An R package for the integrated analysis of metabolomics and spectral data

Christopher Costa^a, Marcelo Maraschin^b, Miguel Rocha^{a,*}

^a CEB – Centre Biological Engineering, University of Minho, Braga, Portugal

^b Plant Morphogenesis and Biochemistry Laboratory, Federal University of Santa Catarina, Florianopolis, Brazil

ARTICLE INFO

Article history:

Received 24 July 2015

Received in revised form

3 January 2016

Accepted 5 January 2016

Keywords:

Metabolomics

Chemometrics

R

Nuclear magnetic resonance

Mass spectrometry

Infrared and UV–visible spectroscopy

ABSTRACT

Recently, there has been a growing interest in the field of metabolomics, materialized by a remarkable growth in experimental techniques, available data and related biological applications. Indeed, techniques as nuclear magnetic resonance, gas or liquid chromatography, mass spectrometry, infrared and UV–visible spectroscopies have provided extensive datasets that can help in tasks as biological and biomedical discovery, biotechnology and drug development. However, as it happens with other omics data, the analysis of metabolomics datasets provides multiple challenges, both in terms of methodologies and in the development of appropriate computational tools. Indeed, from the available software tools, none addresses the multiplicity of existing techniques and data analysis tasks.

In this work, we make available a novel R package, named *specmine*, which provides a set of methods for metabolomics data analysis, including data loading in different formats, pre-processing, metabolite identification, univariate and multivariate data analysis, machine learning, and feature selection. Importantly, the implemented methods provide adequate support for the analysis of data from diverse experimental techniques, integrating a large set of functions from several R packages in a powerful, yet simple to use environment.

The package, already available in CRAN, is accompanied by a web site where users can deposit datasets, scripts and analysis reports to be shared with the community, promoting the efficient sharing of metabolomics data analysis pipelines.

© 2016 Elsevier Ireland Ltd. All rights reserved.

1. Introduction

The field of metabolomics has been growing in interest by the scientific community, given its extensive range of applications in functional genomics and, more globally, in the characterization of biological systems [12]. Indeed, the possible tasks include studying metabolic systems, measuring biochemical phenotypes, understanding and reconstructing networks, discriminating between samples, identifying biomarkers of

disease, analyzing food and beverages, studying plant physiology or fostering drug discovery [12,17,11].

Metabolomics can be defined as the identification and quantification of all intracellular and extracellular metabolites with low molecular mass. Unlike transcriptomics and proteomics which are based on polymers, metabolites have a large variance in chemical structures and properties, making difficult the development of high-throughput techniques and reducing the number of molecules that can adequately be measured in a sample [17]. In recent years, however, an exten-

* Corresponding author. Tel.: +351 253604456.

E-mail address: mrocha@di.uminho.pt (M. Rocha).

<http://dx.doi.org/10.1016/j.cmpb.2016.01.008>

0169-2607/© 2016 Elsevier Ireland Ltd. All rights reserved.

sive range of experimental techniques has been developed to obtain metabolomics data, which are growing in size.

The main experimental approaches for metabolomics data acquisition are nuclear magnetic resonance (NMR), liquid chromatography (LC) and gas chromatography (GC) coupled with mass spectrometry (MS), infrared (IR), and ultraviolet–visible (UV–vis) spectroscopies. GC/LC–MS and NMR are more robust techniques and they are frequently employed in the analysis and quantification of the metabolome. IR and UV–vis techniques have the advantage of being simpler and less expensive, providing a complementary view of that provided by GC/LC–MS or NMR.

Two distinct approaches can be chosen for running a metabolomics experiment. The former, a chemometrics approach or metabolic fingerprinting, makes direct use of spectra or peaks lists, and the analysis typically addresses sample discrimination. The latter, known as metabolic target analysis or profiling, focuses on the identification and quantification of compounds in the sample, using that information to run the analysis. The general workflow of an experiment generally consists in the steps of sample preparation, data acquisition, preprocessing, data analysis and interpretation (see Fig. 1). Once the samples are prepared and data is acquired, it will be preprocessed to correct some issues and improve the performance of the next step, data analysis.

