AlpsNMR: an R package for signal processing of fully untargeted NMR-based metabolomics

Francisco Madrid-Gambin[1,*], Sergio Oller-Moreno[1,*], Luis Fernandez[1,2,*], Simona Bartova[3], Christopher Joyce[3], Francesco Ferraro[3], Ivan Montoliu[3], Sofia Moco[3,+], Santiago Marco[1,2,+]

[1] Signal and Information Processing for Sensing Systems, Institute for Bioengineering of Catalonia (IBEC), The Barcelona Institute of Science and Technology, Baldiri Reixac 10-12, 08028 Barcelona, Spain.

[2] Department of Electronics and Biomedical Engineering, Universitat de Barcelona, Marti i Franqués 1, Barcelona 08028, Spain.

[3] Nestlé Institute of Health Sciences, Nestlé Research, EPFL Innovation Park, H, 1015 Lausanne, Switzerland.

[*] These authors contributed equally and should be considered first authors.

[+] These authors contributed equally as senior authors.

Corresponding author: Luis Fernández, E-mail: lfernandez@ibecbarcelona.eu

Abstract

Summary

NMR-based metabolomics is widely used to obtain a metabolic fingerprint of biological systems. While targeted workflows require previous knowledge of metabolites, prior to statistical analysis, untargeted approaches remain a challenge. Computational tools dealing with fully untargeted NMR-based metabolomics are still scarce or not user-friendly. Therefore, we developed *AlpsNMR* (Automated spectraL Processing System for NMR), an R package that provides automated and efficient signal processing for untargeted NMR metabolomics. *AlpsNMR* includes spectra loading, metadata handling, automated outlier detection, spectra alignment and peak-picking, integration, and normalization. The obtained output can easily be used for further statistical analysis, and it has proven effective in detecting metabolite changes in a test case. The tool allows less experienced users to easily implement this workflow from spectra to a ready-to-use dataset in their routines.

Availability and implementation

The AlpsNMR R package is freely available to download from http://github.com/sipss/AlpsNMR under the GPL-3 license.

Contact: lfernandez@ibecbarcelona.eu or soller@ibecbarcelona.eu or smarco@ibecbarcelona.eu

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Over the last decades, a whole field of research opened for NMR spectroscopy in the analysis of biological samples. Metabolomics-based NMR allows capturing a multitude of metabolites, at their inherent concentrations, present in the same sample, in single experiments. Metabolomics is thus sensitive to biotic and abiotic perturbations and it has been decisive in proposing biomarkers of disease, identifying metabotypes, and disclosing physiological mechanisms (Moco *et al.*, 2007; Vignoli *et al.*, 2019).

In metabolomics, 1D ^{1}H NMR experiments are generally common in biochemical applications. Given the complexity of signals in such a spectrum of a biological sample, with contributions of hundreds of metabolites, spectral signal processing is required. Targeted metabolomics workflows aim to detect (and quantify) a set of pre-defined known metabolites that describe specific biologically-relevant processes. Untargeted approaches aim to comprehensively extract metabolite features derived from the entire spectra, including both known and unknown signatures (Schrimpe-Rutledge *et al.*, 2016). Therefore, typical metabolomics workflows for targeted and untargeted approaches differ in certain steps, such as metabolite identification before or after data acquisition and statistical modelling (Oresic *et al.*, 2009).

Various computational tools have been proposed for targeted or semi-targeted approaches, such as *BATMAN* (Hao *et al.*, 2012), *ASICS* (Tardivel *et al.*, 2017), *rDolphin* (Cañueto *et al.*, 2018), and *AQuA* (Röhnisch *et al.*, 2018). A common feature of these is the use of a predefined metabolite library of reference peak patterns of known metabolites in a particular biofluid (typically plasma/serum). These approaches fail to extract any metabolite or feature not present in this metabolite library, even if present at high concentration and potentially informative. Untargeted approaches attempt to fill this gap, by capturing as many metabolite features as possible, including unknown compounds (or not predefined) (Alonso *et al.*, 2015). Most computational routines for untargeted NMR metabolomics analyses are performed by means of commercial software (Weber *et al.*, 2017), with few exceptions of packages developed in R, as *ChemoSpec* to handle spectra alignment, binning and certain statistical analyses (Hanson, 2016), and *speaq* for alignment and wavelet-based peak detection algorithms (Beirnaert *et al.*, 2018).

