muma, An R Package for Metabolomics Univariate and Multivariate Statistical Analysis

Edoardo Gaude*1, Francesca Chignola1, Dimitrios Spiliotopoulos1, Andrea Spitaleri1, Michela Ghiti1, Jose M Garcia-Manteiga2, Silvia Mari*1,3 and Giovanna Musco1

1Dulbecco Telethon Institute, Biomolecular NMR Laboratory c/o Center for Translational Genomics and Bioinformatics, Ospedale San Raffaele, Milano, Italy; 2Genome Function, Center for Translational Genomics and Bioinformatics, Ospedale San Raffaele, Milano, Italy; 3R4R di Mari Silvia, Rodano, Milano, Italy

Abstract: Metabolomics, similarly to other high-throughput "-omics" techniques, generates large arrays of data, whose analysis and interpretation can be difficult and not always straightforward. Several software for the detailed metabolomics statistical analysis are available, however there is a lack of simple protocols guiding the user through a standard statistical analysis of the data.

Herein we present "muma", an R package providing a simple step-wise pipeline for metabolomics univariate and multivariate statistical analyses. Based on published statistical algorithms and techniques, muma provides user-friendly tools for the whole process of data analysis, ranging from data imputation and preprocessing, to dataset exploration, to data interpretation through unsupervised/supervised multivariate and/or univariate techniques. Of note, specific tools and graphics aiding the explanation of statistical outcomes have been developed. Finally, a section dedicated to metabolomics data interpretation has been implemented, providing specific techniques for molecular assignments and biochemical interpretation of metabolic patterns.

muma is a free, user-friendly and versatile tool suite tailored to assist the user in the interpretation of metabolomics data in the identification of biomarkers and in the analysis of metabolic patterns.

Keywords: Metabolomics, metabonomics, statistical analysis, R package, metabolic pattern, chemometrics, multivariate analysis, univariate analysis.

BACKGROUND

Metabolomics studies and analyzes small molecules or metabolites (e.g. lipids, amino acids, sugars, cofactors and other small compounds) in biological samples such as cells, tissues or biofluids. Metabolomics aims to understand and to identify the metabolic variations in response to genetic or environmental changes, such as pathological vs. physiological conditions, pharmacological treatments or conditionings. It is widely applied in plant biology, metabolic engineering for industrial activities, biomedicine and nutrition studies [1].

The most commonly analytical platforms in metabolomics are nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS) [2], which are high-throughput platforms usually generating complex and high-dimensional datasets. The complexity of metabolomics data requires therefore multivariate and multidimensional statistical analyses to facilitate their understanding and interpretation [3]. Several open-source tools [4-9] and proprietary software (e.g. SIMCA-P, Umetrics; Unscrambler X, CAMO; [10]) for the detailed metabolomics statistical analysis are available (Table 1). All these resources are well organized and user-friendly. However, the enormous array of the possible analytical options they offer does not facilitate the user in the identification of straightforward statistical pipelines or protocols, suitable for the analysis of metabolomics datasets in the first investigation stages.

Herein we propose muma (Metabolomics Univariate and Multivariate Analysis), an R software package guiding the user to a step-wise pipeline for data analysis and interpretation. Unlike the majority of R packages released, muma is designed for those users who are not R experts and all its analytical functions they offer does not facilitate the user following the user-friendly muma Tutorial (Supplementary Information).

