{\text{unlabelled}} - I_{\text{central peak}}^{[\text{U-}^{13}\text{C}]} \quad (1)$$

where I represents a central peak area in each biological replicate exposed to labeled substrate, and $\bar{I}$ is the mean area of the same central peak from all unlabeled replicates. The fractional ^{13}C enrichment for each resonance in the 1D-^{1}H-NMR can be calculated as detailed in Eq. (2):

$$F = \frac{I_{\text{satellites}}^{[\text{U-}^{13}\text{C}]}}{I_{\text{satellites}}^{[\text{U-}^{13}\text{C}]} + I_{\text{central peak}}^{[\text{U-}^{13}\text{C}]}} \approx \frac{\bar{I}_{\text{central peak}}^{\text{unlabelled}} - I_{\text{central peak}}^{[\text{U-}^{13}\text{C}]}}{\bar{I}_{\text{central peak}}^{\text{unlabelled}}} \quad (2)$$

Consequently, by indirect determination of ^{13}C-satellites areas, PEPA goes beyond classical assessment of labeled metabolites by NMR and it significantly widens the coverage of the metabolic network that can be investigated using 1D-^{1}H-NMR.



Scheme 1. **(A)** PEPA's conceptual framework: 1D-^{1}H-NMR spectral resonance of methyl protons in lactate at δ (1.33 ppm) obtained from cell extracts of two identical U2OS osteosarcoma cell cultures grown in either unlabeled glucose (black solid line) or [U-^{13}C]-Glucose (Glc) (red solid line). The latter shows evenly spaced ^{13}C-satellite peaks located at ±^1J(C-H)/2 from the central peak. These ^{13}C-satellite peaks result from heteronuclear (^{13}C-^{1}H) scalar couplings due to the replacement of ^{12}C-atoms in the methyl group of lactate by ^{13}C-atoms from [U-^{13}C]-Glc. The area of ^{13}C-satellite peaks (yellow-shaded) indicates the amount of ^{13}C-labeled methyl group in lactate while the area of the central peak (red line) represents the amount of unlabeled methyl group left. As the labeled substrate [U-^{13}C]-Glc is metabolized into lactate, the area of ^{13}C-satellite peaks increases proportionally with the decay of the central peak (red line). The decayed area of this central peak can be quantified from the total pool of lactate in unlabeled equivalent samples (black line). **(B)** PEPA's workflow: metabolites are extracted from replicates (n≥3) of unlabeled and ^{13}C-labeled (e.g., [U-^{13}C]-Glc) biologically equivalent samples and these measured by 1D-^{1}H-NMR. Next, 1D-^{1}H-NMR spectra are profiled and resonances quantified. Significant decayed areas of central peaks in [U-^{13}C]-Glc spectra by comparison with non-labeled controls are determined via statistical testing. As an example, aspartate β(CH$_2$) resonance at δ(2.82ppm) in three replicates of unlabeled (black) and labeled (red) samples provides the mean value of the area (solid line) and standard deviations (color shaded). Significant central peak decay proves ^{13}C-enrichment in this position indicating metabolic transformation of glucose into aspartate in U2OS osteosarcoma cell lines.

3.2.3.2 Cell Cultures

U2OS-cells human osteosarcoma cell lines were cultured in Dulbeccos modified Eagle Medium (DMEM, Invitrogen 41966) supplemented with 10% (v/v) fetal bovine serum (FBS, Sigma F7524-500ML) and 100U/ml penicillin/streptomycin (Reactiva 01030311B000). Puromycin and Hygromycin were used for selection of stable transformed cells. The cells (in duplicated batches) were cultured in a humidified atmosphere at 5% CO2, 37◦C and grown for 8-9 days in polypropylene bottles; the medium was changed once daily initially and twice daily near the end of the growth period. Then, U2OS-cells were seeded on 15-cm tissue culture plates in triplicates. During 6 hours, cells were supplemented with DMEM powder without glucose and glutamine (Sigma) supplemented with 5 mM glucose and 2 mM L-glutamine that was either unlabeled or either [U-^{13}C]-Glucose or [U-^{13}C]-Glutamine labeled (Cambridge Isotope Laboratories). After these 6 hours, culture medium was removed and cells were collected by trypsinization and 106 cells were pelleted and snap-frozen for further metabolites extraction.

3.2.3.3 Metabolites Extraction

Metabolites were extracted into the extraction solvent by adding 2 mL of a cold mixture of chlorofor- m/methanol (2:1 v/v). The resulting suspension was bath-sonicated for 3 minutes, and 2 mL of cold water was added. Then, 1 mL of chloroform/methanol (2:1 v/v) was added to the samples and bath- sonicated for 3 minutes. Cell lysates were centrifuged (5000 g, 15 min at 4◦C) and the aqueous phase was carefully transferred into a new tube. The sample was frozen, lyophilized and stored at −80◦C until further NMR analysis.

3.2.3.4 NMR Analysis

For NMR measurements, the hydrophilic extracts were reconstituted in 600 μl of D2O phosphate buffer (PBS 0.05 mM, pH 7.4, 99.5 % D2O) containing 0.73 mM of deuterared trisilylpropionic acid (d6-TSP). Samples were then vortexed, homogenized for 5 min, and centrifuged for 15 min at 14000 g at 4◦C. Finally, clear redissolved samples were tranferred into 5 mm NMR tubes. NMR spectra were recorded at 300 K on an Avance III 600 spectrometer (Bruker, Germany) operating at a proton frequency of 600.20 MHz and a carbon frequency of 150.93 MHz using a 5 mm CPTCI triple resonance (^{1}H, ^{13}C, ^{31}P) gradient cryoprobe. One-dimensional ^{1}H pulse experiments were performed using the nuclear Overhauser Effect Spectroscopy (NOESY) presaturation sequence (RD-90◦-t1-90◦-tm-90◦-ACQ) to suppress the residual water peak with mixing time (tm) of 100 ms. Solvent presaturation with irradiaton power of 50 Hz was applied during recycling delay (RD = 5 s) and mixing time. The 90◦ pulse length was calibrated for each sample and varied from 8.06 to 10.26μs. The spectral width was 12 kHz (20 ppm), and a total of 256 transients were collected into 64 k data points for each ^{1}H spectrum. The exponential line broadening applied before Fourier transformation was 0.3 Hz. One-dimensional ^{13}C-NMR spectra were recorded using Inverse Gate decoupled ^{13}C pulse experiment and WALTZ-16 scheme proton presaturation and a relaxation delay of 10 s. The 90◦ pulse length was 14.75 μs. The spectral width was 36 kHz (240 ppm), and a total of 6144 transients were collected into 64 k data points for each ^{13}C spectrum. The exponential line broadening applied before Fourier transformation was 1 Hz. Two-dimensional TOCSY spectra were recorded using MLEV17 spin lock with 63 ms of duration, B1 field strength of 8,33 kHz, and acquisition times of 0.28 s in t2 and 0.071 s in t1. The recycling delay was set to 5s, and 64 scans for each transient were collected into 2k of data points in t2 and 1024 transients were acquired in t1. Data were zero-filled to 4k points, apodized with a shifted Gaussian and a 1 Hz line broadening exponential function in both dimensions prior to Fourier transformation. Two-dimensional gradient ^{1}H-^{13}C-HSQC spectra were recorded using an acquisition time of 0.128 s in t2 and 0.0084 s in t1 and a recycle time of 1.5 s, with the ^{13}C GARP decoupling set to 100 ppm, and the evolution delay set for 145 Hz corresponding approximately to an average value of JCH coupling constants. The data were apodized with an unshifted Gaussian and a 1 Hz line broadening exponential in both dimensions prior to Fourier transformation.

3.2.3.5 NMR Metabolite Profiling and Data Analysis

The frequency domain spectra were phased, baseline-corrected and referenced to TSP signal (δ= 0 ppm) using TopSpin software (version 2.1, Bruker). Dolphin 1D[10] was used to profile metabolites identified in the ^{1}H-NMR spectra. Dolphin is a spectral profiling approach based on decomposition of the NMR spectrum as linear combination of a set of individual pure reference signals obtained from known compounds in a reference library. Poorly resolved and partially overlapped resonances are resolved using line shape fitting and consequently, their area can be properly quantified (Figure C3.6). Profiled areas were row-wise normalized using ERETIC[11] signal. Area of each profiled metabolite in unlabeled experiments were compared with the same area in experiments using isotopically enriched substrates using t-test ($p<0.05$ were considered for statistical significance). Data analysis was performed using R version 3.2.0.

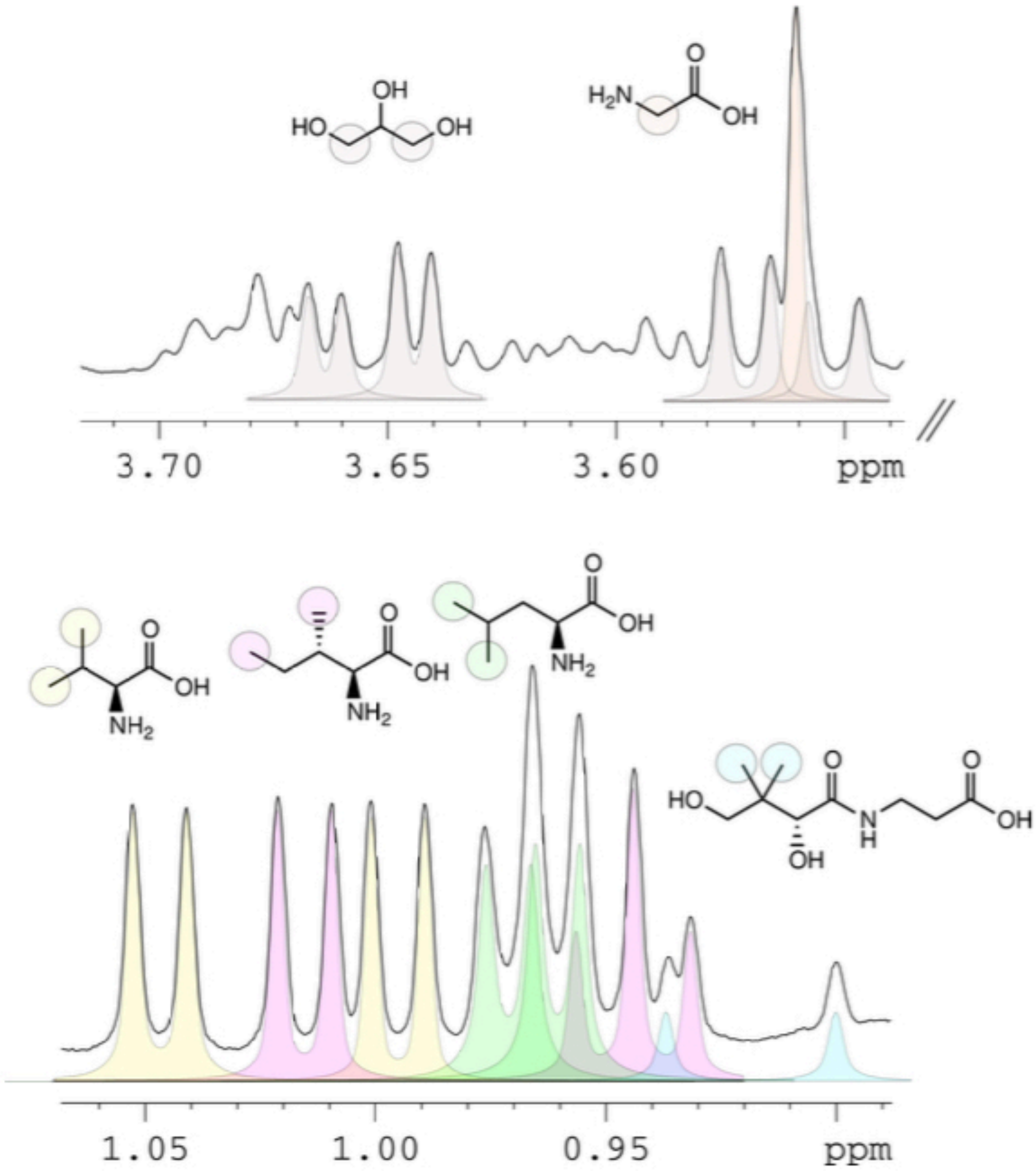


Figure C3.6: Example of Dolphin profiling in two ^{1}H-NMR different regions of the same U2OS cellular extract spectrum highlighting fittings for pantothenate (cyan), leucine (green), valine(yellow), isoleucine (pink), glycine(orange) and glycerol(gray).

3.2.4 Results and Discussion

U2OS osteosarcoma cells were used as a cellular model for the implementation of PEPA. U2OS cells were subjected during 6 hours to culture medium containing ^{13}C-enriched substrates (either uniformly labeled glucose [U-^{13}C]-Glc or uniformly labeled glutamine [U-^{13}C]-Gln) or to culture medium with unlabeled substrates (each condition was run in triplicate, see Supplementary Information for details). Overall, 84 different resonances underlying 46 polar metabolites were identified in 1D-^{1}H-NMR spectra. The area of 1D-^{1}H-NMR resonances were quantified using DOLPHIN[13], a spectral matching and deconvolution tool specially developed for quantifying partially overlapped and poorly resolved 1D-^{1}H-NMR resonances (Figure C3.6). Among the 84 proton resonances profiled in the ^{1}H-NMR spectra, we determined via statistical testing that 37 resonances showed significant decays in the central peak area in samples that were fed a stable isotopically labeled substrate. Figure C3.7 depicts fractional enrichments calculated according to Eq. 2 for these 37 resonances. According to PEPA, metabolic fates of [U-^{13}C]-Glc were found in the riboside C1', glucose anomeric C1'' and uracil ring C5 and C6 positions of UDP-Glucose, UDP-Glucuronate, UDP-GlcNAc and UDP-GalNAc. The C1'-riboside position was also found enriched in cytidine and adenosine nucleotides (AXP, where X refers to the number of phosphate groups). AXP, in addition, showed ^{13}C-labeling in C2 and C8 positions of the adenine ring. Finally, the metabolic fate of labeled glucose was also found in glycine(CH_2); glycerol(CH_2-OH); guanidoacetate(CH_2-OH); glutamate γ(CH_2); acetates(CH_3-); aspartate β(CH_2); o-phosphocholine(-CH_2-O/CH_3-); fumarate(CH) and glutathione γ(CH_2)-Glu/α(CH) and β(CH_2)-Cys. Thus, by using PEPA we were able to monitor the carbon flux of [U-^{13}C]-Glc into the glycolysis, tricarboxylic acid cycle, pentose phosphate pathway, glycine metabolism, hexosamine pathway and both purine and pyrimidine biosynthetic pathways. On the other hand, the main metabolic fates of [U-^{13}C]-Gln were found in fumarate (CH) and succinate (CH_2), both accounting for 99% and 79% fractional enrichment in mean, respectively. ^{13}C-atoms from [U-^{13}C]-Gln were also incorporated in aspartate β(CH_2) and in C5 and C6 positions of uracil in pyrimidine nucleotides, proving *de novo* synthesis of pyrimidine ring from aspartate. Finally, the fate of [U-^{13}C]-Gln was detected in glutamate γ(CH_2)/α(CH) and glutathione α(CH)/β(CH_2), indicating activation of glutaminolysis and biosynthesis of glutathione in this cancer cell line.

