BMC Bioinformatics

SOFTWARE Open Access

Identifying and quantifying metabolites by scoring peaks of GC-MS data

Raphael BM Aggio[1,2*], Arno Mayor[2], Sophie Reade[2], Chris SJ Probert[2] and Katya Ruggiero[1]

Abstract

Background: Metabolomics is one of most recent omics technologies. It has been applied on fields such as food science, nutrition, drug discovery and systems biology. For this, gas chromatography-mass spectrometry (GC-MS) has been largely applied and many computational tools have been developed to support the analysis of metabolomics data. Among them, AMDIS is perhaps the most used tool for identifying and quantifying metabolites. However, AMDIS generates a high number of false-positives and does not have an interface amenable for high-throughput data analysis. Although additional computational tools have been developed for processing AMDIS results and to perform normalisations and statistical analysis of metabolomics data, there is not yet a single free software or package able to reliably identify and quantify metabolites analysed by GC-MS.

Results: Here we introduce a new algorithm, PScore, able to score peaks according to their likelihood of representing metabolites defined in a mass spectral library. We implemented PScore in a R package called MetaBox and evaluated the applicability and potential of MetaBox by comparing its performance against AMDIS results when analysing volatile organic compounds (VOC) from standard mixtures of metabolites and from female and male mice faecal samples. MetaBox reported lower percentages of false positives and false negatives, and was able to report a higher number of potential biomarkers associated to the metabolism of female and male mice.

Conclusions: Identification and quantification of metabolites is among the most critical and time-consuming steps in GC-MS metabolome analysis. Here we present an algorithm implemented in a R package, which allows users to construct flexible pipelines and analyse metabolomics data in a high-throughput manner.

Keywords: Metabolomics, Identification, GC-MS, Data analysis

Background

Metabolomics, the popular modern approach to screening large numbers of low molecular mass compounds in biological samples, has been successfully applied in drug discovery [1], food science [2] and systems biology [3] studies. The three most commonly used analytical platforms for the identification and quantification of metabolites in biological samples are perhaps gas chromatography-mass spectrometry (GC-MS), nuclear magnetic resonance (NMR) and liquid chromatography-mass spectrometry (LC-MS) [4]. While none of these is stand-alone in the sense that it provides complete coverage of a sample's metabolome, GC-MS is among the most widely applied because of its ability to separate complex mixtures of metabolites with high efficiency and at low cost [5].

The Automated Mass Spectral Deconvolution System (AMDIS) is the most popular freeware available for metabolite identification and quantification in biological samples analysed by GC-MS [6]. Originally developed for the identification of chemical weapons and related compounds in complex chemical mixtures [7], it is now used in environmental chemistry [8] and metabolomics studies [9]. AMDIS is linked to the NIST standard reference database: one of the most popular mass spectral databases for metabolite identification.

While AMDIS performs well in the identification and quantification of target metabolites within a single biological sample, it does not, in general, use a common reference ion mass fragment (IMF) to quantify the same metabolite across different samples [6]. This limits

*Correspondence: ragg005@aucklanduni.ac.nz
[1]The University of Auckland, 3A Symonds Street, 1142 Auckland, New Zealand
[2]Department of Gastroenterology, Institute of Translational Medicine, University of Liverpool, Nuffield Building, Crown Street, L693BX Liverpool, UK

the reproducibility of the intensity data generated by AMDIS and, therefore, its direct utility for comparative metabolomics studies. Such data may, for example, lead to erroneous identification of chemical signatures (i.e. biomarkers) and, potentially, to the misinterpretation of the activity of metabolic pathways. AMDIS is also known to yield a high rate of false identifications of metabolites, referred to simply as the *false positive rate* [10]. Furthermore, AMDIS reports different results according to the zoom level applied to the chromatogram under analysis. Some compounds are only correctly identified when a smaller portion of the chromatogram is analysed. Finally, the layout of metabolomics data preprocessed by AMDIS is such that it requires further manipulation before it is amenable to subsequent processing and analysis [11]. The necessary manual curation of AMDIS-generated datasets can, therefore, potentially require months to complete.

Recent years have seen exponential growth in the number of metabolomics studies. At the same time, spectral libraries have themselves continued to grow in size, thereby enabling an ever-increasing number of target metabolites to be identified within individual GC-MS-analysed samples. Additionally, high impact scientific journals have raised their standards with respect to the validation of results from metabolomics studies, requiring higher numbers of samples and technical replicates. The net result has been an explosion in the amount of GC-MS-generated data [4], making manual curation post-processing by AMDIS impracticable. An algorithm which more reliably identifies and quantifies metabolites analysed by GC-MS and which is implemented in a software package that reports results in a format that facilitates further data processing without manual intervention is urgently needed.

Numerous programs and software packages to automate processes for the analysis of metabolomics data have become available in the last couple of years. These tools enable quick data normalisation, statistical analysis and the production of graphs for data visualisation [6,12]. Among them is web-based XCMS Online ([13]; https://xcmsonline.scripps.edu/). It is widely used for the comparative analysis (i.e. comparisons between pairs of experimental conditions) of the abundances of *unidentified* IMFs in raw GC-MS data. While XCMS Online enables the identification of metabolites present at significantly different levels across experimental conditions, it is important to note that this involves manual processing. Thus, although XCMS Online can be particularly useful when searching for potential biomarkers, it does not fit the requirements of high-throughput identification and quantification of GC-MS data. Consequently, despite AMDIS's limitations, it remains the most popular software for the identification and quantification of metabolites in raw GC-MS metabolomics datasets.

We introduce here a new algorithm, PScore, which we have developed for the identification and quantification of metabolites in biological samples analysed by GC-MS. PScore scores the metabolites contained in a pre-defined spectral library according to their likelihood of being associated with a specific chromatographic peak; the higher the score, the greater the similarity between the expected (i.e. defined in the spectral library) and observed spectra and RTs (i.e. measured in the biological sample). For a given metabolite: (1) the closer its fragments' detected peaks are to its expected RT, (2) the more closely its fragments' relative intensities follow those defined in the spectral library, and (3) the higher the correlation between the intensities of its fragments, the higher its score. PScore enables the use of threshold scores based on the certainty requirements of each metabolomics experiment, with higher threshold scores resulting in greater precision in compound identification.

PScore is implemented in our new R package, MetaBox, which generates an integrated list of identified metabolites and their corresponding intensities from replicate samples analysed by GC-MS. MetaBox includes functions for removing specific ion mass fragments from GC-MS files and for the generation of graphical outputs. The reports generated by MetaBox can be directly applied to other tools, such as MetaboAnalyst [12] and the R package Metab [6], in order to perform further data processing and statistical analyses. In addition, MetaBox accepts spectral libraries built using AMDIS, including the original formats in which they were generated. Furthermore, MetaBox's use of pop-up dialog boxes makes it more accessible to novice R users. Finally, being an R package, MetaBox is open-source, allowing users to adapt it to their own pipelines for data analysis.

