Ph.D. in Electronic and Computer Engineering
Dept. of Electrical and Electronic Engineering
University of Cagliari

Information Management and Multivariate Analysis Techniques for Metabolomics Data

Piergiorgio Palla

Advisor: Prof. Giuliano Armano

Curriculum: ING-INF/05 SISTEMI DI ELABORAZIONE DELLE INFORMAZIONI

XXVII Cycle
April, 2015

Ph.D. in Electronic and Computer Engineering
Dept. of Electrical and Electronic Engineering
University of Cagliari

Information Management and Multivariate Analysis Techniques for Metabolomics Data

Piergiorgio Palla

Advisor: Prof. Giuliano Armano
Curriculum: ING-INF/05 SISTEMI DI ELABORAZIONE DELLE INFORMAZIONI

XXVII Cycle
April, 2015

Acknowledgments

I would like to express my deep gratitude to Dr. Patricia Rodriguez-Tomé for her valuable and constructive suggestions and for giving me the opportunity to experience this doctoral journey while I was working for the Center for Advanced Studies, Research and Development in Sardinia (CRS4).

My grateful thanks are also extended to Professor Giuliano Armano, my advisor, for his useful critiques and for his patience throught this venture.

I take this opportunity to express my sicere thanks to Professor Luigi Atzori and Federica Murgia for their suggestions and help in doing the metabolomics data analysis.

A special mention must go to Dr. Amit Kumar for all the help and his valuable advice. I also thank all the colleagues that, in one way or another, provided insight and expertise that greatly assisted my research activities.

Thanks to my parents and my partner for the unceasing encouragement, support and patience during these years.

I also place on record, my sense of gratitude to one and all, who directly or indirectly, have lent their helping hand in this venture.

Abstract

Among the so-called "omics" disciplines, metabolomics has been receiving considerable attention over the last few years. Metabolomics is the large-scale study of metabolites that are small molecules within cells, biofluids and tissues, produced as a result of metabolism.

The growing interest in metabolomics has been encouraged by rapid advances in metabolic profiling techniques and by technological developments of the diverse analytical platforms, including proton Nucleic Magnetic Resonance (1H NMR), Gas Chromatography-Mass Spectrometry (GC-MS) and Liquid Chromatography-Mass Spectrometry (LC-MS), used for extracting metabolic profiles. The output generated from these experimental techniques results in the production of a huge amount of data and information.

This thesis attempts to provide an overview of the analytical technologies, the resources and databases employed in this emerging discipline, and is mainly focused on the following two aspects: (i) the challenges of handling the large amounts of data generated and managing the complex experimental processes needed to produce them; (ii) the techniques for the multivariate analysis of metabolomics data, with a special emphasis on methods based on the random forest algorithm.

To this aim, a detailed description and explanation of QTREDS, a software platform designed for managing, monitoring and tracking the experimental processes and activites of "omics" laboratories is provided.

In addition, a thorough elucidation of the software package RFmarkerDetector, available through the Comprehensive R Archive Network (CRAN), and a description of the multivariate analysis techniques it implements, is also given.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis illustrates the research activities I have conducted during the years of my PhD. Part of the work has been performed in partnership with the Center for Advanced Studies, Research and Development in Sardinia (CRS4) and the Department of Biomedical Sciences of the University of Cagliari.

The thesis deals with metabolomics, the newest of the so-called "omics" disciplines, and the multivariate data analysis strategies applied to metabolomics data. It consists of six chapters, each focused on a specific topic related to metabolomics.

A brief review of the history of metabolomics and of its main applications is given in the Introduction.

Chapter 2 illustrates the main databases used in metabolomics, dividing them into the following categories:

- Comprehensive Metabolomic Databases;
- Metabolic Pathway Databases;
- Spectral Databases;
- Disease and Physiology Databases;
- Compound-Specific Databases.

QTREDS, our software platform designed for managing and tracking all the activities and processes of an "omics" laboratory, is described in detail in Chapter 3.

Chapter 4 provides an overview of the main spectroscopic techniques employed in metabolomics with the aim of identifying and quantifying metabolites in different biological samples.

Chapter 5 describes the multivariate analysis strategies implemented in our software package RFmarkerDetector and illustrates a concrete case study.

Lastly, Chapter 6 briefly summarizes the findings of the research activities.

1.1 Metabolomics or Metabonomics

In the post-genomic era, molecular biology has increasingly focused its attention on disciplines like Transcriptomics and Proteomics. The former refers to the determination of multiple protein expression changes in a cell or tissue, while the latter to the determination of multiple gene-expression changes at the RNA level. Similar developments have been taking place at metabolite small-molecule level, leading to the increasing expansion in studies now termed Metabolomics. The main purpose of these techniques is to gather new insights and a better understanding of the biological functioning of a cell or organism [27, 53].

The interpretation of transcriptome and proteome data is not easy, due to the problem of relating observed gene-expression fold changes or protein-level changes to conventional disease and pharmaceutically relevant end-points. Indeed, changes in the transcriptome and proteome do not always result in altered biochemical phenotypes.
Metabolomics, unlike other "omics" disciplines, is a powerful approach because it can provide the most "functional" information. The metabolome is the complete set of metabolites within a cell or biological sample at any given time point. It can be seen as the last stage in the flow of events from genes to metabolism, and the metabolic profile is the most direct indication of the actual biological state of an organism. Thus metabolomics best represents the molecular phenotype.

Metabolites have a well-defined function in the life of biological systems reflecting the surrounding environment. Thus, quantitative global analysis of endogenous metabolites from cells, tissues, fluids, etc. is becoming an integral part of functional genomics [30] effort as well as a tool for discovering diagnostic biomarkers [53, 75].

The terms *metabonomics* and *metabolomics* appeared for the first time at the end of the

90's and are often used interchangeably although their exact definitions are slightly different [83].

The term *metabolomics* was coined by O. Fiehn [29] and defined as a comprehensive analysis in which all metabolites of a biological system were identified and quantified [53, 28]. The similar term *metabonomics* [63] was introduced earlier and is defined as the quantitative measurement of the dynamic multi- parametric metabolic response of living systems to pathophysiological stimuli or genetic modification [83]. Since, the two expressions are often employed indifferently in practice, in the rest of the thesis it will be used the term *metabolomics*.

Metabolomics is a rapidly maturing field: it is increasingly being applied to many areas of biomedical research, such as toxicology studies, nutritional effects, inborn errors of metabolism, diabetes, cancer diagnostics, and diagnosing of neurological diseases.

The analysis of metabolites is not a completely new field; early studies on the metabolic profiles date back to 1950, but they have been limited to relatively small numbers of target analytes as in the study of a particular metabolic pathway.

The concept that individuals might have a "metabolic profile" that would be reflected in the constituents of their biological fluids was first developed and tested by Roger Williams during the late 1940s and early 1950s. Utilizing data from over 200,000 paper chromatograms, Williams was able to show that characteristic metabolic patterns in urine and saliva were associated with diseases such as schizophrenia [35, 53]. The work of Williams and his group and his ideas about the utility of metabolic pattern analysis, remained essentially unexpressed until the late 1960s, when gas chromatography-mass spectrometry (GC-MS) [81, 76] and liquid chromatography-mass spectrometry(LC-MS) [47] were sufficiently advanced to allow such studies to be carried out with considerably less effort. In fact, it was only through technological advancements in the 1960s and 1970s that it became feasible to quantitatively measure metabolic profiles. The expression "metabolic profile" to refer to qualitative and quantitative analyses of complex mixtures of physiological origin, was coined by Horning that with his group led the development of gas chromatography-mass spectrometry (GC-MS) methods to monitor the metabolites present in urine through the 1970s [40].

Almost from the birth of GC, people involved in organic MS saw the potential advantage of separating complex mixtures into its components followed by structural analysis by MS. It soon became evident that GC-MS was different from both GC and MS. Specifically three issues had to be addressed:

1. the large amount of gas leaving the column (working with packed columns), while MS separates the ions in high vacuum condition;
2. the need for rapid mass spectral acquisition;
3. the enormous amount of data collected during a GC-MS analysis.

A device named "jet separator" solved the first problem, eliminating most of the carrier gas in a selective manner [74]. For the second issue, it can be said that when GC- MS was at its beginning, magnetic sector mass spectrometer did not have a rapid data acquisition capability. After the commercialisation of the first magnetic sector instrument built as a GC-MS in the mid '60s, the problem of rapid acquisition was rapidly solved. On the other hand this instrument did not deal with the third issue related to the amount of data acquired in a GC-MS analysis. A light beam oscilloscope was used to record the mass spectra that were manually selected for recording. Minicomputers were also developed in the mid '60s, and in few years the automated collection of GC-MS data became possible.

GC-MS became a routinely used technology with the introduction of quadrupolar instruments. Quadrupole technology, which includes the transmission quadrupole (TQ) and the quadrupole ion trap (QIT), was studied by Paul.

The first GC-QTMS was developed in the late '60s and rapidly replaced the magnetic sector based instruments [32], because of its simplicity and the continuous advancement of the data station.

One of the major drawbacks of GC identification is the need for thermostable, volatile analytes; derivatization of the polar functional group can improve volatility, but a derivatiation step introduces bias and it is not always possible. This limitation is overcome by LC, which is virtually suitable for the separation of all kind of molecules. During the same period, Nuclear Magnetic Resonance (NMR) spectroscopy was rapidly evolving. In 1974, a study conducted by Seeley et al. demonstrated the utility of using NMR to detect metabolites in biological samples [45].

Nuclear magnetic resonance spectroscopy and mass spectrometry are two fundamental analytical techniques for the identification and quantification of a large set of metabolites present in a given biological system. Each technology shows advantages and disadvantages, as we will see in the following chapters, but they are essentially complementary [68].

1.2 Application of Metabolomics

Metabolomics can bring enomous new insights on metabolic fluxes and a more comprehensive understanding of a cell's environment [53]. Applications of metabolomics can be seen in many clinical or pharmaceutical areas such as drug discovery, clinical toxicology and human diseases.

Over the past few years metabolomics has also emerged as a field of increasing interest to food scientists.

1.2.1 Applications within Food Industry

Foods are now being analysed with more chemical detail leading to hundreds or even thousands of distinct chemical identities being detected or identified.

Metabolomic applications within the food industry are diverse ranging from profiling of plant species to studying the effects of stresses on plants [65].

Food component analysis traditionally involved the identification and the classification of food components into broad categories such as carbohydrates, proteins, fats, vitamins. The development of metabolomics allows the identification of hundreds of distinct molecules being detected and or identified in certain foods with considerably more chemical detail [65].

Future trends will involve the use of discriminative and predictive metabolomics as the ultimate tool for quality control. The metabolite profile of products meeting minimum standards can be used as a baseline for quality acceptance.

1.2.2 Toxicity Assessment

Metabolic profiling (especially of urine or blood plasma samples) can be used to detect the physiological changes caused by toxic insult of a chemical (or mixture of chemicals) [94].

This can be of particular relevance to pharmaceutical companies wanting to test the toxicity of potential drug candidates: if a compound can be eliminated before it reaches clinical trials on the grounds of adverse toxicity, it saves the enormous expense of the trials [94].

1.2.3 Applications in Oncology

The main current applications and challenges of metabolomics in cancer research are:

- biomarkers for diagnosis, staging and monitoring of the disease and therapeutic response;
- protein expression profiling of tumours;
- protein microarrays;
- pharmacoproteomics

All these applications continue to benefit from further technological advances such as high-resolution, high-speed and high-sensitivity Mass Spectrometry and advanced bioinformatics for data handling and interpretation [54].

1.2.4 Applications in Genetics

Metabolomics is an important "omic" science to fill the gap between genomics and proteomics. It involves the determination of multiple metabolites simultaneously in biofluids, tissues and tissue extracts and these all have some levels of genetic involvement [53].

Metabolomics can be an excellent tool for determining the phenotype caused by a genetic manipulation, such as gene deletion or insertion. More interesting is the prospect of predicting the function of unknown genes by comparison with the metabolic perturbations caused by deletion/insertion of known genes [94].

Chapter 2

Databases and tools in Metabolomics

2.1 Introduction

Metabolomics technologies yield many insights into basic biological research in areas such as systems biology and metabolic modeling , pharmaceutical research, nutrition and toxicology [42]. When combined with genomic, transcriptomic and proteomic studies, metabolomics can also help in the interpretation and understanding of many complex biological processes. Indeed, metabolomics is now widely recognized as being a cornerstone to all of systems biology [103].

However, to exploit the full potential of metabolomics, researchers need access to data and knowledge to compare, contrast and make inferences from the results they obtain in their experiments.

The metabolome refers to the total complement of small-molecule chemicals (metabolites) present within a biological sample under given genetic, nutritional or environmental conditions. Since such conditions can vary dramatically, metabolomics has to combine different disciplines such as molecular biology, chemistry and physiology to accurately reflect the underlying diversity and complexity. Therefore there is a need for not just one type of database, but a wide variety of electronic resources [103].

Currently, there are at least five types of databases used in metabolomics research. These include:

1. comprehensive, organism- specific metabolomic databases;
2. metabolic pathway databases;
3. spectral databases;
4. disease/physiology databases
5. compound-specific databases; [103].

Far from being complete a description of some important databases and tools in metabolomics is presented in the following pages.

2.2 Comprehensive Metabolomic Databases

2.2.1 HMDB

First introduced in 2007, the Human Metabolome Database (HMDB) is currently the world's largest and most comprehensive, organism-specific metabolomics database. It contains spectroscopic, quantitative, analytic and physiological information about human metabolites, their associated enzymes or trans-porters, their abundance and their disease-related properties [102].

The HMDB combines the data-rich molecular biology content that can be found in curated sequence databases such as SwissProt and UniProt [16] with the equally rich data found in KEGG [48] (about metabolism) and OMMBID [79] (about clinical conditions). Furthermore it collects a large amount of experimental data, including NMR spectra, MS spectra, solubility data and validated metabolite concentrations. [104].

The latest release of the HMDB provides detailed information on over 40 000 metabolites, representing an expansion of nearly 600% over what was previously contained in the database [102] .
A detailed content comparison between the HMDB (release 1.0 and 2.0) versus the HMDB (release 3.0) is provided in Table 2.1.

This growth is mainly a result of the important expansion of both 'detected' metabolites (divided into two categories: (i) detected and quantified and (ii) detected not quantified) and

’expected’ metabolites (those for which biochemical pathways are known but the compound has yet to be detected in the body).

Among the ’detected’ metabolites, the number has grown from 4413 (in version 2.0) to 20900 (in version 3.0), or roughly by 450%. While among the ’expected’ metabolites, their numbers have grown much more significantly, from 1995 (in version 2.0) to more than 19000 (in version 3.0). This amount includes more than 450 dipeptides, over 1500 drugs and drug metabolites, over 13000 foodderived compounds and more than 2000 other compounds [102].

A key feature that differentiates the HMDB database from other metabolic resources is its extensive support for higher level database searching and selecting functions. In fact, in addition to the data viewing and sorting features provided, the HMDB also offers a chemical structure search utility, a local BLAST search [15] that supports both single and multiple sequence queries, a boolean text search based on GLIMPSE [59], a relational data extraction tool, an MS spectral matching tool and an NMR spectral search tool (for identifying compounds via MS or NMR data from other metabolomic studies) [104].

The structure similarity search tool (ChemQuery) allows users to draw chemical structures or paste a SMILES string [101] of a compound into the ChemQuery window. Submitting

Database feature or content status	HMDB (version 1.0)	HMDB (version 2.0)	HMDB (version 1.0)
Number of metabolites	2180	6408	40153
Number of unique metabolite synonyms	27700	43882	199668
Number of compounds with disease links	862	1002	3105
Number of compounds with biofluid or tissue concentration data	883	4413	5027
Number of compounds with chemical synthesis references	220	1647	1943
Number of compounds with experimental reference 1H and or 13C	385	792	1054
NMR spectra Number of compounds with reference MS/MS spectra	390	799	1249
Number of compounds with reference GC-MS reference data	0	279	1220
Number of human-specific pathway maps	26	58	442
Number of compounds in Human Metabolome Library	607	920	1031
Number of HMDB data fields	91	102	114
Number of predicted molecular properties	2	2	10
Metabolite search/browse	yes	yes	yes
Pathway search/browse	no	yes	yes
Disease search/browse	no	yes	yes
Chemical class search/browse	no	yes	yes
Biofluid browse	no	yes	yes
Metabolite library browse	no	yes	yes
Protein/transporter browse	no	no	yes

Table 2.1: Comparison between the HMDB releases

the query launches a structure similarity search tool that looks for common substructures from the query compound that match the HMDB's metabolite database. The ChemQuery tool allows to quickly discover whether their compound of interest is a known metabolite or chemically related to a known metabolite. In addition to these structure similarity searches, the ChemQuery utility also supports compound searches on the basis of chemical formula and molecular weight ranges.

The BLAST search (SeqSearch) allows users to search through the HMDB on the basis of the sequence similarity. A given gene or protein sequence may be searched against the HMDB's sequence database of metabolically important enzymes and transporters by pasting the FASTA formatted sequence (or sequences) into the SeqSearch query box.

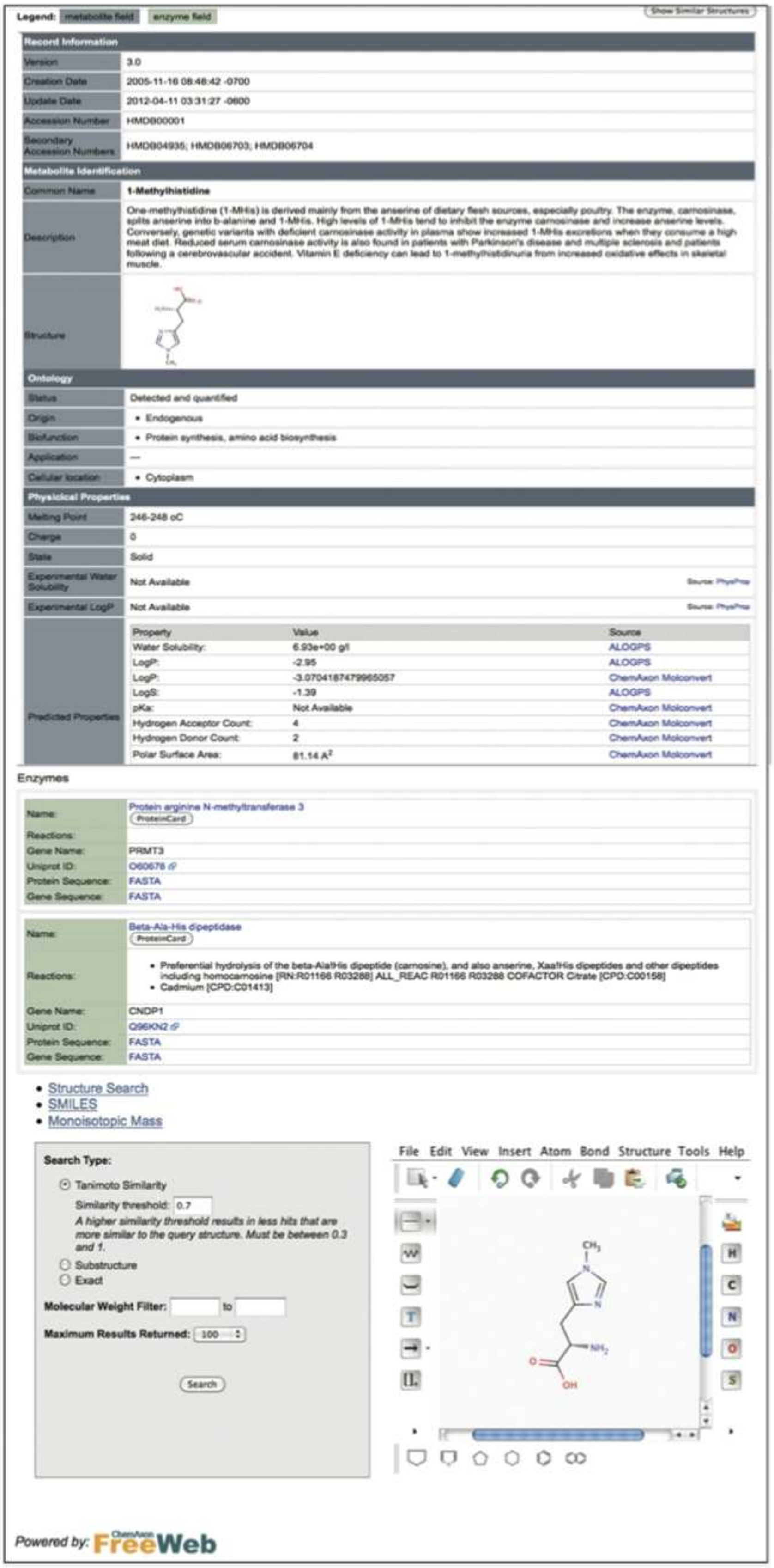
The HMDB's spectral search utilities allow both pure compounds and mixtures of compounds to be identified from their MS or NMR spectra via peak matching algorithms. The NMR spectral matching algorithm uses a simple peak matching rule with pre-defined chemical shift tolerances. Query spectra are scored on the number of peak matches to the database spectra. The MS/MS spectral matching algorithm uses a peak matching and spectral scoring.

Perhaps the most relevant features of the HMDB from the perspective of a medical geneticist or a clinical chemist are its rich content and extensive linkage to metabolic diseases, to normal and abnormal metabolite concentration ranges (in many different biofluids), to mutation/SNP data and to the genes, enzymes, reactions and pathways associated with many diseases of interest [104].

2.2.2 MetaboLights

MetaboLights is the first general-purpose, open-access repository for metabolomics studies, their raw experimental data and associated metadata, maintained by one of the major open-access data providers in molecular biology [42]. The development of this tool was driven by the needs to:

- provide a single point of access to worldwide data and knowledge in metabolomics;
- facilitate the development and adoption of a common data sharing format;
- ensure data traceability and reproducibility;



Record Information	
Version	3.0
Creation Date	2005-11-16 08:48:42 -0700
Update Date	2012-04-11 03:31:27 -0600
Accession Number	HMDB00001
Secondary Accession Numbers	HMDB04935; HMDB06703; HMDB06704
Metabolite Identification	
Common Name	**1-Methylhistidine**
Description	One-methylhistidine (1-MHis) is derived mainly from the anserine of dietary flesh sources, especially poultry. The enzyme, carnosinase, splits anserine into b-alanine and 1-MHis. High levels of 1-MHis tend to inhibit the enzyme carnosinase and increase anserine levels. Conversely, genetic variants with deficient carnosinase activity in plasma show increased 1-MHis excretions when they consume a high meat diet. Reduced serum carnosinase activity is also found in patients with Parkinson's disease and multiple sclerosis and patients following a cerebrovascular accident. Vitamin E deficiency can lead to 1-methylhistidinuria from increased oxidative effects in skeletal muscle.
Structure	
Ontology	
Status	Detected and quantified
Origin	• Endogenous
Biofunction	• Protein synthesis, amino acid biosynthesis
Application	—
Cellular location	• Cytoplasm
Physicical Properties	
Melting Point	246-248 oC
Charge	0
State	Solid
Experimental Water Solubility	Not Available Source: PhysProp
Experimental LogP	Not Available Source: PhysProp

Predicted Properties

Property	Value	Source
Water Solubility:	6.93e+00 g/l	ALOGPS
LogP:	-2.95	ALOGPS
LogP:	-3.0704187479965057	ChemAxon Molconvert
LogS:	-1.39	ALOGPS
pKa:	Not Available	ChemAxon Molconvert
Hydrogen Acceptor Count:	4	ChemAxon Molconvert
Hydrogen Donor Count:	2	ChemAxon Molconvert
Polar Surface Area:	81.14 A^2	ChemAxon Molconvert

Enzymes

Name:	Protein arginine N-methyltransferase 3 ProteinCard
Reactions:	
Gene Name:	PRMT3
Uniprot ID:	O60678
Protein Sequence:	FASTA
Gene Sequence:	FASTA

Name:	Beta-Ala-His dipeptidase ProteinCard
Reactions:	• Preferential hydrolysis of the beta-Ala!His dipeptide (carnosine), and also anserine, Xaa!His dipeptides and other dipeptides including homocarnosine [RN:R01166 R03288] ALL_REAC R01166 R03288 COFACTOR Citrate [CPD:C00158] • Cadmium [CPD:C01413]
Gene Name:	CNDP1
Uniprot ID:	Q96KN2
Protein Sequence:	FASTA
Gene Sequence:	FASTA

- Structure Search
- SMILES
- Monoisotopic Mass

Search Type:

● Tanimoto Similarity

Similarity threshold: 0.7

A higher similarity threshold results in less hits that are more similar to the query structure. Must be between 0.3 and 1.

○ Substructure

○ Exact

Molecular Weight Filter: to

Maximum Results Returned: 100

Search

File Edit View Insert Atom Bond Structure Tools Help
T
H
C
N
O
S
CH_3
N
N
NH_2
O
OH
Powered by: ChemAxon FreeWeb

Figure 2.1: **HMDB user interface**. A screenshot showing several of HDMB's search and data dispaly tools (from [102]).

- progressively promote interoperability across existing resources.

MetaboLights consists of two distinct layers: a repository, enabling the metabolomics community to share findings, data and protocols for any form of metabolomics study, and a reference layer of curated knowledge about metabolites. It is not intended to replace specialist resources but is specifically designed to build on prior art and extensively collaborate with the existing databases to ensure that data are exchanged and that assimilation efforts target gaps in worldwide available knowledge.