This protocol provides a pipeline for the whole process of metabolomics statistical analysis, from data preprocessing, to dataset exploration and visualization, to the identification of potentially interesting variables (or metabolites). To this aim, we implemented multivariate techniques commonly applied in metabolomics data analysis, such as principal component analysis (PCA), partial least squares discriminant analysis (PLS-DA) and orthogonal projection to latent structures (OPLS-DA). In addition to multidimensional methods well-established univariate statistical techniques are provided, supporting the interpretation of results with clear and ready-to-use graphical outcomes. We also implemented a useful statistical method to evaluate the performance of PCA.
| Commercial software | | | |
|---------------------|-----------------|---------------|
| | Main features | | NMR specific; Spectra management and preprocessing; Dataset exploration and classification |
| | Advantages | | Reference database of 1D, 2D and J-Res spectra; Interaction between statistical results and spectra; Direct link with on-line repositories (e.g. HMDB, KEGG) |
| | Disadvantages | | Less suited for not-NMR data; No univariate analysis; Compound library customization not well developed |
| Name | CAMO Unscrambler X | link | http://www.camo.com/rt/Products/Unscrambler/unscrambler.html |
| | Main features | | Data pretreatment; Exploratory data analysis; Regression and Classification; Design of Experiment |
| | Advantages | | Univariate techniques; PCA Rotation methods; L-PLSR; Possibility of on-line implementation |
| | Disadvantages | | Non exhaustive array of techniques for model validation |
| Name | PLS-Toolbox | link | http://www.eigenvector.com/software/pls_toolbox.htm |
| | Main features | | Spectral and data preprocessing; Model building and validation; Extensive Design of Experiment |
| | Advantages | | Possibility to build ad hoc scripts; Multiway PCA; User-friendly pipeline |
| | Disadvantages | | Requires MatLab licence (or higher user upfront analysis); No univariate analysis |
| Name | Umetrics SIMCA | link | http://www.umetrics.com/simca |
| | Main features | | Data visualization and data pretreatment; Data modeling and validation; Interactivity between different techniques |
| | Advantages | | O2PLS and O2PLSDA; Wavelet denoising; Extensive model; Validation techniques; Graphical outcomes |
| | Disadvantages | | Univariate techniques |

| Open-source or web-based tools | | | |
|-----------------------------|-----------------|---------------|
| Name | MeltDB [19] | link | https://meltdb.cebitec.uni-bielefeld.de/cgi-bin/login.cgi |
| | Main features | | Raw LC- or GC-MS; spectra management; Integrated R functions (e.g. T-test/ANOVA, PCA, clustering) |
| | Advantages | | Linked with other -omics DB; On-line forum for real time discussion |
| | Disadvantages | | Not complete array of statistical techniques; Mass spec data only; Web-based only |
Open-source or web-based tools

<table>
<thead>
<tr>
<th>Name</th>
<th>MetaboAnalyst [20]</th>
</tr>
</thead>
<tbody>
<tr>
<td>link</td>
<td><a href="http://www.metaboanalyst.ca/MetaboAnalyst/faces/Home.jsp">http://www.metaboanalyst.ca/MetaboAnalyst/faces/Home.jsp</a></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Main features</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metabolomic specific data analysis tools; Data filtering, editing and normalization; Data quality check (batch effect, temporal drift) Univariate and Multivariate modeling</td>
<td>Time course analysis; Functional interpretation (Over Representation Analysis, Metabolic Pathway Analysis …); Link to DBs and pathway visualization</td>
<td>Lack of univariate techniques for not-normally distributed variables; Not straightforward analysis pipeline; Web-based only</td>
</tr>
</tbody>
</table>

Name | MetaP [21] |
link | http://compute1.lsrc.duke.edu/softwares/MetaP/metap.php |

<table>
<thead>
<tr>
<th>Main features</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data exploration and classification; Hypothesis tests</td>
<td>Ready-to-use PDF report</td>
<td>No data pre-processing; Not complete array of statistical techniques; Internet connection needed</td>
</tr>
</tbody>
</table>

Command line tools – R packages

Name | Bioconductor |
link | http://www.bioconductor.org/ |

<table>
<thead>
<tr>
<th>Main features</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wide array of diverse R packages for almost every step of biological data analysis and interpretation</td>
<td>Constantly updated; Possibility to edit scripts for ad hoc purposes</td>
<td>Most of metabolomics packages are mass spec-specific; Need of R language expertise</td>
</tr>
</tbody>
</table>