We validated the results produced by PScore through MetaBox via a two-step approach. First, we compared its performance against AMDIS's when identifying and quantifying volatile organic compounds (VOCs) present in standard mixtures of metabolites. MetaBox yielded a smaller proportion of misidentifications and higher accuracy in quantification. Second, we used XCMS Online to generate reference datasets for comparing MetaBox's performance against AMDIS's when identifying compounds present at different levels in faecal samples from female and male mices. MetaBox yielded a higher percentage of metabolites matching XCMS Online's results.

Implementation

PScore: The algorithm

PScore is a GC-MS-based retention time (RT) scoring algorithm used to assess the likelihood that the observed RTs in a biological sample correspond to known metabolites within a user-defined spectral library.

Metabolite identification and quantification by GC-MS

GC-MS instruments usually generate a single file per biological sample, each file containing a list of mass spectra together with their corresponding RTs. These spectra are commonly shown on a chromatogram represented by RT on the horizontal axis and signal intensity on the vertical axis. Peaks in intensity on the chromatogram correspond to putative metabolites in the analysed sample. PScore performs metabolite identification based on a spectral library containing the RT and fragmentation patterns of potential target metabolites.

Spectral library requirements

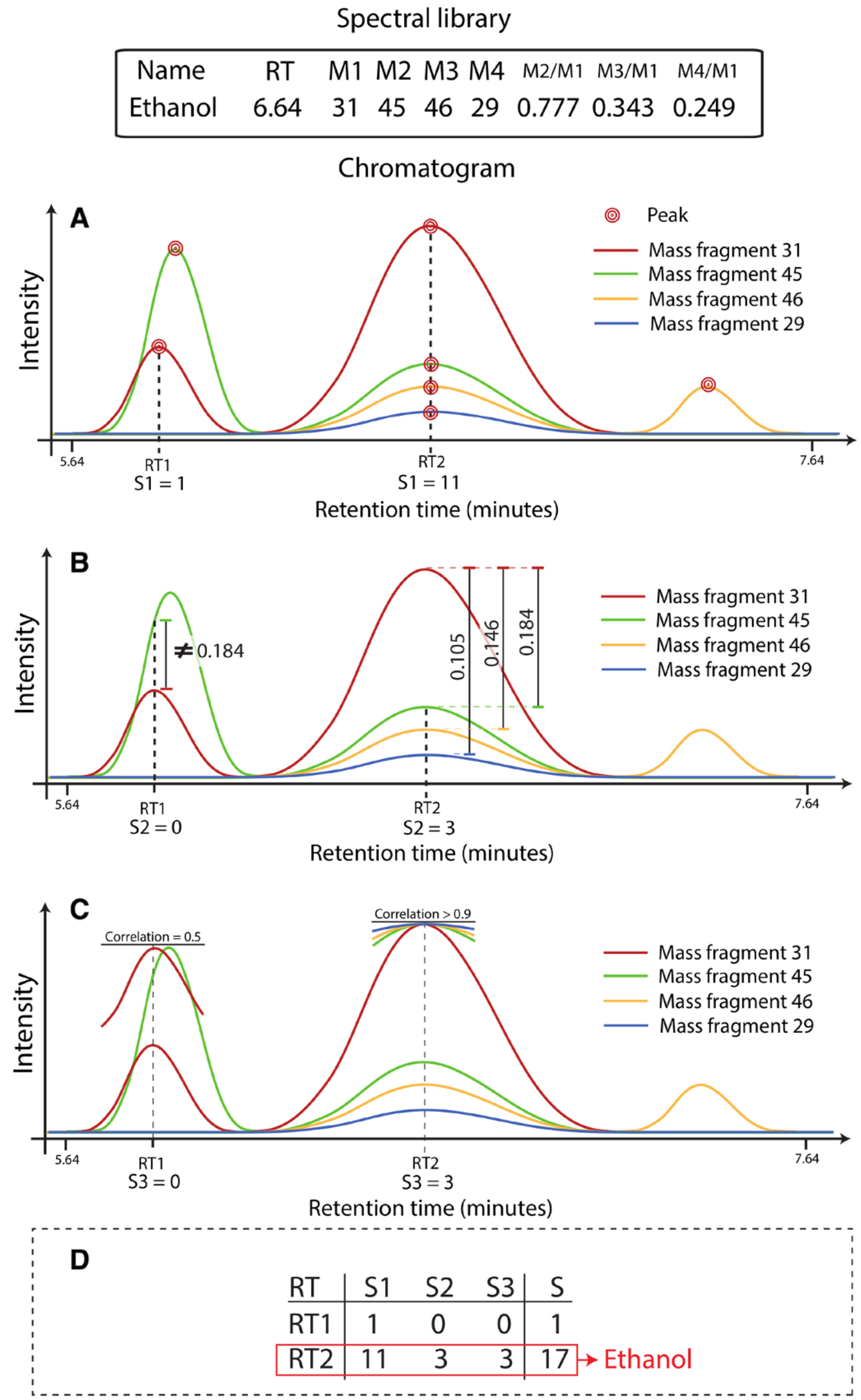
Metabolite identification and quantification require a spectral library containing reference information against which observed spectra can be compared. PScore requires that for each metabolite, M say, in a spectral library, L say, information is included about its expected retention time, E_{RT}, and typically its four most abundant IMFs' mass-to-charge (m/z) ratios, which we will denote by M_i ($i = 1, 2, 3, 4$). Additionally, PScore requires that L contains the intensity ratios $R_i = I_{i'}/I_1$ ($i' = 2, 3, 4$), where $I_{i'}$ denotes the expected intensity of IMF $M_{i'}$, i.e. R_i is the intensity of $M_{i'}$ relative to that of M_1. We will refer to relative intensities simply as intensity *ratios*. For example, consider the first row of the spectral library shown in Table 1, corresponding to the compound ethanol. It has an expected retention time of 6.64 minutes; its four most abundant IMFs have m/z ratios of 31, 45, 46 and 29; the intensities of the last three of these IMFs, relative to the first, are 0.777, 0.343 and 0.249, respectively.