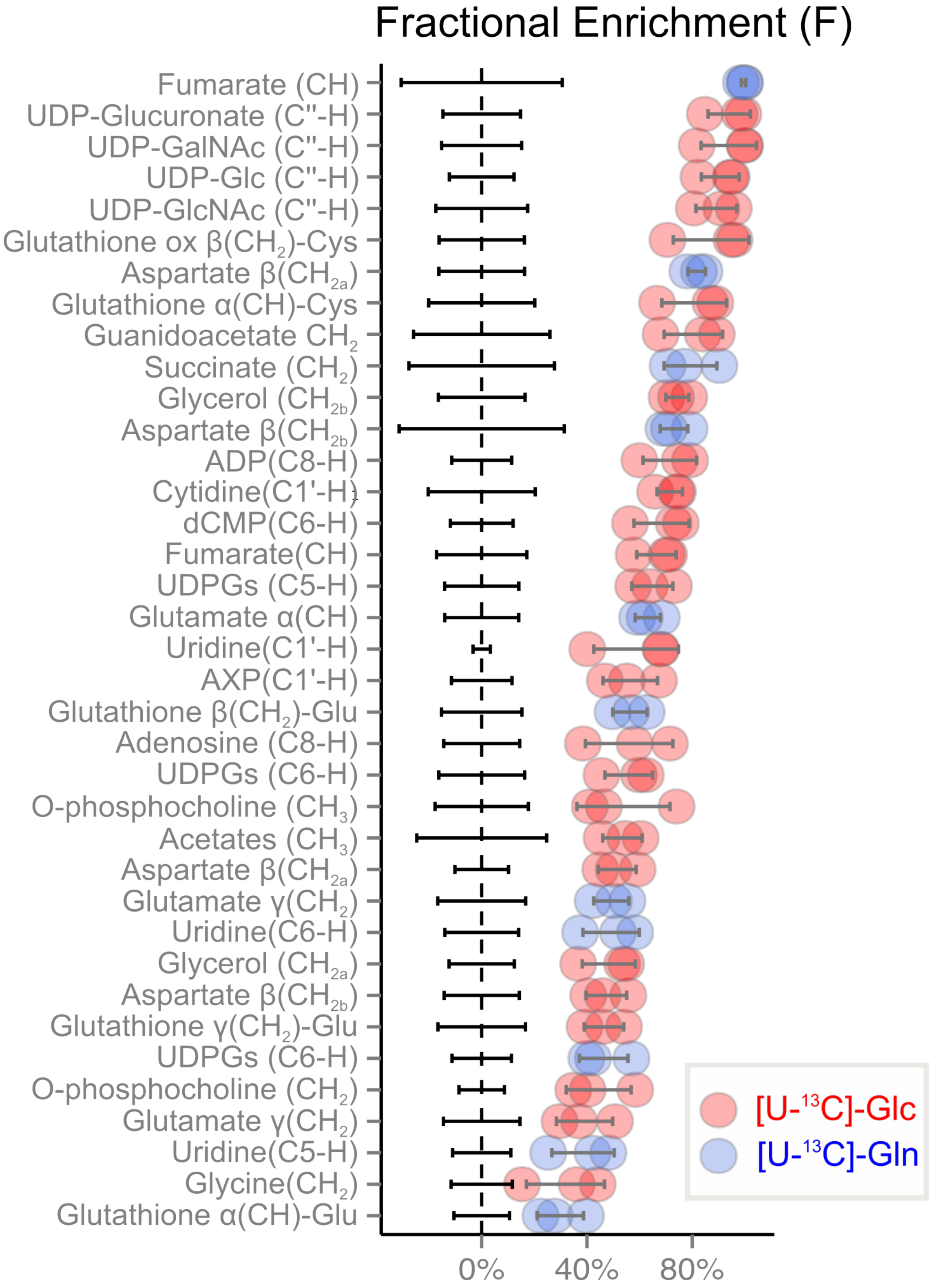


Figure C3.7. PEPA predicted significant fractional enrichments (F) calculated following Eq.2. Red and blue dots represent individual F values calculated for each of the three replicate samples using [U-^{13}C-Glc] or [U-^{13}C Gln], respectively. Grey lines indicate standard deviations in labeled samples. Black line represents the mean centered deviation about mean in unlabeled samples.

In order to confirm PEPA's results, ^{13}C-NMR and 2D-edited NMR experiments (TOCSY, HSQC and HMBC) were acquired on representative sample extracts grown either in unlabeled or labeled substrates. Figure C3.8 shows the validation by TOCSY analysis of enrichments originally detected by 1D-^{1}H-NMR and PEPA in C1' position of uridine and cytidine, and C5 position of uridine at δ[5.89-5.93ppm] (Figure C3.8A). In [U-^{13}C]-Glc experiments (Figure C3.8B), the cross-peaks patterns around the central unlabeled peaks of H1'-H3'-cytidine (gray) and H1'-H2'-uridine (green) correlations, confirmed the predicted isotopic enrichment of such metabolites in C1' position by PEPA. In [U-^{13}C]-Gln experiments, the ^{13}C-cross peak pattern around the central H5-H6 correlation peak of uridine in the TOCSY spectrum (Figure C3.8C) confirms the isotopic enrichment of C5 position of the uracil ring determined by PEPA. Altogether, uridine serves as a good example that dispersal of carbon flux across a range of metabolic substructures may determine the regulation of many metabolites that are assembled from these substructures[14].

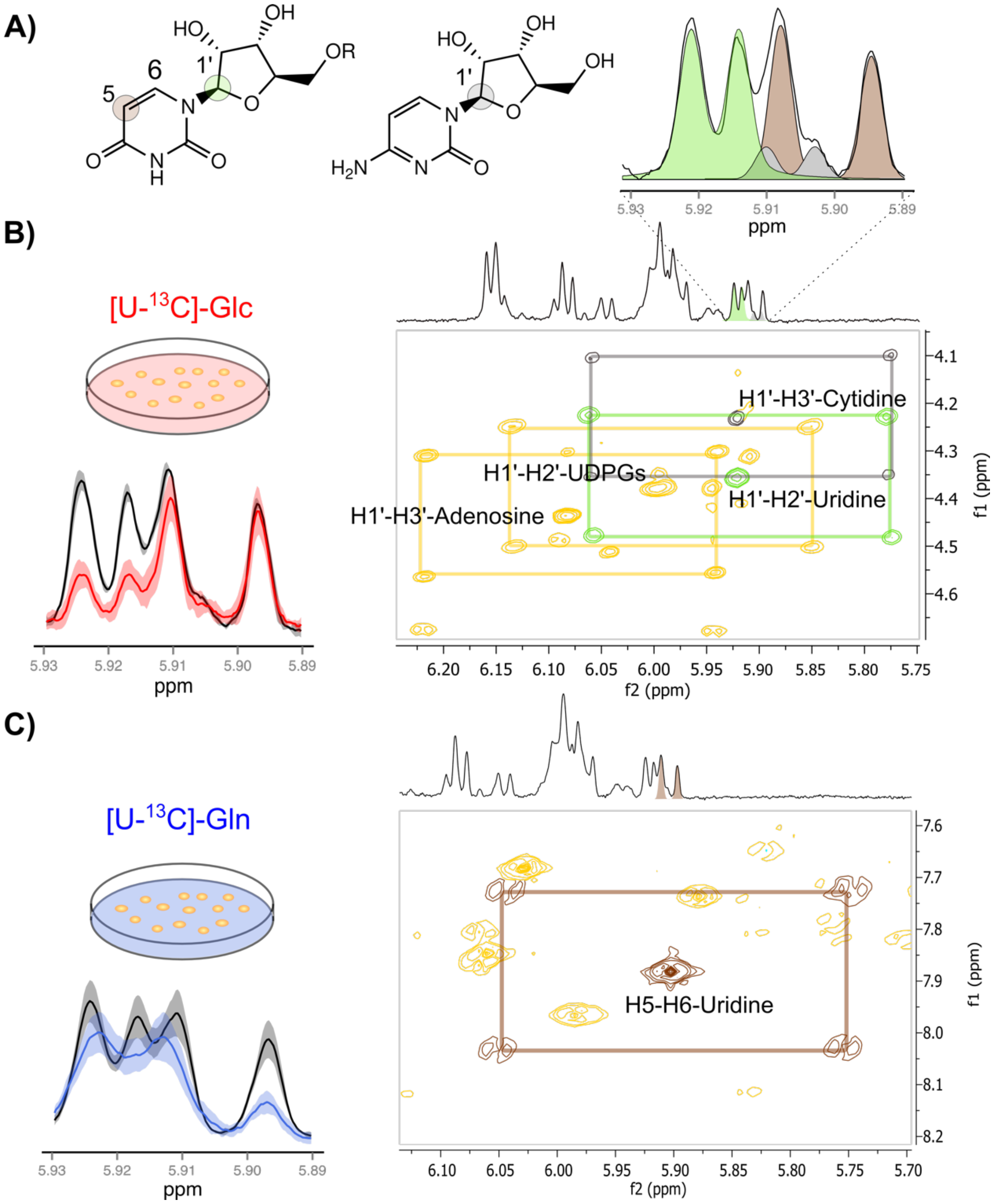


Figure C3.8. Validation of PEPA using TOCSY **(A)** 1D-^{1}H-NMR spectral region at δ[5.89-5.93ppm] shows the deconvolution of overlapping proton resonances in C1'-H position of uridine (d,J=4.4Hz, green) and cytidine (d,J=4.1Hz, gray), and C5-H position of uracil in the uridine structure (d,J=8.5Hz, brown). **(B)** Left panel: mean area (solid line) and standard deviation (color-shaded) of 1D-^{1}H-NMR spectra acquired on unlabeled (black) and [U-^{13}C]-Glc (red) cell extracts showing a significant decayed area of C1'-H uridine and cytidine doublets. Right panel: TOCSY spectrum shows ^{13}C cross-peaks traced around their corresponding central unlabeled signals in green (H1'-H2' correlation of uridine) and gray (H1'-H3' correlation of cytidine). Additional ^{13}C cross-peaks for H1'-H3' correlation of adenosine and H1'-H2' of UDPG-derived compounds are also indicated. **(C)** Left panel: mean area (solid line) and standard deviation (color-shaded) of 1D-^{1}H-NMR spectra acquired on unlabeled (black) and [U-^{13}C]-Gln (blue) cell extracts showing a significant decayed area for C5-H position of uracil in the uridine structure. Right panel: TOCSY spectrum shows ^{13}C cross-peaks traced around the corresponding central unlabeled signal in brown (H5-H6 correlation of uridine).

Noteworthy, each of the 37 positional ^{13}C-enrichments observed by PEPA could only be confirmed using a combination of ^{13}C-NMR and 2D-edited NMR experiments. Many positional enrichments in Figure C3.7 were confirmed using ^{13}C-NMR. The analysis of splitting patterns (i.e., multiplicities in ^{13}C-NMR and ^{13}C-cross-peaks in TOCSY) spectra serves to highlight a constraint of PEPA, namely, the composition of positional isotopomers can not be elucidated using PEPA. For instance, the analysis of ^{13}C-NMR splitting patterns of aspartate-β(CH2) and glutamate(γCH2/βCH2) from cells cultured in [U-^{13}C]-Glc revealed various ^{13}C-positional isotopomers. This information enables elucidating the number of turns of the TCA cycle. It is nonetheless also true that PEPA detected significant ^{13}C-enrichments in metabolites with either symmetric or non-correlated protons, including fumarate, succinate, guanidoacetate and acetates, an information that can not be otherwise detected using TOCSY experiments. The carbon labels of these metabolites were confirmed using HSQC spectra acquired on labeled sample extracts and contrasting them with HSQC spectra from unlabeled equivalent replicates. Although all positional ^{13}C-enrichments in Figure C3.7 were confirmed using ^{13}C-NMR, TOCSY, HSQC or HMBC, the enrichment of the methyl carbon position of lactate (d, δ 22.9 ppm) and alanine (d, δ 18.9 ppm) from cells fed with [U-^{13}C]-Glc observed in the ^{13}C-NMR spectra were not anticipated using PEPA. Although there was decay of the central peak corresponding to the methyl resonances in the 1D-^{1}H-NMR spectra acquired on labeled samples relative to unlabeled equivalent replicates, differences did not reach statistical significance. Since PEPA relies on statistical comparison of profiled resonances to establish whether or not there is ^{13}C-enrichment, the biological variability of the cellular model and a limited number of replicates may lead to the conclusion that there is no enrichment in certain metabolites. In our experience, increasing the number of cell culture replicates, and standardizing cell cultures procedures to reduce biological variability can prevent this.

3.2.5 Conclusions

In summary, PEPA takes advantage of the sensitivity, robustness, quantitativeness, high-throughput capabilities and ease-to-implement of 1D-^{1}H-NMR to study the position of carbon labels in large sample sets. Here we proved that PEPA span the range of isotopically enriched metabolites detected in cellular extracts over ^{13}C-NMR and 2D-edited experiments in NMR-based SIRM studies. PEPA greatly simplifies NMR untargeted carbon flux monitoring and complements the information derived from 2D-NMR. Altogether, PEPA is called to accelerate the implementation of NMR-based SIRM in cell metabolism studies.

3.2.6 Acknowledgments

We thank Dr. Reza M. Salek for helpful discussions. We thank the financial support from the Spanish Biomedical Research Centre in Diabetes and Associated Metabolic Disorders (CIBERDEM), an initiative of Instituto de Investigacion Carlos III (ISCIII), and the Spanish Ministry of Economy and Competitiveness grants SAF2011-30578 and BFU2014-57466 (to OY).

3.3 BASELINE METABOLOMIC SIGNATURE PREDICTS IMMUNOLOGICAL CD4^{+} T-CELL RECOVERY AFTER 36 MONTHS OF VIROLOGICALLY SUCCESSFUL ART IN ADULT HIV-INFECTED PATIENTS: A PILOT STUDY

3.3.1 Abstract

Background Poor immunological recovery in treated HIV-infected patients is associated with greater morbidity and mortality. To date, predictive biomarkers of this incomplete immune reconstitution have not been established. For this reason, we aimed to discover such biomarkers that can identify a discordant response to ART.

Methods This was a multi-centre, prospective cohort study in ART–naïve and a pre-ART low nadir (<200 cells/µl) HIV-infected patients. We obtained clinical data and metabolomic profiles for each individual, in which low molecular weight metabolites, lipids and lipoproteins (including particle concentrations and sizes) were measured by nuclear magnetic resonance spectroscopy. We differentiated patients as individuals who have not arrived to 250 CD4^{+} T-cells/µL after 36 months of virologically successful ART against those who have passed this threshold. We used univariate comparisons, Random Forest test and ROC curves for baseline predictive factors of discordant immunology after treatment.

Findings HIV-infected patients with a baseline metabolic pattern characterized by high levels of large HDL particles, HDL cholesterol and larger sizes of LDL particles had a better immunological recovery after treatment. On the contrary, patients with high ratios of non HDL lipoprotein particles did not experience this full recovery. Medium VLDL particles and glucose increased the classification power of the multivariate model despite not showing any significant differences between the two groups.

Interpretation In HIV- infected patients, pre ART lipid metabolism intermediates, mainly HDL-related lipoprotein parameters, and glucose levels can accurately predict immunological recovery in treated HIV infection.

3.3.2 Introduction

Since the introduction of effective antiretroviral therapy (ART) the fatal course of HIV infection has been prevented. ART decreases viral replication, increases CD4$^+$ T-cells count and consequently, improves the immune system function[15]. Unfortunately, 25 to 30% of HIV-infected patients still fail to restore their CD4$^+$ T-cell number despite optimal treatment and sustained virological suppression[16]. This group of patients is referred to as "immunodiscordants" or "immunological non-responders" (INR) and are at a higher risk of clinical progression and death[17,18].

Traditional factors associated with poor immune recovery are delayed diagnosis, advanced age, co-infection with hepatitis C virus (HCV), thymic dysfunction, immune activation, and genetic factors among others. However, none of them provides a full explanation of the lack of total immune reconstitution[19–21]. In addition, no predictive biomarkers of this immunological recovery in HIV-infected patients are currently available.

In the present study, we used a comprehensive metabolomic approach to plasma samples from HIV-infected individuals before starting ART with the aim of identifying a "metabolomic signature" that might predict immunological recovery measured after 36 months.

Metabolomics techniques such as nuclear magnetic resonance (NMR) have emerged as a powerful method for discovering new biomarkers for disease diagnosis, prognosis and risk prediction. One of its most important advantages is that it can be used to identify disease-related patterns through accurate detection of numerous metabolic changes in biological samples[22,23].