Name | Metabonomic [22] |
link | Not available |

<table>
<thead>
<tr>
<th>Main features</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMR spectra management and preprocessing; Data pretreatment; Exploration and classification</td>
<td>Artificial Neural Network (ANN); Group-specific spectral differences</td>
<td>Spectra management not friendly; NMR specific; no longer available on CRAN repository</td>
</tr>
</tbody>
</table>

in terms of clustering power [11]. muma has been designed for the analysis of metabolomics data generated with diverse analytical platforms (NMR, MS, NIR), but it provides additional tools specifically designed to help the user in the interpretation of NMR data, such as Statistical TOtal Correlation Spectroscopy (STOCSSY) [12] and for Ratio Analysis of NMR spectroscopy (RANSY) [13]. Overall muma can be a useful, though not exhaustive tool, that can assist the researcher in the first steps of a metabolomic project. However, in this context, it is worth mentioning that the statistical analyses of metabolomic datasets is a challenging task and requires accurate technical choices to avoid errors and pitfalls. The discussion of all these aspects goes beyond the scope of muma. For a thorough discussion of all these aspects we suggest the reader to refer to seminal reviews quoted in the Tutorial (Supplementary Material).

IMPLEMENTS

In this section we provide a brief introduction to the statistical and graphical tools implemented within muma. Far from being an exhaustive explanation of the statistics behind each tool provided in the package, this introductory part aims to depict the rationale behind the application of each statistical tool to the analysis of metabolomics data. An overview picture of the analytical workflow proposed by muma is represented in Fig. (1).

Data pre-processing and exploration. The first function recalled by muma provides both dataset reading and different pre-processing methods. The dataset is first screened for missing values and four imputation options are available (mean, minimum, half-minimum and zero). Metabolomics measurements resulting with negative values are usually considered as spectral artifacts or noise, therefore these val-
ues are transformed automatically into zeroes before any other treatment of the data. Normalization is then performed on total spectral area, transforming every single variable into a fraction of the total intensity of the “spectrum”. This type of normalization has been chosen because of its wide application in literature and in different analytical platforms (NMR and MS) as well as in other “–omic” techniques (transcriptomics, proteomics). Briefly, normalization helps to flatten the global differences between samples (or spectra), allowing the comparison of the same variable within an array of different spectra.

The user can then chose among different variable scaling methods, including Pareto, Range, Vast, Auto and Median scaling [14]. These data preprocessing procedures can be useful when comparing multiple variables with different levels of intensity and/or distributions, a situation commonly found in most metabolomic studies. To this aim, each variable is mean-centered, i.e. each variable is distributed around the zero, and it is then scaled according to the selected method (for an introduction into scaling methods refer to muma Tutorial in Supplementary Material). Both normalization and scaling are not mandatory in the pipeline, and the analyses can be performed directly on the raw data.

As a first statistical exploratory method principal component analysis (PCA) is applied on the pre-processed data. PCA is an unsupervised multivariate technique able to find patterns of variation within a multidimensional dataset: within a graphic called Scores plot PCA is able to scatter experiment samples according to the variance among them. If the perturbation applied during the metabolomic experiment has influenced the global variance of the system, samples should be scattered according to the experimental condition (pharmacological treatment, disease, genetic background, etc.). In the same way, PCA is able to provide information regarding the variables (or metabolites) that lead to the system’s variance under observation, i.e. to highlight