Many algorithms applied for identifying metabolites analysed by GC-MS, such as AMDIS and X-Rank [14], for example, make use of more than 4 ion mass fragments, if available, when calculating the similarity between two mass spectra. Our experience analysing GC-MS data suggests that the 4 most abundant ion mass fragments and the RT are generally the key factors defining the identity of an analyte. For many compounds, the remaining fragments are generally close to or at the noise level, which increases their variability across samples and may reduce the accuracy in identification. In addition, in the way PScore was developed, every additional fragment to be analysed requires additional computer power, which may considerably increase the analysis' time. Compounds showing less than 4 fragments in their spectra may have the existent fragments recycled. For example, a compound X containing only the fragments 58 and 106 in their spectra would have these fragments analysed twice by PScore. In this case, the row of the ion library defining compound X would have its most abundant fragment defined as M1 and M3 in the ion library and the second most abundant fragment defined as M2 and M4.

In the remainder of this section we describe *PScore*, a peak scoring method which utilises the information available within a single GC-MS sample to score observed peaks occurring within a range of RTs and that are potentially associated with a metabolite, M, in the spectral library, L. The highest scoring peak is inferred as belonging to M. We describe the PScore algorithm according to the four stages shown in Figure 1.

Table 1 This table shows an example of the mass spectral library required by Pscore, which contains each standard compound's name (Compound), its expected RT (E_{RT}) in minutes, the m/z ratio of its four (generally) most IMFs ($M_1, M_2, M_3 \text{and} M_4$) and the relative intensities, R'_i, of each M'_i ($i' = 2, 3, 4$) to that of M_1

		IMF m/z ratio				Intensity relative to M_1		
Compound	E_{RT}	M_1	M_2	M_3	M_4	R_2	R_3	R_4
Ethanol	6.64	31	45	46	29	0.777	0.343	0.249
Acetone	7.37	43	58	42	39	0.262	0.076	0.044
Isopropyl alcohol	7.58	45	41	27	39	0.107	0.090	0.072
Acetonitril	7.90	41	40	39	38	0.546	0.223	0.137
Ethyl acetate	10.59	43	45	70	61	0.137	0.116	0.105
1-butanol	13.38	56	41	43	31	0.720	0.543	0.346
2-pentanone	13.95	43	86	41	71	0.249	0.127	0.109
Pyridine	16.42	79	52	51	50	0.564	0.275	0.205
1,2-dimethylbenzene	20.39	91	106	77	51	0.327	0.080	0.077
1,3-dimethylbenzene	20.69	91	106	105	77	0.533	0.223	0.115
1,4-dimethylbenzene	21.80	91	106	105	77	0.488	0.189	0.109
Benzaldehyde	25.71	106	105	77	51	0.990	0.935	0.404
Indole	38.63	117	90	89	63	0.414	0.313	0.103



Spectral library

Name	RT	M1	M2	M3	M4	M2/M1	M3/M1	M4/M1
Ethanol	6.64	31	45	46	29	0.777	0.343	0.249

Chromatogram

RT	S1	S2	S3	S
RT1	1	0	0	1
RT2	11	3	3	17

Figure 1 PScore - algorithm. PScore searches a GC-MS file for metabolites contained in a defined mass spectral library. It analyses a region of the chromatogram searching for chromatographic peaks representing a metabolite and scores retention times (RT) potentially representing a metabolite if: **(A)** peaks of the IMFs expected to originate from this specific metabolite are present at the same RT and if their intensities are equal to the highest intensity observed for each IMF; if **(B)** these IMFs are detected at the expected proportions defined in the mass spectral library; and **(C)** if the intensities of these IMFs show positive correlation. Finally, **(D)** PScore calculates the final score associated to each potential RT, it assigns the metabolite searched to the RT showing the highest score and registers the intensity of the most abundant mass fragment associated with this metabolite.

Stage 1: Scoring peaks associated with IMFs M_1 – M_4

When a metabolite elutes from the gas chromatography column and enters the mass spectrometer, it is bombarded by electrons and fragmented into ionised components, or IMFs. In theory, the IMFs from the parent metabolite, M, should almost simultaneously reach the mass spectrometer's detector, where their intensities and RTs are recorded. This information is commonly used to build both their individual chromatograms and their cumulative or *total ion chromatogram*. Ideal process would result in entire complement of IMFs yielding a set of overlapping peaks centered precisely on a single expected RT. In practice, however, RT shifts may be observed depending on the type of sample being analysed and the variability across GC-MS runs. Consequently, a metabolite's IMF peaks may occur in the vicinity of, but not precisely at, its expected RT. Thus, a search must be conducted across a window of RTs spanning the region of the chromatogram which most plausibly contains the IMF peaks corresponding to the metabolite.

Consider a metabolite M in spectral library L with expected retention time E_{RT}. We define a RT window $\mathscr{W} = E_{RT} \pm w$, with the window parameter, w, being user-defined. The region $\mathscr{W}$ is searched for groups of peaks potentially corresponding to IMFs $M_1, \ldots, M_4$ belonging to M. The jth group's observed peak intensities are recorded as $\mathscr{P}_j = (\hat{I}_{1j}, \hat{I}_{2j}, \hat{I}_{3j}, \hat{I}_{4j}; t_j)$, where $\hat{I}_{ij}$ is the observed intensity of IMF M_i and t_j is the RT at which M_1's peak is observed. Letting $\hat{I}_{\max} = \max\{\hat{I}_{ij}\}$, each observed intensity, $\hat{I}_{ij}$, in $\mathscr{P}_j$ is scored according to

$$S_{1j}\left(M_i|\mathscr{P}_j\right) = \begin{cases} 3, \text{ if } \hat{I}_{ij} \text{ occurs at time } t_j \pm 1s \text{ and } \hat{I}_{ij} = \hat{I}_{\max} \\ 2, \text{ if } \hat{I}_{ij} \text{ occurs at time } t_j \pm 1s \text{ but } \hat{I}_{ij} < \hat{I}_{\max} \\ 1, \text{ if } 0 < \hat{I}_{ij} < \hat{I}_{\max} \text{ but does not occur at } t_j \pm 1s \\ 0, \text{ otherwise} \end{cases}.$$

The total score for $\mathscr{P}_j$ is the sum over the scores assigned to each of its IMFs, i.e.

$$T_{1j} = \sum_{i=1}^{4} T_{1j}\left(M_i|\mathscr{P}_j\right),$$

allowing a maximum possible score of 12.

