rDolphin: a GUI R package for proficient automatic profiling of 1D ^{1}H-NMR spectra of study datasets

Daniel Cañueto[1]*, Josep Gómez[2], Reza M. Salek[3], Xavier Correig[1,4], Nicolau Cañellas[1,4]*

[1]Metabolomics Platform, IISPV, DEEEA, Universitat Rovira i Virgili, Campus Sescelades, Carretera de Valls, s/n, 43007 Tarragona, Catalonia, Spain, [2]Intensive Care Unit, Joan XXIII University Hospital, IISPV, Carrer Dr. Mallafrè Guasch, 4, 43005 Tarragona, Catalonia, Spain, [3]European Bioinformatics Institute (EMBL-EBI), European Molecular Biology Laboratory, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, [4]CIBERDEM, Spanish Biomedical Research Centre in Diabetes and Associated Metabolic Disorders, Madrid, Spain

*To whom correspondence should be addressed.

ABSTRACT

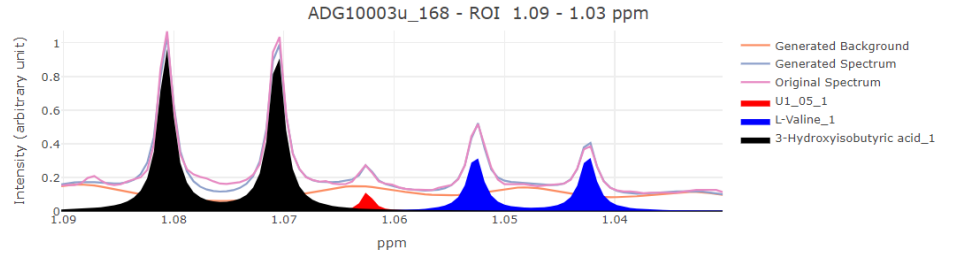
Despite the emergence of automatic profiling tools for 1D ^{1}H-NMR based metabolomics, their adoption and usage still lag behind manual profiling or fingerprinting approaches. Existing tools may not be able to fully contribute the necessary information, flexibility and interactivity to adapt to study dataset properties. Dolphin, a project to perform automatic profiling supervised through GUI, has been redesigned in R language to fully integrate these needs. This redesign incorporates novel techniques to optimize exploratory analysis, metabolite profiling approach, identification and validation of the results in a study, resulting in an application with the best balance between accuracy, reproducibility and ease of use.

INTRODUCTION

^{1}H-NMR is a high throughput analytical technique that allows the quantification of metabolites in biofluids, tissues or cell culture extracts in a reliable and reproducible manner. However, variability in the sample properties and preparation and during the spectra acquisition and preprocessing incorporates complexity to the generated spectra (Sokolenko et al. 2013). Examples of this complexity are baseline artefacts due to macromolecules, broad signals originated from lipids, signal overlappings and misalignments, and variability of signal shapes. Profiling approaches can provide more resilience to these sources of variability than fingerprinting ones (Weljie et al. 2006).

In comparison to manual profiling (e.g. through Chenomx), with high variability depending on the user (Tredwell et al. 2011); automatic profiling promises more robustness as well as.lower time-consuming demands.Several tools that perform automatic 1D 1H-NMR profiling have been already published like Dolphin, BATMAN or BAYESIL (Gómez et al. 2014; Hao et al. 2014; Ravanbakhsh et al. 2015; additional options are available on Spicer et al. 2017). Nonetheless, they may require extensive knowledge of the dataset properties or expertise in programming languages to be fully utilized. Other possible drawbacks are the reliance on strict parameter settings for sample preparation and data acquisition, or the inability to handle unknown signals. Most automatic tools may lose performance when applied to a large number of complex biological samples: the complexity to monitor may become too demanding to be efficiently controlled in a single profiling iteration, therefore several iterations with different parameters are needed to avoid wrong annotations and suboptimal quantifications. These issues hamper the reproducibility of the profiling based approaches compared to fingerprint based ones. The challenges become exacerbated if the user does not have previous expertise with the matrix in hand.