3.3.3 Methods

3.3.3.1 Study Design

This was a multi-centre and prospective cohort study comprising all adult HIV-1 infected individuals who started their first ART between 2009 and 2011 and were followed-up at the HIV outpatient clinics of the participating hospitals: Hospital Joan XXIII (Tarragona), Hospital de la Santa Creu i Sant Pau (Barcelona) and Hospital Virgen del Rocío (Sevilla). Of the initial cohort (n=491), 98 participants fulfilled the following inclusion criteria: >18 years, presence of HIV-1 infection and pre-ART low nadir CD4 counts (<200 cells/μl). Exclusion criteria were pre-ART nadir of CD4 counts >200 cells/μl, HCV co- infection, the presence of active opportunistic infections, current inflammatory diseases or conditions, consumption of drugs with known metabolic effects, type 2 diabetes mellitus, acute or chronic renal failure, pregnancy, history of vaccination during the previous year and plasma C- reactive protein >1mg/dL. From those patients who fulfilled the inclusion criteria, we handled a subset of 84 whose stored plasma samples, drawn when enrolled, were available. In a final step, we excluded 20 samples due to their poor spectral quality, which resulted in a final cohort of 64 patients. Patients were categorized into two groups based on their CD4 T-cell count at 36 months after ART: immunological responders (IR group) if their CD4 T-cell count was higher than 250 cells/μL or immunological non-responders (INR group) if they did not reach this threshold. Figure C3.9 provides a flowchart with patient selection and enrolment. The ethics committee from each recruiting centre reviewed and approved this consent procedure before the study began, and all volunteers provided their written informed consent.

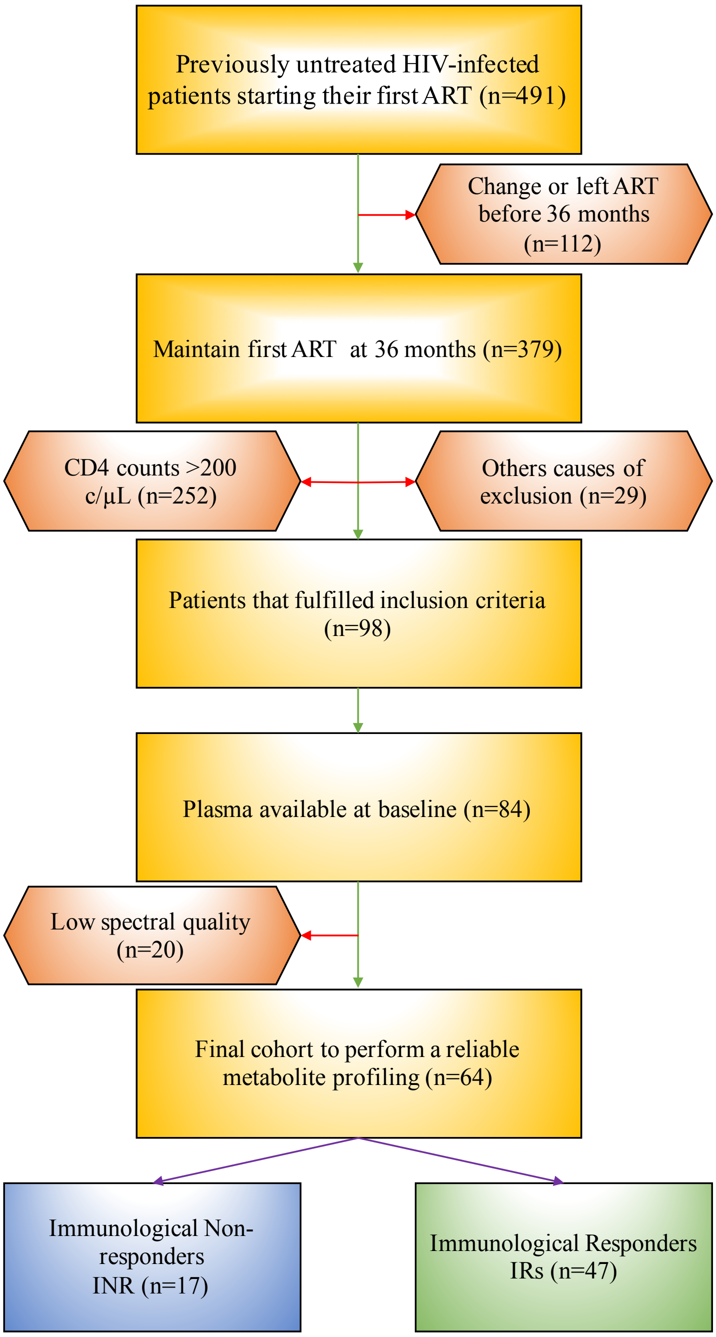

Previously untreated HIV-infected patients starting their first ART (n=491)
Change or left ART before 36 months (n=112)
Maintain first ART at 36 months (n=379)
CD4 counts >200 c/µL (n=252)
Others causes of exclusion (n=29)
Patients that fulfilled inclusion criteria (n=98)
Plasma available at baseline (n=84)
Low spectral quality (n=20)
Final cohort to perform a reliable metabolite profiling (n=64)
Immunological Non-responders INR (n=17)
Immunological Responders IRs (n=47)

Figure C3.9. Flowchart of patient selection and enrollment.

3.3.3.2 Data Collection

Relevant clinical and demographic data were extracted from electronic predefined database specially defined for this study. Fasting venous blood samples were collected in EDTA tubes and centrifuged immediately for 15 min at 4°C and 1,500 g. Plasma samples were then stored at 80°C until further analysis.

3.3.3.3 NMR Measurements

For NMR measurements, 430 µl of plasma was transferred to 5-mm NMR tubes. A double tube system was used. The external reference tube (o.d. 2 mm, supported by a Teflon adapter) containing the reference substance (9.9 mmol/l sodium 3-trimethylsilyl[2,2,3,3-d4] propionate (TSP), 0.47 mmol/l MnSO4 in 99.9% D2O) was placed coaxially into the NMR sample tube (o.d. 5 mm). This double tube system was kept at 4ºC in the sample changer until the moment of analysis.

All ^{1}H NMR spectra were recorded at 310 K on a Bruker Avance III 600 spectrometer operating at a proton frequency of 600.20 MHz and using a 5 mm CPTCI triple resonance (^{1}H, ^{13}C, ^{31}P) gradient cryoprobe.

^{1}H spectra of low molecular weight metabolites (LMWMs) were performed using the Carr-Purcell-Meiboom-Gill sequence (CPMG spin-spin T2 relaxation filter) (RD–90º– [τ–180º–τ-n–ACQ FID), with a 0.4 ms of echo time (τ) to allow elimination of J modulation and 500 loops (n) for a total time filter of 410 ms, that attenuate the signals of macromolecules to a residual level. Pre-saturation of the water signal was applied during the recycling delay (RD) period of 5s. The spectral width was 20 ppm, and a total of 64 transients were collected during acquisition time (ACQ) of 2.73 s into 64 k data points for each CPMG spectrum. The total CPMG experiment time was 9 min per sample.

^{1}H spectra of macromolecules were measured using a diffusion-edited pulse sequence with bipolar gradients and the longitudinal eddy-current delay (LED) scheme with two spoil gradients (ledbpgp2s1d Bruker ® pulse RD-90º-G1-180º-(-G1)-90º-Gs-D-90º-G1-180º-(-G1)-90º-Gs-τ-90º-acquire FID). The relaxation delay (RD) was set to 2 s, and the FIDs were collected into 64K, complex data points. 64 scans were acquired for each sample with a gradient pulse strength (G1) of 3.23 Gauss per cm and an eddy current delay (τ) of 5 ms. A diffusion time of 116 ms and bipolar sine-shaped gradient pulses of length 2.6 ms were applied to obtain the lipoprotein profile without the low-molecular weight metabolites signals. The total diffusion experiment time was 4.5 min per sample.

The acquired NMR spectra were phased, baseline-corrected, and referenced to the chemical shift of the α-glucose anomeric proton doublet taken at 5.233 ppms, except for the diffusion-edited spectra, in which the spectral reference (SR) offset value from a referenced ^{1}H co-acquisition on the same sample was used (in diffusion-edited spectra the α-glucose signal is attenuated and cannot be used as a reference). Additionally, an electronic reference signal ERETIC was introduced for quantification purposes.

3.3.3.4 Quantitative Variables

Lipid concentrations (i.e., triglycerides and cholesterol), sizes, and particle numbers for VLDL (38.6–81.9 nm), LDL (18.9–26.5 nm), and HDL classes (7.8–11.5 nm), as well as the particle numbers of nine subclasses, namely large, medium, and small VLDL, LDL, and HDL of frozen EDTA plasma specimens were measured by nuclear magnetic resonance (NMR) spectroscopy using the Liposcale test [24]. This test is based on 2D spectra from diffusion-ordered NMR spectroscopy (DOSY) experiments. Briefly, cholesterol and triglyceride concentrations of the main lipoprotein fractions were predicted using partial least squares (PLS) regression models. Then, the methyl proton resonances of the lipids in lipoprotein particles were decomposed into nine Lorentzian functions representing nine lipoprotein subclasses and the mean particle size of every main fraction (VLDL, LDL, and HDL) was derived by averaging the NMR area of each fraction by its associated size. Finally, the particle numbers of each lipoprotein main fraction were calculated by dividing the lipid volume by the particle volume of a given class and the relative areas of the lipoprotein components used to decompose the NMR spectra were used to derive the particle numbers of the nine lipoprotein subclasses.

A target set of eleven low molecular weight metabolites (LMWMs) was identified and quantified by NMR spectroscopy in the 1D Carr-Purcell-Meiboom-Gill sequence (CPMG) spectra using Dolphin [10,13]. Each metabolite was identified by checking for all its resonances along the spectra, and then quantified using line-shape fitting methods on one of its signals. The quantification units corresponding to the area under the curve of each metabolite were normalized by the mean of each of them along all samples, being the final units a reflection of the fold change of each sample over the mean of the dataset. In addition to this set, we incorporated the measure of a peak related with glycoprotein concentration in blood [25] and the quantification of two EDTA peaks that have been previously reported as indicators of calcium and magnesium levels in blood [26,27].

3.3.3.5 Satistical Analyses

A descriptive analysis of patients' characteristics was carried out using frequency tables for categorical variables and median and interquartile ranges for continuous variables. Differences in socio-demographic and clinical characteristics between INR and IR groups were assessed through the non-parametric Mann-Whitney test for continuous variables and the chi-squared test for independence for categorical variables.

In order to find metabolic differences between the two groups, we compared their metabolic patterns at baseline using different methods. Univariate comparisons were made through the non-parametric Mann-Whitney U test, where variables with a P-value < 0.05 were determined as significantly altered between the two groups. In addition, the fold change of each variable was calculated as 'A/B', where 'A' was the variable mean in the IR group and 'B' was the variable mean in the INRs group. Multivariate statistics were also used to improve the refining and distilling of all the metabolic baseline data and for pattern recognition purposes. In this sense, Random Forest analysis was applied, which is a supervised classification technique based on an ensemble of decision trees and provides an unbiased selection of variables that make the largest contributions to the classification. For this analysis, apart from the metabolic variables, the variables of age and $CD4^+$ T-cell count at baseline were included in order to evaluate their importance as predictors in the classification between the two groups. Finally, logistic regression analysis and receiver operating characteristic (ROC) curves were generated, using as input those metabolites considered important discriminators of $CD4^+$ T-cell recovery, obtained from the Mann-Whitney U test (P values < 0.05) and the Random Forest analysis (largest contributions in the classification model) and adjusted for confounders (age and baseline $CD4^+$ T-cell count). The statistical software used included the program 'R' (http://cran.r-project.org) and the SPSS 21.0 package (IBM, Madrid, Spain).

3.3.3.6 Role of the Funding Source

The study sponsors had no role in the study design, data collection, data analysis, data interpretation, or writing of the report manuscript. The correspondings authors have full access to all the data in the study and had final responsibility for the decision to submit for publication.

3.3.4 Results

After 36 months, 17/64 individuals (27%) did not arrive at 250 CD4$^+$ T-cell count / µL while 47/64 (73%) reached this threshold. **Table C3.1** contains baseline clinical details of the two subsets analysed. The differences in age and baseline CD4$^+$ T-cell count between groups, despite not being statistically significant, were in agreement with the literature[20,21], suggesting that older people with a low nadir CD4$^+$ T-cell count are associated with a minor recovery capacity.

Table C3.1. Baseline characteristics of the HIV-1–infected subjects studied categorized according to their immunological response after 36 months of ART.

	Study cohort (n=64)		
Variable	**INR (n=17)**	**IR (n=47)**	***P***
Age (years)	44 (39-55)	38 (34-50)	0.075
Male (%)	76.5	78.7	1.000
AIDS (%)	100.0	85.1	0.175
HIV-1 risk factor (%)			
Injecting drug user	0.0	8.5	0.566
Homosexual	47.1	44.7	1.000
Heterosexual	41.2	42.6	1.000
Other/Unknown	11.8	4.3	0.285
CD4+ T-cell count (cells/μL)			
Baseline	60 (28-122)	92 (48-166)	0.068
Current	188 (117-219)	378 (323-469)	<0.001
Plasma viral load (log copies/mL)			
Baseline	5.5 (4.8-5.7)	5.4 (4.8-5.7)	0.915
Current	1.3 (1.3-1.7)	1.3 (1.3-1.6)	0.355
ART Received (%)			
2NRTi + NNRTi	47.1	42.6	0.782
2NRTi + PI	52.9	57.4	0.782

Quantitative variables are expressed as median (interquartile range). Qualitative variables are expressed as percentages. INR, immunological non-responders; IR, immunological responders.

Figure C3.10A shows a heat map of the fold change of the 40 metabolomic variables used in this study. All HDL particles, including HDL cholesterol and HDL triglycerides, were incremented in the metabolism of the IR group at baseline point, while all VLDL particles, including VLDL cholesterol and VLDL triglycerides, and almost all LDL particles, including LDL cholesterol and LDL triglycerides, were incremented in the INR group. This particle balance makes the two ratio variables 'LDL particles / HDL particles' and 'Total particles / HDL particles' the most accentuated in the INR group at the baseline point. The size of all kinds of particles (HDL, LDL, VLDL) were higher in the IR group. From the LMWM balance, the IR group presented a higher concentration of most of the amino acids (histidine, glutamine, valine, creatine, tyrosine and leucine), while the INR group showed a higher concentration of a few (alanine and isoleucine). All acids (lactate, formate and acetate), the two EDTA peaks, the glycoprotein peak and glucose were **lower** in the IR group. **Figure C3.10B** presents the notched box-plots of the five variables that altered significantly in the metabolism at a baseline point in the comparison between groups[28]. Large HDL particles (p=0.002), LDL particle size (p=0.029) and HDL cholesterol (p=0.045) were all significantly higher in the IR group, while the ratios of total particles / HDL particles (p=0.029) and LDL particles / HDL particles (p=0.049) were incremented in the INR group. The univariate results therefore suggest that high levels of HDL particles (especially the subclass 'large') including HDL cholesterol and larger LDL particle sizes favoured immunological recovery. On the other hand, high ratios of lipid particles, where HDL particles were the denominator, did not.

Random Forest analysis revealed large HDL particles as the primary differentiator in a ranked list of metabolites in order of their importance in the classification scheme (**Figure C3.11**). It is important to highlight that large HDL particles were also the most important variable in the univariate test, which suggests they are the most important pre-ART indicator of CD4$^+$ T-cell recovery over time. The next three variables in order of importance were: medium VLDL particles, glucose and all non-HDL particles, which despite not appearing significantly different in the univariate analysis had strong classification power in the multivariate model, and thus were selected for the logistic regression and ROC analyses.

The consequent assessment was the validation of our candidate biomarkers for their clinical usefulness. Accordingly, the ROC curve revealed the diagnostic accuracy of these signature biomarkers, obtained in the Mann-Whitney and the Random Forest analyses. Our results show that the area under the curve (AUC) of each analyte was less than 0.8 **(Supplementary Material S2**). For this reason, we used a multivariate logistic regression model that combined each potential biomarker mentioned before. This model displayed an AUC value of 0.901 and correctly classified 84.4% of patients with 80% of sensitivity and 82.4% specificity (**Figure C3.12, Model A**). Moreover, given that age and baseline CD4$^+$ T-cell counts are considered confounders, we adjusted their model. In this case, the AUC value increased by only 0.006, specificity increased by 5.8% and it did not improve the percentage of classification (84.4%) (**Figure C3.12, Model B**). We therefore propose the combination of these metabolic variables as a possible diagnostic panel for the prediction of immunological recovery.