Stage 2: Similarity scoring of theoretical and observed spectra

If metabolite M is present in a GC-MS-analysed sample, not only do we expect a group of peaks to be observed at its expected RT, we also expect its observed intensity ratios to be identical to their corresponding theoretical values in L. However, due to variability across GC-MS runs and the possible convolution of metabolites, the values of the observed and theoretical ratios may differ from one another. Thus, at Stage 2 we compute the intensity ratios $\mathscr{R}_j = (\hat{R}_{2j}, \hat{R}_{3j}, \hat{R}_{4j})$ from the jth group's observed peak intensities, $\mathscr{P}_j$, where $\hat{R}_{i'j} = \hat{I}_{i'j}/\hat{I}_{1j}$ $(i' = 2, 3, 4)$. It follows that if the observed intensities in $\mathscr{P}_j$ are from metabolite M then we expect $\hat{R}_{i'j} = R_{i'}$ or, equivalently, $\hat{R}_{i'j}/R_{i'} = 1$.

We make allowance for variability between observed and theoretical intensity ratios by introducing a *match factor* f $(0 < f < 1)$ which we use to construct intervals around each theoretical ratio, $R_{i'}$, associated with metabolite M. The lower and upper limits of this interval are given by $L_{i'} = fR_{i'}$ and $U_{i'} = (2-f)R_{i'}$, respectively, with the value of f chosen to yield sufficiently narrow intervals such that only observed peaks from a group of IMFs corresponding to M will lie within them. To reflect this, we give each observed ratio $\hat{R}_{i'j}$ a score of 1 if it falls within its match factor interval $[L_{i'}, U_{i'}]$. The total score for $\mathscr{R}_j$ is given by the sum over all of its ratios' scores, i.e.

$$T_{2j} = \sum_{i'=2}^{4} \mathbf{1}_{\left\{\hat{R}_{i'j} \in [L_{i'}, U_{i'}]\right\}},$$

where

$$\mathbf{1}_{\left\{\hat{R}_{i'j} \in [L_{i'}, U_{i'}]\right\}} \begin{cases} 1, \text{ if } \hat{R}_{i'j} \in [L_{i'}, U_{i'}] \\ 0, \text{ otherwise} \end{cases}.$$

allowing a maximum possible score of 3.

Stage 3: Scoring the correlation between IMFs' intensities

The ion chromatogram of each IMF originated from a single compound is expected to form an approximately bell-shaped curve over a range of RTs $t_j \pm \Delta$, where Δ is chosen to capture the non-zero intensities with magnitudes that are dependent on RT. We represent this by expressing the intensity of IMF M_i of M $(i > 1)$ as a function of retention time t, i.e. $\hat{I}_{ij}(t)$. If the IMFs corresponding to the intensities in $\mathscr{P}_j$ are perfectly aligned, then theoretically their intensity ratios would be expected to be constant across $t \in t_j \pm \Delta$, i.e. $r_{ij}(t) = \hat{I}_{ij}(t)/\hat{I}_{1j}(t) = c_{ij}$, where c_{ij} denotes the proportionality constant in the linear relationship between $\hat{I}_{ij}$ and $\hat{I}_{1j}$ and independent of RT. In other words, IMFs originating from the same compound are expected to have highly correlated intensities, as they are expected to increase and decrease at the same time.

At stage 3 we compute the correlation between the intensities $\hat{I}_i$ and I_1, of M_1, $(i = 2, 3, 4)$, across the retention time window $t_j \pm \Delta$, denoted by $\rho_{i1|t_j}$ which is calculated using Pearson's correlation coefficient. In our experience, the optimal neighborhood of t_j is $\Delta = 0.07$. Ideally, $\rho_{i1|t_j} = 1$. However, this is not always the case. Metabolite coelution, for example, may affect the correlation between IMFs' intensities. Thus, we define a correlation threshold, ct, such that $0 < ct < 1$. We then give metabolite M a score of 1 for each of its observed

IMFs at t_j which have $\rho_{i1|t_j} \geq ct$; that is, the value of the Pearson's correlation is greater or equal to the correlation threshold ct. The Stage 3 score function is then given by

$$S_{3j} = \sum_{i=2}^{4} k_{\{\rho_{i1|t_j\pm\Delta}|ct\}},$$

where

$$k_{\left\{\rho_{i1|t_j\pm\Delta}|ct\right\}} = \begin{cases} 1 \text{ if } \rho_{i1|t_j\pm\Delta} \geq ct \\ 0 \text{ otherwise} \end{cases}.$$

Metabolites found at similar RTs, e.g. $RT_{Ma} - RT_{Mb} \leq |0.1|$ where RT_{Ma} is the RT of metabolite a and RT_{Mb} is the RT of metabolite b, and sharing IMFs, e.g. $Ma_{M_1} = Mb_{M_1}$ where Ma_{M_1} is the m/z of IMF M_1 originated from metabolite Ma and Mb_{M_1} is the m/z of IMF M_1 originated from metabolite Mb, may have lower $\rho_{i1|t_j}$ and, potentially, lower scoring at stage 3. Three pairwise correlations are scored in Stage 3, which allows a maximum possible score of $S_{3j} = 3$.

Stage 4 - Defining the RT and the abundance of metabolite M

We calculate the score S_M of metabolite M at time t_j by

$$S_{M_{(t_j)}} = S_{1_{t_j}} + S_{2_{t_j}} + S_{3_{t_j}}.$$

Then, we obtain the intensity of M_1 at the t_j associated with the highest score, $S_{M_{(t_j)}}$, and with the lowest difference to the expected RT, E_{RT}. This intensity represents the abundance of M_1.

Stages 1, 2, 3 and 4 are performed for every metabolite M in library L. After all metabolites in L are analysed, it may happen that different metabolites were associated to the same time t_j. In these cases, we select for each time t_j only the metabolite showing the highest score $S_{M_{(t_j)}}$ and the lowest difference between time t_j and the E_{RT}.

Implementing PScore in MetaBox

We have implemented our PScore algorithm in an R package named MetaBox. For each GC-MS sample, it generates a list of metabolites, M, with their respective abundances, $P_{M(j)}$, their unique RT, t_j, at which they were identified and their calculated score $S_{M_{(t_j)}}$. MetaBox then merges the results of individual GC-MS samples into a single R data frame called *Total* using metabolite's names as reference (Additional file 1: Table S1). Optionally, the data frame Total can be exported to a csv file.

Ideally, $S_{M_{(t_j)}} = 18$ when metabolite M is actually present in the analysed sample. However, it is not always the case. A specific compound's spectrum may vary slightly from sample to sample as a result of GC-MS variation, matrix effect and metabolite coelution. Therefore, we define a score threshold s_t, such that $8 \leq s_t \leq 18$. MetaBox then selects metabolites that have a calculated score $S_{M_{(t_j)}} \geq s_t$ and stores them in a second R data frame called *cutOff*, containing the name of each metabolite in the first column and their respective abundances in each GC-MS sample in the following columns (Additional file 1: Table S2). Optionally, the data frame cutOff can be exported to a csv file.