Dolphin (Gómez et al. 2014) is a project that uses the region of interest (ROI) concept (Lewis et al. 2009) to allow the flexible and time-affordable automatic profiling of 1D ^{1}H-NMR spectra. Dolphin implements an approach based on the calculation of the baseline and of the signal parameters (chemical shift, intensity, half bandwidth, j-coupling) values which maximize the lineshape fitting of the signals in the analyzed ROI (Figure 1). The supervision of adequate starting estimates for this calculation is performed through GUI. The challenges observed during the profiling of study datasets of complex matrices emphasized the need for a new framework with novel solutions to optimize the exploratory analysis and metabolite identification of study datasets. This framework should also enable the interactive validation and optimization of the performed annotations and quantifications without the need to generate new profiling iterations. It should as well allow the loading and analysis of metabolite profiling sessions on any computer: the enhanced reproducibility may help heighten the standardization and quality of profiling approaches for any matrix (Rocca-Serra et al. 2016). Standardisation would also be heightened by the availability of information about



Here you have some indicators of the quantification.

Show 10 entries Search:

	Quantification (arbitrary unit)	Fitting Error	Signal/total area ratio
U1_05_1	0.0004	0.0201	13.2574
L-Valine_1	0.0019	0.0185	44.2583
3-Hydroxyisobutyric acid_1	0.0051	0.0299	74.4261

Figure 1: Example of line-shape fitting with baseline in the 1.09-1.03 ppm region of the human urine MTBLS1 dataset. Information of each quantified signal area and of some indicators of fit quality are shown below the interactive Plotly figure. The predefined ROI profile can be modified if necessary to optimize the line shape fitting on the studied spectrum.

metabolite identification in previous studies in the matrix analyzed. Here we present a redesign implementation of Dolphin workflow using open-source R software, called rDolphin. rDolphin has been specifically redesigned in order to satisfy these needs during the profiling of study datasets of complex matrices.

MATERIALS AND METHODS

The Dolphin workflow has been extensively tested with around 4050 different spectra across 25 different metabolomics studies with varying biological complexity, ranging from urine, blood to aqueous and lipidic cell extract. The workflow requires as input preprocessed spectra (in Bruker or CSV format) which can be then normalized or aligned in different ways through the tool.Then, ROI profiles are adapted to the dataset and metabolite relative concentrations are quantified and outputted (and can be converted to absolute ones through TSP or ERETIC quantification). In order to show the improvements achieved during the adaptation of rDolphin to common issues found in such datasets, we have selected two publicly available studies from the MetaboLights repository (Haug et al. 2013):

- MTBLS1: This study contains 132 spectra of human urine samples.
- MTBLS237: This study contains 114 spectra of human faecal extract.

Supplementary Material contains sample preparation and spectra acquisition and preprocessing details.

The redesign of the original MATLAB-based Dolphin workflow to elaborate strategies to solve common challenges in study datasets recommended the use of R language. This programming language enables the adoption of state-of-the-art statistical approaches freely available on current and future packages. In addition, the use of open-source language facilitates loading the profiling information in .RData format on any computer in order to be revised and updated. In addition, the package designed incorporated a Shiny GUI with interactive Plotly figures and interactive data tables that ease the use of its capabilities by users not proficient in programming languages.

RESULTS AND DISCUSSION

Improvement and time reduction of exploratory analysis

To assist during exploratory data analysis, rDolphin incorporated the visualization through interactive Plotly figures for a subset of representative spectra, which capture the variance within the entire dataset, selected through the affinity propagation algorithm (Bodenhofer et al. 2011). In addition, the median spectrum for each group of spectra (according to metadata specified by the user) is visualized (Figure 2). Both kinds of dataset visualization were combined with fingerprint analysis information. This information points to the regions that vary across the different studied sample groups, giving possible important information on where to focus the profiling efforts on. These tools greatly ease the intricate process of exploratory analysis of different kinds of information in order to solve doubts in the choice or annotation of signals to profile (because of signal overlap and metabolite concentration variability). The eased dataset visualization also helps monitor fluctuations in the chemical shift of signals (caused by pH and ionic strength variability) or in other signal parameters.