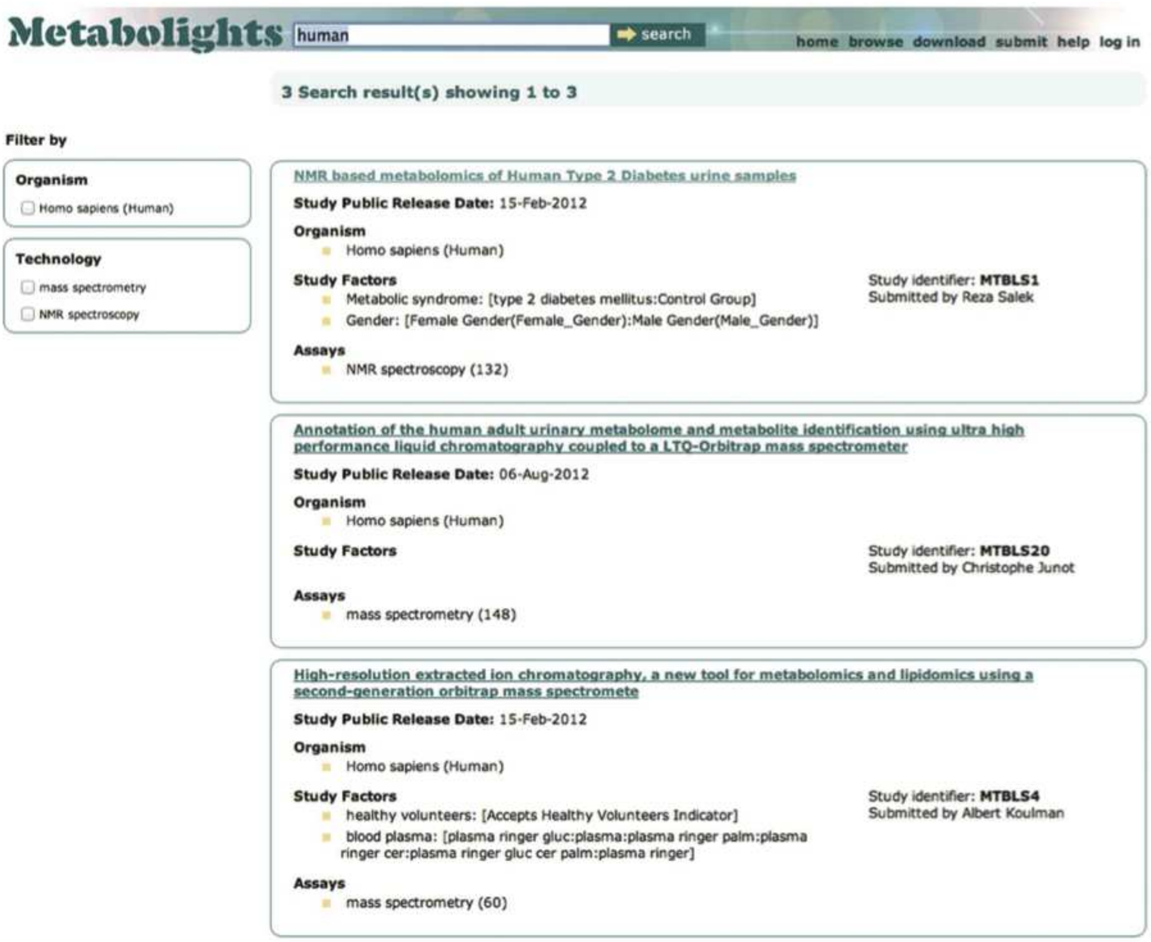
The system stores and display an extensive set of associated information which includes submitter and author information, publication references, the study design, protocols applied, names of data files included, platform information and metabolite information. For each metabolite it includes a description, external database identifiers, formula and intensity or concentration, and where the metabolite was identified in the sample.

Essentially the MetaboLights is a web application running on an Apache Tomcat server. Data are stored on a backend Oracle database, whose implementation is based on the ISA framework [72].

The online search tool allows users to submit a query using free text through most of the underlying data fields, including the study description, study title, protocols, metabolites and authors. The search result page, as illustrated in Figure 2.2, shows general study information like the submitter of the study, the study title, organisms, study design and platform [42].

The MetaboLights provides users with the ability to browse the complete list of all the public studies available which are also downloadable as ISA-Tab [72] metadata files with associated data files directly from the online study details page and from the MetaboLights download page. For those users who are registered, there is also the possibility to get information on additional private studies.

MetaboLights allows to submit experimental studies and data in ISA- Tab format, which can be created by the ISAcreator editor tool (Figure 2.3). ISAcreator is a standalone Java desktop application that enables researchers to report experimental information, associate raw and processed data files, and submit the collated in- formation to the MetaboLights database (Figure 2.4).



3 Search result(s) showing 1 to 3

Filter by

Organism
□ Homo sapiens (Human)

Technology
□ mass spectrometry
□ NMR spectroscopy

NMR based metabolomics of Human Type 2 Diabetes urine samples

Study Public Release Date: 15-Feb-2012

Organism
- Homo sapiens (Human)

Study Factors
- Metabolic syndrome: [type 2 diabetes mellitus:Control Group]
- Gender: [Female Gender(Female_Gender):Male Gender(Male_Gender)]

Assays
- NMR spectroscopy (132)

Study identifier: **MTBLS1**
Submitted by Reza Salek

Annotation of the human adult urinary metabolome and metabolite identification using ultra high performance liquid chromatography coupled to a LTQ-Orbitrap mass spectrometer

Study Public Release Date: 06-Aug-2012

Organism
- Homo sapiens (Human)

Study Factors

Assays
- mass spectrometry (148)

Study identifier: **MTBLS20**
Submitted by Christophe Junot

High-resolution extracted ion chromatography, a new tool for metabolomics and lipidomics using a second-generation orbitrap mass spectromete

Study Public Release Date: 15-Feb-2012

Organism
- Homo sapiens (Human)

Study Factors
- healthy volunteers: [Accepts Healthy Volunteers Indicator]
- blood plasma: [plasma ringer gluc:plasma:plasma ringer palm:plasma ringer cer:plasma ringer gluc cer palm:plasma ringer]

Assays
- mass spectrometry (60)

Study identifier: **MTBLS4**
Submitted by Albert Koulman

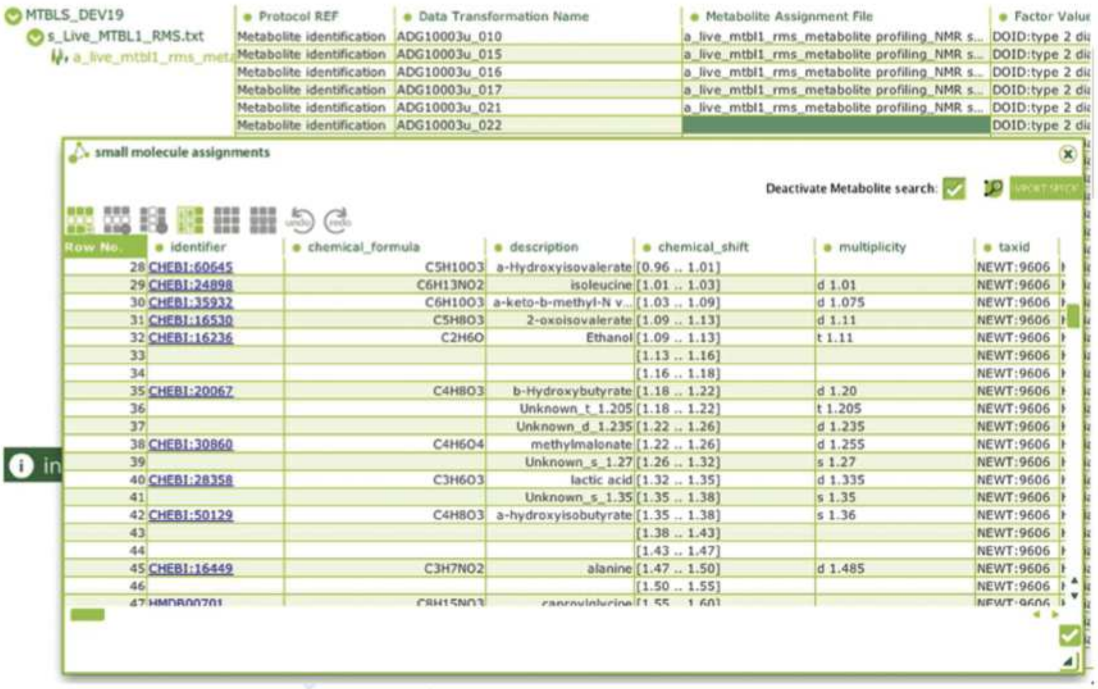
Figure 2.2: **MetaboLights user interface (from [42]).**

2.2.3 BiGG

The Biochemically, Genetically and Genomically database (BiGG) is a metabolic reconstruction of human metabolism designed for systems biology simulation and metabolic flux balance modeling. Figure 2.5 illustrates its database schema.

BiGG integrates several published genome-scale metabolic networks into one resource with standard nomenclature which allows components to be compared across different organisms. It can be used to browse model content, visualize metabolic pathway maps, and export Systems Biology Markup Language (SBML) files of the models for further analysis by external software packages [78].

BiGG accounts for the functions of 1496 Open Reading Frames (ORFs), 20004 proteins, 2766 metabolites, and 3311 metabolic and transport reactions. It was assembled from build



Protocol REF	Data Transformation Name	Metabolite Assignment File	Factor Value
Metabolite identification	ADG10003u_010	a_live_mtbl1_rms_metabolite profiling_NMR s...	DOID:type 2 di
Metabolite identification	ADG10003u_015	a_live_mtbl1_rms_metabolite profiling_NMR s...	DOID:type 2 di
Metabolite identification	ADG10003u_016	a_live_mtbl1_rms_metabolite profiling_NMR s...	DOID:type 2 di
Metabolite identification	ADG10003u_017	a_live_mtbl1_rms_metabolite profiling_NMR s...	DOID:type 2 di
Metabolite identification	ADG10003u_021	a_live_mtbl1_rms_metabolite profiling_NMR s...	DOID:type 2 di
Metabolite identification	ADG10003u_022		DOID:type 2 di

Deactivate Metabolite search: ☑

Row No.	identifier	chemical_formula	description	chemical_shift	multiplicity	taxid
28	CHEBI:60645	C5H10O3	a-Hydroxyisovalerate	[0.96 .. 1.01]		NEWT:9606
29	CHEBI:24898	C6H13NO2	isoleucine	[1.01 .. 1.03]	d 1.01	NEWT:9606
30	CHEBI:35932	C6H10O3	a-keto-b-methyl-N v...	[1.03 .. 1.09]	d 1.075	NEWT:9606
31	CHEBI:16530	C5H8O3	2-oxoisovalerate	[1.09 .. 1.13]	d 1.11	NEWT:9606
32	CHEBI:16236	C2H6O	Ethanol	[1.09 .. 1.13]	t 1.11	NEWT:9606
33				[1.13 .. 1.16]		NEWT:9606
34				[1.16 .. 1.18]		NEWT:9606
35	CHEBI:20067	C4H8O3	b-Hydroxybutyrate	[1.18 .. 1.22]	d 1.20	NEWT:9606
36			Unknown_t_1.205	[1.18 .. 1.22]	t 1.205	NEWT:9606
37			Unknown_d_1.235	[1.22 .. 1.26]	d 1.235	NEWT:9606
38	CHEBI:30860	C4H6O4	methylmalonate	[1.22 .. 1.26]	d 1.255	NEWT:9606
39			Unknown_s_1.27	[1.26 .. 1.32]	s 1.27	NEWT:9606
40	CHEBI:28358	C3H6O3	lactic acid	[1.32 .. 1.35]	d 1.335	NEWT:9606
41			Unknown_s_1.35	[1.35 .. 1.38]	s 1.35	NEWT:9606
42	CHEBI:50129	C4H8O3	a-hydroxyisobutyrate	[1.35 .. 1.38]	s 1.36	NEWT:9606
43				[1.38 .. 1.43]		NEWT:9606
44				[1.43 .. 1.47]		NEWT:9606
45	CHEBI:16449	C3H7NO2	alanine	[1.47 .. 1.50]	d 1.485	NEWT:9606
46				[1.50 .. 1.55]		NEWT:9606
47	HMDB00701	C8H15NO3	caproylglycine	[1.55 .. 1.60]		NEWT:9606

Figure 2.3: **MetaboLights ISAcreator with the Metabolite Identification Plugin (from [42]).**

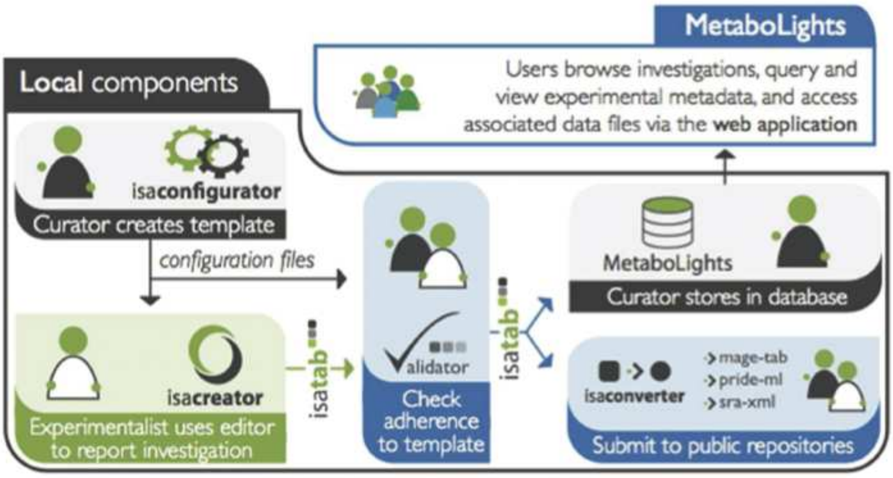

MetaboLights
Users browse investigations, query and view experimental metadata, and access associated data files via the web application
Local components
isaconfigurator
Curator creates template
configuration files
isacreator
Experimentalist uses editor to report investigation
isatab
validator
Check adherence to template
isatab
MetaboLights
Curator stores in database
isaconverter
mage-tab
pride-ml
sra-xml
Submit to public repositories

Figure 2.4: **The ISA framework for reporting information and submitting it to the MetaboLights database (from [42]).**

35 of the human genome [4].

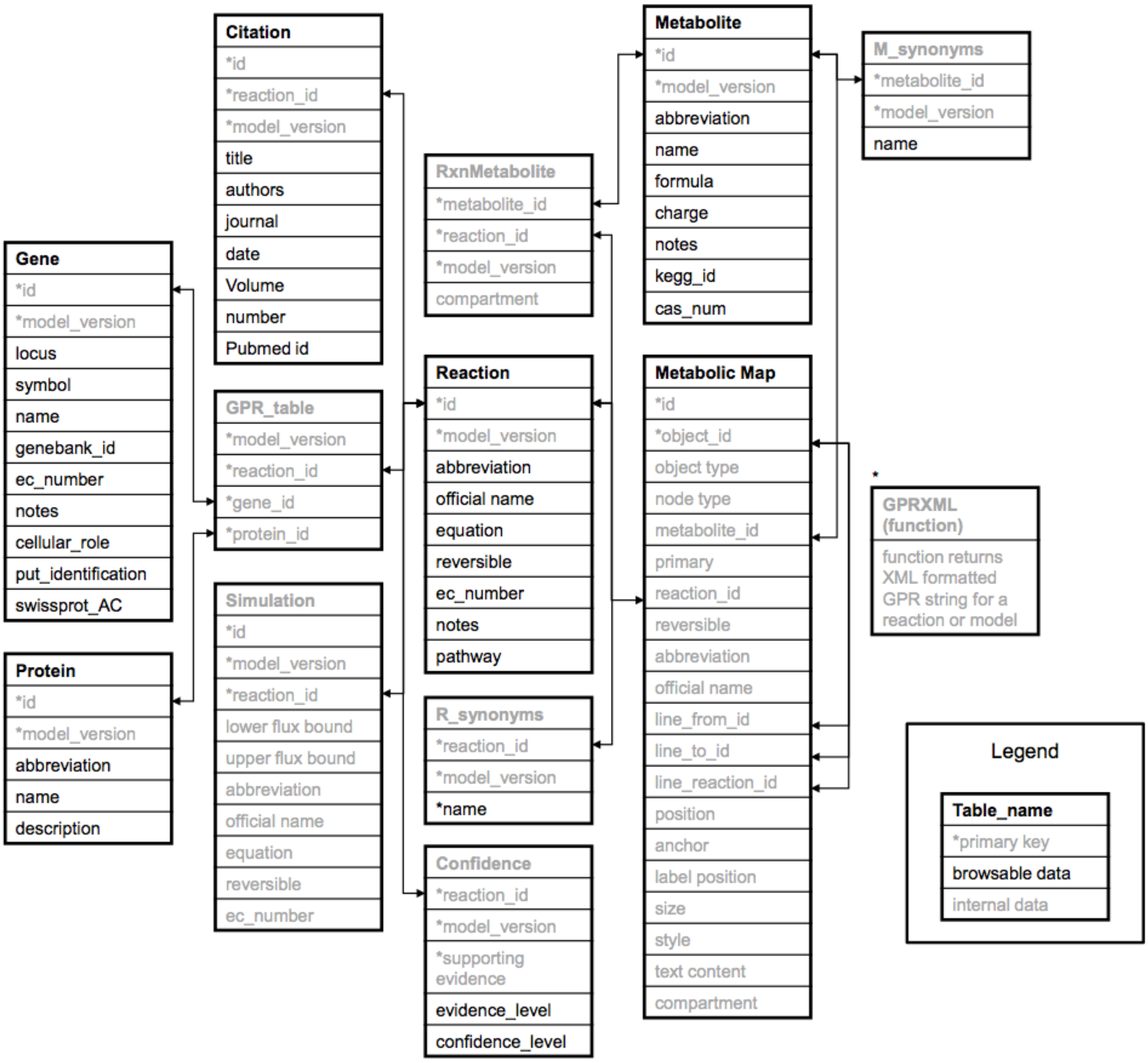


Figure 2.5: **BiGG database schema** (from [78]).

2.3 Metabolic Pathway Databases

2.3.1 KEGG

KEGG (Kyoto Encyclopedia of Genes and Genomes) is a bioinformatics resource for understanding higher-order functional meanings and utilities of the cell or the organism from its genome information [48].

KEGG is an integrated database resource consisting of 15 main databases (see Table 2.2), containing metabolic pathways (372 reference pathways) from a wide variety of organisms (over 700) [49, 4]. Its overall architecture consists of various data objects, called KEGG objects, which according to the developers, are a computer representation of biological systems.

Category	Database	Content
Systems information	KEGG PATHWAY	Pathway maps
	KEGG BRITE	Functional hierarchies
	KEGG MODULE	KEGG modules
	KEGG DISEASE	Human diseases
	KEGG DRUG	Drugs
	KEGG ENVIRON	Crude drugs, etc
Genomic information	KEGG ORTHOLOGY	KO groups
	KEGG GENOME	KEGG organisms
	KEGG GENES	Genes in high-quality genomes
Chemical information	KEGG COMPOUNDS	Metabolites and other small molecules
	KEGG GLYCAN	Glycans
	KEGG REACTION	Biochemical reactions
	KEGG RPAIR	Reactant pairs
	KEGG RCLASS	Reaction class
	KEGG ENZYME	Enzyme nomenclature

Table 2.2: KEGG databases.

KEGG has been developed as a reference knowledge base to assist in the process of gathering knowledge from information. In particular, the KEGG pathway maps are widely used for biological interpretation of genome sequences and other high-throughput data [50].

2.3.2 Reactome

Reactome is a manually curated open-source open-data resource of human pathways and reactions including metabolic pathways and signaling pathways. It includes several types of reactions in its pathway diagram collection including experimentally confirmed, manually inferred and electronically inferred reactions [25, 4]. Reactome describes over 7000 human proteins, participating in almost 7000 reactions based on data extracted from more than 15000 research publications with PubMed links [25].

At the cellular level, life is a network of molecular interactions that can be organized into higher order interconnected pathways. Molecules are synthesized, degraded, undergo a bewildering array of temporary and permanent modifications, are transported from one location to another, and form complexes with other molecules [25, 10]. By annotating all of these processes in a single, consistent reaction-pathway format, the Reactome Knowledgebase systematically links human proteins to their molecular functions, providing a resource that functions both as an archive of biological processes and as a tool for discovering unexpected functional relationships in data.

The goal of the Reactome knowledgebase is to represent human biological processes, many of which have not been directly studied in humans. Rather, a human event has been inferred from experiments on material from a model organism. In such cases, the model organism reaction is annotated in Reactome, the inferred human reaction is annotated as a separate event, and the inferential link between the two reactions is explicitly noted [10].

The Reactome data model consists of classes (frames) that describe the different concepts (e.g., reaction, simple entity). Knowledge is captured as instances of these classes. Classes have attributes (slots) which hold properties of the instances [10].

2.3.3 MetaCyc

MetaCyc is a highly curated nonredundant reference database of small-molecule metabolism, describing metabolic pathways and enzymes from all domains of life [23]. It contains chemical compounds, genes, enzymatic reactions, enzymes and metabolic pathways. The information about enzymes includes many elements like substrate specificity, kinetic properties, activators, inhibitors, cofactor requirements and links to sequence and structure databases [53] . MetaCyc contains more than 2000 pathways derived from over 37000 publications.

Besides its role as a general reference on metabolic processes, MetaCyc can be employed in conjunction with the PathoLogic component of the Pathway Tools software for the prediction the metabolic network of any organism that has a sequenced and annotated genome.

2.3.4 BioCyc

BioCyc is a collection of over 3000 organism-specific Pathway/Genome Databases (PGDB), each containing the full genome and predicted metabolic network of one organism, including metabolites, enzymes, reactions, metabolic pathways and pathway-hole fillers [23].

The organization of the databases within the BioCyc collection is divided into tiers [1] according to the amount of manual review and updating they have received:

- **Tier 1** have been created through intensive manual efforts, and receive continuous updating. It includes the following databases
  - EcoCyc,
  - MetaCyc,
  - AraCyc,
  - HumanCyc,
  - LeishCyc,
  - YeastCyc
- **Tier 2** includes 39 PGDBs computationally generated by the PathoLogic program;
- **Tier 3** contains nearly 3000 PGDBs that includes computationally predicted metabolic pathways, as well as predictions as to which genes code for missing enzymes in metabolic pathways, and predicted operons.

The BioCyc Web site offers a variety of tools for querying and analysis of PGDBs, including Omics Viewers and tools for comparative analysis [23].

2.4 Spectral Databases

2.4.1 Golm Metabolome Database

The Golm Metabolome Database (GMD) started as a collection of annotated and non-annotated mass spectra from biological samples and was extended to contain, in addition, retention index information (RI) [46], becoming soon a reference library dedicated to metabolite profiling experiments.

Metabolite profiling has extensive applications in discovering the mode of action of drugs and in explaining the effect of altered gene expression on metabolism [56]. A fundamental step in metabolite profiling is the unambiguous identification of metabolites in complex metabolite preparations from biological samples. Collections of mass spectra, containing frequently observed metabolites, represent one of the most effective ways to combine the identification efforts currently performed in many laboratories around the world [56].

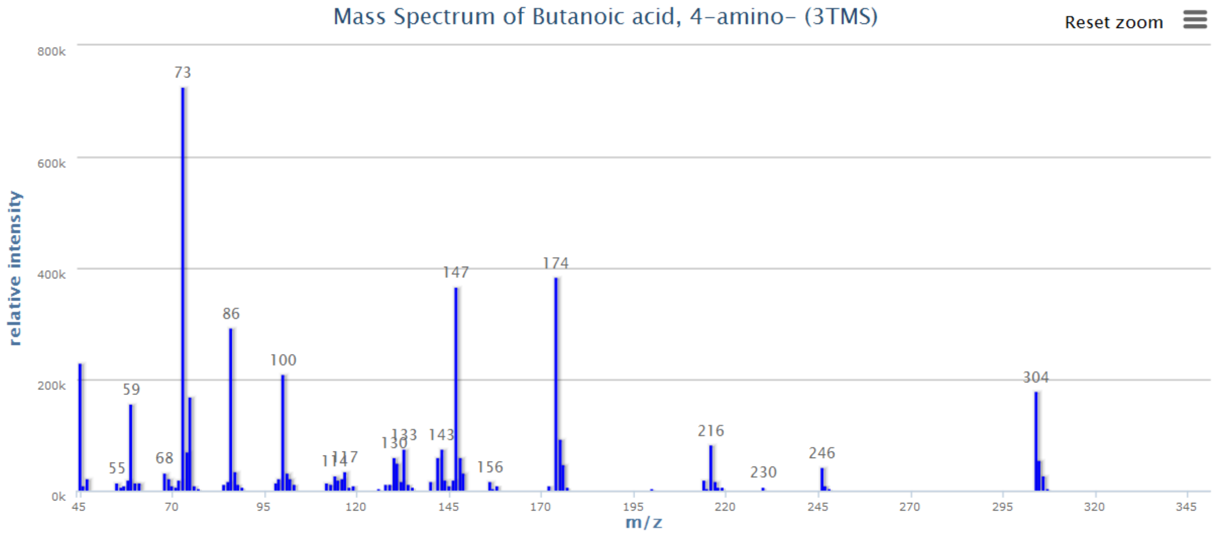


Figure 2.6: **GMD reference mass spectrum**.

The Golm Metabolome Database provides public access to custom mass spectral libraries, metabolite profiling experiments as well as additional information and tools. These libraries of mass spectral and retention time indices can be used in conjunction with software tools to identify metabolites according their spectral tags and RI's [4].