The RT index is an excellent system for obtaining reproducible results within and across labs. It is currently implemented in AMDIS and other tools such as TagFinder [15]. However, PScore was initially developed to use only the RT. The possibility to use the RT index will most probably be implemented in a further version of MetaBox.

Validation

As we implemented PScore in the R package MetaBox, we compared MetaBox's performance against AMDIS's in identifying and quantifying VOCs present in standard mixtures of metabolites and in faecal pellets of female and male mice.

Methods

Standard mixtures

A single standard mixture containing 13 metabolites (Table 1) was prepared and divided into 10 aliquots: 5 aliquots of 50 μL and 5 aliquots of 100 μL. Each 50 μL aliquot was diluted by adding 50 μL of water, resulting in a final volume of 100 μL. Each aliquot was then warmed in an incubator oven at 60°C for 30 minutes, then VOCs were adsorbed onto a solid phase microextraction fiber CAR-PDMS 85 μm (Sigma-Aldrich) for 20 minutes and analysed by a Perkin Elmer (Clarus-500) GC-MS using solvent delay, 6 min; temperature program (40°C), 1 min; ramp of 5°C/min to 220°C; finally held at 220°C for 4 min (total run time 41 min). The MS was operated in EI positive mode scanning mass ions in the range 10 to 300 (6–41 min). Room and lab air were used as controls.

Metabolite identification

Metabolites were identified using a mass spectral library built using AMDIS and NIST (Version 2.0) (Table 1) (NB. The library used by AMDIS contains additional ions than shown in Table 1). We first characterised algorithm performance on a per-sample basis, calculating the percentage of false positive and false negative metabolite identifications, defining the percentage of false positives as $100p_i^+\%$, where p_i^+ is the proportion of misidentified metabolites (in relation to the total number of identified compounds) in the ith standard sample, and the percentage of false negatives as $100p_i^-\%$, where p_i^- is the proportion of unidentified metabolites in the ith standard sample. For example, consider the standard sample described above containing 13 metabolites. If an algorithm identifies

100 metabolites, including 10 of which are in the standard sample, it is reported as having 23.1% of false negatives (i.e. 100 × 3/13) and 90% of false positives (i.e. 100 × 90/100).

High percentages of both false positives and false negatives may lead to erroneous inferences being drawn from the data. Optimal metabolite identification tool is one which yields the smallest percentages of both false positives and false negatives. We evaluate the performances of AMDIS and MetaBox over all $n = 10$ with these criteria in mind.

The match factor used by AMDIS may affect the number of false negatives and positives reported. Therefore, AMDIS was applied using the match factor values of 70, 80 and 90. MetaBox was applied using match factor of 70, correlation of 0.95 and score cut of 13.

Metabolite quantification

All aliquots from the standard mixture were analysed by both AMDIS and MetaBox. For AMDIS, its 'Base Peak' values were reported for the metabolite intensities. A reference dataset (Reference), containing the intensity of each metabolite's most abundant IMF, was manually obtained for each sample using the R package XCMS [16]. The abundances reported by MetaBox, AMDIS and Reference for each metabolite are expected to be very similar. We confirmed this by performing a hierarchical cluster analysis (HCA) and a principal component analysis (PCA) on the combined datasets.

Mice samples

Five female and five male five-week old inbred wild-type C57BL/6 mice were purchased from Charles River Laboratories (Margate, UK) and acclimated to standard animal house conditions at the University of Liverpool for a minimum of 1 week. The mice were individually housed for a total of 8 weeks, when one ten-pellet faecal sample was taken from a clean cage. Mice were then sacrificed under Schedule 1 Animals Act 1986. Mice were used in accordance with local ethics approved from the University of Liverpool. Each ($n = 10$; Female = 5; Male = 5) ten-pellet sample was then analysed by GC-MS using the same configuration described in *Standard mixtures*. The mice samples were analysed using AMDIS and MetaBox, using a mass spectral library built using AMDIS and NIST database (Version 2.0) (Additional file 1: Table S3). In order to remove potential false positives, we only analysed those metabolites present in at least 2 samples per experimental condition (i.e. Female and Male).

It is difficult to generate a reference or control when analysing mice samples, as the identity and concentrations of metabolites in these samples are unknown. Therefore, we applied an approach used for biomarker discovery [16]. We used XCMS Online to generate a reference dataset containing the list of IMFs present at significantly different levels between female and male samples (Welch t-test; p-value < 0.05), including the RT where the peak of each IMF is detected. Then, we used our spectral library (Additional file 1: Table S3), which contains the expected RT and the IMFs of each metabolite, to identify the IMFs reported by XCMS Online. We then conducted a Welch's t-test on the AMDIS and MetaBox datasets comparing males and females for each listed metabolite and compared these algorithms' performances against the t-test results from XCMS Online. For clarity, compounds found at significantly different levels between female and male mice samples will be called as biomarkers. (NB. All chromatograms were left untreated and no data normalisations were applied to metabolite abundances.)

The CAS numbers of all metabolites used in this study are available in Table S7 of the Additional file 1.

Results and discussion

Standard mixtures

For clarity, aliquots of 50 μL of standard mixture + 50 μL of water will be described simply as 50 μL samples, while aliquots of 100 μL will be described as 100 μL samples.

Metabolite identification

To enable the comparison of AMDIS's and MetaBox's efficacies in metabolite identification, we calculated the percentages of false positives and false negatives reported by each algorithm when analysing 10 samples of a standard mixture of metabolites (i.e. 5 samples of 50 μL and 5 of 100 μL), using match factors of $f = 70, 80$ and 90 for AMDIS; and match factor of $f = 70$ and score cut of 13 for MetaBox. Every compound reported by AMDIS was considered in the analysis, including multiple identifications for a single RT. For $f = 70$, AMDIS reported an average ± SE ($n = 10$) of 32.8% ± 1.8% of false positives and an average of 6.9% ± 0.8% of false negatives. $f = 80$ and 90 resulted in 30.3% ± 1.9% and 27.8% ± 1.0% of false positives, respectively, and 6.2% ± 1.0% and 4.6% ± 1.3% of false negatives, respectively (Figure 2). MetaBox performed overwhelming better than AMDIS, reporting no false positives and no false negatives.