The analysis of metabolomics datasets needs to address important challenges regarding the extraction of relevant knowledge from large amounts of data [16]. As it is the case with other types of omics data, metabolomics data analysis has promoted the development and adaptation of numerous methods for data analysis from fields such as statistics, machine learning or computer science.

Some of the main available tools for metabolomics data analysis are summarized in Table 1. Most of the tools reviewed in the table, while providing useful functionalities, are limited to specific types of metabolomics or spectral data (as it is the case with hyperSpec, ChemoSpec, speaq or Automics). Other tools offer a limited portfolio of data analysis tools (e.g. metabolomics package or MeltDB), not offering an extensive range of solutions for the construction of metabolomics data analysis pipelines. The most comprehensive tool found was the web-based suite MetaboAnalyst [18], which provides

Fig. 1 – General workflow of a metabolomics experiment.

a large number of tools with a user-friendly interface. It is, however, limited to the provided closed user interface and available configuration parameters, being difficult to extend and define personalized workflows. Also, it does not cover spectral data.

On the other hand, a number of R packages have been developed, from which we emphasize *xcms* and *MAIT* [20] (from Bioconductor), *hyperSpec* [2] and *ChemoSpec* [9], which, although limited to specific data types, provide valuable functions for many types of metabolomics data analysis. However, these packages are not well integrated with each other, and therefore it is not straightforward to combine their functionalities.

Taking into account the aforementioned context, the proposed work aims to develop an integrated script-based software for the analysis of metabolomics data in the format

Table 1 – Available free tools for metabolomics data.

Name	URL	Short description
MetaboAnalyst	http://www.metaboanalyst.ca	Web application to analyze metabolomic data
<i>hyperSpec</i>	http://hyperspec.r-forge.r-project.org	R package to handle spectral data and metadata
<i>ChemoSpec</i>	http://cran.r-project.org/web/packages/ChemoSpec	R package to handle spectral data
Metabolomic package	http://cran.open-source-solution.org/web/packages/Metabonomic/	R-package + GUI for metabolomic profiles
<i>speaq</i>	https://code.google.com/p/speaq/	Integrated workflow for robust alignment and quantitative analysis of NMR
<i>Automics</i>	https://code.google.com/p/automics/	NMR-based spectral processing and analysis
MeltDB	https://meltdb.cebitec.uni-bielefeld.de	Web-based system for data analysis
Metabolomics	http://cran.r-project.org/web/packages/metabolomics	R functions for metabolomics statistical analysis
metaP-Server	http://metabolomics.helmholtz-muenchen.de/metap2/	Web application to analyze metabolomic data
Bioconductor	http://bioconductor.org/packages/release/BiocViews.html#___Metabolomics	Bioconductor R packages for metabolomics (include <i>xcms</i>)

Fig. 2 – Modules developed in the package.

of a package for the R scientific computing environment. This package will incorporate many available functions provided by metabolomics oriented R packages (as highlighted above), but also more general-purpose data analysis R functions. The developed tool will address at this stage metabolomics and spectral data from GC–MS, LC–MS, NMR, IR, and UV–vis experiments.

The framework was developed to address a wide variety of common tasks on metabolomics data analysis, providing a general workflow that can be adapted for specific case studies, addressing tasks as data loading, pre-processing, normalization, metabolite identification, univariate and multivariate statistical analysis, clustering, machine learning and feature selection. Thus, it provides a tool that not only covers the main metabolomics and spectral data types, but also addresses a full range of tasks in data analysis, allowing to create flexible and powerful analysis pipelines for specific case studies, not losing the capability of interacting with the multitude of functions provided by the R project's community.

2. Development of the package and provided functions

To achieve the objectives of this work, a package with features covering the main steps of the metabolomics data analysis workflow was developed, containing functions for data reading and dataset creation, preprocessing, metabolite identification and data analysis. Fig. 2 shows the main modules developed.

The package contains functions that are easy to use, i.e. with few mandatory parameters, but also very flexible, since most functions have default parameters, but they also have a large number of parameters that can be used to change the default behavior. The package integrates many functions imported and sometimes adapted/extended from other packages, integrating various packages over a unique interface. The package's functions were meant to provide abundant graphical visualization options of the results. The idea is to minimize the complexity of the code needed to make an analysis pipeline over a dataset, but also to easily allow creating variants for this analysis.