Fig. (1). Workflow for the statistical analysis provided by muma. Different types of metabolomics data are suited for muma’s analysis, ranging from spectral bins, integrated peaks to metabolite concentrations. According to the type of input data different pre-processing and pre-treatment methods can be applied before dataset exploration. Notably, none of these steps is mandatory for subsequent analyses. Dataset exploration can be performed either with a PCA or with the automated comprehensive decisional algorithm for univariate analysis, leading to the identification of potential biomarkers or metabolic changes between experimental conditions. Classification analyses for dataset denoising and discriminant interpretations are provided, as well as useful tools for biochemical and structural analysis of metabolic features. During each step of the analysis detailed reports on statistical tests, scores and p-values, computed matrices for multivariate techniques (PCA, PLS and OPLS) as well as each graphical outcome recalled by muma are stored in the working directory and are ready to use for publications, presentations or subsequent analyses.
metabolites implicated in the biological response induced by the experimental perturbation. This information is indicated in a different graphical output called PCA Loadings plot. Thanks to PCA’s ability in grasping the variance of the system, this technique can be used as a dataset exploration method for assessing the presence of unexpected sources of variation (e.g., presence of outliers). In order to help the user to select the most appropriate pair of PCs, we implemented a tool able to calculate the statistical significance of cluster separation obtained from each pair of PCs [11] (see Results section).

PCA Scores and Loadings plots corresponding to PC1 and PC2 computed on the pareto-scaled table are recalled. The typical outcome of these first steps of analysis is shown in Fig. (2).

**Discriminant Analysis.** PCA is a very robust technique thanks to its unsupervised features, but it might not be able to grasp the information corresponding to a potential biological response when diverse sources of variance are present in the dataset, such as batch effects, time drifts, high noise levels, etc. that can mask the experimental information.

In these situations, statistical techniques taking advantage of additional information, regarding sample classification or dataset structure (supervised techniques), may be required for extracting the information corresponding to the biological experiment.

As supervised analysis algorithms, *muma* provides both Partial Least Squares Discriminant Analysis (PLS-DA) and Orthogonal Projection to Latent Structures Discriminant Analysis (OPLS-DA), with proper graphical outcomes, as Score plots, W*c plot and S-plot. These two techniques share a similar algorithm and they are able to find correlations between sample classification and variable variation patterns: by using the information regarding the groups of samples they can highlight variables (where present) able to discriminate the experimental conditions. In this way, these techniques may aid to eliminate the undesired or unexpected sources of variation from the dataset, focusing on the “good” part of the information. In this context it should be mentioned that several pitfalls are associated to these techniques: sometimes a potential difference highlighted by PLS or OPLS algorithms may be an artifact of the special attitude of these techniques to find correlations between variables and

---

**Fig. (2).** Graphical outcome of data set exploratory analysis obtained with *muma*. A) Overview of the Scores plots of the first ten principal components (PCs) calculated by PCA for the MetaBc dataset. Square, diamonds, stars, triangles, rectangles represent NMR spectra deriving from the four time points of LPS treatment plus the control group (five groups in total). B) Results of the quantitative analysis of cluster separation performed on each pair of PCs, presented as ranking of the five best-separating PCs with statistical significance of separation and corresponding proportion of variance explained by each pair of PCs. C) PCA Scores plot of the best-separating pair of PCs (PC1 vs. PC2), reporting the control group (black stars) together with Day1 (dark grey triangles), Day2 (grey diamonds), Day3 (light grey dots), Day4 (light grey squares) upon LPS treatment. D) PCA loading plot reporting all the variables (1D-1H NMR peaks) in the data set, with peaks showing high positive or negative values being responsible for cluster separation represented in the PCA Scores plot (C).
group of samples. For a better understanding of potential pitfalls associated with supervised techniques, please refer to muma Tutorial in Supplementary Material.

To our knowledge muma is the only free available metabolomics R-package performing OPLS-DA. The advantage of this supervised technique, compared to PLS-DA, resides in distinguishing the information corresponding to the experimental perturbation (such as pathological conditions, pharmacological treatments) from the portion of the data which is orthogonal, i.e. independent from the experimental perturbation, thereby OPLS-DA results in a better clustering ability and identification of the features characterizing the experimental groups [15].