The main goal of the GMD is to act as an exchange platform for experimental research activities and bioinformatics to develop and improve metabolomics by a multidisciplinary approach.

2.4.2 MassBank

MassBank is the first public repository of mass spectra of small chemical compounds (< 3000 Da) for life sciences for sharing them among scientific research community [44].

More than 13000 high precision and accurate mass spectra of biologically endogenous and exogenous substances are available. It offers various query methods (e.g. mass spectral search by exact mass-to-charge ratio m/z) for standard spectra obtained from Keio University, RIKEN PSC, and other Japanese research institutions [4, 44].

MassBank data are useful for the chemical identification and structure elucidation of chemical compounds detected by mass spectrometry [5]. The number of accesses to the MassBank is increasing every year as shown in Figure 2.7.

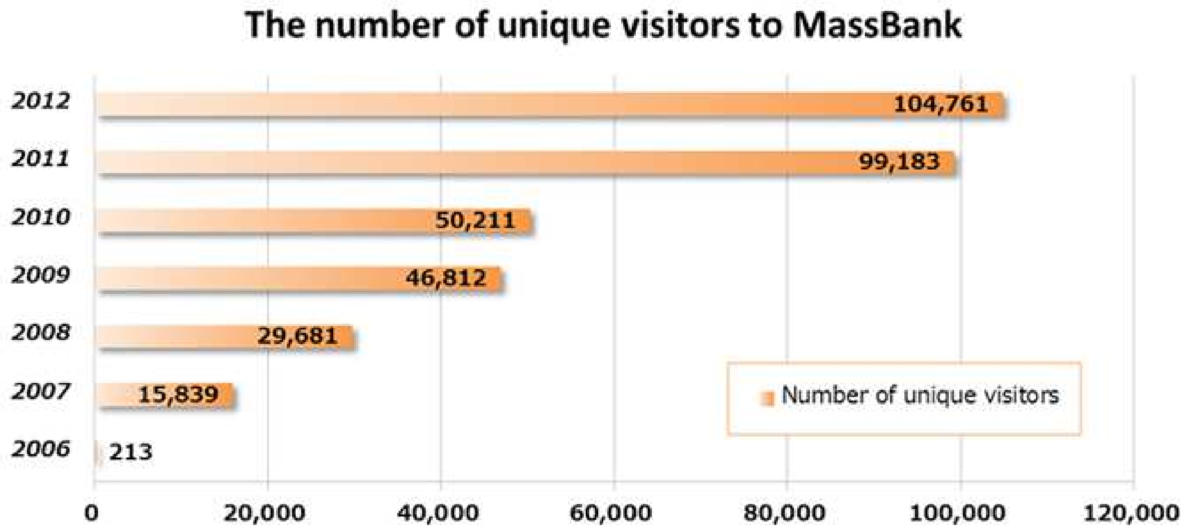


Figure 2.7: **Number of accesses to the MassBank**. (from [5])

2.4.3 METLIN

METLIN is a public, web-based database developed to assist in a variety of applications in the field of metabolomics and to simplify metabolite identification through mass analysis. The data repository has been designed for the archiving, visualization, and analysis of metabolite data [82].

METLIN contains over 60000 high resolution MS/MS spectra and over 240000 metabolites [7], providing the following information from multiple biologic sources [82]:

- Structural and physical data on known endogenous metabolites and drug metabolites;
- High-accuracy Fourier Transform Mass Spectrometry (FTMS) data from reference biofluid samples;
- Reference tandem MS data from known metabolites and metabolite derivatives;
- LC/MS profiles from primarily human and some model organisms

Spectral data (MS/MS, LC/MS and FTMS) can be searched by peak lists, mass range, biological source or disease.

2.5 Disease and Physiology Databases

2.5.1 OMMBID

The On-Line Metabolic and Molecular Basis to Inherited Disease (OMMBID), originally developed by Charles Scriver at McGraw-Hill, is a web-accessible resource describing the genetics, metabolism, diagnosis and treatment of hundreds of metabolic disorders contributed from hundreds of experts. It also contains detailed pathways, chemical structures, physiological data and extensive reviews, that are particularly useful for clinical biochemists.

2.5.2 MetaGene and RAMEDIS

MetaGene is a repository for comprehensive information on over 400 genetic metabolic diseases, including information on differential diagnoses, clinical and laboratory findings [91]. It is designed to be used as a tool to support diagnosis and treatment of patients with rare metabolic disorders. Database entries can be searched by disease name, symptoms, patient information and clinical study author.

The Rare Metabolic Disease Database (RAMEDIS) [90] is a manually curated resource that collects detailed patient information on rare metabolic diseases. It was developed in close cooperation with clinical partners to allow them to collect information on rare metabolic diseases with extensive details (occurring symptoms, laboratory findings, therapy and molecular data).

Thus far, 818 patients have been published with 93 different genetic metabolic diseases. As a universal resource, RAMEDIS allows researchers to extract a diversity of standardized data types, including clinical, biochemical, and molecular [9].

2.6 Compound-Specific Databases

In the category of Compound or Compund-specific databases we can surely count **PubChem**, **ChEBI** and **ChemSpider**.

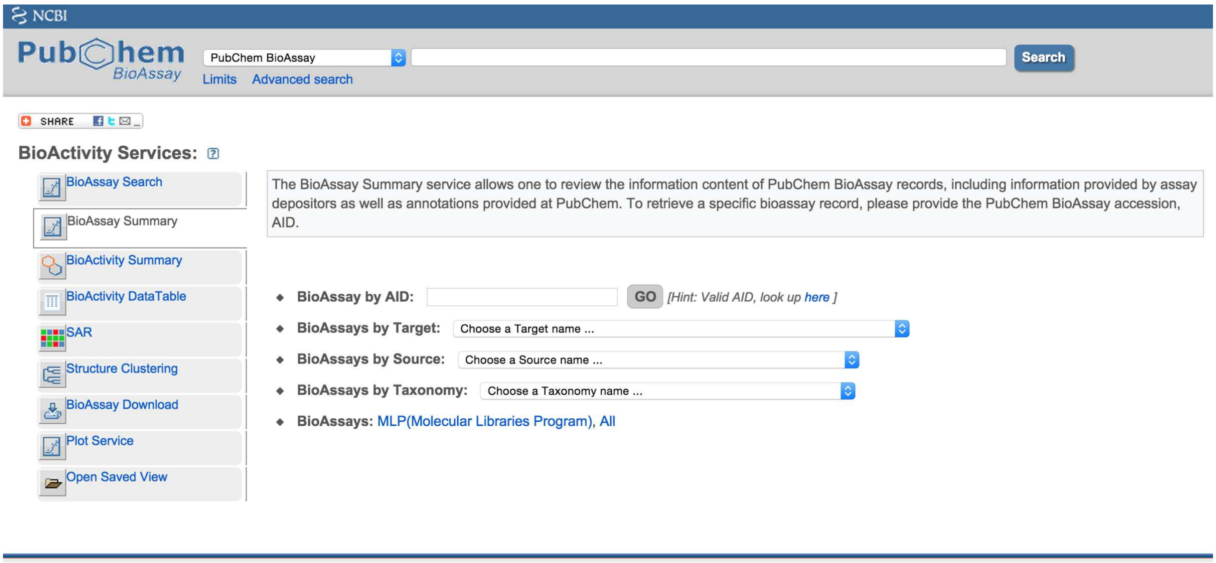
PubChem

PubChem is an open repository for chemical structures and their biological test results, launched in September 2004 as part of a research program under the NIH Molecular Libraries Roadmap Initiative [100].

It comprises the following related databases:

- Substance;
- Compound;
- BioAssay

The Substance database contains more than 180 million records of contributed sample descriptions provided by depositors, whereas the Compound database contains more than 63 million unique chemical structures derived from the substance depositions. The PubChem BioAssay database contains over 1 million bioactivity screens of chemical substances described in PubChem [100, 8].



The BioAssay Summary service allows one to review the information content of PubChem BioAssay records, including information provided by assay depositors as well as annotations provided at PubChem. To retrieve a specific bioassay record, please provide the PubChem BioAssay accession, AID.

- BioAssay by AID: [Hint: Valid AID, look up here]
- BioAssays by Target: Choose a Target name ...
- BioAssays by Source: Choose a Source name ...
- BioAssays by Taxonomy: Choose a Taxonomy name ...
- BioAssays: MLP(Molecular Libraries Program), All

Figure 2.8: **PubChem BioActivity Analysis Service**. It provides a central entry point for accessing bioassay records.

The primary goal of PubChem is to give biomedical researchers access to all this information in a very simple and straightforward way. To accomplish this goal it provides a wide range of web-services with tools for data retrieval, integration and comparison of biological screening results, exploratory structure-activity analysis, and target selectivity examination.

ChEBI

Chemical Entities of Biological Interest (ChEBI) is a freely available dictionary of either natural products (metabolites) or synthetic products involved in the processes of living organisms, focused on small chemical compunds [26]. Molecules directly encoded by the genome (such as nucleic acids or proteins) are not included in ChEBI, as these are amply represented in other databases.

Natural and synthetic products are part of the so-called "molecular entities". The term "molecular entity" refers to any constitutionally or isotopically distinct atom, molecule, ion, ion pair, radical, radical ion, complex, conformer identifiable as a separately recognizable entity.

In addition to these molecular entities, ChEBI contains groups (parts of molecular entities) and classes of entities. Furthermore it includes an ontological classification, whereby the relationships between molecular entities or classes of entities and their parents and/or children are specified [26]. Currently it contains over 40000 fully annotated compounds.

ChEBI uses nomenclature approved by the International Union of Pure and Applied Chemistry (IUPAC) and the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). All the data in ChEBI is non-proprietary or derived from a non-proprietary source and is therefore freely available to anyone.

ChemSpider

ChemSpider is a free chemical structure database providing fast access to over 34 million structures, properties, and associated information linked out to almost 500 separate data sources on the Web, owned by the Royal Society of Chemistry [69, 2].

It was designed to:

- bring together compound data on the web
- make this data accessible and reusable

- provide a publishing platform for the addition and preservation of data

ChemSpider adopt a a crowdsourcing approach: it can be updated with user contributions including chemical structure deposition, spectra deposition and user curation. Curation, which involves ensuring the accuracy of the data in a digital database (7), is an essential problem for any reference source. ChemSpider allows registered users to enter information and annotate and curate the records. The requirement to register and login is to prevent anonymous acts of vandalism [69].

Chapter 3

QTREDS for Omics data management

Quality and TRacEability Data System (QTREDS) is a software platform for cross-omics data management, originally developed to address the specific needs of the Sequencing and Genotyping Platform (SGP) at the Center for Advanced Studies, Research and Development in Sardinia (CRS4) where I have conducted most of my research activities.
Tracking and monitoring all the phases of the laboratory activities can help to identify and troubleshoot problems more quickly, reducing the risk of process failures and their related costs. QTREDS has been designed with these goals in mind to meet the specific requirements of the SGP laboratory, where it has been successfully used for over a year. Thanks to its flexibility the system can be easily adapted to meet the requirements of other omics laboratories. Currently the system is undergoing an optimization process in order to adapt it for metabolomics laboratories.

3.1 Introduction

High throughput technologies give the opportunity to study the genome, proteome and metabolome at a global scale. This opens up new possible scenarios in desease diagnosis. At the same time, the rapid development of these technologies has produced two main consequences: a large amount of data generated, new and complex laboratory procedures [64, 86, 61].

A Laboratory Management Information System (LIMS) is designed considering the need to carry out the research in an efficient and transparent manner allowing the implementation of different quality control strategies and improving the accessibility of the instruments. The improvement of the laboratory activities involves three primary factors: technology, information and people. In order to develop an effective LIMS all the three resources must be recognized and a thorough study of the laboratory processes must be taken into consideration.

Till the late 1970s all the activities concerning the management of laboratory samples, associated analysis and reporting were time-consuming and error prone due to manual processes [71]. This gave some organizations impetus to optimize data collection and laboratory procedures. Initially some custom in-house solutions have been developed, while some analytical instrument manufacturers, at the same time began to develop some commercial systems to run on their instruments.

The term LIMS entered the commercial world in the early 1980s to describe systems used in the pharmaceutical and related industries as Quality Assurance and Quality Control tools [71, 37].

In 1982 the first generation of LIMS was introduced in the form of single centralized minicomputer provided with automated reporting tools. Second generation LIMS became available in 1988 and used third-party commercial relational databases to provide application-specific solutions. Most of them relied on minicomputers [18]. Third generation LIMS began in 1991, as personal computers became more powerful and prominent. They combined the personal computer's easy to use interface and standardized desktop tools with the computational power and reliability of minicomputer servers in a client/server configuration. By 1995 fourth generation LIMS came into the picture decentralizing the client/server architecture further, optimizing resource sharing and network throughput by enabling process to be performed anywhere on the network [71].

From 1996 to 2002 additional features and functionalities were included in LIMS, from wireless networking capabilities and geo-referencing of samples, to the adoption of XML standards [71].

In the latest generation LIMS the adoption of web oriented software technologies assumes a key role [86] together with a rising interest in the Software as a Service (SaaS) distribution model through which the customers can save the expense of license fees and the costs of hardware and maintenance.

This chapter provides an introduction to QTREDS, a software platform initially born to address the specific needs of the CRS4 sequencing laboratory. The main purpose of our in-house solution was to set up a system that provides the researchers with a complete knowledge of the laboratory processes at each step, managing and verifying the:

- workflow creation;
- samples traceability;
- diverse experimental protocol definitions;
- inventory of reagents;
- users' roles and privileges.

Why develop a LIMS from scratch rather than buy a commercial one? A great number of proprietary LIMS have been developed. STARLIMS [14], Exemplar LIMS [77], LABVANTAGE SAPPHIRE citelabvan11 just to name a few, allow customers to benefit from vendors long-established experience and valuable resources.

On the other hand, most often these commercial solutions are large, complex and feature-rich products designed to be sold to large laboratories. Their license fees are usually prohibitive and each extra feature or module they provide might come at additional costs [105]. Furthermore the laboratories have to buy also the servers, peripherals, storage devices and other software licenses (such as databases, load balancers, etc...). Most small or medium-sized laboratories cannot afford this expense [52].

Many commercial LIMS vendors are now offering rented, hosted and SaaS-based LIMS solutions. The rental approach is almost identical to the purchased one, except that the laboratory rents the software rather than purchasing the license. All the other purchases (hardware and additional software) remain the same, as do other costs [52]. The major difference is a staged payment for the software.

Some LIMS vendors provide hosted thin-client solutions. A thin-client LIMS is an architecture which offers full application functionality that can be accessed through a simple web browser. Rather than requiring the customer to purchase hardware and software, the customer simply uses the software running at the vendor's site. However, hosted software providers often do not rewrite their products to take advantage of new Internet-based technologies, but simply put a different front-end onto dated systems.

Another approach is the cloud-based model. While it bears some resemblance to the hosted model, the cloud-based SaaS model is usually built from the ground up using a service-oriented architecture. They are designed for multi-tenancy, where multiple customers share the same instance of the application running on the same operating system, on the same hardware, with the same data-storage mechanism.

These software applications are designed to virtually partition their data and configurations, so that customers do not see each other's data and work within their customized virtual application instances. According to this model, data are stored on the servers of the service provider and this fact can raise a number of issues if data confidentiality is critical, as it often happens in the biomedical field [52, 17].

Many open-source LIMS are now available, but some of them had not been published when we started the development of QTREDS in early 2011 [17, 93, 92].

Before starting the development phase, we tried some of the solutions available at the time: we tested Open-LIMS [55] by installing it on our server but it was very unstable, in fact it was not recommended by the developers to use it in any productive environment; we also tried Bika Lims [19] which is one of the leading open source LIMS, with a wide range of applications from agriculture to environmental monitoring. It offers many functionalities for free, but optional modules at a cost. It is based on Python and the Plone content management system. We programmed web services with Python Zope and Plone, and our experience is that it is not a trivial software stack. Furthermore Plone performs better on a dedicated server and that could represent an hidden cost. Other systems we have looked at, but not considered because we felt they did not correspond well to our needs are: LabKey Server [62] a very much oriented to data analysis tool. In our case, the experiments are done "as a service", and the results are given to the researchers. The analysis are not done in the laboratory.

SLIMS [96] a Sample-based Laboratory Information Management System with a web-based interface to create, edit and view sample information. SLIMS is designed to store and manage biological data in fact it features a micro-plate annotation tool and supports SDS-

PAGE gels. It can also generate and export reports, but it does not provide any inventory management system and its web interface does not include the latest web technologies.

GNomEX [64] a very complete platform that includes a next generation sequencing/microarray LIMS, an analysis project center, and an application for annotating and programmatically distributing genomic data. It is much more complex than QTREDS and it was designed for large research centers and clinics. Because of that it does not meet the needs of a relatively small entity like the CSGP laboratory. We tested also other solutions, but some of them were in a very early development stage or they were buggy and crashing and not stable enough to run in a production environment [92, 39].

The most important factor for the development of an in-house solution, even more than the license fees or the confidentiality issues, was the fact that the application had to be developed to meet the specific needs of the researchers of the CSGP laboratory. When we started to develop QTREDS, the main project in our laboratory was related to the DNA sequencing of 2100 individuals from Sardinia [80]. At the same time other projects concerning RNA and exome sequencing of a large part of the same set were in their early stages.

While for the DNA sequencing the techniques and procedures in use were well defined and standardized, in the case of RNA and exome sequencing, the methodologies and the protocols had not been decided yet, so we began to develop QTREDS not only to collect the data, trace and manage each lab activity, but also to help researchers choose the best protocol to implement for their experiments.

We designed a system flexible and responsive enough to keep up with the speed at which the laboratory evolves.

3.2 Implementation

3.2.1 Development methodologies

QTREDS has been developed adopting an Agile software development approach. Indeed, we have worked closely and continuously with researchers, operators, managers and other stakeholders involved in the project. In particular, we followed a Behavior-Driven Development (BDD) strategy, asking questions focused on the behavior of the platform before and during the development stages, to avoid or at least reduce misunderstandings between stakeholders.

Requirements were written down in the form of user stories, which described the expected use of each part of the application. User stories, a lightweight approach to use case analysis, have been compiled and continuously refined, in nontechnical language allowing all stakeholders to be involved in the process of creation and prioritization of the requirements.

Starting from a general description of the needs of the CRS4 SGP and the main functional requirements that the system was expected to have, we created a working but incomplete prototype, refining it constantly through a continuous interaction with the researchers and the personnel of the laboratory until the achievement of the desired results.

3.2.2 Software architecture and design patterns

QTREDS is a web application with a client-server architecture developed in the Ruby programming language, using the framework Rails [43].

The application, according to the architectural pattern known as Model-View-Controller (MVC), has been organized dividing the code into three kinds of components (Figure 3.1). Models implement business logic and are concerned with the manipulation of the data: how to store it, to change it or move it. Typically for each type of entity managed by the application, we have created a corresponding model that encapsulates it. Views serve as the interface between application users and model data.

They contain information about the models with which users can interact and manage how to display it. Controllers have the role of intermediaries between views and models in both directions: when a user interacts with a view, a precise controller action corresponding to that activity is invoked and it saves or updates data from the user to the model. On the other hand, the controller makes the model data available to the view so that it can be displayed to the user. One important job of the Model is to persist data which requires that some correspondence must be established between the operations on a model object in memory and how it is manipulated in the storage tier.

Models implement the Active Record architectural pattern, providing an Object Relational Mapping (ORM) layer which supports a wide variety of Relational Database Manage-

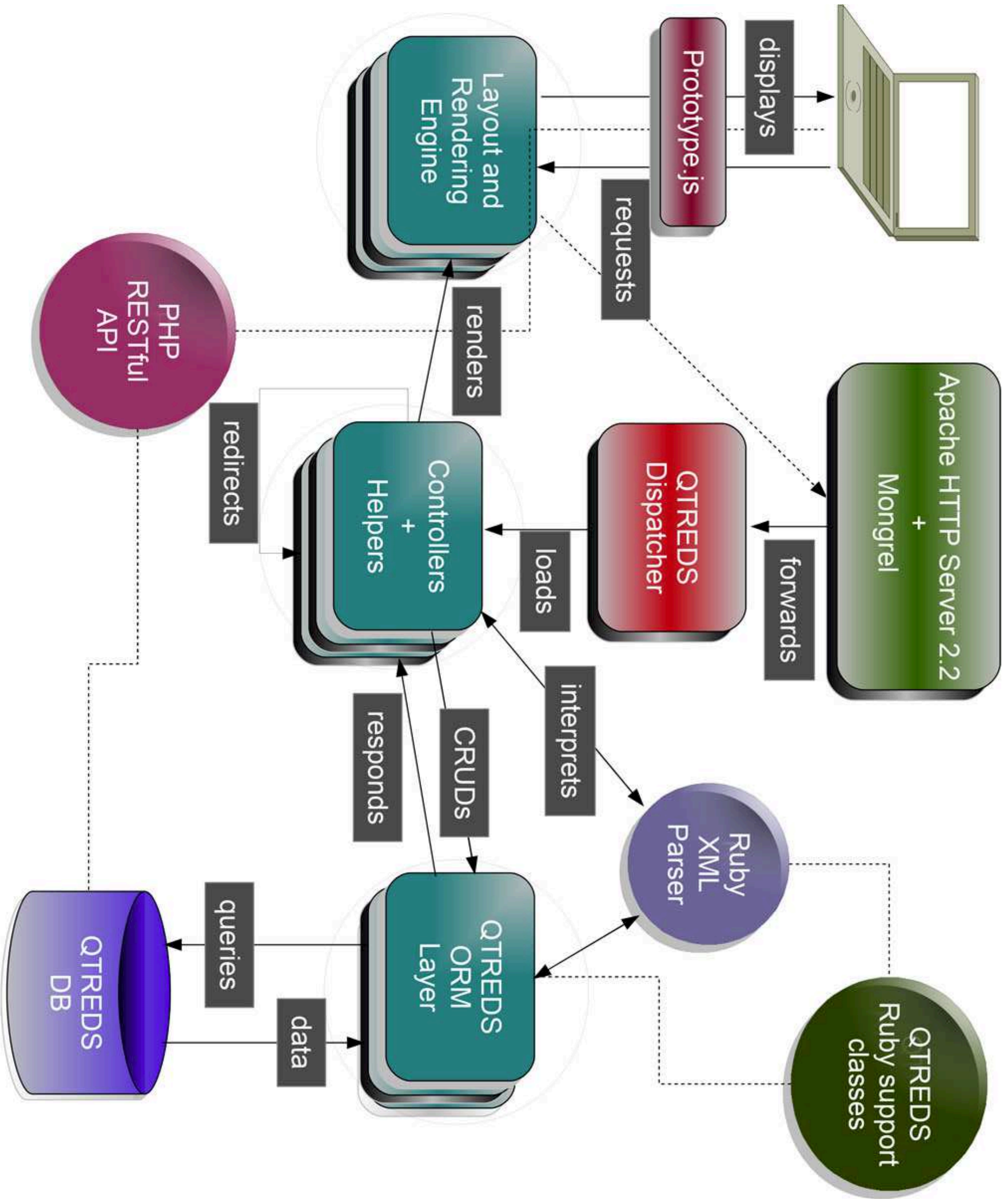


Figure 3.1: **QTREDS architectural overview**. QTREDS has been developed according to the MVC software architecture pattern. The Protocol Parser and the internal class libraries have a key role for the generation of the experimental workflows.

ment System (RDBMS). For the QTREDS persistence tier we have chosen the MySQL RDBMS [66]. Each instance of a model class corresponds to a single row in a specific table of the MySQL database. The model object has built-in behaviors that allow to directly operate on the tables of the storage layer of the application.

The implementation of QTREDS also relies on the use of different open-source programming libraries. The web user interface has been developed combining the Ruby's built-in erb templating system with the Prototype JavaScript Framework [85] that enabled us to deal with the Asynchronous JavaScript And XML (AJAX) [34] technology in a very easy and efficient way.

Furthermore the use of the script.aculo.us [33] set of Javascript libraries provides us with a visual effects engine, that we used to enhance the interactive user experience with the application.

3.3 Functional overview

All the activities and operations allowed by the QTREDS platform can be assigned to four different functional blocks:

- workflow management system
- sample handler
- inventory management system
- authorization system.

3.3.1 Workflow management system

The workflow management system is a key component of our application. Figure 3.2 illustrates the main concepts related to this functional block: it has the responsibility for defining and verifying a protocol and to convert it into the sequence of steps and tasks that represent a particular procedure or experiment.

A protocol, in our system, is a formal representation of an experimental procedure, expressed in the XML language (see Listing 5.1) that has to be compiled according to a strict set of rules that we defined and collected in an XML Schema Definition (XSD) document. This task can be accomplished manually by an authorized member of the laboratory with basic informatics knowledge. But writing down a protocol manually can be a very long, boring and error-prone task that requires the observation of precise syntactic and semantic rules. To reduce the probability of error and to allow users with no technical background to create an experimental protocol, we have developed a user-friendly visual tool, which we describe later in this article.

The XML protocol is interpreted and checked by the protocol parser module that processes the document, extracting and sending information to some support classes. Coordinating the activities of these classes and of the experiment-related controllers and views, it provides the system with all the information needed to graphically represent the experiment workflow as a sort of "state diagram" that guides the user step by step, enabling him to manage and monitor the progression of his experiment.