2.1. Development tools and installation

The package was developed using the R¹ environment, a free software environment for data manipulation, scientific and statistical computing and graphical visualization. RStudio² was the environment chosen to develop the scripts and assemble the package. Reports were made using a plugin named RMarkdown,³ which can create easily dynamic HTML reports from annotated R code.

The package can be accessed and installed through the CRAN repository, being available under the name of *specmine* in the URL <https://cran.r-project.org/web/packages/specmine/>, being also available a set of documents providing full documentation for all functions. The package's most recent development version can also be installed directly in RStudio using the *install.bitbucket* function of the *devtools* package, using the following repository as an argument: "chris-bcl/metabolomicsPackage".

2.2. Data reading and dataset structure

The types of metabolomics data supported are NMR, UV–vis, IR, and (GC- or LC-)MS data. A number of different file formats are supported, including comma (or tab) separated values (CSV or TSV) files, (J)DX spectra files, NetCDF, mzDATA and mzXML MS data. Metadata can be given as a CSV/TSV file. The core reading functions for (J)DX are provided by *ChemoSpec* [9]. For LC/GC–MS spectra data (NetCDF, mzDATA, and mzML), the package *xcms* [19] was used. It is also possible to load data as a peaks list, using peak alignment functions to reach a dataset in a standard tabular form. Fig. 3 represents a scheme of the data reading processes.

The basic structure of a dataset is general-purpose, independent of the type of data and source. A dataset is an R list consisting of the following fields: description of the dataset, type of data, the data matrix, the metadata data frame and the labels for the x- and y-axis. In Fig. 4, a graphical representation of the dataset structure is provided.

¹ <http://www.r-project.org>.

² <http://www.rstudio.com>.

³ <http://rmarkdown.rstudio.com>.

Fig. 3 – Scheme of the data reading.

2.3. Exploratory analysis and data pre-processing

The package includes a number of functions that allow to calculate global statistics and others to provide graphical visualization of the data. A basic visualization function allows to see the distribution of values for (a subset of) the variables in the dataset in the form of boxplots. There are also functions to plot spectral data (where variables are represented by numerical values). For visualization functions, the base graphics system of R was used, but some functions also rely on the *ggplot2* package.

Various preprocessing methods were implemented for the different types of data. To extract relevant parts of a dataset, a number of functions were developed that allow to extract (or remove) relevant subsets of samples, data and metadata variables.

To perform spectral pre-processing, methods are available for different types of corrections or spectral processing. These include functions for shifting correction, multiplicative scatter correction, first derivative, baseline (based on the methods available for the baseline R package), offset and background corrections (both implemented from the *hyperSpec* package).

Fig. 4 – Representation of the structure of the data in a dataset.

A few methods for smoothing interpolation are available, such as bin or loess smoothing, as well as Savitzky–Golay filters.

To treat missing values, various imputation methods were implemented. Also, methods for data normalization, transformation and scaling are provided. Sample normalization methods includes normalization by sum, median, by a reference sample or feature. Cubic root and logarithmic transformation methods are available, and as for scaling methods, auto, range and pareto scaling methods are provided. Finally, to remove variables with low variance, flat pattern filters are provided with distinct metrics and parameters.

2.4. Metabolite identification

A method from the MAIT (<http://bioconductor.org/packages/release/bioc/html/MAIT.html>) package is currently available to perform metabolite identification, at this stage only for LC–MS data.

The method provides peak annotation with the CAMERA package. It can also be used to obtain the significant features from the spectra, or the user can supply the features and proceed with metabolite identification. The MAIT package provides a method called biotransformation that helps to improve the metabolite identification process and also provides its own database [20]. The results are stored on a folder with the query mass and the identification information associated with them.

2.5. Univariate analysis

A number of distinct univariate analysis methods, all based in the stats R package, were implemented covering the following types of analysis: correlations, fold changes, t-tests, analysis of variance (ANOVA) and regression analysis. Indeed, there are functions to calculate the correlations between variables or samples, and the resulting matrix can be visualized as a heatmap.

