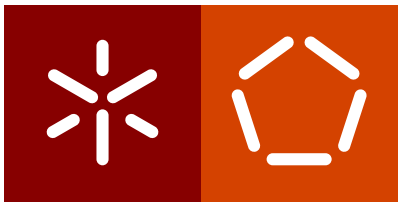
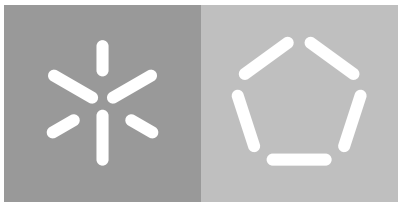


Universidade do Minho
Escola de Engenharia
Departamento de Informática

Sara Manso de Sousa Cardoso

Development of web-based tools for metabolomics data analysis and mining

December 2017



Universidade do Minho
Escola de Engenharia
Departamento de Informática

Sara Manso de Sousa Cardoso

Development of web-based tools for metabolomics data analysis and mining

Master dissertation
Master Degree in Bioinformatics

Dissertation supervised by
Miguel Francisco de Almeida Pereira da Rocha
Marcelo Maraschin, Universidade Federal de Santa Catarina

December 2017

AGRADECIMENTOS

Quero aqui deixar o meu agradecimento a todas as pessoas que, directa ou indirectamente, contribuiram para a concretização deste trabalho.

Em primeiro lugar, quero agradecer ao professor e orientador Miguel Rocha, por me ter dado a oportunidade de desenvolver este trabalho, pela partilha de conhecimentos e por todas as oportunidades proporcionadas que surgiram deste trabalho.

Quero também agradecer ao Marcelo Maraschin, da Universidade Federal de Santa Catarina, pela orientação, por concender dados de metabolómica para análise e pela colaboração na análise desses mesmos dados, mais especificamente, os dados de cascas de banana, que resultou no caso de estudo da presente dissertação e num artigo publicado.

Quero também agradecer a todos os meus amigos por me terem sempre apoiado ao longo dos anos, mas especialmente neste último ano.

Por último, e não de todo menos importante, quero agradecer a toda a minha família por todo o apoio dado não só ao longo deste ano, mas desde sempre. Mas especialmente ao meu pai, por me ter ensinado muito do que sou hoje e por muito mais; à tia São, por me ter sempre apoiado, mas especialmente por ter estado sempre ao meu lado ao longo deste ano; e, mais especialmente, à minha mãe, que sempre foi quem mais me apoiou em tudo e quem mais me incentivava a crescer profissionalmente e pessoalmente. A ela dedico este trabalho.

Mais uma vez, obrigado a todos.

ABSTRACT

The recent advances in metabolomics experimental techniques have provided novel approaches for many research issues in the biological fields. Indeed, the ability to identify and quantify numerous compounds in biological samples provides significant advances in functional genomics, biomarker identification, sample characterization or drug discovery and development. To take full advantage of these data advanced bioinformatics methods for data analysis and mining have been required.

A number of methods and tools for metabolomics data analysis have been put forward recently, being one of the major limitations still faced the lack of integrated frameworks for extracting relevant knowledge from these data and being able to integrate these data with previous biochemical knowledge. Also, the lack of reproducibility in many data analyses or data mining processes is a strong obstacle for biological discovery.

In recent work from the host group, *specmine*, a metabolomics and spectral data analysis/mining framework, in the form of a package for the R system, has been developed to address some of these issues.

In this thesis, an integrated web-based platform for metabolomics data analysis and mining, named *WebSpecmine*, was designed and developed, based on the *specmine* package, thus providing an easier and friendly user interface. This website provides means for analysing metabolomics data from different formats, including tasks such as pre-processing, univariate and multivariate analysis and metabolite identification. This web-based platform was developed collaboratively and, therefore, this work focused mainly in data from nuclear magnetic ressonance and mass spectrometry.

Also, the package faced some limitations regarding types of analysis not yet provided, such as metabolite identification for other data formats besides Mass Spectrometry coupled to Liquid Chromatography. Therefore, the extension of the metabolite identification feature was addressed, by implementing such analysis for Nuclear Magnetic Ressonance data in the *specmine* package, as well as making it available in the website.

The website was validated by applying it to reproduce the pipelines from previous studies that made use of the *specmine* package. Furthermore, a case study involving banana peels and the analysis of their characteristics and potential made use of the newly created website to further validate its functionality. All the analyses here executed were stored and are available in the web application, as public projects.

RESUMO

Os mais recentes avanços nas técnicas experimentais metabolómicas têm levado a novas abordagens de muitas questões na investigação em áreas biológicas. De facto, a capacidade de identificar e quantificar os inúmeros compostos presentes nas amostras biológicas veio provocar enormes avanços na genómica funcional, identificação de biomarcadores, caracterização de amostras e descoberta e desenvolvimento de drogas. Para tirar maior partido destes dados, é necessário a existência de métodos avançados de bioinformática para a análise e mineração de dados.

Vários métodos e ferramentas que permitem a análise de dados metabolómicos têm vindo a ser apresentadas, tendo no entanto como grande limitação a falta de extracção de conhecimento relevante destes dados e integrá-los com conhecimento bioquímico anterior. Para além disto, a falta de reprodutibilidade de muitas análises de dados ou de processamentos de mineração é um grande obstáculo à descoberta biológica.

Em trabalhos recentes do grupo de acolhimento foi desenvolvido um package para o sistema R por forma a abordar algumas destas questões. Este package, denominado *specmine*, permite a análise e mineração de dados espectrais e de metabolómica.

Na presente tese, uma plataforma web integrada para a análise e mineração de dados de metabolómica, denominada *WebSpecmine*, foi desenvolvida, baseada no package *specmine*, fornecendo assim uma interface simples e fácil para o usuário. Este site permite a análise de dados de metabolómica de formatos diferentes, incluindo pre-processamento, análises univariada e multivariada, e identificação de metabolitos. Esta plataforma web foi desenvolvida de forma colaborativa e, deste modo, o presente trabalho focou-se maioritariamente em dados provenientes das técnicas espectrometria de massa e ressonância magnética nuclear.

Para além disto, o package apresentava algumas limitações no que toca a tipos de análise ainda não disponíveis, como é o caso da identificação de metabolitos para outros formatos de dados que não a espectrometria de massa acoplada com cromatografia líquida. Assim, a funcionalidade de identificação de metabolitos no *specmine* package foi estendida a dados de ressonância magnética nuclear, bem como também implementada no website.

O site foi validado através da sua aplicação para reproduzir pipelines de estudos anteriores que fizeram uso do package *specmine*. Para além disto, um estudo de caso envolvendo cascas de banana e a análise das suas características e potencial, fez uso do site recentemente criado para também validar a sua funcionalidade. Todas as análises aqui executadas foram guardadas na aplicação web, estando disponíveis para consulta, como projectos públicos.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

R^2 coefficient of determination. 14
1D-NMR one-dimensional NMR. viii, 9, 13, 14, 17, 18, 25, 26, 75, 76, 78, 80, 84
2D-NMR two-dimensional NMR. 9, 14, 63

AE Aqueous Extracts. 73, 74
ANOVA Analysis of Variance. viii, ix, 23, 24, 56, 57, 60, 63, 68, 73, 75, 78, 80, 84
AUC Area Under the ROC Curve. 14, 23

ChEBI Chemical Entities of Biological Interest. 18, 19
COW Correlation Optimized Warping. 11
CSS Cascading Style Sheets. vi, 28, 29

DTW Dynamic Time Warping. 11

FAAH Fatty Acid Amide Hydrolase. 65, 66, 70
FDR False Discovery Rate. ix, 57, 60, 67, 78

GC Gas Chromatography. 7, 8
GC-MS Gas Chromatography-Mass Spectrometry. 6, 20
GC/LC-MS Gas Chromatography/Liquid Chromatography-Mass Spectrometry. 10–12, 21, 22

HMDB Human Metabolome Database. 13, 17, 18, 46–48, 61, 80

IR Infrared Spectroscopy. 7, 22

KEGG Kyoto Encyclopedia of Genes and Genomes. 18, 19

LC Liquid Chromatography. 7, 8
LC-MS Liquid Chromatography-Mass Spectrometry. 6, 13, 20, 23, 25, 40, 41, 46, 65, 66, 84
LDA Linear Discriminant Analysis. 15, 43, 84

MMCD Madison Metabolomics Consortium Database. 13, 18
MS Mass Spectrometry. vi, vii, ix, 7, 8, 10, 13, 17–20, 22–24, 28, 30, 34, 42, 46, 84

NAE N-Acyl Ethanolamine. 65, 70
NMR Nuclear Magnetic Resonance. v, vii, ix, 6–10, 12–14, 17–20, 22–28, 30, 40–42, 47, 48, 73, 74, 82, 84, 85

NN Neural Networks. 15

PCA Principal Components Analysis. vii, viii, 23, 24, 41, 49, 57, 58, 73, 75, 79, 84
PLS Partial Least Squares. vii, 16, 25, 43, 49, 59, 60, 63, 64, 84
PLS-DA Partial Least Squares Discriminant Analysis. 15, 21
PLS-r Partial Least Squares Regression. 15

RMSE Root Mean Square Error. 14
ROC Receiver Operating Characteristic. 44

SNP Single Nucleotide Polymorphism. 17
SVMs Support Vector Machines. 15, 16, 21, 43, 84

UV-Vis Ultraviolet-Visible. 7, 22, 28, 30, 84

1

INTRODUCTION

1.1 CONTEXT

Omics technologies are responsible for analysing a global set of molecules and their interactions at a large scale. These technologies have revolutionized the way biological research is conducted, being very important in areas such as functional genomics, characterization of biological systems, biotechnology and biomedical research (Costa, 2014; Villas-Bôas et al., 2006).

Two of these technologies are transcriptomics, which represents the study of the total set of mRNA present in the cell, and proteomics, consisting in the analysis of all proteins present in the cell (Villas-Bôas et al., 2006).

Metabolomics is a more recent omics technology and, as the name suggests, analyses all or part of the metabolome, which corresponds to the set of all metabolites used or formed by the cell under study. The metabolome is separated into two major groups: endometabolome, which is the group of intracellular metabolites, and exometabolome, the group of secreted metabolites (Villas-Bôas et al., 2006).

Metabolites are compounds of low molecular weight (less than 1000 Daltons), such as glucose, being the biochemical reaction intermediates, playing a role in connecting the pathways that occur in the cell. The metabolites are divided into primary and secondary metabolites. The former are directly involved in the normal cellular growth, development and division, whereas the latter are not directly involved in these processes, although they have very important functions, both in the cell (e.g. in regulatory or signalling pathways) and biomedical research (as is the case of antibiotics).

Therefore, metabolites represent essential information about the cell function, helping to understand and define the cell and tissue phenotype, in response to genetic or environmental changes. Thus, metabolomics data help in the study of metabolic systems, sample discrimination and identification of biomarkers, for example (Costa, 2014; Villas-Bôas et al., 2006).

Although transcriptomics and proteomics are now in rapid development, with high throughput analysis methods being used, the methods used in metabolomics are far less

common and there is not a single one that allows the analysis of the whole metabolome. This is partially due to the fact that metabolites are compounds with great chemical diversity, that can go from hydrophobic lipids to volatile alcohols, and with fast turnover, i.e., many metabolites are rapidly consumed/transformed as soon as they are formed, thus having low concentrations in the cell. Also, since the same metabolite can participate in many different pathways, interpretation of metabolomics data can be difficult (Villas-Bôas et al., 2006).

Still, metabolomics has a wide range of applications, such as plant biology, nutrition, drug discovery and the study of human diseases, which can consist of finding diagnostic and prognostic biomarkers, environmental factors on human health and predict treatment response (Alonso et al., 2015).

A number of computational tools have been put forward over the last years for metabolomics data analysis, targeting broader purposes or more specific tasks, as well as covering a wider or smaller range of experimental techniques. Many of these tools that analyse metabolomics data require programming skills, as they come in the form of packages to be used in different programming systems, with few ones based on a web service, which facilitates the analysis of these type of data for people who do not have programming skills.

The existing web services lack some tools regarding metabolomics data analysis, such as the lack of training models diversity, or not allowing the user to save data to be used later or shared. Accordingly, following the development of the R package *specmine* by the host research group (Costa et al., 2016), a metabolomics and spectral data analysis/mining framework that addresses the development of customizable data analysis pipelines, covering different types of metabolomics and spectral data, a website supporting this package was created.

1.2 OBJECTIVES

Given the context described above, the main aim of this work was the design and development of a web-based computational platform for metabolomics data analysis and knowledge extraction, based on the *specmine* R package previously developed on the host group. The work aimed to address the exploration and integration of data from distinct experimental techniques, focusing on Nuclear Magnetic Resonance (NMR) and Mass Spectroscopy (MS), and, as regards to analysis tools, on machine learning, feature selection and metabolite identification, as the website was created colaboratively.

More specifically, the work aimed to address the following scientific/ technological goals:

- To review the state of the art regarding metabolomics data analysis, including available computational tools;

- To design and implement adequate web-based interfaces for metabolomics data analysis and data mining pipelines, based on the functions provided by the *specmine* package.
- To design and implement novel functions for metabolomics data analysis/mining extending the functionality of *specmine*, focusing on metabolite identification.
- To validate the tools developed with several case studies of interest for the host groups in the analysis of the biomedical potential of natural products, including for instance bananas or propolis.
- To write scientific publications with the results of the work.

1.3 DISSERTATION ORGANIZATION

The present dissertation is divided into six chapters. This first chapter consists in a brief introduction to the subject of the present work, by providing a context, reasons and objectives for the work developed. After this chapter, the second one aims to present the state of the art of the metabolomics field, by providing an explanation on the main steps of a metabolomics experiment, as well as current metabolomics tools available, not only in analysis of such data but also in databases that provide metabolomics data and associated information.

After these two chapters, the ”Development” chapter explains how the work was conducted, by showing the development strategy applied and tools used in the process. Furthermore, it is also given a detailed information on how the website works and why and how the new features were added to the *specmine* package.

In the following chapter, two use cases are presented, to show how previously developed studies can be reproduced by using the developed site. The next chapter contains a case study developed during this work that made use of the developed technologies, focusing on the analysis of NMR data for bananas.

The final chapter contains the main conclusions taken from this work and an analysis on future work that might be necessary to further improve both *specmine* package and *WebSpecmine* site.

2

STATE OF THE ART

This chapter covers the state of the art of the metabolomics field, addressing the main steps of a metabolomics experiment.

Usually, a metabolomics experiment is performed following a certain methodology, beginning with the sample preparation to be studied, involves the acquisition of data, which can be quantification and/or qualification of the metabolites present, according to the type of technique chosen, and ends in the treatment, analysis and interpretation of those data.

The **Table 1** shows some of the metabolomics studies that use chemical analytical technologies, data treatment and analysis methods covered in this chapter.

2.1 METHODOLOGIES

The metabolome analysis, which corresponds to the identification and quantification of the metabolites in biological samples, is done making use of, in general, four main methodologies. Metabolic Fingerprinting aims to know the fingerprint of the metabolites produced in the cell, without providing information about specific metabolites. Therefore, it is a technique utilized to group/discriminate the samples under study, such as sorting mutants or types of growth media, according to the metabolites present. Metabolic Footprinting analyses the exometabolome through the analysis of specific metabolites or fingerprints, similarly to what happens with Fingerprinting. Metabolite Profiling is a semi-quantitative technique involved in analysing specific groups of metabolites, whose acquired data may be further used for metabolic models. Lastly, Metabolite Target Analysis quantifies the metabolites involved in a certain part of the metabolism (Villas-Bôas et al., 2006; Costa, 2014).

Even though Metabolic Fingerprinting and Footprinting can help in the discrimination of the samples under study, the data obtained from these two methodologies is not so useful for integration with other omics data (transcriptomics and proteomics), where quantitative data is required (Villas-Bôas et al., 2006).

Table 1: Metabolomics studies regarding MS and NMR techniques.

Study Description	Technique	Pre-processing	Analysis	Reference
Analyze transgenic maize effects on the development of seeds	NMR	Normalization; Baseline Correction	Machine Learning	Castro and Manetti (2007)
Classification of rapeseed oils	NMR	Baseline Correction	Variable Selection	Chen et al. (2010)
Predict cancer-associated skeletal muscle wasting	NMR	Logaritmic Transformation; Normalization	Machine Learning	Eisner et al. (2010)
Metabolic profiling in Crohn's disease	NMR	Baseline Correction; Normalization; Mean Centering	Biomarker Analysis	Fathi et al. (2014)
Identification of the farm origin of salmon	NMR	Peak Alignment; Normalization	Machine Learning	Martinez et al. (2009)
Confirmation of wild and farmed salmon and their origins	NMR	Peak Alignmetn	Machine Learning	Masoum et al. (2007)
Metabolic profiles of tomato flesh and seeds during fruit development	NMR	Baseline correction; Peak Alignment	Machine Learning	Mounet et al. (2007)
Propolis classification	NMR	Normalization	Machine Learning	Papotti et al. (2010)
Evaluation of tubulointerstitial lesions' severity in patients with glomerulonephritides	NMR	Baseline Correction	Machine Learning	Psihogios] et al. (2007)
Study the metabolic response of the earthworm Eisenia fetida to two pesticides	NMR and GC-MS	Peak Alignment; Normalization	Metabolite Identification; Biomarker Analysis	McKelvie et al. (2009)
Identify novel biomarkers and pathways activated in myocardial ischemia	LC-MS	Logaritmic Transformation; Normalization	Pathway Analysis; Biomarker Analysis	Sabatine et al. (2005)
Development of an optimized extraction and derivatization protocol, using experimental design theory, for analyzing the human blood plasma metabolome	GC-MS	Baseline Correction; Alignment	Machine Learning	Jiye et al. (2005)
Exposure end effect markers of fruit and fruit fibre intake	LC-MS	Normalization	Biomarker Analysis	Kristensen et al. (2012)
Whether the metabolic signature in serum from patients with acute lymphoblastic leukemia is different from the healthy ones	LC-MS	Normalization	Metabolite Identification; Biomarker Analysis; Pathway Analysis; Enrichment Analysis	Bai et al. (2014)

2.2 SAMPLE PREPARATION

Sample Preparation for metabolomics analysis is very important, since it has to ensure that the samples are prepared to be a true representation of the original samples. This step differs with the organism in question and the cells structure, as well as the type of metabolites to be analysed. As mentioned above, metabolites have a high turnover, mainly the primary ones, which makes the quenching method an important step that should be done simultaneously or right after obtaining the samples that will be analysed. Quenching is a sudden and instant change of the samples temperature or pH, to stop all the biochemical processes that are taking place in the concerned cells and, therefore, obtain the metabolites present in that exact moment. It is noteworthy that the techniques used in the quenching vary according to the type of cells being studied (Villas-Bôas et al., 2006).

Next, to obtain the desired metabolites, the method varies according to their physico-chemical properties nature and localization, if intracellular or extracellular. If the aim is to analyse both intracellular and extracellular metabolites, without any distinction between these two types, the cells membranes and/or walls should be disrupted, followed by the separation of the compounds from the biological matrix. On the other hand, if the desire is to analyse both extracellular and intracellular metabolites separately, or only one of these, it is necessary to separate the cells from the extracellular medium and, thereafter, extract the intracellular metabolites from the cells and the extracellular ones from the medium (Villas-Bôas et al., 2006).

A final step is sample concentration, by removing totally or partially the solvents of the samples. This is done through freeze-drying the solvents (lyophilization) and removing them through sublimation. Removing water from aqueous samples through this method can also prevent heat degradation (Villas-Bôas et al., 2006).

2.3 TECHNIQUES

There are various techniques used in the analysis of the metabolome, mostly MS, normally coupled to chromatography techniques such as Gas Chromatography (GC) and Liquid Chromatography (LC), as well as NMR technique, Infrared Spectroscopy (IR), and Ultraviolet-Visible (UV-Vis) (Villas-Bôas et al., 2006).

Next, we will cover MS and NMR techniques, which lead to the types of data approached in this work.

2.3.1 *Mass Spectrometry (MS)*

Mass Spectrometry determines the mass charge ratio (m/z) of charged compounds, whether they are molecules, groups (clusters) of molecules, complexes or fragments, or any combination of these. To be able to know the mass/charge ratio of the different metabolites in the sample, usually it is necessary to analyse the metabolites separately. This separation is achieved by the techniques GC or LC, performed before the MS technique. Both chromatographies are based on the interaction of the different metabolites in the sample with the absorbent materials inside the chromatography column, where the metabolites with different chemical properties result in a different crossing time along the column. GC is only used in volatile samples. The different metabolites are then ionized, so that they are charged and, therefore, to be possible to determine their mass/charge ratio. The results, i.e., spectral data, are obtained in the form of MS spectrum, where the x-axis is the mass/charge ratio and the y-axis the intensity/quantity of the different ions originated by the metabolite, as shown in **Figure 1**. Thus, the peak patterns of each graph allow the metabolite identification, since each metabolite has a characteristic pattern (Alonso et al., 2015; Villas-Bôas et al., 2006).

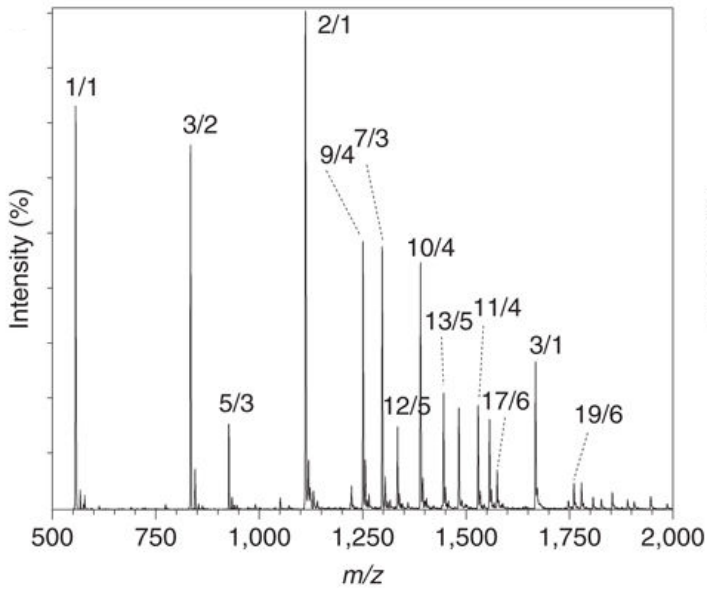


Figure 1: An example of an MS Spectrum, adapted from Bleiholder et al. (2011).

2.3.2 *Nuclear Magnetic Ressonance (NMR)*

NMR spectroscopy exploits the magnetic properties of certain atomic nuclei to determine the chemical and physical properties of those atoms or molecules. This happens because

the variation of the external magnetic field causes the absorption and re-emission of energy by the atomic nuclei to vary. Therefore, this shift is calculated as the difference between the resonance and the reference substance frequencies, divided by the operating frequency of the spectrometer.

It must be kept in mind that the analysis of different atomic nuclei generates different metabolomics data, being the most used one the hydrogen (1H-NMR), by virtue of being the most abundant in biological samples. Other atoms, less frequent, such as carbon and phosphorus, can also be the target of this technique. The spectra obtained display the chemical shifts on the x-axis and the signal intensity on the y-axis (Costa, 2014; Alonso et al., 2015).

Two types of spectra are more commonly obtained with NMR. The 1D-NMR spectrum, shown in **Figure 2**, is the most used in metabolomics studies, while the two-dimensional NMR (2D-NMR) spectrum is only used for characterization of compounds that are impossible to identify through the 1D-NMR spectrum, since the second dimension of this spectrum allows to separate overlapping peaks. The spectral data obtained from NMR not only enables metabolite quantification, but also provides information on the chemical structure of the metabolites, allowing their identification. This last information is acquired by analysing the spectral peaks pattern, since each metabolite has a characteristic pattern (Costa, 2014; Alonso et al., 2015).

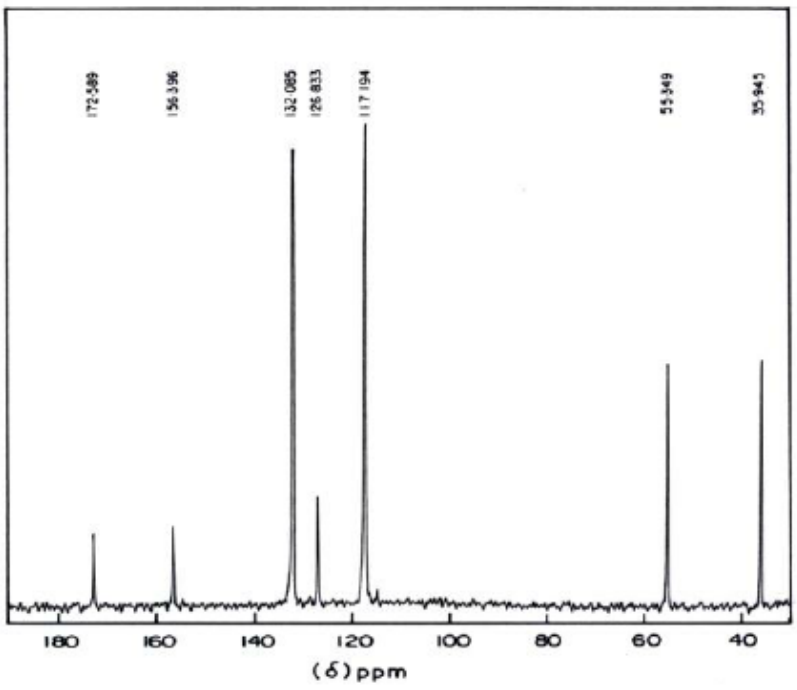


Figure 2: An example of an 1D-NMR spectra, adapted from Anandan et al. (2012).

MS has higher sensitivity than NMR, which enables a quantitative measure of a wider range of metabolites and identification of unknown and unexpected compounds, despite being slower and more complex than NMR. Furthermore, NMR is a non-destructive technique and highly reproducible, although it requires a relatively large number of samples and the equipment costs are much higher. Also, NMR metabolite quantification may be difficult, due to shifting peak positions, peak overlap and noise from the biological matrix. NMR has been widely applied to obtain information on metabolite profiles of complex biological mixtures, such as biological fluids and cells extracts (Costa, 2014; Alonso et al., 2015; Kosmides et al., 2013; Hao et al., 2012).

2.4 Data Pre-Processing and Pre-Treatment

Pre-processing and pre-treatment are two terms that can be considered different. The first one is often taken to include the techniques used when extracting the data or preparing it for analysis, and the latter to include the methods applied on the dataset that make samples analysable and comparable (Liland, 2011).

Nevertheless, the processing/treatment of NMR and GC/LC-MS data is normally achieved by treatment of missing values and outliers, peak spectra processing, normalization and scaling.

A missing value consists of an observation where a variable has no value. There are several ways to do the processing of missing values. One of these involves the removal of the variable or sample that contains the missing value, whilst the substitution of the variable with a certain value is another way. This "default" value may result from the data column or row mean, as well as from other methods that use information from the nearest values, such as K-nearest neighbours, which selects the most frequent value from the k nearest values or their mean. Outliers are data points that are distant from the other observations, whether due to variations in data measurement or experimental errors, and are usually excluded from the dataset (Batista and Monard, 2002).

Processing spectral data from GC/LC-MS and NMR techniques consists of peak alignment, baseline, offset and background corrections and smoothing.

Normally, spectral peaks can suffer dislocations through the x-axis, which can be caused by changes in the chemical environment of the sample like ionic strength, pH or protein content, in NMR, or in the chromatographic column, in the MS technique. Correcting these displacements before the data analysis is very important, so that the peaks appear where expected and it will be possible to study the multiple samples, specially to compare metabolic features between spectra (Alonso et al., 2015).

Peak Alignment methods can be separated into warping and segmenting methods. The first ones are based on the application of a non-linear transformation to the ppm, in NMR,

and retention time, in GC/LC-MS, axis, in order to increase the correlation between spectra. Then, the alignment is done through the enlarging or shortening of the spectra segments until the maximum correlation is achieved. Correlation Optimized Warping (COW) and Dynamic Time Warping (DTW) are two of these methods.

On the other hand, segmenting methods apply a constant shift to all spectra points, then aligning the overall spectra or splitting it and aligning all the resulting segments independently. The easiest way to correct the dislocations is by using shifting correction, which divides the spectrum into a certain number of local windows, moving the peaks to match along the spectrum. Although it is fast due to the fact that it is done locally, this method leads to some alignment errors, when peaks fall into the wrong local window, existing more robust methods, such as the aforementioned COW. Also, there is the Icoshift procedure, a newer and quicker method that uses fast Fourier transforms (Liland, 2011; Alonso et al., 2015).

Data correction, which can either be baseline, offset or background, aims to eliminate the effect of certain signal variations through the spectra and background noise during the sample analysis, caused by either experimental or instrumental variation, because most analysis techniques cannot distinguish between noise and signal (Alonso et al., 2015; Villas-Bôas et al., 2006). This correction is very important, for example, in metabolite identification, so that the analysis does not reveal metabolites that are not there (Liland, 2011). Lastly, smoothing consists in the reduction of the data random noise, which helps the robustness of the analysis and visual interpretation, especially in cases where the ratio signal-to-noise is high or the analysis methods to be used are sensitive to noise. There are three different methods to perform this technique: Savitzky-Golay, the most used, binning, useful when there are various measurements per spectrum and has the potential to correct small peak shifts, and loess (Varmuza and Filzmoser, 2010; Liland, 2011).

Sometimes, it is preferable to look at the relative differences between samples, instead of the absolute values. This allows to make different data comparable and consistent and, thus, allow a correct measurement of the features in the metabolomics analysis. This type of correction is known as normalization and, therefore, consists of row-wise transformation. The most common normalization method is the subtraction of the row mean to each data value and dividing by the standard deviation (Villas-Bôas et al., 2006; Alonso et al., 2015; Varmuza and Filzmoser, 2010).

Data mean-centering is a column-wise transformation and consists of subtracting the mean spectrum to each sample, so that all columns have zero mean. Mean-centering is focused on emphasizing the differences and not the similarities of the data, being applied with scaling methods (Varmuza and Filzmoser, 2010; van den Berg et al., 2006). Scaling consists of a column-wise transformation so that all have the same variance, by dividing each variable by a factor, the scaling factor. Scaling also allows samples with variables mea-

sured on different scales more comparable. Auto-scaling, used when comparing metabolites based on correlations is required, Pareto scaling, used to reduce the relative importance of large values, while keeping the data structure relatively intact, and Range scaling, which enables the comparison between different metabolites with regard to biological response range (van den Berg et al., 2006; Liland, 2011).

Finally, transformations, both logarithmic and power, are non-linear conversions of the data and are usually applied to correct, for example, heteroscedasticity, which is the existence of absolute noise that increases with the rising of the signal intensity. Whilst the logarithmic transformation does not deal with null values, the power transformation can, in addition to having effect on heteroscedasticity (Kvalheim et al., 1994; van den Berg et al., 2006).

2.5 Analysis of the processed data

To analyse metabolomics data, it is possible to do metabolite identification, univariate analysis, unsupervised and supervised multivariate analysis. Univariate analysis studies a data variable at a time, which is easy to perform and interpret. Multivariate analysis uses all or many metabolomics features simultaneously to identify relations between them. The unsupervised multivariate analysis is normally used to summarize data and thus detect patterns that can be related to biological variables or experimental ones. Supervised multivariate analysis consists of machine learning and feature selection (Alonso et al., 2015).

2.5.1 *Metabolite Identification*

Metabolite identification is done using NMR and GC/LC-MS data and is an essential step to give biological meaning to the study.

Metabolite identification of GC/LC-MS data starts with the detection of the existing peaks in the spectra, followed by discrimination of which peaks belong to the same source metabolite and, finally, each of these groups of peaks, which have a certain mass and were acquired under a certain chemical environment (ionization mode, for example), are compared to the peaks of each metabolite on a predefined database. These metabolites are acquired under similar conditions to those of the samples analysed. The identification is possible because each metabolite has a characteristic spectrum under different conditions.

The identification of metabolites from NMR data is achieved by matching the spectra of measured NMR peaks against the ones from each reference metabolite, acquired under similar conditions to those of the samples analysed. One of the approaches taken can also include a discrimination of which peaks belong to the same source metabolite before match-

ing each group of peaks obtained with the peaks of each one of the reference metabolites (Jacob et al., 2013).

The reference metabolites that reveal to have more peaks that better match to the ones in the samples, or to the groups of peaks obtained, are the identified metabolites.

The quantity, as well as quality, of the reference spectra in metabolite spectral databases is essential for a good metabolite identification (Alonso et al., 2015; Fernández-Albert et al., 2014).

There are several R packages that were developed to achieve metabolite identification. *MAIT* is an R package that allows metabolite identification of data from the LC-MS technique. It includes peak detection in the LC-MS samples files, peak annotation, which increases the biological and chemical information of the dataset and improves the metabolite identification, statistic analysis, revealing which samples features are statistically significant and their predictive power, and tables and graphics creation. The *MAIT* package uses the packages *xcms* and *CAMERA* to detect and align the peaks and the Human Metabolome Database (HMDB) as the database of metabolites to compare to (Fernández-Albert et al., 2014; Kuhl et al., 2012; Tautenhahn et al., 2012).

The Bayesian AuTomated Metabolite Analyser for NMR spectra, *batman*, is an R package that deconvolves the 1D-NMR spectrum peaks, automatically assigning them specific metabolites and estimating their concentration, with a reduced average estimation error when compared with other conventional numeric integration methods. The Bayesian model uses information about characteristic peak patterns of metabolites, which enables the recognition of overlapped signal(s), and has into account the shifts that may occur in the position of the peaks that normally are found in NMR spectra of biological samples. These characteristic peak patterns of metabolites are obtained through the database HMDB (Hao et al., 2012, 2014).