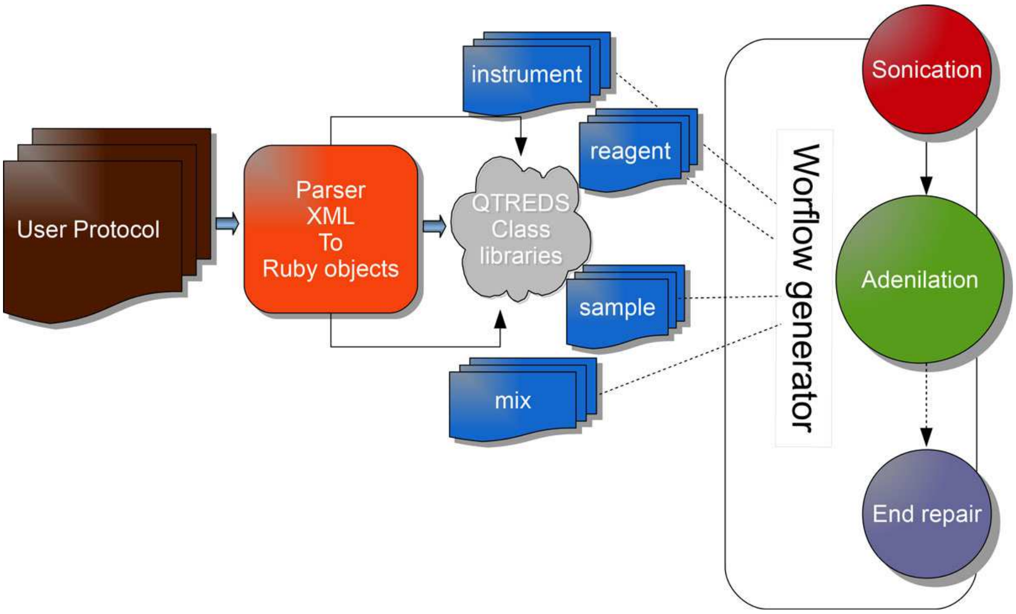


Figure 3.2: **Workflow management system - the protocol parser**. The parser checks and interprets the experimental protocols written in the XML format and with the help of the internal class libraries provides the controllers with the information needed to generate the experimental.

Figure 3.3 illustrates the steps of an exome library preparation workflow of a running experiment. Exome library preparation is one of the procedures performed within the exome sequencing technology. The workflow is graphically represented as a sequence of different color balls. Each labeled ball describes a single step of the laboratory procedure (sonication, end repair, adenilation, etc...) and its color defines its state: a green ball represents a completed activity, an orange ball an activity ready to be executed or in progress and not completely carried out; a red ball indicates that the corresponding activity has been terminated abnormally for some reason, and that the workflow cannot be carried out. Grey balls represent steps of the workflow not yet available that require the completion of previous activities to be performed (Figure 3.3). When a user clicks on the ball of the step he wants to begin - order as we said, is mandatory at the moment - he will get a web page with forms to enter data and information related to that particular step.

If default values have been set in the protocol/workflow definition then these will be already filled in the form. The user will then only have to fill what is different from the default, and then start the process. The complexity and level of detail of each of these web pages depends on how the users have defined that step in the protocol: it can be general or describe precisely every single phase of the process. It is up to the "workflow supervisor", i.e. the user

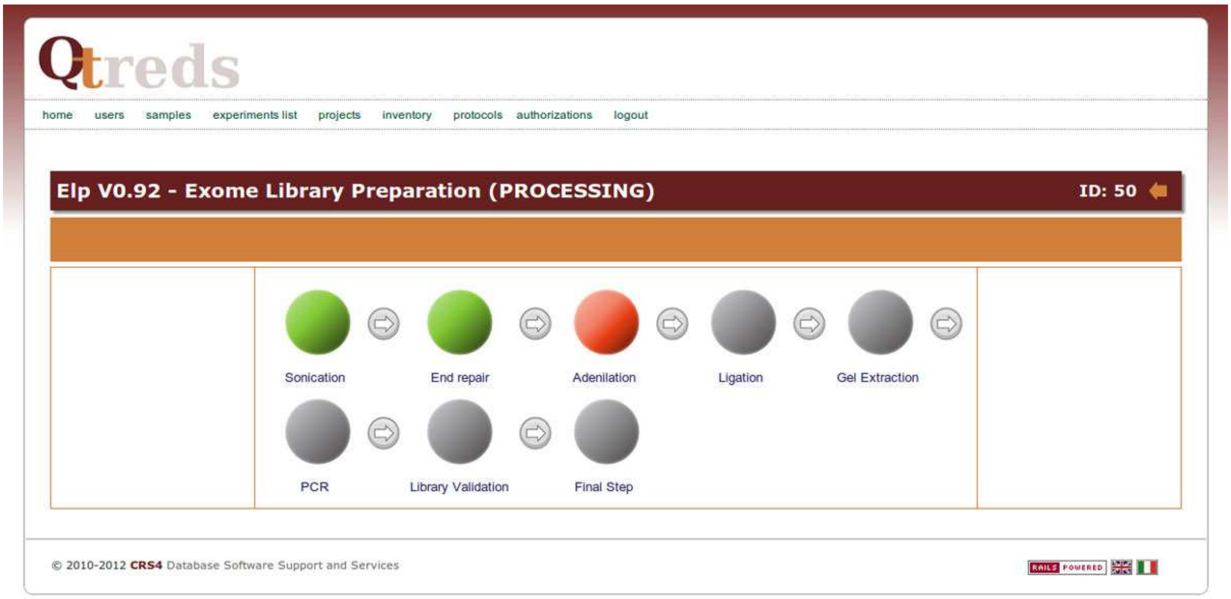


Figure 3.3: **Experiment workflow**. QTREDS represents the experiment workflow as a sort of state diagram that guides the user step by step, enabling him to perform and monitor each phase of the experiment.

authorized to create workflows, to decide the level of granularity of the information and the community standards to be used (e.g., MIBBI [89], ISA-Tab [72]).

```xml
<protocol xmlns="http://www.w3schools.com"
          xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
          xsi:schemaLocation="http://www.w3schools.com protocols.xsd" name='Exome Library Preparation'
     output='P,S' version='0.9' input='U'>
    <activities>
        <activity name='Sonication' id='1'>
            <instrument name="Covaris S-Series">
                <input name='Device parameters' id='1.1'>
                    <attribute type='decimal' key='duty cycle'>8</attribute>
                    <attribute type='decimal' key='intensity'>3</attribute>
                    <attribute type='decimal' key='cycles per burst'>200</attribute>
                    <attribute type='decimal' key='second frequency sweeping'>60</attribute>
                    <attribute type='decimal' key='number of cycles'>2</attribute>
                </input>
            </instrument>

            <input name='Note' id='1.2'>
                <attribute type='textarea' key='note'/>
            </input>
            <comment title='Alternative procedure'/>
        </activity>
        <activity name='End repair' id='2'>
            <input name='Schema Multiwell-Samples' id='2.1'>
                <attribute type='file' key='Path'/>
            </input>
            <instrument required='true' name='Centrifuge' category='centrifuga' id='2.1'>
                <input name='Device parameters' id='2.2'>
                    <attribute unit='rpm' type='decimal' key='Cycles per minute'>600</attribute>
                    <attribute unit='seconds,minutes' type='decimal' key='Duration'>5</attribute>
                </input>
            </instrument>
```

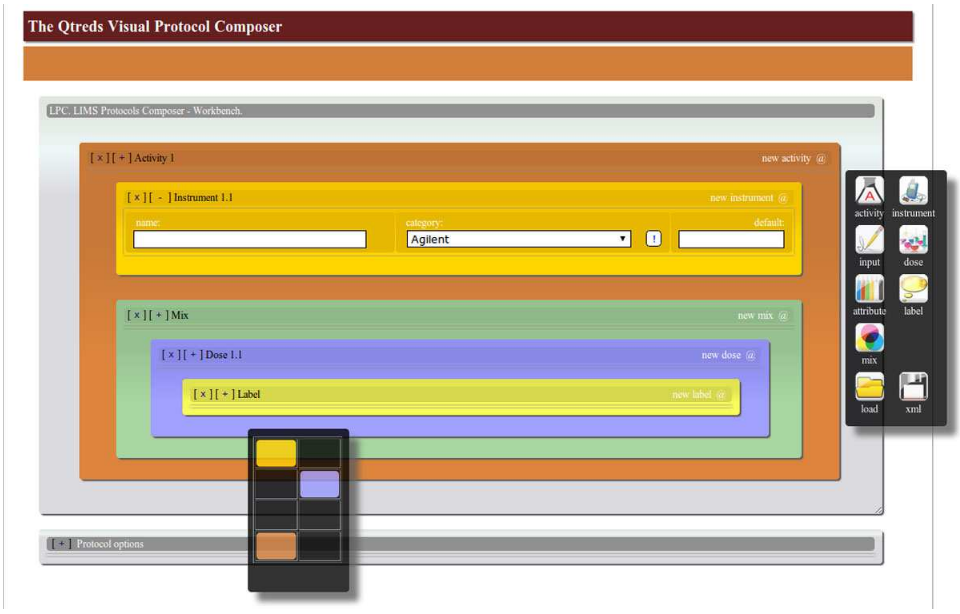
Listing 3.1: Excerpt of the Exome library preparation protocol

Workflows can be created directly in XML or using the Visual Tool. Plugins can be implemented for the Visual Tool to check for the required information of the chosen standard. Plugins can also be written to export the data to various formats for inclusion in submissions to public databases. None of the laboratories we are collaborating with, is equipped with robots that can transfer samples and reagents between machines; because of that human intervention is always required between different steps of the workflow. So far the workflows that have been implemented reflect this and do not automatically activate the next step.

3.3.2 Visual Tool

Whenever an authorized user creates a new protocol he has to upload the related XML protocol description file to the system. At this point, the system checks the file for syntactical correctness and semantic coherence and it stops when the document does not follow the rules defined in the XSD document.

As already mentioned, the process of defining and writing down an experimental protocol can be very complex and annoying. In order to simplify this task, we developed a special tool for creating protocols: it allows the user to drag-and-drop graphical objects to create experimental protocols in the XML format (Figure 3.4). Each visual object has the aspect of a box and can be filled up with other objects, according to the rules defined in the XSD document. This reproduces the hierarchical structure of the XML protocols written down manually.



category: Agilent

Figure 3.4: **Overview of the graphical user interface to design experimental protocols.**. The visual tool allows the user to drag graphical objects from the right-most floating palette and drop them on the workbench. Combining those objects the user can create experimental protocols and export them to XML files.

The interface is made up of two main components: a workbench in which the user can combine all graphical elements, and a floating palette in which he can find different elements needed to define an experimental protocol: an activity object, that represents a single step of an experiment, an instrument object, which can identify any device or machine present in the laboratory, a dose object to describe a particular reagent to use and so on. The user can combine all these elements, organize them in the appropriate hierarchical order and set all the parameters that are needed. The result of this graphical representation can be easily exported in the XML format and used by the workflow management module of the system. Through the visual tool the user can also import an existent XML protocol, convert it to a graphical representation and manipulate it with the editing tools provided.

The tool has been implemented in pure HTML5 and JavaScript. HTML5 defines an event-based mechanism and additional markup for natively supporting drag and drop operations. This allowed us to develop a faster and more responsive tool, without the support of any other JavaScript library or framework.

3.3.3 Sample handler

QTREDS enables the users to enter either one single sample or multiple samples at a time using an Excel spreadsheet-based wizard. In the first case the user should fill in a web form providing some mandatory information, for instance a unique sample identifier (sample id).

In the second case, the user loads a group of samples through an Excel file: the wizard allows the mapping of each column of the spreadsheet to one of the attributes used by the system to describe a sample. After a sample is entered into the system, a new record is saved to the database with its defined set of attributes.

If the number of columns of the spreadsheet mapped exceeds the number the samples' attributes or if the user needs to associate a sample with some extra parameters, the system will store them in a different table. To characterize each sample, we have defined two attributes, the original id that corresponds to the identifier with which a sample is submitted to the laboratory and the lab id that is an internal parameter used by the system for the sample tracking process. Samples may be inputs of an experiment in which they are processed to generate new samples. The output samples created, keep their relationship with the inputs, holding the same original id value, while they change their lab id in relation to the particular experiment in which they were involved.

QTREDS checks for the uniqueness of the combination of the two attributes, refusing samples with the same original id and lab id.

Depending on the experimental procedure carried out, the system internally associates to each sample an attribute called state, which describes the current status of the sample (for instance, a DNA sample could be processed to construct a DNA library; in this case the value of the attribute state will change from "untreated" to "library"). The value of this attribute is exploited by the system to identify which processing activity can be done and the class of experiment each sample can be associated with.

3.3.4 Inventory management system

The Inventory Management module allows the tracking of all the reagents and items used by the researchers for their experiments.

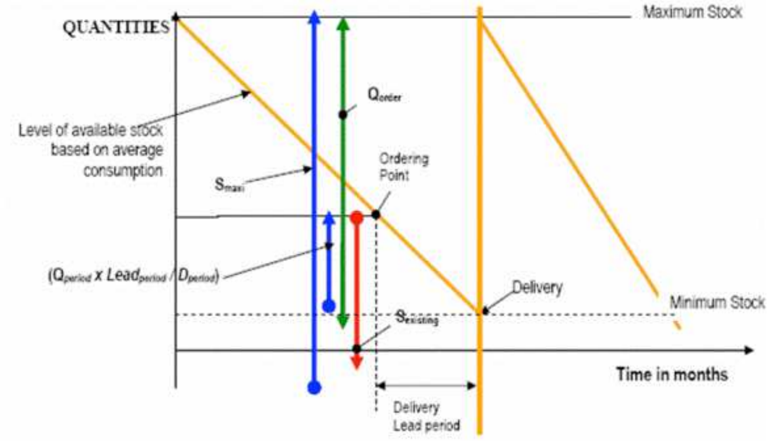


Figure 3.5: **Minimum inventory level**. Minimum stock levels can be handled to avoid shortages of essential products.

It includes four different components:

- *catalog*: all items (consumables, reagents, tubes, etc...) involved in some laboratory process, are represented in QTREDS as abstract entities that we defined as categories. A category is not a physical item that can be found inside the laboratory, but it is an abstract description of a set of objects that share some features. The catalog collects all these categories, allowing the basic CRUD operations on them;

- *stock*: the smallest physical instance or unit of a particular category is referred to as stock. A stock indicates an item physically present in the laboratory and it specifies its quantity. To prevent the danger that a running experiment may be interrupted due to shortage of reagents or other consumables, the system provides a mechanism of "real time" assessment of stock levels, warning the researchers if some item goes below a defined threshold (Figure 3.5);

- *personal stock*: before starting an experiment, QTREDS lists all reagents and consumables needed to conduct it. The personal stock is a sort of "shopping cart" in which each researcher must insert all the items required to perform his experiment. Each experiment is represented as a sequence of consecutive steps called activities. The system does not allow the user to begin his experiment if his personal stock does not contain at least the reagents needed to perform the first activity;

- *topology*: starting from a simple YAMAL file, QTREDS builds a hierarchical map of the laboratory modeled as a rooted tree. The root of the tree is the whole laboratory, the subsequent nodes are the different rooms, then the freezers, going down to the granular level of the shelves, racks, etc. This representation is used by the system to track sample location.

3.3.5 Authorization system

QTREDS is a web-based multi-user application. Many users can access the system simultaneously, define their own projects, experiments and manage the inventory. Within this context, it is very important the definition of user privileges and roles.

The authorization module defines different user roles, each with a different access profile; each role includes a set of features and privileges to which the assigned user have access. So far, we have implemented six main roles: administrator, supervisor, simple user, inventory manager, analyzer and viewer. Depending on the role assigned, each user is allowed to perform different levels of operations and access different kinds of information. For example a simple user can see only data related to his experiments or to the projects in which he is involved, while the administrator has a complete view of all the activities and data processing operations in the laboratory.

A user can have different roles in different projects. The core of the authorization module includes a set of database tables in which is stored all of the information about user roles and privileges, and a centralized authorization function. This function provides access rights and privileges to each user according to:

- user identity (*user_id*);
- specific action to be performed (*auth_id*);
- some additional parameters.

The response this function returns can be a boolean value, which tells if the user is allowed or not to perform that action or a SQL query that is used by the system to extract all the information a user can access to, according to his role.

Each user's request to gain access to a specific resource, involves a call to the centralized authorization function, passing along some arguments (for example, the *user_id* and the *auth_id*) to it. To retrieve these parameters, the system has to perform some queries on those database tables that are related to the authorization mechanism. In order to reduce duplicate queries and repeated function calls, we have implemented a caching strategy that allowed us to improve the performances of the system in terms of responsiveness and reactivity.

3.4 Conclusions

QTREDS has been designed to facilitate data management for multiple "omics" experiments. It has been developed starting from the needs of the CRS4 SGP [67], where it has been used since late 2011 to make almost one hundred DNA library preparation and sequencing experiments, processing thousands of samples. We received two different kinds of reaction from the users of the QTREDS system: the ones working in team fully adopted the tool for their daily activities, providing us continuously feedback for the development of new features; on the other side some of those working on individual assignments had more difficulties to accept it.

A positive point for the users of QTREDS has been the fact that it has an iPad optimized user interface: all users of our laboratory are equipped with tablets and can enter data into our system as they would with a paper notebook while moving around for the experiments. Another point of satisfaction has been the implementation of the "personal stock" tool in the inventory management system. It warns users about all consumables and items needed, helping them in the smooth run of the experiments by preventing an abrupt stop due, for example, to lack of a given reagent. When a first demo version of QTREDS was released, some users complained the absence of simple computational tools to convert measurement units or to calculate some common physical quantities like mass, concentration and so on.

The requests have been addressed and satisfied in a later stable release introducing new elements and attributes in the XSD file and enriching the XML protocols with new functionalities.

As a whole, most of the users appreciated the way QTREDS improved the management of information especially when there was a huge increase of the number of samples being treated.

A new version of QTREDS is currently being tested and it is going to be released. The upcoming version is provided with an efficient Application Programming Interface (API) in order to allow a smart and automated access to information. The API has been implemented according to the REpresentational State Transfer (REST) architecture [31]. Using this API any authorized user or system can retrieve resources and information via a standard Hypertext Transfer Protocol request, appending the appropriate query parameters to the URL.

The RESTful web service, based on a dedicated web server, handles requests from clients, processes and then returns the appropriate response as an XML document. The API can also be used to insert data into the QTREDS database tables, creating this way, a bidirectional communication channel between our system and any other external application or tool. The new release will also provide a complete reporting system to visualize and export data in different file formats.

Thanks to its flexibility the system can be easily adapted to address the issues and the needs of other kinds of laboratories; therefore I am actively involved in the development of an implementation of that for a research group in the field of Metabolomics with whom I am intensely collaborating.

Chapter 4

Analytical technologies

4.1 Introduction

The metabolome is very complex entity in terms of both chemical diversity and quantities of each metabolite. Metabolome analysis aims to the identification of all these metabolites in a large number of small samples and possibly even to quantify the amount of each of them. Currently it is not possible to analyze the entire range of metabolites by a single analytical method, not even from the simplest organisms.

The main analytical techniques that are used for metabonomic studies are based on nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS). The last-mentioned technique requires a pre-separation of the metabolic components using either gas chromatography (GC) after chemical derivatisation[1], or liquid chromatography (LC), with the newer method of ultra-high-pressure LC (UPLC) being used increasingly [57].

Other less common techniques such as Fourier transform infra-red (FTIR) spectroscopy and arrayed electrochemical detection have been used in some cases.

To choose the most suitable analytical strategy, it is important to consider the following aspects: (i) the kind of information needed, (ii) the kind of chemistry expected and (iii) the analytical facilities available [99].

In general, in metabolomics we can distinguish three different strategies:

- Fingerprinting In this strategy, a metabolic "signature" or mass profile of the sample

[1] Derivatization is a technique used in chemistry which transforms a chemical compound into a product of similar chemical structure, called a derivative.

of interest is generated and then compared in a large sample population to screen for differences between the samples. When signals that can significantly discriminate between samples are detected, the metabolites are identified and the biological relevance of that compound can be elucidated [73];

- Profiling Metabolite profiling aims at the analysis of a larger set of compounds, both identified and unknown with respect to their chemical nature. Profiling is typically done by chromatography in combination with MS or by capillary electrophoresis (CE) combined with MS [73, 99];

- Target Target analysis has been applied for many decades and includes the determination and quantification of a small set of known metabolites (targets) using one particular analytical technique of best performance for the compounds of interest [73].

Sometimes these strategies share some analytical approaches, but typically they are implemented quite differently. Usually fingerprinting is mostly based on direct spectrometric measurement by NMR, or mass spectrometers (MS), while profiling and target analyses require, in general, a separation of the compounds by gas or liquid chromatography (GC or LC) prior to the spectrometric detection by UV, NMR, or MS [99].

As already stated, due to the complex nature of the metabolome, no single methodology can detect the complete metabolome in one step.

NMR and MS approaches are highly complementary, and use of both is often necessary for full molecular characterisation. MS can be more sensitive with lower detection limits provided the substance of interest can be ionised, but NMR spectroscopy is particularly useful for distinguishing isomers, for obtaining molecular conformation information and for studies of molecular dynamics, and given the now increasing use of cryoprobes, it is becoming ever more sensitive [57].

To properly select an analytical methodology, the following parameters should be taken into account: (i) Chemistry (polarity, concentration, volatility, etc.); (ii) Concentration (trace or massive amounts); (iii) Matrix (interference from coextracted substrate or may be from major components in the sample) [99].

In this chapter I will provide a brief overview of the principles of these key analytical techniques.

4.2 NMR

High-resolution NMR spectroscopy is a non-destructive technique, widely used in chemistry, that provides detailed information on molecular structure, both for pure compounds and in complex mixtures as well as information on absolute or relative concentrations [57]. It was developed by exploiting the phenomenon of nuclear magnetic resonance for recording the magnetic properties of atomic nuclei. The NMR phenomenon was soon later demonstrated for protons.

Since its discovery in the 1940s, Nuclear Magnetic Resonance (NMR) Spectroscopy underwent rapid technical growth. In the late 1960s Fourier transform NMR spectroscopy entered the scene and next in the 1970s the implementation of superconducting magnets permitted the beginning of the application of NMR spectroscopy for the metabolite profiling of biofluids [83].

The use of ^{1}H NMR for metabolic studies was described as early as 1977 when it was shown that ^{1}H signals could be observed from a range of compounds in a suspension of red blood cells, including lactate, pyruvate, alanine and creatine. A great deal of metabolic information can be derived from such metabolic studies and it was soon recognized that ^{1}H NMR of body fluids has a considerable role to play in areas of pharmacology, toxicology and the investigations of inborn errors of metabolism [57].

Further NMR technical improvements in the 1990s, namely stronger magnetic fields and introduction of cryo-cooled NMR probes, have led to an enormous boost in NMR sensitivity; the signal to noise ratio has increased significantly and still improves. Today, the detection limit of metabolite concentration is of the order of μM [83].

4.2.1 Theoretical Principles

Matter is composed of molecules built of atomic nuclei with a characteristic proton/neutron composition. Nuclei are surrounded by electronic "clouds"[57].

Subatomic particles (electrons, protons and neutrons) can be imagined as spinning on their axes. Besides charge and mass, a further property of these particles is an angular momentum known as spin [11].

The total spin of a nucleus depends on its nucleon content. In many atoms (such as ^{12}C) these spins are paired against each other, such that the nucleus of the atom has no overall spin, while in other atoms ^{1}H and ^{13}C, the nucleus does possess an overall spin.

The rules for determining the net spin of a nucleus can be summarized as follows:

- If the number of neutrons and the number of protons are both even, then the nucleus has spin equal to zero;
- If the number of neutrons plus the number of protons is odd, then the nucleus has a half-integer spin;
- If the number of neutrons and the number of protons are both odd, then the nucleus has an integer spin

In the absence of an external magnetic field, these orientations are of equal energy. If a magnetic field is applied, then the energy levels split as shown in Figure 4.2. Each level is given a magnetic quantum number, m. When the nucleus is in a magnetic field, the initial populations of the energy levels are determined by thermodynamics, as described by the Boltzmann distribution. This is very important, and it means that the lower energy level will contain slightly more nuclei than the higher level. It is possible to excite these nuclei into the higher level with electromagnetic radiation. The frequency of radiation needed is determined by the difference in energy between the energy levels[11].

The nucleus of an atom like ^{1}H has a positive charge and is spinning. This generates a small magnetic field and therefore the nucleus possesses a magnetic moment, m, which is proportional to its spin I:

$$\mu = \frac{\gamma I h}{2\pi}$$

γ is the gyromagnetic ratio of the atomic nucleus, while h is Planck's constant.

The energy of each level is given by:

$$E = -\frac{\gamma h}{2\pi} m B_0$$

where B_0 is the magnitude of the magnetic field. The transition energy (the difference in energy between levels) is expressed by the following equation:

$$\Delta E = \frac{\gamma h B_0}{2\pi} \tag{4.1}$$

This means that ΔE grows proportionally with the strength of the magnetic field B_0.

To describe the interaction of the magnetic field with the nucleus, we will assume that the nucleus acts as a charged particle in a magnetic field. If a nucleus (of spin 1/2), spinning on its axis, is exposed to magnetic field B_0, its axis of rotation will precess around the magnetic field as shown in Figure 4.1. The frequency of precession is termed the *Larmor frequency* and it is identical to the transition frequency ($\Delta E = \frac{\gamma B_0}{2\pi}$).

The potential energy of the precessing nucleus is given by:

$$E = -mB_0 cosq$$

where q is the angle between the direction of the applied field and the axis of nuclear rotation. At this point the lower energy level contains slightly more nuclei than the higher level.