To perform fold change analysis, there are functions to calculate fold changes of values considering two groups of samples (as defined by a metadata variable), being possible

to visualize the results in tabular and graphical forms. Also, *t*-tests can be performed and visualized in a similar way. The package also implements one-way ANOVA, with the Tukey HSD post hoc test, and multifactorial ANOVA. There are functions to summarize the main results of complex analyses providing *p*-values and percentage of variation explained by the different factors. Both *t*-tests and ANOVA implementations are prepared to handle multiple testing by default, being the provided *p*-values adjusted using the false discovery rate (FDR) method proposed by Benjamini and Hochberg [3].

Non-parametric tests were also considered, through the implementation of the Kruskal–Wallis and Kolmogorov–Smirnov tests. On the other hand, linear regression analysis can be performed, being available functions that summarize the coefficients for the different factors and interactions, as well as the respective *p*-values.

2.6. Unsupervised multivariate analysis

The package includes functions to perform principal component analysis (PCA) using two methods: classical and robust. Robust PCA uses the grid search algorithm to compute the desired number of principal components, using the *pcaPP* R package [5]. It also provides a number of ways to visualize the results through scree plots, scores plots, biplots and pairs plots.

Additionally, two clustering methods were implemented: *k*-means clustering and hierarchical clustering. The distance method can be chosen according to the available option from the *dist* function, as well the method used in the case of the hierarchical clustering. Clustering results can also be plotted. There are functions that create a dendrogram of the hierarchical clustering results and functions that allow to visualize the results from *k*-means.

2.7. Machine learning and feature selection

The package provides a number of functions to train, use and evaluate machine learning methods, being mostly based in the R package *caret*, covering both classification and regression methods. Also, there are functions to evaluate the importance of each variable in the models. A list of possible models and tunable parameters can be seen in: <http://topepo.github.io/caret/modelList.html>.

Validation methods are available to estimate the error metrics, which include *k*-fold cross-validation, leave-one-out cross-validation and resampling, among others. The error metrics available include accuracy, area under the ROC curve (AUC) and kappa statistic for classification, and root mean square error (RMSE) and the coefficient of determination (R^2) for regression.

The implemented functions enable to do model training and the posterior use of the model to predict new data. Also, it is possible to optimize a set of model parameters (e.g. number of components in partial least squares, or hidden nodes in a neural network) testing a set of possible values and evaluating those according to the selected validation method and error metrics. The results from a process of model optimization includes the best model obtained, the performance of this model, the variables' importance, the results of all tested

combinations of parameters, the confusion matrices (for classification), among other statistics. These models can be used later for predictions or visualization, as it happens in the partial least squares (PLS) case.

Also, from the same package, there are a number of functions to perform feature selection, i.e. determine which attributes are more valuable when applying different machine learning methods. Both filter and wrapper methods are provided for feature selection, which can be combined with selected models from the list mentioned above. Regarding the wrappers available, the most commonly used is recursive feature elimination (RFE) [22]. This will test different subsets of features, iteratively reducing the number of features and verifying which configuration provides the best performance.

2.8. Web site

Complementary to the R package developed, the authors have also created a web site that aims to provide a means for sharing datasets and data analyses among the community. This web site, which is still provided in a preliminary version, allows users to deposit datasets, scripts and analysis reports for metabolomics data. The information made public by the users can be freely navigated by anyone, making available the data and metadata for a specific task, the R scripts used in the analysis and the data analysis reports that were generated using R markdown.

The current version of the site is provided in the URL <http://darwin.di.uminho.pt/metabolomics>, being at this stage only available the datasets referring to the case studies described in the following section of this paper.

3. Case studies

Four case studies using real data will be presented to test the package and provide meaningful data analysis pipelines. The first is the analysis of metabolites concentrations of urine samples from control and cachexic cancer patients, the next is the discrimination of propolis samples from southern Brazil (using NMR and UV–vis data), the third is the analysis of the effect of postharvest physiological deterioration (PPD) of cassava samples (IR data), and, finally, the last illustrates the use of LC–MS data with spectra of two groups of samples from spinal cord of wild-type and mutant mice.

For one of the case studies, some selected results will be shown as a demonstration of some of the package's capabilities. Full data analysis reports for all cases, following distinct analysis pipelines are further provided in the web site mentioned above.