There are also web tools specialized in performing metabolite identification. MetaboHunter (`http://www.nrcbioinformatics.ca/metabohunter/`) is a web tool that automaticaly identifies metabolites, based on spectra or peak lists from the 1D-NMR technique, through three different methods (Tulpan et al., 2011). This web server uses manually curated data from the publicly available databases HMDB and Madison Metabolomics Consortium Database (MMCD). Bayesil is another web tool specialized in metabolite identification from the NMR technique. This tool focuses on identifying metabolites and their concentrations present in a human's biofluids, the metabolic profile, due to the fact that many diseases can cause changes in the metabolic profile (Ravanbakhsh et al., 2015).

This and other tools used for metabolite identification from MS and NMR techniques are listed in **Table 2**

Table 2: Metabolite identification tools.

Tool	Type	Spectral Data	URL
MAIT	R	GC-MS	https://www.bioconductor.org/packages/release/bioc/html/MAIT.html
BATMAN	R	1D-NMR	http://batman.r-forge.r-project.org/
rNMr	R	NMR	http://rnmr.nmrfam.wisc.edu/
MetaboHunter	Web tool	1D-NMR	http://www.nrcbioinformatics.ca/metabohunter/
Bayesil	Web tool	NMR	http://bayesil.ca/
MetaboMiner	Software	2D-NMR	http://wishart.biology.ualberta.ca/metabominer/index.html

2.5.2 *Machine Learning*

Supervised multivariate analysis, or Machine Learning, consists in the creation of predictive models that predict a determined output from a certain data input. From certain examples, normally part of the given data, that lead to a particular known output, it is possible to form a general model that not only can predict the given examples, but also other data whose outputs are unknown, in an accurate way. There are two approaches to reach a predictive model. The first one, called classification, involves a discrete output variable, i.e., the models seek to predict to which category a certain sample belongs, from a set of output categories/classes possible. The second one, regression, involves a numerical output variable, where the model will be, typically, a set of mathematical expressions that lead to the output value (Rocha et al., 2008).

So that the model can be evaluated, it is necessary to use error metrics. The simplest one is accuracy, which is the proportion of samples correctly classified, while the Kappa Statistic compares the observed accuracy with the expected, determining how close the classification from the model is to real classification. The metric Area Under the ROC Curve (AUC) measures the performance of a two classes classifier. These metrics are used in a classification problem. With regard to regression problems, there are essentially two error metrics: Root Mean Square Error (RMSE) and coefficient of determination (R^2) (Rocha et al., 2008).

The error metrics are then used in model validation methods, which evaluate the model performance based on data, test examples, whose output is already known, but not used to build the model. The holdout method, the most popular, divides the available data into these two parts. The k-fold cross validation method also divides the data this way, but k different times, where test sets are mutually exclusive. In the resampling method, train examples are randomly chosen, allowing the existence of data repetitions, and test examples are those that are not selected as train examples. Model validation is crucial to

choose the best model to utilize, out of a set of selected models, having to pay attention that, nevertheless, the model with the least error may not be the best one (Rocha et al., 2008). The pipeline of machine learning is elucidated in **Figure 3**.

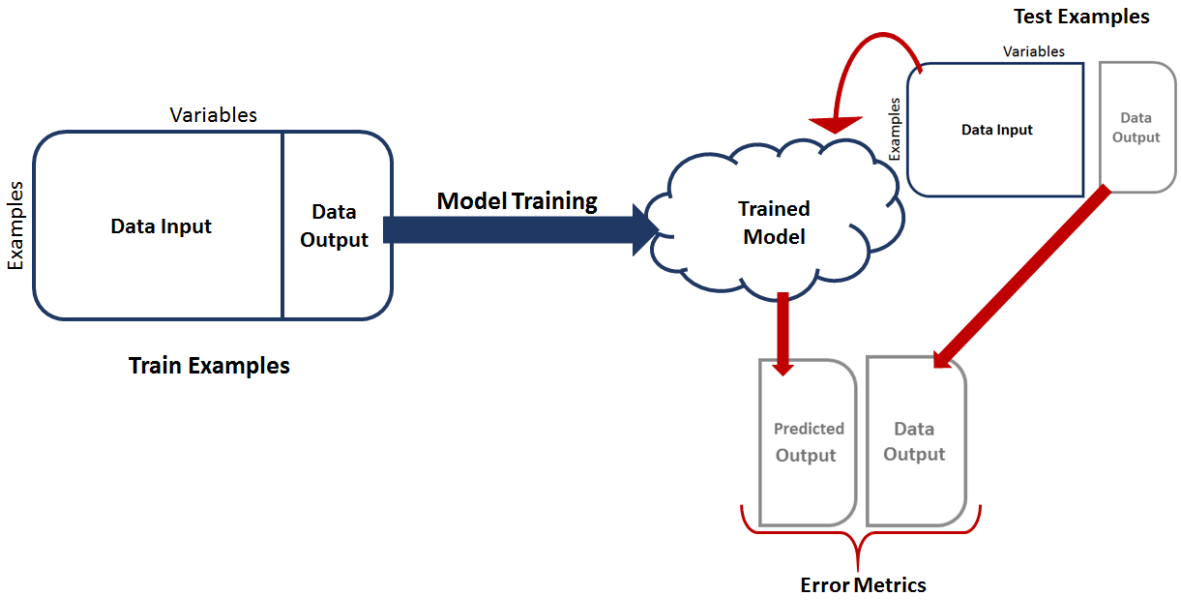


Figure 3: Machine Learning Pipeline.

About the actual models, trees and functional models can be applied in both classification and regression problems. In trees, each example is predicted by a downward path through the tree, starting at the root node and ending in one of the leaves, whose value is the output prediction for that example. Each tree node is a variable and its possible values are the branches that leave the node. In classification problems, the model is called decision tree, while in regression problems is named regression tree (Rocha et al., 2008).

Random forests consist of a set of decision trees whose nodes are chosen randomly (Liland, 2011). Similar to tree models is Rule-Base Classifier, which is a model with a set of classification rules, where each rule leads to one of the possible output values. The output of an example to classify depends on what rule or rules the example follows (Rocha et al., 2008).

Functional models attempt to model a problem through a function, whose arguments are the input values and the output variable is the one intended to be predicted. Some of these models, such as Linear Discriminant Analysis (LDA), Partial Least Squares Discriminant Analysis (PLS-DA) for classification problems and Partial Least Squares Regression (PLS-r) for regression problems, utilize a mathematical formula known *a priori* that is adjusted to better predict data.

Other models are more complex and general, with non-linear components, capable of modelling any type of function, such as Neural Networks (NN) and Support Vector Ma-

chines (SVMs). The first one tries to mimic the learning process of the brain, by having an initial layer that has as input the examples, an hidden layer that does calculations that lead to the output, the third layer. The latter, SVMs, have vectors that are the examples considered to be on the margin between two possible outputs and are used to distinguish the outputs (Liland, 2011).

There are also many other models, such as K-Nearest Neighbours, where a predicted value corresponds to the most common value on a set of k nearest examples to the example to predict, from a dataset whose outputs are known; and Naïve-Bayes, that estimates the probability of each possible output value for a certain example, according to what is known by a set of examples whose outputs are known (Costa, 2014; Rocha et al., 2008; Liland, 2011).

2.5.3 *Feature Selection*

Many of the input variables have null impact (or even negative) on the output to predict. Therefore, feature selection seeks to select the set of the most important input variables, allowing the model learning process to be easier, improving its prediction performance, robustness and generalization. Thus, the models derived from the selected variables can be faster and more cost-effective (Rocha et al., 2008; Guyon and Elisseeff, 2003).

In addition to the potential benefits from using feature selection listed above, the data visualization and comprehension can be facilitated, as well as the measurement and storage requirements and training and utilization time reduced (Guyon and Elisseeff, 2003).

The feature selection algorithms can be divided into two main groups. Wrapper Algorithms use the prediction performance of a given model to evaluate the usefulness of a previous selected subset of variables and compare it to other subsets of variables created.

There are, then, three main questions to answer when developing a wrapper algorithm. Firstly, how the search of the possible variable subsets to consider should be done. For this, an exhaustive search could be used. However, for a large number of variables, it is not the best method. There are more efficient methods, such as genetic algorithms, which use populations of solutions that evolve during a set of generations; forward selection, which begins with one or few features and adds iteratively more important variables; and backward selection, which begins with all variables and removes less important variables iteratively.

Secondly, what model validation methods should be used to access the model performance, by estimating error measures, in order to evaluate the subsets of variables. Lastly, what learning model should be used. The most popular include decision trees, naïve Bayes, PLS and SVMs (Guyon and Elisseeff, 2003).

Filter Algorithms, on the other hand, do not depend on a model, which can be seen as a pre-processing method. The best subset of variables is selected through the evaluation of the variables in the dataset, using univariate statistics, such as correlation between the input variables and the variable to predict, or quality information gain, leaving behind the least interesting variables. Then, the selected subset of variables is used in a selected model and its prediction performance is assessed (Rocha et al., 2008; Guyon and Elisseeff, 2003; Costa, 2014).

Sometimes, these two strategies are used together.

2.6 METABOLOMICS DATABASES

So that is possible to perform analyses such as metabolite identification, pathway analysis, enrichment analysis and biomarker identification, it is important that researchers have an easy access to information such as metabolite characteristics, metabolic networks, metabolite spectra and datasets of metabolomics experiments. This information has been easier to access due to the increasing number of open and web accessible databases, some mentioned in **Table 3**.

The Human Metabolome Database (HMDB), `http://www.hmdb.ca`, was first introduced in 2007 and it is now the world's largest and most comprehensive web-accessible organism-specific metabolomics database. HMDB is designed for biochemists, clinical chemists, physicians, medical geneticists, members of the metabolomics community and general education (Wishart et al., 2007, 2013).

Each metabolite entry in the database has many separate data fields, including compound description, names and synonyms, structural information, physical, chemical and biological data, such as biofluid concentrations, disease associations and pathway information. It also contains NMR and MS spectra from many reference metabolites. Furthermore, it can also provide enzyme data, gene sequence data, Single Nucleotide Polymorphism (SNP) and mutation data, and links to images, references and other public databases (Wishart et al., 2007, 2009).

Metabolights, `http://www.ebi.ac.uk/metabolights/`, was introduced in 2012 and it is an open-access repository for metabolomics studies, providing not only the raw experimental data, but also the respective metadata. This repository is cross-species and cross-technique, such as NMR spectroscopy and MS. Lastly, it also comprises metabolite structures, their reference spectra and biological roles, location and concentrations (Haug et al., 2013). The total number of studies made public arises to more than 200.

Mery-b, `http://services.cbib.u-bordeaux.fr/MERYB/`, was first introduced in 2009 and it is a web-accessible repository for metabolomics studies, metabolites information and spectra of 17 different plants, obtained by the 1D-NMR technique. The number of spectra

Table 3: Databases of metabolomics data.

Database	Description	Type[1]	URL
BioCyc	Freely accessible data collection of metabolic pathways and genomic information.	MPath	https://biocyc.org/
ChEBI	Freely accessible data collection of small molecules, such as metabolites.	MInf	http://www.ebi.ac.uk/chebi/
CHEMBL	Freely accessible data collection of manually curated drug-like bioactive compounds.	MInf	https://www.ebi.ac.uk/chembl/
ChemSpider	Freely accessible data collection of chemical compounds, not only metabolites.	MInf	http://www.chemspider.com/
DROP met	Being part of the web technology Prime, it allows the download of plant datasets and metadata from the MS technique.	MDat	http://prime.psc.riken.jp/?action=drop_index
HMDB	Freely accessible data collection of human metabolites.	MInf, MPath, MSpec	http://www.hmdb.ca/
KEGG	Freely accessible data collection of genes, genomes, genes products, metabolic and regulatory pathways from various species.	MPath	http://www.genome.jp/kegg/
MassBank	Freely accessible data collection of MS pure metabolite spectra.	MSpec	http://www.massbank.jp/
MassBase 1.0	Freely accessible data collection of metabolomics plant datasets from the MS technique for download.	MDat	http://webs2.kazusa.or.jp/massbase/index.php/
Mery-b	Web-accessible repository for metabolomics studies, metabolites information and spectra of plants, obtained by the 1D-NMR technique	MInf, MSpec, MDat	http://services.cbib.u-bordeaux.fr/MERYB/
MetaboLights	Freely accessible data collection ofmetabolomics datasets and metadata from the techniques MS and NMR for download, and metabolite information derived from those experiences.	MInf, MSpec, MDat	http://www.ebi.ac.uk/metabolights/
Metabolome Express	Freely accessible data collection of metabolomics datasets, centered in plants, from the MS technique.	MDat	https://www.metabolome-express.org/
Metabolomics Workbench	Public repository for metabolomics experimental data from various species and from the NMR and MS techniques, and metabolite information.	MInf, MDat	http://www.metabolomicsworkbench.org/
METLIN	Freely accessible data collection of metabolite information and MS/MS spectra, centered in humans.	MInf, MSpec	https://metlin.scripps.edu/
MMCD	Freely accessible data collection of NMR metabolite spectra.	MSpec	http://mmcd.nmrfam.wisc.edu/
PubChem	Freely accessible data collection of general information about chemical compounds.	MInf	https://pubchem.ncbi.nlm.nih.gov/
Reactome	Freely accessible data collection of previously curated metabolic pathways.	MPath	http://www.reactome.org/

[1]The column *Type* covers the types of metabolite information that the databases may provide: MInf-Metabolite Information, MPath-Metabolic Pathways, MSpec-Metabolite Spectra, MDat-Datasets from metabolomics experiments

publicly available accounts to more than 1900, with 34 studies made public (Ferry-Dumazet et al., 2011).

Metabolomics Workbench, http://www.metabolomicsworkbench.org/, is a public repository for metabolomics experimental data, and respective metadata, from various species and from the NMR and MS techniques. This database also provides information on more than 60 000 metabolites, regarding structure, physicochemical properties, taxonomy and database links (Sud et al., 2016).

The Chemical Entities of Biological Interest (ChEBI) database, present in http://www.ebi.ac.uk/chebi/, was first introduced in 2007 and provides information about natural or synthetic small molecules, such as metabolites, that intervene in the living organisms processes. This information includes molecules' 2D and 3D structure, definition, molecular formula, onthology, synonyms, links for other databases, among others (Degtyarenko et al., 2008; Hastings et al., 2013).

Another important database is Pubchem, https://pubchem.ncbi.nlm.nih.gov/, which was first introduced in 2004 as a component of the Molecular Libraries Roadmap Initiatives of the US National Institutes of Health (NIH) and provides information about chemical substances and their biological activities. This database consists on three inter-linked databases: Substance, which contains chemical information; Compound, with the chemical structures; and BioAssay, with the biological activity of the chemical substances (Kim et al., 2015).

One last database worth mentioning is the Kyoto Encyclopedia of Genes and Genomes (KEGG), a database project initiated in 1995 under the Human Genome Program of the Ministry of Education, Science, Sports and Culture in Japan that comprises information from three different databases, present in http://www.genome.jp/kegg/. The Genes database is a collection of gene catalogues for completely sequenced and partial genomes. The Ligand database consists in a collection of chemical compounds in the cell, enzyme molecules and enzymatic reactions. Lastly, the Pathway database contains graphical representations of cellular processes, such as metabolism, membrane transport, signal transduction and cell cycle. The metabolic pathways part is the best organized one of the Pathway database, with approximately 90 graphical diagrams of reference metabolic patahways (Ogata et al., 1999).

2.7 WEB TOOLS TO ANALYSE METABOLOMICS DATA

Web tools that allow the processing and analysis of metabolomics data facilitate the comprehension and extraction of knowledge from metabolomics data, without enforcing the user to have programming skills, as many tools currently available that help in this analysis require these skills. There are many web tools that have been created throughout the years, and some of them are reviewed in the **Table 4**. The website MetaboAnalyst3 is the most remarkable one.

Table 4: Web tools for metabolomics data analysis.

Web tool	Covered Techniques	Features[1]	URL
Bayesil	NMR	MId	http://bayesil.ca/
IMPaLA	-	PA, EA	http://impala.molgen.mpg.de/
MassTrix	-	PA	http://masstrix3.helmholtz-muenchen.de/
MetaboAnalyst 3.0	NMR, MS	MId, ML, FS, PA, EA, BI	http://www.metaboanalyst.ca/
MetExplore	-	PA, L	http://metexplore.toulouse.inra.fr/
MetaboHunter	NMR	MId	http://www.nrcbioinformatics.ca/metabohunter/
Metabolites Biological Role (MBRole2.0)	-	EA	http://csbg.cnb.csic.es/mbrole2/
Metabolome Express	GC-MS	MId, L	https://www.metabolome-express.org/
Metabolomics Workbench	NMR, MS	ML, FS, PA, L	http://www.metabolomicsworkbench.org/
MetDAT	MS	MId, ML, PA, L	https://smbl.nus.edu.sg/METDAT2/
Metldb2.0	GC-MS	MId, ML, EA, L	https://meltdb.cebitec.uni-bielefeld.de/
Paintomics	-	PA, L	http://bioinfo.cipf.es/paintomics/
XCMX online	LC-MS	MId, PA, L	https://xcmsonline.scripps.edu/

[1]The column *Features* covers the features that the web tools may have: MId-Metabolite Identification, ML-Machine Learning, FS-Feature Selection, PA-Pathway Analysis, EA-Enrichment Analysis, BI-Biomarker Identification, L-Login

MetaboAnalyst, http://www.metaboanalyst.ca/, was first introduced in 2009 and it allows now the comprehensive analysis, visualization and interpretation of metabolomics data that come from NMR and MS techniques. This website provides various types of metabolomics analysis and processing. Besides providing various statistical analysis, it is also available enrichment analysis, pathway analysis, machine learning and feature selection, metabolite identification, biomarker analysis, among others (Xia et al., 2009, 2012, 2015).

However, some of these analyses may be very limited in terms of methods diversity. For instance, regarding machine learning, it is only possible to perform the PLS-DA, random forests and SVMs algorithms. Furthermore, classification of new samples using the models created is not available, as well as the impossibility to perform feature selection separately from the model creation, as the model creation from each algorithm has a specific feature selection method associated.

XCMS online is another web tool, `https://xcmsonline.scripps.edu/`, providing a simple and user friendly web based version of the XCMS software. This tool provides analysis for metabolomics data that come from the GC/LC-MS technique, such as statistical analysis and data visualization (Tautenhahn et al., 2012).

There are other web tools that provide metabolomics data processing and analysis, and some of them are reviewed in the **Table 4**.

3

DEVELOPMENT

In this chapter, the details about the website and its development process will be covered, as well as the technologies used in this process. Furthermore, improvements done to the *specmine* package, used as the core package for the development of the website, will also be covered.

3.1 *specmine* PACKAGE

The *specmine* package was created for the R environment by the host research group and has functions that allow pre-processing and analysis of metabolomic data from techniques such as GC/LC-MS, NMR, IR and UV-Vis (Costa et al., 2016). This is the package that constitutes the base for the development of the website proposed in this work.

The data that can be processed by this package must have one of these formats: CSV files for NMR and MS peaks lists, CSV or TSV files for metadata files and concentrations files, (J)DX spectra files, and NetCDF, mzDATA or mzXML for MS data. The dataset then formed is an R list with the following fields: description of the dataset, the data type, data matrix, the metadata data frame and the x and y axis labels, as shown in **Figure 4**.

To process MS spectral data, baseline, offset and background corrections, and peak alignment are available.

Missing values can be replaced by values that are calculated through K-nearest neighbours, linear approximation, mean or median, or that are given by the user. They can also be eliminated by removing the sample or the variable.

Normalization can be achieved by the sum, median, sample reference or feature reference. Four different approaches for scaling are provided: pareto, auto, range and interval scaling. Mean-centering and cubic root or logarithmic transformations can also be done.

Finally, it is possible to remove variables with low variance using flat pattern filters, using methods such as interquantile range, relative standard deviation, standard deviation, median absolute deviation, median and mean. The removal of the low variance variables is achieved through a chosen percentage or a given treshold value. It is also possible to remove specific samples and data and metadata variables, as well as remove samples and

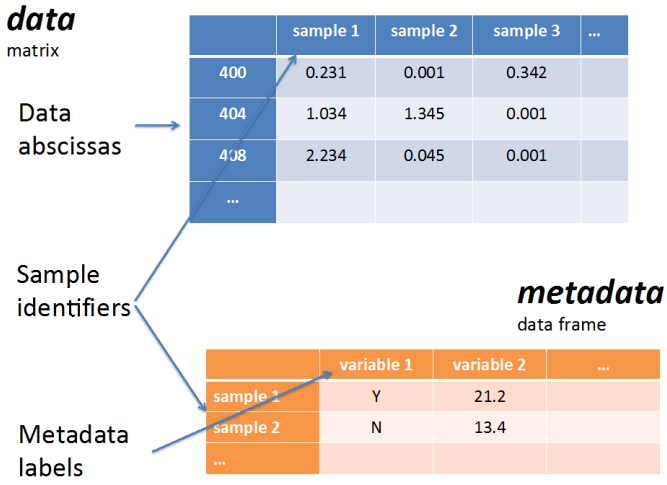


Figure 4: Representation of the data structure in a dataset formed in the *specmine* package (Costa et al., 2016).

data variables according to the amount of missing values present in the samples or data variables, respectively, or according to the amount of missing values in the metadata.

Metabolite identification was made only available for LC-MS data, using the *mait* package.

For supervised multivariate analysis, there are functions to train models and predict samples separately, but also one that does these two things. These functions are based on the *caret* R package. There are many validation methods available, such as k-fold cross validation, leave-one-out cross validation, resampling, and others. The error metrics available are AUC, kappa statistics, and others.

For feature selection, it is possible to perform filter and wrapper selections, separately from the machine learning functions.

There are also functions that allow the execution of univariate analysis, such as t-test, one-way and mutifactor ANOVA, Kruskal-Wallis and Komolgorov-Smirnov tests, fold change analysis, and regression and correlation analysis; and unsupervised multivariate analysis, such as PCA, and clustering analysis.

Finally, there are functions that allow the user to update dataset fields, do various different plots, among others.

The *specmine* functions that allow the MS and NMR data pre-processing and analysis listed above and others are present in **Table 5**.

Many were the applications that the *specmine* package was used for. Cachexia, characterized by loss of muscle with or without loss of fat mass, is a complex metabolic syndrome. Therefore, the metabolites produced from tissue breakdown may be a good indicator for this disease. The package was used to analyse urine samples, as many end products of

Table 5: Specmine functions that allow MS and NMR data pre-processing and analysis.

		Specmine Function
Data Reading	Read folder with peak list files	*read_csvs_folder*
	Read metabolite concentrations file	*read_dataset_csv*
	Read MS spectra File	*read_ms_spectra*
	Read metadata File	*read_metadata*
Pre-processing spectral data	Baseline, offset and background correction	*data_correction*
	Peak alignment	*group_peaks*
	Smoothing (Savitzky-Golay; binning; and loess)	*smoothing_interpolation*
Other Pre-processing	Missing values	*missingValues_imputation*
	Normalization	*normalize*
	Mean-centering	*mean_centering*
	Scaling (auto; pareto; and range)	*scaling*
	Logarithmic and cubic root transformations	*transform_data*
	Flat pattern filters	*flat_pattern_filter*
	Create subset of a dataset by data variables interval	*subset_x_values_by_interval*
	Remove specific samples,	*remove_samples*
	data or metadata	*remove_data_variables*
	variables	*remove_metadata_variables*
	Remove samples or data	*remove_samples_by_nas*
	variables according to	*remove_variables_by_nas*
	ammount of missing values	*remove_samples_by_na_metadata*
	Aggregate samples	*aggregate_samples*
MS Metabolite identification		*MAIT_identify_metabolites*
Machine learning	Train models	*train_models_performance*
	Predict samples	*predict_samples*
	Train models and predict samples	*train_and_predict*
Feature selection		*feature_selection*
Univariate analysis	T-tests	*ttests_dataset*
	One-way ANOVA	*aov_all_vars*
	Multifactor ANOVA	*multifactor_aov_all_vars*
	Kruskal-Wallis tests	*kruskalTest_dataset*
	Kolmogorov-Smirnov tests	*ksTest_dataset*
	Fold change	*fold_change*
	Regression analysis on one variable	*line_regression_onevar*
	Regression analysis on two or more variables	*linreg_all_vars*
	Correlation analysis	*correlations_dataset*
PCA	Classical PCA	*pca_analysis_dataset*
	Robust PCA	*pca_robust*
Clustering analysis	Hierarchical	*hierarchical_clustering*
	K-means	*kmeans_clustering*

muscle catabolism are excreted in urine, from 77 patients, 47 of whom had cachexia and 30 were control patients. These data were obtained from 1D-NMR spectra of the urine samples, and then the different metabolites were detected and each concentration measured. The obtained data were firstly proccessed through logarithmic transformation and auto scaling. Besides univariate analysis and unsupervised multivariate analysis being made, machine learning was also performed. In the last analysis, 5 different models were trained and the algorithm PLS was revealed to be the one that better predicted the data according to the accuracy metric (Costa et al., 2016; Costa, 2014).

Propolis, produced by bees from the collected exudates of plants, has been highlighted with pharmacological properties by recent studies, such as antimicobial, anti-oxidative, anti-viral, anti-tumoral or anti-inflammatory. As the chemical composition of propolis may be strongly influenced by environmental factors and seasoning, the *specmine* package was used to analyse 59 samples of NMR peak lists data, collected from propolis produced in the Santa Catarina state, southern Brazil, in the four different seasons of the 2010 year. Data regarding the agroecological region of the collected samples was further retained. The obtained data were firstly pre-proccessed. Peak alignment was made, followed by missing values treatment, logarithmic transformation and auto scaling, besides univariate analysis and unsupervised multivariate analysis being made, machine learning was also performed, to train models to be capable of discriminating propolis samples by seasons and by agroecological regions (Costa et al., 2016; Maraschin et al., 2016).

3.2 IMPROVEMENTS TO THE *specmine* PACKAGE

The first improvement performed on the package regards the metabolite identification for NMR peak data, as the package only provided this type of analysis for Liquid Chromatography-Mass Spectrometry (LC-MS) spectral data.

The overall pipeline for the NMR metabolite identification starts with the clustering of the peaks in the dataset according to a correlation. After clustering, the peaks are separated in the respective clusters based on a minimum correlation that each peak inside a cluster must have with the others on the same cluster. The value of this correlation can be set by the user or calculated, where the optimal value is the one that leads to the larger number of clusters. The code developed for the clustering of peaks made use of the R code proposed in Jacob et al. (2013). Furthermore, when calculating the optimal correlation value, this code setted 40 as the maximum number of peaks in each cluster. However, the code was adapted so that people could choose the value wanted or choose it to be the number of peaks of the larger reference metabolite, but by still having 40 as the default value. To perform the clustering steps, the R package *igraph* is required. Each of these clusters is considered a

potential metabolite, as it can be assumed that peaks coming from the same molecule show similar behaviour across all samples and, therefore, correlate strongly with each other.

After setting the library of the reference metabolites, each cluster is compared with each reference metabolite, using the Jaccard index to score the match. This index is used to compare the similarity between sets, as it is defined by the division of the size of intersection by the size of the union of the sets: $J(A, B) = |A \cap B| \div |A \cup B|$.

The main function that performs the identification, by calling the other functions developed, is named *nmr_identification* and returns the results for the top reference metabolites with the best score, for each cluster formed. The default number of metabolites in each set is 5, but this parameter can be changed. These results come in the form of a list, with the following items for each cluster obtained:

- *cluster.peaks*: the peaks of the cluster, with respective intensities;
- *summary*: scores (Jaccard index) of each top reference metabolites matched with the cluster;
- *metabolites.matched*: list with full results for each top reference metabolite matched;
  - *score*: Jaccard index score;
  - *matched_peaks_ref*: matched peaks of the reference spectra;
  - *matched_peaks_clust*: matched peaks of the cluster;
  - *reference_peaks*: all the reference spectra peaks

The names of the functions in charge of performing the different identification steps mentioned are listed in **Table 6**.

To construct the library of reference metabolites, the XML spectra files from the Human Metabolome Database (HMDB), version 3.6, mentioned previously in the **section 2.6**, were downloaded and parsed. Two data variables were created. One, called *nmr_1d_spectra*, is the list with all the 1D-NMR spectra peaks provided. Each spectrum is a dataframe with the values of the peaks, in ppm, and the respective intensities. Each of these spectra has certain characteristics associated, with which the library is filtered to generate the set of reference metabolites that are used in the metabolite identification. These characteristics are stored in the other variable, a dataframe named *nmr_1d_spectra_options*. Here, each line corresponds to a spectrum and each column to a characteristic. These characteristics are the frequency of the spectra, the atomic nuclei, and the solvent, pH and temperature of the samples used. Only the NMR spectra that were obtained using frequencies greater or equal to 400 MHz were considered, as the validity of metabolomics experiments using smaller frequencies have been questioned, given the lower capacity for detection of the nuclei in question.

Table 6: Table summarizing the new functions added that help in the metabolite identification from NMR peaks data.

Function	Description
choose_nmr_references	Returns the reference spectra, according to the characteristics chosen (frequency, nucleus, solvent, pH and temperature). Only the frequency and nucleus characteristics are mandatory.
find_corr	Takes a dataset and calculates the optimum correlation value, which leads to the maximum number of clusters.
nmr_clustering	Takes a dataset and performs clustering of variables, according to a correlation. The variables will be separated into different clusters, according to a minimum correlation between variables and the minimum number of peaks wanted for each cluster. Each cluster will correspond to a metabolite.
jaccard_index	Calculates the Jaccard index, i.e., the similarity between cluster and reference metabolite. A ppm tolerance can be chosen.
nmr_identification	This function performs metabolite identification on a dataset of NMR peaks, finding the top reference metabolites that matched with each cluster of variables formed.

The function developed to filter the library according to the characteristics wanted is named *choose_nmr_references*, as seen in **Table 6**, and it is mandatory to choose the frequency and atomic nuclei, while the other characteristics may not be defined.

To integrate the new functions into the *specmine* package, the functions were stored in a R file named *NMR_metabolite_identification*, saved in the *R* folder of the package. Furthermore, the constructed data, *nmr_1d_spectra* and *nmr_1d_spectra_options*, were stored in the *data* folder of the package, as RDA files. This was acomplished by using the *devtools* package. This package was also required, alongside with the *roxygen2* package, to document the newly created functions and create the respective Rd files, stored in the *man* folder of the package, for each function. Due to the clustering functions, the package now needs to import the *igraph* package too.

Further improvements were done regarding the function used to train models, specified in **Table 5**. This function now allows the user to give the set of parameter values to be tested in the parameter optimization of the respective models trained.

This new version of *specmine* is the one used in the development of the created website and can only be installed for now and loaded in R from the bitbucket repository, with the following commands:

```
library(devtools)
install_bitbucket("chrisbcl\metabolomicspackage")
library(specmine)
```

However, the new version of this package will be released in CRAN, where the previous version is already available.

3.3 WEBSITE DEVELOPMENT STRATEGIES AND TOOLS

The website was developed collaboratively with Afonso (2017). With this, although the data that can be analysed includes data from NMR, MS, infrared, Raman and UV-Vis spectroscopies or concentrations data, the developed work refers to data from NMR and MS techniques, and concentrations data. Furthermore, as regards to the analyses provided, the ones developed in this work were metabolite identification, machine learning, including model training and samples prediction, and feature selection. Finally, the website layout and other functionalities, such as visualization of the data being analysed, pre-processing and workspaces were also developed in this work. Therefore, these will be the main focus on this chapter.

To achieve the proposed goals regarding the creation and development of the website, the *shiny* package for the R environment, more specifically the RStudio, was used (`https://shiny.rstudio.com/`). *Shiny* was chosen due to the fact that, besides allowing building interactive web applications, it is a package developed for the R environment, like the *specmine* package, which facilitates the usage of this package for the construction of the target application.

Applications developed using *shiny* have two components, the user-interface script (ui.R) and the server script (server.R). The user-interface script is in charge of controling the layout and appearance of the application, and the server script controls what happens in the website, as it has the instructions needed to build the app. In this work, the server script was divided into several R files, that were then called in the main server file, to facilitate the construction of the website and the reading/organization of the code. Each of these files was in charge of a specific set of instructions. For example, code related to each type of analysis was written in their respective separate files. A scheme with all the files used to develop the website is present in **Figure 5**.

Another useful aspect about *shiny* is that it comes with reactive programming, where the different reactive values can change over time or in response to the user through the reactive expressions, which can access other reactive values and reactive expressions, leading to a propagation of the changes over the different values. This aspect simplifies the creation of interactive user interfaces.

Furthermore, *shiny* allows the usage of Javascript and CSS languages to style the app in the wanted way. This was used in the development of the website, when needed, to achieve a better look for the website. The CSS files created, as well as the images used in the website were stored in the folder "www", seen in **Figure 5**.