The necessary exploratory analysis to minimize the influence of extraneous factors in the dataset becomes only harder when studying not familiar matrices.

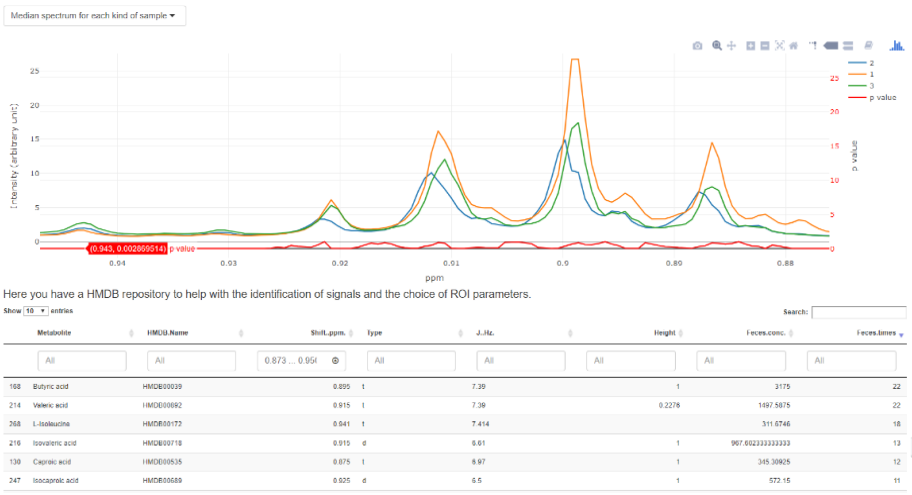


Figure 2: exploratory analysis of human faecal extract MTBLS237 dataset with rDolphin. Differences between the median spectrum of three kinds of sample in the 0.94-0.88 ppm region are shown on an interactive figure. The user can analyze in the below table the most relevant signals for feces in this region according to HMDB data and add them to the list of signals to profile.

When there is not extensive information related to the matrix, the process suffers from high uncertainty liable to researcher's resources and biases (e.g. identifying a metabolite not typical from the biofluid). To further assist in the exploratory analysis of not familiar matrices, a repository of signals was created based on ^{1}H-NMR reference data of the Human Metabolome Database (HMDB) (Forsythe & Wishart⁋009) (Figure 2). This repository is able to filter the metabolites signals by the studied matrix avoiding assignments to metabolites not expected to be found in this matrix. In addition, metabolite signals can be ranked by the times the metabolite has been reported in the matrix according to the HMDB and by the mean concentration reported. These options ensure the correct identification of a signal when several options are possible.

Thanks to these novel interactive assistance options, the preparation of the necessary information to profile the MTBLS237 dataset (a dataset from a previously not studied matrix) only endured three hours on a standard computer. The consistency of the performed metabolite identification with the available HMDB data suggests that our framework may enable a quicker standardisation in the metabolite profiling of emerging matrices. In the case of the MTBLS1 dataset, most efforts during exploratory analysis were focused on the modification of chemical shift information (in order to control for the buffer influence) and on the identification of some metabolite signals. Evaluation of signals to add and remove from the dataset in order to maximize insights was also performed with the help of the Plotly figures. The process endured less than two hours.

For the MTBLS1 dataset, 40 metabolites were quantified, a standard number of metabolites to profile in human urine (Supplementary Figure 1, a). In the case of the MTBLS237 dataset, 34 metabolites were profiled (Supplementary Figure 1, b), being able to expand the number of metabolites from the 13 analyzed ones in the original study. Profiling results show that common challenges found during exploratory analysis were efficiently monitored thanks to the novel options provided. On Supplementary Material and rDolphin Github website, we share examples in .RData format of the profiling sessions performed on both datasets.