3.1. Cachexia

Cachexia is a complex metabolic syndrome associated with an underlying illness (such as cancer) and characterized by loss of muscle with or without loss of fat mass [7]. Improved approaches for detecting the onset and evolution of muscle wasting would help to manage wasting syndromes and facilitate early intervention [6]. As metabolites produced from tissue breakdown are likely to be a sensitive indicator of

muscle wasting, urine samples were collected since several end products of muscle catabolism are specifically excreted in urine [6]. A total of 77 urine samples were collected being 47 of them patients with cachexia, and 30 control patients. All one-dimensional NMR spectra of urine samples were acquired and then the metabolites were detected and quantified, i.e. for each metabolite its concentration was measured.

3.2. Propolis

Propolis is a substance produced from the collected exudates of plants (resin) by bees. The resin is masticated, salivary enzymes are added, and the partially digested material is mixed with beeswax and used in the hive to seal the walls, strengthen the borders of combs, and embalm dead invaders. Recently, this product has been the subject of studies highlighting its pharmacological properties, such as the antimicrobial [4,1], anti-oxidative [10], anti-viral [8], anti-tumoral [13,14] or anti-inflammatory [4,1].

It has long been known that propolis' chemical composition might be strongly influenced by environmental factors peculiar to the sites of collection of a given geographic region of production, as well as by seasoning. The aim of this case study is to gain insights of important features associated to chemical composition, harvest season, and geographic origin of propolis produced in the Santa Catarina state, southern Brazil. The propolis samples used in this study for NMR data analysis were collected in the autumn (AU), winter (WI), spring (SP), and summer (SM) of 2010 from hives located in Santa Catarina state. A total of 59 samples were collected, with the distribution of samples by seasons being: SM – 16 samples, AU and SP – 15 samples, WI – 13 samples. Also, three agroecological regions were defined for the different apiaries, distributed as follows: highlands – 12 samples, plain – 11 samples, plateau – 36 samples.

After preprocessing, data analysis was conducted and some of results are presented below. The one-way ANOVA results in Table 2a indicate that compounds with anomeric structural moieties appear to have a significant effect on the discrimination of propolis samples over the seasons, because all the main resonances selected occur at the anomeric spectral region (3.00 ppm–5.50 ppm).

PCA results revealed a low percentage of variance explained by the first principal components. Fig. 5a shows that there is significant overlapping on the first two components. On the other hand, results from machine learning models trained to discriminate the harvest seasons are shown in Table 2b with 10-fold cross-validation with 10 repetitions. The accuracy, kappa statistic and AUC results were quite good, which can reveal the features that are significant to discriminate the harvest seasons and to successfully predict new samples. A 3D plot of the first 3 components of the PLS model is shown in Fig. 5b, showing a clear separation of the classes.

3.3. Cassava

Cassava is widely cultivated in tropical and subtropical regions for its starchy tuberous root, a great source of carbohydrates. It has a variety of applications, like animal feeding, culinary or alcoholic beverages. As it is a tropical root, it undergoes PPD, which is characterized by streaks of blue/black in the root vascular tissue, which with time spread and cause a brown discoloration. PPD begins quickly within 24 h postharvest and, because of that, the roots need to be rapidly consumed. Some studies revealed that the deterioration is caused mostly from wound-healing responses [15]. This study was conducted to identify changes and discriminate cassava samples from different regions during PPD with the aid of supervised and unsupervised methods of data analysis.

3.4. Mice spinal cord

This case study consists of 12 LC–MS samples of spectra in the netCDF format, from mice spinal cord divided into 2 groups: the wild type and the knock-out group. The data was obtained from the MetaboAnalyst site, originating from a study which describes a general strategy for identifying endogenous substrates of enzymes by untargeted LC–MS analysis of tissue metabolomes from wild-type and enzyme-inactivated organisms [22].

On this case study, the metabolite identification methods were applied over these data and the resulting hits are shown on the report available on the web site.

Table 2 – (a) ANOVA results for discriminating harvest seasons, with the chemical shifts (ppms), *p*-values, the logarithm of the *p*-values, the false discovery rate and the Tukey's HSD results. (b) Classification models result for discriminating the harvest seasons.