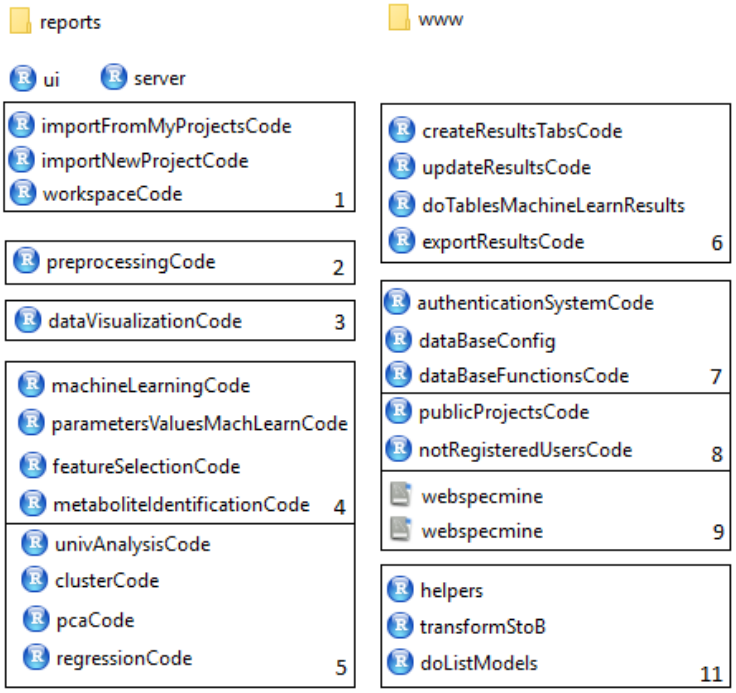


Figure 5: Scheme of the files used to develop the website. The "reports" folder stores the different *RMarkdown* files used as basis to create the different reports that can be saved/downloaded. The "www" folder stores the CSS files created to customize the website, as well as other images used in the website. The main files are "ui.R" and "server.R". The latter calls the rest of the R files present in the scheme. Files numbered with 1 are in charge of submitting the wanted data to analysis, while the files numbered 2 and 3 have, as the names suggest, the code that is in charge of pre-processing and visualizing the data, respectively. The files numbered 4 and 5 include the code that does the different analysis provided by the website, while the files numbered 6 are in charge of the exposure of the obtained results in the website. Files numbered 7 and 8 are responsible for actions related to the database. The files numbered 11 include minor functions. Finally, the files numbered 9 are the SQL files developed to create the database.

There are some R packages, used in the development of the website, worth mentioning:

- *ShinyWidgets* package, which improved the appearance of some parts of the website, such as radio buttons, and allowed the creation of pop-up windows to inform the user that a particular task has been successfully done, like saving reports, for example;
- *shinyBS* package, that allowed the development of pop-up windows to allow the user to perform certain tasks, such as the submission of new projects, loading and saving workspaces, among others;
- *shinyjs* package, that allowed the website to hide/show certain parts of the website dinamically, according to the tasks performed by the user;

- *DT* package, which allows the construction of tables that can be scrolled down and right when the table exceeds the fixed size given to the table. While scrolling down, the names of the columns will remain in place, so that the user is still able to know what each column stands for while going through the table;

- *RMarkdown* package, which is used to create the basis for the reports of the results and data visualization, all in HTML format, as well as HTML files with results' tables. In the beggining of each report, after the main title, the user is provided with the information of when the report was created. All base reports are stored in the folder "reports" (**Figure 5**).

Finally, besides being able to obtain reports of the results and data visualization, the user can also obtain the different tables in the results in CSV and MS EXCEL formats, besides HTML format.

The website can be accessed through the url `http://darwin.di.uminho.pt:3838/webspecmine/` and the respective code is available at the following GitHub repository: `https://github.com/TelmaAfonso/webspecmine`.

3.4 WEBSITE ARCHITECTURE AND LAYOUT

The overall website arquitecture consists in providing means of analysing metabolomics data from the NMR, MS, infrared, Raman and UV-Vis spectroscopies, as well as allowing the sharing of metabolomics experimental data between users. The name chosen for the website is based on the name of the core package *specmine* where functionalities are implemented: *WebSpecmine*.

The website starts with a home page, where users can enter their user account or do the analysis without logging in, although some features will not be available in the last scenario, as shown in **Table 7**. These features would mainly consist on saving, into the account, experimental data, so it can be used later, reports and the current work (data and results), that the user could later return to and continue the analysis being made. These differences will be explained in further detail along this chapter.

Table 7: Table summarizing the main the differences and similarities between a logged out and logged in user.

Features	Logged in User	Logged out User
Store experimental data	✓	
Analyse metabolomics data	✓	✓
See Public Projects	✓	✓
Analyse public Projects	✓	
See results	✓	✓
Save reports	✓	
Save the current data and results	✓	

The R package *shinydasboard* was used to obtain the general appearance wanted for the site, which consists of a sidebar, header and main panels. The layout is shown in **Figure 6**.

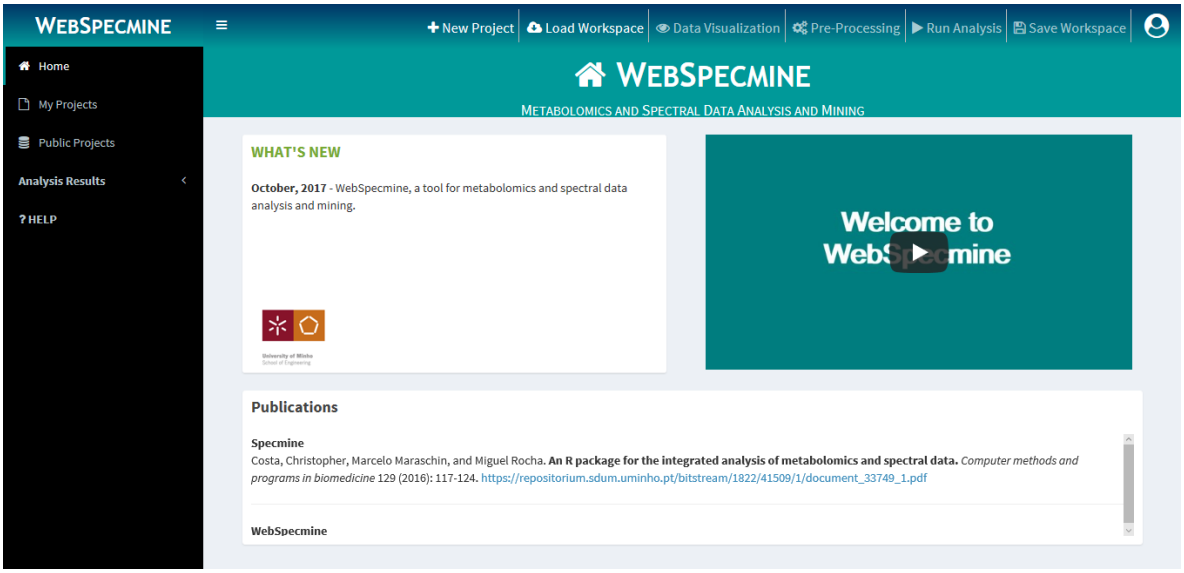


Figure 6: Layout of the WebSpecmine Analysis App.

The header panel contains the website name and a button that allows to show and hide the sidebar. Furthermore, in this panel, it can be seen the links to log in into the user account and to register a new account, or, in case the user is logged in, links to change the name and password, delete the account, and to log out. Also, all pages that carry out "actions" are made available through the header panel: *New Project* or *Choose Project*, *Load Workspace*, *Data Visualization*, *Pre-Processing*, *Run Analysis*, *Save Workspace*.

The sidebar panel has five tabs, named *Home*, *My Projects*, *Public Projects*, *Analysis Results* and *Help*. All these tabs lead to pages that show the respective information. The *My Projects* and *Public Projects* show the projects saved by the user into the account and the public projects made available, respectively. The main panel presents the content for each page, when selected.

The content and development of each page will be explained in detail along this chapter.

3.5 CHOOSE THE DATA TO WORK WITH

There are three different ways of choosing data for analysis. The "New Project" one, available for users not logged in, the "Choose Files", for logged in users, and "Load Workspace", for both types of mentioned users.

3.5.1 *New Project*

The "New Project" feature is available through the header panel, when the user is not logged in. When the user clicks the "New Project" button, a pop-up window to submit the files for analysis appears, like the one in **Figure** 7.

In this window, according to the the data type chosen on the top of the window, either "MS Spectra (.mzXML, .netCDF, mzData)", "NMR or MS peaks lists", "Concentrations" or "Spectral Data", the user has to set some options about the data and metadata files, present under the data type choices. Optional information can also be given.

Because this is a feature for logged out users, the user is here asked to submit the files for analysis. For MS Spectra and NMR or MS peaks lists, a .zip file with the data files and a metadata file must be submitted, while for concentrations data, a data file and metadata file is all what is needed.

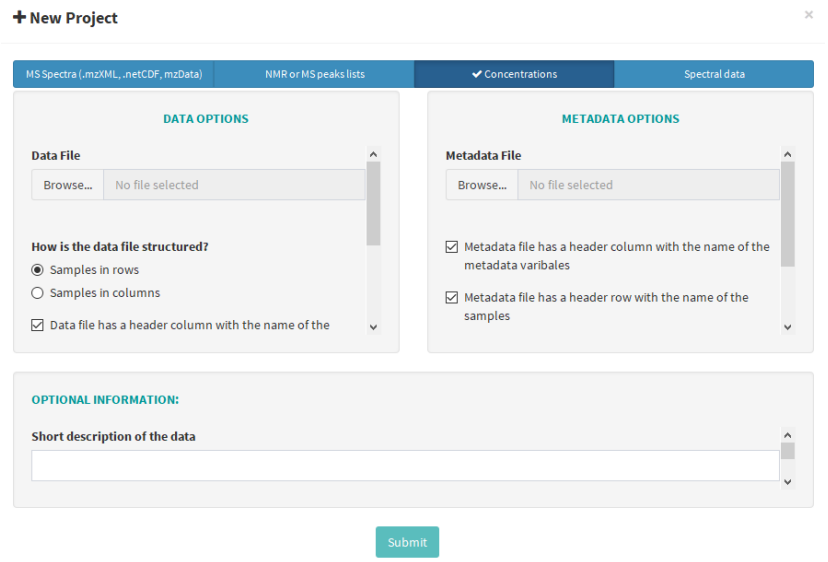


Figure 7: Layout of the submission of a new project of concentrations data.

The "Submit" button present at the bottom of the window is only enabled when both the data and metadata are submitted. After clicking the button "Submit", the files are processed and the data is stored in a reactive variable called "data", under the name "OriginalData", where all the variables related to the data will be stored. These variables consist not only in the original dataset submitted but also in the different datasets generated by the processing pipelines, explained later.

Then, the window will disappear and the page "Run Analysis" will appear. All the other buttons in the header panel will be made available, except for the "Save workspace" one, only available for logged in users. Also, the tab "Dataset being used" appears on the sidebar

panel, with one selected option, ”OriginalData”, which means that this is the dataset being currently used by the server. This feature will be further explained later.

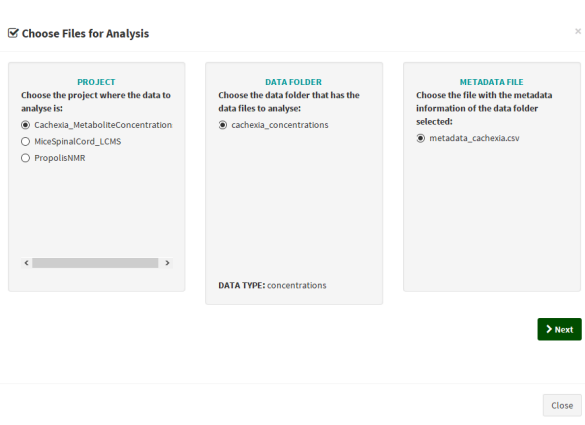
Finally, each time a new submission of files for analysis is done, all the work that the user may have done before is lost.

Whenever a user submits files for analysis this way, a temporary folder on the server side will be created, to store the data files, as it might be necessary to use them in certain data analysis. However, any time a user leaves the website or submits new files, the created data folder will be removed.

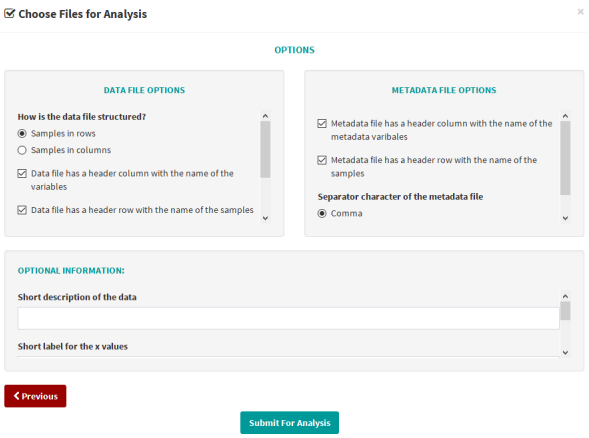
3.5.2 *Choose Files*

The ”Choose Files” feature is available through the header panel, only when the user is logged in. When the user clicks the ”Choose Files” button, a pop-up window to choose the files for analysis, from all the files present in the user's account, appears, like the one that can be observed in **Figure 8**.

These files are stored in projects. These projects are organized in a way so that, for each project, there are data folders stored in an overall *Data* folder, with each one of the data folders with data files that are used in an analysis; and a *Metadata* folder with metadata files, where each one can be used in an analysis. These projects can either be private, where only the user is able to access that project, or public, where not only the user accesses the project, but also all website users, with an account or not. When the user is logged in, he/she can copy a public project from the database to his account and use the project as the other private ones. For each project, the reports generated by the analysis of a certain data are stored in the *Reports* folder of the project in question.



(a) Initial panel to choose the files for analysis from the user's account.



(b) Layout of concentrations options in this feature

Figure 8: Layout of ”Choose Project” feature.

Further details on how the database and the public and private projects work are present in Afonso (2017).

Initially, three boxes appear on the window (**Figure 8a**), one to choose the project to work with, another one to choose the data folder from the chosen project that contains the data files to analyse, and another one to choose the metadata file from the chosen project that contains the metadata information about the data to analyse. The website was constructed so that only the projects that do not have empty Data and Metadata folders can be selected for analysis.

After doing so, the user will have to set some options regarding the data type in question (**Figure 8b**). After this, the user is able to submit the chosen files for analysis, by clicking the "Submit for Analysis" button.

Everytime a different project is chosen, the workspace regarding the previous project is lost, unless the user saves it first.

3.5.3 *MS Spectra Options*

The *specmine* functions used to read MS data are *read_ms_spectra* for data files and *read_metadata* for the metadata file.

The data options made available concern the feature (peak) detection in the chromatographic time domain. The user must choose the profile generation method ("bin", "binlin", "binlinbase" or "intlin"); the full width at half maximum (fwhm) of matched filtration gaussian model peak, commonly 30 for LC-MS spectra and 4 for GC-MS spectra; the bandwidth (standard deviation or half width at half maximum) of the Gaussian smoothing kernel, to apply to the peak density chromatogram, commonly 30 for LC-MS spectra and 5 for GC-MS spectra; and the peak intensity measure ("Integrated area of original (raw) peak", "Integrated area of filtered peak", "Maximum intensity of original (raw) peak" or "Maximum intensity of filtered peak").

The available options for the metadata file concern the way how this file is formatted. The user must say if the submited metadata file has a header column with the name of the metadata variables, if it has a header row with the name of the samples, and if it is a comma or white space that separates the data. The user can also provide, optionally, a short description of the data.

3.5.4 *NMR or MS peaks lists Options*

The *specmine* functions used to read peaks lists data are *read_csvs_folder* for data files and *read_metadata* for the metadata file.

The user must provide the type of data peaks submitted (either ”NMR”, ”GC-MS” or ”LC-MS” peaks) and some information regarding how the data files are formatted, such as the occurrence, or not, of a header row with the names of the data variables, the separator character, and the character used for decimal points.

Regarding the metadata file, the options are similar to those mentioned before. Optional information can also be given by the user, such as a short description of the data and short labels for the x and y values.

When this type of data is chosen, there is no ”Submit”/”Submit for Analysis” button initially, but a ”Next” button, which, in the ”New Project” feature, is enabled when both data and metadata files are submitted. When the user clicks next, a set of options to do the alignment of peaks, after processing the files, is provided. This alignment is done by using the *specmine* function *group_peaks*. The user is able to choose between the MetaboAnalyst and Specmine algorithms. When the specmine algorithm is chosen, the user must give the size of the step, in ppms. On the other hand, when the MetaboAnalyst method is chosen, the metadata variable to be used can be chosen. After this, the user is able to press the ”Submit”/”Submit for Analysis” button, so that the data is further processed and available to analysis through the website.

3.5.5 *Concentrations Options*

The *specmine* functions used to read concentrations data are *read_dataset_csv* for data files and *read_metadata* for the metadata file.

The data options made available concern the way how the file is formated. The user must say if the samples are distributed over the rows or columns. According to what he responds to this option, the user must say if the file has a header column with the names of the variables or samples and a header row with the names of the samples or variables. If the user says that the file does not have the samples names, he will be asked to give them, by writing them, separating each one by a comma. Finally, the user must specify the data separator character (”Comma” or ”White space”).

Regarding the metadata file, the options are similar to those mentioned before for metadata files. Moreover, the user is able to provide, optionally, a short description of the data and short labels for the x and y values.

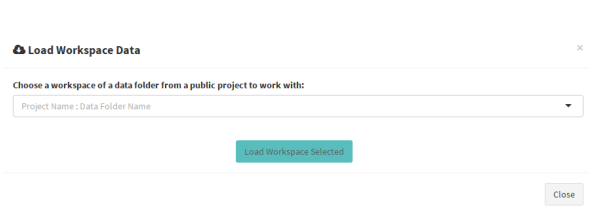
3.5.6 *Load and Save Workspaces*

First of all, a workspace consists on all data the user is working on at the moment and the possible results that have already been obtained. The website code was developed so that each data folder from a project can have a workspace associated. Therefore, only users with

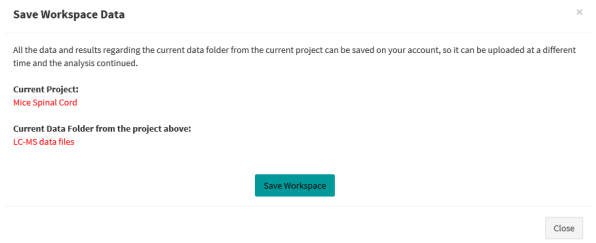
an account can save workspaces. These workspaces are stored in two Rdata files, one will store the ”data” reactive variable and the other one the ”results” variable. These files are stored in the respective project's folder of the user.

On the other hand, not only users that hold an account can load a workspace, but also users with no account. The difference remains in the fact that logged in users upload workspaces from their account or from public projects, even if the project to where they belong was not yet copied to the account, whilst logged out users can only upload public workspaces. After loading a public workspace, the logged in user has to copy the public project associated with the public workspace loaded to be alowed to save the workspace, as a warning message appears in the pop-up window of ”Save Workspace” if the user has not yet copied the public project. Only the user that ”owns” the public project can save and, therefore, change the actual associated public workspace(s).

The load workspace and save workspace features are both available through the header panel. After clicking in one of these buttons, the respective window appears above the website, as seen in **Figure 9**. In the load workspace window, the available workspaces to load are presented to the user grouped by the respective type of data, in order to make easier the search for the wanted data. Furthermore, each workspace is identified by the project and data folder that it corresponds to.



(a) Load workspace window.



(b) Save workspace window.

Figure 9: Layout of new samples prediction results page.

Similarly to what happens with the submission or choice of different projects, loading a different workspace implies loosing the workspace the user is working at that time, unless it is saved.

3.6 PRE-PROCESSING

The ”Pre-Processing” feature is available through the header panel, only when a dataset is available for analysis. When the user clicks the ”Pre-Processing” button, the page that allows the user to pre-process datasets appears.

This page was organized into two columns with the different types of pre-processing in each box: ”Missing Values”, ”Data Transformation”, ”Scaling”, ”Convert to Factor”, ”Mean

Centering", "Create subset by interval", "Remove data", "Remove data by NAs", "Aggregate samples", "Flat Pattern Filter", "Data Normalization", and, only accessible when it comes to spectral data, "Correction", "Smoothing Interpolation", "First Derivative", "Multiplicative Scatter Correction" and "Low-level data fusion". At the end of the page, at the center, the user is asked to give a name to the dataset that will be generated by applying the pre-processing tasks. The layout of this page can be observed in **Figure 10**.

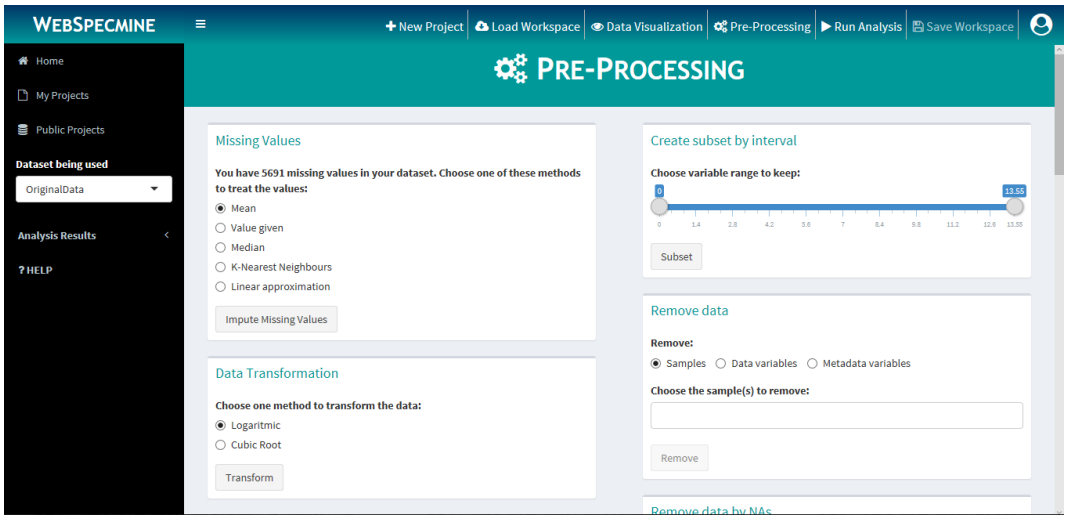


Figure 10: Layout of "Pre-Processing" page.

The processing is done over the dataset being currently used, and it can be done in any desired order, applying the wanted tasks. Various datasets can be generated, with different pre-processing pipelines, which allows to compare different results of the same analysis, according to the processing pipeline applied.

At the end of the page, the "Finish" button, only enabled when the dataset name input is filled, allows the user to indicate that the processing pipeline is defined. After naming the pre-processing, the name of the dataset will appear on the sidebar panel, in the section "Dataset being used", so that the user can choose the new dataset for further analysis.

If the chosen name already exists, the site will not allow the user to save the pre-processing done when he clicks the "Finish" button and the message "A dataset with that name already exists! Please choose a different one" will appear, giving the opportunity of renaming the dataset.

In the "Missing Values" box, a message saying "Your data has no missing values" appears if the dataset in use does not have missing values. If it has, the options to treat missing values will appear. These include replacing the missing values by the mean or median of the variables, by a given value defined by the user, using K-Nearest Neighbours, or a linear approximation. When the "Value given" or "K-Nearest Neighbours" choice is selected, an input will appear so that a value can be specified by the user for the relevant parameter.

The two data transformation methods made available are logaritmic and cubic root ones. The scaling methods are auto, pareto and range scaling. Normalization of the data can be done by the sum of a constant number given by the user, or the median.

Six functions are available for selection in the ”Flat Pattern Filters” processing box, including interquantile range, relative standard deviation, standard deviation, median absolute deviation, mean, and median. The values can be filtered by the percentage or through a treshold value. If the percentage option is chosen, a slider input with numbers between 0 and 100 appears, while an numeric value is asked when the threshold option is chosen.

Samples and data variables can be removed according to the missing values they have, in the box ”Remove data by NAs”. The user has an initial radio buttons input to choose between removing samples or data variables. According to what it is chosen in this input, a second radio buttons input changes its options, so that the user can choose how to remove the data in question accordingly. In both cases, the data can be removed by the number or percentage of missing values. When the first is chosen, an input will appear, asking the user to give the maximum number of missing values a sample or data variable can have. On the other hand, when the percentage option is chosen, a slider input will appear, allowing the user to set the maximum percentage of missing values a sample or data variable can have. Furthermore, the samples have an additional option of removing samples if they have missing values in the respective metadata variables.

All the pre-processing features available made use of *specmine* functions, which are shown in **Table 5**.

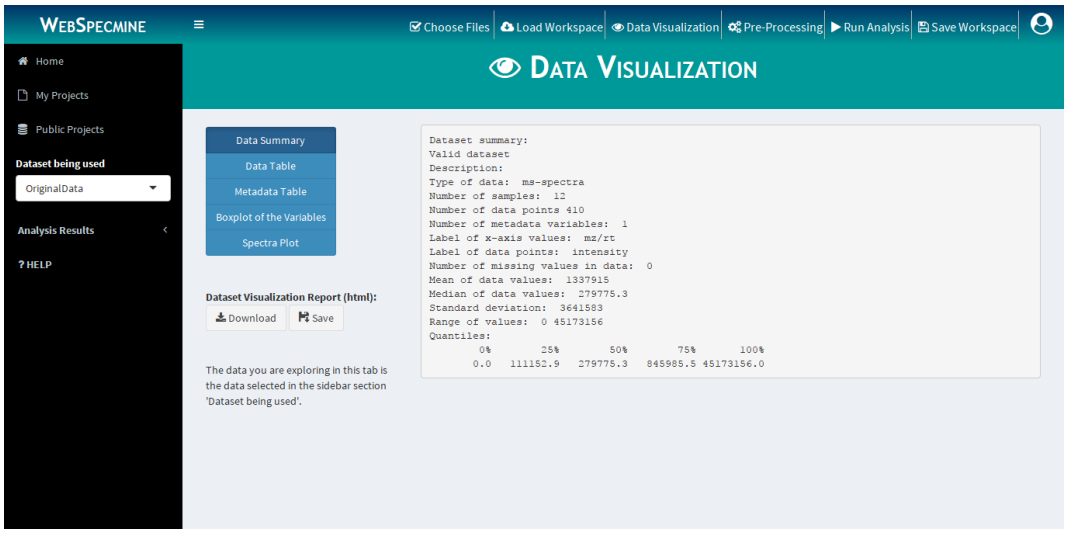
3.7 DATA VISUALIZATION

The ”Data Visualization” feature is available through the header panel. When the user clicks the ”Data Visualization” button, the page that allows the user to see the data and some of its characteristics appears, as it can be seen in **Figure 11**.

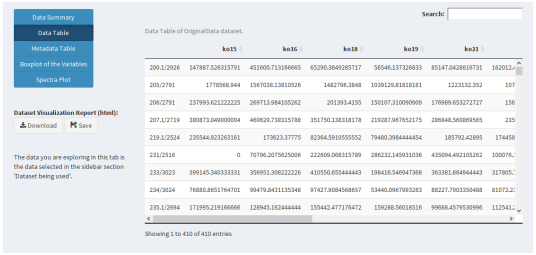
The layout of this page consists in two columns. One has the buttons that correspond to what the user can see about the dataset in question in a vertical list. Below this list, the possibility to download or save the data visualization report is made available. The second column shows the content that belongs to the last button clicked by the user.

The website was developed in such a way that the information the user is able to see in this page corresponds to the dataset being currently used, chosen in the sidebar tab ”Dataset being used”.

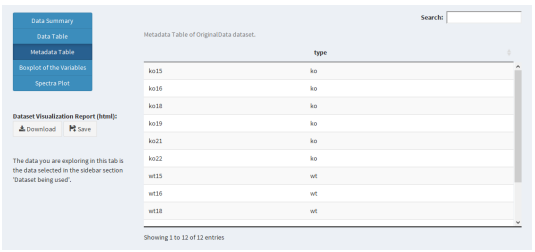
A dataset summary is available (**Figure 11a**), with information such as a short description of the data that was provided by the user while submiting the project for analysis; the type of data; number of samples, data points and metadata variables; the xx and yy axis labels; the number of missing values in the dataset; and some statistics over the data, such as mean,



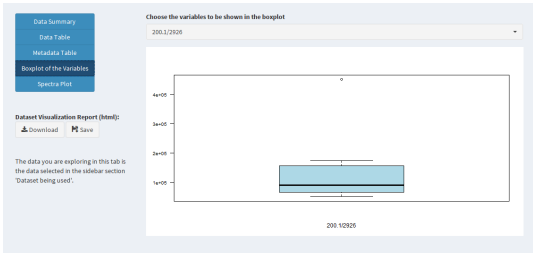
(a) Summary of the selected dataset.



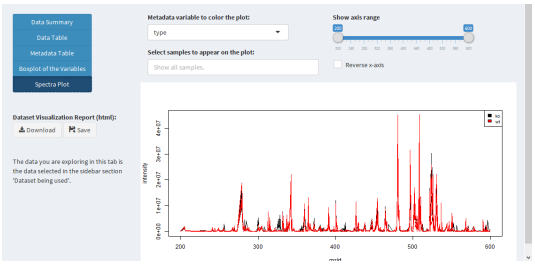
(b) Data table.



(c) Metadata table



(d) Boxplot of the dataset variables.



(e) Spectra plot of the dataset.

Figure 11: Layout of "Data Visualization" page.

median, standard deviation, range and quantiles. This summary is obtained by using the *specmine* function *sum_dataset*. The user can also see the data and metadata tables from the dataset (**Figures 11b and 11c**).

A boxplot of one or more variables variables can also be viewed (**Figure 11d**), by making use of the *specmine* function *boxplot_variables*. The variables to appear in the boxplot can be chosen by the user, through the select input above the plot. The code was developed so that if no variable in the input is selected, all variables in the boxplot will appear. A select input was chosen over a radio button input, due to the fact that the latter could take most of the window if the number of variables were high.

A spectra plot is also available, only for datasets of spectral type, by making use of the *specmine* function *plot_spectra*. This plot shows the spectra of one or more selected samples from the dataset. Again, if no samples are chosen in the select input, the spectra for all samples is plotted. The user is also able to color the plot according to the different values of the metadata variable that he chooses in the respective select input. Furthermore, the user is able to choose, through a slider input, the range of the xx axis and if the values in this axis appear in descendent or ascendent order.

As regards to the reporting, the plots appear in the report like they are in the page in the moment of the report file creation. This means that only the variables that are selected to appear in the boxplot at the time of the report file creation will appear in the file plot, for example, and the same happens with other inputs that may change the plots.

3.8 ANALYSIS

3.8.1 *Run Analysis*

The "Run Analysis" feature is available through the header panel and leads to the page that allows the user to do the analysis of the datasets.

There are a total of 7 types of analysis provided. Each is represented by a box, from which the respective analysis is accessed, as it can be observed in **Figure 12**.

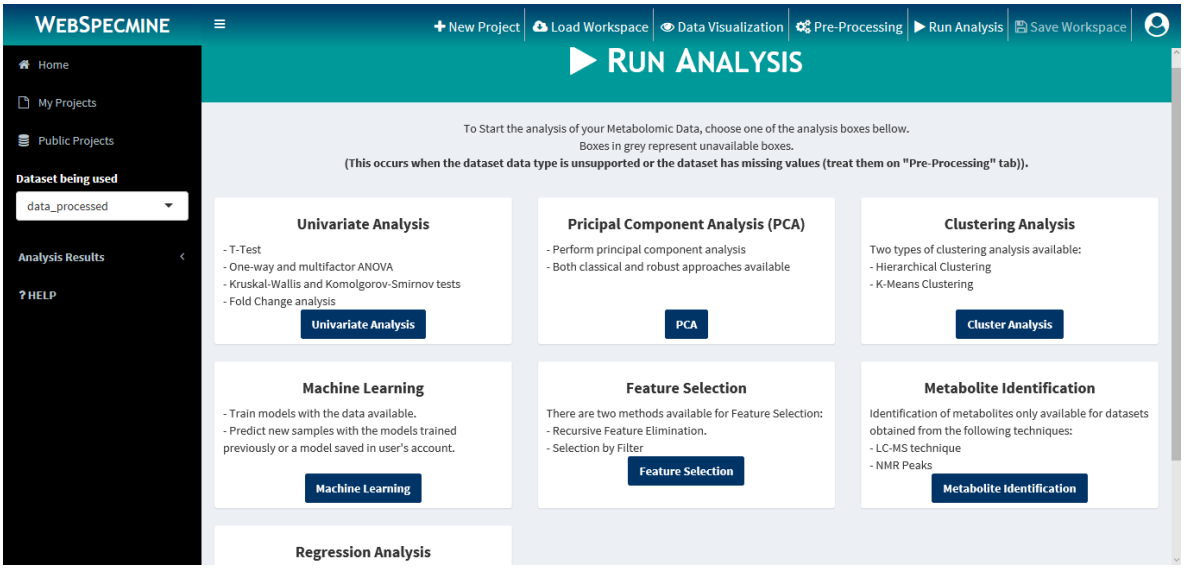


Figure 12: Layout of the "Run Analysis" page.

- *Metabolite Identification*, only available for spectral data from the LC-MS technique or peaks lists data from the NMR technique;

- *Univariate Analysis*, where t-tests, one-way and multifactor analysis of variance (ANOVA), Kruskal-Wallis and Komolgorov-Smirnov tests, and fold change analysis can be done;
- *Regression Analysis*, where regression and correlation analysis are made available.
- *PCA*, both classical and robust approaches;
- *Clustering Analysis*, where hierarchical and k-means clustering are available;
- *Machine Learning*, where it is possible to train models and predict new samples;
- *Feature Selection*, where two methods are available, namely recursive feature elimination and selection by filter;

The analysis boxes might not be accessible if the dataset currently in use contains missing values. The box "Metabolite Identification" might also be unaccessible if the dataset type is not supported, i.e., if the dataset is not spectral data from the LC-MS technique or NMR peaks lists data. In these cases, the respective boxes remain in grey, inaccessible, unless a dataset with no missing values is chosen or, in case of the "Metabolite Identification" box, the data type is supported and the dataset selected has no missing values.