Although, AMDIS performed reasonably well in terms of low percentages of false negatives, it was a poor performer with respect to its high reporting of false positives. It may be that AMDIS is actually performing as expected given the primary motivation for its development, single-sample analyses of complex chemical mixtures to identify any signs of potential target compounds or chemical weapons [7]. In this context a low false negative rate is crucial and AMDIS's performance meets this requirement. However, the primary motivation for most metabolomics experiments, is the identification and quantification of the highest possible number of metabolites present in

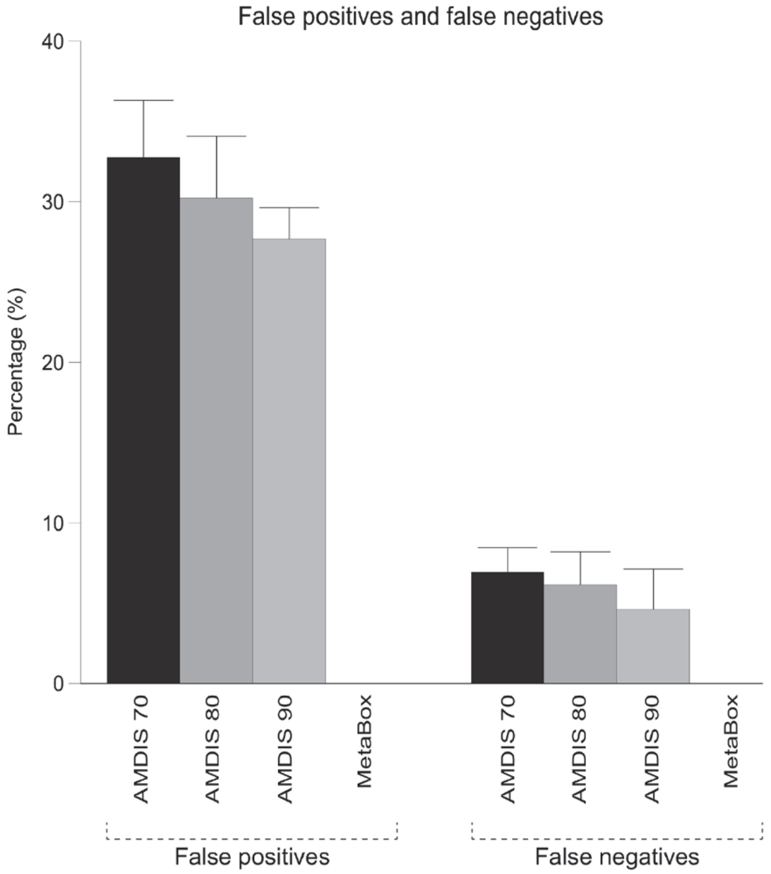


Figure 2 Average percentages of false positives and false negatives. A standard mixture containing 13 metabolites was divided in 10 aliquots and analysed by GC-MS. Each sample was then processed by MetaBox and AMDIS using match factors of 70, 80 and 90. Shown are the average percentages, plus error bars representing two times the standard error, of false positives and false negatives produced by each tool. False positives are compounds that are misidentified, while false negatives are unidentified compounds that are present in the standard mixtures.

biological samples for the comparisons of their abundances, or relative abundances, across experimental conditions. It is non-targeted analysis generally limited only by the metabolites represented in the spectral library. The biological interpretation is then achieved based on the metabolite profile generated by each sample. In this case, the percentages of both false negatives and false positives are crucial for biologically meaningful interpretations of the data. A high percentage of false negatives represents potential losses of biological evidence, while a high percentage of false positives may provide misleading evidences. Therefore, results generated by AMDIS should be manually curated and critically assessed in order to achieve sound biological interpretations.

Metabolite quantification

Average-linkage hierarchical cluster analysis (HCA) (Figure 3A) and principal component analysis (PCA) (Figure 3B) were performed on the metabolite abundances reported by AMDIS and MetaBox (Additional file 1: Table S4). The HCA yielded two main nodes, or clusters: one containing the 50 μL samples and the other the 100 μL samples. Within samples, the MetaBox and reference datasets always clustered together under the same node in the first agglomeration round and this node excluded the corresponding AMDIS dataset. This is indicative of MetaBox-generated abundances being closer in value to those in the reference datasets than the AMDIS-generated ones. The PCA yielded results consistent with those from the HCA, i.e. the 50 μL samples clustered together around negative values of the first principal component (PC 1) while the 100 μL samples clustered around positive values of PC 1. The 50 μL samples varied little in the direction of the second principal component (PC 2), indicating that AMDIS and MetaBox yielded datasets that were similar to one another and to the reference datasets. Samples corresponding to MetaBox-based datasets were always adjacent to the matching reference dataset, showing once again the high degree of agreement between the MetaBox and reference datasets. The 100 μL samples showed separation of datasets in the direction of PC 2. The reference and MetaBox datasets derived