Here, we present the R package *AlpsNMR* for NMR signal processing capable of: importing ^{1}H NMR spectra, handling metadata, performing spectrum interpolation, automated outlier detection, spectra alignment, peak-picking, integration and normalization, and delivering a combined reduced dataset ready for statistical analysis or machine learning. This user-friendly step-by-step workflow allows data analysts and NMR users to treat NMR-based metabolomics datasets in a fully untargeted manner.

2 IMPLEMENTATION AND MAIN FUNCTIONS

The workflow of *AlpsNMR* includes a sequential procedure from data import to final dataset (Figure 1A). In the loading step (step 1), the AlpsNMR_nmr_read_samples_dir function imports the directory containing multiple spectra, for instance, a Bruker data directory. Then, the nmr_meta_add function automatically matches external and/or analytical metadata from a Microsoft Excel file (step 2) with the NMR dataset. At this point, the spectra are ready for interpolation (step 3) and exclusion of the solvent regions (*e.g.* water, methanol, or other; step 4). The *AlpsNMR* package includes robust principal component analysis (rPCA) for outlier detection (step 5), with a proposed threshold, based on quantiles, for Q residual and T2 score values, less sensitive to extreme intensities due to, for instance, outliers (Hubert and Engelen, 2004). The nmr_pca_outliers_filter function automatically detects, reports and removes, if applicable, outliers from the NMR dataset. These can be visualised by the user for further exploration. The nmr_detect_peaks_plot function (step 6) provides visualization of signals, so that users can edit the sensitivity of the algorithm to obtain either a more exhaustive or a reduced peak table as a final output (step 10). The peak alignment step (step 7) is based on hierarchical cluster-based peak alignment (CluPA) (Vu *et al.*, 2011) by means of a *speaq* function (Beirnaert *et al.*, 2018). Normalization by an internal calibrant and/or the probabilistic quotient normalization (PQN) (Dieterle *et al.*, 2006) may be performed by the nmr_normalize function (step 8). The plot_interactive function produces an interactive plot created with *plotly* in a HTML file that may be edited, zoomed and downloaded as a *PNG* figure. The integration of peaks (step 9) may be performed automatically, from the detected peaks, or manually, in which users can select a specific region of interest, after which the dataset is generated (step 10) for further data analysis. Finally, the tool has functions to export data in several formats compatible with packages such as *BATMAN, ASICS* and *ChemoSpec* to run, for example, other targeted and/or machine learning analyses.

AlpsNMR is based on R environment. The package is hosted on github.com/sipss/AlpsNMR and a version will be available as an R package on CRAN. *AlpsNMR* was created on standard CRAN and Bioconductor characteristics, and thus, allows the flexibility and cooperation with other R based packages. It can be executed under Windows, Mac and Linux operating systems. *AlpsNMR* supports 1D NMR spectral data files from formats JCAMP-DX and/or the vendor Bruker, R data format and text file as input. Various 1D NMR pulse sequences are supported, such as NOESY-1D (Nuclear Overhauser Effect SpectroscopY) 1D and CPMG (Carr-Purcell-Meiboom-Gill sequence).

3 CASE STUDY

AlpsNMR was applied on a publicly available dataset (MTBLS242) found in the repository MetaboLights (Haug *et al.*, 2013). This dataset of 106 ^{1}H NMR CPMG spectra of human serum were obtained from a longitudinal study on severe obese after bariatric surgery (Gralka *et al.*, 2015), Supplementary Material. After importing the NMR directory and metadata (Figure 1B, step 3), *AlpsNMR* detected one outlier (Figure 1B, step 5 and Figure S1) prior to spectral alignment. After peak detection (Figure 1B, step 6), alignment, PQN-normalization (Figure 1B, step 8) and peak integration, the final dataset consisted of 686 signals. This reduction of dimensionality was proven essential in discovering significant metabolites, when challenged to a test case. In fact, *AlpsNMR* proved successful in replicating the reported results obtained by of Gralka *et al*, as shown in Table 1S, Supplementary Material, after biostatistical analysis. The workflow took 20 min to analyse this dataset using a twelve-core workstation.

4 CONCLUSION

In sum, the R package *AlpsNMR* was developed to perform pre-processing of 1D NMR-based metabolomics datasets. It includes a workflow organised in steps, after which a reduced output of signal intensities according to chemical shift is generated. This peak list can be then taken for further statistical analysis and metabolite identification, feasible to be imported by other tools or packages. Using the dataset MTBLS242, as a test case, our package succeeded in identifying significant metabolites, in accordance to previously reported. We believe this tool to be of use for the NMR metabolomics community, as it allows for automated and integrated data processing in a fully untargeted way.