All analyses done must have a name associated to them, given by the user. These names must differ, otherwise the user cannot execute the analysis wanted, due to the fact that the button that leads to such analysis is disabled while these requirements are not met. However, the input text box where this name has to be given comes with a default value.

The *specmine* functions used to run the analyses mentioned below are listed in **Table 5**.

Metabolite Identification

In case of entering this analysis with a LC-MS spectral dataset, as it can be seen in **Figure 13**, besides the analysis name, the user only needs to choose the column of metadata that may help in the identification of metabolites. All the other parameters are already set by default and the user cannot change them, like the peak tolerance and mass tolerance ones, which are set to 0.005 and 0.5, respectively.

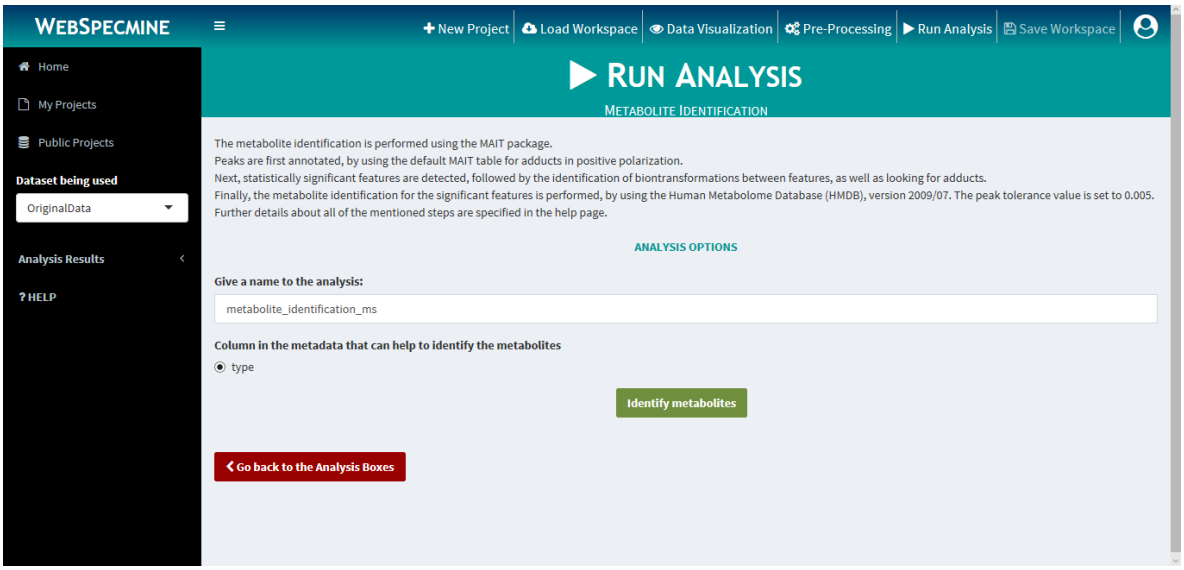


Figure 13: Layout of the metabolite identification in the "Run Analysis" page for MS dataset.

On the other hand, if the user enters this box with an NMR peak lists dataset, the user will have different parameters to set, besides the one regarding the analysis name, as it can be seen in **Figure 14**. The user must set the ppm tolerance used in the matching between cluster and reference peaks and the number of top metabolites matched to show in the results.

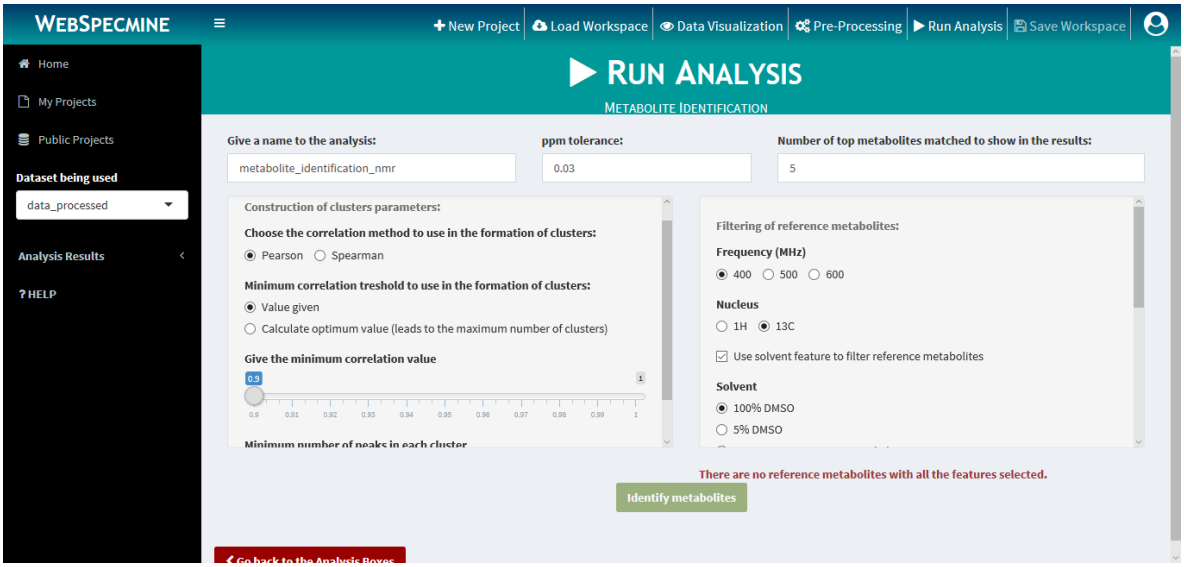


Figure 14: Layout of the metabolite identification in the "Run Analysis" page for NMR peaks lists dataset.

Furthermore, there is a box for the construction of clusters, where the user must set the correlation method, that might either be "Pearson" or "Spearman", set the minimum

number of peaks that the clusters must have, and choose if he/she wants to give the value for the minimum correlation in the formation of clusters or allow the website to calculate the optimum value. If the option to calculate the optimum value is chosen, the user is able to choose if the maximum number of peaks a cluster can have while searching for this value is the number of the largest reference metabolite or give a number.

Finally, the user is able to filter the library of reference metabolites in a box at the right of the previous one. The user must choose the frequency, whose possible values must be 400, 500 and 600, and the nucleus, either "1H" or "13C", of the reference spectra metabolites. If wanted, the user can also filter the library through the solvent, pH or temperature. The possible values for solvent include water, DMSO, D_2O, among others. To give the pH, the user is provided with a slider input, where the user can choose the minimum and maximum value for the interval of pH wanted, or choose one single value of pH, by setting both values as the same value. Finally, there are only two temperatures available, which are 25 and 50 degrees Celsius. While the user is setting these parameters, the website checks if the combination of the chosen parameters lead or not to no reference metabolites. If so, the warning message "There are no reference metabolites with all the features selected" appears under the box to alert the user and disables the button to do the identification.

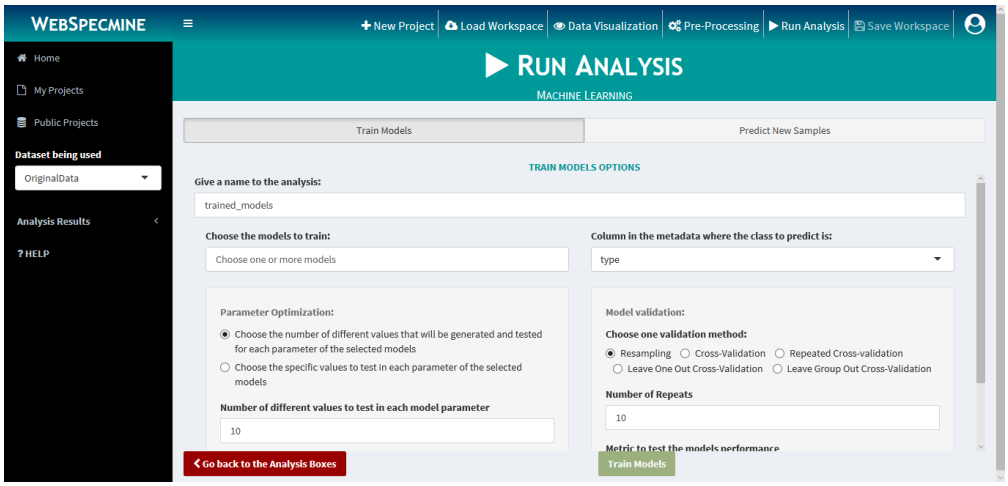
Machine Learning

At the top of the Machine Learning page, there are two buttons that can lead to either model training, if the button "Train Models" is clicked, or to samples prediction, if the button "Predict New Samples" is chosen. The options that can be chosen for each type of analysis appear below these buttons. By default, the page related to "Train Models" appears the first time this box is accessed, while the "Predict New Samples" option is not accessible if no model was previously trained. In fact, if the user has not yet trained a model using the dataset currently chosen, and tries to access the page "Predict New Samples", a warning message will show: "Feature only available when you have trained models". The layout of this page is seen in **Figure 15**.

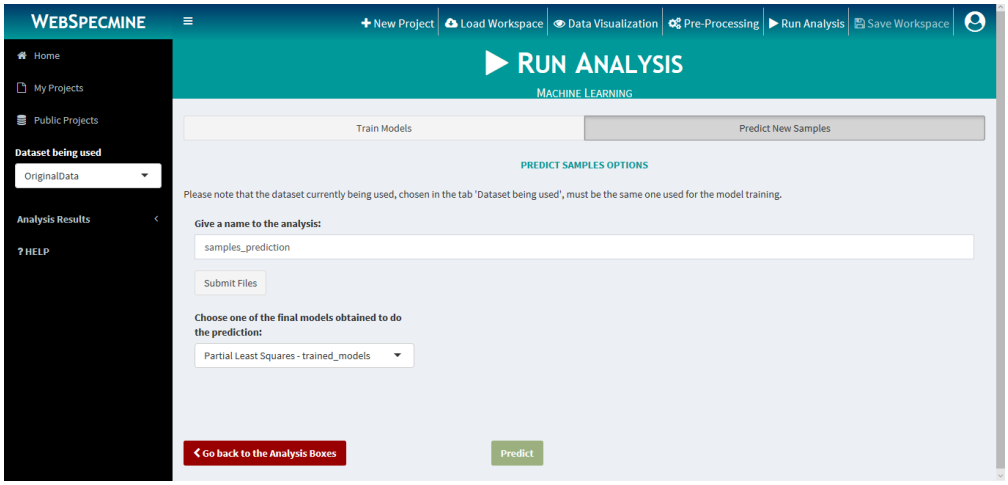
There are several options that must be set so that the model training is done, as seen in **Figure 15a**. It is possible to choose one or more types of models to train, which include PLS, Decision Tree (C4.5), Rule-Based Classifier, SVMs with linear kernel, Random Forests, LDA and Neural Network. The name of the metadata variable to predict must also be given.

Regarding the optimization of models' parameters, the user can choose to give all the values that will be tested for each parameter of each chosen model, or only define the number of different values that will be tested for each parameter, whose values will be set automatically.

Finally, the user must also set the model validation options. The available model validation methods are resampling, cross-validation, repeated cross-validation, leave one out



(a) Model Training.



(b) New Samples Prediction

Figure 15: Layout of the machine learning in the "Run Analysis" page.

cross-validation and leave group-out cross-validation. If the selected method is resampling, the number of resampling iterations will be asked by the website. On the other hand, if one of the other methods is selected, the number of validation folds is asked, as well as the number of repeats, if the selected method is repeated cross-validation. The validation metric to be used can either be accuracy or Receiver Operating Characteristic (ROC).

As regards to new samples predictions, as seen in **Figure 15b**, the user has to submit the new samples file(s) and choose one of the available trained models to do the prediction. Only the trained models originated from the dataset currently in use will be available. When the user clicks the button to start the submission of the file(s), a pop-up window will appear, with the options to proccess the data file(s), according to the type of data that must be submitted, which is the same type of data used to train the models. After submitting the files, the user will be asked to treat the missing values, with the same options present

in the pre-processing page, if the new data has missing values. After that, the data will be further pre-processed similarly to the data used to train the model chosen. With this, a brief summary of the new samples, similar to the one observed in the "Data Visualization" page, will appear. The user cannot execute this analysis while no new samples files have been submitted, besides the already mentioned need to give a correct name for the analysis.

Feature Selection

As shown in **Figure 16**, there are several options that must be set so that this analysis is done. The user has to choose the metadata variable that will be predicted and one of the two of the feature selection methods available, which are recursive feature selection and selection by filter. For both methods, the available functions for model fitting, prediction and variable importance/filtering include random forests, linear regression, bagged trees, linear discriminant analysis and naive-bayes.

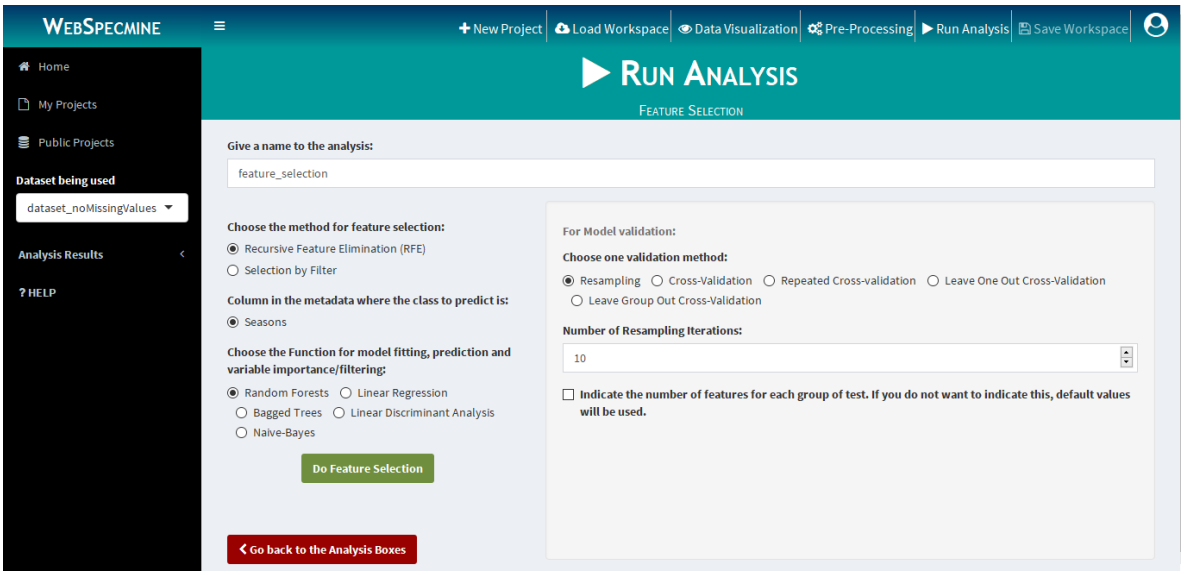


Figure 16: Layout of the feature selection in the "Run Analysis" page.

Finally, it is also possible to set the options regarding model validation. These consist on choosing the validation method, whose options are the same as the ones provided in the model training section, and some options regarding each type of method selected. If the selected method is resampling, the number of resampling iterations will be asked by the website. On the other hand, if one of the other methods is selected, the number of validation folds is asked, as well as the number of repeats, if the selected method is repeated cross-validation. The user can also choose if he wants to manually set the number of features that will be tested in each group test. If the user chooses to do so, he must give the size of each group test, separated by a comma. If not, the groups' sizes will be generated by default.

3.8.2 *Analysis Results*

Each time an analysis is finished, the user is redirected to the respective results page. All the obtained results related to the data available for analysis are accessible through the sidebar panel, in the tab called "Analysis Results". This tab has subtabs that correspond to each type of analysis made. These subtypes have the links to the respective analysis results' pages, represented by the names given by the user.

These results are stored in a reactive variable called "results", which is a list where each item is a performed result. Each item is named after the names given by the user to the analysis performed and is a list whith two items, called *result* and *options*. The actual results obtained by the analysis are stored in the *result* item, while options related to the analysis performed are stored in *options*. This last item is another list, whose content varies according to the type of analysis performed.

Metabolite Identification

In a metabolite identification results page for LC-MS spectral data, as seen in **Figure 17**, it is possible to see the options chosen and used in the respective metabolite identification analysis. These options are accessed through a circular button placed at the top left corner of the results page. By cliking it, the user can see such options, which consist on the name of the dataset used for the identification and the name of the metadata variable used to help in the metabolite identification.

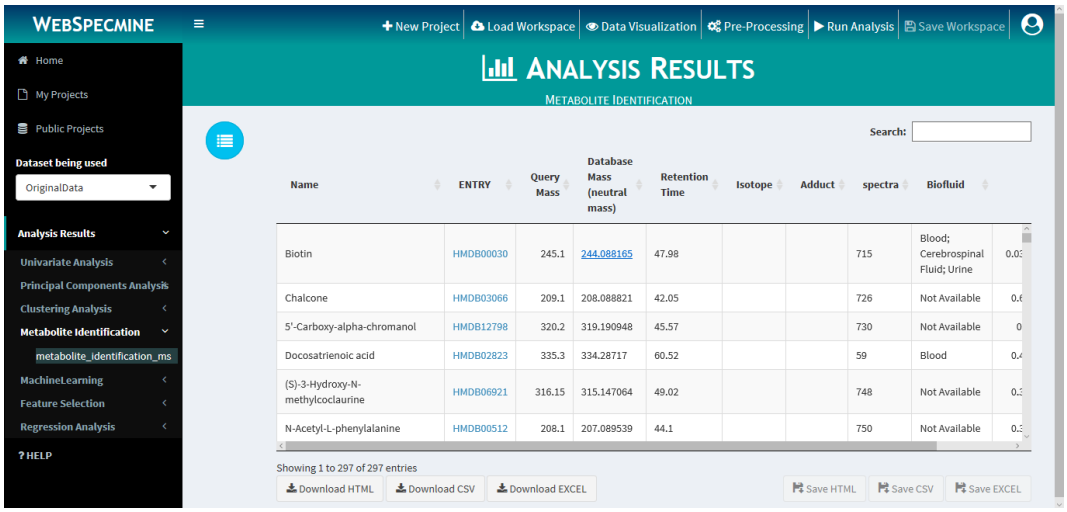


Figure 17: Layout of a metabolite identification results page for MS data.

The actual results are available in the form of a results table, where each line corresponds to an identified metabolite. For each of these metabolites, there is information on the HMDB entry number, with a link to the HMDB webpage of the respective metabolite, the

query and theoretical masses, the retention time, isotopes, adducts, spectra, biofluids and the adjusted p-value.

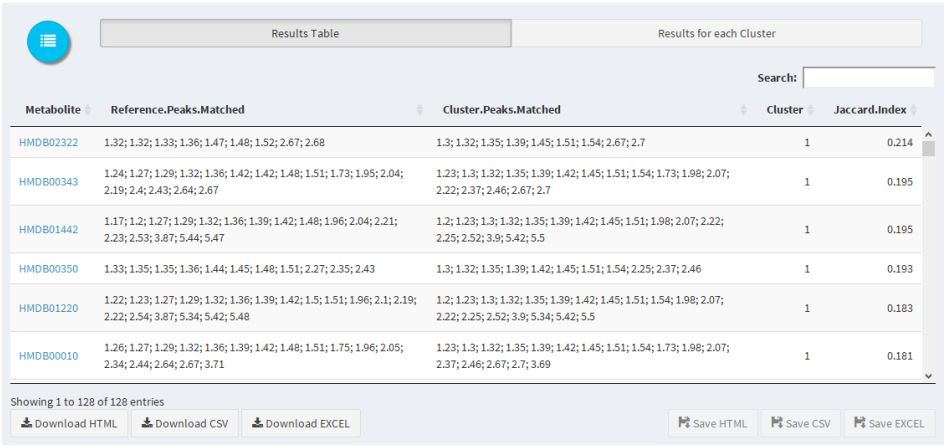
For an NMR metabolite identification results page, the available options (**Figure 18**) include the dataset used for the identification; the ppm tolerance used; the chosen number of top metabolites matched per cluster to show; the parameters used in the construction of clusters, which include the correlation method, the correlation value, if the value was provided by the user, the minimum number of peaks chosen for a cluster, and the maximum number of peaks in a cluster, if provided by the user for the calculation of the optimum correlation value; and the parameters, if used, regarding the filtering of the library of reference metabolites, such as the frequency, nucleus, solvent, pH and temperature.

Figure 18: Layout of the options in a metabolite identification results page for NMR peaks lists data.

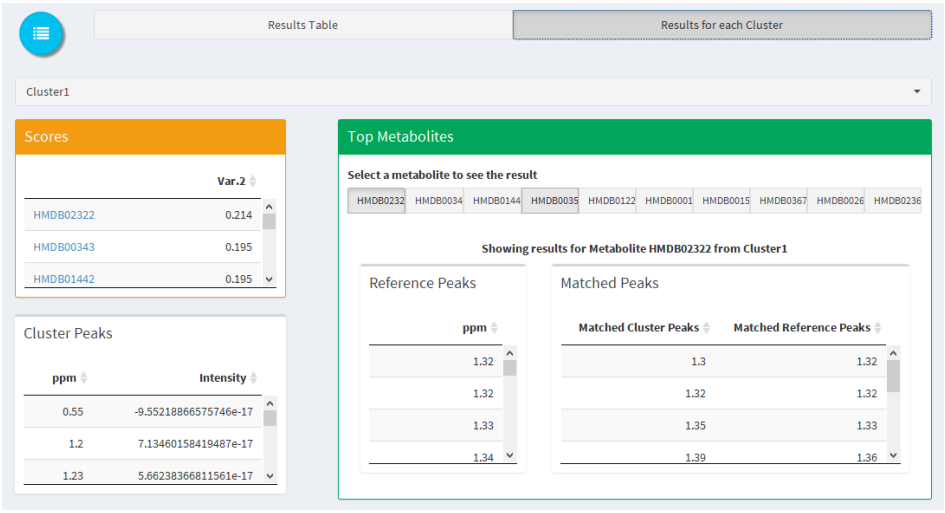
At the right of the options button, there are two buttons, which allow the user to see the different types of results obtained, shown below these buttons. The "Results table" option leads to a table where each line corresponds to an identified metabolite, as shown in **Figure 19a**. All metabolites identified in each cluster are here present, which can lead to repetitions if the same metabolite matched different clusters. The information here provided for each identified metabolite includes the HMDB entry number, with a link to the HMDB webpage of the respective metabolite, the cluster and reference peaks matched, the cluster, and the Jaccard Index score.

On the other hand, the "Results for each Cluster" option leads to more detailed information on the matches obtained in each cluster, as shown in **Figure 19b**. When this option is clicked, a select input with all the clusters obtained is available, so that the user can choose one of the cluster and all the results regarding the respective cluster appear below.

These results are organized in three boxes, disposed in two columns. The first column has the two boxes whose content refers to the scores of the top matches and the peaks of



(a) Layout of the general results table section.



(b) Layout of the section of the results per cluster.

Figure 19: Layout of a metabolite identification results page for NMR peaks lists data.

the cluster in question, respectively. Each of these matches is represented by the HMDB entry number, with a link to the HMDB webpage of the respective metabolite. The other box, which fills the whole second column, contains further information on the matches of each metabolite. At the top of this box, there are buttons, where each refers to one of the top metabolites that matched the cluster. By clicking on one of these buttons, information about the metabolite peaks and the peaks that matched between the cluster and reference metabolite are given below the buttons.

Furthermore, in case no metabolites matched a certain cluster, only two boxes, side by side, will appear. One with the cluster peaks and the other with the message "No metabolites matched this cluster". However, so that the user does not need to enter in each cluster to know if it got macthes or not, the clusters with no matches are followed by the message "No matches" in the select input.

Machine Learning

In a model training results page, shown in **Figure 20**, it is possible to see the options chosen and used in the respective analysis. These options are accessed through a circular button placed at the top left corner of the results page. By clicking it, the user is able to see such options, which include the name of the dataset used, the name of the metadata variable to be predicted, the validation method and the validation metric.

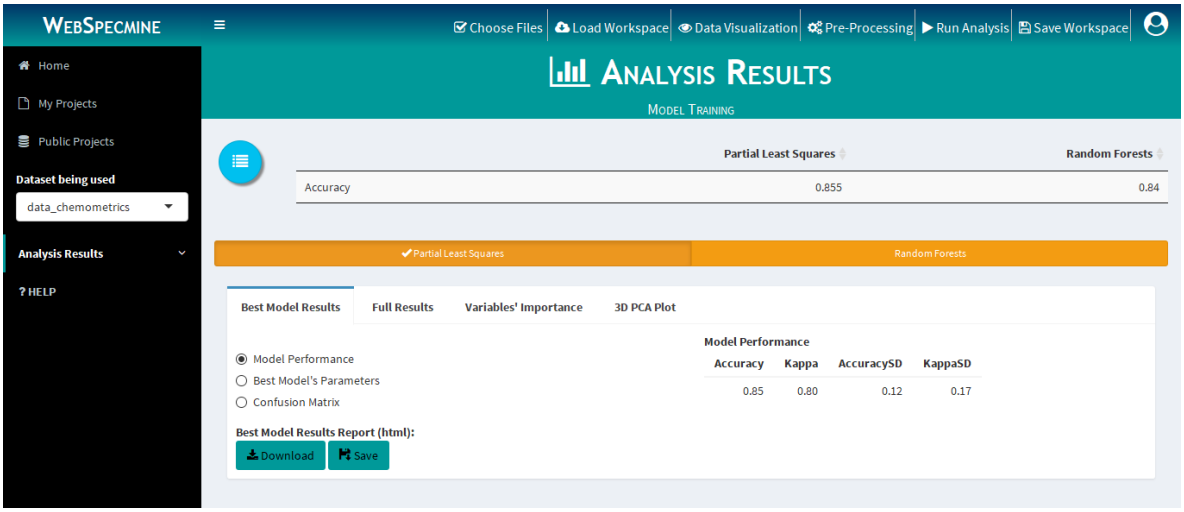


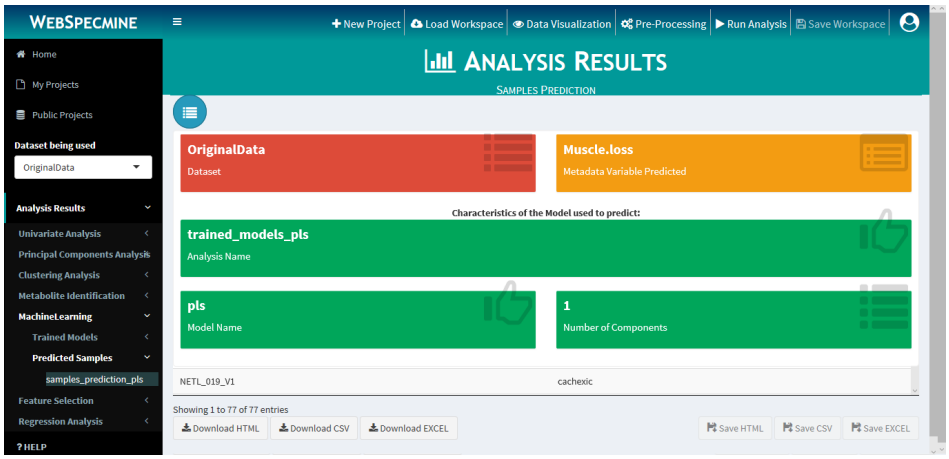
Figure 20: Layout of a model training results page.

At the right of this button, there is a summary table where it is possible to observe the accuracies of all models trained in a certain analysis, so users can have a quick overview of each model. If only one model is trained, this table will not appear.

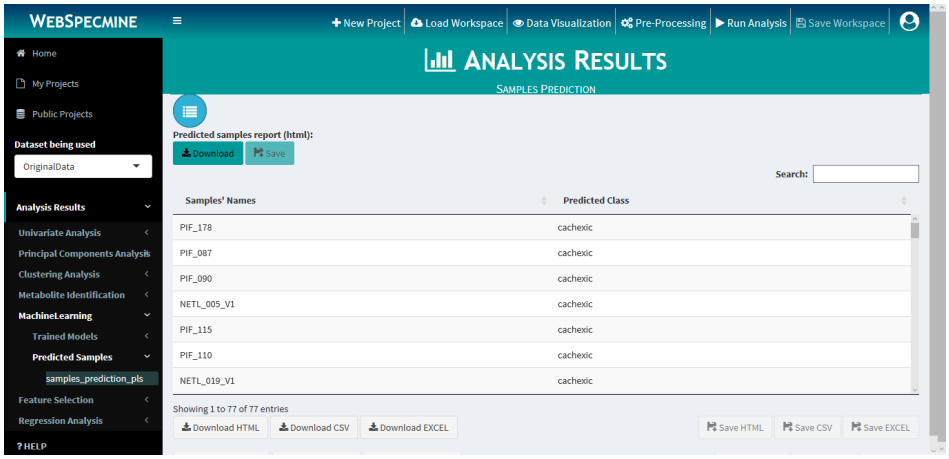
Below this, there are one or more buttons, each representing a trained model. By clicking in one of the buttons, all the results regarding the respective model are shown below. By default, the results that are shown when the user is redirected to this page are the ones of the first model. The results are shown in the form of a tabbed panel. The first tab contains information about the best model obtained: the performance, the parameters and the confusion matrix.

A report of these results is available, which includes the options' parameters. The second tab provides a table with all the results for the trained model, with the values of accuracy, kappa and respective standard deviations for each combination of parameter values tested. The third tab provides a table with the importance of each variable used in the model training. Finally, if the selected model is a PLS model whose best model has 3 or more components, an aditional tab with the 3D PCA plot appears. In this tab, the user can choose in a select input three of the components formed to appear on the plot. This plot appears bellow this input and the data points are colored according to the metadata variable predicted.

In a samples prediction results page, as shown in **Figure 21**, it is possible to access the options chosen and used in the respective analysis through a circular button placed at the top left corner of the results page. By clicking it, the user is able to see such options, which consist on the name of the dataset used to train the model used to predict the new samples, the name of the metadata variable predicted, and the characteristics of the model used for prediction: the name of the analysis from which the model comes from, the name of the model, and the values of the model's parameters.



(a) New Samples Prediction options.



(b) New Samples Prediction results.

Figure 21: Layout of new samples prediction results page.

Below this button, the user can see the results table with the predicted class for each of the new samples submitted. Above this table, a report of these results is available. This report not only contains the full table with the obtained results, but also the options chosen to run the analysis and the characteristics of the model used to do the prediction.

Feature Selection

In a feature selection results page, as seen in **Figure 22**, it is possible to see the options chosen and used in the respective analysis. Such options consist on the name of the dataset used, the name of the metadata variable to be predicted, the selection method, the function and the validation method.

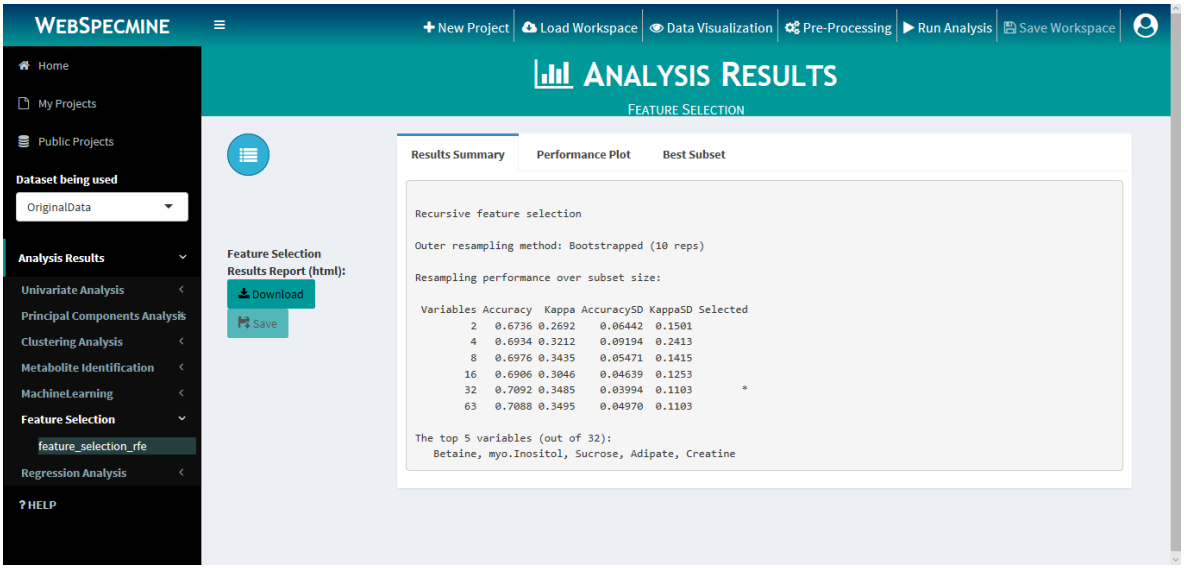


Figure 22: Layout of a feature selection results page.

The actual results are disposed in a tabset panel, located to the right of the options button. One of the tabs contains a brief summary of the results, and another one has the names of the variables that make up the best subset. In case the recursive feature selection method is the one used, another tab will appear with a performance plot, which shows the accuracy across the different subset sizes. In this tab, the plot image can be saved or downloaded in pdf format. To accomplish this, the *RMarkdown* package was used.