Resonant absorption by nuclear spins will occur only when electromagnetic radiation of frequency equal to the Larmor precession rate is being applied to match the energy difference between the nuclear spin levels in a constant magnetic field of the appropriate strength. Thsi radiation can excite nuclei in the lower energy level into the higher level.

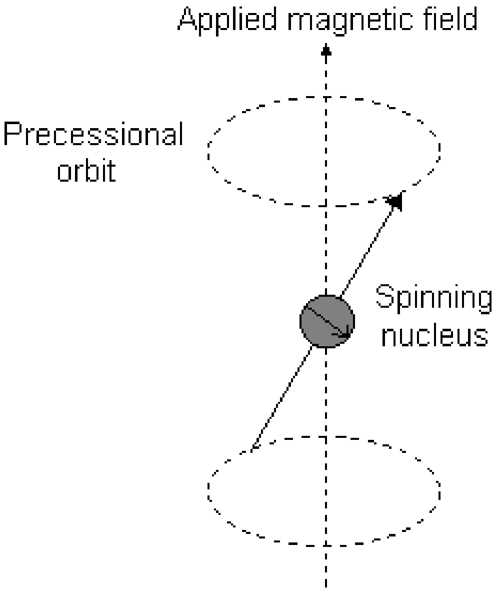


Figure 4.1: **NMR: precession of the nuclear magnetic moment**.

How do nuclei in the higher energy state return to the lower state? Emission of radiation at radio frequencies is negligible because the probability of re-emission of photons varies with the cube of the frequency. The main process in this case is based on thermodynamics and is called population relaxation.

Two major relaxation processes can be distinguished:

1. Spin - lattice or longitudinal magnetic relaxation;
2. Spin - spin or transverse relaxation

The term lattice refers to the biological sample (to analyze) in which the nuclei are held. Nuclei in the lattice are in vibrational and rotational motion, which creates a complex magnetic field. This field (called lattice field) has many components, some of which will be equal in frequency and phase to the Larmor frequency of the nuclei of interest. These components of the lattice field can interact with nuclei in the higher energy state, and cause them to lose energy (returning to the lower state). The energy that a nucleus loses increases the amount of vibration and rotation within the lattice resulting in a tiny rise in the temperature of the sample.

The relaxation time, T_1[2] depends on the gyromagnetic ratio of the nucleus and the mobility of the lattice. As mobility increases, the vibrational and rotational frequencies increase, making it more likely for a component of the lattice field to be able to interact with excited nuclei. However, at extremely high mobilities, the probability of a component of the lattice field being able to interact with excited nuclei decreases.

Spin - spin relaxation describes the interaction between neighbouring nuclei with identical precessional frequencies but differing magnetic quantum states. In this situation, the nuclei can exchange quantum states; a nucleus in the lower energy level will be excited, while the excited nucleus relaxes to the lower energy state. There is no net change in the populations of the energy states, but the average lifetime of a nucleus in the excited state will decrease. This can result in line-broadening.

The magnetic field at the nucleus is not equal to the applied magnetic field because electrons around the nucleus shield it from the applied field. Electrons, similar to the nucleus, are also charged and rotate with a spin to produce a magnetic field opposite to the magnetic

[2] The average lifetime of nuclei in the higher energy state

field produced by the nucleus. In general, this electronic shielding reduces the magnetic field at the nucleus (which is what determines the NMR frequency). The difference between the applied magnetic field and the field at the nucleus is termed the *nuclear shielding.* As a result the energy gap is reduced, and the frequency required to achieve resonance is also reduced. This shift in the NMR frequency due to the electronic molecular orbital coupling to the external magnetic field is called chemical shift. NMR would not be very valuable if all protons absorbed at the same frequency. You would see only a signal that indicates the presence of hydrogens in your sample. What makes it useful is that different protons usually appear at different chemical shifts. Chemical shift is a function of the nucleus and its environment. It is measured relative to a reference compound. For ^{1}H, the reference is usually tetramethylsilane.

NMR signal detection

In NMR-spectroscopy, as we have seen, samples of liquid or solid material are exposed to an external static and homogeneous magnetic field referred to as B_0. The direction of this field is usually defined along z. The magnetic moments in the sample align along B0 according to a Boltzmann distribution [57]. Quantum mechanics shows that magnetic moments due

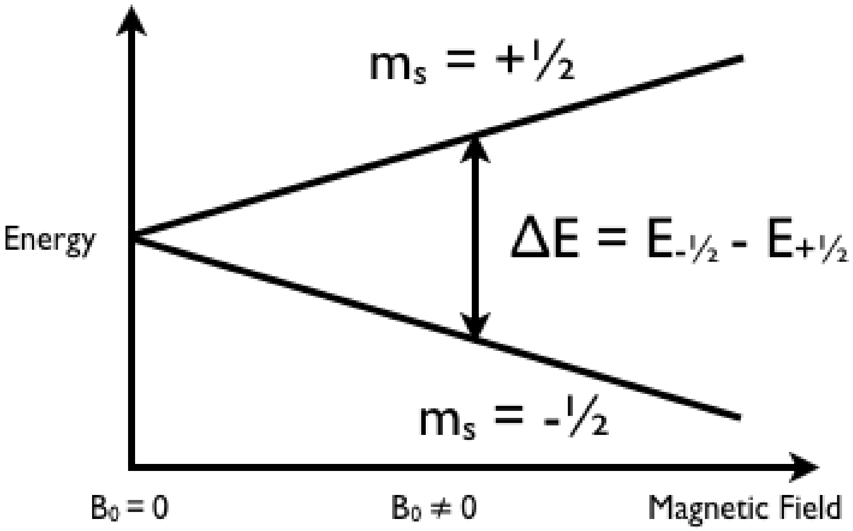


Figure 4.2: **NMR - Splitting of nuclei spin states in an external magnetic field**.

to spin -1/2 particles can only align parallel or anti-parallel (called down) with respect to this external field, these two states have a difference of energy given by equation 4.1. The magnetic moment of any macroscopic sample can be described like a classical macroscopic magnetic moment $\vec{M}$.

When the sample is in a magnetic field the lower energy level will contain slightly more nuclei than the higher level, therefore the magnetization of the sample is aligned with $\vec{B}_0$. In this configuration we call the magnetization $\vec{M}_0$. For the detection of the NMR signal, $\vec{M}_0$ is flipped orthogonal to B_0 by use of a high-frequency magnetic field B_1 orthogonal to z, applied for a defined time period (B_1 - pulse) [57]. $\vec{M}_0$ now aligned along the x direction will precess with a (resonance) frequency given by:

$$f_0 = \frac{\gamma B_0}{2\pi}$$

The sample to be analysed is positioned inside a detection coil. After short application of a high-frequency B_1 field, the precessing magnetization will induce a voltage U_i modulated with f_0. The amplitude of this voltage is directly proportional to $\vec{M}$ and thus with the number of spins rotating with f_0 located inside the observe volume of the apparatus. The signal detected is called a Free Induction Decay (FID).

4.3 Mass Spectrometry

Mass spectrometry (MS) is an analytical chemistry tool that helps identify the amount and type of chemicals present in a sample by measuring the mass-to-charge ratio and abundance of gas-phase ions [84].

In MS experiments, a sample, which may be solid, liquid, or gas, is ionized, for example by bombarding it with electrons. This may cause some of the sample's molecules to break into charged fragments. These ions are then separated according to their mass-to-charge ratio, typically by accelerating them and subjecting them to an electric or magnetic field: ions of the same mass-to-charge ratio will undergo the same amount of deflection [84]. The ions are detected by a mechanism capable of detecting charged particles, such as an electron multiplier. Results are displayed as spectra of the relative abundance of detected ions as a function of the mass-to-charge ratio (*m/z*).

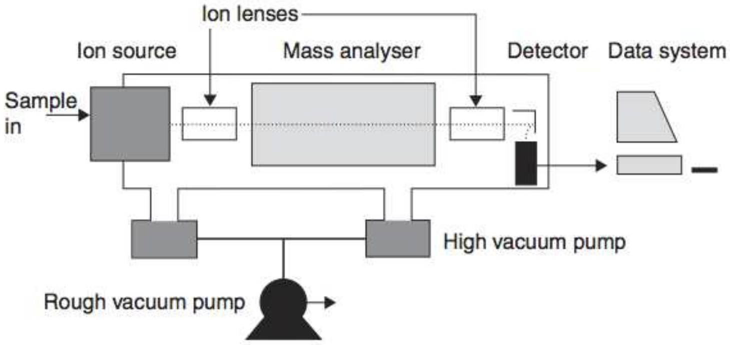


Figure 4.3: **Simplistic view of a mass spectrometer** (from [99]).

Mass spectra can be used to determine the elemental composition of a sample, the masses of particles and of molecules, and to elucidate the chemical structures of molecules, such as peptides and other chemical compounds.

Like NMR, mass spectrometry is widely used in metabolic fingerprinting and metabolite identification as well as being an important technique in the pharmaceutical industry for identification and quantitation of drug metabolites [57].

4.3.1 Principles

The mass spectrometer is the instrument that allows to performs all the required processes for mass spectrometric analysis starting from a sample in either a gas or a liquid phase: ionisation/transfer of sample to the gas phase and transfer to vacuum, separation according to mass-to-charge ratio (m/z), detection of ions and processing, and presenting the data in a usable format [99].

Simplistically, a mass spectrometer (Figure 4.3) consists of:

- an ion source;
- a mass analyser;
- a detector;
- a data system.

Ion Source. Sample molecules are introduced into the ion source, where they become ionised. Usually, before the ion source there is separating inlet device in which complex mixtures can be separated prior to admission to the mass spectrometer.

The inlet device is normally either a capillary gas chromatography (GC) column or a high-performance liquid chromatography (HPLC) column, although capillary electrophoresis and thin-layer chromatography can be interfaced with mass spectrometry [41].

The main processes in the ion source are: (i) transfer of the sample to the gas phase, (ii) ionisation, and (iii) transfer to vacuum. The order of these processes can vary depending on the sample type (gas or liquid) and ionisation method [99]

For metabolite analysis a number of different types of ionisation methods are available. The most common techniques are electron impact ionisation (EI) used with gas chromatography, and electrospray ionisation (ESI) used either with direct sample infusion or combined with liquid chromatography.

Other relevant ionisation techniques used are atmospheric pressure chemical ion- isation (APCI), atmospheric pressure photoionisation (APPI), desorption electrospray ionisation (DESI), liquid secondary ion mass spectrometry (LSIMS), matrix-assisted laser desorption/ionisation (MALDI) and fast atom bombardment (FAB) [41].

Mass Analyser. The ions coming out from the ion source are in the gas phase. They are separated according to their mass-to-charge ratio (m/z) in the mass analyser. The mass-to-charge ratio is evaluated with a combination of electric and/or magnetic fields. It is important that the ions produced in the ionisation chamber have a free run and do not collide with uncharged molecules or with each other. For this reason the mass analysers (and often also the ion source) are mantained in high vacuum.

There are several types of mass analyzers currently available, the better known of which include quadrupoles , time-of-flight (TOF) analysers, magnetic sectors , and both Fourier transform and quadrupole ion traps [13].

These mass analysers have different features:

- the m/z range covered;
- the achievable resolution;
- the mass accuracy;

- the compatibility with different ionisation methods.

It is worth to mention the tandem (MS-MS) mass spectrometers that have more than one analyser and so can be used for structural and sequencing studies.

Detector. The detector measures the amount of ions or their number as a function of time. It monitors the ion current, amplifies it and the signal is then transmitted to the data system.

Data System The data system should be considered as the fourth leg of the mass spectrometer and it is as important as the other parts [99]. The signal generated by the detector is recorded in the form of mass spectra in the data system. The m/z values of the ions are plotted against their intensities to show the number of components in the sample, the molecular mass of each component, and the relative abundance of the various components in the sample [99, 13].

4.3.2 Basics of Chromatography

Chromatography is a very efficient technique for the separation of compounds or mixtures. It involves a sample being dissolved in a mobile phase which may be a gas, a liquid or a supercritical fluid. The mobile phase is then forced through an immobile, immiscible stationary phase. The phases are chosen such that components of the sample have differing solubilities in each phase. A component which is quite soluble in the stationary phase will take longer to travel through it than a component which is not very soluble in the stationary phase but very soluble in the mobile phase. As a result of these differences in mobilities, sample components will become separated from each other as they travel through the stationary phase [12].

All chromatographic techniques utilize small differences in distribution coefficient to separate compounds in a two-phase system. HPLC. (High Performance Liquid Chromatography) and GC (Gas Chromatography) use columns - narrow tubes packed with stationary phase, through which the mobile phase is forced. The sample is transported through the column by continuous addition of mobile phase. This process is called elution. The average rate at which an analyte moves through the column is determined by the time it spends in the mobile phase [12, 99].

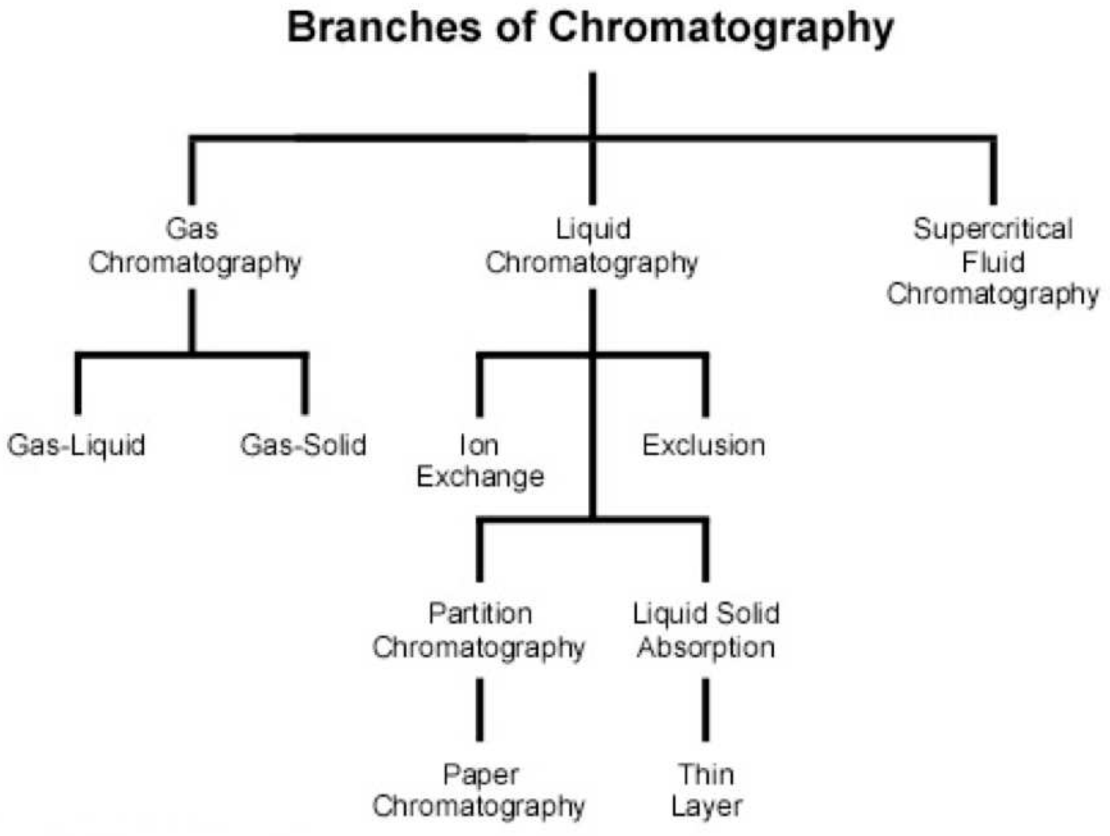


Figure 4.4: **Categories of Chromatography** (from [38]).

Chromatography developed dramatically between the 1960s and the 1990s mostly because of the improvements of columns, detectors, and electronics.

Metabolomics, where many small metabolites have to be separated, is almost always based on high-performance chromatographic separation with either a gas or a liquid as the mobile phase. [99]

4.3.3 GC-MS

The use of a mass spectrometer as the detector in gas chromatography was developed during the 1950s. Originally the use of these devices was limited. In the 1990s rapid developments in both the engineering of GC-MS systems and in the power and of computing systems has helped in the simplification of the use of this instrument, as well as allowed great improvements in the amount of time it takes to analyze a sample. These improvements enabled biological laboratories to perform GC-MS analysis on a routine basis [95].

As the name suggests, the GC-MS is composed of two major building blocks: the gas chromatograph and the mass spectrometer. The latter utilizes a capillary column providing an efficient and high resolution separation method. The different chemical properties of the molecules in a mixture and their relative affinity for the stationary phase of the column allow the separation of the molecules as the sample travels the length of the column. The molecules are kept into the column and then come off at different times, going towards the mass spectrometer downstream.

The mass spectrometer breaks each molecule into ionised fragments and detects them using their mass-to-charge ratio.

In GC-MS there are essentially two kinds of ionisation: electron impact ionisation (EI) and chemical ionisation (CI) [57].

The EI technique involves the bombardment of gas-phase sample molecules (M) with high-energy electrons (e^-), usually of 70 eV energy. This process generates $[M]^+$ ions and thermal energy free electrons (e^-) [41]

$$M(g) + e^- \longrightarrow M^+(g) + 2e^- \tag{4.2}$$

The molecular ions $[M]^+$ often are unstable and split to generate more stable products:

$$M^+(g) \longrightarrow A^+(g) + B(g) \tag{4.3}$$

Before using the EI techinque, the sample to be ionised must be in the gas phase. This has led to the extensive development of derivatisation chemistry to allow the vaporisation of many small biomolecules without their decomposition [41].

Chemical Ionisation differs from Electron Ionisation in that analyte ionisation is achieved via proton attachment rather than electron ejection [41]. In chemical ionization a reagent gas, typically methane or ammonia is introduced into the Ion Source of the mass spectrometer. The reagent gas becomes ionised by EI and acts as a proton donor to the analyte:

$$CH_4(g) + e^- \longrightarrow CH_4^+(g) + 2e^- \tag{4.4}$$

$$CH_4^+(g) + CH_4(g) \longrightarrow CH_5^+(g) + CH_3(g) \tag{4.5}$$

$$CH_5^+(g) + M(g) \longrightarrow MH^+(g) + CH_4(g) \tag{4.6}$$

The resulting ion MH^+ is an even-electron protonated molecule, which is more stable than the equivalent odd-electron molecular ion M^+ generated by EI [41]. One of the main benefits of using chemical ionization is that a mass fragment closely corresponding to the molecular weight of the analyte of interest is produced

CI is a lower energy process than electron ionization. The lower energy yields less fragmentation, and usually a simpler spectrum.

4.3.4 LC-MS

Liquid chromatography-mass spectrometry (LC-MS) is an analytical chemistry technique that combines the physical separation features of liquid chromatography with the mass analysis capabilities of mass spectrometry. Its field of application is usually oriented towards the separation, detection and identification of chemicals of particular masses in complex mixtures. While the coupling of the separation technique and the spectrometer in GC-MS has proven to be relatively straightforward, the hyphenation of liquid chromatographic separations with mass spectrometers was technically more difficult [57]. One of the main hurdles to overcome for LC-MS-based techniques has been the incompatibility of the liquid eluent coming from the column and the vacuum of the mass spectrometer [99].

Initially direct liquid introduction of the solvent (at very low flow rates) into the EI source was tried, but even very powerful vacuum pumps performed rather poorly. Also the use of techniques based on separation of analytes from solvents did not succeed. The introduction of atmospheric ionization techniques in the mid-1980s, especially electrospray ionization (ESI) enabled a significant advance for LC-MS, which now has become one of the most important analytical techniques in biotechnology [99].

ESI works well with moderately polar molecules and is thus well suited to the analysis of many metabolites. As shown in Figure 4.5 liquid samples are pumped through a metal capillary maintained at 3 to 5 kV and nebulised at the tip of the capillary to form a fine spray of charged droplets [51]. The capillary is usually orthogonal to, or off-axis from, the entrance to the mass spectrometer in order to minimise contamination. The droplets are rapidly evaporated by the application of heat and dry nitrogen, and the residual electrical charge on the

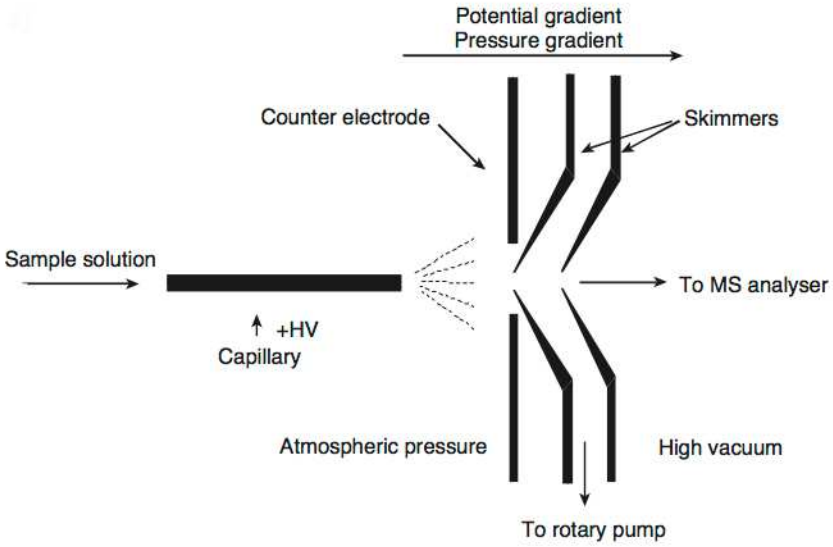


Figure 4.5: **Electrospray Ionisation** (from [41]).

droplets is transferred to the analytes. The ionised analytes are then transferred into the high vacuum of the mass spectrometer via a series of small apertures and focusing voltages [70].

The ion source and subsequent ion optics can be operated to detect positive or negative ions, and switching between these two modes within an analytical run can be performed.

Under normal conditions, ESI is considered a "soft" ionisation source, meaning that relatively little energy is imparted to the analyte, and hence little fragmentation occurs. This is in contrast to other MS ion sources, for example the electron impact source commonly used in GC-MS, which causes extensive fragmentation.

MS using ESI and other ionisation methods can be applied to a much wider range of biological molecules than GC-MS. LC-MS provides superior specificity and sensitivity compared to direct injection methods. Another advantage of LC-MS assays is the capacity to multiplex several analytes within a single analytical run with minimal incremental cost. This has the potential to simplify laboratory set up (e.g. creation of test panels) and provide additional useful information (e.g. metabolite profiles) [70].

Chapter 5

RFmarkerDetector: a tool for multivariate analysis of metabolomics data

5.1 Introduction

Over the last decades biomedical research has undergone profound changes. The search for single genes, transcripts, proteins, or metabolites has been replaced by the coverage of the entire genome, transcriptome, proteome, and metabolome [60].

Metabolomics, defined as the quantitative measurement of the multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification [58], is still a recent discipline compared to other "omics" fields, but its particular features and the improvement of both analytical techniques and pattern recognition methods has contributed greatly to its increasingly use.

Metabolites are the end product of all cellular processes and directly reflect all functional activities, transient effects, as well as endpoints of biological processes determined by the sum of genetic features, regulation of gene expression, protein abundance and environmental influences [87]. Thus, metabolites are more proximal to a phenotype or disease than either genetic or proteomic information [36]. This occurs because a simple change in the expression level of a gene or protein does not necessarily correlate directly with a variation in the activity level of a protein, but an alteration in a metabolite only occurs through such a change [60]. Consequently, for its non-invasive nature and its close link to the phenotype, metabolomics is an ideal tool for the pharmaceutical, preventive healthcare, and agricultural industries, among others.

Biomarker discovery is another example where metabolomics has already enabled informed decision making. In metabolomics, biomarkers are metabolites that can be used to distinguish two groups of samples, typically a disease and control group. For example, a metabolite reliably present in disease samples, but not in healthy individuals would be classified as a biomarker. Samples of urine, saliva, or cerebrospinal fluid (CSF) can contain highly informative metabolites, and can be readily analysed through metabolomics fingerprinting or profiling, for the purpose of biomarker discovery.

The analysis of metabolomic data has multiple issues and complications. In fact, every single organism, organ or tissue is essentially unique. Therefore each sample is characterized by an high degree of variability that makes it difficult to identify the few chemical features against the large and complex background of metabolites that uniquely define the system.

The identification of these features is further complicated by the fact that all biological systems are easily perturbed by various experimental or environmental factors, such as age, diet, growth phase, pH, sex.

It is also important to consider the unavoidable fluctuations in spectral data, such as changes in peak position or peak width that are caused by instrument instability and variability in sample conditions [60].

Herein we introduce RFmarkerDetector, a software package that provides a set of tools to carry out a complete multivariate analysis of metabolic profiles, exploiting the strengths of the Random Forest algorithm. RFmarkerDetector has been developed as an R package and it is distributed through the Comprehensive R Archive Network (CRAN) under the GNU General Public License (version 3).

5.2 Implementation

5.2.1 Datasets

We have seen in the previous chapter that the choice of the analytical method to use to carry out a metabolomics experiment is influenced by many factors. For instance, the main advantage of mass spectrometry (MS) is sensitivity. Coupling MS with gas chromatography (GC) or with liquid chromatography (LC) enables the measurement of hundreds of individual species within a single sample [97]. Conversely, two of the major weaknesses of MS in metabolomics are quantification and the time-consuming sample preparation techniques.

The major weaknesses of MS are the major strengths of NMR spectroscopy. In fact, high-resolution ^{1}H NMR requires limited sample preparation, is quantitative, non-destructive and may detect compounds that are too volatile for GC [83].