(a)				
ppm	<i>p</i> -values	Logs	fdr	Tukey
4.66	9.585e–26	25.018	2.319e–23	sm-au; sp-sm; wi-sm
4.58	3.385e–17	16.470	4.096e–15	sm-au; sp-sm; wi-sm
4.55	6.092e–14	13.215	4.915e–12	sm-au; sp-au; wi-au; sp-sm; wi-sm
4.63	1.044e–13	12.981	6.316e–12	sm-au; sp-sm; wi-sm
4.71	2.083e–13	12.681	1.008e–11	sm-au; sp-sm; wi-sm
(b)				
Model	Accuracy	Kappa statistic		AUC
PLS	86.78%	82.10%		96.94%
Random forest	84.56%	79.08%		94.51%

Fig. 5 – (a) PCA scores plot (PC1 and PC2) grouped by the harvest seasons. (b) 3D plot of the first 3 components of the PLS model with the harvest seasons.

4. Conclusions

To face the numerous challenges brought by metabolomics data, an R package was developed, which includes features and methods for a variety of important aspects, starting with data loading into a defined structure, various methods of pre-processing and visual exploration of the data, methods for metabolite identification, as well as methods for data analysis and machine learning. All these methods have been implemented and demonstrated in distinct real-world case studies.

Since it provides the possibility of quickly creating and visualizing the results of analysis pipelines, the package can be used by anyone with or without a computational background. The package is quite flexible, with functions that are easy to use having default configurations, but with the possibility of configuring more specific details as users get acquainted with the potential of the functions. A web site has also been

created to complement the package functionalities, providing a way for the community to exchange datasets, analysis scripts and reports of the analysis pipelines performed.

Therefore, the resources put forward in this work will be valuable tools for researchers in the growing field of metabolomics. In current and future work, the team will extend the framework with other methods for the covered tasks, but also address new functionalities. In the first case, the task of metabolite identification is still an area where the package can be improved covering other types of data and methods, but also machine learning possibilities can be enhanced by considering other possible frameworks. In the second case, the team will focus in extending the framework with tools for extracting peaks from NMR raw data, handling 2D NMR datasets, metabolite enrichment analysis, pathway analysis and in the implementation of methods for data fusion, which can allow the integrated analysis of different types of data for the same case study.

Conflict of interest

The authors declare they have no conflict of interests regarding this article.

Acknowledgments

The work is partially funded by Project 23060, PEM – Technological Support Platform for Metabolic Engineering, co-funded by FEDER through Portuguese QREN under the scope of the Technological Research and Development Incentive system, North Operational and by Project PropMine, funded by the agreement between Portuguese FCT and Brazilian CNPq.