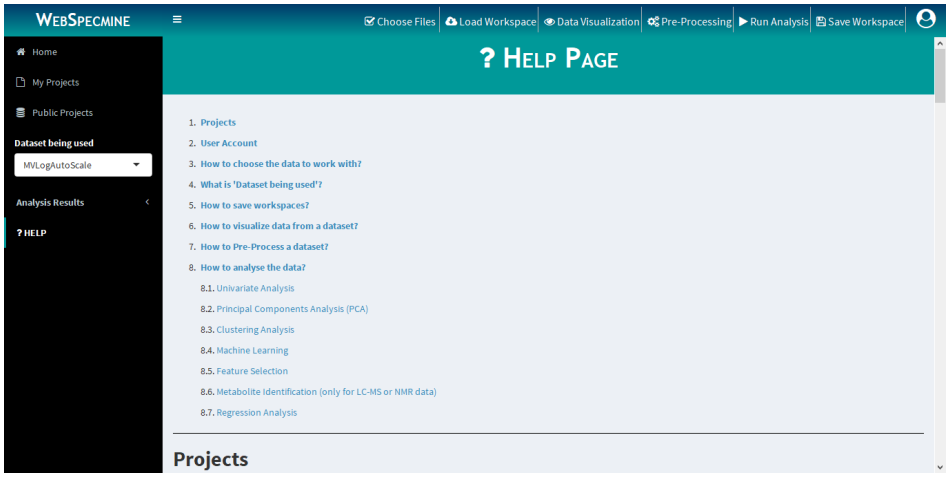
Below the options button, a report with all the results shown in the tabset panel and the options chosen to run the analysis is available.

3.9 HOME AND HELP PAGES

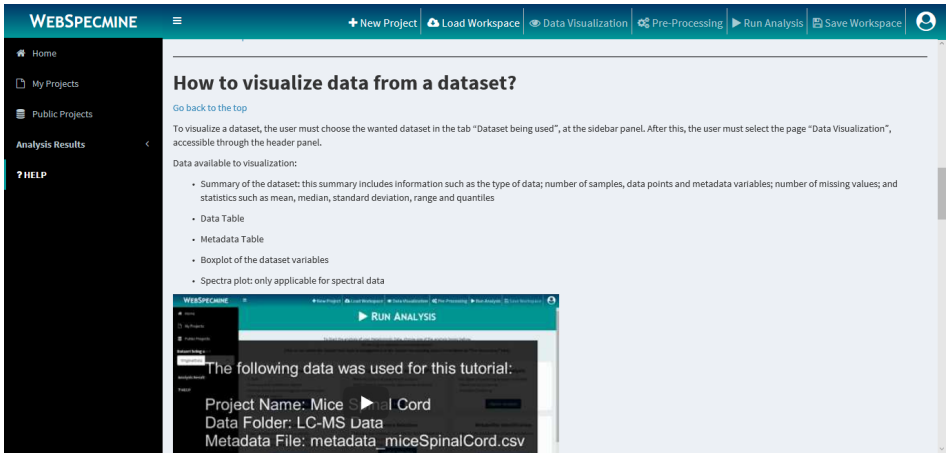
The "Home" and "Help" pages are both accessed through the sidebar panel. The latter is considered the main page, where the user is firstly directed to when accessing the web application. In this page, the user is provided with a video giving a brief information of how the website works, being redirected for more detailed information to the "Help" page, and the publications made on *specmine* and *WebSpecmine*. Furthermore, in the future, it will

be here where the users will have access to brief news about new functionalities for the website (**Figure 6**).

As regards the "Help" page, it is here where the user has access to detailed information on how to use the website, including information about how to choose data for analysis, how the user accounts work, how to pre-proccess data, how to analyse the data, among other possible frequent doubts that may arise to the users. At the top of this page (**Figure 23a**), there is an index from which the user is able to access the different information provided. Each section of information has a brief text explaining/answering the topic in question. When needed, the text is followed by a video explaining the covered task (**Figure 23b**).



(a) Index section.



(b) Section of the "Help" page.

Figure 23: Layout of the "Help" page.

This page was developed making use of the *RMarkdown* package, as it allows the construction of HTML files that can be integrated in a shiny website.

4

USE CASES

In this chapter, it will be demonstrated how some previously developed studies can be reproduced by using the *WebSpecmine* site. These will be made available as public projects, which may be imported by any user of the platform.

4.1 NMR DATA: PROPOLIS

4.1.1 *Introduction*

Propolis, or bee glue, is a substance produced from the collected buds or exudates of plants (resin) by bees (*Apis mellifera L.*), where they mix it with bee wax and use it to seal the walls of the hive, strengthen the borders of combs and embalm dead invaders (Wollenweber et al., 1990). This substance is known to have antimicrobial, anti-viral, anti-tumoral, anti-inflammatory and anti-neurodegenerative properties. Propolis' chemical composition might be strongly influenced by environmental factors and seasoning. Thereore, the study here reproduced aimed to get insights of important features associated with the chemical composition, seasons and geographical origin of the propolis produced in the Santa Catarina state, in southern Brazil (Maraschin et al., 2016).

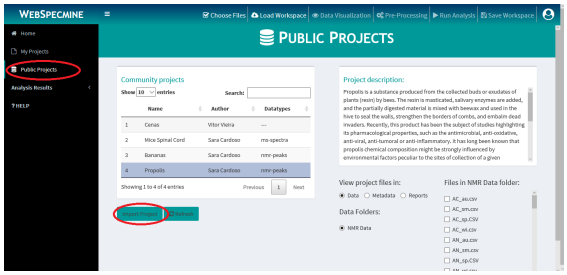
The samples used in this study, acquired using the NMR technique, were stored in the public project called *Propolis*, under the data folder *NMR Data*. Regarding the metadata, the file *metadata_propolis* is given.

There are a total of 59 samples, 15 from autumn (AU) and spring (SP), 13 from winter (WI) and 16 from summer (SM). They were collected in 2010 from *Apis mellifera* hives located in southern Brazil (Santa Catarina State). The samples are also separated in three agroecological regions for the different apiaries: 12 samples from the Highlands, 11 from the Plain, and 36 from the Plateau.

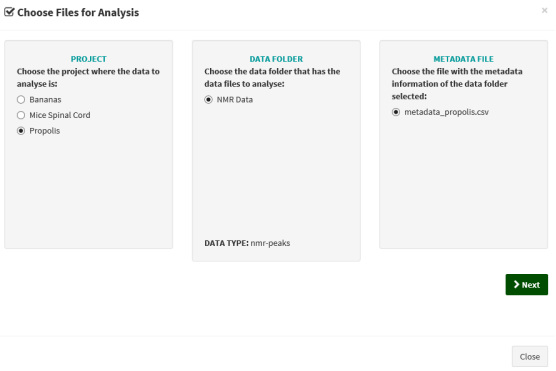
The analysis pipeline here demonstrated followed one available in `http://pubs.acs.org/doi/suppl/10.1021/acs.jnatprod.5b00315/suppl_file/np5b00315_si_001.pdf`, regarding the type of analysis developed in this work. This was conducted using R and the specmine package.

4.1.2 *Choosing files for analysis and pre-processing*

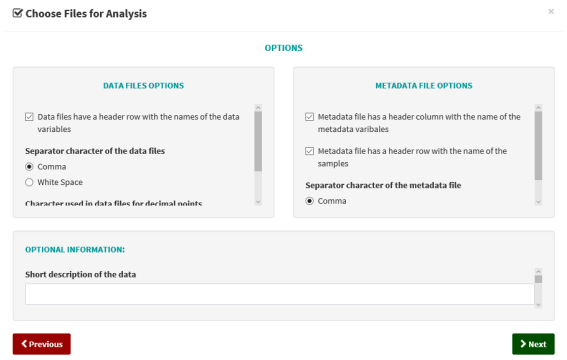
After entering the account, the user must copy the public project in question, named *Propolis*, into his/her account. This is done through the page "Public Projects", accessible through the sidebar panel (**Figure 24a**). After this, the user must click in the "Choose Files" button, present in the header panel, and choose the project in question, and the mentioned data folder and metadata file for analysis (**Figure 24b**). Then, the user must click the "Next" button, which will lead him/her to the window where the options regarding the data and metadata files are set. In this case, the options are the default ones, so no change is needed (**Figure 24c**). Again, the user must then click the "Next" button, so it is possible to set the options for the alignment of peaks (**Figure 24d**). In this case, the default ones are also used, which consist in the specmine algorithm as the method used and 0.03 ppm as the size of the step. With this, the user is able to click the button "Submit For Analysis" to finalize the submission of the data to analyse.



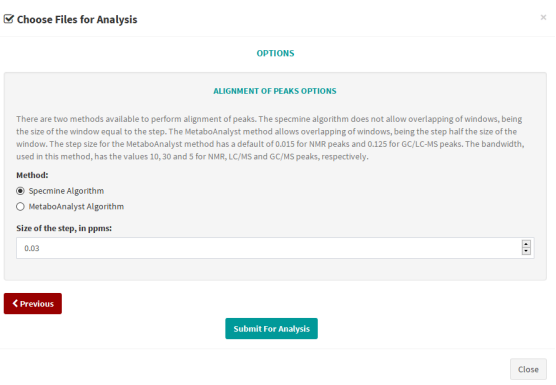
(a) To copy the *Propolis* project to the user account, he/she must go to the "Public Projects" page, select the project in the table and click the button "Import Project".



(b) Choose the project and respective data and metadata files for analysis.



(c) Options regarding the data files and metadata file.

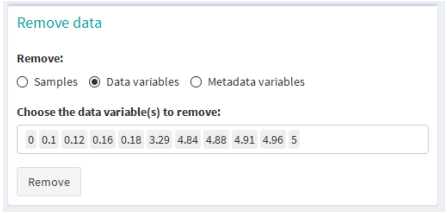


(d) Options regarding the alignment of peaks.

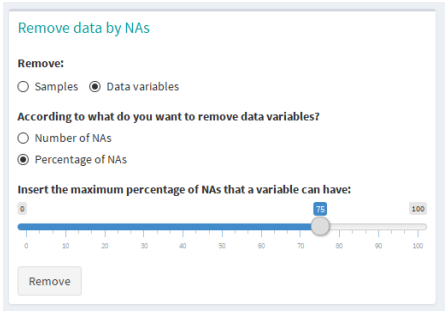
Figure 24: Demonstration of how to choose the files from the *Propolis* project for analysis.

After the data files are processed, the user is redirected to the "Run Analysis" page. Here, the user will notice that no box is accessible at this point. This happens because the dataset created by processing the files (*OriginalData*) has missing values, which makes impossible to proceed with the analysis. Thus, pre-processing of the dataset is needed. This task is performed through the page "Pre-Processing", accessed through the header panel.

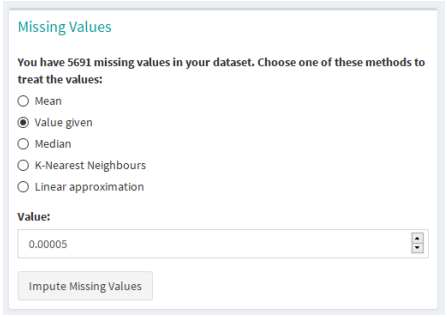
Two different pre-processing pipelines were applied, one to be used in the chemometrics analysis, named *data_chemometrics* (**Figure 25**), and the other one for metabolite identification, named *data_ID*.



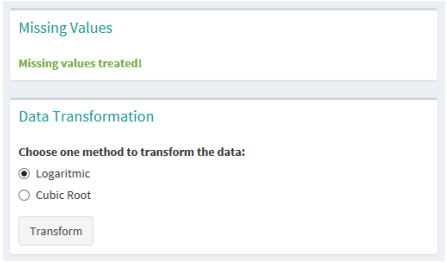
(a) Removing the peaks in the intervals 0-0.19, 3.29-3.31 and 4.84-5.00.



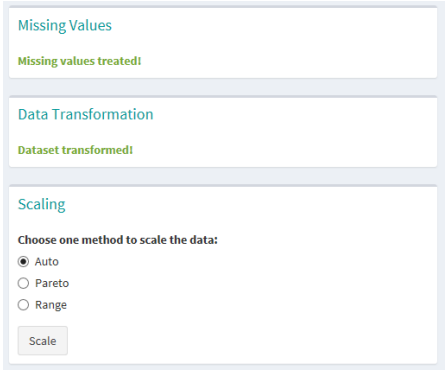
(b) Removing data variables with more than 75% of missing values.



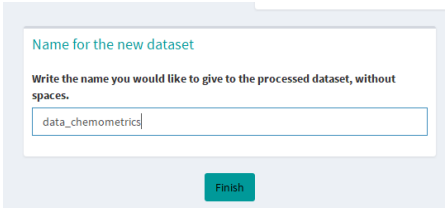
(c) Missing values treatment with replacement by a given value of 0.00005.



(d) Logaritmic transformation of the dataset.



(e) Autoscaling of the dataset.



(f) Naming the dataset *data_chemometrics*.

Figure 25: Demonstration of how to pre-proccess the dataset for the chemometrics analysis.

First, data variables between the ppm values 0-0.19, 3.29-3.31 and 4.84-5.00 (**Figure 25a**) were extracted, followed by removing any variables with more than 75% of missing values (**Figure 25b**). After removing these data, the missing values on the dataset were treated, by replacing them with the given value of 0.00005 (**Figure 25c**). This is followed by a logaritmic transformation of the data (**Figure 25d**) and data autoscaling (**Figure 25e**). To finalize this pre-processing pipeline, the dataset was given a name (*data_chemometrics*) and the button "Finish" clicked (**Figure 25f**). With this, the dataset being currently in use will automatically change to the newly created dataset and, by entering the "Run Analysis" page once again, the user will notice that the boxes are now available.

However, the user still needs to create another dataset, appropriate for the metabolite identification. Therefore, still in the "Pre-Processing" page, the user must set the dataset being used back to *OriginalData* to start the new processing. The user will only need to remove the data variables between the ppm values 0, 3.29-3.31 and 4.84-5.00, treat the missing values by replacing them with the given value of 0.00005, and perform logaritmic transformation and auto scaling, so that data variables can be better clustered. Finally, a name to the new dataset (*data_ID*) must be given.

4.1.3 *One-way ANOVA Analysis*

The analysis started with one-way ANOVA, along with TuckeyHSD test, by using the metadata variable *seasons*. This analysis was performed by entering the "Univariate Analysis" box in the "Results Analysis" page, while the dataset being used is *data_chemometrics*, and accessing the "One-Way Analysis of Variance (ANOVA)" tab, in the tab box located at the left of the page, so that the options regarding this type of analysis appear at the rigth, as shown in **Figure 26**. Besides the options regarding the analysis, the user can also set the options for the results plot, which in this case are the p-value threshold set to 0.05 and the reverse of the axis, which was not needed. The name given to this analysis was *oneANOVA*.

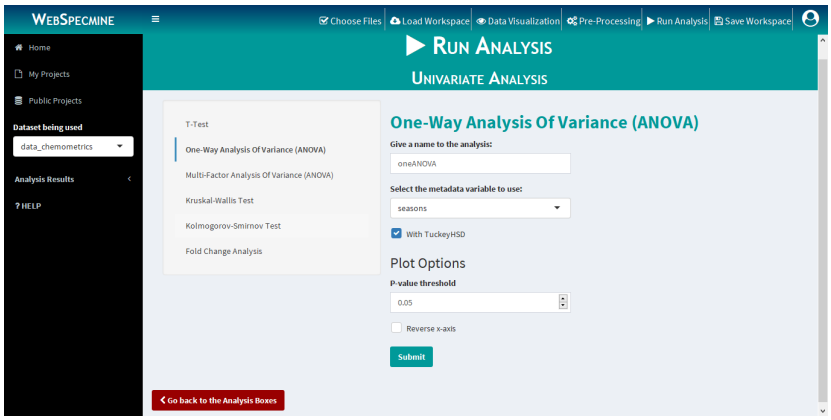


Figure 26: Options of the one-way ANOVA on the *data_chemometrics* dataset.

Once this analysis is finished, the website redirects the user to the corresponding results page, where the user can see the table with the ANOVA results, where different variables are ordered by decreasing corrected p-value (FDR method), and the results plot, as shown in **Figure 27**.

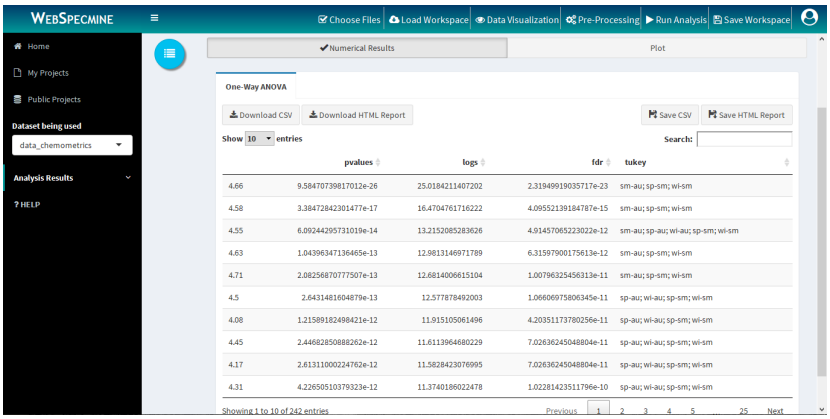


Figure 27: Results of the one-way ANOVA on the *data_chemometrics* dataset.

Further details on what this analysis consists on and how this part of the website is implemented are in Afonso (2017).

4.1.4 *Principal Components Analysis*

The next analysis performed was PCA. To perform this analysis, the user must go back to the "Run Analysis" page and select the "Principal Component Analysis (PCA)" box. Here, the user must select the "Normal PCA" tab in the tab box at the left of the page, so that the options regarding this analysis appear at the rigth. The only options needed to set are the name of the analysis, named *PCA*, and if the variables must be scaled and centered, which in this case they do, as shown in **Figure 28**.

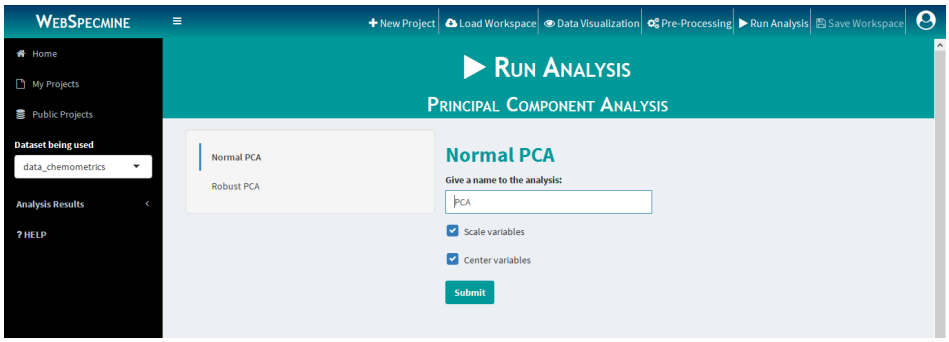
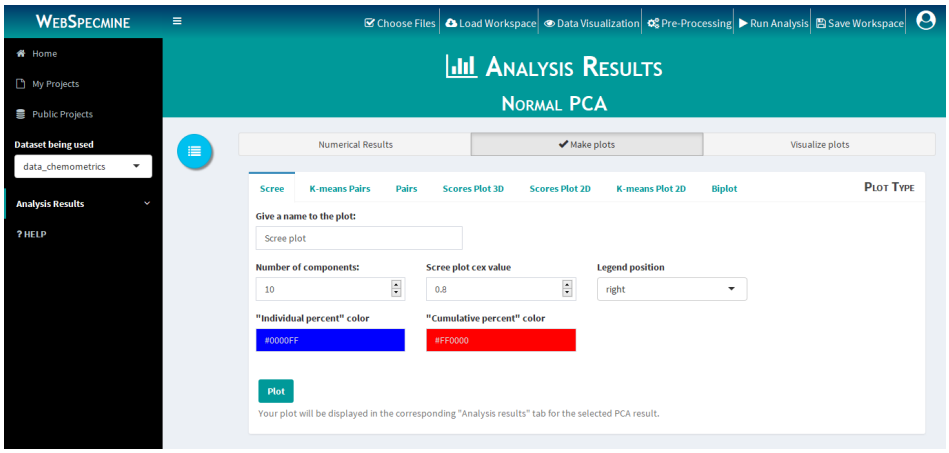
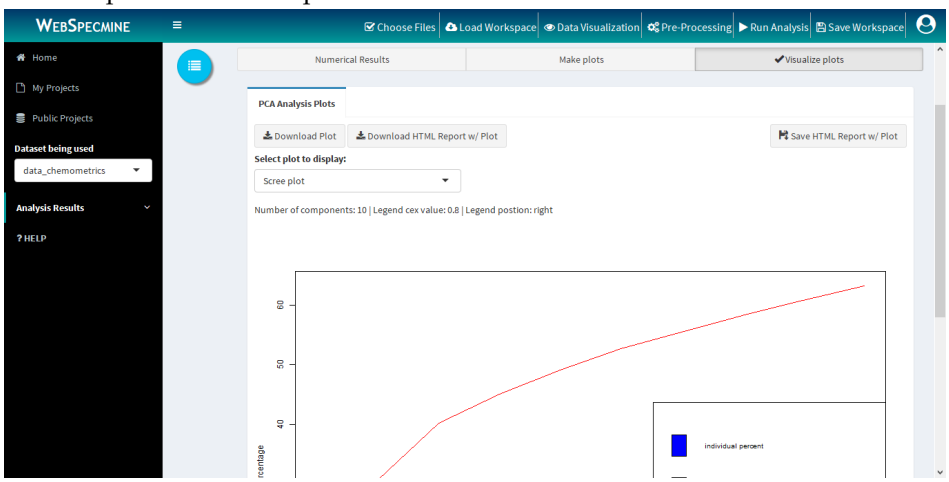


Figure 28: Options of normal PCA on the *data_chemometrics* dataset.

Once this analysis is finished, the website redirects the user to the corresponding results page, where the user can see the numerical results, including the components' importance, the scores matrix and the variable loadings, and see the different plots available to visualize the results, as shown in **Figure 29**. These plots include the scree, k-means pairs and pairs plots, the scores plot 3D and 2D, the k-means plot 2D and the biplot. The plots have to be done in the section *Make Plots* of the page (**Figure 29a**) and can be observed in the section *Visualize Plots* (**Figure 29b**).



(a) Construction of the results plots is made in *Make Plots* section, here with the example for the scree plot.



(b) The plot made can be seen in *Visualize Plots*, sections.

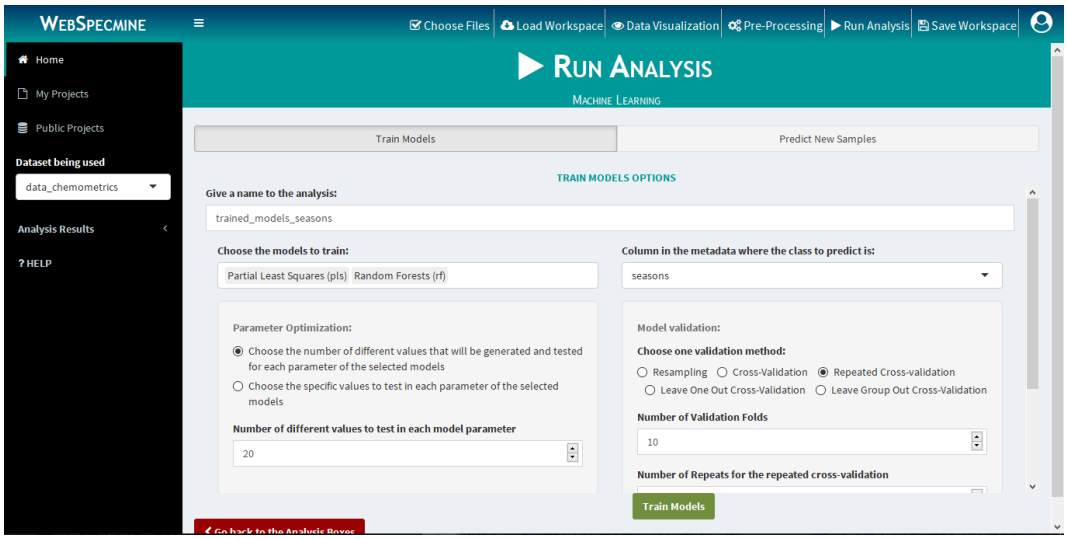
Figure 29: Results of normal PCA on the *data_chemometrics* dataset.

Further details on what this analysis consists on and how this part of the website is implemented are in Afonso (2017).

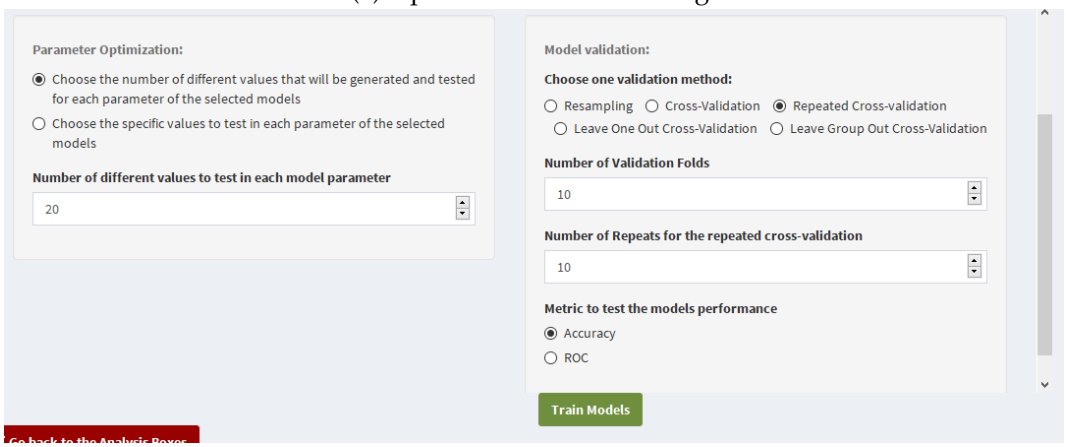
4.1.5 *Machine Learning*

Finally, to build models to discriminate samples by seasons, the user must go back to the "Run Analysis" page and select the "Machine Learning" box.

In the model training performed, where the metadata class to predict represented the *seasons*, the two models used in the study were chosen, including PLS and random forests. For hyper-parameter optimization, 20 different values were chosen to test for each parameter of the selected models. For model validation, the repeated cross-validation method was selected, using 10 validation folds and 10 repeats, and the metric used was accuracy. The setting of these options is shown in **Figure 30**. The name given to this analysis was *trained_models_seasons*.



(a) Options for model training.



(b) Parameter optimization and model validation options.

Figure 30: Model training of the two models: PLS and random forests, using the *data_chemometrics* dataset, for the metadata class *seasons*.

After clicking the "Train Models" button, the site redirects to the results page when the analysis is concluded. Here (**Figure 31**), the user is able to see all the results regarding the trained models, by having a summary table at the top where he/she can check what model shows the best results and see more extensive results for each model trained.

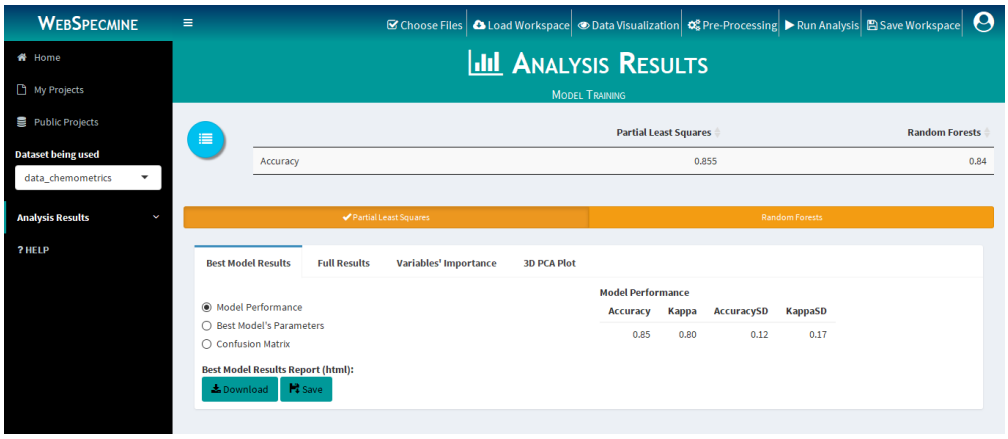


Figure 31: Results of the model training performed on the *data_chemometrics* dataset.

The chemometrics results here obtained revealed to be very similar to the ones obtained in the article. As regards to the one-way ANOVA, the 20 best peaks obtained in the article correspond to the ones obtained through the reproduced pipeline, although the numerical values, like p-value and FDR corrected p-value, changed slightly, as well as the TuckeyHSD test. As regards to the machine learning results, they show some differences, especially at the number of components in the best PLS model, which are 8 in the reproduced pipeline and 19 in the article. However, the results regarding the variables' importance are very similar.

These observed differences may come from the fact that the pre-processing pipeline applied in this work was not entirely equal to the one used in the article. In fact, the removal of specific data variables was performed before peak alignment by the authors of the article, as opposed to in this work, due to the website being developed in such a way that the peak alignment is performed right after reading the submitted data files.

The chemometrics analysis here demonstrated can also be done for the metadata variable regions.

4.1.6 *Metabolite Identification*

To perform the identification of metabolites present in the samples, the user must change the selected "Dataset being used", in the sidebar panel, to the dataset *data_ID*. With this done, the user can now go to the "Run Analysis" page and enter the box "Metabolite Identification" and set the options to perform this analysis, as shown in **Figure 32**.

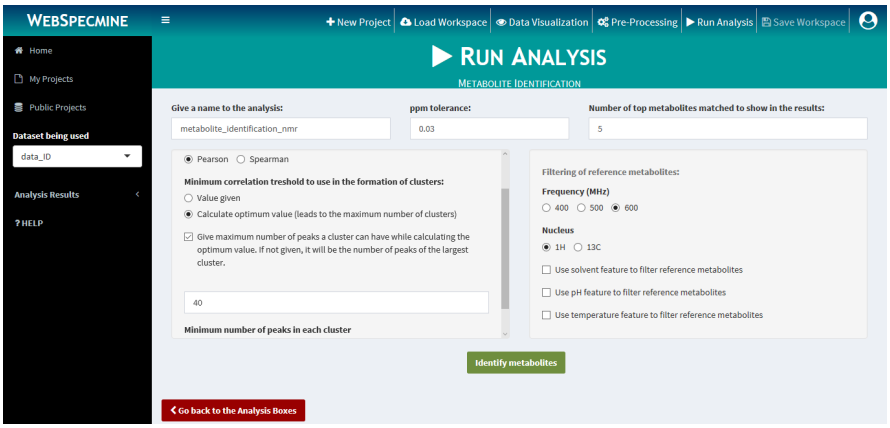


Figure 32: Options of metabolite identification on the *data_ID* dataset.

A ppm tolerance of 0.03 was chosen, as well as a maximum of 10 top metabolites matched for each cluster. Regarding the parameters for the construction of clusters, the Pearson correlation was chosen and it was set a minimum of 1 peaks in each cluster, in order to also obtain the metabolites that were present in the samples and only have one peak in their respective spectra. The optimum minimum correlation threshold to use in the cluster formation was calculated, by setting the maximum number of peaks in a cluster at 40. As regards to the filtering of reference metabolites, the only options set were the frequency (600 MHz) and the nucleus (1H). The name given to this analysis was *metabolite_identification_nmr*.

After clicking the "Identify Metabolites" button, the website redirects the user to the respective results page when the analysis is concluded. Here (**Figure 33**), the user is able to see all the results obtained from the identification of metabolites.

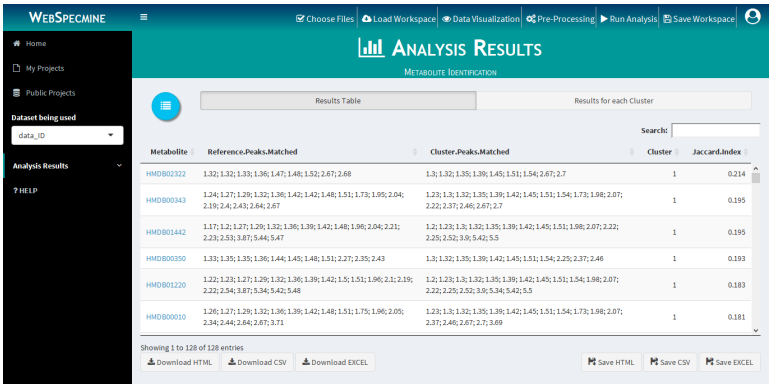


Figure 33: Results of the metabolite identification performed on the *data_ID* dataset.

In **Table 8** is present the metabolites with the best scores identified by the website, after filtering the metabolites that did not make sense to be present in plant data, as the reference metabolites were extracted from HMDB, a database that only takes into consideration metabolites that occur in the human body, as mentioned before.

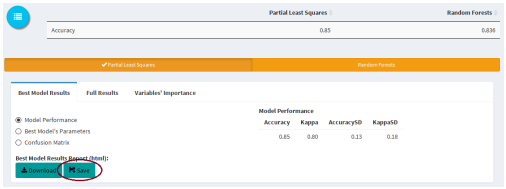
Table 8: Identified metabolites in propolis samples with the best scores.