Avoidance of possible suboptimal quantifications

rDolphin, thanks to interactive data tables with indicators of quality and reliability for each quantification, allows the finding and modification of suboptimal quantifications and annotations caused by signal misalignment or by the interference of neighboring signals. The indicators of quality and reliability are: fitting error, ratio between quantified signal area and spectrum region area, expected chemical shift, expected half bandwidth, expected intensity. The last three indicators are novel information sources enabled by machine learning based prediction of these signal parameters according to information extracted from correlated signals to the one of interest. In Supplementary Material, we show how this information enables the finding of wrong annotations e.g. in carnitine quantification the MTBLS1 dataset (Supplementary Figure 2). When suboptimal quantifications or wrong annotations are found, quantifications can be loaded and optimal lineshape fittings for the signal can be performed (through the choice of new starting estimates for the signal parameters or through the manual edition of the lineshape fitted signals and baseline in the ROI).

Limited resolution, metabolite concentration variability and signal misalignment create wrong annotations of overlapping signals and limit the effectiveness of automatic approaches to accurately annotate and quantify the signals of interest in all spectra. The provided quality information of quantifications added to

the easy optimization of profiling output facilitated by the rDolphin GUI provides the necessary framework to increase the quality and robustness to NMR limitations of profiling strategies. This increase will be even more important when promising improvements in NMR sensitivity and resolution enable the increase of profiled metabolites in study datasets. These improvements will increase signal overlap and, in the case of pure shift NMR, remove the multiplicities that eased identification: optimal approaches for reliable identification and deconvolution of signals will become even more necessary. In addition, the possibility of storing all profiling information in open-source format enables a much higher reproducibility of the profiling process performed in the study dataset.

Identification of unknown or wrongly identified signals

rDolphin provides the option of metabolite identification through STOCSY (Cloarec et al. 2005) or RANSY (Wei et al. 2011). However, these tools might be sometimes limited by factors such as signal misalignment, baseline, or correlated metabolites. In order to maximize metabolite identification capabilities, rDolphin incorporates dendrogram heatmaps of quantification and chemical shift of the profiled signals in order to help in their identification. Signals show correlations in several signal parameters according to e.g. chemical structure (in the case of chemical shift) or biological pathway (in the case of quantification). These correlations can be exploited to explore the clustering of signals of not identified metabolites with signals from identified ones. The clustering of the unidentified signal with known metabolites provides insightful chemical and biological information to help identify the metabolite. In addition, the evaluation of clusters is not limited to quantified metabolites: evaluation of spectra clusters is also possible.

This novel tool helped solve two inconsistencies between metabolite identification in the original MTBLS1 study and in previous datasets of the same matrix performed by our research group. In Supplementary Material, we show how these inconsistencies were evaluated through the analysis of signals with similar patterns to the signal to identify (Supplementary Figure 3). In addition, we demonstrate that a broad triplet at 4.04 ppm present in some human urine datasets is a signal closely related to creatinine observable in spectra of internal standards of creatinine (Supplementary Figure 4). To our knowledge, this identification represents a novelty in the profiling of human urine datasets.

Metabolite identification remains one of the biggest bottlenecks in metabolomics (van der Hooft and Rankin 2016). rDolphin incorporates a novel technique to the metabolomics community to help in the right identification of metabolites in complex matrices.

FUTURE WORK

The ROI approach, currently implemented in rDolphin, does not relate signals from the same metabolite placed in different ROIs. However, a novel approach to relate these signals is already available in a developmental stage on the package. Additionally, rDolphin cannot deconvolute signals that are more complex than quadruplets (although these can be decomposed in substructures and then fitted). The information of metabolite identification and concentration for matrices is limited to human biofluids present in the HMDB database. In addition, concentration and identification information become less accurate the less studied is the biofluid. Nonetheless, we are happy to study and share the optimal ROI profiles for any matrix of interest for the metabolomics community.