Automated Comprehensive Univariate Analysis. muma provides a decisional tree algorithm for univariate analysis that can automatically perform rigorous hypothesis tests on each individual variable. One of the main issue when comparing a single variable between two or more groups concerns the normality of its distribution: hypothesis tests designed for normally distributed variables may not be suited for not-normally distributed variables and vice versa. To address this issue we implemented the algorithm previously described by [16] testing the normality of each variable through Shapiro Wilk’s test; depending on the p-value of this test (threshold=0.05), Welch’s T test or Wilcoxon-Mann Whitney (WMW) test are performed in a completely automatic manner. For each tested variable the algorithm reports the p-value corresponding to the appropriate test (Welch or WMW). Moreover, as metabolomics dataset are usually multidimensional, a multiple testing correction can be applied using the Benjamini-Hochberg method.

In addition, volcano plots will be visualized to screen and boxplots of each tested variable will be created and saved in a dedicated directory. For a brief introduction into interpretation of volcano and box plots generated by this function, please refer to muma Tutorial in Supplementary Material.

To the best of our knowledge muma is the only freely available software providing an automated decisional algorithm for univariate testing.

Structural/Biochemical interpretation. To assist metabolomics data interpretation, correlation heatmaps can be created to help the user in the identification of biochemical relationships between different metabolites and potentially highlighting the involvement of metabolite patterns (or pathways) in the context of a biological response. In particular, metabolites showing correlations higher than 0.85 have been proposed to indicate a biochemical (or pathway) relationship [17]. When dealing with NMR spectral data this correlation maps have been referred to as Statistical TOTal Correlation SpectroscopyY (STOCSY) [12] and correlations higher than 0.95 usually indicate a structural relationship between NMR peaks belonging to the same molecule [17]. When a high number of variables is used these heatmaps may result chaotic and of difficult interpretation. To address this issue it is possible to set specific thresholds for positive and negative correlations, in order to visualize the most important relationships only (see muma Tutorial in Supplementary Material).

A unique muma feature is the implementation of Ratio Analysis NMR SpectroscopyY (RANSY), a novel tool which has shown higher abilities and power in the identification of structural relationships between NMR peaks, as compared to existing tools [13]. This innovative tool is based on the concept that peaks belonging to the same molecule should show similar ratios with all the other peaks, across a set of spectra.

Reports. muma provides a detailed report of the analysis. Each muma function generates a directory in which all the data (matrices, p-values, etc) and plots are written and saved, thus allowing at any time the recovery of all the results of a given analysis, in a simple and effective manner (see muma Tutorial in Supplementary Material).

RESULTS AND DISCUSSION

muma package includes an example data set [18] and a tutorial with detailed description of all the functions and of the executable commands (See Tutorial in Supplementary Information). The example dataset (MetaBc dataset) includes 98 NMR peaks deriving from a murine B cell line grown in culture medium and collected before and after the addition of LPS at 4 different time points for a total of N=25 NMR spectra (5 NMR spectra for each time point) [18].

Here we present some results deriving from a typical analysis using muma, with a special focus on novel functions and tools that are uniquely implemented and provided by muma and that are not present in other software dedicated to the statistical analysis of metabolomics data sets.

DATASET EXPLORATION

The first step in muma analysis consists of data preprocessing and dataset exploration. Fig. (2) shows the results deriving from the NMR analysis of the MetaBc data set. As a first tool for dataset exploration muma provides a graphical outcome reporting the first ten principal components of the computed PCA, thus providing a general overview of the dataset (Fig. 2A). To guide the user through this first step, muma performs statistical analysis of cluster separation obtained with different pairs of principal components (PCs) and reports the results in a clear ranking (Fig. 2B). In this output the first five best-separating PCs are reported, with p-values of statistical cluster separation and with corresponding proportion of variance explained by the single PCs. This tool can be also exploited to quantitatively compare the effects of the different scaling types on PCA performance. For a deeper understanding of this specific tool and in particular of how to avoid potential pitfalls while using it, please refer to muma Tutorial in Supplementary Material.