HMDB code	Name	Reference Peaks Matched	Cluster Peaks Matched	Score
HMDB03099	1-Methyluric acid	3.28	3.27	1
HMDB05807	Gallic acid	7.034	7.03	1
HMDB02428	Terephthalic acid	7.879	7.87	1
HMDB01394	Heparin	3.36; 3.36; 3.39; 3.4; 3.44; 4.01; 4.09; 4.19; 4.22; 4.25; 4.33; 4.37; 4.38	3.33; 3.36; 3.39; 3.42; 3.45; 4.02; 4.08; 4.17; 4.2; 4.25; 4.34; 4.38; 4.41	0.277
HMDB00058	Cyclic AMP	4.31; 4.31; 4.52; 4.55	4.28; 4.31; 4.55; 4.58	0.25
HMDB03066	Chalcone	7.43; 7.53; 7.53; 7.62; 7.64; 7.74; 7.75; 7.78; 8.07	7.42; 7.53; 7.56; 7.63; 7.67; 7.71; 7.78; 7.81; 8.08	0.237
HMDB00045	Adenosine monophosphate (AMP)	4.01; 4.02; 4.35; 4.36; 4.5; 4.5; 6.11; 6.12	4.02; 4.05; 4.34; 4.38; 4.5; 4.53; 6.09; 6.13	0.222
HMDB00108	Ethanol	3.63; 3.64	3.6; 3.66	0.222
HMDB02322	Cadaverine	1.32; 1.32; 1.33; 1.36; 1.47; 1.48; 1.52; 2.67; 2.68	1.3; 1.32; 1.35; 1.39; 1.45; 1.51; 1.54; 2.67; 2.7	0.214
HMDB00229	Nicotinamide ribotide	4.02; 4.02; 4.05; 4.19; 4.19; 4.22; 4.47; 4.61; 9.33; 9.58	4.02; 4.05; 4.08; 4.17; 4.2; 4.25; 4.5; 4.63; 9.35; 9.59	0.204
HMDB01003	Adenosine phosphosulfate	4.23; 4.23; 4.39; 4.4; 4.49; 4.5; 6.12; 6.13	4.2; 4.25; 4.38; 4.41; 4.5; 4.53; 6.09; 6.13	0.19
HMDB02040	4-Methoxycinnamic acid	3.83; 6.96; 7.53; 7.53; 7.61; 7.64	3.81; 6.99; 7.53; 7.56; 7.63; 7.67	0.188
HMDB01220	Prostaglandine E2	1.22; 1.23; 1.27; 1.29; 1.32; 1.36; 1.39; 1.42; 1.5; 1.51; 1.96; 2.1; 2.19; 2.22; 2.54; 3.87; 5.34; 5.42; 5.48	1.2; 1.23; 1.3; 1.32; 1.35; 1.39; 1.42; 1.45; 1.51; 1.54; 1.98; 2.07; 2.22; 2.25; 2.52; 3.9; 5.34; 5.42; 5.5	0.183
HMDB00095	Cytidine monophosphate (CMP)	3.99; 4.03; 4.05; 4.23; 4.23; 4.32; 6.12; 6.13	4.02; 4.05; 4.08; 4.2; 4.25; 4.34; 6.09; 6.13	0.182
HMDB00902	NAD	4.2; 4.214; 4.221; 4.35; 4.353; 4.397; 4.484; 4.502; 6.084; 6.117; 8.142; 9.332	4.17; 4.2; 4.25; 4.34; 4.38; 4.41; 4.5; 4.53; 6.09; 6.13; 8.14; 9.35	0.182
HMDB03403	Amylose	3.59; 3.66	3.6; 3.66	0.182
HMDB00153	Estriol	1.19; 1.2; 1.27; 1.29; 1.32; 1.38; 1.4; 1.42; 1.71; 2.08; 2.19; 2.22; 2.67; 2.69	1.2; 1.23; 1.3; 1.32; 1.35; 1.39; 1.42; 1.45; 1.73; 2.07; 2.22; 2.25; 2.67; 2.7	0.177
HMDB01413	Citicoline	4.16; 4.17; 4.17; 4.25; 4.31; 4.37; 4.38; 6.11; 6.12	4.13; 4.17; 4.2; 4.25; 4.34; 4.38; 4.41; 6.09; 6.13	0.176
HMDB00269	Sphinganine	1.29; 1.31; 1.32; 1.39; 1.4; 1.42; 1.53; 1.54; 2.71; 3.71	1.3; 1.32; 1.35; 1.39; 1.42; 1.45; 1.51; 1.54; 2.7; 3.69	0.175
HMDB02364	Oleanolic acid	1.2; 1.23; 1.3; 1.32; 1.34; 1.39; 1.41; 1.43; 1.48; 1.51; 2.73; 5.16	1.2; 1.23; 1.3; 1.32; 1.35; 1.39; 1.42; 1.45; 1.51; 1.54; 2.7; 5.13	0.174
HMDB01273	Guanosine Triphosphate (GTP)	4.25; 4.28; 4.58; 4.59	4.28; 4.31; 4.55; 4.58	0.174
HMDB01886	3-Methylxanthine	3.51	3.48	0.167
HMDB01644	D-Xylulose	4.03; 4.04; 4.17; 4.18; 4.36; 4.37; 4.38	4.02; 4.05; 4.17; 4.2; 4.34; 4.38; 4.41	0.163
HMDB00192	L-cystine	3.36; 3.37; 3.39; 3.39; 4.09; 4.1	3.33; 3.36; 3.39; 3.42; 4.08; 4.13	0.158
HMDB05794	Quercetin	6.887; 7.537; 7.54; 7.682; 7.686	6.88; 7.53; 7.56; 7.67; 7.71	0.156

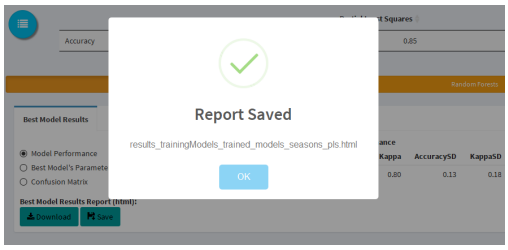
By comparing the results obtained, only two of the identified metabolites in the article seem to be identified in the website, gallic acid and quercetin. This can be due to the use of a different reference library of metabolites, as the authors in the article used spectra that were being obtained by the group previously to the study, with the exact same conditions as the ones used to obtain the propolis data, and 2D-NMR spectra to help in the identification. Furthermore, some metabolites identified in this article may not be identified through the website, as they may not occur in the human body. However, many other metabolites were identified that were not present in the article results, due to the fact that identification here performed took into consideration all samples peaks and the reproduced study only analysed the most important peaks, obtained through the ANOVA analysis. And, in fact, by taking a closer look at these metabolites, some of them were identified in propolis samples in previous studies (Marcucci, 1995; Ankovaa et al., 2000; Bittencourt et al., 2015; Mujica et al., 2017), such as chalcone derivates, 4-Methoxycinnamic acid, D-Xylitol and Benzaldehyde.

4.1.7 *Save Reports*

If wanted, the user is able to save reports and results tables of different sets of results. When possible, buttons to perform this task will be available near the respective result, as it can be seen in **Figure 34a**, where a report of the model training was saved (**Figure 34b**).



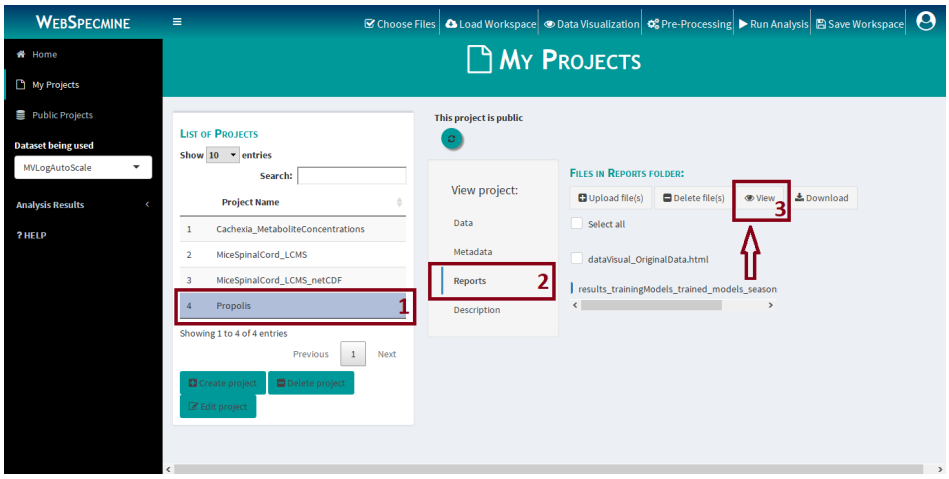
(a) The user needs to click the save button at the bottom of the *Best Model Results* tab of the PLS model to save the HTML report.



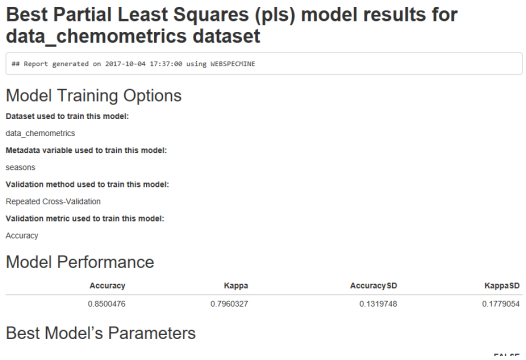
(b) After the report is fully generated, the website sends a pop-up window informing that the report was created successfully.

Figure 34: Demonstration on how to save a report, with the example for the best model results obtained for the PLS model in the machine learning analysis.

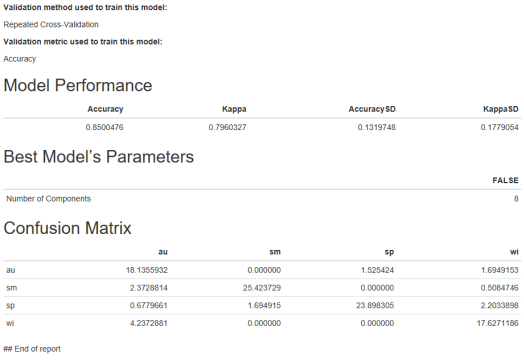
After accessing the project in the users' account, in the "My Projects" page (**Figure 35a**), this report can be seen (**Figure 35b** and **Figure 35c**).



(a) 1-The user must select the *Propolis* project. 2-Then, select the "Reports" tab. At the right, all the reports will appear and the report created has the name mentioned in the pop-up window. 3-The user must select this report and then click in the "View" button.



(b) First half of the created report.



(c) Second half of the created report.

Figure 35: Demonstration of how to see a report, with the example for the results obtained from the best PLS model in machine learning.

4.1.8 *Conclusions*

The chemometrics results obtained by performing the article pipeline through the website are very similar to the original ones, allowing to obtain the same general conclusions about the collected data. A reason that may explain the observed differences is that the pre-processing pipeline applied in this work was not entirely equal to the one used in the article, as stated before.

As regards to metabolite identification, although only two of the identified metabolites in the article were also present in the website results, many other metabolites were identified, including some metabolites that have been discovered before to be part of the propolis composition.

In general, this use case helped validating the website and demonstrated to help accomplish more detailed and aditional results in the metabolite identification.

4.2 LC-MS DATA: MICE SPINAL CORD

4.2.1 *Introduction*

The Saghatelian et al. (2004) study aimed to identify the endogenous substrates of the Fatty Acid Amide Hydrolase (FAAH) enzyme. This enzyme is involved in degrading several neural signaling lipids in the central nervous system (CNS), including the endogenous cannabinoid N-arachidonoyl ethanolamine (anandamide) (Cravatt et al., 2001). Furthermore, this enzyme is emerging as a therapeutic target for treatment of pain (Cravatt and Lichtman, 2003) and neuropsychiatric disorders, such as anxiety (Gaetani et al., 2003).

In fact, previous studies using targeted LC-MS analysis have shown that mice without this enzyme possess highly elevated brain levels of anandamide and other N-Acyl Ethanolamine (NAE) substrates (Cravatt et al., 2001; Clement et al., 2003).

To know all the substrates used by this enzyme, this study aimed to identify the metabolites in both wild-type strains and organisms whose FAAH enzyme was inactivated. With this, the metabolites that showed greater concentration in the second group were considered candidate endogenous substrates for the mentioned enzyme.

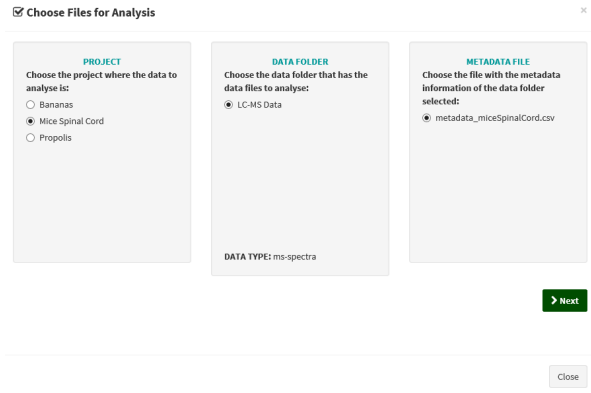
For this, wild-type mice and mice lacking the FAAH enzyme were used, by comparing LC-MS metabolite profiles of spinal cord tissues of these two groups. For each group considered, six samples were taken from six different mice, collecting a total of 12 samples.

The LC-MS data files produced and anlysed in the mentioned study were obtained from the Metaboanalyst website, mentioned previously in **section 2.7**. These consist of 12 CDF files, representing the 12 mentioned samples. These samples were stored in the public project called *Mice Spinal Cord*, under the data folder *LC-MS Data*. Regarding the metadata, there is only one class, named *type*, which consists in distinguishing which samples become from wild-type strains (wt) or not (ko), and the file given is named *metadata_miceSpinalCord*.

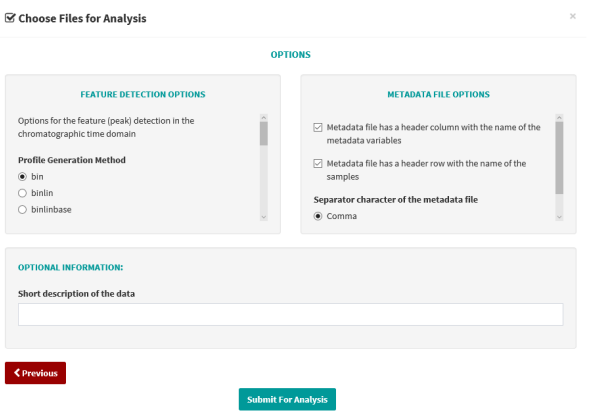
4.2.2 *Choosing files for analysis and pre-processing*

Once again, the user must enter his/her account to copy the public project in question, named *Mice Spinal Cord*, in a similar way to that of the previous use case. After this, the user must click in the "Choose Files" button, present in the header panel, and choose the project in question, and the mentioned data folder and metadata file for analysis (**Figure 36a**). Then, the user must click the "Next" button, which will lead him/her to the window where the options regarding the data and metadata files are set. In this case, the options

are the default ones, so no change is needed (**Figure 36b**). With this, the user is able to click the button "Submit For Analysis" to finalize the submission of the data to analyse.



(a) Choose the project and respective data and metadata files for analysis.



(b) Options regarding the data and metadata files.

Figure 36: Demonstration of how to choose the files from the *Mice Spinal Cord* project for analysis.

After the data files are processed, the user is redirected to the "Run Analysis" page. Here, the user will notice that, in this case, all boxes are accessible. Following this, no pre-processing is applied, as no processing was conducted by the present study and no missing values were encoutered.

4.2.3 *Data Analysis*

The study here reproduced aimed to identify which metabolites were more or less present in the samples from the mutated mice, by comparing the intensities of peaks between both groups. If the differences were significant, it meant that the metabolites that those intensities represent were a substrate or product of the enzyme.

To perform a similar pipeline to the one adopted by this study using the website developed, the user must perform a T-test on the dataset and metabolite identification for LC-MS technique. Thus, the identified metabolites whose query mass corresponds to one of the variables that had a median significantly different between the two groups of samples correspond to the metabolites that are substrates or products of the FAAH enzyme.

T-Test

To perform this analysis, the user must enter in the "Univariate Analysis" box. Once inside, he/she must select the "T-Test" tab, in the tab box located at the left of the page, so that the options regarding this type of analysis appear at the rigth (**Figure 37**). With this, the user can set options including the metadata variable to use, which in this case is *type*. The name

given to this analysis was *TTest_MiceSpinalCord*. The options regarding the results plot can also be chosen, which in this case the p-value treshold was set to 0.05. After seting the options, the user must click the button "Submit", to start the analysis wanted.

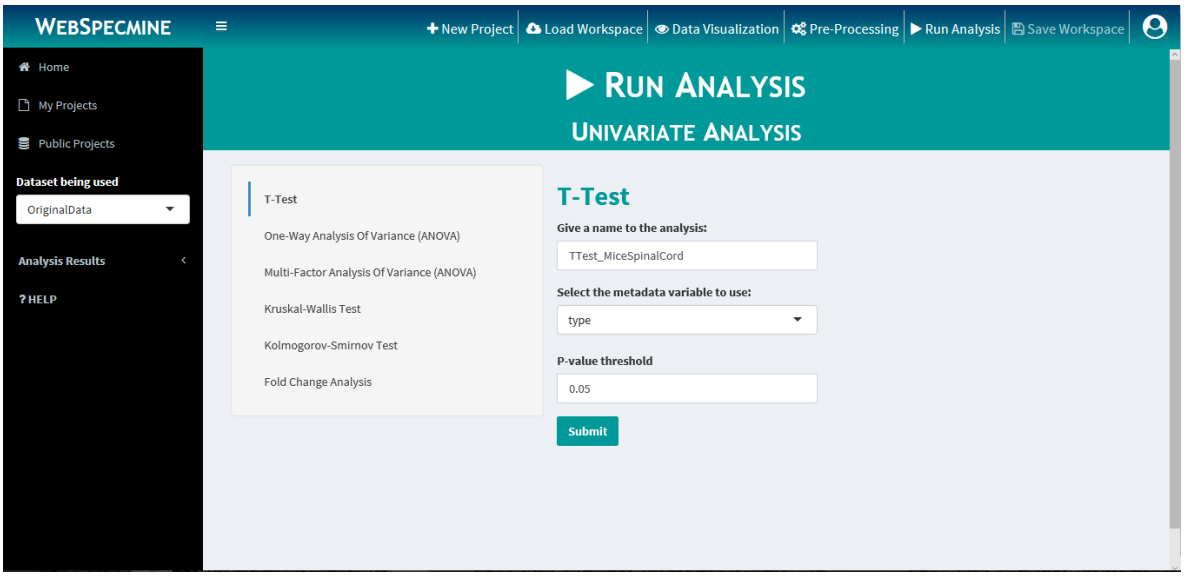


Figure 37: Options of the T-Test on the dataset.

Once this analysis is finished, the website redirects the user to the corresponding results page, seen in **Figure 38**, where the user can see a table with the T-Test results, where the different variables are ordered by decreasing corrected p-value (FDR method), and the results plot.

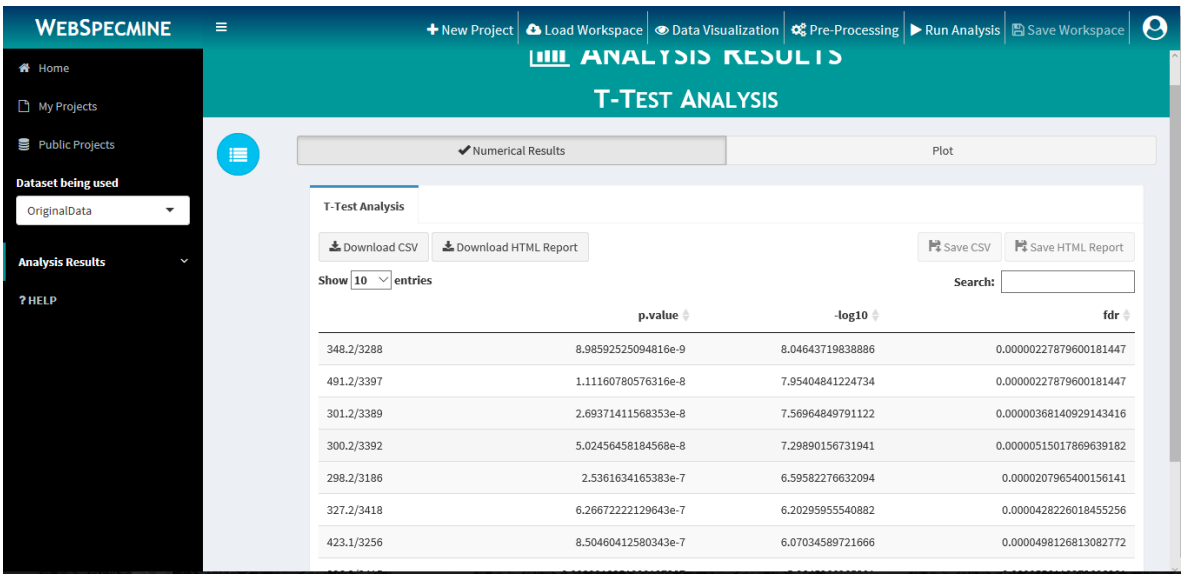
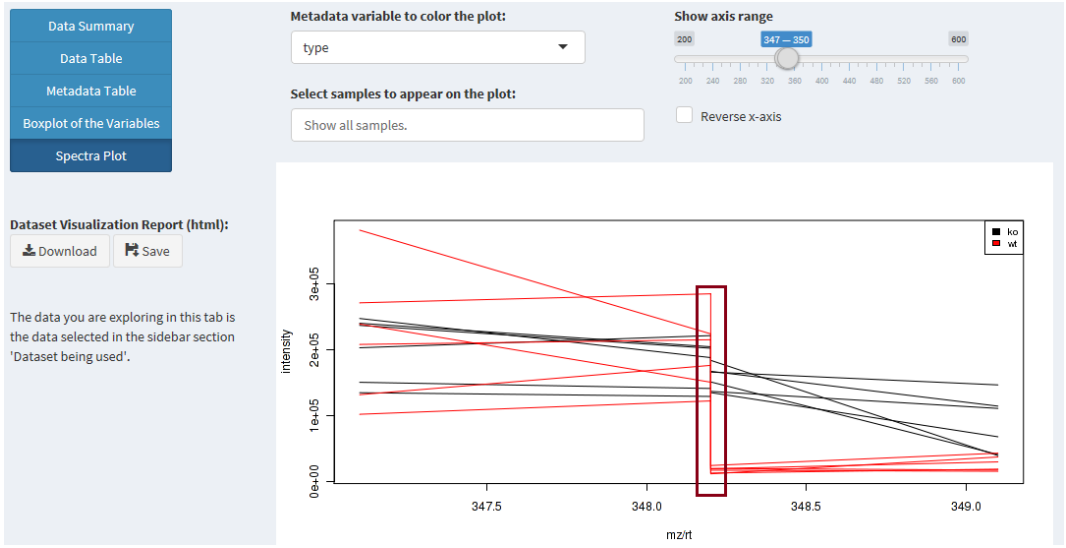


Figure 38: Results of the T-Test on the dataset.

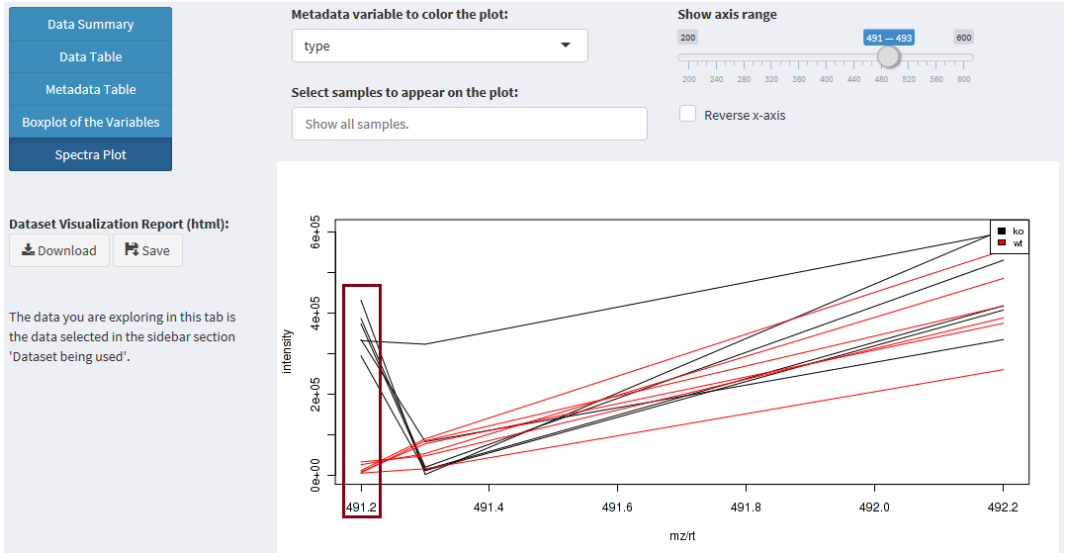
Further details on what this analysis consists on and how this part of the website is implemented are in Afonso (2017).

So that the user has a visual perception on how much the variables are different in terms of intensity between the two groups of samples, he/she can go to the "Data Visualization" page, accessed through the header panel, and select the tab "Spectra Plot". Here, the user can choose the axis range to see the intensity values of the variables in all samples.

In **Figure 39**, it can be seen the intensity values for the top two variables in the T-Test.



(a) Spectra plot of the dataset between the 347 and 350 masses, with the 348.2 variable highlighted in the red rectangle.



(b) Spectra plot of the dataset between the 491 and 293 masses, with the 491.2 variable highlighted in the red rectangle.

Figure 39: Spectra plots of the dataset, highlighting the two top variables from the one-way ANOVA test.

Metabolite Identification

To perform metabolite identification, the user must go back to the "Run Analysis" page and the analysis boxes and enter the "Metabolite Identification" box. Here, the only option to be set is the metadata class that can help in the identification of the metabolites, which in this case can only be *type*. The name given to this analysis was *MID_MiceSpinalCord*. The setting of these options can be seen in **Figure 40**. The user must then click the "Identify Metabolites" button to start the analysis.

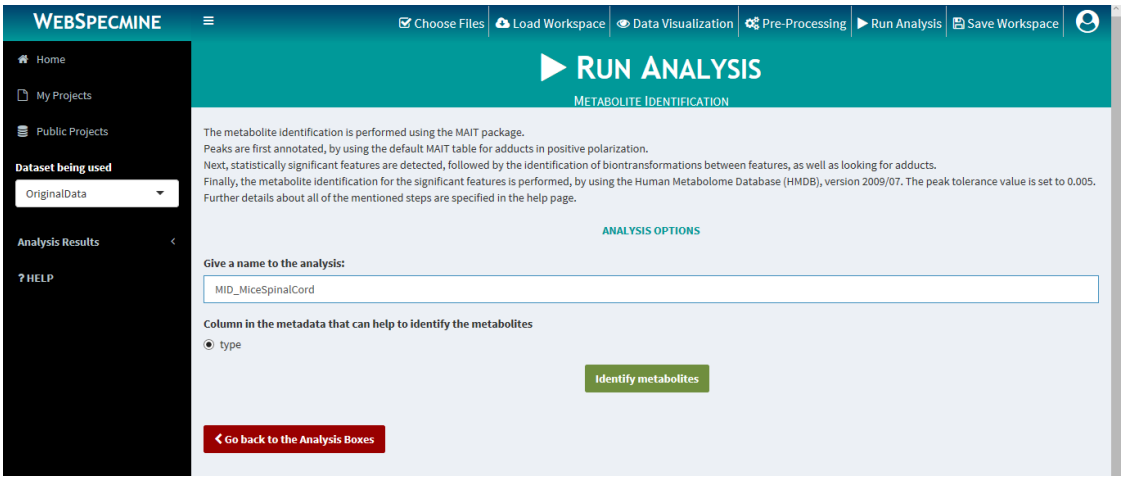


Figure 40: Options of the metabolite identification on the dataset.

After the identification is concluded the website redirects to the corresponding results page, where the user is able to see the table with the results (**Figure 41**).

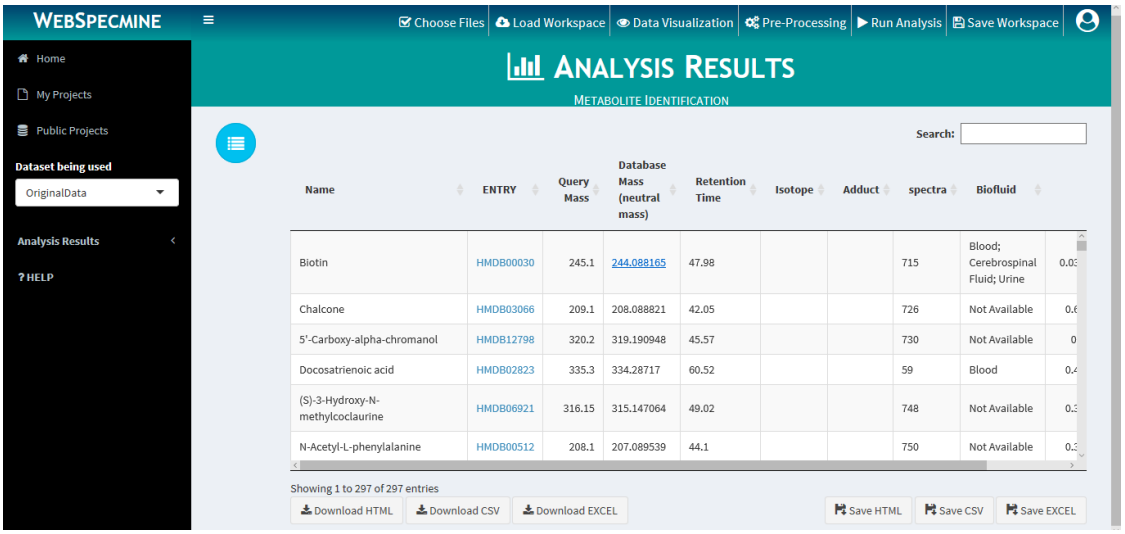


Figure 41: Results of the metabolite identification on the dataset.

4.2.4 *Conclusions*

By comparing the obtained results with the ones present in the article and respective supporting information, it can be concluded that they are very similar. In fact, besides identifying most of the metabolites mentioned in the article, the website provides many more identified metabolites, many from the group of fatty acyls. However, most of these metabolites, not mentioned in the article, are not important as regards to the study in question, as they do not correspond to masses that had intensities significantly different between the two groups of samples.

On the other hand, the identified compound pterin corresponds to an important variable mass obtained from the t-test (322.1). To have a more visual perception on how this mass varied, a closer look was taken to the spectra plot on the area of this variable, similarly to what was done to two variables, mentioned before. This mass revealed to have more intensity in the samples that lacked the FAAH enzyme and have intensities very close to zero in the wild type samples. This could be an important information, as this compound is known to be able to function as a cofactor in enzyme catalysis.

Furthermore, ceramides and phospholipids were also identified, as in the paper, and revealed that, in fact, they did not pratically changed between samples. As regards to monoacylglycerols, many were identified, besides the ones mentioned in the article. The last ones corresponded to the masses obtained in the t-test, corroborating the information present in the article.

Finally, no N-acyl taurines and NAEs seemed to be identified through the website, which were expected to correspond to the majority of the variables obtained in the t-test. However, this may be due to the reference library not having reference spectra for these metabolites, as the majority of the masses obtained in the t-test were not matched with a metabolite.

5

CASE STUDY: BANANA (*MUSA SPP*)

In this chapter, a case study making use of a wide range of the package and website functionalities was performed, not only to further validate the implemented website, but also to gain insights on the product here analysed.

Banana peels are well recognized as a source of important bioactive compounds, such as phenolics, carotenoids, biogenic amines, among others. As such, they have recently started to be used for industrial purposes. However, its composition seems to be strongly affected by biotic or abiotic ecological factors. Thus, this study aimed to investigate banana peels chemical composition, not only to get insights on eventual metabolic changes caused by the seasons, in southern Brazil, but also to identify the most relevant metabolites for these processes. To achieve this, a Nuclear Magnetic Resonance (NMR)-based metabolic proling strategy was adopted, followed by chemometrics analysis, using the specmine package for the R environment, and metabolite identification. The results showed that the metabolomic approach adopted allowed identifying a series of primary and secondary metabolites in the aqueous extracts investigated. Besides, over the seasons the metabolic profiles of the banana peels showed to contain biologically active compounds relevant to the skin wound healing process, indicating the biotechnological potential of that raw material.

This study was published in Practical Applications of Computational Biology & Bioinformatics (PACBB) Proceedings and the extended version, here reproduced, was accepted in the Journal of Integrative Bioinformatics (JIB).

5.1 INTRODUCTION

Musa is a genus, part of the *Musaceae* family, that includes dessert bananas and plantains. Bananas are one of the leading fruit crops as source of energy worldwide and mainly for people living in the humid tropical regions (Pereira and Maraschin, 2015; Padam et al., 2014). In this scenario, Brazil is one of the top producers of banana in the world (Padam et al., 2014).

Traditionally, bananas are known as source of bioactive compounds able to promote wound healing, mainly from burns, and to help overcome or prevent illnesses such as

depression (Balbach, 1945; Kumar et al., 2012). Although biologically active compounds typically occur in small quantities in plant biomasses (Kris-etherton et al., 2002), bananas have been recognized as a source of important pharmacologically-active compounds.

These compounds include phenolics, such as gallic acid and derivatives (Nguyen et al., 2003), which are secondary metabolites needed for normal cell growth and development (Szajdek and Borowska, 2008). Indeed, phenolic compounds are bioactive compounds known for their health benefits (Cook and Samman, 1996), as they have antioxidant properties and important biological effects, such as antibacterial and antiviral, vasodilatory, and protection against ultra-violet radiation, among others (Szajdek and Borowska, 2008; Cook and Samman, 1996). Phenolics are vital for human health, due to their antioxidant and chelating properties, and antimutagenic and antitumoral effects (Pereira and Maraschin, 2015). However, in excess, they can limit the bioavailability of proteins, in the gastrointestinal tract (Szajdek and Borowska, 2008).