FUNDING

This work was supported by Nestle Research. We would like to acknowledge, the Departament d'Universitats, Recerca i Societat de la Informació de la Generalitat de Catalunya (expedient 2017 SGR 1721); the Comissionat per a Universitats i Recerca del DIUE de la Generalitat de Catalunya; and the European Social Fund (ESF). Additional financial support has been provided by the Institut de Bioenginyeria de Catalunya (IBEC). IBEC is a member of the CERCA Programme/Generalitat de Catalunya.

Conflict of Interest: The authors S.B., C.J., F.F., I.M. and S.M. are Nestle Research employees.

ACKNOWLEDGEMENT

All authors thank Dr. Leonardo Tenori for clarification on the metadata and Prof. Dr. med. Bernd Schultes for clarification on the ethical status of the case study used in this manuscript.

REFERENCES

Alonso,A. *et al.* (2015) Analytical Methods in Untargeted Metabolomics: State of the Art in 2015. *Front. Bioeng. Biotechnol.*, **3**.

Beirnaert,C. *et al.* (2018) speaq 2.0: A complete workflow for high-throughput 1D NMR spectra processing and quantification. *PLoS Comput. Biol.*

Cañueto,D. *et al.* (2018) rDolphin: a GUI R package for proficient automatic profiling of 1D1H-NMR spectra of study datasets. *Metabolomics*, **14**, 1–5.

Dieterle,F. *et al.* (2006) Probabilistic Quotient Normalization as Robust Method to Account for Dilution of Complex Biological Mixtures. Application in ^{1} H NMR Metabonomics. *Anal. Chem.*, **78**, 4281–4290.

Gralka,E. *et al.* (2015) Metabolomic fingerprint of severe obesity is dynamically affected by bariatric surgery in a procedure-dependent manner. *Am. J. Clin. Nutr.*, **102**, 1313–1322.

Hanson,B.A. (2016) ChemoSpec: Exploratory Chemometrics for Spectroscopy. **R package**.

Hao,J. *et al.* (2012) BATMAN--an R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a Bayesian model. *Bioinformatics*, **28**, 2088–2090.

Haug,K. *et al.* (2013) MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res.*, **41**, D781–D786.

Hubert,M. and Engelen,S. (2004) Robust PCA and classification in biosciences. *Bioinformatics*, **20**, 1728–1736.

Moco,S. *et al.* (2007) Metabolomics technologies and metabolite identification. *TrAC Trends Anal. Chem.*, **26**, 855–866.

Oresic,M. *et al.* (2009) Metabolomics, a novel tool for studies of nutrition, metabolism and lipid dysfunction. *Nutr. Metab. Cardiovasc. Dis.*, **19**, 816–24.

R Core Team (2015) R: A language and environment for statistical computing. *R Found. Stat. Comput.*

Röhnisch,H.E. *et al.* (2018) AQuA: An Automated Quantification Algorithm for High-Throughput NMR-Based Metabolomics and Its Application in Human Plasma. *Anal. Chem.*, **90**, 2095–2102.

Schrimpe-Rutledge,A.C. *et al.* (2016) Untargeted Metabolomics Strategies—Challenges and Emerging Directions. *J. Am. Soc. Mass Spectrom.*, **27**, 1897–1905.

Tardivel,P.J.C. *et al.* (2017) ASICS: an automatic method for identification and quantification of metabolites in complex 1D1H NMR spectra. *Metabolomics*, **13**, 1–9.

Vignoli,A. *et al.* (2019) High-Throughput Metabolomics by 1D NMR. *Angew. Chemie Int. Ed.*, **58**, 968–994.

Vu,T.N. *et al.* (2011) An integrated workflow for robust alignment and simplified quantitative analysis of NMR spectrometry data. *BMC Bioinformatics*, **12**, 405.

Weber,R.J.M. *et al.* (2017) Computational tools and workflows in metabolomics: An international survey highlights the opportunity for harmonisation through Galaxy. *Metabolomics*, **13**, 12.