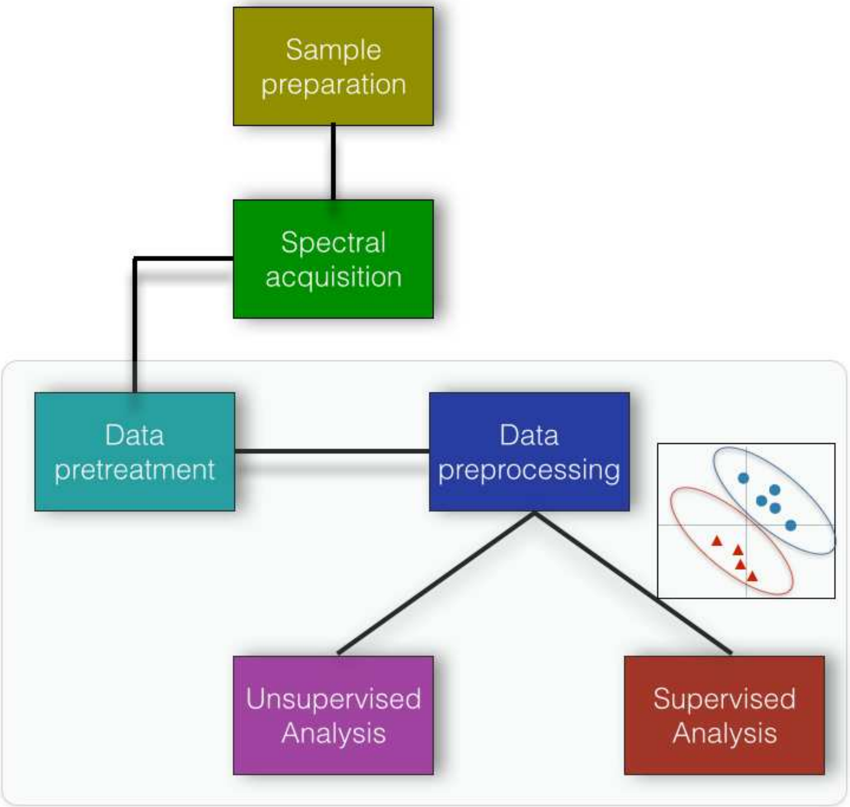


Figure 5.1: **Metabolomics data analysis pipeline**.

All these analytical platforms generate a huge amount of data, often characterized by a large number of variables. In Metabolomics, it is quite common to deal with data sets known as large p small n data, since for this type of data, the number of observations (or samples) n is much less than the number of variables p.

Conventional statistical techniques are mainly applied to situations in which the number of observations is of the same order of magnitude or exceeds the number of variables. Also traditional classification techniques like k-nearest neighbours, logistic regression, often fail on this kind of data, mainly due to the fact that the condition $p \ggg n$ leads to ill-posed problems and thereby the inability of those methods to even have a solution. Linear regression methods for instance, are infeasible on this kind of data, as the dataset is singular (i.e. no longer invertible) and no unique least-squares solution exists. Consequently, analysis of metabolomics data requires the use of multivariate analysis techniques capable of dealing with this kind of dataset often characterised by a relevant collinearity.

5.3 Data preprocessing

RFmarkerDetector provides a set of tools for the whole process of multivariate analysis of metabolomics data (Figure 5.1) including preprocessing, exploration, visualization, calibration and validation of models and identification of potentially relevant biomarkers.

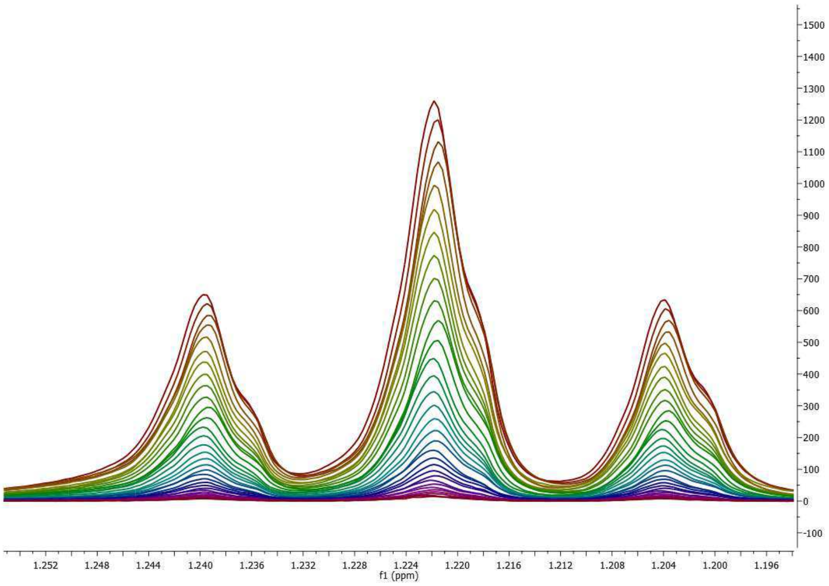


Figure 5.2: **Chemical shift variability across spectra**.

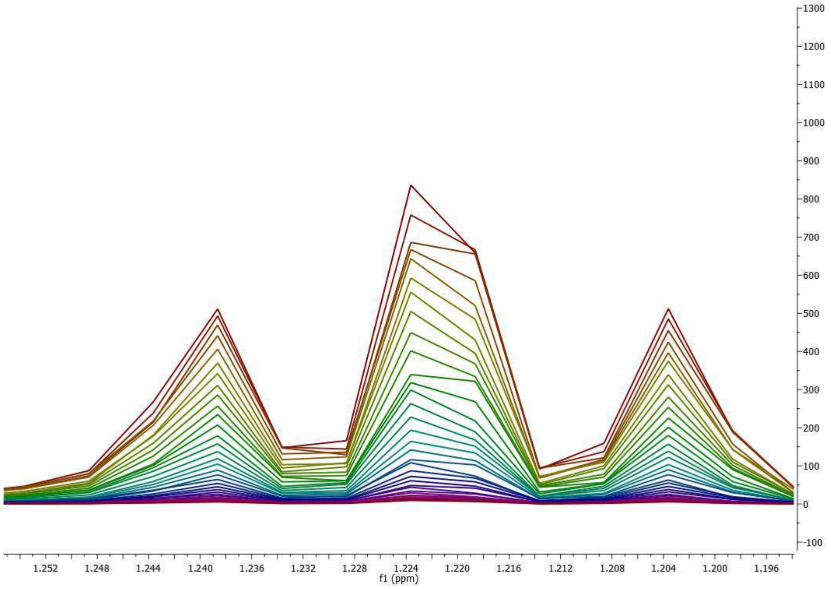


Figure 5.3: **Chemical shift variability mitigated through 'binning'**.

Data typically come from NMR experiments and can be of two types:

- concentration matrices;
- binned NMR spectra data.

Concentration matrices represent the concentrations of the metabolites in a biofluid analyzed with NMR spectroscopy, while binned data are the result of the bucketing or binning methodology applied to NMR spectra. NMR spectra of biological samples are usually poorly aligned due to wide changes in chemical shift arising from temperature, pH, ionic strength, and other factors.

The binning procedure is the most widely used method of addressing this chemical shift variability across spectra and consists in segmenting a spectrum into small areas (called buckets or bins) and taking the area under the spectrum for each segment [106].

The size of the bins should be large enough so that a given peak remains in its bin despite small spectral shifts across the spectra, but not so large as to include peaks belonging to multiple compounds within a single bin.

Moreover NMR spectra contains thousands of variables. Binning can also be used to reduce the data dimensionality [83].

The main goal of data preprocessing is to transform the data in order to ease and improve the data analysis. To this aim, RFmarkerDetector includes several pretreatment methods that can be divided into three categories: (i) scaling, (ii) centering and (iii) filtering.

Scaling methods are data pretreatment approaches that divide each feature (variable) of the dataset by a scaling factor, different for each of them. Depending on the type of scaling factor used, we can distinguish two subclasses within scaling. The first class exploits a measure of data dispersion (such as standard deviation), while the second uses a size measure.

RFmarkerDetector scaling methods focuses on the approaches belonging to the first class. The function *autscale()* uses the standard deviation of each feature as the scaling factor (for this reason this method is also called unit variance scaling). After autoscaling, all metabolites have a standard deviation of one. In this case data are compared on the basis of correlations.

The method *paretoscale()* is very similar to autoscaling. The scaling factor in this case is represented by the square root of the standard deviation of each feature. Pareto scaling is intermediate between centering and autoscaling and partially preserves data structure.

meanCenter() adjusts for differences between high-concentrated and low-concentrated metabolites by converting all values to vary around zero instead of around the mean of each variable. Centering is often combined with scaling methods.

The filtering methods *rsdFilter()* and *lqvarFilter()* allow to remove irrelevant features from the dataset: the former removes the predictor variables with a relative standard deviation less than a user-defined threshold, while the second eliminates those variables with a percentage of zero-values above a tunable limit.

Table 5.1 summarizes the preprocessing methods included in the software package.

RFmarkerDetector Method	Expression	Goal	Pro	Cons
autoscale	$x_{ij} = \frac{x_{ij} - \bar{x_i}}{\sigma_i}$	Compare variables based on correlations	All variables equally important	Inflate baseline noise
paretoscale	$x_{ij} = \frac{x_{ij} - \bar{x_j}}{\sqrt{\sigma_j}}$	Preserve data structure, reducing the relative importance of large features	Increase the weight of medium features without inflating baseline noise	Sensitive to large fold changes
meanCenter	$\bar{x_{ij}} = x_{ij} - \bar{x_j}$	Emphasize the differences in the data	Remove the offset from data	Big features dominate
lqvarFilter	-	Filters variables with a relevant percentage of zero-values	Can remove irrelevant variables from the dataset	Discriminative features can be removed
rdsFilter	$rsd = \frac{\sigma_j}{\bar{x_j}}$	Remove variables with a relative standard deviation less than or equal a defined threshold	Exclude near constant variables	Raw filtering approach

Table 5.1: RFmarkerDetector preprocessing methods

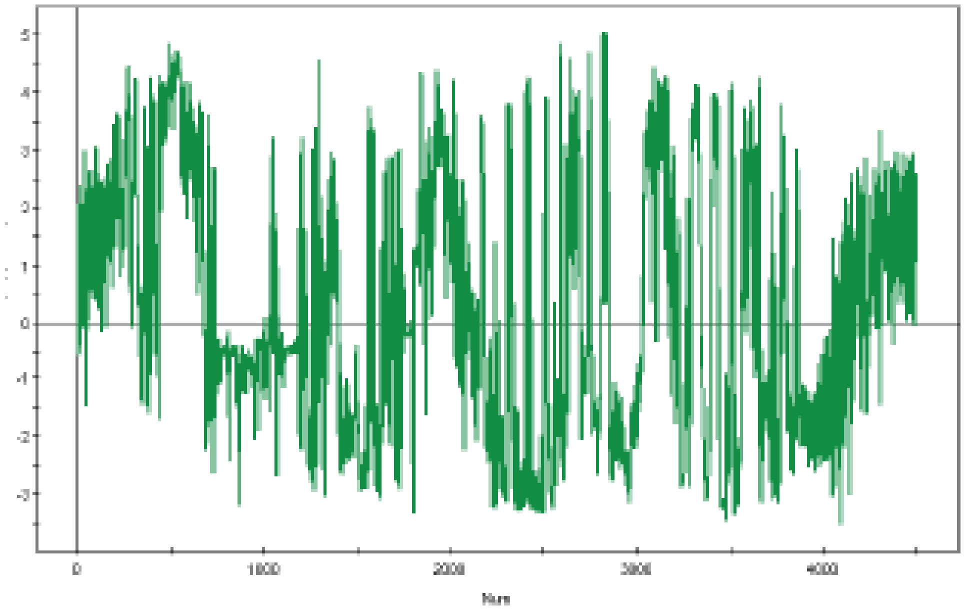


Figure 5.4: **Scaling effect on NMR spectra: Autoscaling**.

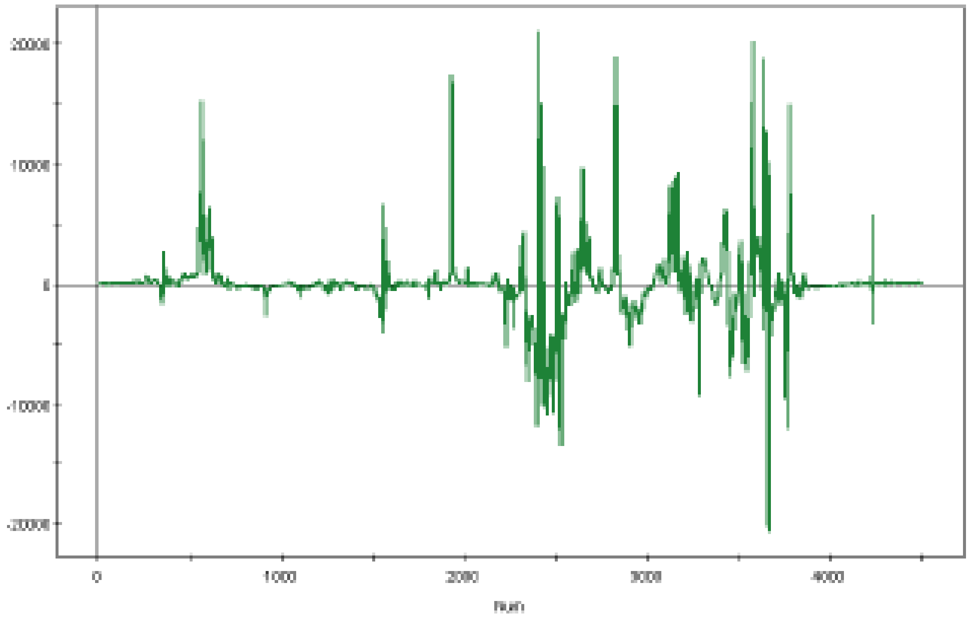


Figure 5.5: **Scaling effect on NMR spectra: Pareto scaling**.

5.4 Exploratory Data Analysis

RFmarkerDetector includes a set of functions to perform exploratory data analysis, a crucial step that helps to analyze data sets, summarizing the main traits with the use of visual methods.

Exploratory Data Analysis (EDA) is an approach based on a variety of techniques mainly aimed to:

- identify significant variables;
- reveal underlying structures;
- detect outliers;
- possibly develop parsimonious models.

One of the techniques we employ is Principal Component Analysis (PCA), an unsupervised approach that can assist in the identification of patterns in high dimensional data sets and can help to express the data in such a way as to highlight their similarities and differences.

PCA is a well known method that uses an orthogonal transformation to convert a set of samples of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables.

One of the advantages of PCA is that you can compress the data by reducing the number of dimensions, without much loss of information. PCA provides a direct mapping of the (often high-dimensional) original dataset into a lower-dimensional space containing most of the original information.

The coordinates of the samples in the new space are called *scores*, while the dimensions of the new space are linear combinations of the original variables called *loadings*.

Ultimately we can say that PCA "simplifies" data breaking down a large matrix in two smaller ones.

The function *pca()* of the RFmarkerDetector package returns the scores and the loading matrices and also the variances associated with the principal components. These matrices can be used both to obtain the scores plot (*plot.pca.scores()*) (Figure 5.6), a projection of data onto a low dimensional subspace that helps to identify possible relationships between observations, and the loading plot (*plot.pca.loadings()*) (Figure 5.7) that provides information regarding the variables (metabolites), which can explain the relationships revealed in the scores plot.

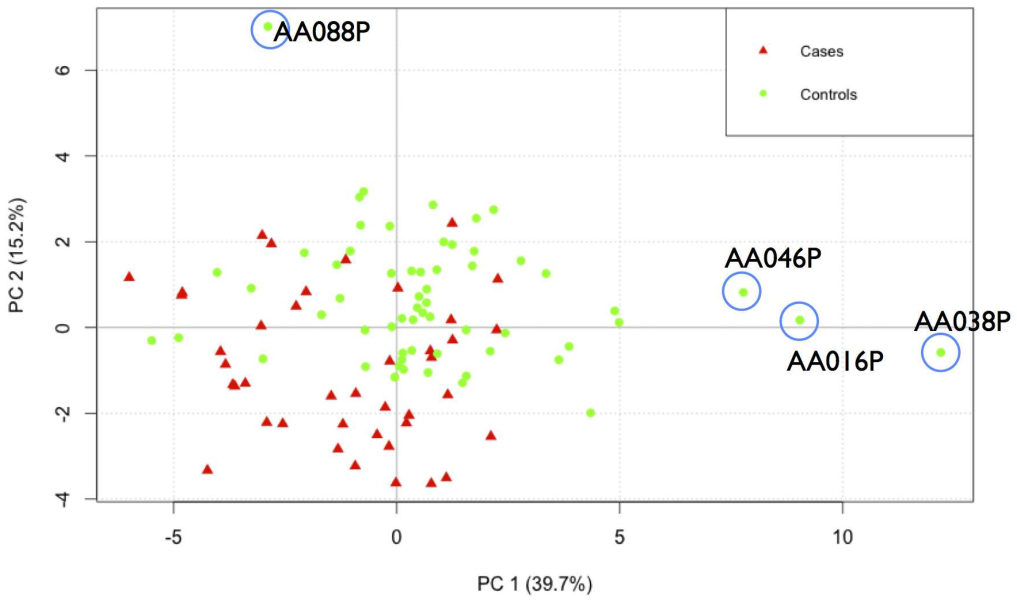


Figure 5.6: **PCA scores on PCs 1 and 2**.

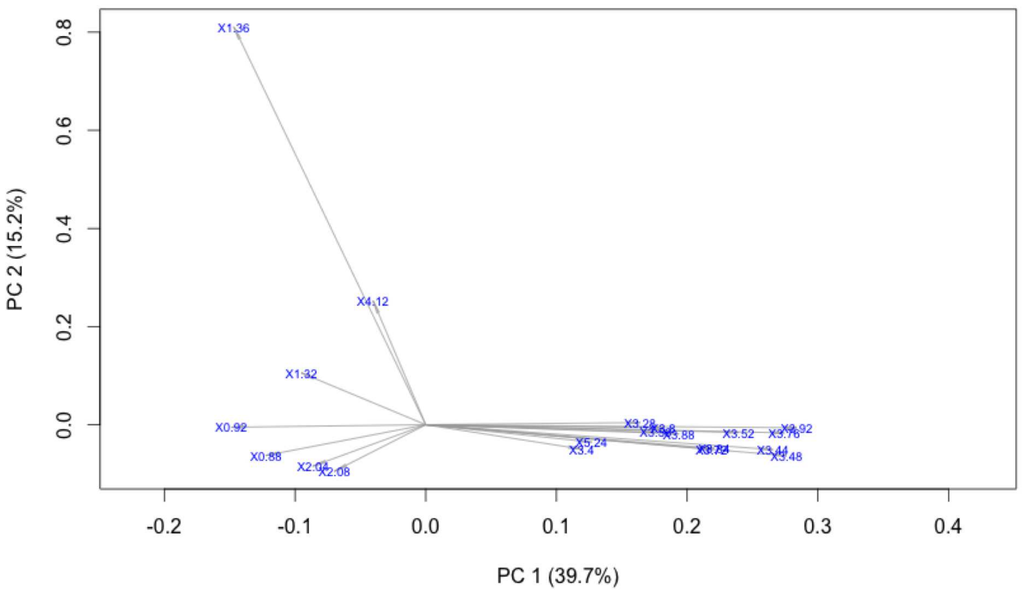


Figure 5.7: **PCA loadings on PCs 1 and 2**.

The *screeplot()* returns a graphical display of the variance explained by each principal component that is useful for determining the appropriate number of principal components to be used (Figure 5.8).

Another exploratory data analysis technique available in RFmarkerDetector is the Multidimensional Scaling (MDS) based on the Random Forest proximity matrix (Figure 5.10).

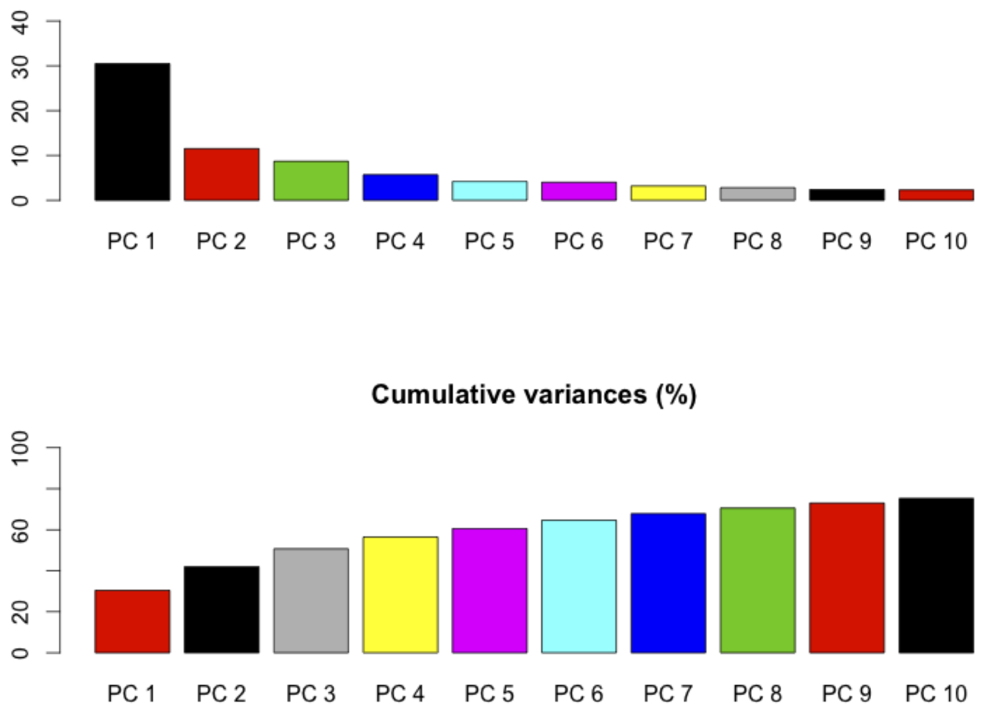


Figure 5.8: **Scree plot showing the fraction of total variance in the data as explained by each PC**.

As already said, RFmarkerDetector is based on the Random Forest algorithm [20], an ensemble of weak learners used for solving both classification and regression problems.

Given a set of training data $\chi_t = \{(\mathbf{x_m}, y_m), m = 1,2,...,M\}$ where $\mathbf{x_m}$ is an M-dimensional input vector and y_m is the predictor output, we can define a weak learner as a predictor $f(\mathbf{x}, \chi_t)$ with low bias and high variance [21, 98].

A collection of weak learners $f(\mathbf{x}, \chi_t, \theta_k)$ can be obtained by randomly sampling from the set χ_t. The random vector θ_k selects the data points for the kth weak learner $f(\mathbf{x}, \chi_t, \theta_k)$. By applying bootstrap sampling to generate θ_k, for example, almost two-thirds of the observations are used by each weak learner, while the remaining are out of the bootstrap sample or out-of-bag (OOB) [98]. Bootstrap samples are independent and identically distributed.

It can be shown that creating a committee by combining independent and identical distributed weak learners, you can keep the bias approximately unchanged and reduce the variance by a factor equal to the mean value of the correlation between each weak learner [20].

In other words we can say that a random forest is a collection of decision trees (used as weak learners) created following an efficient strategy aimed at increasing the diversity between the trees. Each weak learner is an unpruned classification or regression tree, created

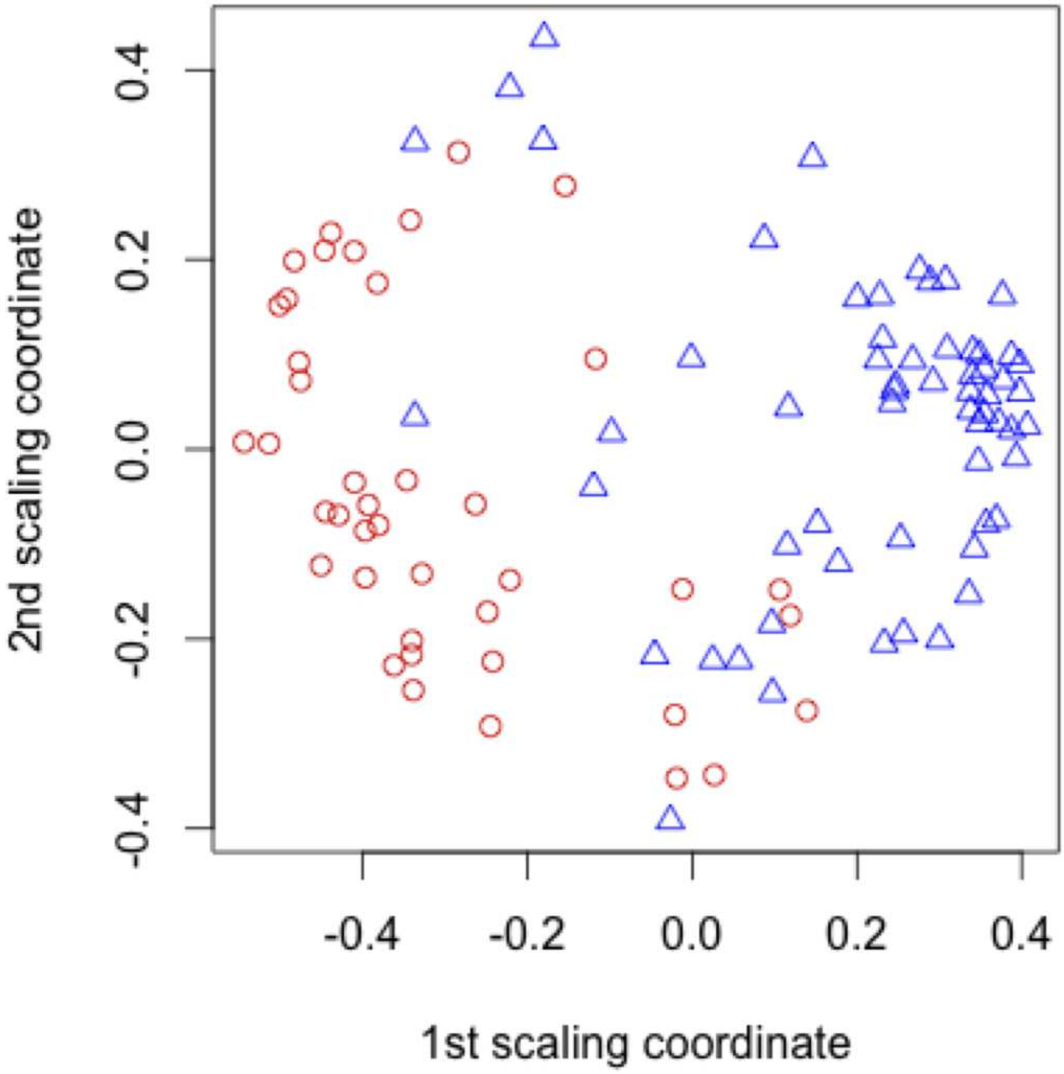


Figure 5.9: **MDS plot using Random Forests proximities**.

by using bootstrap samples of the training data and random feature selection.