REFERENCES

- [1] A.H. Banskota, Y. Tezuka, S.H. Kadota, Recent progress in pharmacological research of propolis, *Phytother. Res.* 15 (2001) 561–571.
- [2] C. Beleites, *hyperSpec* Introduction, University of Trieste Spectroscopy – Imaging, IPHT Jena e.V, 2014, May, CENMAT and DI3, <http://cran.r-project.org/web/packages/hyperSpec/vignettes/introduction.pdf>.
- [3] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, *J. R. Stat. Soc. Ser. B* 57 (2015) 289–300.
- [4] G.A. Burdock, Review of the biological properties and toxicity of bee propolis (propolis), *Food Chem. Toxicol.* 36 (1998) 347–363.
- [5] C. Croux, P. Filzmoser, M. Oliveira, Algorithms for projection-pursuit robust principal component analysis, *Chemom. Intell. Lab. Syst.* 87 (2007) 218–225.
- [6] R. Eisner, C. Stretch, T. Eastman, J. Xia, D. Hau, S. Damaraju, R. Greiner, D.S. Wishart, V.E. Baracos, Learning to predict cancer-associated skeletal muscle wasting from H-NMR profiles of urinary metabolites, *Metabolomics* 7 (2010) 25–34.
- [7] W.J. Evans, J.E. Morleya, J. Argilésa, C. Balesa, V. Baracosa, D. Guttridgea, A. Jatoia, K. Kalantar-Zadeha, H. Lochsa, G. Mantovania, D. Marksa, W.E. Mitcha, M. Muscaritolia, A. Najanda, P. Ponikowskia, F.R. Fanellia, M. Schambelana, A. Scholsa, M. Schustera, D. Thomas, R. Wolfea, S.D. Anker, Cachexia, A new definition, *Clin. Nutr.* 27 (2008) 793–799.
- [8] G. Gekker, S. Hu, M. Spivak, J.R. Lokensgard, P.K. Peterson, Anti-HIV-1 activity of propolis in CD4+ lymphocyte and microglial cell cultures, *J. Ethnopharmacol.* 102 (2005) 158–163.
- [9] B.A. Hanson, *Chemospec: An R Package for Chemometric Analysis of Spectroscopic Data and Chromatograms*, 2003 <https://cran.r-project.org/web/packages/ChemoSpec/vignettes/ChemoSpec.pdf>.
- [10] S. Kumazawa, R. Ueda, T. Hamasaka, S. Fukumoto, T. Fujimoto, T. Nakayama, Antioxidant prenylated flavonoids from propolis collected in Okinawa, Japan, *J. Agric. Food Chem.* 55 (2007) 7722–7725.
- [11] F. Mozzi, M.E. Ortiz, J. Bleckwedel, L.D. Vuyst, M. Pescuma, *Metabolomics as a Tool for the Comprehensive Understanding of Fermented and Functional Foods With Lactic Acid Bacteria*, Food Research International, 2012.
- [12] J. Nielsen, M.C. Jewett, *Metabolomics: A Powerful Tool in Systems Biology*, Springer, 2007.
- [13] J.M. Sforcin, Propolis and the immune system: a review, *J. Ethnopharmacol.* 113 (2007) 1–14.
- [14] K. Tan-No, K.T. Nakajima, T. Shoi, O. Nakagawasai, F. Nijima, M. Ishikawa, Y. Endo, T. Sato, S. Satoh, K. Tadano, Anti-inflammatory effect of propolis through nitric oxide production on carrageenin-induced mouse paw edema, *Biol. Pharm. Bull.* 29 (2006) 96–99.
- [15] V.G. Uarrotta, R. Moresco, B. Coelho, E. da Costa Nunes, L.A. Martins Peruch, E. de Oliveira Neubert, M. Rocha, M. Maraschin, Metabolomics combined with chemometric tools (PCA, HCA, PLS-DA and SVM) for screening cassava (*Manihot esculenta* Crantz) roots during postharvest physiological deterioration, *Food Chem.* 161 (2014) 67–78.
- [16] K. Varmuza, P. Filzmoser, *Introduction to Multivariate Statistical Analysis in Chemometrics*, CRC Press, 2008.
- [17] S. Villas-Boas, U. Roessner, M.A.E. Hansen, J. Smedsgaard, J. Nielsen, *Metabolome Analysis: An Introduction*, Wiley, 2007.
- [18] J. Xia, R. Mandal, I.V. Sinelnikov, D. Broadhurst, D.S. Wishart, *Metaboanalyst 2.0 – a comprehensive server for metabolomic data analysis*, *Nucleic Acids Res.* 40 (2012) 127–133.
- [19] C.A. Smith, E.J. Want, G. O'Maille, R. Abagyan, G. Siuzdak, XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching and identification, *Anal. Chem.* 78 (2006) 779–787.
- [20] F. Fernandez-Albert, R. Llorach, C. Andres-Lacueva, A. Perera, An R package to analyse LC/MS metabolomic data: MAIT (metabolite automatic identification toolkit), *Bioinformatics (Oxf., Engl.)* 30 (13) (2014) 1937–1939, <http://dx.doi.org/10.1093/bioinformatics/btu136> <http://bioinformatics.oxfordjournals.org/cgi/content/long/30/13/1937>.
- [21] A. Saghatelian, S.A. Trauger, E.J. Want, E.G. Hawkins, G. Siuzdak, B.F. Cravatt, Assignment of endogenous substrates to enzymes by global metabolite profiling, *Biochemistry* 43 (2004) 14332–14339.
- [22] Recursive Feature Elimination, 2015 <http://topepo.github.io/caret/rfe.html#rfe>.