Exploration of MetaBc dataset highlights PC1 and PC2 as the best-separating pair of PCs, with a cumulative separation p-value of 0.046, accounting for the 88.9% of system’s variance (Fig. 2B). At this stage of data exploration, it can be useful to specifically visualize PC1 and PC2 in terms of PCA score and loading plots (Fig. 2C, D). PCA Scores plot (Fig. 2C) shows clustering of the different days of B cells proliferation and differentiation in response to LPS treatment. PCA loading plot (Fig. 2D) indicates that NMR peaks corresponding to glucose (X3.72), lactate (X1.32 and X4.11), glutamine (X2.11 and X2.44) and glu-
tamate (X2.34) are the main metabolites responsible for B cells’ response to LPS exposure.

In summary, herein we have shown that within few straightforward steps muma is able to deliver a first general overview of the data set, giving an unbiased interpretation of the system’s variance.

DISCRIMINANT ANALYSIS

The PCA exploration of MetaBc dataset gave clear indications about potential metabolic differences among experimental time points. Nevertheless, it might be interesting to perform supervised discriminant analysis either to confirm those metabolic variations or to potentially highlight additional ones. Results deriving from the supervised interpretation of MetaBc dataset are presented in (Fig. 3). PLS-DA Scores plot (Fig. 3A) of B cell medium subjected to LPS treatment shows clear separation between different time points, similarly to what was obtained with PCA Scores plot. PLS-DA loading (or w*c) plot (Fig 3B) confirmed the metabolic variations previously highlighted by PCA with a similar pattern.

In the OPLS-DA analysis we present only two of the five groups of the MetaBc dataset (no LPS and day2), as this technique can only be performed on a two-group data set. OPLS-DA Scores plot (Fig. 3C) shows optimal separation between groups and highlights the presence of intragroup variations within the control group (No LPS) that are orthogonal (i.e. unrelated) to experimental perturbation. (Fig. 3D) shows the corresponding S-plot, which identifies the metabolites correlating with groups clustering in the Scores plot. In agreement with previous analyses, chemical shifts corresponding to lactate and glucose are mainly responsible for the separation between time points 0 and day2.

UNIVARIATE ANALYSIS

In accord with the PCA analysis, chemical shift corresponding to lactate (1.32ppm and 4.11ppm), glucose (3.42ppm), glutamine (2.11ppm and 2.44ppm) and glutamate (3.72ppm) have significant p-values and distinguish Day0 vs. Day2 (Fig. 4A). Variable boxplots are automatically created representing the median value, the interquartile range and the minimum/maximum values for that variable, in each experi-

Fig. (3). Supervised multivariate statistical analyses presented as PLS-DA and OPLS-DA outcomes. PLS-DA score scatter plot of the MetaBc dataset, and B) corresponding PLS-DA loading plot. OPLS-DA score scatter plot of group 1 (black stars) vs. group 2 (grey diamonds), showing clear separation. S-plot showing the main variables (upper right and lower left corners) responsible for the separation presented in the OPLS-DA score plot.
An R Package for Metabolomics Univariate and Multivariate

Current Metabolomics, 2013, Vol. 1, No. 2

Fig. (4). Graphical outcome for the interpretation of univariate analysis automatically generated with muma. Volcano plot of group 1 vs. 3 showing all the variables tested with corresponding fold changes and p-values. In black are highlighted those variables that reached significance (p<0.05, horizontal line) and showed a fold change higher than 1.2 (right vertical line) or lower than 0.8 (left vertical line). B) Boxplot showing median value (grey line), interquartile range (box edges) and range (whiskers) of the 1D-1H NMR peak at 1.32ppm (corresponding to lactate) of control B cells medium (black) or different days of LPS exposure (shown as grey scale). C) PCA loading plot showing loading values of each variable as multivariate information together with a color-code indicating the significance of each variable as univariate information. Black circles indicate variables that reached significance (p-value < 0.05) in at least one pairwise comparison. D) Snapshot of the directories automatically created by muma after the execution of the univariate command; the directory created by the univariate analysis is highlighted, containing results from all the performed statistical tests.