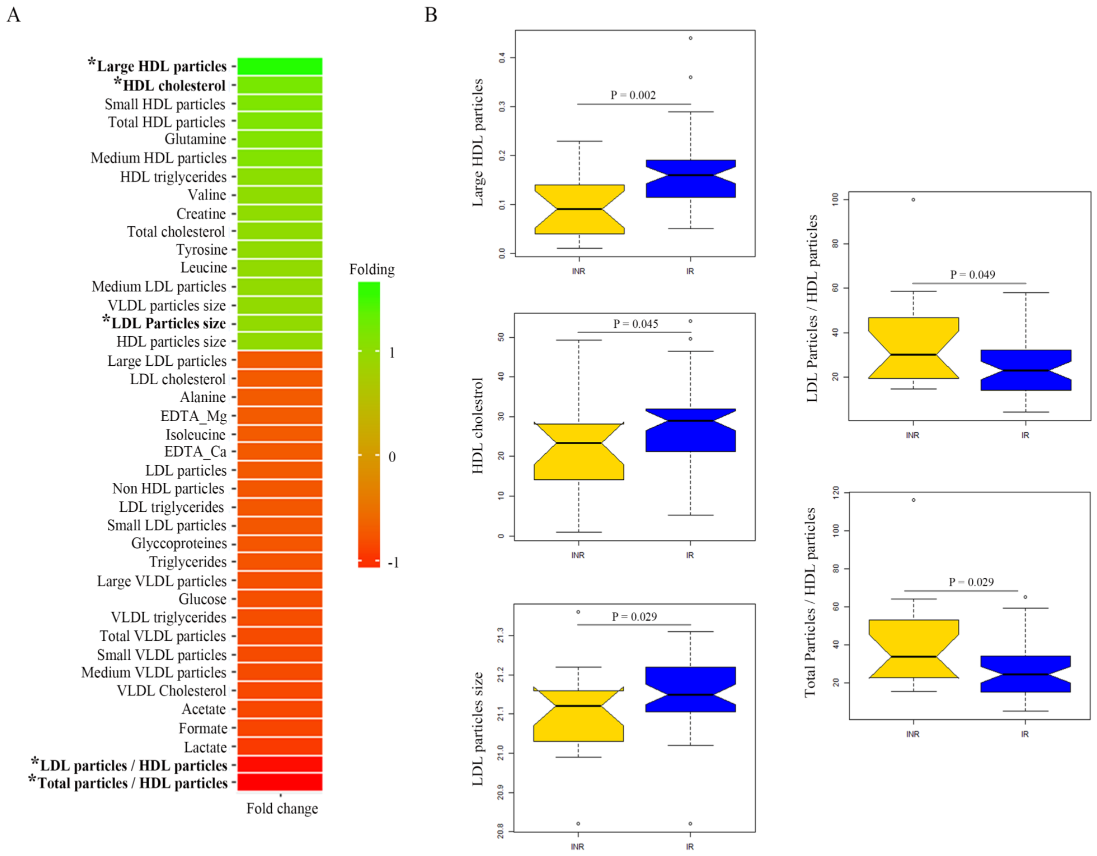


Figure C3.10. A) Fold-change heat map of the relative plasma concentrations of measured metabolites at baseline. Positive folding (green) means higher concentrations in the responders group, while negative folding (red) means the opposite. The asterisk highlights those variables with a significant P-value able to distinguish responders and non-responders HIV-patients. B) Notched box-plots of statistically significant altered metabolites, where the notch shows the 95% confidence interval (CI) for the median, given by m ± 1.58 x IQR/√n.

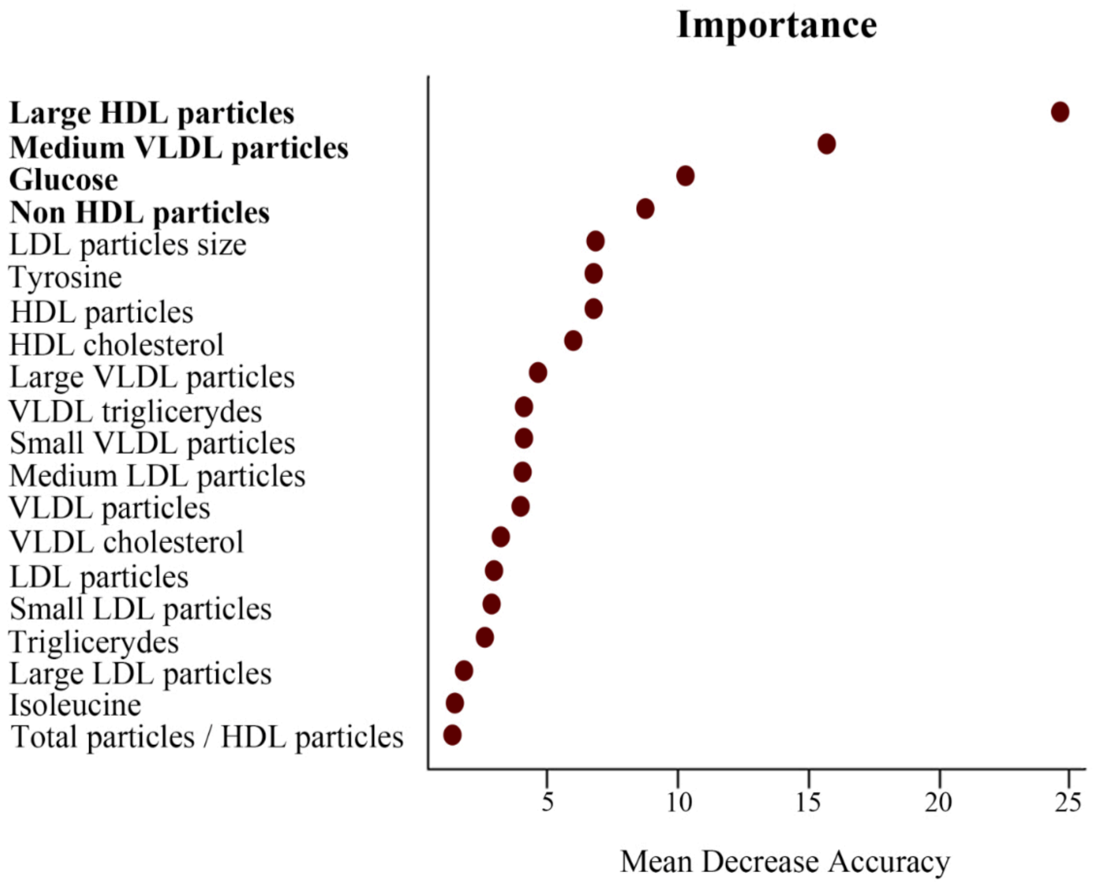
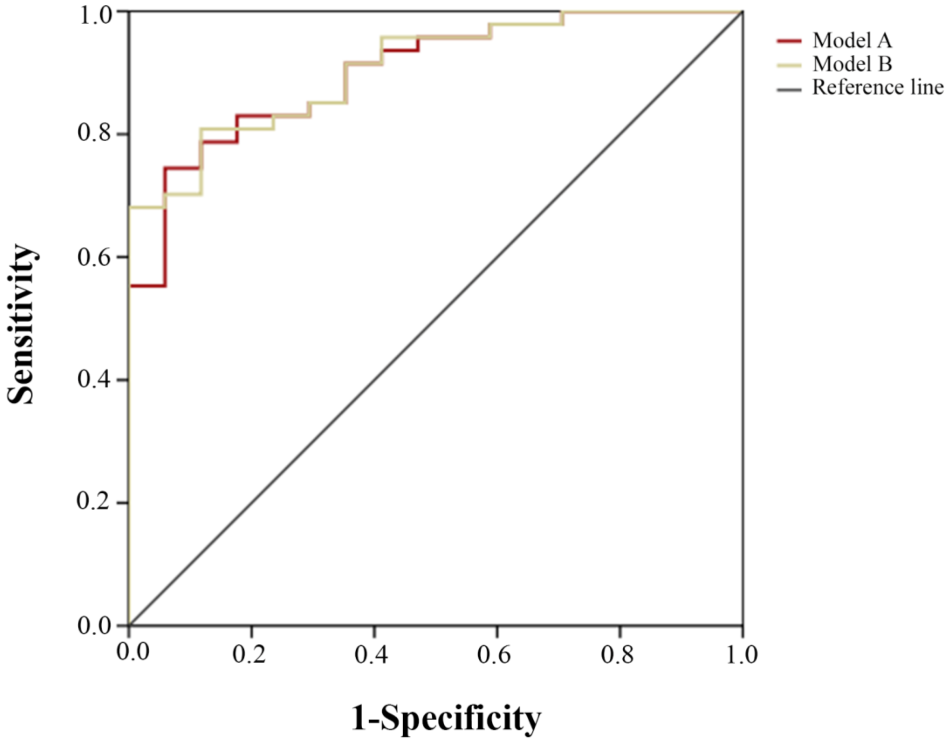
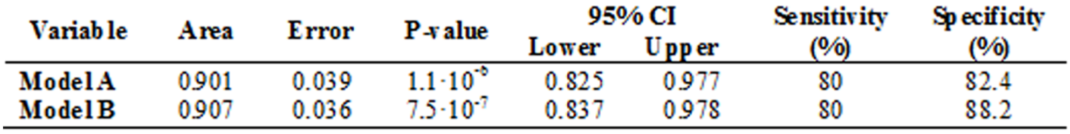


Figure C3.11. Variable importance plot of the Random Forest analysis resulting from a large number of models built around immunological response to ART. The variables are ordered top-to-bottom as most-to-least important in classifying between responders and non-responders. The ranked list of variables tells us the importance of each variable in classifying data. The figure shows the top 20 variables in importance of classification from a total of 42, including age and CD4$^+$ T-cell count, and only the top four (bold) were considered for the ROC curve analysis.





Variable	Area	Error	P-value	95% CI Lower	 Upper	Sensitivity (%)	Specificity (%)
Model A	0.901	0.039	$1.1 \cdot 10^{-6}$	0.825	0.977	80	82.4
Model B	0.907	0.036	$7.5 \cdot 10^{-7}$	0.837	0.978	80	88.2

Figure C3.12. Using receiver operator characteristic curves, we assessed multi-metabolite biomarker models that could accurately predict a discordant response to HIV-infection treatment. **Model A** inputs: Large HDL particles, LDL particle size, Total particles / HDL particles, LDL Particles / HDL particles, Glucose, Medium VLDL particles and Non-HDL particles**.** For **Model B** we used the same metabolomic inputs but the model was adjusted for age and baseline CD4$^+$ T-cell count.

3.3.5 Discussion

Our findings reveal that there are metabolomic differences between pre-ART HIV-infected individuals with a low nadir of CD4$^+$ T-cell count at baseline and that these differences can predict future response after 36 months of treatment. In general, the IR group presents a healthier metabolomic profile than the INR group. Even considering that only five metabolomic indicators significantly altered in the univariate test, the fold change of all kinds of HDL particles and the size of all the lipoprotein classes are higher in the IR group, while all kinds of VLDL and most LDL particles are higher in the INR group. Moreover, the random forest test corroborates that large HDL particles and LDL particle size are key metabolomic features that differentiate IR and INR. This test also shows that medium VLDL and non-HDL particles, despite not being significantly different between groups, contribute to that. Putting all of these factors together, we report a multivariate model with all of these variables that can accurately predict the immunological recovery in HIV-patients with a low nadir $CD4^+$ T-cell count after 36 months of ART.Incomplete immunological recovery during ART is a relevant clinical problem; indeed, in our multi-centre, prospective cohort study, 24% of the 64 treated HIV-infected patients who had viral suppression were considered INR using a restrictive definition ($CD4^+$ T-cell<250 cels/µL after 36 months of successful treatment). Several studies have tried to elucidate the impact of HIV-induced immunological changes on metabolism[29,30], and others focused on finding metabolomic differences between HIV-infected subjects and healthy controls[29,31]. However, only a small number of studies have investigated baseline indicators of $CD4^+$ T-cell recovery before ART. In those studies, older age, a lower nadir $CD4^+$ T-cell count, higher immune activation, viral load, and HCV co-infection have all been proposed as the most relevant predictive factors for a immunodiscordant response even though individually they were not able to predict treatment response, so their combined effect remains elusive[32,33]. The present study has included the variables of age and $CD4^+$ T-cell count at baseline in the statistical analyses but none of them has presented a significant P-value in the univariate comparisons (Table 1), showed importance in the classification between the two groups in the Random Forest model (Figure 3) or improved the logistic regression analysis model (Figure 4). Consequently, our results suggest that even if differences in age and immunological parameters influence the metabolomic profile, such influence is weak enough to discard bias. Therefore, there is still the need to identify specific predictive factors for an early recognition and classification of immunological discordant individuals in order to propose the appropriate therapy to each situation.

Some studies have characterized the metabolomic profile of HIV/AIDS biofluids using metabolomic techniques such as proton nuclear magnetic resonance spectroscopy and mass spectrometry and demonstrated the ability to detect metabolites affected by infection and treatment[34,35]. However, none of them have been carried out using metabolism as the target for immunological recovery biomarker identification. For this reason, the present study could be considered novel research work aimed at identifying a useful metabolomic signature for the early prediction of immune response after ART in HIV-infected patients with a low nadir CD4$^+$ T-cell count.

Most of the metabolomic biomarkers obtained in the present study rely directly on the number of HDL particles. The role of HDL in immunity has been studied fairly extensively and several beneficial effects have been attributed to this lipoprotein class[36,37]. Notably, proteomics studies have revealed that HDL components exert regulatory functions on the immune system[37]. Accordingly, it has been accepted that HDL has an important role in host defence, contributing to both innate and adaptive immunity. As an example, apoliprotein A-I (apoA-I), the principal protein of HDL, impairs HIV fusion, thus preventing HIV cell penetration[38]. Moreover, this protein positively correlates with CD4$^+$ T-cell count[39]. However, the most described function of HDL lipoproteins is its anti-atherogenic role due to its ability to transport excess cellular cholesterol to the liver for excretion. A number of studies have described HIV-infected patients as having lower HDL-C and fewer HDL particles, especially large HDL. On the contrary, VLDL and LDL particles have the opposite behaviour, as they contribute to the risk of developing cardiovascular disease (CVD), which is higher in subjects with chronic inflammatory diseases such as HIV[40–42]. In this sense, smaller sizes of LDL particles have also been linked to a higher risk of CVD in epidemiological studies[43–45]. This is in accordance with our results, since bigger LDL sizes increased further in the IR metabolism at baseline, suggesting a relationship between a lower atherogenic lipid profile and the prediction of immunological recovery.

During inflammation and infection, serum triglycerides and VLDL levels increase, which in turn has an effect on other lipoproteins, such as an increase in the production of small, dense LDL and a decrease in the production of HDL[46]. This metabolomic mechanism completely agrees with our results, showing a positive correlation between the medium subclass of VLDL particles and the ratios of ‘total/HDL particles’ and ‘LDL/HDL particles’, high values of which are an indicator of non-recovery. Moreover, a meaningful negative link between ‘LDL/HDL particles’ ratios and CD4$^+$ T-cell recovery has already been reported in a previous study, in which total particles (including VLDL class) were not measured[33]. Even if several beneficial effects to immunity of HDL particles, as well as damaging effects of LDL and VLDL particles have already been reported, no predictive role have been attributed to them so far.

HIV-infected patients also have alterations in glucose metabolism. Several studies have reported glucose-associated disorders including insulin resistance and diabetes mellitus. Glucose metabolism plays a fundamental role in supporting the growth, proliferation and effector functions of T cells[47,48]. Several studies have demonstrated that HIV-infected patients have an increased glycolytic metabolism in CD4$^+$ T-cells because activated immune cells consume glucose at an extremely high rate. Consequently, high plasma levels of this metabolite are associated with a low CD4$^+$ T-cell count[49–51]. Glucose could therefore be a confounder to immunological recovery due to its correlation with CD4$^+$ T-cell count.

In summary, our experimental data establishes that HIV-infected patients with a baseline metabolomic pattern characterized by high levels of HDL particles (especially the subclass 'large'), including HDL cholesterol and big sizes of LDL particles, will have a better immunological recovery after treatment. On the other hand, patients with high ratios of non-HDL lipoprotein particles, high levels of VLDL particles (especially the subclass 'medium') and high concentrations of glucose will not fully recover CD4$^+$ T-cells. However, given the relatively small sample size in the present study, further studies with larger cohorts are needed to confirm the strength of the proposed predictive model.