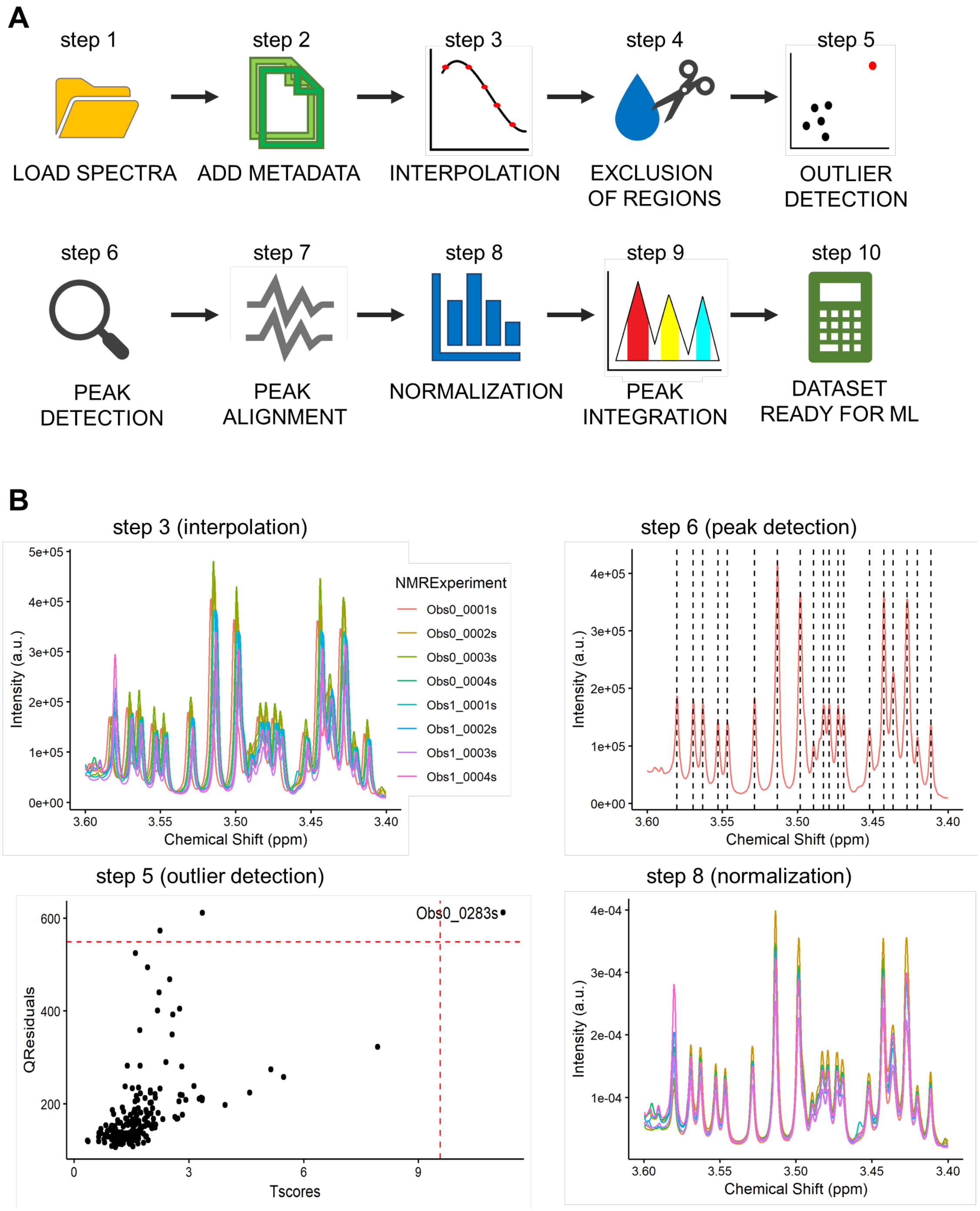

Figure 1. A, Workflow for NMR-based metabolomics data pre-processing using *AlpsNMR* to obtain a dataset ready for machine learning (ML) or statistical analysis. B, MTBLS242 dataset was submitted to the AlpsNMR workflow as a case study: step 3 (top left plot, interpolation), overlay of 8 unprocessed ^{1}H NMR spectra on 3.60-3.40 ppm region, after interpolation; step 5 (bottom left plot, outlier detection), diagnostic of outlier detection based on robust principal component analysis (outliers are labelled); step 6 (top

right plot, peak detection), spectral region of detected peaks on the chosen reference spectrum (Obs1_0323s); step 8 (bottom right plot, normalization), overlay of the same 8 ^{1}H NMR spectra on 3.6-3.40 ppm region, after normalisation.