To get low bias, trees are grown to maximum depth (unpruned), while to achieve low correlation, randomization is applied at different levels:

- Each tree of the forest is grown on a bootstrap sample drawn from the training set;

- For each tree "node", n variables are randomly selected from the set of all variables and evaluated for their ability to split the data. Only the variable that provides the best split is used out of the n selected.

Outputs of all trees are then aggregated to produce one final prediction $\hat{Y}$. For classification problems, $\hat{Y}$ is the class predicted by the majority of trees, while in regression it is the average of the individual tree predictions.

Random Forest allows to calculate the proximity between samples which assumes high values if samples are similar. These values can be used to perform a Multi Dimensional Scaling (MDS) that is a means of visualizing the level of similarity of individual samples of a dataset.

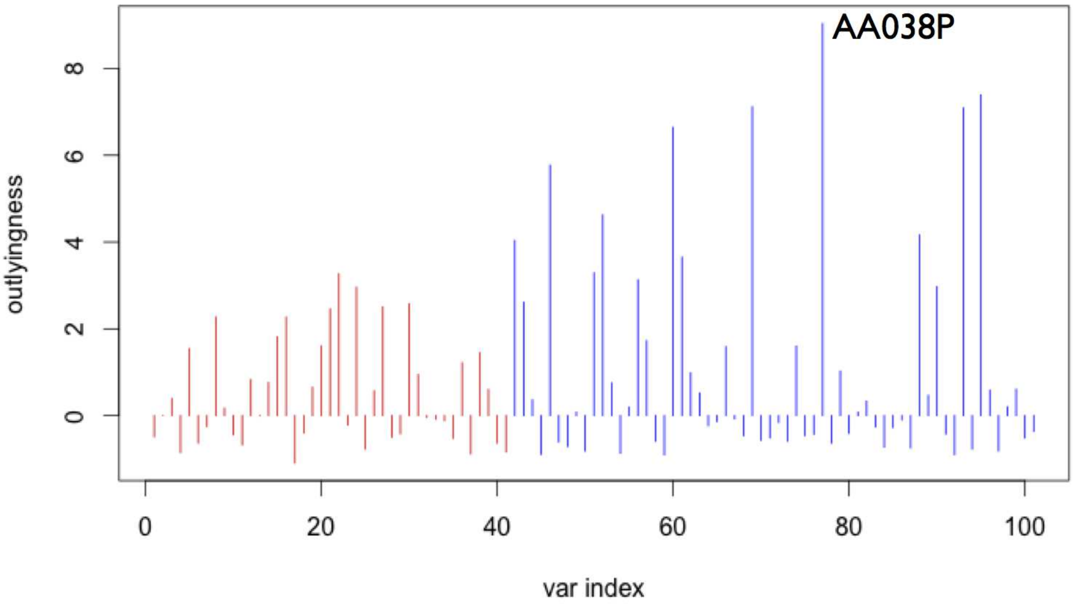


Figure 5.10: **Proximity measures and outliers**.

The proximity or similarity between any two samples in a dataset is calculated as the number of times the two samples end up in the same terminal node of a tree divided by the number of trees in the forest. The Random Forest algorithm calculates the proximities between samples and then arranges them in a matrix termed proximity matrix.

The method *plot.mds()* of the RFmakerDetector package, employs these proximity scores to build two-dimensional MDS plots that provide a means for visualizing the similarity between samples, represented as the distances between data points.

MDS plots can also be used for the identification of outliers or mislabelled samples that can be recognized as those samples whose proximity to all other samples of the same class is small.

5.5 Tuning Model Parameters

The number of input variables randomly chosen at each split (often referred to as *mtry*) and the number of trees (*ntree*) in the forest, are the two main parameters for the random forest algorithm.

The RFmarkerDetector package provides two methods to tune and optimize these parameters:

- *tuneMTRY()*;
- *tuneNTREE()*

tuneMTRY() attempts to identify the optimal mtry, testing a user-defined sequence of values. For each mtry value, the function builds several Random Forest models (the number can be selected by the user) providing their performances in terms of out-of-bag (OOB) errors and arranging them in a matrix.

In the original implementation of the random forest algorithm, each tree is trained on about 2/3 of the total training data. As the forest is built, each tree can thus be tested on the samples not used in building that tree. This is the OOB error estimate, an internal error estimate of a random forest as it is being constructed.

Comparing OOB performance estimation and k-fold cross-validation, has shown that they are in good agreement [88]. Thus, being the OOB error the default output of the random forest algorithm, we decided to use this estimate to compare the models within the *tuneMTRY()* method.

The matrix returned by this method can be passed as argument to the function *plotOOBvsMTRY()* for visualizing the trend of the average OOB error as a function of mtry (see Figure 5.11). *tuneNTREE()* follows a similar approach to optimize the number of trees in the forest.

5.6 Potential Biomarker Identification

The identification of potential biomarkers from matrices of metabolic profiles involves a double cross validation scheme: one to optimize the model complexity given for each candidate subset of variables, and the other to assess the final model performance.

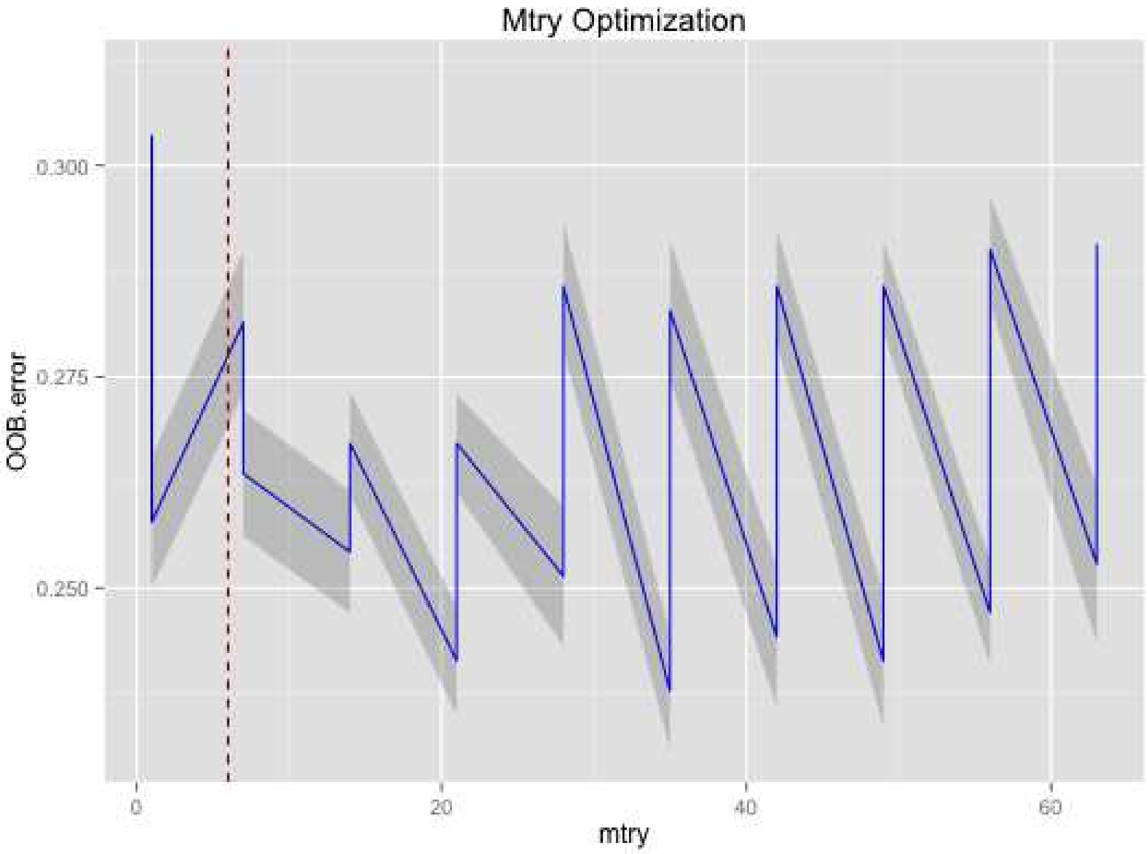


Figure 5.11: **The average OOB error expressed as a function of the mtry parameter. The shaded area represents the 95% confidence interval**.

RFmarkerDetector includes the method *rfMCCV()* that implements this scheme consisting of two steps:

- **step 1**: the whole dataset is randomly split into training and test sets;
- **step 2**: the biomarker selection is iteratively performed using only the training set. Each candidate variable-subset random forest model is evaluated using the out-of-bag error estimate (this is the inner cross validation loop);
- **step 3**: the best parsimonious random forest model is then validated on the test set. The performance of this model is evaluated on the basis of he ROC curve analysis of the test data;
- **step 4**: the first three steps are then repeated *N* times such that *N* parsimonious model evaluations can be performed.

Examining the probability of selection of the features selected across the N parsimonious models, it will be possible to determine whether a consistent group of metabolites has been found.

```r
library("WilcoxCV")
library("randomForest")
rfmccv <- function(data, nsplits, test_prop, opt_params, nvar, ranking = "MDA") {
    NTREE = 1000
    MTRY = floor(sqrt(ncol(data) - 2))
    if ((nvar < 2 | nvar > ncol(data) - 2)) {
        nvar = ncol(data) - 2
        stop("argument nvar out of range.")
    }
    if (hasArg(opt_params)) {
        if (!is.null(opt_params[["ntree"]])) {
            NTREE = opt_params$ntree
        }
        if (!is.null(opt_params[["mtry"]])) {
            MTRY = opt_params$mtry
        }
        if (!is.null(opt_params[["ref_level"]])) {
            ref_level = opt_params$ref_level
            labels <- levels(data[, 2])
            if (!(ref_level %in% labels))
                stop("A problem occurred in opt_params: check ref_level parameter")
            data[, 2] <- relevel(data[, 2], ref = ref_level)
        }
    }
    levels(data[, 2]) <- c(0, 1)
    ntest <- floor(test_prop * nrow(data))
    set.seed(1234)
    test.index.matrix <- generate.split(n = nrow(data), niter = nsplits, ntest = ntest)
    m <- matrix(nrow = ntest, ncol = nsplits)
    predictions <- data.frame(m)
    labels <- data.frame(m)
    models <- list()
    for (i in 1:nrow(test.index.matrix)) {
        indexes <- test.index.matrix[i, ]
        testset <- data[indexes, ]
        trainingset <- data[-indexes, ]
        trained_model <- randomForest(x = trainingset[, 3:ncol(trainingset)], y = trainingset[, 2], mtry = MTRY, ntree = NTREE,
            importance = T)
        if (ranking == "MDA") {
            tmp_var = importance(x = trained_model, type = 1)
        }
        else {
            tmp_var = importance(x = trained_model, type = 2)
        }
        top_var <- tmp_var[order(tmp_var, decreasing = T), , drop = F]
        top_var <- top_var[1:nvar, ]
        headers <- append(names(data)[1:2], names(top_var))
        trainingset <- trainingset[, headers]
        testset <- testset[, headers]
        model <- randomForest(x = trainingset[, 3:ncol(trainingset)], y = trainingset[, 2], xtest =
  testset[, 3:ncol(testset)],
            ytest = testset[, 2], mtry = MTRY, ntree = NTREE)
        predictions[, i] <- model$test$votes[, 2]
        labels[, i] <- testset[, 2]
        models[[i]] <- model
    }
    res <- RFmarkerDetector::mccv(predictions, labels, models)
}
```

Listing 5.1: Source code of the function rfMCCV.

For the identification of potential biomarkers, RFmarkerDetector provides an implementation of the AUC-RF algorithm [22] based on optimizing the area under the ROC curve (AUC) of a random forest model. The strategy performs an iterative backward elimination process based on the initial ranking of variables.

The functions *aucMCV()* and *plotVarFreq()* can help to detect the candidate biomarkers by providing a graphical representation of them.

5.7 Case Study

In the following section, I will illustrate the predictive and interpretational benefits of some of the outlined methodologies using a real data sets.

5.7.1 Epilepsy dataset

Epilepsy is a group of neurological disorders characterized by epileptic seizures [24]. Epileptic seizures are episodes that can vary from brief and nearly undetectable to long periods of vigorous shaking.

The dataset investigated has been obtained selecting patients affected by partial epilepsy pharmacologically well controlled or pharmaco-resistant, enrolled in the Epilepsy Diagnostic and Treatment Centre of Cagliari. Blood samples were collected from 2 groups of fasted patients : (i) 35 affected by epilepsy and (ii) 35 healthy subjects.

The epileptic patients included two subgroups: 18 patients classified as responders (R) and the remaining 17 as nonresponders (NR) according to their response to therapy (Table 5.2).

Classes	Age (mean + SD)/ [range]	Gender (F/M)
Controls (n = 35)	(44.38 + 17.19)/ [6-76]	24/11
Responders (n = 18)	(47.5 + 16.86)/[27-80]	11/7
Non Responders (n = 17)	(52.17 + 9.57)/[41-71]	11/6

Table 5.2: Characteristics of the classes enrolled in the study

5.7.2 Sample Preparation

Plasma samples were thawed on ice and then centrifuged at 2500 g for 10 min at 4°C. An aliquot of 800 µl of plasma was used and a solution of chloroform/methanol 1:1 (2400 µl) plus 350 µl of distilled water was added. Samples were stirred for 1 minute and centrifuged for 30 min, at 1700 g at RT. After the centrifugation, hydrophilic and hydrophobic phases were obtained. The first was concentrated overnight using a speed vacuum instrument and then re-suspended in 630 µl of D2O and 70 µl TSP (Trimethylsilyl propanoic acid) 5.07 mM.

TSP was added to provide an internal reference for the chemical shifts of the spectrum obtained with the NMR analysis.

NMR experiments were performed on a Varian UNITY INOVA 500 spectrometer operating at 499 MHz equipped with a 5mm triple resonance probe with z-axis pulsed field gradients and autosampler.

One-dimensional 1H-NMR spectra were collected at 300 K with a noesy pulse sequence to suppress the residual water signal by using 0.100 ms of mixing time.

Spectra were manually phased, baseline corrected and chemical shifts referred to the internal standard TSP (at δ = 0.0 ppm) using MestReNova software [6]. Metabolite identification was carried out by using the library of metabolite NMR spectra from the Chenomx NMR Suite (version 7.1) [3]. The Chenomx NMR Suite software allows to fit the spectral signatures (singlets, doublets, triplets etc) of a compound from an internal database of reference spectra to the experimental NMR spectrum.

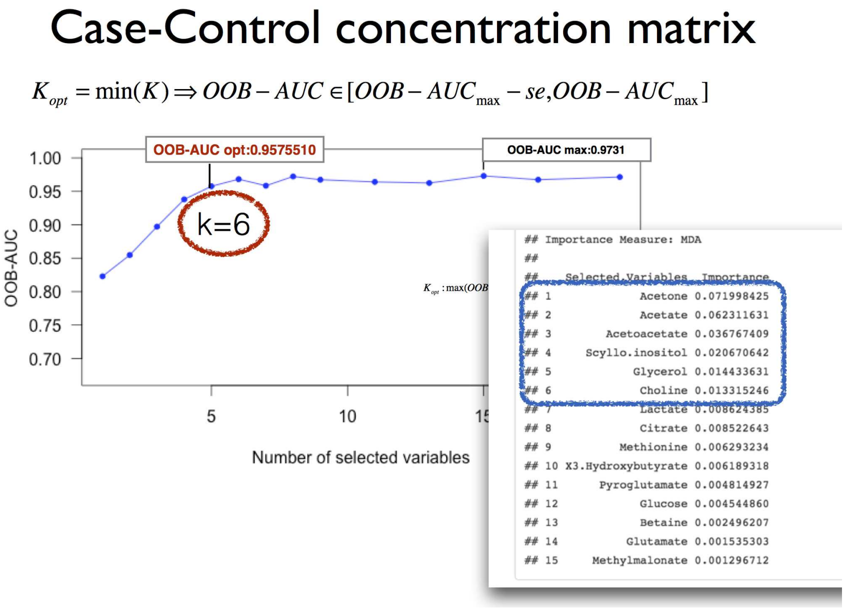
Also for the determination of the concentrations of individual metabolites (quantification) Chenomx NMR Suite has been used. The matrix obtained had 70 rows and 20 columns. Each row represented a sample, while each column the concentration of one of the metabolites identified.

The samples have been divided into two groups: 35 labeled as healthy control individuals and the remaining 35 as epileptic patients.

5.7.3 Data Analysis Strategy

The goal of the analysis was to build a classification model able to correctly discriminate between healthy controls and patients and to attempt to identify a panel of potential biomarkers.

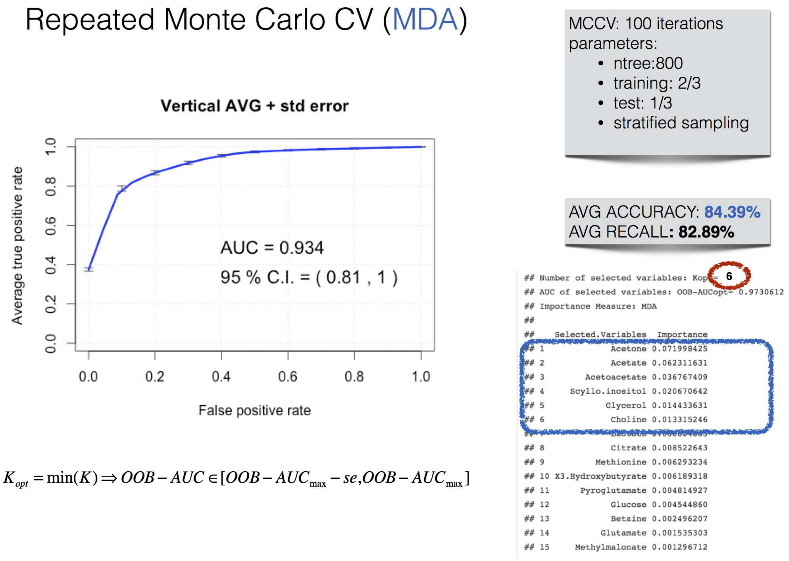
After a raw filtering step, the concentration matrix has been scaled using the Pareto approach, in order to emphasize all metabolite signals and reduce the noise.



```
## Importance Measure: MDA
##
##     Selected.Variables  Importance
## 1               Acetone 0.071998425
## 2               Acetate 0.062311631
## 3           Acetoacetate 0.036767409
## 4        Scyllo.inositol 0.020670642
## 5              Glycerol 0.014433631
## 6               Choline 0.013315246
## 7               Lactate 0.008624385
## 8               Citrate 0.008522643
## 9            Methionine 0.006293234
## 10 X3.Hydroxybutyrate 0.006189318
## 11         Pyroglutamate 0.004814927
## 12              Glucose 0.004544860
## 13              Betaine 0.002496207
## 14            Glutamate 0.001535303
## 15       Methylmalonate 0.001296712
```

Figure 5.12: **Identification of the optimal parsimonious model**.

The biomarker selection process has been performed using a repeated Monte Carlo cross validation (MCCV) scheme based on random forest models. In each MCCV, two thirds (2/3) of the samples have been used to evaluate the feature importance by using the importance measure of the random forest algorithm. The top important features have been used to build classification models validated on the 1/3 of the samples that were left out. The procedure have been repeated multiple times to calculate the performance and confidence interval of each model.



MCCV: 100 iterations
parameters:

- ntree:800
- training: 2/3
- test: 1/3
- stratified sampling

AVG ACCURACY: **84.39%**
AVG RECALL**: 82.89%**

```
## Number of selected variables: Kopt= 6
## AUC of selected variables: OOB-AUCopt= 0.9730612
## Importance Measure: MDA
##
##    Selected.Variables  Importance
## 1             Acetone 0.071998425
## 2             Acetate 0.062311631
## 3        Acetoacetate 0.036767409
## 4     Scyllo.inositol 0.020670642
## 5            Glycerol 0.014433631
## 6             Choline 0.013315246
## 7             Lactate [illegible]
## 8             Citrate 0.008522643
## 9          Methionine 0.006293234
## 10 X3.Hydroxybutyrate 0.006189318
## 11       Pyroglutamate 0.004814927
## 12             Glucose 0.004544860
## 13             Betaine 0.002496207
## 14           Glutamate 0.001535303
## 15      Methylmalonate 0.001296712
```

$$K_{opt} = \min(K) \Rightarrow OOB-AUC \in [OOB-AUC_{\max} - se, OOB-AUC_{\max}]$$

Figure 5.13: **Evaluation of the performance of the parsimonious model**.

Figure 5.12 illustrates the structure of the optimal parsimonious model and the candidate biomarker identified, while in Figure 5.13 is reported the the average ROC curve generated by Monte Carlo Cross-Validation.

Chapter 6

Concluding remarks

In this thesis we investigated the problem of metabolomics data management in conjunction with the application of machine learning techinques and chemometric analysis.

The focus has been put on multivariate strategies for metabolomics data analysis and on the validation of classification and prognostic models for the identification of potential biomarkers that could aid the diagnosis, monitoring and the prediction of a disease or the outcome of a therapy.

Metabolomics deals with the global assessment of the metabolites present in a biological system to evaluate the progress of a disease and provide insights into the underlying pathophysiology. It can be seen as a complementary tool to genomics and proteomics: in fact, while Genomics and Proteomics provide extensive information regarding the genotype but provide limited insights about phenotype, the metabolites are the closest link to the phenotype of the biological system studied.

The growing interest in metabolomics has been encouraged by rapid advances in metabolic profiling techniques and by technological developments of the diverse analytical platforms, including proton Nucleic Magnetic Resonance (1H NMR), Gas Chromatography-Mass Spectrometry (GC-MS) and Liquid Chromatography-Mass Spectrometry (LC-MS), used for generating metabolic profiles. The result is the production of a huge amount of data and information.

To efficiently handle the data generated and optimize the complex experimental processes needed to produce them, we designed and developed a software platform called QTREDS (Quality and TRacEability Data System).

QTREDS is a Laboratory Information Management System (LIMS), which is a software infrastructure to support the integrated management of multiple data types and the activities of "omics" laboratories.

The software application was designed to provide researchers with a complete knowledge of the laboratory processes at each step, in order to manage and verify the: (I) workflow creation, (II) samples traceability, (III) diverse experimental protocol definitions, (IV) inventory of reagents and (V) users' roles and privileges.

The software platform has been developed to address the specific needs of the Sequencing and Genotyping Laboratories of the CRS4 research center, where it has been tested and used to carry out almost one hundred DNA library preparation and sequencing experiments. Thanks to its flexibility QTREDS is currently undergoing an optimization process to adapt it to the requirements of metabolomics laboratories. Another topic I have investigated in this thesis, concerns the multivariate analysis of metabolomics data. The following aspects have been covered and discussed:

- data preprocessing and pretreatment;
- exploratory analysis;
- biomarker discovery and selection using the Random Forest algorithm.

The data used in our experiments were mainly ^{1}H - NMR spectra of blood plasma samples from epileptic patients, provided by the Department of Biomedical Sciences of the University of Cagliari. The first step of data preprocessing has been baseline removal, carried out by using a robust estimation procedure with the help of the researchers of the Department of Biomedical Sciences of the University of Cagliari. Baseline distortions in fact, can affect the quantification of metabolites and the consequent statistical analysis.

After baseline corrections, several other preprocessing methods has been applied (spectral regions suppression, alignment, binning) in order to eliminate spurious signals or reduce the chemical shift problem. Different normalization and filtering techniques (autoscaling, Pareto scaling) have been investigated and compared to evaluate their impact on the subsequent statistical analysis.

In order to explore and discover the overall structure of the data, find trends and groupings, several exploring techniques have been studied and implemented: Principal Component Analysis, Multi Dimensional Scaling based on Random Forest proximity matrices, K-means just to name a few.

Most of the research activities and the experiments has been focused on the identification and selection of potential biomarkers. Choosing the most suitable machine learning algorithm for biomarkers discovery was not an easy task, because different requirements had to be fulfilled:

- high prediction accuracy;
- ability to handle high dimensional datasets;
- interpretability of the prediction model.