3.3.6 Conclusions

This study confirms the potential of metabolomics for biomarker discovery in critical illnesses. We have identified a metabolomic signature for HIV-infected patients, mainly HDL-related parameters such as large HDL particles, that relate to their differential immunological response after ART treatment. Therefore, our work links the baseline metabolomic profile to the immunological response after ART treatment, suggesting that an adjustment in the baseline metabolomic pattern could improve the immunological outcome of HIV-infected patients with a low nadir of CD4$^+$ T-cell count. Accordingly, this study provides new insights into HIV pathogenesis and may point the way to the development of new diagnostic, prognostic and therapeutic strategies for HIV, such as lipoprotein composition, structure and function.

3.3.7 Acknowledgments

This work was partially supported by grants from the Fondo de Investigacion Sanitaria (PI10/2635, PI13/0796, PI13/01912, PI14/01693, PI14/0700, PI14/0063 and PI16/00503) Instituto de Salud Carlos III; Fondos Europeos para el Desarrollo Regional (FEDER); Programa de Suport als Grups de Recerca AGAUR (2014SGR250 and 2009 SGR 1061); Gilead Fellowship Program GLD13/00168 and GLD14/293; the Junta de Andalucía, Consejería de Economía, Innovación, Ciencia y Empleo (Proyecto de Investigación de Excelencia; CTS2593); Red de Investigación en Sida (RIS): RD12/0017/0002, RD12/0017/0005, RD17/0017/0014, RD17/0017/0029 and RD16/0025/0006) Instituto de Salud Carlos III, Spain. FV and PD are supported by a grant from the Programa de Intensificación de Investigadores, Instituto de Salud Carlos III (INT11/240, INT12/ 282 and INT12/383, INT13/232, INT15/226 IT15/00140). Y.M. Pacheco was supported by the Fondo de Investigación Sanitaria through the ‘‘Miguel Servet’’ program [CPII13/00037], and by the Consejería de Salud y Bienestar Social of Junta de Andalucía through the ‘‘Nicolás Monardes’’ program [C-0010/13].

3.4 REFERENCES

1. Patti, G. J., Yanes, O. & Siuzdak, G. Metabolomics Science Rev.
2. Zamboni, N., Saghatelian, A. & Patti, G. J. Defining the Metabolome: Size, Flux, and Regulation. *Mol. Cell* **58,** 699–706 (2015).
3. DeBerardinis, R. J. *et al.* Beyond aerobic glycolysis: transformed cells can engage in glutamine metabolism that exceeds the requirement for protein and nucleotide synthesis. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 19345–50 (2007).
4. Zamboni, N., Fendt, S.-M., Rühl, M. & Sauer, U. (13)C-based metabolic flux analysis. *Nat. Protoc.* **4,** 878–92 (2009).
5. Liu, W. *et al.* Reprogramming of proline and glutamine metabolism contributes to the proliferative and metabolic responses regulated by oncogenic transcription factor c-MYC. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 8983–8 (2012).

6. Fan, T. W.-M. & Lane, A. N. Structure-based profiling of metabolites and isotopomers by NMR. *Prog. Nucl. Magn. Reson. Spectrosc.* **52,** 69–117 (2008).

7. Fan, T. W.-M. & Lane, A. N. NMR-based stable isotope resolved metabolomics in systems biochemistry. *J. Biomol. NMR* **49,** 267–80 (2011).

8. Lane, A. N. & Fan, T. W.-M. Quantification and identification of isotopomer distributions of metabolites in crude cell extracts using 1H TOCSY. *Metabolomics* **3,** 79–86 (2007).

9. Le, A. *et al.* Glucose-independent glutamine metabolism via TCA cycling for proliferation and survival in B cells. *Cell Metab.* **15,** 110–21 (2012).

10. Gómez, J. *et al. 9th International Conference on Practical Applications of Computational Biology and Bioinformatics*. *9th International Conference on Practical Applications of Computational Biology and Bioinformatics Advances in Intelligent Systems and Computing* **375,** (Springer International Publishing, 2015).

11. Akoka, S., Barantin, L. & Trierweiler, M. Concentration Measurement by Proton NMR Using the ERETIC Method. *Anal. Chem.* **71,** 2554–7 (1999).

12. Haug, K. *et al.* MetaboLights--an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res.* **41,** D781–6 (2013).

13. Gómez, J. *et al.* Dolphin: a tool for automatic targeted metabolite profiling using 1D and 2D (1)H-NMR data. *Anal. Bioanal. Chem.* **406,** 7967–76 (2014).

14. Moseley, H. N. B., Lane, A. N., Belshoff, A. C., Higashi, R. M. & Fan, T. W. M. A novel deconvolution method for modeling UDP-N-acetyl-D-glucosamine biosynthetic pathways based on (13)C mass isotopologue profiles under non-steady-state conditions. *BMC Biol.* **9,** 37 (2011).

15. Battegay, M., Nüesch, R., Hirschel, B. & Kaufmann, G. R. Immunological recovery and antiretroviral therapy in HIV-1 infection. *Lancet. Infect. Dis.* **6,** 280–7 (2006).

16. Corbeau, P. & Reynes, J. Immune reconstitution under antiretroviral therapy: the new challenge in HIV-1 infection. *Blood* **117,** 5582–90 (2011).

17. Pacheco, Y. M. *et al.* Increased risk of non-AIDS-related events in HIV subjects with persistent low CD4 counts despite cART in the CoRIS cohort. *Antiviral Res.* **117,** 69–74 (2015).

18. Engsig, F. N. *et al.* Long-term Mortality in HIV-Positive Individuals Virally Suppressed for &gt;3 Years With Incomplete CD4 Recovery. *Clin. Infect. Dis.* **58,** 1312–1321 (2014).

19. Mocroft, A. *et al.* Risk factors and outcomes for late presentation for HIV-positive persons in Europe: results from the Collaboration of Observational HIV Epidemiological Research Europe Study (COHERE). *PLoS Med.* **10,** e1001510 (2013).

20. Pacheco, Y. M. *et al.* Risk factors, CD4 long-term evolution and mortality of HIV-infected patients who persistently maintain low CD4 counts, despite virological response to HAART. *Curr. HIV Res.* **7,** 612–9 (2009).

21. Massanella, M., Negredo, E., Clotet, B. & Blanco, J. Immunodiscordant responses to HAART--mechanisms and consequences. *Expert Rev. Clin. Immunol.* **9,** 1135–49 (2013).

22. Wishart, D. S. Emerging applications of metabolomics in drug discovery and precision medicine. *Nat. Rev. Drug Discov.* (2016). doi:10.1038/nrd.2016.32

23. Zhang, A. *et al.* Metabolomics for Biomarker Discovery: Moving to the Clinic. *Biomed Res. Int.* **2015,** 1–6 (2015).

24. Mallol, R. *et al.* Liposcale: a novel advanced lipoprotein test based on 2D diffusion-ordered 1H NMR spectroscopy. *J. Lipid Res.* **56,** 737–46 (2015).

25. Otvos, J. D. *et al.* GlycA: A Composite Nuclear Magnetic Resonance Biomarker of Systemic Inflammation. *Clin. Chem.* **61,** 714–23 (2015).

26. Nicholson, J. K., Buckingham, M. J. & Sadler, P. J. High resolution 1H n.m.r. studies of vertebrate blood and plasma. *Biochemical Journal* **211,** 605–615 (1983).

27. Barton, R. H. *et al.* The influence of EDTA and citrate anticoagulant addition to human plasma on information recovery from NMR-based metabolic profiling studies. *Mol. Biosyst.* **6,** 215–24 (2010).

28. Krzywinski, M. & Altman, N. Points of Significance: Visualizing samples with box plots. *Nat. Methods* **11,** 119–120 (2014).

29. Hadigan, C. *et al.* Metabolic abnormalities and cardiovascular disease risk factors in adults with human immunodeficiency virus infection and lipodystrophy. *Clin. Infect. Dis.* **32,** 130–9 (2001).

30. Bergersen, B. M., Schumacher, A., Sandvik, L., Bruun, J. N. & Birkeland, K. Important differences in components of the metabolic syndrome between HIV-patients with and without highly active antiretroviral therapy and healthy controls. *Scand. J. Infect. Dis.* **38,** 682–9 (2006).

31. Mondy, K. *et al.* Metabolic syndrome in HIV-infected patients from an urban, midwestern US outpatient population. *Clin. Infect. Dis.* **44,** 726–34 (2007).

32. Moore, D. M. *et al.* Effect of baseline CD4 cell counts on the clinical significance of short-term immunologic response to antiretroviral therapy in individuals with virologic suppression. *J. Acquir. Immune Defic. Syndr.* **52,** 357–63 (2009).

33. Azzoni, L. *et al.* Metabolic and anthropometric parameters contribute to ART-mediated CD4+ T cell recovery in HIV-1-infected individuals: an observational study. *J. Int. AIDS Soc.* **14,** 37 (2011).

34. Hollenbaugh, J. A., Munger, J. & Kim, B. Metabolite profiles of human immunodeficiency virus infected CD4+ T cells and macrophages using LC-MS/MS analysis. *Virology* **415,** 153–9 (2011).

35. Sitole, L. J., Williams, A. A. & Meyer, D. Metabonomic analysis of HIV-infected biofluids. *Mol. Biosyst.* **9,** 18–28 (2013).

36. Pirillo, A., Catapano, A. L. & Norata, G. D. in *Handbook of Experimental Pharmacology* **224,** 483–508 (2015).

37. Yu, B., Wang, S., Peng, D. & Zhao, S. HDL and immunomodulation: an emerging role of HDL against atherosclerosis. *Immunol. Cell Biol.* **88,** 285–290 (2010).

38. Srinivas, R. V *et al.* Antiviral effects of apolipoprotein A-I and its synthetic amphipathic peptide analogs. *Virology* **176,** 48–57 (1990).

39. Rose, H. *et al.* HIV infection and high density lipoprotein metabolism. *Atherosclerosis* **199,** 79–86 (2008).

40. Kontush, A. HDL particle number and size as predictors of cardiovascular disease. *Front. Pharmacol.* **6,** 218 (2015).

41. Duprez, D. A. *et al.* Lipoprotein particle subclasses, cardiovascular disease and HIV infection. *Atherosclerosis* **207,** 524–9 (2009).

42. Mora, S. *et al.* Lipoprotein particle profiles by nuclear magnetic resonance compared with standard lipids and apolipoproteins in predicting incident cardiovascular disease in women. *Circulation* **119,** 931–9 (2009).

43. Lamarche, B. *et al.* Small, dense low-density lipoprotein particles as a predictor of the risk of ischemic heart disease in men. Prospective results from the Québec Cardiovascular Study. *Circulation* **95,** 69–75 (1997).

44. Gardner, C. D., Fortmann, S. P. & Krauss, R. M. Association of small low-density lipoprotein particles with the incidence of coronary artery disease in men and women. *JAMA* **276,** 875–81 (1996).

45. Stampfer, M. J. *et al.* A prospective study of triglyceride level, low-density lipoprotein particle diameter, and risk of myocardial infarction. *JAMA* **276,** 882–8 (1996).

46. Feingold, K. R. & Grunfeld, C. *The Effect of Inflammation and Infection on Lipids and Lipoproteins*. *Endotext* (2000).

47. Palmer, C. S., Ostrowski, M., Balderson, B., Christian, N. & Crowe, S. M. Glucose metabolism regulates T cell activation, differentiation, and functions. *Front. Immunol.* **6,** 1 (2015).

48. Newsholme, E. A., Crabtree, B. & Ardawi, M. S. The role of high rates of glycolysis and glutamine utilization in rapidly dividing cells. *Biosci. Rep.* **5,** 393–400 (1985).

49. Misra, R. *et al.* Relationship of ethnicity and CD4 Count with glucose metabolism among HIV patients on Highly-Active Antiretroviral Therapy (HAART). *BMC Endocr. Disord.* **13,** 13 (2013).

50. Palmer, C. S. *et al.* Increased glucose metabolic activity is associated with CD4+ T-cell activation and depletion during chronic HIV infection. *AIDS* **28,** 297–309 (2014).

51. McKnight, T. R. *et al.* A combined chemometric and quantitative NMR analysis of HIV/AIDS serum discloses metabolic alterations associated with disease status. *Mol. Biosyst.* **10,** 2889–97 (2014).

CHAPTER IV

GENERAL DISCUSSION

The high complexity of NMR spectra of biofluids difficults metabolite identification and quantification tasks. Usually, users have to deal with signal overlapping, signal position shifting and distortions in the signal shapes among other characteristic issues inherent in NMR spectra of biofluids. Even if all the steps of the NMR-based metabolomics analysis workflow previous to the final spectra have been carefully optimized and executed, the final extraction of the metabolite profile in NMR samples will present, at least, signal overlap.

Metabolite identification process is carried out by checking all the resonances belonging to the same metabolite signature. Due to signal overlap, some of these resonances can be masked by neighboring signals, leading to the loss of their characteristics in the final shape of the spectra. Moreover, to quantify a metabolite, the area of one of its resonances has to be calculated, and the only way to do so is correctly modelling its shape. In this process, if the most isolated signal of a metabolite signature is in a crowed region, bucket integration of the signal region may produce quantification errors since the area of other resonances is contributing the final shape, and to correctly fit our signal of interest, line-shape fitting algorithms are needed. Due to the issues mentioned above, we aimed to design a strategy to achieve the first main objective of this thesis, which consisted in developing and evaluating a set of algorithms to match and quantify a set of target signals in a completely automated manner combining 1D (NOESY and CPMG) spectra with its own 2D JRES complementary spectra.

Dolphin is a line-shape-fitting package for high throughput automated metabolite profiling in NMR spectra. Dolphin takes profit of the 2-dimensional JRES spectra, where signals are much less superposed, in order to improve the identification and quantification processes. In those spectra, the congestion produced by the multiplicity of signals disappears, since they are projected in an orthogonal dimension. The orthogonal cut at the position of the signal of interest allows the algorithm to compare the multiplicity and j-coupling of the signal placed there with an internal library, and if all the resonances in the library match with a determined compound, a metabolite is considered identified.

In a subsequent step, Dolphin focalizes the most isolated signal of a targeted metabolite and finds the neighboring signals that can affect its quantification. In this process, Dolphin does an automated annotation of the neighboring signals in terms of position, multiplicity and j-coupling. This process is done despite the neighboring signals identity (known or unknown compounds). Finally, it performs a line-shape fitting of that region, adjusting the intensities of all the signals present and modelling their shapes as lorentzian-gaussian functions.

Dolphin's performance is evaluated comparing its results against two of the most used methods in NMR metabolite profiling: reference deconvolution (using Chenomx NMR Suite) and bucket integration (using Bioref AMIX). More concretely, we compared their capability to correctly profile a set of 15 metabolites in a pull of standards and in liver aqueous extracts of rats, and a set of 5 metabolites in human urine samples.

Results by the automated profiling of Dolphin were in good agreement with those produced by the manual profiling using Chenomx. The two packages performed well in regions with signal overlap, where bucket integration gave the worst results due to the integration of part of the neighboring signals surrounding the signal of interest. However, bucket integration successfully quantified isolated peaks with intensities clearly higher than background noise. In fact, due to the shape constraints of its spectral library, the reference deconvolution of Chenomx failed in most of these cases, while the flexibility of the lorentzian-gaussian proportions included in Dolphin's line-shape fitting algorithm shown a better performance.

In summary, Dolphin has proven to be an excellent solution for the automated profiling of metabolites in isolated and partially overlapped regions of NMR spectra of tissue extracts and in the same kind of regions in urine. Moreover, it was able to adjust the position of citrate in urine samples, which is known for its capacity of shifting position along samples. However, the main advantage of Dolphin is also its main limitation, since it needs a 2-dimensional extra spectrum for each sample to perform the automated metabolite profiling. Moreover, even if the second dimension reduces signal overlapping produced by the multiplicity of resonances, it is still very high in regions with a huge number of resonances. To overcome this, more complex filters of signal discriminators need to be implemented now that the algorithm's strategy has demonstrated to work well.

Once at this point, we decided to change Dolphin's strategy to do not depend on extra acquisitions and to allow the profiling of metabolites in highly crowded regions, where current filters of automated signal annotation using the 2D JRES spectra were not able to solve the profiling. To achieve this, we aimed at developing and evaluating a set of algorithms to allow the user interacting with the spectra and to use different automated quantification modes under a user-friendly and intuitive GUI, avoiding the necessity of 2D JRES complementary spectra to perform the profiling. This new strategy includes finding an agreement between editability and automation, minimizing user-subjectivities but avoiding black-box processes at the same time in a useful and versatile tool. Moreover, we decided to include functions able to combine fingerprinting approaches and to import reference spectra from public databases to perform the most reliable NMR profiling.