Carotenoids, such as β-carotene and xanthophylls (Subagio et al., 1996), are accessory pigments in photosynthesis and have been claimed to cause immune-enhancement and reduction of the risk of developing degenerative diseases, for instance (Krinsky and Johnson, 2005; Tapiero et al., 2004; Voutilainen et al., 2006). Besides, carotenoids such as α- and β-carotene and β-cryptoxanthin are relevant constituents of the diet in less favoured populations in certain countries in south hemisphere (Erdman et al., 1993). In bananas, carotenoids are usually found in higher concentrations in the peel than in the pulp (Rodriguez-Amaya, 2001).

Biogenic amines, such as dopamine (DOPA) and L-3, 4 dihydroxyphenylalanine (L-DOPA), play an important role as neurotransmitters in the hormonal regulation of the glycogen metabolism in mammals (Kanazawa and Sakakibara, 2000; Kimura, 1968). Moreover, DOPA, present in both banana pulp and peel, can be used to prevent or treat the Parkinsons neurodegenerative disease (Pereira and Maraschin, 2015).

Other bioactive compounds found in bananas include anthocyanins (delphinidin and cyaniding - Seymor (1993)), catechins (gallocatechin and epigallocatechnin - Someya et al. (2002)), and sterols and triterpenes, such as β-sitosterol, stigmasterol, campesterol, and 24-methylene cycloartanol (Knapp and Nicholas, 1969).

Unlike the pulp, banana peels, the main residual biomass of the processing industry, are normally used for animal feeding, as organic fertilizers or simply discarded (Charrier et al., 2014). The latter can cause serious environmental problems, as this product is a rich source of nutrients, like nitrogen and phosphorus, that can lead to imbalances in the soil and aquatic environments where they are discarded (González-Montelongo et al., 2010). However, banana peels, which represent about 30% of the fruit, have recently started to be used for industrial purposes. Depending on the technology employed, they can be used

either as ingredients for products with therapeutic activity, as functional compounds in human nutrition, prevention and health care (Pereira and Maraschin, 2015).

The use of banana peels for industrial purposes depends on its chemical composition, a trait strongly affected by climatic factors, orchard manage practices, genotype, and harvest time, among others.

In this context, this case study investigated the banana peels chemical composition over the seasons in Southern Brazil, aiming to gain insights regarding eventual metabolic changes occurring along the harvest times of that fruit. For that, it was adopted a typical NMR-based metabolic profiling strategy coupled to metabolite identification and chemometrics tools, including both univariate (ANOVA) and multivariate (PCA and clustering) statistical analysis.

5.2 DATA COLLECTION AND PROCESSING

5.2.1 *Chemicals*

Ultra-pure water was obtained through a reverse-osmosis system (Permution E-10, Curitiba, Brazil). The deuterated solvent D2O was purchased from TediaBrazil (Rio de Janeiro, Brazil) and 3-trimethylsilyl propionic-2, 2, 3, 3-d4 acid sodium salt (98 atom % D - TSP) and deuterium chloride solution (35 wt. % in D2O, 99 atom % D) were obtained from Sigma-Aldrich (Saint Louis, MO, USA).

5.2.2 *Samples*

Thirteen banana peels samples were collected from an agro-ecologically managed orchard, in Biguaçú County (27 29 39 S; 48 39 20 W, altitude 2m), Santa Catarina State, southern Brazil. Three samples were harvested in the autumn (March, April, and May-2011), four in winter (June-2011, July-2010/2011, and August-2011), five in spring (September-2010/2011, October-2010/2011, and November-2010), and one in summer (February-2011). The producing region is characterized for well-marked seasons.

The sampled biomass was collected from ripe fruits, showing a yellow colour throughout the peel. Fruits were washed in tap water and the peels immediately immersed into liquid N_2 to guarantee the metabolic quenching. Further Aqueous Extracts (AE) of the banana peels were obtained as described in Pereira (2014) and lyophilized.

5.2.3 *One-dimensional NMR Spectroscopy*

Lyophilized AE were added of 700 uL D_2O, containing 0.024g % of 3-trimethylsilyl propionic-2, 2, 3, 3-d4 acid sodium salt (98 atom % D - TSP) as internal standard, vortexed (3x), and centrifuged (4000 rpm/10min), followed by recovering the supernatant (650 uL) and transferring it to 5mm-NMR tubes. The pH of the samples was adjusted to 3.45 with a deuterium chloride solution (35 wt. % in D_2O, 99 atom % D). The unidimensional NMR spectra (1H-NMR) were recorded in a Varian Inova 500 MHz NMR spectrometer and the chemical shifts (δ, ppm) were referenced to the TSP peak at δ(1H) 0.00ppm. Data acquisition used a Dell workstation and the VNMRJ software, running on Windows 7 platform. Briey, 1H-NMR spectra acquisition parameters were as follows: 300 K, no spinning, spectral window 5995.7 Hz, acquisition time 4s, complex points 32983, scans 32, steady state 4, receiver gain 10, relaxation delay 6s, observe pulse 8.18us at a power compression 59/0.98, mixing time 100ms for saturation of water (δ=4.87ppm, Watergate pulse), and digital resolution 0.08657Hz.

5.2.4 *Data Processing*

The 1H-NMR spectra were processed using the ACD/NMR processor software (Advanced Chemistry Development, release 12.0) consisting of zero filling, Fourier transforming the 32K data points, and automatically phased (Pho and Ph1). The baseline was manually corrected and all spectra referenced to the internal standard (TSP, 1H - 0.00ppm). The spectroscopy information of interest was exported as a CSV le containing a matrix with the chemical shifts (1H ppm) and a peak intensity list. Typical resonance regions of the water and internal standard (TSP) signals removed from the dataset for further analysis.

5.3 DATA ANALYSIS

The metadata considered in this study were the seasons of collection. However, as it was only possible to obtain one sample for the summer period, only 3 seasons were considered for data analysis. The season variable was therefore assigned as spring for samples from September and October (years 2010 and 2011) and November-2010; as summer/autumn for the samples from February, March, April, and May-2011; and, finally, as winter for the June-2011, July-2010/2011, and August-2011 samples.

Both files and results were stored in a public project, *Bananas*, in the *Webspecmine* site. Each obtained sample was stored in a CSV file, with the chemical shifts, in ppms, in the first column, and the respective intensities in the second one, under the data folder named *banana_nmr_1*. Each file is named with the month and year of sample collection. A CSV file, named *metadata_nmr_1.csv*, with the metadata was also generated, providing the season

to which each sample belongs. Furthermore, a workspace is available to load, with all the results obtained in this study.

A report generated using the *RMarkdown* package, with all the analysis performed in this study, is present in http://darwin.di.uminho.pt/pacbb2017/banana-nmr, given as supplementary material by the study article here reproduced.

During data submission for analysis, the alignment of peaks used a moving window of 0.03 ppm.

5.3.1 *Chemometrics Analysis*

To perform this type of analysis, the data submitted (under the dataset name *OriginalData*) was processed with missing values imputation, with a constant value of 0.0005, followed by logaritmic transformation and auto-scaling. This new dataset, generated after the processing pipeline was applied, was stored under the name *processed_data*. It was with this last dataset that the chemometrics analysis was performed.

The analysis started with hierarchical clustering on samples, by using the euclidean distance between samples and the agglomeration method "complete" and stored under the name *processed_data_HClust*, to evaluate how close samples are inside each season and between seasons. The former parameter was chosen because it tends to find compact clusters of approximately equal diameters and does not force clusters together due to single elements being close to each other, like in single linkage clustering. After that, K-means clustering was performed, by separating samples in three (*processed_data_K-Means_3*) or four groups (*processed_data_K-Means_4*), in order to observe if samples would be well grouped according to their seasons, either by the three seasons considered or by the four actual seasons. After this, ANOVA was performed, in conjunction with TukeyHSD, to test the difference in means between the seasons for each one of the variables. The analysis was stored under the name *processed_data_OneWay_ANOVA*. Finally, PCA analysis was conducted, whose name is *processed_data_PCA*, by scaling and centering the variables too.

By observing the spectral profiles obtained for each sample collected, we could see that the samples have, approximately, the same peaks. When constructing the mean 1D-NMR spectra for each group considered (**Figure 42**), we observed that the peaks did not vary much between seasons. Despite this, the intensity of some peaks varied. Thus, although the banana samples collected seems to have the same metabolites, their concentrations vary from season to season.

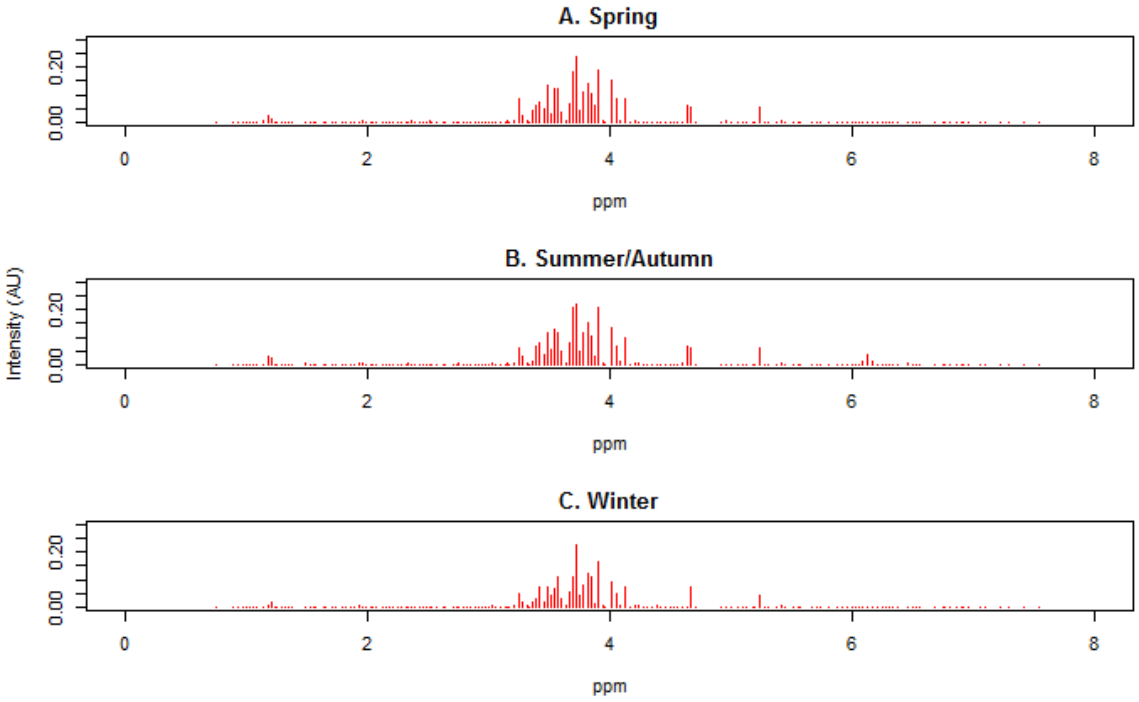


Figure 42: 1D-NMR mean spectra plots for each season. Each plot was obtained from the ppm mean of the different samples for each season. A - Spring season. B - Summer/Autumn season. C - Winter season.

Looking at the hierarchical clustering performed, shown in **Figure 43**, the samples were grouped in the considered seasons quite well. The samples from winter grouped very closely, except for the sample from July 2011, that was closer to the samples from May 2011 (summer/autumn) and September 2011 (spring). Adding this, the other three winter samples seem to be closer to the samples from April 2011, from summer/autumn, and October 2011, spring, than these samples are to their groups. Furthermore, the spring samples were also well grouped, apart from the already mentioned ones from September 2011 and October 2011. On the contrary, the summer/autumn group revealed to be the most difficult one to group, not just regarding the summer sample in comparison to the autumn samples, but also between the autumn samples analysed.

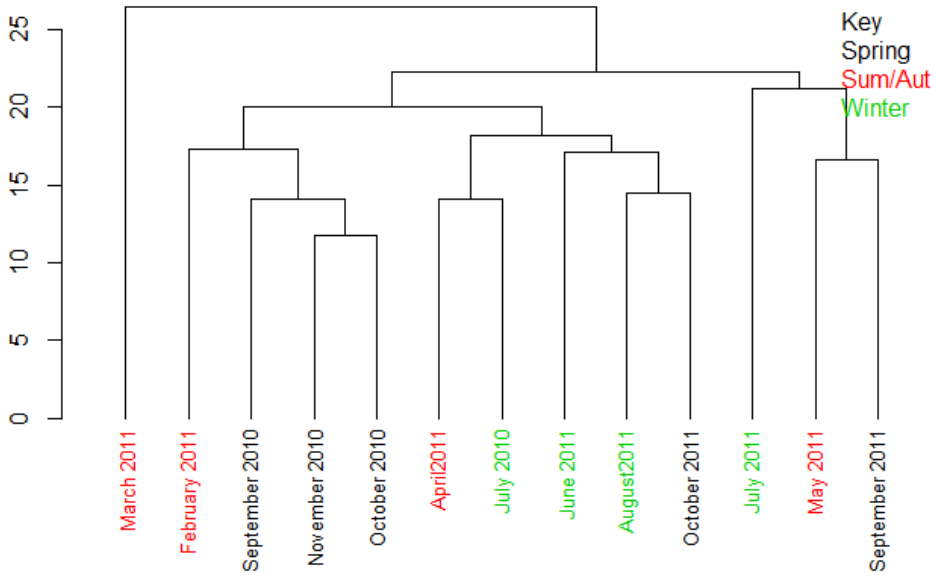


Figure 43: Dendrogram plot of the hierarchical clustering, with euclidean distance between samples. Spring samples are in black, Summer/Autumn samples in red and Winter samples in green.

Another clustering method performed was K-means clustering. Two executions of this method were performed, one grouping the samples into three groups and the other into four. The results from both these approaches are present in **Figure 9**.

Table 9: K-means clusters for clustering into 3 groups (K=3) and into 4 groups (K=4). Spring samples are in black, Summer/Autumn samples in red and Winter samples in green.

K-means Clustering	K=3	K=4
Clusters	June 2011 | July 2010 July 2011 | August 2011 October 2011	June 2011 | July 2010 July 2011 | August 2011 October 2011
	May 2011 | September 2011	May 2011 | September 2011
	February 2011 | March 2011 April 2011 | September 2010 October 2010 | November 2010	February 2011 | March 2011
		April 2011 | September 2010 October 2010 | November 2010

When performing K-means clustering, all winter samples were grouped together, along with the sample from October 2011 (spring). This was an expected result, after analysing the previous results from the hierarchical clustering. The samples from May 2011 and September 2011 were also grouped together in one of the clusters. This similarity was also previously observed in the hierarchical clustering.

When performing K-means clustering to separate samples in three different groups, the remaining summer/autumn and spring samples were grouped together. This is a result that could be expected, as autumn and spring are two seasons that normally might show similar climate conditions in southern Brazil. Furthermore, the difference for the K-means clustering for four different groups lies on the separation of this last cluster into two clusters: February 2011 and March 2011 form a separate cluster from the other samples.

Some of these inconsistencies in separating the samples in the right groups may be due to abnormal climatic conditions from the months in question. In fact, according to CPTEC/INPE (Center for Weather Forecasting and Climate Research/National Institute of Space Research) (http://clima1.cptec.inpe.br/monitoramentobrasil/pt) in October 2011 were recorded abnormally lower precipitation, while the months of June and July showed abnormally higher precipitation, revealing values very similar, which are not typical. Therefore, this might have been one of the factors to why October 2011, a spring sample, was grouped closer to the winter samples in both clustering methods.

Looking at the results from ANOVA, few were the peaks that showed a low corrected p-value, with the FDR method. Three peaks showed a corrected p-value below 0.1, as it can be observed in **Table 10**.

Table 10: ANOVA results for the peaks with the best corrected p-values (FDR method). The first column contains the considered peaks, the second one the respective corrected p-value, and the final column the result of the Tukey's test, which consists on the pair of groups that were signicantly dierent in terms of means for each peak.

Peaks	FDR	Tukey Result
1.89	0.0809	Spring - Winter; Summer/Autumn-Winter; Summer/Autumn-Spring
4.01	0.0809	Spring-Winter; Summer/Autumn-Winter
4.05	0.0809	Spring-Winter; Summer/Autumn-Winter

Furthermore, all peaks have means significantly different between the pairs of groups spring and winter, and summer/autumn and winter. Only the peak at 1.89 ppm showed a mean significantly different between summer/autumn and spring. This seems to corroborate some observations taken from the previous analysis performed i.e., the winter and spring samples are the ones that are better distinguished, and the summer/autumn samples are not so well distinguished from the spring samples. The three peaks occur in the aliphatic (1.89 ppm) and anomeric (4.01 and 4.05 ppm) regions of the 1D-NMR spectrum.

On the other hand, the PCA analysis showed that the first 3 components generated led to a cumulative explanation of more than 50% of the data variability, as it can be seen in **Figure 44**. Only the first component is already able to explain more than 20% of the data variability.

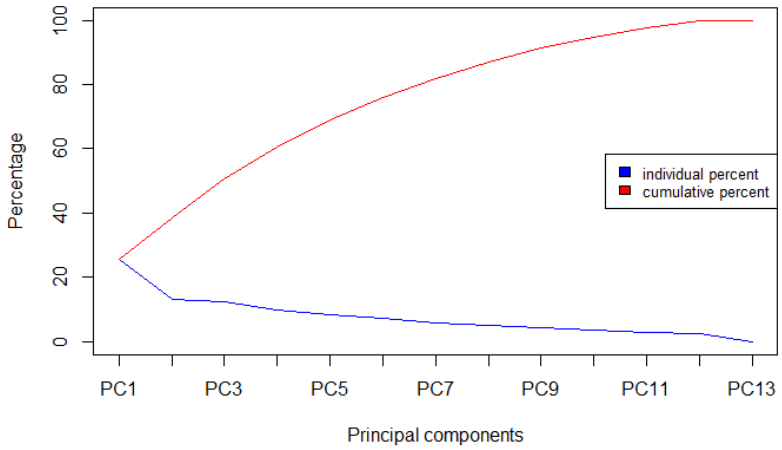


Figure 44: Screeplot of the PCA, showing the percentage of explained data variability for each principal component obtained. The blue line corresponds to the individual percentage and the red one to the cumulative percentage.

By further observing the pairs plot generated by these first 3 components (**Figure 45**), we can realize that PC1 allows the distinction of the winter and spring groups from summer/autumn, while PC2 leads to the distinction of the groups spring and summer/autumn from winter, and the PC3 discriminates the winter and summer/autumn seasons from spring.

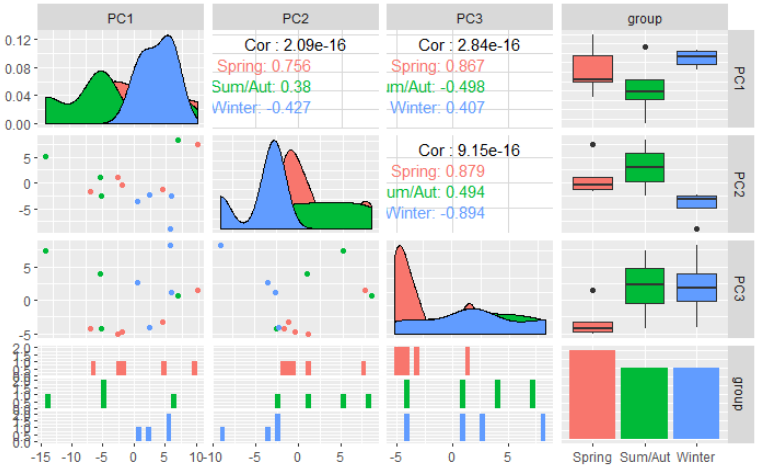


Figure 45: PCA pairs plot of the first 3 components. The variables in pink correspond to the spring group, the green ones to the summer/autumn group and the blue to the winter group.

5.3.2 *Metabolite Identification*

To perform metabolite identification, the *OriginalData* dataset was differently processed. For this, the missing values were replaced with a constant value of 0.0005, generating the dataset used in the identification, named *noMissingValues_dataset*.

A peak tolerance of 0.03 ppm was used to find the metabolites in the samples as in Jacob et al. (2013) it showed to be the best threshold value to use with the reference metabolites from the HMDB database. Furthermore, the optimum correlation value for the formation of clusters was calculated and the Pearson correlation used. Only the metabolites' spectra from 1D-NMR technique (nucleus feature chosen was "1H") obtained with a frequency of 500 MHz from the library were used. For each cluster, the top 5 reference metabolites with the best score were obtained. The name given to this analysis was *metabolite_ID_500_1H*.

Additionally, metabolite identification was performed by visual inspection of the resonances both in unidimensioal (^{1}H-NMR) and bidimensional (^{1}H/^{1}H, TOCSY and ^{1}H/^{13}C, HSQC) NMR spectra.

Twenty-three metabolites have been detected as shown in **Table 11**.

None of the three peaks that revealed the lowest corrected p-value in the ANOVA analysis were found in a cluster. Because it was chosen to only take into consideration the clusters with two or more related peaks, this could mean that each one of these peaks belongs to metabolites whose spectra is only composed by one peak. In fact, after submitting each of these peaks to HMDB, the peak 1.89 matched, with a ppm tolerance of 0.03, with the peak of the acetic acid spectra (1.91 ppm). For the other two peaks, 4.05 and 4.01, the best match found was β-D-fructose (4.08 ppm, m, 3-CH; 4.05 ppm, m, 5-CH) as previously described (Maraschin et al., 2016). However, the correlation between these two peaks was not considered enough, as the optimum correlation value calculated by the code was 0.9, and their correlation was lower. Thus, those resonances have been tentatively assigned to that monosaccharide.

Taking into account a biochemical approach, the polar fraction of the banana peel metabolome recovered by the aqueous extracts revealed to be a chemically complex matrix, containing a series of distinct primary and secondary metabolites, e.g., carbohydrates and derivatives (sugar alcohols, e.g.), amino acids, nitrogenous bases, organic and phenolic acids, vitamin B6 (pyridoxine), and polyamines. Since the number of identified compounds allow one discussing several biochemical relevant issues to, e.g., nutritional, agronomic and pharmacological areas, in line with previous studies of the research group (Pereira and Maraschin, 2015; Pereira, 2014), focus will be put on the metabolites with claimed effects on the wound healing process of cutaneous tissues. Thus, banana peels collected over the seasons in southern Brazil contain important bioactive compounds of interest to human health. In fact, in

Table 11: 1H and 13C chemical shifts and proton multiplicity for assigned compounds found in aqueous extracts of banana peels (cv. Prata Anã) produced in southern Brazil (Santa Catarina State).

Metabolite	HMDB ID	Cluster Peaks Matched	Reference peaks matched ppm δ 1H [multiplicity, assignment - J(H, H) in Hz]/13C	Jaccard Index
Acetic acid	HMDB00042	1.92	1.91 (s, 2-CH3)/24.17 (C-2)	0.5
Adenosine	HMDB00050	6.00; 8.10	6.01 (d, 2-CH, 6.01); 8.09 (s, 7-CH)	0.0909
L-α-aminobutyric acid	HMDB00452	0.99; 1.92	0.97 (t, 7-CH3, 7,60); 1.89 (q, 3-CH2)	0.1538
L-Arabitol	HMDB01851	3.54; 3.57; 3.63; 3.66; 3.69; 3.72; 3.75; 3.78; 3.81; 3.87	3.56 (dd, 3-CH, 8.26, 1.50); 3.65; 3.66 (m, 1-CH2); 3.73; 3.75; 3.82; 3.84 (dd, 5-CH2, 11.70, 2.60)	0.3704
Caffeic acid	HMDB01964	6.34	6.34 (d, 2-CH, 15.80)/115.10 (C-2)	0.0909
α/β-Cellobiose (reducing)	HMDB00055	3.27; 3.39; 3.42; 3.46; 3.48; 3.54; 3.57; 3.63; 3.66; 3.69; 3.72; 3.75; 3.78; 3.81; 3.87; 4.64; 5.24	3.28 (m, 2-CH); 3.37; 3.41; 3.43; 3.45; 3.51; 3.54 (5-CH); 3.60; 3.63 (3-CH); 3.66 (4-CH); 3.74; 3.75; 3.76; 3.81; 3.84 (m, 6-CH); 4.67(d, 1-CH, 7.89); 5.23 (d, 1-CH, 3,64)	0.3148
Citric acid	HMDB00094	2.7	2.67(d, 2, 5-CH2, 15.14)/44.51 (C-2, 5)	0.2
Dimethylglycine	HMDB00092	2.91	2.91 (s,4, 5-CH3)/46.20 (C-4,5)	0.3333
α-L-Fucose	HMDB00174	3.42; 3.46; 3.48; 3.63; 3.66; 3.69; 3.72; 3.75; 3.78; 3.81; 3.87; 5.24	3.44; 3.45; 3.63; 3.64; 3.66; 3.75; 3.76; 3.77; 3.78 (m, 2, 4-CH); 3.85 (dd, 3-CH); 5.21 (d, 1-CH, 3.88)	0.3158
Galactaric acid	HMDB00639	4.24	4.25 (s, 2, 5-CH)	0.25
4-Hydroxybenzoic acid	HMDB00500	6.90; 6.92	6.90; 6.92 (d, 2, 6-CH, 8.75)/133.04 (C-2)	0.5
Itaconic acid	HMDB02092	5.88; 6.34	5.85 (s, CH2); 6.33 (s, CH2)	0.25
Maleic acid	HMDB0000176	6.03	6.04 (s, 2.3-CH)	0.3333
Malic acid	HMDB00744	2.39; 4.30	2.36 (dd, 2-CH2, 15.33, 10.12)/181.20 (C-2); 4.28 (dd, 3-CH, 10.15, 2.82)/73.10 (C-3)	0.1667
Oxaloacetic acid	HMDB0000223	2.39	2.38 (s, 2-CH2)	0.5
Pyridoxine/ pyridoxal/ pyridoxamine	HMDB00239/ HMDB01545/ HMDB01431	2.44; 4.32; 6.55	2.45 (s, CH3); 4.32 (s, 11-CH2); 6.54 (d, 9-CH, 1.95)	0.25/ 0.2/ 0.1
Spermidine	HMDB01257	1.55; 1.57; 1.65; 2.57	1.58; 1.60; 1.63 (m, 3-CH2); 2.56 (m, 4, 2, 6, 9-CH2)	0.1818
Succinic acid	HMDB00254	2.39	2.39 (s, 2-CH3)/33.89 (C-2, 3)	0.5
Succinic acid	HMDB00254	2.42	2.39 (s, 2, 3-CH2)/36.03 (C-2, 3)	0.5
Sucrose	HMDB00258	4.21; 4.24; 5.42	4.20; 4.22 (d, 3-CH, 8.71); 5.42 (d, G1H)/92.97 (C-1)	0.1154
Syringic acid	HMDB02085	3.84	3.84 (s, CH3)	0.25
Trehalose	HMDB00975	3.39; 3.42; 3.46; 3.63; 3.66; 3.72; 3.75; 3.78; 3.81; 3.87	3.42 (4-CH); 3.44; 3.46; 3.63 (dd, 5-CH); 3.74; 3.75 (6-CH); 3.76; 3.79; 3.84 (2-CH)	0.3571
Uracil	HMDB00300	7.54	7.51 (d, 6-CH, 7.65)	0.1667

the Brazilian folk medicine banana peel has a history of utility to promote wound healing when used topically (Balbach, 1945; Pereira and Maraschin, 2015).

An important class of biologically active molecules in banana peels are the phenolic compounds. In the present study, NMR spectroscopy coupled to bioinformatics tools have been able to detect caffeic acid (3, 4-dihydroxycinnamic acid), 4-hydroxybenzoic acid and syringic acid (3, 5-dimethoxybenzoic acid) in the aqueous extracts investigated. The former is originated from the mevolanate-shikimate biosynthesis pathways as the benzoic acids are produced via the loss of a two-carbon moiety from phenylpropanoids. These phenolic compounds possess antioxidant activity and anti-inflammatory properties, as well as are cytotoxic toward certain tumor cell lines (Gaglione et al., 2013). Caffeic acid, for instance, is recognized as a potent antioxidant due to the delocalization of an unpaired electron caused by the extended conjugated side chain. Besides, its o-dihydroxyl group forms a hydrogen bond, which creates a more stable configuration after breaking the O-H bond (Son and Lewis, 2002).

Another class of secondary metabolites found in bananas are the natural amines. At least fifteen bioactive amines and derivatives have been identified in bananas so far. Among these, the polyamines spermidine, spermine and putrescine are commonly found, being spermidine as herein shown the prevalent one, followed by its precursor putrescine (Glória, 2005). These are secondary metabolites with well-known metabolic and physiological functions on the growth and development of plants, as well as presenting neuroactive, psychoactive and vasoactive effects in mammals. Importantly, polyamines are required in higher contents during periods of wound healing and post-surgery recovery (Gould et al., 2008), what could explain in certain extension the positive effects found by the research group (Pereira, 2014) in using standardized aqueous extracts of banana peels in the wound healing of mechanically damage-cutaneous tissue of isogenic Balb/C mice - Mus musculus. In fact, arginine, a dibasic amino acid and its metabolites nitric oxide, proline and polyamines (spermidine, spermine and putrescine, e.g.) are quite important in wound healing, affecting all phases of the process. Through the arginase pathway, L-ornithine, proline and polyamines are produced, the latter directly stimulating cell proliferation (Gould et al., 2008) and favouring the regeneration of damaged tissues. Besides, more recently, it has been shown that expression levels of heparan sulphate in human skin and wounded tissues of mouse are well correlated with polyamine contents (Imamura et al., 2016).

Finally, vitamin B6 is long known to play a critical role in protein metabolism. All the three forms of that water-soluble vitamin, i.e., pyridoxine, pyridoxal and pyridoxamine occur in the human body, being stable under acidic conditions (Liepa et al., 2007). Regarding the effect of pyridoxine on the skin wound healing process, studies in rats have shown that, upon deficiency, a decrease in the content, synthesis and maturation of the skin collagen is

found, with negative effects on the healing of excision and incision wounds (Lakshmi et al., 1988).

5.4 CONCLUSIONS

The results revealed that the distinction of the banana peels metabolic composition according to the seasons is possible, mostly due to the peak intensity, i.e, the concentrations of the metabolites across the different seasons. This distinction is more noticeable in the winter and spring groups, as these were the ones that better grouped in both cluster analysis and showed more signicant differences in means regarding the ANOVA analysis. Furthermore, the PCA analysis revealed that it is only necessary three principal components to explain more than 50% of the data variability.

These results show that the different conditions of the seasons can, in fact, influence the composition of the banana peels. The NMR-based metabolomic analytical strategy herein shown seems to be capable of identifying the chemical heterogeneity of banana peels over the harvest seasons, allowing to obtain standardized extracts for further industrial applications. Finally, as regards to metabolite identification a series of primary and secondary metabolites has been detected, shading some light on the chemical complexity of the aqueous extracts investigated. As herein shown, one could easily determine through the metabolomic approach adopted the potential of banana peels as source of active compounds on the skin wound healing (e.g.) upon the effects of seasoning. However, it must be taken into consideration that the metabolite identification pipeline adopted does not consider the clusters only composed by one peak. Therefore, some valuable information may have been lost, as there are many metabolites whose spectra only has one peak. Further analysis on this topic should be done. Furthermore, as the peaks intensities seem to be the main aspect that changes according to the seasons, further analysis on the calculation of concentrations of the identified metabolites should be taken into account as well.

6

CONCLUSIONS AND FUTURE WORK

Many of the tools that nowadays allow metabolomics data analysis require programming skills, being few the ones that are based on a web service, facilitating the analysis of these types of data for people who do not have programming skills. However, the existing web services are specific for a type of analysis or data format, or simply lack diversity in the tools provided.

With this in mind, the development of the present website is a valuable tool to gather various types of analyses in one place. This site provides simple, interactive and easy-to-use tools to perform the analysis of different types of metabolomics data, such as NMR, MS, infrared, Raman, UV-Vis and concentrations data.

The analyses provided include univariate analysis, like t-test and ANOVA, regression and correlation analysis, and unsupervised multivariate analysis, like PCA and clustering.

As regards to supervised multivariate analysis, the users can perform machine learning tasks, which include model training, with models such as PLS, Decision Tree (C4.5), Rule-Based Classifier, Random Forests, SVMs with linear kernel, LDA and Neural Network available, and prediction of new samples, using models previously trained with similar samples. Hyper-parameter optimization of the models is possible. Feature selection is also available, including recursive feature elimination and selection by filter. In both machine learning and feature selection, various validation methods are available, such as resampling, cross-validation, repeated cross-validation, leave one out cross-validation and leave group-out cross-validation.

Finally, regarding metabolite identification, it is possible to identify metabolites from LC-MS samples and, with the newly added functions to the *specmine* package, from 1D-NMR peak lists, by firstly separating the peaks into clusters representing probable metabolites and comparing them to the reference metabolites.

Adding this, users can also perform various pre-processing tasks, visualize data, create reports of the results obtained, get CSV and/or MS EXCEL files of the results tables, as well as save and, if wanted, publicly share projects and work done.

To validate the website created and developed in this work, as well as to allow an explanation on how a pipeline of an analysis can be carried out, previously developed studies

were reproduced, with success. Furthermore, one case study, with the purpose of getting insights on bananas' peels composition and its properties along the seasons, was performed by making use of a wide range of the website and package functionalities, revealing interesting results. This case also helped to further validate the website and functionalities developed. All the analyses performed were stored in projects and made public in the *Webspecmine* site.

With this, the main objectives for this work were accomplished, although there are still many improvements that could be done in the website and functionalities to add the *specmine* package. For instance, functionalities that allow a more extensive interpretation of results obtained are an important feature to add to the *specmine* package and the website, such as pathway analysis, biomarker identification and enrichment analysis. Furthermore, adding more methods to existing analysis is also important, such as further methods to perform feature selection, like the wrapper methods genetic algorithm and forward selection. Also, although the metabolite identification feature was extended to NMR data, the quantification of metabolites identified should also be implemented. It is also needed to add support to read NMR spectral data and subsequent detection of peaks.