BIOCHEMICAL INTERPRETATION AND MOLECULAR ASSIGNMENT

One critical step in almost every metabolomics analysis consists in the assignment of unknown peaks to the corresponding metabolite and in the relationships among different variables that can highlight potential biochemical interaction in the context of the biological response of interest. We applied STOCSY method to the MetaBc dataset without specifying the positive/negative correlation thresholds (Fig. 5A) or visualizing correlations higher than 0.95 or lower than -0.85 (Fig. 5B) in order to simplify the interpretation. Interestingly, NMR peaks corresponding to lactate (1.32 and 4.11) show high positive correlations, indicating a structural relationship. Moreover, lactate peaks showed negative correlations with NMR peaks corresponding to glucose (Fig. 5B). Notably, this observation is not only in accordance with PCA results presented above, but is also indicative of the cellular glycolytic activity able to consume glucose from the media for the production of lactate.

A comparison between mono-dimensional STOCSY vs. RANSY is provided in Fig. (5C-D), where structural relationships with lactate peak (1.32ppm) are sought: as it is shown, STOCSY (Fig. 5C) is able to highlight the correlation between the methyl group (1.32ppm) and the aliphatic hydrogen of C2 carbon of lactate (4.11ppm), but several other peaks show high positive correlations with the lactate’s methyl group. On the other hand RANSY (Fig. 5D) universally identifies the resonances belonging to the same molecular system (lactate), showing the high power of RANSY in the successful identification of structural relationships.

CONCLUSION

In this paper we have presented muma, an R package which assists the user in the first stage of a statistical analysis of a metabolomic data set, guiding him through a simple, step-wise interpretation pipeline. muma has been designed to work on data sets of variable size to address diverse analytical needs including data exploration and visualization, biomarker identification, diagnostic or prognostic metabolite patterns, generation of plots/figures for publications, identification and assignment of new molecular species and biochemical interpretation of metabolic patterns. Herein, we have presented the application of muma on an NMR data set, the software however is very versatile and can be easily applied to MS datasets. As muma is not a web-based tool and works from command line, it can rapidly provide analysis.
results: once the dataset has been correctly formatted, the whole statistical protocol (pre-processing, exploration, multivariate, univariate and correlation/ratio analyses) of a typical metabolomics data set (30x300 matrix) requires less than two minutes on a standard workstation. For these reasons muma could be suitable for on-line implementations within diverse software for spectral management/processing, in the context of an automated metabolomics serial analysis procedure, potentially useful for diagnostic/clinical and industrial quality control applications.

CONFLICT OF INTEREST

The authors declared no conflicts of interest.

ACKNOWLEDGEMENTS

GM thanks the Dulbecco Telethon Institute for financial support (TCP99035). EG thanks the Gruppo Italiano Discus-ione Risonanze Magnetiche for the “Anna Laura Segre” fellowship.

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publishers Web site along with the published article.

ABBREVIATIONS

NMR = nuclear magnetic resonance
MS = mass spectrometry
PCA = principal component analysis
PLS-DA = partial least squares discriminant analysis
OPLS-DA = orthogonal projection to latent structures
NIR = Near Infrared
STOCSY = Statistical TOtal Correlation Spectroscopy

Fig. (5). Application of STOCSY and RANSY for molecular identification. STOCSY correlation heatmap without threshold specification and showing correlations higher than 0.95 (light grey) or lower than -0.85 (black). 1D-STOCSY showing correlations of all variables with the peak at 1.32ppm in order to highlight structural correlations. D) RANSY showing the ratios between NMR peak at 1.32ppm and the other variables, univocally identifying the peak at 4.11ppm as belonging to the same molecule (lactate).
RANSY = Ratio Analysis of NMR spectroscopy
WMW = Wilcoxon-Mann Whitney
LPS = lipopolisaccharide

REFERENCES