Whale is the package version that allows users approaching metabolomics NMR data in different ways depending on the final goal of their studies. Whale was born to reach the second and third main objectives of this thesis, and it incorporates several functions for exploring data and manually adjusting quantification parameters in order to run automated metabolite profiling along samples. The core of the tool is based on automated quantification of targeted metabolites through easily editable ROI patterns. To facilitate metabolite assignments, it offers metabolite suggestions depending on the spectral regions, and a repository panel where the user can compare target spectra with reference spectra from public databases. To facilitate metabolite quantification in large datasets, it offers up to four different quantification modes in order to optimize the time-span of the analysis. It is also easy to interpret Whale outputs, allowing users to detect where the algorithm failed and re-run the analysis only for those spectra where the algorithm failed.

All Whale functions were detailed and illustrated with examples, and two studies where the tool was applied were exposed. The first presents a novel approach for NMR-based stable isotope tracer studies called PEPA. In Whale's algorithm, each resonance is quantified without taking into account the intensity ratios between the other resonances that belong to the NMR signature of the pure compound under normal conditions, in contrast with the algorithms that are commonly used in other metabolite profiling packages that are based in reference deconvolution. In this sense, neither Chenomx nor BATMAN could have performed this analysis, since in their strategy all the resonances of a determined compound are necessarily linked to each other in terms of intensity signal ratios. Of course, BAYESIL could not be used not only for the same reason, but also because is not suited for biofluids that are not ultra-filtered plasma, serum or CSF. Thus, it is important to highlight that due to the nature of PEPA, where the intensity loss of each resonance of a unique compound can be affected by a different ratio of proton enrichment, quantification based in reference deconvolution would not have worked, being Whale's strategy a crucial part of the process.

All positional ^{13}C-enrichments calculated by PEPA using Whale quantifications in ^{1}H-NMR spectra were confirmed using ^{13}C-NMR, TOCSY, HSQC or HMBC, which endorse the capability of Whale to successfully profile metabolites in ^{1}H-NMR spectra. Only two enrichments observed in the ^{13}C-NMR spectra were not anticipated using PEPA, due to the differences in the quantifications of these protons did not reach statistical significance. In these cases, is important to highlight that the biological variability of the model and a limited number of replicates may produce false negatives since PEPA relies on statistical comparisons to establish whether or not there is an enrichment.

The other study consists in a clinical study where NMR-based metabolomics is the methodology used for finding novel biomarkers of CD4$^+$ T-cell recovery in adult HIV-infected patients with pre-ART low nadir of CD4$^+$ T-cell count after 3 years of successfully ART treatment. Even if metabolic data of only 64 patients were used in this study due to the exclusion criteria, Whale was used to obtain the metabolite profile of more than 200 human plasma samples. Anecdotally, using the function 'correlate samples to metadata variable' included in the fingerprinting panel of Whale, we detected a labeling error in samples in the initial database, that could be corrected taking advantage of the high correlation that normally exist between glucose resonances in NMR spectra and glucose levels calculated by biochemical parameters in the lab.

In this case, and due to the low quality of the dataset spectra (human plasma samples without any dilution neither centrifugation that came from different hospitals), only eleven metabolites, a peak related to glycoprotein concentration in blood and two EDTA peaks previously reported as indicators of calcium and magnesium levels in blood were profiled.

A set of 8 baseline metabolic variables was able to classify patients with pre-ART low nadir of CD4$^+$ T-cell count according to their immunological response after 3 years of treatment. The fact that none of the metabolic variables presented enough classification power by itself suggest that there is a complex panel of metabolic interactions underlying the relation between metabolism and immunological recovery. However, the nature of these variables indicates that the most important interactions rely on the number of HDL particles and the glucose metabolism.

In our cohort, these metabolic differences resulted more important than two of the most reported evidences related to the immune-discordant drug response which are age and number of CD4$^+$ T-cell count at baseline. This work was a pilot study where we applied metabolomics for the discovery of new biomarkers in diagnosis and prognosis of HIV infection. In that sense, we proposed some baseline metabolic variables which in a multivariate model could predict the discordant response to ART in adult HIV-infected patients with pre-ART low nadir of CD4$^+$ T-cell count. The main limitation of this study is the relatively small size of the cohort used.

CHAPTER V

GENERAL CONCLUSIONS

During this thesis, several algorithms for extracting valuable metabolic information from ^{1}H-NMR spectra have been developed, evaluated and embedded into two software packages: Dolphin and Whale. The first package, Dolphin, aimed to find and quantify a set of target signals in a completely automated manner combining 1D (NOESY and CPMG) spectra with its own 2D JRES complementary spectra. The second package, Whale, aimed to be an interactive package for approaching data using different fingerprinting functions and giving users automated modes of signal quantification for metabolite profiling.

Dolphin's strategy proved to be an excellent solution for the automated annotation of signals present in regions with moderated signal congestion. It was able to discriminate signals according to their exact position, their multiplicity and their j-coupling using 2 dimensional JRES spectra, and perform an automated annotation of all of them for a posterior deconvolution of the region in 1 dimensional ^{1}H-NMR spectra using its line-shape fitting algorithm. It is important to highlight that in all this process all signals surrounding the target signal were taken into account even if they were unknown compounds. Its quantification accuracy was evaluated using a pull of standards at known concentrations and a spike-in experiment in human urine samples. Moreover, its performance in biological samples was compared with two of the most used methods for metabolite quantification, manual reference deconvolution and bucket integration, represented by two of the most established packages, Chenomx and AMIX respectively. In all cases Dolphin demonstrated high quantification accuracy, validating in this way the power of its line-shape fitting algorithm. However, due to the extremely high complexity of some highly congested regions of ^{1}H-NMR spectra, more complex signal discriminator algorithms need to be developed and implemented in order to reach the number of metabolites that can be manually profiled.

This necessity of expanding the number of metabolites able to be profiled in an automated manner inspired the design of Whale. As an automated annotation of all the metabolites present in biological samples was almost impossible to achieve in a short period of time, a new strategy was developed and implemented based on automated functions for signal quantification under a user-friendly GUI where users could edit profiling options and apply fingerprinting approaches.

In the first study where Whale was applied (chapter 3 section 3.6.1), the package was used to develop a new methodology called PEPA, a one-dimensional ^{1}H-NMR approach for ^{13}C stable isotope resolved metabolomics. Whale was able to quantify a set of 84 different resonances underlying 46 polar metabolites in ^{1}H-NMR spectra of aqueous extracts of human cancer cells. Metabolites were graphically identified by expert users and a total of 55 ROI patterns were generated using a mean spectrum before running the quantifications in the whole dataset. The images of the spectral deconvolution in each ROI generated by the package allowed users to graphically check the capability of the tool to correctly adjust the shape of each resonance. Moreover, the results obtained in PEPA using Whale's quantifications were confirmed using ^{13}C-NMR, TOCSY, HSQC or HMBC. These confirmations demonstrated that Whale is a tool able to perform reliable automated quantifications in ^{1}H-NMR spectra using the ROI pattern files generated by users in a semi-automated strategy for open-targeted metabolite profiling. In this study, PEPA proved to be an excellent approach to study the position of carbon labels in large datasets, taking advantage of the easy-implementation of high-throughput analyses using only ^{1}H-NMR, instead of other 2D-edited experiments. This new methodology simplifies NMR untargeted carbon flux monitoring and can be positioned as a key element of NMR-based SIRM in cell metabolism.

In the second study where Whale was applied (chapter 3 section 3.6.2), the package was used to profile low molecular weight metabolites of adult HIV-infected individuals with the aim of finding a metabolomic signature able to predict their immunological response to ART. More concretely, Whale quantified a target set of 11 resonances underlying 11 metabolites in ^{1}H-NMR CPMG spectra of human plasma samples. Apart from these 11 metabolites and taking advantage of the tool versatility, two EDTA peaks and another peak of interest were also quantified. The results obtained using Whale were put together with the results obtained using the Liposcale test, and a total number of 40 metabolomic variables were used to perform the statistical analyses. For these analyses, several algorithms of univariate and multivariate statistics such as fold change calculations, unpaired mann-withney U test, noxched boxplots, linear regression models, random forest analyses and ROC curves were implemented in R in order to convert metabolite quantifications into significant and interpretable results. Finally, a metabolomics signature conformed by HDL-related parameters and glucose, with great emphasis in the number of large HDL particles as maximum differentiator, was identified for adult HIV-infected patients in relation to their differential response to ART. This study confirmed the potential of metabolomics for biomarker discovery in critical illnesses and suggested that an adjustment of the baseline metabolomic pattern could improve the immunological outcome of adult HIV-infected patients with pre-ART low nadir of CD4^{+} T-cell counts.

In summary, the algorithms presented in this thesis demonstrated to be necessity to improve metabolite profiling in ^{1}H-NMR samples. On one hand, Dolphin presented a new methodology for automated metabolite profiling combining 2D JRES data with 1D ^{1}H-NMR data, an original approach that can be improved with advances in NMR acquisitions and in algorithmic filters able to discriminate signals in highly overlapped regions. On the other hand, Whale presented a different strategy based on an agreement between user-interactivity and automatisms with promising results for the profiling of metabolites in several datasets of ^{1}H-NMR spectra of different biofluids. Future work include coding Whale in R language.

ANNEXES

WHALE'S USER'S MANUAL

1 Importing data:

1r and 2rr Bruker files can be imported by our tool using the following kind of inputs (all the following excel files must have only one sheet):

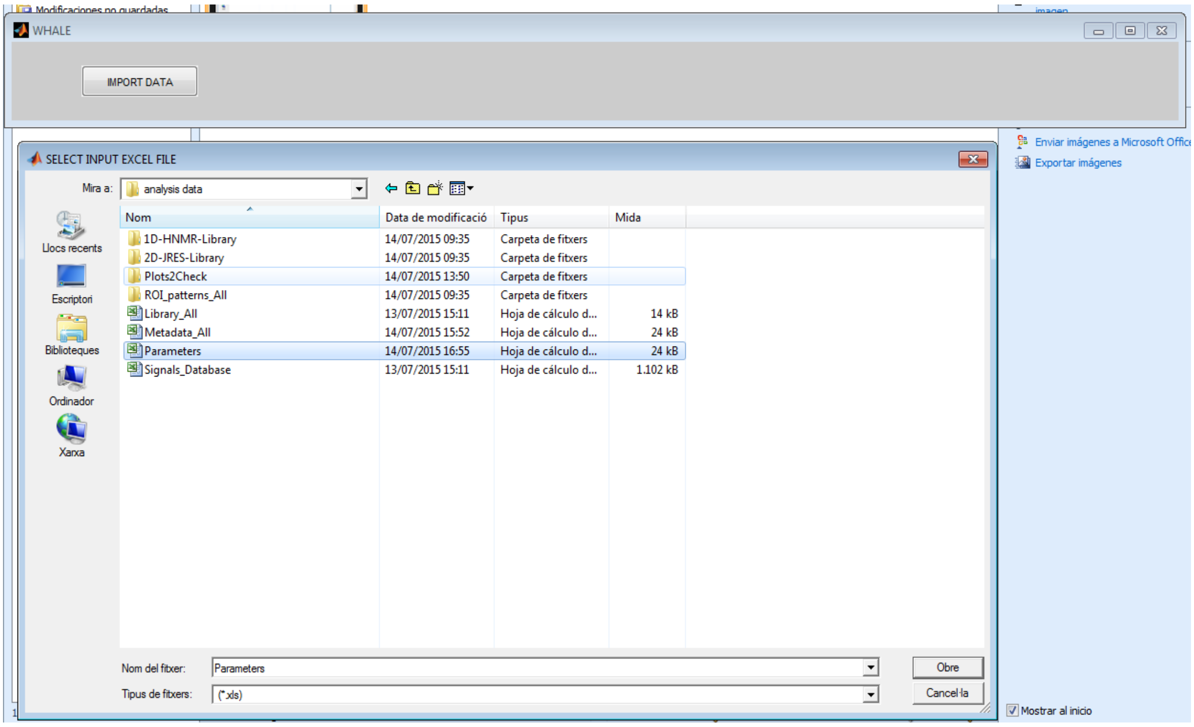
- Metadata: Excel file with at least 2 columns, the first of which containing the sample name and the rest with metadata information about the experiment (use numbers, not characters). The first row of each column will be processed as a title. If there is no metadata available, the second column has to be filled e.g. with random numbers.
- Metabolite library: Excel file with 1 column containing the names of target metabolites (the package includes different default libraries optimized depending upon the target biological matrix, but users can adapt the library to the experiment needs). More information in section 4 (Editing libraries and ROI patterns).
- Folder with ROI Patterns: 1 Excel file per ROI. More information in section 4 (Editing libraries and ROI patterns).
- Parameters: Excel file which contains the necessary parameters to import data contained in Bruker files, with additional parameters to fit the tool to users' needs. This file has two columns, the first of which contains the parameter name and the second the parameter value filled by users. A more detailed description is shown in Table A1:

Parameter	**Value of the parameter**
nmr folder path	Path which contains the 'nmr' folder where the Bruker files are contained. Samples must be contained in a folder called 'nmr'.
Metadata path (xls format)	'path\metadata_file_name.xls'
1D data index	Number of the data index (10, 11, 12...)
Library path (xls format)	'path\library_file_name.xls'

ROI_patterns folder path	Path which contains the 'ROI_patterns' folder
Plots2Check folder path	Path where you want to create the 'Plots2Check' folder
Fitting error threshold (in %)	Error value above which saving an image (10, 20…). More info in section 8 (The output).
Signal area threshold (in %)	Percentage of contribution of the signal to the total fitted area under which saving an image (30, 50…). More info in section 8 (The output).
Normalization (1=Eretic, 2=TSP, 3=No)	Value between 1-3 according to the choice
Alignment (1=Glucose, 2=TSP, 3=Formate)	Value between 1-3 according to the choice
Suppression (1=Water, 2=EDTA, 3=Urea, 4=Glucose)	Value or values (separated by comma) between 1-4 according to the choice
Spectrometer Frequency (Hz)	Spectrometer frequency value (600, 700…)
2D-JRES mode (1=No, 2=One/Sample, 3=One as reference)	Value between 1-3 according to the choice
2D-JRES data index	Number of the data index (10, 11, 12...)
2D-JRES (2rr file) reference path (Only if 2D-JRES mode is set to 3)	Path which contains the J-Res-NMR spectrum used as reference.
2D-JRES library path	Path where the library of J-Res-NMR spectrum of pure compounds is contained.
1D-HNMR library path	Path where the library of ^{1}H-NMR spectrum of pure compounds is contained.
Signals repository path (xls format)	Path where the excel file with a database of signals is contained.
Bucket resolution	Bucket resolution (in ppms).

Table A1: Importing parameters

To start running the package, users must press the button ‘IMPORT DATA’ in the main panel. Once the browser is open, users must search for the Parameters file and select it (Figure A1).



Mira a: analysis data

Nom	Data de modificació	Tipus	Mida
1D-HNMR-Library	14/07/2015 09:35	Carpeta de fitxers	
2D-JRES-Library	14/07/2015 09:35	Carpeta de fitxers	
Plots2Check	14/07/2015 13:50	Carpeta de fitxers	
ROI_patterns_All	14/07/2015 09:35	Carpeta de fitxers	
Library_All	13/07/2015 15:11	Hoja de cálculo d...	14 kB
Metadata_All	14/07/2015 15:52	Hoja de cálculo d...	24 kB
Parameters	14/07/2015 16:55	Hoja de cálculo d...	24 kB
Signals_Database	13/07/2015 15:11	Hoja de cálculo d...	1.102 kB

Nom del fitxer: Parameters

Tipus de fitxers: (*.xls)

Figure A1: selection of the Parameters file that contains the necessary information.

Once all data have been imported successfully, six new buttons will appear in the main panel (‘FINGERPRINT’,’REPOSITORY, ‘ROIs TESTING’, ‘AUTO RUN’, ‘SAVE RESULTS’ and ‘EXIT’) (Figure A2) and a new folder called ‘Plots2Check’ in the path selected within the parameters file.