BIBLIOGRAPHY

(Afonso, 2017) Telma Afonso. Development of web-based tools for spectral data analysis and mining. Master thesis, University of Minho, 2017.

(Alonso et al., 2015) Arnald Alonso, Sara Marsal, and Antonio Julià. Analytical methods in untargeted metabolomics: state of the art in 2015. *Frontiers in bioengineering and biotechnology*, 3(March):23, 2015. ISSN 2296-4185. doi: 10.3389/fbioe.2015.00023.

(Anandan et al., 2012) P. Anandan, S. Vetrivel, S. Karthikeyan, R. Jayavel, and G. Ravi. Crystal growth, spectral and thermal analyses of a semi organic nonlinear optical single crystal: L-tyrosine hydrochloride. *Optoelectronics and Advanced Materials, Rapid Communications*, 2012. ISSN 18426573.

(Ankovaa et al., 2000) Vassya S B Ankovaa, Solange L D E C Astrob, and Maria C M Arcuccic. Propolis : recent advances in chemistry and plant origin. *Apidologie*, 2000.

(Bai et al., 2014) Yunnuo Bai, Haitao Zhang, Xiaohan Sun, Changhao Sun, and Lihong Ren. Biomarker identification and pathway analysis by serum metabolomics of childhood acute lymphoblastic leukemia. *Clinica Chimica Acta*, 436:207–216, sep 2014.

(Balbach, 1945) Alfons Balbach. *As frutas na medicina doméstica.* MVP, São Paulo, 21 edition, 1945.

(Batista and Monard, 2002) Gustavo E. A. P. A. Batista and Maria C. Monard. A study of k-nearest neighbour as an imputation method. *Frontiers in Artificial Intelligence and Applications*, 2002.

(Bittencourt et al., 2015) Mara L F Bittencourt, Paulo R. Ribeiro, Rosana L P Franco, Henk W M Hilhorst, Renato D. de Castro, and Luzimar G. Fernandez. Metabolite profiling, antioxidant and antibacterial activities of Brazilian propolis: Use of correlation and multivariate analyses to identify potential bioactive compounds. *Food Research International*, 2015. ISSN 09639969. doi: 10.1016/j.foodres.2015.07.008.

(Bleiholder et al., 2011) Christian Bleiholder, Nicholas F. Dupuis, Thomas Wyttenbach, and Michael T. Bowers. Ion mobilitymass spectrometry reveals a conformational conversion from random assembly to β-sheet in amyloid fibril formation. *Nature Chemistry*, 2011. ISSN 1755-4330. doi: 10.1038/nchem.945.

(Castro and Manetti, 2007) Cecilia Castro and Cesare Manetti. A multiway approach to analyze metabonomic data: a study of maize seeds development. *Analytical biochemistry*, 371(2):

194–200, dec 2007. ISSN 0003-2697. doi: 10.1016/j.ab.2007.08.028. URL http://www.ncbi.nlm.nih.gov/pubmed/17904514.

(Chagoyen and Pazos, 2013) Monica Chagoyen and Florencio Pazos. Tools for the functional interpretation of metabolomic experiments. *Briefings in bioinformatics*, 14(6):737–44, 2013. ISSN 1477-4054. doi: 10.1093/bib/bbs055. URL http://www.ncbi.nlm.nih.gov/pubmed/23063930.

(Charrier et al., 2014) André Charrier, Michel Jacquot, Serge Hamon, and Nicolas Dominique. *L'amélioration des plantes tropicales.* CIRAD-Centre de Coopération Internationale en Recherche Agronomique Pour le Développement, France, 2014.

(Chen et al., 2010) Xiaojing Chen, Han Li, Di Wu, Xinng Lei, Xiangou Zhu, and Anjiang Zhang. Application of a hybrid variable selection method for the classification of rapeseed oils based on 1H NMR spectral analysis. *European Food Research and Technology*, 230 (6):981–988, mar 2010. ISSN 1438-2377. doi: 10.1007/s00217-010-1241-7. URL http://www.springerlink.com/index/10.1007/s00217-010-1241-7.

(Clement et al., 2003) Angela B Clement, E Gregory Hawkins, Aron H Lichtman, and Benjamin F Cravatt. Increased seizure susceptibility and proconvulsant activity of anandamide in mice lacking fatty acid amide hydrolase. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 2003. ISSN 1529-2401.

(Cook and Samman, 1996) N.C. Cook and S. Samman. FlavonoidsChemistry, metabolism, cardioprotective effects, and dietary sources. *The Journal of Nutritional Biochemistry*, 7(2):66–76, 1996. ISSN 09552863. doi: 10.1016/S0955-2863(95)00168-9.

(Costa et al., 2016) Christopher Costa, Marcelo Maraschin, and Miguel Rocha. An R package for the integrated analysis of metabolomics and spectral data. *Computer Methods and Programs in Biomedicine*, 129:117–124, jun 2016.

(Costa, 2014) Christopher Borges Costa. Development of an integrated computational platform for metabolomics data analysis and knowledge extraction. Master thesis, University of Minho, 2014.

(Cottret et al., 2010) Ludovic Cottret, David Wildridge, Florence Vinson, Michael P Barrett, Hubert Charles, Marie-France Sagot, and Fabien Jourdan. MetExplore: a web server to link metabolomic experiments and genome-scale metabolic networks. *Nucleic acids research*, 38(Web Server issue):W132–7, jul 2010. ISSN 1362-4962. doi: 10.1093/nar/gkq312.

(Cravatt et al., 2001) B. F. Cravatt, K. Demarest, M. P. Patricelli, M. H. Bracey, D. K. Giang, B. R. Martin, and A. H. Lichtman. Supersensitivity to anandamide and enhanced endogenous

cannabinoid signaling in mice lacking fatty acid amide hydrolase. *Proceedings of the National Academy of Sciences*, 2001. ISSN 0027-8424. doi: 10.1073/pnas.161191698.

(Cravatt and Lichtman, 2003) Benjamin F. Cravatt and Aron H. Lichtman. Fatty acid amide hydrolase: An emerging therapeutic target in the endocannabinoid system, 2003. ISSN 13675931.

(Degtyarenko et al., 2008) Kirill Degtyarenko, Paula De matos, Marcus Ennis, Janna Hastings, Martin Zbinden, Alan Mcnaught, Rafael Alcntara, Michael Darsow, Mickal Guedj, and Michael Ashburner. Chebi: A database and ontology for chemical entities of biological interest. *Nucleic Acids Research*, 36(SUPPL. 1), 2008.

(Eisner et al., 2010) Roman Eisner, Cynthia Stretch, Thomas Eastman, Jianguo Xia, David Hau, Sambasivarao Damaraju, Russell Greiner, David S. Wishart, and Vickie E. Baracos. Learning to predict cancer-associated skeletal muscle wasting from 1H-NMR profiles of urinary metabolites. *Metabolomics*, 7(1):25–34, aug 2010. ISSN 1573-3882. doi: 10.1007/s11306-010-0232-9. URL http://www.springerlink.com/index/10.1007/s11306-010-0232-9.

(Erdman et al., 1993) John Erdman, Tiffany Bierer, and Eric Gugguer. Absorption and Transport of Carotenoids. *Annals of the New York Academy of Sciences*, 691:76–85, 1993.

(Fathi et al., 2014) Fariba Fathi, Laleh Majari-Kasmaee, Ahmad Mani-Varnosfaderani, Anahita Kyani, Mohammad Rostami-Nejad, Kaveh Sohrabzadeh, Nosratollah Naderi, Mohammad Reza Zali, Mostafa Rezaei-Tavirani, Mohsen Tafazzoli, and Afsaneh Arefi-Oskouie. 1H NMR based metabolic profiling in Crohn's disease by random forest methodology. *Magnetic resonance in chemistry : MRC*, 52(7):370–6, jul 2014. ISSN 1097-458X. doi: 10.1002/mrc.4074. URL http://www.ncbi.nlm.nih.gov/pubmed/24757065.

(Fernández-Albert et al., 2014) Francesc Fernández-Albert, Rafael Llorach, Cristina Andrés-Lacueva, and Alexandre Perera. An R package to analyse LC/MS metabolomic data: MAIT (Metabolite Automatic Identification Toolkit). *Bioinformatics*, 30(13):1937–1939, 2014. ISSN 14602059. doi: 10.1093/bioinformatics/btu136.

(Ferry-Dumazet et al., 2011) Hélène Ferry-Dumazet, Laurent Gil, Catherine Deborde, Annick Moing, Stéphane Bernillon, Dominique Rolin, Macha Nikolski, Antoine de Daruvar, and Daniel Jacob. MeRy-B: a web knowledgebase for the storage, visualization, analysis and annotation of plant NMR metabolomic profiles. *BMC Plant Biology*, 2011. ISSN 1471-2229. doi: 10.1186/1471-2229-11-104.

(Gaetani et al., 2003) Silvana Gaetani, Vincenzo Cuomo, and Daniele Piomelli. Anandamide hydrolysis: A new target for anti-anxiety drugs?, 2003. ISSN 14714914.

(Gaglione et al., 2013) Maria Gaglione, Gaetano Malgieri, Severina Pacifico, Valeria Severino, Brigida D'Abrosca, Luigi Russo, Antonio Fiorentino, and Anna Messere. Synthesis and biological properties of caffeic acid-PNA dimers containing guanine. *Molecules*, 2013. ISSN 14203049. doi: 10.3390/molecules18089147.

(García-Alcalde et al., 2011) Fernando García-Alcalde, Federico García-López, Joaquín Dopazo, and Ana Conesa. Paintomics: A web based tool for the joint visualization of transcriptomics and metabolomics data. *Bioinformatics*, 27(1):137–139, jan 2011.

(Glória, 2005) MBA Glória. Bioactive amines. In YH Hui and F Sherkat, editors, *Handbook of Food Science, Technology and Engineering.*, volume 4. Taylor and Francis, Boca Raton, 2005.

(González-Montelongo et al., 2010) Rafaela González-Montelongo, M. Gloria Lobo, and Mónica González. Antioxidant activity in banana peel extracts: Testing extraction conditions and related bioactive compounds. *Food Chemistry*, 119(3):1030–1039, 2010. ISSN 03088146. doi: 10.1016/j.foodchem.2009.08.012.

(Gould et al., 2008) A. Gould, C. Naidoo, and Geoffrey C. Candy. Arginine metabolism and wound healing. *Wound Healing Southern Africa*, 1:48–55, 2008.

(Guyon and Elisseeff, 2003) Isabelle Guyon and André Elisseeff. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research (JMLR)*, 3(3):1157–1182, 2003.

(Hao et al., 2012) Jie Hao, William Astle, Maria De Iorio, and Timothy M D Ebbels. BATMAN–an R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a Bayesian model. *Bioinformatics (Oxford, England)*, 28(15):2088–90, aug 2012. ISSN 1367-4811. doi: 10.1093/bioinformatics/bts308. URL http://www.ncbi.nlm.nih.gov/pubmed/22635605.

(Hao et al., 2014) Jie Hao, Manuel Liebeke, William Astle, Maria De Iorio, Jacob G Bundy, and Timothy M D Ebbels. Bayesian deconvolution and quantification of metabolites in complex 1D NMR spectra using BATMAN. *Nature protocols*, 9(6):1416–27, 2014. ISSN 1750-2799. doi: 10.1038/nprot.2014.090. URL http://www.ncbi.nlm.nih.gov/pubmed/24853927.

(Hastings et al., 2013) Janna Hastings, Paula De Matos, Adriano Dekker, Marcus Ennis, Bhavana Harsha, Namrata Kale, Venkatesh Muthukrishnan, Gareth Owen, Steve Turner, Mark Williams, and et al. The chebi reference database and ontology for biologically relevant chemistry: Enhancements for 2013. *Nucleic Acids Research*, 41(D1), 2013.

(Haug et al., 2013) Kenneth Haug, Reza M. Salek, Pablo Conesa, Janna Hastings, Paula De Matos, Mark Rijnbeek, Tejasvi Mahendraker, Mark Williams, Steffen Neumann, Philippe Rocca-Serra, Eamonn Maguire, Alejandra González-Beltrán, Susanna Assunta Sansone, Julian L.

Griffin, and Christoph Steinbeck. MetaboLights - An open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Research*, 41(D1), jan 2013.

(Imamura et al., 2016) M Imamura, K Higashi, K Yamaguchi, K Asakura, T Furihata, Y Terui, T Satake, J Maegawa, K Yasumura, A Ibuki, T Akase, K Nishimura, K Kashiwagi, R Linhardt, K Igarashi, and T Toida. Polyamines release the let-7b-mediated suppression of initiation codon recognition during the protein synthesis of ext2. *Scientific Reports*, 6, 2016. doi: 10.1038/srep33549.

(Jacob et al., 2013) Daniel Jacob, Catherine Deborde, and Annick Moing. An efficient spectra processing method for metabolite identification from 1h-nmr metabolomics data. *Analytical and Bioanalytical Chemistry*, 405(15):5049–5061, 2013. doi: 10.1007/s00216-013-6852-y.

(Jiye et al., 2005) A Jiye, Johan Trygg, Jonas Gullberg, Annika I. Johansson, Pär Jonsson, Henrik Antti, Stefan L. Marklund, and Thomas Moritz. Extraction and GC/MS Analysis of the Human Blood Plasma Metabolome. *Analytical Chemistry*, 77(24):8086–8094, 2005. ISSN 0003-2700. doi: 10.1021/ac051211v. URL http://pubs.acs.org/doi/abs/10.1021/ac051211v.

(Kamburov et al., 2011) Atanas Kamburov, Rachel Cavill, Timothy M D Ebbels, Ralf Herwig, and Hector C. Keun. Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. *Bioinformatics*, 27(20):2917–2918, oct 2011.

(Kanazawa and Sakakibara, 2000) Kazuki Kanazawa and Hiroyuki Sakakibara. High content of dopamine, a strong antioxidant, in cavendish banana. *Journal of Agricultural and Food Chemistry*, 48(3):844–848, 2000. ISSN 00218561. doi: 10.1021/jf9909860.

(Kankainen et al., 2011) Matti Kankainen, Peddinti Gopalacharyulu, Liisa Holm, and Matej Orešič. MPEA-metabolite pathway enrichment analysis. *Bioinformatics*, 27(13):1878–1879, jul 2011.

(Kim et al., 2015) Sunghwan Kim, Paul A. Thiessen, Evan E. Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A. Shoemaker, Jiyao Wang, Bo Yu, Jian Zhang, and Stephen H. Bryant. Pubchem substance and compound databases. *Nucleic Acids Research*, 44(D1):D1202, 2015. doi: 10.1093/nar/gkv951. URL http://dx.doi.org/10.1093/nar/gkv951.

(Kimura, 1968) Mieko Kimura. Fluorescence histochemical study on serotonin and catecholamine in some plants. *The Japanese Journal of Pharmacology*, 18(2):162–168, 1968.

(Knapp and Nicholas, 1969) FF Knapp and HJ Nicholas. The sterols and triterpenes of banana peel. *Phytochemistry*, 8(1):207–214, 1969.

(Kosmides et al., 2013) Alyssa K Kosmides, Kubra Kamisoglu, Steve E Calvano, Siobhan A Corbett, and Ioannis P Androulakis. Metabolomic fingerprinting: challenges and opportunities. *Critical reviews in biomedical engineering*, 41(3):205–21, jan 2013.

(Krinsky and Johnson, 2005) Norman I. Krinsky and Elizabeth J. Johnson. Carotenoid actions and their relation to health and disease, 2005. ISSN 00982997.

(Kris-etherton et al., 2002) P.M. Kris-etherton, K.D. Hecker, A. Bonanome, S. M. Coval, A.E. Binkoski, and K.F. Hilpert. Bioactive Compounds in Foods: Their Role in the Prevention of Cardiovascular Disease and Cancer. *American Journal of Chemistry*, 13(9):71–88, 2002.

(Kristensen et al., 2012) Mette Kristensen, Søren B. Engelsen, and Lars O. Dragsted. LC-MS metabolomics top-down approach reveals new exposure and effect biomarkers of apple and apple-pectin intake. *Metabolomics*, 8(1):64–73, feb 2012.

(Kuhl et al., 2012) Carsten Kuhl, Ralf Tautenhahn, Christoph Böttcher, Tony R. Larson, and Steffen Neumann. CAMERA: An integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Analytical Chemistry*, 84(1):283–289, 2012. ISSN 00032700. doi: 10.1021/ac202450g.

(Kumar et al., 2012) K. P. Sampath Kumar, Debjit Bhowmik, S. Duraivel, and M. Umadevi. Traditional and medicinal uses of banana. *Journal of Pharmacognosy and Phytochemistry*, 2012. ISSN 2278-4136.

(Kvalheim et al., 1994) Olav M Kvalheim, Frode Brakstad, and Yizeng. Liang. Preprocessing of analytical profiles in the presence of homoscedastic or heteroscedastic noise. *Analytical Chemistry*, 66:43–51, 1994. URL http://pubs.acs.org/doi/abs/10.1021/ac00073a010.

(Lakshmi et al., 1988) AV Lakshmi, PV Diwan, and MS Bamji. Wound healing in pyridoxine deficiency. *Nutrition Research*, 8:1203–1206, 1988.

(Liepa et al., 2007) GU Liepa, C Ireton-Jones, H Basu, and CR Baxter. B vitamins and wound healing. In JA Molnar, editor, *Nutrition and Wound Healing*. CRC Press, Boca Raton, 2007.

(Liland, 2011) Kristian Hovde Liland. Multivariate methods in metabolomics from preprocessing to dimension reduction and statistical analysis. *TrAC Trends in Analytical Chemistry*, 30(6):827–841, jun 2011. ISSN 01659936. doi: 10.1016/j.trac.2011.02.007. URL http://linkinghub.elsevier.com/retrieve/pii/S0165993611000914.

(López-Ibáñez et al., 2016) J López-Ibáñez, F Pazos, and M Chagoyen. MBROLE 2.0-functional enrichment of chemical compounds. *Nucleic Acids Res*, 44(W1):W201–W204, 2016. ISSN 0305-1048. doi: 10.1093/nar/gkw253. URL http://www.ncbi.nlm.nih.gov/pubmed/27084944.

(Maraschin et al., 2016) Marcelo Maraschin, Amélia Somensi-Zeggio, Simone K. Oliveira, Shirley Kuhnen, Maíra M. Tomazzoli, Josiane C. Raguzzoni, Ana C.M. Zeri, Rafael Carreira, Sara Correia, Christopher Costa, and Miguel Rocha. Metabolic Profiling and Classification of Propolis Samples from Southern Brazil: An NMR-Based Platform Coupled with Machine Learning. *Journal of Natural Products*, 2016. ISSN 15206025. doi: 10.1021/acs.jnatprod.5b00315.

(Marcucci, 1995) M.C. Marcucci. Propolis: chemical composition, biological properties and therapeutic activity. *Apidologie*, 1995. ISSN 0044-8435. doi: 10.1051/apido.

(Martinez et al., 2009) I. Martinez, I.B. Standal, D.E. Axelson, B. Finstad, and M. Aursand. Identification of the farm origin of salmon by fatty acid and HR 13C NMR profiling. *Food Chemistry*, 116(3):766–773, oct 2009. ISSN 03088146. doi: 10.1016/j.foodchem.2009.03.026. URL http://linkinghub.elsevier.com/retrieve/pii/S0308814609003100.

(Masoum et al., 2007) Saeed Masoum, Christophe Malabat, Mehdi Jalali-Heravi, Claude Guillou, Serge Rezzi, and Douglas Neil Rutledge. Application of support vector machines to 1H NMR data of fish oils: methodology for the confirmation of wild and farmed salmon and their origins. *Analytical and bioanalytical chemistry*, 387(4):1499–510, feb 2007. ISSN 1618-2642. doi: 10.1007/s00216-006-1025-x. URL http://www.ncbi.nlm.nih.gov/pubmed/17200859.

(McKelvie et al., 2009) Jennifer R. McKelvie, Jimmy Yuk, Yunping Xu, Andre J. Simpson, and Myrna J. Simpson. 1H NMR and GC/MS metabolomics of earthworm responses to sublethal DDT and endosulfan exposure. *Metabolomics*, 5(1):84–94, 2009.

(Mounet et al., 2007) Fabien Mounet, Martine Lemaire-Chamley, Mickaël Maucourt, Cécile Cabasson, Jean-Luc Giraudel, Catherine Deborde, René Lessire, Philippe Gallusci, Anne Bertrand, Monique Gaudillère, Christophe Rothan, Dominique Rolin, and Annick Moing. Quantitative metabolic profiles of tomato flesh and seeds during fruit development: complementary analysis with ANN and PCA. *Metabolomics*, 3(3):273–288, may 2007. ISSN 1573-3882. doi: 10.1007/s11306-007-0059-1. URL http://link.springer.com/10.1007/s11306-007-0059-1.

(Mujica et al., 2017) Verónica Mujica, Roxana Orrego, Jorge Pérez, Paula Romero, Paz Ovalle, Jessica Zúñiga-Hernández, Miguel Arredondo, and Elba Leiva. The Role of Propolis in Oxidative Stress and Lipid Metabolism: A Randomized Controlled Trial. *Evidence-based Complementary and Alternative Medicine*, 2017. ISSN 17414288. doi: 10.1155/2017/4272940.

(Nguyen et al., 2003) Thi Bich Thuy Nguyen, Saichol Ketsa, and Wouter G. Van Doorn. Relationship between browning and the activities of polyphenol oxidase and phenylalanine

ammonia lyase in banana peel during low temperature storage. *Postharvest Biology and Technology*, 30(2):187–193, 2003. ISSN 09255214. doi: 10.1016/S0925-5214(03)00103-0.

(Ogata et al., 1999) Hiroyuki Ogata, Susumu Goto, Kazushige Sato, Wataru Fujibuchi, Hidemasa Bono, and Minoru Kanehisa. KEGG: Kyoto encyclopedia of genes and genomes, jan 1999.

(Padam et al., 2014) Birdie Scott Padam, Hoe Seng Tin, Fook Yee Chye, and Mohd Ismail Abdullah. Banana by-products: an under-utilized renewable food biomass with great potential, 2014. ISSN 09758402.

(Papotti et al., 2010) Giulia Papotti, Davide Bertelli, Maria Plessi, and Maria Cecilia Rossi. Use of HR-NMR to classify propolis obtained using different harvesting methods. *International Journal of Food Science & Technology*, 45(8):1610–1618, jul 2010. ISSN 09505423. doi: 10.1111/j.1365-2621.2010.02310.x. URL http://doi.wiley.com/10.1111/j.1365-2621.2010.02310.x.

(Pereira, 2014) Aline Pereira. *Determinação do perl químico e da atividade cicatrizante de extratos de casca de banana cultivar prata anã (Musa sp.) e o desenvolvimento de um curativo para pequenas lesões.* PhD thesis, Universidade Federal de Santa Catarina, Florianopolis, 2014.

(Pereira and Maraschin, 2015) Aline Pereira and Marcelo Maraschin. Banana (Musa spp) from peel to pulp: Ethnopharmacology, source of bioactive compounds and its relevance for human health. *Journal of Ethnopharmacology*, 160:149–163, 2015. ISSN 18727573. doi: 10.1016/j.jep.2014.11.008.

(Psihogios et al., 2007) Nikolaos G Psihogios, Rigas G Kalaitzidis, Sofia Dimou, Konstantin I Seferiadis, Kostas C Siamopoulos, and Eleni T Bairaktari. Evaluation of tubulointerstitial lesions' severity in patients with glomerulonephritides: an NMR-based metabonomic study. *Journal of proteome research*, 6(9):3760–70, sep 2007. ISSN 1535-3893. doi: 10.1021/pr070172w. URL http://www.ncbi.nlm.nih.gov/pubmed/17705523.

(Ravanbakhsh et al., 2015) Siamak Ravanbakhsh, Philip Liu, Trent C. Bjordahl, Rupasri Mandal, Jason R. Grant, Michael Wilson, Roman Eisner, Igor Sinelnikov, Xiaoyu Hu, Claudio Luchinat, and et al. Accurate, fully-automated nmr spectral profiling for metabolomics. *PLoS ONE*, 10(5), 2015.

(Rocha et al., 2008) Miguel Rocha, Paulo Cortez, and José Maia Neves. *Análise inteligente de dados: algoritmos e implementação em Java*. FCA, 2008.

(Rodriguez-Amaya, 2001) Delia B Rodriguez-Amaya. *A Guide to Carotenoid Analysis in Food.* ILST Press, Washington, 21 edition, 2001.

(Sabatine et al., 2005) Marc S. Sabatine, Emerson Liu, David A. Morrow, Eric Heller, Robert McCarroll, Roger Wiegand, Gabriel F. Berriz, Frederick P. Roth, and Robert E. Gerszten.

Metabolomic identification of novel biomarkers of myocardial ischemia. *Circulation*, 112 (25):3868–3875, dec 2005.

(Saghatelian et al., 2004) Alan Saghatelian, Sunia A. Trauger, Elizabeth J. Want, Edward G. Hawkins, Gary Siuzdak, and Benjamin F. Cravatt. Assignment of endogenous substrates to enzymes by global metabolite profiling. *Biochemistry*, 2004. ISSN 00062960. doi: 10.1021/bi0480335.

(Samudrala et al., 2015) Devasena Samudrala, Brigitte Geurts, Phil A. Brown, Ewa Szymańska, Julien Mandon, Jeroen Jansen, Lutgarde Buydens, Frans J M Harren, and Simona M. Cristescu. Changes in urine headspace composition as an effect of strenuous walking. *Metabolomics*, 11(6):1656–1666, may 2015.

(Seymor, 1993) G Seymor. Banana. In G Seymor, J Taylor, and G Ticker, editors, *Biochemistry of fruit ripening.*, pages 95–98. Chapman and Hall, London, 1993.

(Someya et al., 2002) Shinichi Someya, Yumiko Yoshiki, and Kazuyoshi Okubo. Antioxidant compounds from bananas (Musa Cavendish). *Food Chemistry*, 79(3):351–354, 2002. ISSN 03088146. doi: 10.1016/S0308-8146(02)00186-3.

(Son and Lewis, 2002) S Son and B a Lewis. Free radical scavenging and antioxidative activity of caffeic acid amide and ester analogues: structure-activity relationship. *J Agric Food Chem*, 2002. ISSN 0021-8561. doi: 10.1021/jf010830b.

(Subagio et al., 1996) Achmad Subagio, Naofumi Morita, and Shigeo Sawada. Carotenoids and Their Fatty-Acid Esters in Banana Peel. *Journal of Nutricional Science and Vitaminology*, 42 (6):553–566, 1996. ISSN 0301-4800. doi: 10.3177/jnsv.42.553.

(Sud et al., 2016) Manish Sud, Eoin Fahy, Dawn Cotter, Kenan Azam, Ilango Vadivelu, Charles Burant, Arthur Edison, Oliver Fiehn, Richard Higashi, K. Sreekumaran Nair, Susan Sumner, and Shankar Subramaniam. Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Research*, 2016. ISSN 13624962. doi: 10.1093/nar/gkv1042.

(Suhre and Schmitt-Kopplin, 2008) Karsten Suhre and Philippe Schmitt-Kopplin. MassTRIX: mass translator into pathways. *Nucleic acids research*, 36(Web Server issue):W481–4, jul 2008. ISSN 1362-4962. doi: 10.1093/nar/gkn194.

(Szajdek and Borowska, 2008) Agnieszka Szajdek and E. J. Borowska. Bioactive compounds and health-promoting properties of Berry fruits: A review, 2008. ISSN 09219668.

(Tapiero et al., 2004) H. Tapiero, D. M. Townsend, and K. D. Tew. The role of carotenoids in the prevention of human pathologies, 2004. ISSN 07533322.

(Tautenhahn et al., 2012) Ralf Tautenhahn, Gary J. Patti, Duane Rinehart, and Gary Siuzdak. XCMS online: A web-based platform to process untargeted metabolomic data. *Analytical Chemistry*, 84(11):5035–5039, jun 2012.

(Tulpan et al., 2011) Dan Tulpan, Serge Lger, Luc Belliveau, Adrian Culf, and Miroslava Cuperlovi-Culf. Metabohunter: an automatic approach for identification of metabolites from 1h-nmr spectra of complex mixtures. *BMC bioinformatics*, 12(1):400, 2011. URL http://www.biomedcentral.com/1471-2105/12/400.

(van den Berg et al., 2006) Robert a van den Berg, Huub C J Hoefsloot, Johan a Westerhuis, Age K Smilde, and Mariët J van der Werf. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC genomics*, 7:142, 2006. URL http://www.ncbi.nlm.nih.gov/pubmed/16762068.

(Varmuza and Filzmoser, 2010) K. Varmuza and P. Filzmoser. *Introduction to multivariate statistical analysis in chemometrics*, volume 64. 2010. ISBN 9781420059472. doi: 10.1366/000370210791114185.

(Villas-Bôas et al., 2006) Silas G. Villas-Bôas, Ute Roessner, Michael A E Hansen, Jørn Smedsgaard, and Jens Nielsen. *Metabolome Analysis: An Introduction*. John Wiley and Sons, jun 2006.

(Voutilainen et al., 2006) Sari Voutilainen, Tarja Nurmi, Jaakko Mursu, and Tiina H. Rissanen. Carotenoids and cardiovascular health, 2006. ISSN 00029165.

(Wishart et al., 2007) David S. Wishart, Dan Tzur, Craig Knox, Roman Eisner, An Chi Guo, Nelson Young, Dean Cheng, Kevin Jewell, David Arndt, Summit Sawhney, Chris Fung, Lisa Nikolai, Mike Lewis, Marie Aude Coutouly, Ian Forsythe, Peter Tang, Savita Shrivastava, Kevin Jeroncic, Paul Stothard, Godwin Amegbey, David Block, David D. Hau, James Wagner, Jessica Miniaci, Melisa Clements, Mulu Gebremedhin, Natalie Guo, Ying Zhang, Gavin E. Duggan, Glen D. MacInnis, Alim M. Weljie, Reza Dowlatabadi, Fiona Bamforth, Derrick Clive, Russ Greiner, Liang Li, Tom Marrie, Brian D. Sykes, Hans J. Vogel, and Lori Querengesser. HMDB: The human metabolome database. *Nucleic Acids Research*, 35 (SUPPL. 1), jan 2007.

(Wishart et al., 2009) David S. Wishart, Craig Knox, An Chi Guo, Roman Eisner, Nelson Young, Bijaya Gautam, David D. Hau, Nick Psychogios, Edison Dong, Souhaila Bouatra, Rupasri Mandal, Igor Sinelnikov, Jianguo Xia, Leslie Jia, Joseph A. Cruz, Emilia Lim, Constance A. Sobsey, Savita Shrivastava, Paul Huang, Philip Liu, Lydia Fang, Jun Peng, Ryan Fradette, Dean Cheng, Dan Tzur, Melisa Clements, Avalyn Lewis, Andrea De souza, Azaret Zuniga, Margot Dawe, Yeping Xiong, Derrick Clive, Russ Greiner, Alsu Nazyrova, Rustem

Shaykhutdinov, Liang Li, Hans J. Vogel, and Ian Forsythei. HMDB: A knowledgebase for the human metabolome. *Nucleic Acids Research*, 37(SUPPL. 1), 2009.

(Wishart et al., 2013) David S. Wishart, Timothy Jewison, An Chi Guo, Michael Wilson, Craig Knox, Yifeng Liu, Yannick Djoumbou, Rupasri Mandal, Farid Aziat, Edison Dong, Souhaila Bouatra, Igor Sinelnikov, David Arndt, Jianguo Xia, Philip Liu, Faizath Yallou, Trent Bjorndahl, Rolando Perez-Pineiro, Roman Eisner, Felicity Allen, Vanessa Neveu, Russ Greiner, and Augustin Scalbert. HMDB 3.0-The Human Metabolome Database in 2013. *Nucleic Acids Research*, 41(D1), jan 2013.

(Wollenweber et al., 1990) E Wollenweber, BM Hausen, and Greenaway W. Phenolic constituents and sensitizing properties of propolis, poplar balsam and balsam of peru. *Bulletin de Groupe Polyphenol*, 15:112–120, 1990.

(Xia and Wishart, 2010) Jianguo Xia and David S. Wishart. MSEA: A web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Research*, 38(SUPPL. 2), may 2010.

(Xia et al., 2009) Jianguo Xia, Nick Psychogios, Nelson Young, and David S. Wishart. MetaboAnalyst: A web server for metabolomic data analysis and interpretation. *Nucleic Acids Research*, 37(SUPPL. 2), 2009.

(Xia et al., 2012) Jianguo Xia, Rupasri Mandal, Igor V. Sinelnikov, David Broadhurst, and David S. Wishart. MetaboAnalyst 2.0-a comprehensive server for metabolomic data analysis. *Nucleic Acids Research*, 40(W1), jul 2012.

(Xia et al., 2013) Jianguo Xia, David I Broadhurst, Michael Wilson, and David S Wishart. Translational biomarker discovery in clinical metabolomics: an introductory tutorial. *Metabolomics : Official journal of the Metabolomic Society*, 9(2):280–299, apr 2013. ISSN 1573-3882. doi: 10.1007/s11306-012-0482-9.

(Xia et al., 2015) Jianguo Xia, Igor V. Sinelnikov, Beomsoo Han, and David S. Wishart. MetaboAnalyst 3.0-making metabolomics more meaningful. *Nucleic Acids Research*, 43(W1):W251–W257, 2015.