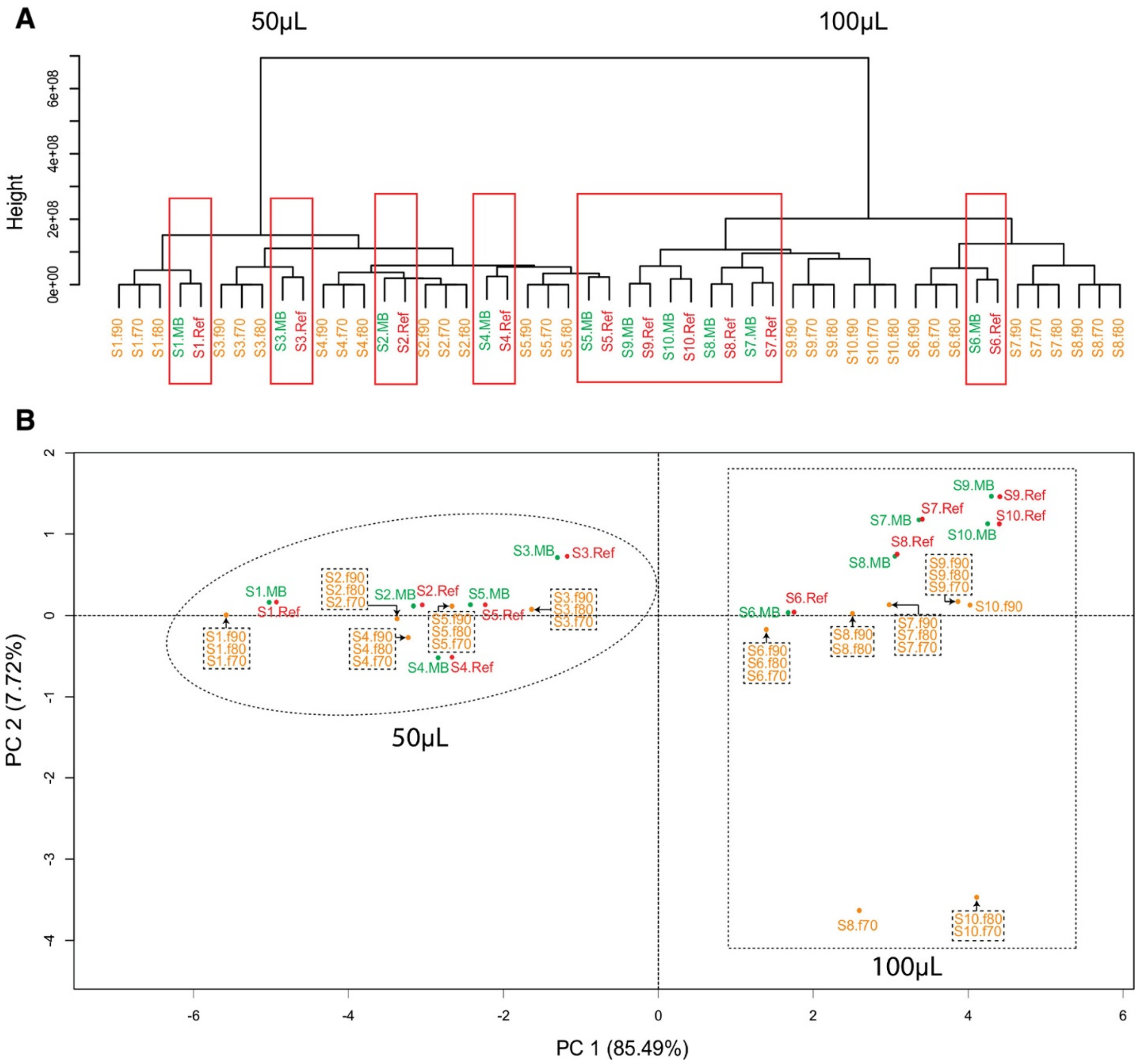


Figure 3 Hierarchical cluster analysis (HCA) and principal component analysis (PCA). (A) Dendrogram from HCA (euclidean distance; average linkage) and **(B)** scatterplot of first two principal components from PCA on data resulting from the application of AMDIS and MetaBox to the raw data from 10 GC-MS-analysed standard mixture samples (5 × 50 μL+50 μL water and 5 × 100 μL aliquots). Reference datasets (Control) were obtained using the R package XCMS. Samples are labeled using a combination of sample number (e.g. S1 = sample 1) and the algorithm applied (MB = MetaBox, Ref = reference, f# = AMDIS using match factor #=70, 80 or 90).

from the same sample consistently yielded approximately equal values for PC 2, once again showing a high degree of similarity between the two sets of data. AMDIS, on the other hand, yielded datasets with PC 2 values less than or equal to zero, demonstrating that only when a high match factor is used will AMDIS yield datasets containing abundances approaching values close to those in the reference datasets.

Part of the dissimilarity between the AMDIS and the reference datasets may be a result of background noise subtraction performed by AMDIS and/or the use of different IMFs when deconvoluting and quantifying the same metabolite across samples. The potential use of different IMFs for metabolite quantification by AMDIS is another indication of its development without a view to comparing the same metabolite across different samples, and yet this is a fundamental concern of metabolomics studies. Further evidence lies in the format it uses for reporting results. AMDIS can generate two types of reports: individual reports or a single report (batch report) for several samples by simply appending results sample-by-sample without actually matching metabolites identified in the different samples. Furthermore, AMDIS reports multiple potential identities associated to a single RT. Consequently, when applied to metabolomics studies, AMDIS's results must be manually cleaned (i.e. the correct hit for each RT must be manually selected), the ion mass fragment used to quantify each metabolite must be manually verified and the results produced for different GC-MS files must be combined in a single table or spreadsheet, and this can be enormously time-consuming depending on the number of samples being processed. MetaBox, however, was developed specially for metabolomics studies. Its results are reported in a single spreadsheet containing the identified metabolites and their respective abundances in every analysed sample, and in the format most commonly required for downstream data normalisation and analysis.

Mice samples

To compare the efficacies of AMDIS and MetaBox in identifying potential biomarkers, we evaluated the datasets generated by each against the XCMS Online reference dataset. XCMS Online reported a total of 387 IMFs (features), from which 73 showed significantly different intensities (Welch *t*-test; *p*-value <0.05) between female and male mice faecal samples (Additional file 1). Based on the IMFs and RTs in the spectral library used by AMDIS and MetaBox (Additional file 1: Table S3), we identified 19 compounds associated to the total list (387) of IMFs reported by XCMS Online. Eleven compounds were associated to 47 of the 73 IMFs reported by XCMS Online at significantly different intensities between female and male samples (Additional file 1: Table S5). However, only 4 of these compounds (Table 2) showed IMFs that were both present at significantly different levels according to XCMS Online results and used by AMDIS and MetaBox for metabolite quantification. Therefore, only these 4 compounds were expected to be found as potential biomarkers by AMDIS and MetaBox. AMDIS and MetaBox were able to identify all 19 compounds associated to the XCMS Online results (Additional file 1: Table S6). For all match factors tested, AMDIS identified 3 potential biomarkers, being only one confirmed by XCMS Online (Additional file 1: Table S5). MetaBox identified 4 potential biomarkers, being two confirmed by XCMS Online (Additional file 1: Table S5). In summary, AMDIS was able to report 1 out of 4 potential biomarkers, while MetaBox reported 2 out of 4. Although MetaBox missed the identification of 2 potential biomarkers, its results represent 100% improvement in relation to AMDIS'.

Table 2 List of compounds identified from XCMS Online results as differentially abundant (based on Welch *t*-test) between GC-MS-analysed female (n = 5) and male (n = 5) mice faecal samples

Compound	MetaBox	AMDIS70	AMDIS80	AMDIS90
Benzene*	0.186	0.123	0.123	0.123
Hexanal	*0.003*	0.203	0.203	0.366
Pentanal	*0.012*	0.146	0.189	NA
Propanoic acid	0.077	*0.038*	*0.038*	*0.038*

* Benzene denotes 1,3-bis(1,1-dimethylethyl)benzene. AMDIS analyses were performed using match factors of 70, 80 and 90. *P*-values < 0.05 in italics indicate differentially abundant metabolites.

Conclusions

Identification and quantification of metabolites is among the most critical and time-consuming steps in GC-MS metabolome analysis. The reliability of the biological inferences that can be drawn from metabolomics studies is directly related to the quality of the data upon which they are based. In addition, as the size and number of metabolomics studies conducted by individual laboratories has grown, the time available to analyse each single dataset has reduced. Therefore, to satisfy the criteria of metabolomics studies ideally software must reliably identify and quantify metabolites, and the results must be reported in a format that facilitates further data analysis. Although AMDIS has been widely used in metabolomics, results show that its performance no longer meets the requirements of modern high-throughput analysis of metabolomics experiments.