Support Vector Machines, Artificial Neural Networks, K-Nearest Neighbors, Linear Discriminant Analysis are some of the most widely used machine learning techniques for creating predictive models. However, most of them provide too little insights on the importance of the variables involved in the prediction process. Variable importance measures, besides helping in the interpretation of a prediction model, can be crucial in the discovery and identification of candidate biomarkers.

Therefore the models I developed to carry out the multivariate analysis of metabolomics data were based on the Random Forest algorithm which is probably the closest to having the desired combination of features previously indicated.

I have mainly been concerned with the study of the calibration of the main parameters of the Random Forest algorithm and the development of procedures of crossvalidation in order to achieve two objectives: first, the creation of models that, starting from the metabolic profile were able to diagnose the presence or absence of a disease with a satisfactory degree of accuracy; then, the development of methods for the identification of those elements of the metabolic profile correlated to a particular disease state (biomarkers).

The techniques discussed, together with a variety of chemometrics and machine learning methods have been encoded in the R language and grouped within an open source package, named RFMarkerDetector, freely available online.

Bibliography

[1] BioCyc Database Collection. http://biocyc.org/BioCycUserGuide.shtml. [cited at p. 17]

[2] ChemSpider: Statistics. http://www.chemspider.com. [cited at p. 23]

[3] Chenomx NMR Suite. http://www.chenomx.com/about/about.php. [cited at p. 75]

[4] HMDB - Metabolomics Databases. http://www.hmdb.ca/w/databases. [cited at p. 14, 15, 16, 19]

[5] MassBank. http://www.massbank.jp/en/about.html. [cited at p. 19, 20]

[6] MestReNova. http://mestrelab.com/. [cited at p. 75]

[7] METLIN: Statistics. https://metlin.scripps.edu/statistics/. [cited at p. 20]

[8] PubChem: Statistics. https://pubchem.ncbi.nlm.nih.gov/. [cited at p. 22]

[9] RAMEDIS: Rare Metabolic Disease Database. http://goo.gl/BzXv33. [cited at p. 21]

[10] REACTOME - a curated pathway database. http://www.reactome.org/pages/about/reactome/. [cited at p. 17]

[11] Sheffield Hallam University:Chemistry. http://teaching.shu.ac.uk/hwb/chemistry/tutorials/molspec/nmr1.htm. [cited at p. 45, 46]

[12] Sheffield Hallam University:Chromatography. http://teaching.shu.ac.uk/hwb/chemistry/tutorials/chrom/chrom1.htm. [cited at p. 53]

[13] The University of Leeds: Introduction to Mass Spectrometry. http://www.astbury.leeds.ac.uk/facil/MStut/mstutorial.htm. [cited at p. 52, 53]

[14] Abbott Company. STARLIMS web-based platform for unified laboratory informatics. http://www.starlims.com/en-us/solutions/lims/. [cited at p. 27]

[15] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, Sept. 1997. [cited at p. 9]

[16] A. Bairoch. The Universal Protein Resource (UniProt). *Nucleic Acids Research*, 33(Database issue):D154–D159, 2004. [cited at p. 8]

[17] A. Bauch, I. Adamczyk, P. Buczek, F.-J. Elmer, K. Enimanev, P. Glyzewski, M. Kohler, T. Pylak, A. Quandt, C. Ramakrishnan, C. Beisel, L. Malmstrom, R. Aebersold, and B. Rinn. openbis: a flexible framework for managing and analyzing complex data in biology research. *BMC Bioinformatics*, 12(1):468, 2011. [cited at p. 28]

[18] D. Bentley. Analysis of a Laboratory Information Management System. http://www.umsl.edu/~sauterv/analysis/LIMS_example.html. [cited at p. 26]

[19] Bika Lab Systems. Bika Lims. http://www.bikalabs.com/softwarecenter/bika. [cited at p. 28]

[20] L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001. [cited at p. 68]

[21] L. Breiman and A. Cutler. *RFtools—for predicting and understanding data*. Interface Workshop, 2004. [cited at p. 68]

[22] M. L. Calle, V. Urrea, A.-L. Boulesteix, and N. Malats. AUC-RF: a new strategy for genomic profiling with random forest. *Human heredity*, 72(2):121–132, 2011. [cited at p. 74]

[23] R. Caspi, T. Altman, R. Billington, K. Dreher, H. Foerster, C. A. Fulcher, T. A. Holland, I. M. Keseler, A. Kothari, A. Kubo, M. Krummenacker, M. Latendresse, L. A. Mueller, Q. Ong, S. Paley, P. Subhraveti, D. S. Weaver, D. Weerasinghe, P. Zhang, and P. D. Karp. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic acids research*, 42(Database issue):D459–71, jan 2014. [cited at p. 17, 18]

[24] B. S. Chang and D. H. Lowenstein. Epilepsy. *New England Journal of Medicine*, 349(13):1257–1266, 2003. [cited at p. 74]

[25] D. Croft, A. F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu, M. Caudy, P. Garapati, M. Gillespie, M. R. Kamdar, B. Jassal, S. Jupe, L. Matthews, B. May, S. Palatnik, K. Rothfels, V. Shamovsky, H. Song, M. Williams, E. Birney, H. Hermjakob, L. Stein, and P. D'Eustachio. The Reactome pathway knowledgebase. *Nucleic Acids Research*, 42(Database issue):D472–7, jan 2014. [cited at p. 16, 17]

[26] K. Degtyarenko, P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj, and M. Ashburner. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Research*, 36(Database issue):D344–50, Jan. 2008. [cited at p. 23]

[27] M. V. der Werf, R. Jellema, and T. Hankemeier. Towards replacing closed with open target selection strategies. *J. Ind. Microbiol. Biotechnol.*, 32:234–252, 2005. [cited at p. 2]

[28] W. Dunn, R. Goodacre, L. Neyses, and M. Mamas. Integration of metabolomics in heart disease and diabetes research: current achievements and future outlook. *Bioanalysis*, 3:2205–2222, 2011. [cited at p. 3]

[29] O. Fiehn. Metabolomics - the link between genotypes and phenotypes. *Plant Mol. Biol.*, 48:155–171, 2002. [cited at p. 3]

[30] O. Fiehn, J. Kopka, P. Dormann, T. Altmann, R. Trethewey, and L. Willmitzer. Metabolite profiling for plant functional genomics. *Nat. Biotechnol.*, 18:1157–1161, 2000. [cited at p. 2]

[31] R. Fielding and R. Taylor. Principled design of the modern web architecture. *Acm T Internet Tech*, 2(2):115–150, 2002. [cited at p. 41]

[32] R. E. Finnigan. Quadrupole mass spectrometers. *Analytical Chemistry*, 1994. [cited at p. 4]

[33] T. Fuchs. script.aculo.us. 2005. last visited: 9 May 2013. [cited at p. 32]

[34] J. Garrett. Ajax: A new approach to web applications. 2005. last visited: 10 May 2013. [cited at p. 32]

[35] S. Gates and C. Sweeley. Quantitative metabolic profiling based on gas chromatography. *Clin. Chem.*, 10:1663–1673, 1978. [cited at p. 3]

[36] T. Gebregiworgis and R. Powers. Application of NMR metabolomics to search for human disease biomarkers. *Combinatorial chemistry & high throughput screening*, 15(8):595–610, Sept. 2012. [cited at p. 59]

[37] G. Gibbon. A brief history of lims. *Laboratory Automation & Information Management*, 32(1):1–5, 1996. [cited at p. 26]

[38] J. Ginsbach and F. Dunnivant. *CHED 361-Online book on Basic GC-MS*. ABSTRACTS . . . , 2008. [cited at p. 54]

[39] Goomedic.com. 15 Free and Open source LIMS. http://www.goomedic.com/15-free-and-open-source-lims-laboratory-information-management-system-programs-and-projects.html. [cited at p. 29]

[40] W. Griffiths and Y. Wang. Mass spectrometry: from proteomics to metabolomics and lipidomics. *Chem. Soc. Rev.*, 38:1882–1896, 2009. [cited at p. 3]

[41] W. J. Griffiths. Metabolomics, metabonomics and metabolite profiling, 2008. [cited at p. 52, 55, 56, 57]

[42] K. Haug, R. M. Salek, P. Conesa, J. Hastings, P. de Matos, M. Rijnbeek, T. Mahendraker, M. Williams, S. Neumann, P. Rocca-Serra, E. Maguire, A. González-Beltrán, S.-A. Sansone, J. L. Griffin, and C. Steinbeck. MetaboLights–an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Research*, 41(Database issue):D781–6, Jan. 2013. [cited at p. 7, 10, 12, 13, 14]

[43] Heinemeier Hansson, D. Ruby on Rails Framework. http://rubyonrails.org/. [cited at p. 30]

[44] H. Horai, M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, Y. Ojima, K. Tanaka, S. Tanaka, K. Aoshima, Y. Oda, Y. Kakazu, M. Kusano, T. Tohge, F. Matsuda, Y. Sawada, M. Y. Hirai, H. Nakanishi, K. Ikeda, N. Akimoto, T. Maoka, H. Takahashi, T. Ara, N. Sakurai, H. Suzuki, D. Shibata, S. Neumann, T. Iida, K. Tanaka, K. Funatsu, F. Matsuura, T. Soga, R. Taguchi, K. Saito, and T. Nishioka. MassBank: a public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry*, 45(7):703–714, July 2010. [cited at p. 19]

[45] D. Hoult, S. Busby, D. Gadian, G. Radda, R. Richards, and P. Seeley. Observation of tissue metabolites using 31p nuclear magnetic resonance. *Nature*, 252:285–287, 1974. [cited at p. 4]

[46] J. Hummel, J. Selbig, D. Walther, and J. Kopka. The Golm Metabolome Database: a database for GC-MS based metabolite profiling. In *Metabolomics*, pages 75–95. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007. [cited at p. 18]

[47] R. L. Jolley and M. L. Freeman. Automated carbohydrate analysis of physiologic fluids. *Clinical chemistry*, 14(6):538–547, June 1968. [cited at p. 3]

[48] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori. The KEGG resource for deciphering the genome. *Nucleic acids research*, 32(Database issue):D277–80, Jan. 2004. [cited at p. 8, 15]

[49] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, 40(D1):gkr988–D114, nov 2011. [cited at p. 15]

[50] M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Research*, 42(Database issue):D199–205, jan 2014. [cited at p. 16]

[51] P. Kebarle. A brief overview of the present status of the mechanisms involved in electrospray mass spectrometry. *Journal of mass spectrometry : JMS*, 35(7):804–817, July 2000. [cited at p. 56]

[52] J. Kent. The Right LIMS Delivery Method. http://www.bio-itworld.com/uploadedFiles/Bio-IT_World/Bio-IT_Issues/2009/Jan-Feb/LabAuto_supplement.pdf. [cited at p. 27, 28]

[53] J. Knapp and W. Cabrera. *Metabolomics: Metabolites, Metabonomics, and Analytical Technologies*. Nova Science Publishers, Inc., 2011. [cited at p. 2, 3, 5, 6, 17]

[54] W. Kolch, H. Mischak, and A. R. Pitt. The molecular make-up of a tumour: proteomics in cancer research. *Clinical science (London, England : 1979)*, 108(5):369–383, May 2005. [cited at p. 6]

[55] R. Konertz. Open-LIMS. http://www.open-lims.org. [cited at p. 28]

[56] J. Kopka, N. Schauer, S. Krueger, C. Birkemeyer, B. Usadel, E. Bergmüller, P. Dörmann, W. Weckwerth, Y. Gibon, M. Stitt, L. Willmitzer, A. R. Fernie, and D. Steinhauser. GMD@CSB.DB: the Golm Metabolome Database. *Bioinformatics*, 21(8):1635–1638, Apr. 2005. [cited at p. 18]

[57] J. Lindon, J. Nicholson, and E. Holmes. The handbook of metabonomics and metabolomics, 2011. [cited at p. 43, 44, 45, 49, 50, 51, 55, 56]

[58] J. Lindon, J. Nicholson, E. Holmes, and J. Everett. Metabonomics: Metabolic processes studied by NMR spectroscopy of biofluids. *Concepts in Magnetic Resonance*, 12(5):289–320, 2000. [cited at p. 59]

[59] U. Manber, M. Smith, and B. Gopal. Webglimpse: Combining browsing and searching. In *Proceedings of the Annual Conference on USENIX Annual Technical Conference*, ATEC '97, pages 15–15, Berkeley, CA, USA, 1997. USENIX Association. [cited at p. 9]

[60] M. S. Monteiro, M. Carvalho, M. L. Bastos, and P. G. de Pinho. Metabolomics Analysis for Biomarker Discovery: Advances and Challenges. *Current Medicinal Chemistry*, 20(2):257–271, 2013. [cited at p. 59, 60]

[61] J. Morris, S. Gayther, I. Jacobs, and C. Jones. A perl toolkit for lims development. *Source Code Biol Med*, 3(1):4, 2008. [cited at p. 25]

[62] E. Nelson, B. Piehler, J. Eckels, A. Rauch, M. Bellew, P. Hussey, S. Ramsay, C. Nathe, K. Lum, K. Krouse, D. Stearns, B. Connolly, T. Skillman, and M. Igra. Labkey server: An open source platform for scientific data integration, analysis and collaboration. *BMC Bioinformatics*, 12(1):71, 2011. [cited at p. 28]

[63] J. Nicholson, J. Lindon, and E. Holmes. "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. [cited at p. 3]

[64] D. Nix, T. Di Sera, B. Dalley, B. Milash, R. Cundick, K. Quinn, and S. Courdy. Next generation tools for genomic data generation, distribution, and visualization. *BMC Bioinformatics*, 11(1):455, 2010. [cited at p. 25, 29]

[65] A. O'Gorman. Metabolic profiling and fingerprinting for the detection and discrimination of mechanical damage in mushrooms (Agaricus bisporus) during storage. 2010. [cited at p. 5]

[66] Oracle Inc. MySQL: The world's most popular open source database. http://www.mysql.com/. [cited at p. 32]

[67] P. Palla, G. Frau, L. Vargiu, and P. Rodriguez-Tome. Qtreds: a flexible lims for omics laboratories. *Embnet.journal*, 18(Suppl B):38–39, 2012. [cited at p. 40]

[68] Z. Pan and D. Raftery. Comparing and combining NMR spectroscopy and mass spectrometry in metabolomics. *Analytical and bioanalytical chemistry*, 387(2):525–527, Jan. 2007. [cited at p. 5]

[69] H. E. Pence and A. Williams. ChemSpider: An Online Chemical Information Resource. *Journal of Chemical Education*, 87(11):1123–1124, Aug. 2010. [cited at p. 23, 24]

[70] J. J. Pitt. Principles and applications of liquid chromatography-mass spectrometry in clinical biochemistry. *The Clinical biochemist. Reviews / Australian Association of Clinical Biochemists*, 30(1):19–34, Feb. 2009. [cited at p. 57]

[71] P. Prosad and G. Bodhe. Trends in laboratory information system. *Chemom Intell Lab Syst*, 118:187–192, 2012. [cited at p. 26]

[72] P. Rocca-Serra, M. Brandizi, E. Maguire, N. Sklyar, C. Taylor, K. Begley, D. Field, S. Harris, W. Hide, O. Hofmann, S. Neumann, P. Sterk, W. Tong, and S. Sansone. Isa software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics*, 26(18):2354–2356, 2010. [cited at p. 12, 35]

[73] U. Roessner and J. Bowne. What is metabolomics all about? *Biotechniques*, 2009. [cited at p. 44]

[74] R. Ryhage. Use of a Mass Spectrometer as a Detector and Analyzer for Effluent Emerging from High Temperature Gas Liquid Chromatography Columns. *Analytical Chemistry*, 1964. [cited at p. 4]

[75] M. Sabatine, E. Liu, D. Morrow, E. Heller, R. McCarroll, R. Wiegand, G. Berriz, F. Roth, and R. Gerszten. Metabolomic identification of novel biomarkers of myocardial ischemia. *Circulation*, 112:3868–3875, 2005. [cited at p. 2]

[76] D. H. Sandberg, J. Sjoevall, K. Sjoevall, and D. A. Turner. Measurement of human serum bile acids by gas-liquid chromatography. *Journal of Lipid Research*, 6:182–192, Apr. 1965. [cited at p. 3]

[77] Sapio Sciences. Examplar LIMS. http://www.sapiosciences.com/LIMS/index.html. [cited at p. 27]

[78] J. Schellenberger, J. O. Park, T. M. Conrad, and B. O. Palsson. Bigg: a biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics*, 11(1):213, apr 2010. [cited at p. 13, 15]

[79] C. Scriver. *Online Metabolic and Molecular Basis of Inherited Disease (OMMBID)*. New York: McGraw-Hill, 2006. [cited at p. 8]

[80] C. Sidore, S. Sanna, A. Kwong, H. Kang, R. Cusano, M. Pitzalis, M. Zoledziewska, A. Maschio, F. Busonero, M. Lobina, A. Angius, R. Lyons, B. Terrier, C. Brennan, R. Atzeni, A. Mulas, M. Dei, M. Piras, S. Lai, F. Reinier, R. Berutti, C. Jones, M. Marcelli, M. Urru, M. Oppo, D. Schlessinger, G. Abecasis, and C. F. Whole genome sequencing of 2100 individuals in the founder sardinian population [abstract]. *Abstract volume of the 62nd Annual Meeting of The American Society of Human Genetics: 6-10 November 2012; San Francisco, USA*, page 76, 2012. [cited at p. 29]

[81] J. Sjovall. Separation and determination of bile acids. *Methods of biochemical analysis*, 1964. [cited at p. 3]

[82] C. A. Smith, G. O. Maille, E. J. Want, C. Qin, S. A. Trauger, T. R. Brandon, D. E. Custodio, R. Abagyan, and G. Siuzdak. METLIN: A Metabolite Mass Spectral Database. *Therapeutic Drug Monitoring*, 27(6):747, Dec. 2005. [cited at p. 20]

[83] A. Smolinska, L. Blanchet, L. Buydens, and S. Wijmenga. NMR and pattern recognition methods in metabolomics: From data acquisition to biomarker discovery: A review. *Anal. Chim. Acta*, 750:82–97, 2012. [cited at p. 3, 45, 61, 63]

[84] O. Sparkman. *Mass Spectrometry Desk Reference*. Global View Pub. [cited at p. 50]

[85] S. Stephenson. Prototype JavaScript Framework. http://prototypejs.org/. [cited at p. 32]

[86] G. Stocker, M. Fischer, D. Rieder, G. Bindea, S. Kainz, M. Oberstolz, J. McNally, and Z. Trajanoski. ilap a workflow-driven software for experimental protocol development, data acquisition and analysis. *BMC Bioinformatics*, 10:390, 2009. [cited at p. 25, 26]

[87] K. Suhre. Introduction. In *Genetics Meets Metabolomics*, pages 1–4. Springer New York, 2012. [cited at p. 59]

[88] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston. Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of chemical information and computer sciences*, 43(6):1947–1958, Nov. 2003. [cited at p. 71]

[89] C. Taylor, D. Field, S.-A. Sansone, J. Aerts, R. Apweiler, M. Ashburner, C. Ball, P.-A. Binz, M. Bogue, T. Booth, A. Brazma, R. Brinkman, A. Michael Clark, E. Deutsch, O. Fiehn, J. Fostel, P. Ghazal, F. Gibson, T. Gray, G. Grimes, J. Hancock, N. Hardy, H. Hermjakob, R. Julian, M. Kane, C. Kettner, C. Kinsinger, E. Kolker, M. Kuiper, and N. Novere. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the mibbi project. *Nat Biotechnol*, 26(8):889–896, 2008. [cited at p. 35]

[90] T. Töpel, D. Scheible, F. Trefz, and R. Hofestädt. RAMEDIS: a comprehensive information system for variations and corresponding phenotypes of rare metabolic diseases. *Human mutation*, 31(1):E1081–8, Jan. 2010. [cited at p. 21]

[91] F. Trefz, D. Scheible, H. Götz, T. Töpel, R. Hofestädt, and G. Frauendienst-Egger. METAGENE and RAMEDIS: databases for metabolic diseases and patients with inborn errors on metabolism. *Journal of Inherited Metabolic Disease*, 31, 2008. [cited at p. 21]

[92] T. Triplet and G. Butler. The enzymetracker: an open-source laboratory information management system for sample tracking. *BMC Bioinformatics*, 13(1):15, 2012. [cited at p. 28, 29]

[93] C. Truong, L. Groeneveld, B. Morgenstern, and E. Groeneveld. Molabis - an integrated information system for storing and managing molecular genetics data. *BMC Bioinformatics*, 12(1):425, 2011. [cited at p. 28]

[94] S. Tyagi, S. Raghvendra, and U. Singh. Applications of metabolomics-a systematic study of the unique chemical fingerprints: an overview. *Int. J. Pharm. Sci. Rev. Res.*, 3(1):83–86, 2010. [cited at p. 5, 6]

[95] S. Vaidyanathan, G. Harrigan, and R. Goodacre. Metabolome Analyses:: Strategies for Systems Biology, 2006. [cited at p. 54]

[96] T. Van Rossum, B. Tripp, and D. Daley. Slims–a user-friendly sample operations and inventory management system for genotyping labs. *Bioinformatics*, 26(14):1808–1810, 2010. [cited at p. 28]

[97] T. D. Veenstra. Metabolomics: the final frontier? *Genome Medicine*, 4(4):40, 2012. [cited at p. 61]

[98] A. Verikas, A. Gelzinis, and M. Bacauskiene. Mining data with random forests: A survey and results of new tests. *Pattern Recognition*, 44(2):330–349, 2011. [cited at p. 68]

[99] S. Villas-Boas, J. Nielsen, and J. Smedsgaard. Metabolome analysis: an introduction, 2007. [cited at p. 43, 44, 51, 52, 53, 54, 56]

[100] Y. Wang, J. Xiao, T. O. Suzek, J. Zhang, J. Wang, and S. H. Bryant. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Research*, 37(Web Server issue):W623–33, July 2009. [cited at p. 22]

[101] D. Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer*, 28(1):31–36, 1988. [cited at p. 9]

[102] D. S. Wishart, T. Jewison, A. C. Guo, M. Wilson, C. Knox, Y. Liu, Y. Djoumbou, R. Mandal, F. Aziat, E. Dong, S. Bouatra, I. Sinelnikov, D. Arndt, J. Xia, P. Liu, F. Yallou, T. Bjorndahl, R. Perez-Pineiro, R. Eisner, F. Allen, V. Neveu, R. Greiner, and A. Scalbert. HMDB 3.0–The Human Metabolome Database in 2013. *Nucleic Acids Research*, 41(Database issue):D801–7, Jan. 2013. [cited at p. 8, 9, 11]

[103] D. S. Wishart, C. Knox, A. C. Guo, R. Eisner, N. Young, B. Gautam, D. D. Hau, N. Psychogios, E. Dong, S. Bouatra, R. Mandal, I. Sinelnikov, J. Xia, L. Jia, J. A. Cruz, E. Lim, C. A. Sobsey, S. Shrivastava, P. Huang, P. Liu, L. Fang, J. Peng, R. Fradette, D. Cheng, D. Tzur, M. Clements, A. Lewis, A. De Souza, A. Zuniga, M. Dawe, Y. Xiong, D. Clive, R. Greiner, A. Nazyrova, R. Shaykhutdinov, L. Li, H. J. Vogel, and I. Forsythe. HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Research*, 37(Database issue):D603–10, Jan. 2009. [cited at p. 7, 8]

[104] D. S. Wishart, D. Tzur, C. Knox, R. Eisner, A. C. Guo, N. Young, D. Cheng, K. Jewell, D. Arndt, S. Sawhney, C. Fung, L. Nikolai, M. Lewis, M.-A. Coutouly, I. Forsythe, P. Tang, S. Shrivastava, K. Jeroncic, P. Stothard, G. Amegbey, D. Block, D. D. Hau, J. Wagner, J. Miniaci, M. Clements, M. Gebremedhin, N. Guo, Y. Zhang, G. E. Duggan, G. D. MacInnis, A. M. Weljie, R. Dowlatabadi, F. Bamforth, D. Clive, R. Greiner, L. Li, T. Marrie, B. D. Sykes, H. J. Vogel, and L. Querengesser. HMDB: the Human Metabolome Database. *Nucleic Acids Research*, 35(Database issue):D521–6, Jan. 2007. [cited at p. 8, 9, 10]

[105] S. Wood. Comprehensive laboratory informatics: A multilayer approach. *Am Lab*, 39(16):20–23, 2007. [cited at p. 27]

[106] B. Worley and R. Powers. Multivariate Analysis in Metabolomics. *Current Metabolomics*, 1:92–107, Jan. 2013. [cited at p. 63]

List of Publications Related to the Thesis

Published Works

- P. Palla, G. Frau, L. Vargiu, P. Rodriguez-Tomé, *QTREDS: a Ruby on Rails based platform for omics laboratories* BMC Bioinformatics 2014, volume 15 (Suppl1):S13.
- P. Palla, G. Frau, L. Vargiu, P. Rodriguez-Tomé, *QTREDS: a flexible LIMS for omics laboratories* EMBnet.journal pp 41- 43 vol. 18, Supplement B - 2012.

Submitted Works

- P. Palla, G. Armano *RFmarkerDetector: an R package for multivariate analysis of metabolomics data using Random Forests* (submitted).

Conferences

- P. Palla, G. Frau, L. Vargiu, P. Rodriguez-Tomé, *QTREDS: a flexible LIMS for omics laboratories*, 12th International workshop on Network Tools and Applications in Biology (oral presentation) November 14 - 16, 2012, Como, Italy.