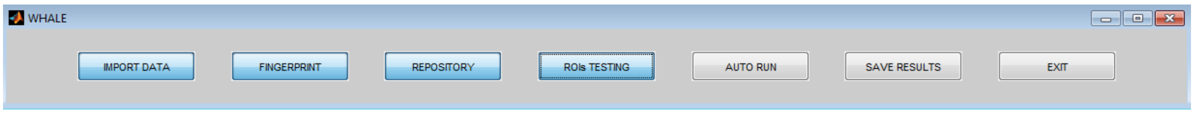


Figure A2: main panel of the program.

2 Fingerprinting:

The Fingerprint panel allows users to explore data using some options before beginning a more exhaustive analysis:

- ‘CORRELATE SAMPLES 2 REGION’: This option allows users evaluating if there are regions in the spectra with high variance correlation with a region of interest through the introduction of the left and right bounds of the region of interest. This is very useful for e.g. checking if a signal belongs to a metabolite by its correlation with other signals of the same metabolite located in other regions of the spectrum.
- ‘CORRELATE SAMPLES 2 METADATA VARIABLE’: This option allows users evaluating if there are regions in the spectra with high variance correlation with metadata selected through the select box on the right. The box on the right contains the metadata variables previously imported in the Metadata file (see section 1 Importing Data). This is very useful for e.g. checking if there is correlation between glucose signals in spectra and glucose quantified by other methods (Figure A3).
- ‘FIND HOTSPOTS BETWEEN GROUPS’: This option allows users detecting regions of the spectrum that present more differences between two subgroups of samples selected through the select boxes on the left and the right of the button. The boxes contain different groups according to the second column of the Metadata file (see section 2 Importing Data).

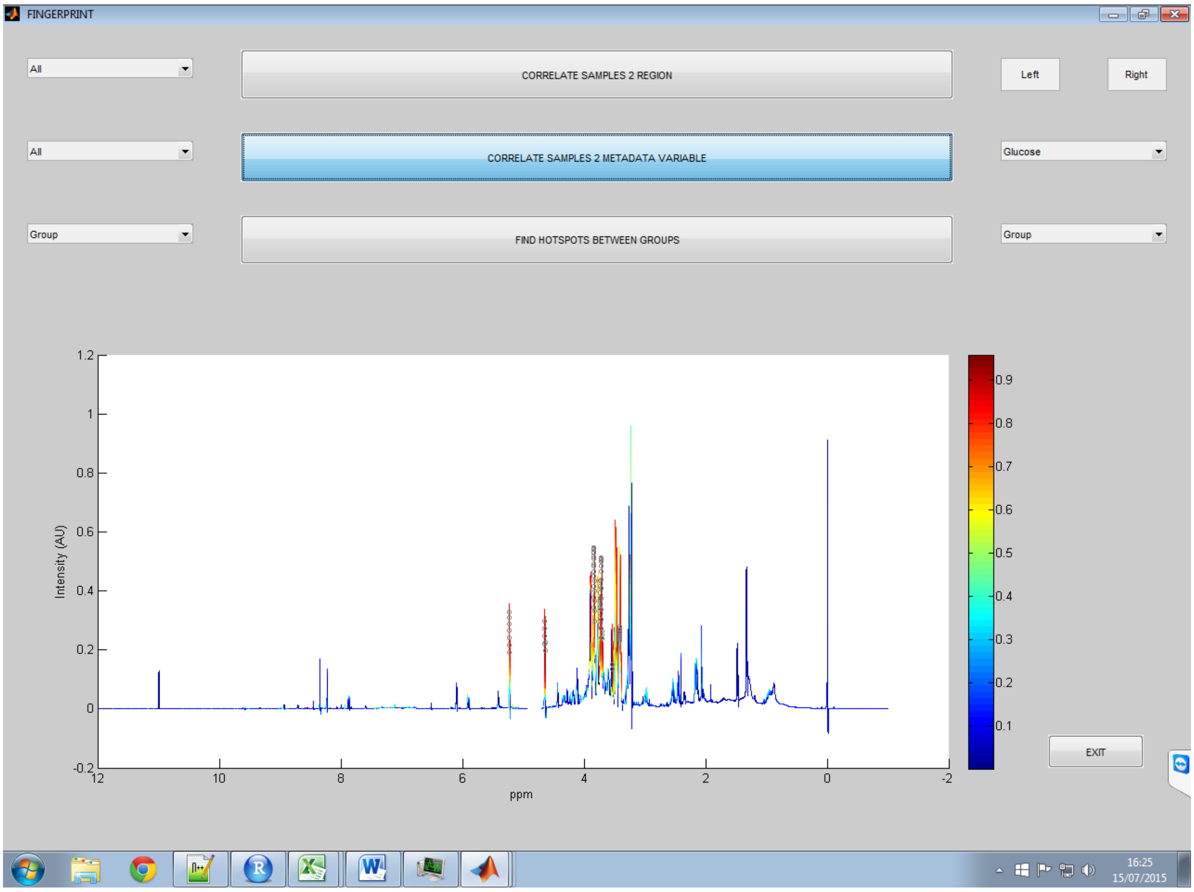


Figure A3: Fingerprint panel. Example of the correlation spectrum of Glucose metadata variable with regions of the spectra dataset.

3 Repository:

The Repository option allows users comparing ^{1}H-NMR and JRES-NMR spectrum of a sample with ^{1}H-NMR and J-RES-NMR spectrum of pure compounds contained in a library.

Users can choose a sample spectrum through the 'Sample' select box and plot the kind of NMR spectrum (^{1}H or JRES) in which they are interested in. If they want to zoom a region, they have to click the left button of the mouse in the region they want to zoom in or select it. If users want to zoom out a region, they have to click Alt and the left button of the mouse. If users want to compare the two kinds of spectrum (e.g. for the identification of the kinds of signals behind a region of the ^{1}H-NMR spectrum, they have to click both 'Plot 1H' and 'Plot JRES' buttons and select the region of interest.

Users can plot the ^{1}H and JRES NMR spectra of pure compounds as a reference to compare them with a sample spectrum in order to facilitate the identification of signals behind an NMR spectrum. They have to select the kind of spectrum they want to use through the 'HNMR Compounds' and 'JRES Compounds' selecting boxes and add other spectra of pure compounds to the generated spectrum through the 'Add 1H Ref' or the 'Add JRES Ref' butons.

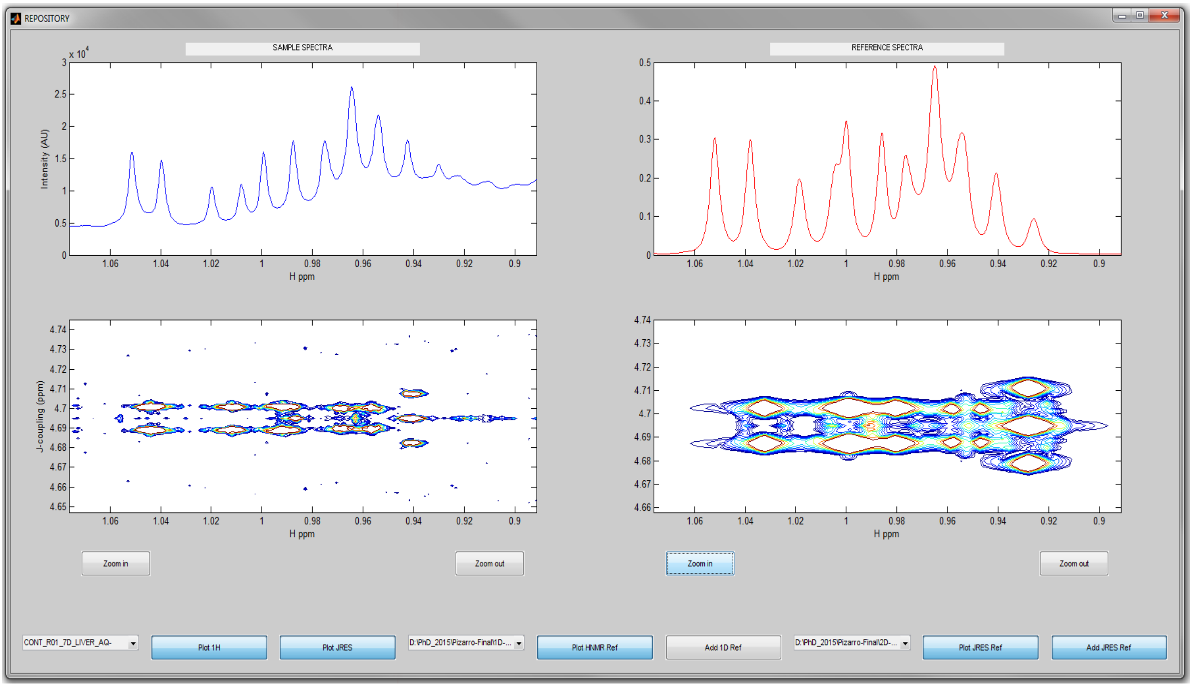


Figure A4: Repository Panel. Example of the Brand-Chained-Amino-Acids (BCAAs) region in our dataset (left axes) and the superposition of the reference spectra of those three BCAAs (Isoleucine, Leucine and Valine) (right axes).

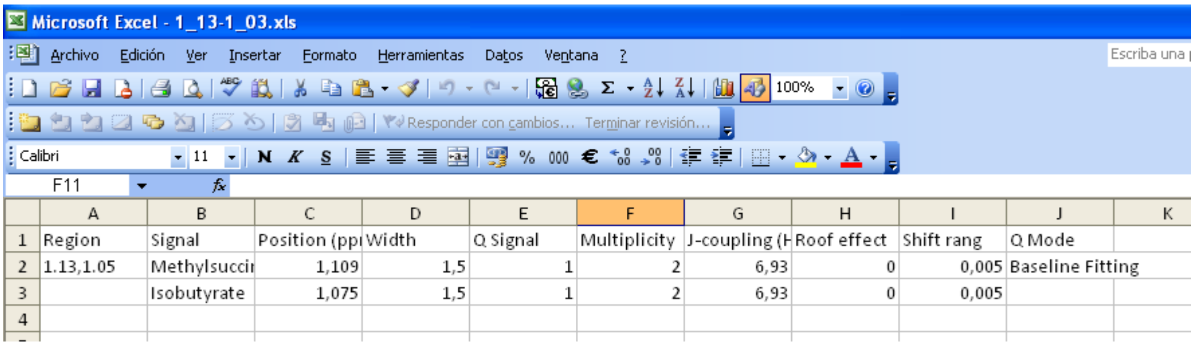
4 Editing libraries and ROI patterns:

Our package comes with a metabolite library and a set of several ROI Patterns prepared for four kinds of matrix: liver extract, brain extract, serum and urine. These libraries and patterns have been built through the analysis of several datasets per matrix using both public and private databases (such as The Human Metabolome Database –HMDB- and Chenomx) and scientific papers (such as The Human Serum Metabolome and The Human Urine Metabolome). Both libraries and ROI patterns are easily editable to adapt the parameters to the target biofluid and the concrete characteristics of samples.

The metabolite library consists in an Excel file where the first column contains the names of all metabolites that users want to quantify (if users want to quantify different signals of the same metabolite, they must give a different name to each one as e.g. Isoleucine1 and Isoleucine2). It is very important that the name of all metabolites or signals in the library matches exactly with those that appear in the ROI pattern files. The first row will be processed as the first metabolite of the library.

ROI pattern files are also Excel files, one per ROI, which contain some parameters of the signals to deconvolve in a ROI. The format of a ROI pattern file is presented in Figure A5. The parameters to give to the program are:

- **Region**: Put left and right bounds of the ROI (in ppm and with period as decimal separator) separated by comma. It is only necessary to fill the cell corresponding to the first signal.
- **Signal**: Put the name of the signal or metabolite. Be sure that the name matches exactly with the name which is annotated in the library.
- **Position (ppm)**: Put the signal position (in ppm and with comma as decimal separator).
- **Width**: Adjust the value of the width of the signal (arbitrary units, with comma as decimal separator).
- **Q Signal**: Put 1 if you want to quantify the signal and 0 if you only need it to adjust the fitting.
- **Multiplicity**: Put the multiplicity of the signal (for now the software supports from singlet (1) to quadruplet (4).
- **J-coupling (Hz)**: Put the J-coupling value (in Hz and with comma as decimal separator).
- **Roof effect**: Put a roof effect value if it is necessary. In some cases (such as citric acid and L-phenylalanine) the signal doesn't follow the conventional proportions, so a positive factor (between 0 and 1) will produce a descending signal and a negative factor (between -1 and 0) will produce an ascending signal.
- **Shift rang**: Give a window (in ppm and with comma as decimal separator) of deviation of the shift from the indicated position. If position is 4 and shift rang 0,005, the program will search the signal between 4.0025 and 3.9975 ppms.
- **Q Mode**: The package works in four modes: Baseline Fitting, Clean Fitting, Baseline Sum and Clean Sum. The "Baseline" mode adjust a putative baseline to perform the quantification; the "Clean" mode will adjust only the signals given by users without including any baseline approximation. The "Fitting" mode will perform a fit of the signals to the spectrum in the determined ROI; the "Sum" mode will integrate the whole region. It is only necessary to fill the cell corresponding to the first signal.


Microsoft Excel - 1_13-1_03.xls
Archivo Edición Ver Insertar Formato Herramientas Datos Ventana ?
Escriba una
100%
Responder con cambios... Terminar revisión...
Calibri
11
F11

	A	B	C	D	E	F	G	H	I	J	K
1	Region	Signal	Position (ppı	Width	Q Signal	Multiplicity	J-coupling (H	Roof effect	Shift rang	Q Mode	
2	1.13,1.05	Methylsuccii	1,109	1,5	1	2	6,93	0	0,005	Baseline Fitting	
3		Isobutyrate	1,075	1,5	1	2	6,93	0	0,005		
4											

Figure A5: Example of an ROI pattern file.

Users can expand and reduce the library and the number of ROI pattern files, and adjust the parameters within the ROI patterns to the experiment needs. We recommend the Sum mode for those regions that contain isolated and not mobile signals because the computation time is severely reduced. Use the Baseline Fitting mode in regions where baseline or broad signals are affecting the final shape of the region.

A Signals Database excel file is provided in order to facilitate editing the metabolite library and the ROI Patterns. This excel file contains information of some parameters for each metabolite present in the HMDB and information about the common compounds in urine, serum, liver extract and brain extract.