We presented here a new algorithm, PScore, which uses a spectral library to analyse GC-MS samples and score retention times according to their probability of representing a metabolite. We implemented PScore in an R package, MetaBox, and compared its performance against AMDIS when analysing standard mixtures of metabolites and mice faecal samples. PScore greatly reduces the percentage of false positives and false negatives, and it considerably improves the quantification of metabolites analysed by GC-MS. In addition, our new R package MetaBox incorporates functions to generate graphical outputs and reports results in a format accepted by other software, such as Metab and MetaboAnalyst, allowing users to perform further data processing and statistical analyses in a high-throughput way. As an R package, MetaBox allows users to construct flexible pipelines for data analysis and allows pop-up dialog boxes, which facilitate its usage by R beginners.

Availability and requirements

Project name: MetaBox
Project home page: http://raphaelaggio.github.io/
Operating system: Platform independent
Programing language: R [17] version 3.0.0 or higher
Other requirements: R packages xcms [16], svDialogs [18], pander [19] and MassSpecWavelet [20]
License: General Public License version 3

Additional files

Additional file 1: Supplementary data. File containing the tables to be used as supplementary data.

Additional file 2: XCMS Online results. File containing the results from the XCMS Online analysis performed on mice samples.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

RA developed the algorithm (PScore), implemented PScore into the R package MetaBox, designed the experiments for validation, analysed the data, created the figures and wrote the manuscript. AM generated the standard mixtures used for validation, assisted on the implementation of PScore into MetaBox and revised the manuscript. SR generated the mice samples used for validation and revised the manuscript. CSJP assisted on the development of PScore and revised the manuscript. KR assisted on the development of PScore, assisted on the implementation of PScore into MetaBox and wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by the School of Biological Sciences at The University of Auckland, and by the Department of Gastroenterology at the University of Liverpool. We would like to thank Paulina Giraldo Perez, Blanca Nubia Perez and Ivan Giraldo Estrada for all the support during the developement of this work.

Received: 19 April 2014 Accepted: 3 November 2014
Published online: 10 December 2014

References

1. Wishart DS: **Applications of metabolomics in drug discovery and development.** *Drugs R D* 2008, **9**(5):307–322.
2. Cevallos-Cevallos JM, Reyes-De-Corcuera JI, Etxeberria E, Danyluk MD, Rodrick GE: **Metabolomic analysis in food science: a review.** *Trends Food Sci Technol* 2009, **20**(11-12):557–566.
3. Feist AM, Thiele I, Palsson BO: *Genome-Scale Reconstruction, Modeling, and Simulation of E. coli's Metabolic Network*. Netherlands: Springer; 2009:149–176.
4. Patti GJ, Yanes O, Siuzdak G: **Innovation: Metabolomics: the apogee of the omics trilogy.** *Nat Rev Mol Cell Biol* 2012, **13**(4):263–269. [<GotoISI>://WOS:000302116200013]
5. Zhang A, Sun H, Wang P, Han Y, Wang X: **Modern analytical techniques in metabolomics analysis.** *Analyst* 2012, **137**(2):293–300. [<GotoISI>://WOS:000297998000001]
6. Aggio R, Villas-Boas SG, Ruggiero K: **Metab: an R package for high-throughput analysis of metabolomics data generated by GC-MS.** *Bioinformatics* 2011, **27**(16):2316–2318. [<GotoISI>://WOS:000293620800026]
7. Stein SE: **An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data.** *J Am Soc Mass Spectrom* 1999, **10**(8):770–781.
8. Furtula V, Derksen G, Colodey A: **Application of automated mass spectrometry deconvolution and identification software for pesticide analysis in surface waters.** *J Environ Sci Health Part B-Pesticides Food Contam Agric Wastes* 2006, **41**(8):1259–1271. [<GotoISI>://WOS:000242363800001]
9. Carneiro S, Villas-Boas SG, Ferreira EC, Rocha I: **Metabolic footprint analysis of recombinant escherichia coli strains during fed-batch fermentations.** *Mol Biosyst* 2011, **7**(3):899–910. [<GotoISI>://WOS:000287367100035]
10. Behrends V, Tredwell GD, Bundy JG: **A software complement to AMDIS for processing GC-MS metabolomic data.** *Anal Biochem* 2011, **415**(2):206–208. [<GotoISI>://WOS:000291904700017]
11. Smart KF, Aggio RBM, Van Houtte JR, Villas-Boas SG: **Analytical platform for metabolome analysis of microbial cells using methyl chloroformate derivatization followed by gas chromatography-mass spectrometry.** *Nat Protoc* 2010, **5**(10):1709–1729. [<GotoISI>://WOS:000282369100011]
12. Xia J, Wishart DS: **Web-based inference of biological patterns, functions and pathways from metabolomic data using MetaboAnalyst.** *Nat Protoc* 2011, **6**(6):743–760. [<GotoISI>://WOS:000291218300003]
13. Tautenhahn R, Patti GJ, Rinehart D, Siuzdak G: **XCMS Online: a web-based platform to process untargeted metabolomic data.** *Anal Chem* 2012, **84**(11):5035–5039. [<GotoISI>://WOS:000304783100056]
14. Mylonas R, Mauron Y, Masselot A, Binz PA, Budin N, Fathi M, Viette V, Hochstrasser DF, Lisacek F: **X-Rank: a robust algorithm for small molecule identification using tandem mass spectrometry.** *Anal Chem* 2009, **81**(18):7604–7610. [<GotoISI>://WOS:000269656700012]
15. Luedemann A, Strassburg K, Erban A, Kopka J: **TagFinder for the quantitative analysis of gas chromatography - mass spectrometry (GC-MS)-based metabolite profiling experiments.** *Bioinformatics* 2008, **24**(5):732–737. [<GotoISI>://WOS:000253746400025]
16. Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G: **XCMS: Processing mass spectrometry data for metabolite profiling using Nonlinear peak alignment, matching, and identification.** *Anal Chem* 2006, **78**(3):779–787.
17. R Core Team: **R: a language and environment for statistical computing.** 2014. [http://www.R-project.org/]
18. Grosjean P: **SciViews-R: A GUI API for R.** 2014. [http://www.sciviews.org/SciViews-R]
19. DarÛczi G: **Pander: an R pandoc writer.** 2013. [http://cran.r-project.org/package=pander]
20. Du P, Kibbe WA, Lin SM: **Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching.** *Bioinformatics* 2006, **22**(17):2059–2065. [<GotoISI>://WOS:000240433100001]