Funding D.C. thanks to the Universitat Rovira i Virgili for providing the Ph.D scholarship and the EMBL-EBI for the visitor programme funding. X.C. and N.C acknowledge the project TEC2015-69076-P financed by the Ministerio de Economía y Competitividad.

Compliance with ethical requirements

Conflict of interest Daniel Cañueto, Josep Gómez, Reza M. Salek, Xavier Correig and Nicolau Cañellas declare no conflict of interest.

Ethical approval and Informed consent This study analysed previously collected data which involved human participants who had provided informed consent. These ethical issues are described in detail in the two primary research papers published by (Salek et al. 2006; Bjerrum et al. 2014).

REFERENCES

Bjerrum, J.T. et al., 2014. Metabonomics of human fecal extracts characterize ulcerative colitis, Crohn's disease and healthy individuals. Metabolomics: Official journal of the Metabolomic Society, 11(1), pp.122–133.

Bodenhofer, U., Kothmeier, A. & Hochreiter, S., 2011. APCluster: an R package for affinity propagation clustering. Bioinformatics , 27(17), pp.2463–2464.

Cloarec, O. et al., 2005. Statistical Total Correlation Spectroscopy: An Exploratory Approach for Latent Biomarker Identification from Metabolic1H NMR Data Sets. Analytical chemistry, 77(5), pp.1282–1289.

Forsythe, I.J. & Wishart, D.S., 2009. Exploring human metabolites using the human metabolome database. Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.], Chapter 14, p.Unit14.8.

Gómez, J. et al., 2014. Dolphin: a tool for automatic targeted metabolite profiling using 1D and 2D (1)H-NMR data. Analytical and bioanalytical chemistry, 406(30), pp.7967–7976.

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(15), pp.2088–2090.

Haug, K. et al., 2013. MetaboLights--an open-access general-purpose repository for metabolomics studies and associated meta-data. Nucleic acids research, 41(Database issue), pp.D781–6.

Lewis, I.A. et al., 2009. rNMR: open source software for identifying and quantifying metabolites in NMR spectra. Magnetic resonance in chemistry: MRC, 47 Suppl 1, pp.S123–6.

Ravanbakhsh, S. et al., 2015. Accurate, fully-automated NMR spectral profiling for metabolomics. PloS one, 10(5), p.e0124219.

Rocca-Serra, P. et al., 2016. Data standards can boost metabolomics research, and if there is a will, there is a way. Metabolomics: Official journal of the Metabolomic Society, 12, p.14.

Salek, R.M. et al., 2006. A metabolomic comparison of urinary changes in type 2 diabetes in mouse, rat, and human. Physiological Genomics, 29, 99–108.

Sokolenko, S. et al., 2013. Understanding the variability of compound quantification from targeted profiling metabolomics of 1D-1H-NMR spectra in synthetic mixtures and urine with additional insights on choice of pulse sequences and robotic sampling. Metabolomics: Official journal of the Metabolomic Society, 9(4), pp.887–903.

Spicer, R. et al., 2017. Navigating freely-available software tools for metabolomics analysis. Metabolomics: Official journal of the Metabolomic Society, 13(9).

Tredwell, G.D. et al., 2011. Between-person comparison of metabolite fitting for NMR-based quantitative metabolomics. Analytical chemistry, 83(22), pp.8683–8687.

van der Hooft, J.J.J. & Rankin, N., 2016. Metabolite Identification in Complex Mixtures Using Nuclear Magnetic Resonance Spectroscopy. In Modern Magnetic Resonance. pp. 1–32.

Wei, S. et al., 2011. Ratio analysis nuclear magnetic resonance spectroscopy for selective metabolite identification in complex samples. Analytical chemistry, 83(20), pp.7616–7623.

Weljie, A.M. et al., 2006. Targeted profiling: quantitative analysis of 1H NMR metabolomics data. Analytical chemistry, 78(13), pp.4430–4442.