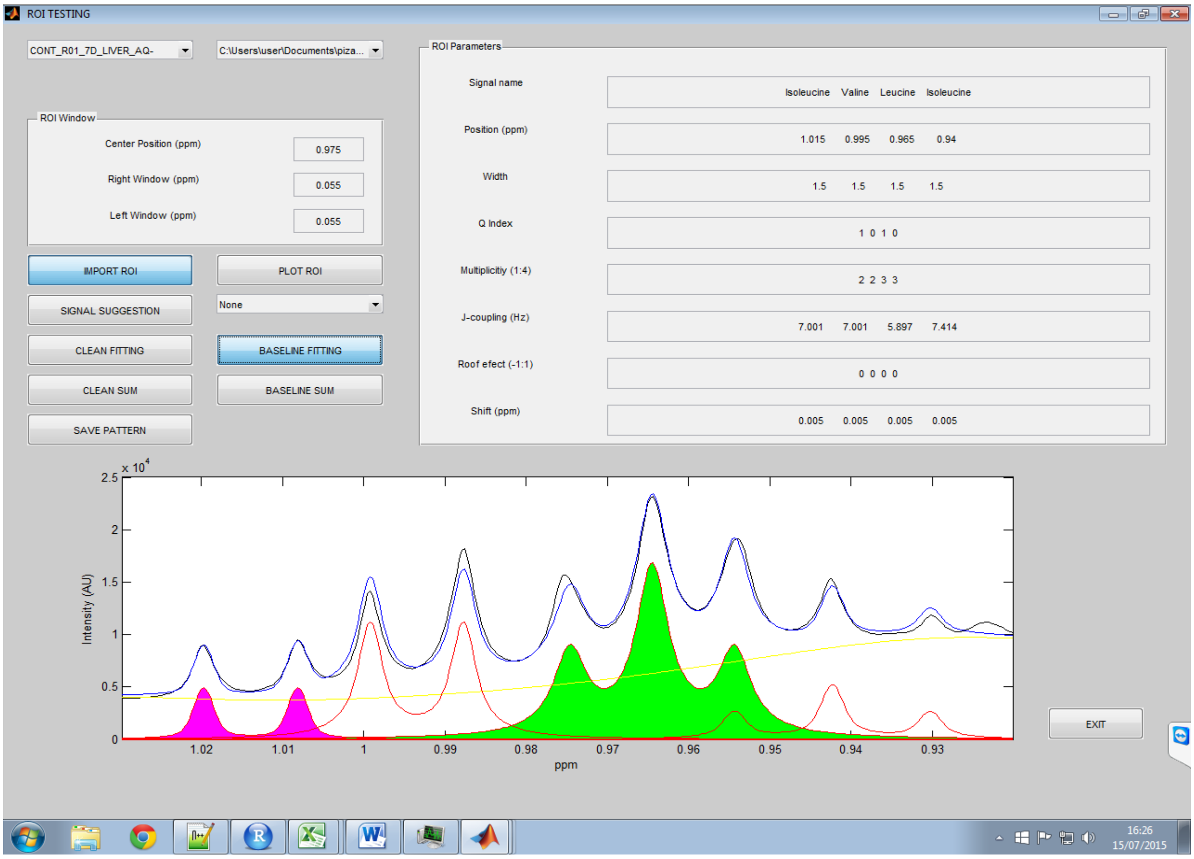
The demo also comes with some examples of ^{1}H (‘1D-HNMR-Library’ folder) and J-RES (‘2D-JRES-Library’ folder) NMR spectra of pure compounds. More ^{1}H spectra can be found in several database websites (BMRB, HMDB, BML-NMR…). More J-RES spectra can be found through the BML-NMR website. Please be careful to prepare the same Bruker file structure when further NMR spectra are added to the library.

5 Testing patterns with the manual panel (ROIs Testing):

Figure A6 shows an example of an ROI pattern test. On the left-upper corner users can select the sample spectrum to test. There is also the option of plotting all spectra, the mean spectrum or the median spectrum of the whole dataset, which can be useful to check alignment behavior in the importing process.

After selecting the sample spectrum, users can use the ‘ROI Window’ panel to generate a region to plot, or can directly choose an ROI pattern to load on the right of where the sample spectrum was chosen. Once a ROI pattern is selected, users can import it through the button ‘Import ROI’ and both Region and Signal Selection panels will be filled automatically. All ROI parameters are the same ones that are contained in the ROI pattern excel files, and are explained in detail in section 4 of this document. With the ‘Plot ROI’ button users can plot the selected spectrum region.

If users are not sure about metabolite assignments, they can test other possibilities through the Signal Suggestion button. These suggestions are based on the presence of signals in the same region for urine, liver, brain extract and liver extract samples. If users are still not convinced, they can check the Signals Database excel file (see Section 4) in order to explore other possibilities.



ROI TESTING

CONT_R01_7D_LIVER_AQ-

C:\Users\user\Documents\piza...

ROI Window

Center Position (ppm): 0.975

Right Window (ppm): 0.055

Left Window (ppm): 0.055

None

ROI Parameters

Signal name: Isoleucine Valine Leucine Isoleucine

Position (ppm): 1.015 0.995 0.965 0.94

Width: 1.5 1.5 1.5 1.5

Q Index: 1 0 1 0

Multiplicitiy (1:4): 2 2 3 3

J-coupling (Hz): 7.001 7.001 5.897 7.414

Roof efect (-1:1): 0 0 0 0

Shift (ppm): 0.005 0.005 0.005 0.005

Figure A6: example of the use of the 'ROIs Testing' panel, in this case for the quantification of branched-chain amino acids through the 'Baseline Fitting' option.

At this point, users can plot the spectra and try which of the four approaches (Baseline Fitting –Figure A4-, Clean Fitting, Baseline Sum or Area Sum) is the best option to quantify the signals of interest. Users has the option to change any of the parameters in 'ROI Window' and 'ROI Parameters' panels to adjust the pattern interactively, and once the performance is optimized, it can be saved through the 'Save Pattern' button and applied to other experiments, or to the whole dataset through the 'Apply4All' button. The changes done interactively through the panel will be lost if users changes or reloads the pattern, so we suggest saving those changes that are useful for most of the spectra of the dataset in the ROI pattern excel file.

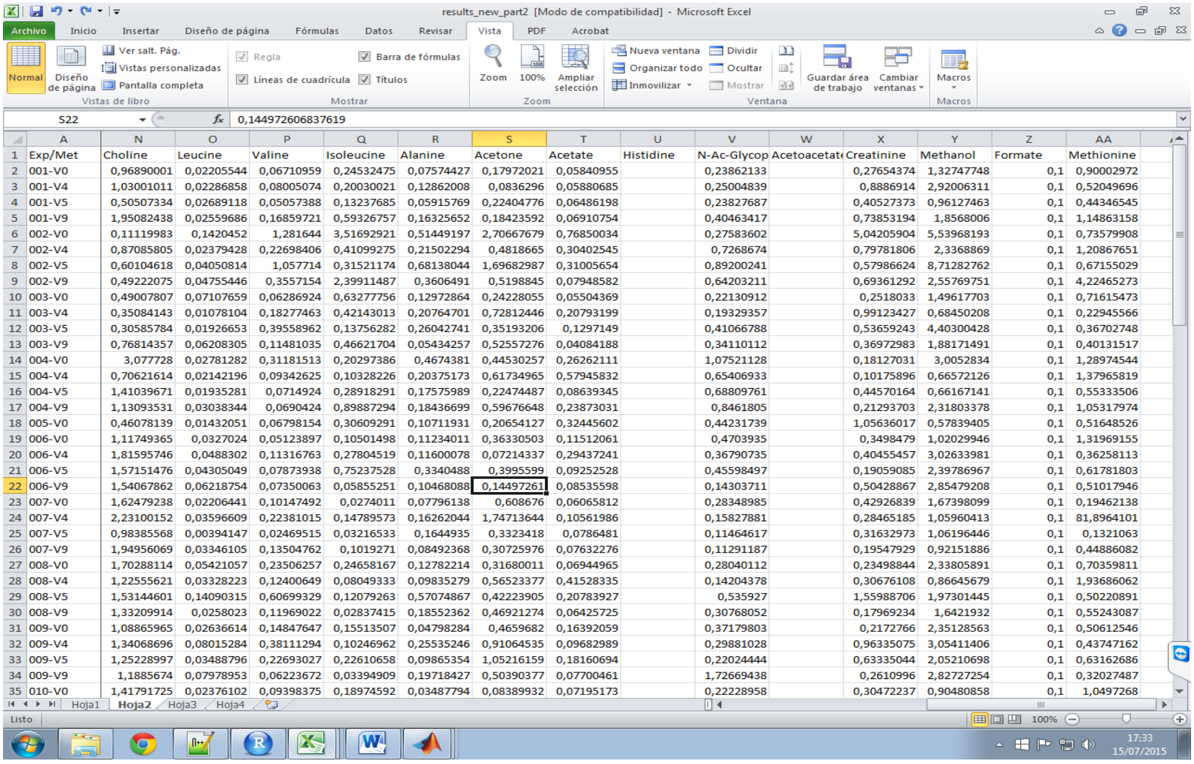
Those signals which don't pass the fitting error or signal area thresholds will generate an image of the plot which will be automatically saved in a folder with the name of the signal within a folder called 'MANRUN' within the folder 'PLOTS2CHECK'. It allows users to check graphically if the results are reliable or not and gives the chance to recalculate wrong quantifications by adjusting some parameters interactively in the ‘ROIs Testing’ Panel.

6 Autorun:

’The package allows users to perform a fully automatic quantification analysis through the button ‘Autorun’. This option is only recommended when users experience and the reliability of the ROI patterns are high (for example as a second analysis of a dataset previously analyzed in a supervised way). Two ‘wait-bars’ will appear giving information about the samples analyzed and the metabolites quantified. This automatic run can be stopped closing any of these two windows. Those signals which don’t pass the fitting error or signal area thresholds will generate an image of the plot which will be automatically saved in a folder with the name of the signal within a folder called ‘AUTORUN’ within the folder ‘PLOTS2CHECK’.

7 The output:

Once the analysis has been finished, users can generate an output file just by pressing the button ‘SAVE RESULTS’, which will open a browser to select the path and the file output name to be saved in xls format. The output file consists in 4 sheets (Figure A7); all of them present a matrix format with the sample titles in the first column and the signal titles in the first row (Figure A7).



S22 f_x 0,144972606837619

	A	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
1	Exp/Met	Choline	Leucine	Valine	Isoleucine	Alanine	Acetone	Acetate	Histidine	N-Ac-Glycop	Acetoacetat	Creatinine	Methanol	Formate	Methionine
2	001-V0	0,96890001	0,02205544	0,06710959	0,24532475	0,07574427	0,17972021	0,05840955		0,23862133		0,27654374	1,32747748	0,1	0,90002972
3	001-V4	1,03001011	0,02286858	0,08005074	0,20030021	0,12862008	0,0836296	0,05880685		0,25004839		0,8886914	2,92006311	0,1	0,52049696
4	001-V5	0,50507334	0,02689118	0,05057388	0,13237685	0,05915769	0,22404776	0,06486198		0,23827687		0,40527373	0,96127463	0,1	0,44346545
5	001-V9	1,95082438	0,02559686	0,16859721	0,59326757	0,16325652	0,18423592	0,06910754		0,40463417		0,73853194	1,8568006	0,1	1,14863158
6	002-V0	0,11119983	0,1420452	1,281644	3,51692921	0,51449197	2,70667679	0,76850034		0,27583602		5,04205904	5,53968193	0,1	0,73579908
7	002-V4	0,87085805	0,02379428	0,22698406	0,41099275	0,21502294	0,4818665	0,30402545		0,7268674		0,79781806	2,3368869	0,1	1,20867651
8	002-V5	0,60104618	0,04050814	1,057714	0,31521174	0,68138044	1,69682987	0,31005654		0,89200241		0,57986624	8,71282762	0,1	0,67155029
9	002-V9	0,49222075	0,04755446	0,3557154	2,39911487	0,3606491	0,5198845	0,07948582		0,64203211		0,69361292	2,55769751	0,1	4,22465273
10	003-V0	0,49007807	0,07107659	0,06286924	0,63277756	0,12972864	0,24228055	0,05504369		0,22130912		0,2518033	1,49617703	0,1	0,71615473
11	003-V4	0,35084143	0,01078104	0,18277463	0,42143013	0,20764701	0,72812446	0,20793199		0,19329357		0,99123427	0,68450208	0,1	0,22945566
12	003-V5	0,30585784	0,01926653	0,39558962	0,13756282	0,26042741	0,35193206	0,1297149		0,41066788		0,53659243	4,40300428	0,1	0,36702748
13	003-V9	0,76814357	0,06208305	0,11481035	0,46621704	0,05434257	0,52557276	0,04084188		0,34110112		0,36972983	1,88171491	0,1	0,40131517
14	004-V0	3,077728	0,02781282	0,31181513	0,20297386	0,4674381	0,44530257	0,26262111		1,07521128		0,18127031	3,0052834	0,1	1,28974544
15	004-V4	0,70621614	0,02142196	0,09342625	0,10328226	0,20375173	0,61734965	0,57945832		0,65406933		0,10175896	0,66572126	0,1	1,37965819
16	004-V5	1,41039671	0,01935281	0,0714924	0,28918291	0,17575989	0,22474487	0,08639345		0,68809761		0,44570164	0,66167141	0,1	0,55333506
17	004-V9	1,13093531	0,03038344	0,0690424	0,89887294	0,18436699	0,59676648	0,23873031		0,8461805		0,21293703	2,31803378	0,1	1,05317974
18	005-V0	0,46078139	0,01432051	0,06798154	0,30609291	0,10711931	0,20654127	0,32445602		0,44231739		1,05636017	0,57839405	0,1	0,51648526
19	006-V0	1,11749365	0,0327024	0,05123897	0,10501498	0,11234011	0,36330503	0,11512061		0,4703935		0,3498479	1,02029946	0,1	1,31969155
20	006-V4	1,81595746	0,0488302	0,11316763	0,27804519	0,11600078	0,07214337	0,29437241		0,36790735		0,40455457	3,02633981	0,1	0,36258113
21	006-V5	1,57151476	0,04305049	0,07873938	0,75237528	0,3340488	0,3995599	0,09252528		0,45598497		0,19059085	2,39786967	0,1	0,61781803
22	006-V9	1,54067862	0,06218754	0,07350063	0,05855251	0,10468088	0,14497261	0,08535598		0,14303711		0,50428867	2,85479208	0,1	0,51017946
23	007-V0	1,62479238	0,02206441	0,10147492	0,0274011	0,07796138	0,608676	0,06065812		0,28348985		0,42926839	1,67398099	0,1	0,19462138
24	007-V4	2,23100152	0,03596609	0,22381015	0,14789573	0,16262044	1,74713644	0,10561986		0,15827881		0,28465185	1,05960413	0,1	81,8964101
25	007-V5	0,98385568	0,00394147	0,02469515	0,03216533	0,1644935	0,3323418	0,0786481		0,11464617		0,31632973	1,06196446	0,1	0,1321063
26	007-V9	1,94956069	0,03346105	0,13504762	0,1019271	0,08492368	0,30725976	0,07632276		0,11291187		0,19547929	0,92151886	0,1	0,44886082
27	008-V0	1,70288114	0,05421057	0,23506257	0,24658167	0,12782214	0,31680011	0,06944965		0,28040112		0,23498844	2,33805891	0,1	0,70359811
28	008-V4	1,22555621	0,03328223	0,12400649	0,08049333	0,09835279	0,56523377	0,41528335		0,14204378		0,30676108	0,86645679	0,1	1,93686062
29	008-V5	1,53144601	0,14090315	0,60699329	0,12079263	0,57074867	0,42223905	0,20783927		0,535927		1,55988706	1,97301445	0,1	0,50220891
30	008-V9	1,33209914	0,0258023	0,11969022	0,02837415	0,18552362	0,46921274	0,06425725		0,30768052		0,17969234	1,6421932	0,1	0,55243087
31	009-V0	1,08865965	0,02636614	0,14847647	0,15513507	0,04798284	0,4659682	0,16392059		0,37179803		0,2172766	2,35128563	0,1	0,50612546
32	009-V4	1,34068696	0,08015284	0,38111294	0,10246962	0,25535246	0,91064535	0,09682989		0,29881028		0,96335075	3,05411406	0,1	0,43747162
33	009-V5	1,25228997	0,03488796	0,22693027	0,22610658	0,09865354	1,05216159	0,18160694		0,22024444		0,63335044	2,05210698	0,1	0,63162686
34	009-V9	1,1885674	0,07978953	0,06223672	0,03394909	0,19718427	0,50390377	0,07700461		1,72669438		0,2610996	2,82727254	0,1	0,32027487
35	010-V0	1,41791725	0,02376102	0,09398375	0,18974592	0,03487794	0,08389932	0,07195173		0,22228958		0,30472237	0,90480858	0,1	1,0497268

Figure A7: Example of Excel file saved, with the first column representing the samples and the other columns representing the data of every metabolite contained in the library Excel file. The sheet represented here is the sheet with the information about the fitting error.

The first sheet contains the quantification values of each signal in each sample in arbitrary units, corresponding to the area under the curve calculated. The second sheet contains the fitting errors of each signal in each sample in percentage and is calculated according to this equation:

$$Fitting\ Error = \sqrt{\frac{(SA - FA)^2}{SA^2}} \cdot 100$$

where SA is the spectrum area in the window where the signal is located and FA is the total fitted area in the window where the signal is located.

The third sheet contains the position of each signal in each sample in ppm units. The last sheet contains the percentage of the fitted area that is represented by the fitting of the signal. This parameter is calculated according to this equation:

$$Signal\ Area\ Ratio = \frac{SFA}{FA} \cdot 100$$

where FA is the total fitted area in the window of the ROI where the signal is located (contains the baseline and all the signals) and SFA is the fitted area of the signal in the window of the ROI where the signal is located (does not contain the fitting of the baseline and/or the other signals)

All this information joined to the saved images gives users information about the final results that can become useful before entering into statistical analysis.