Accepted Manuscript

Structure/reaction directed analysis for LC-MS based untargeted analysis

Miao Yu, Mariola Olkowicz, Janusz Pawliszyn

PII: S0003-2670(18)31304-7

DOI: https://doi.org/10.1016/j.aca.2018.10.062

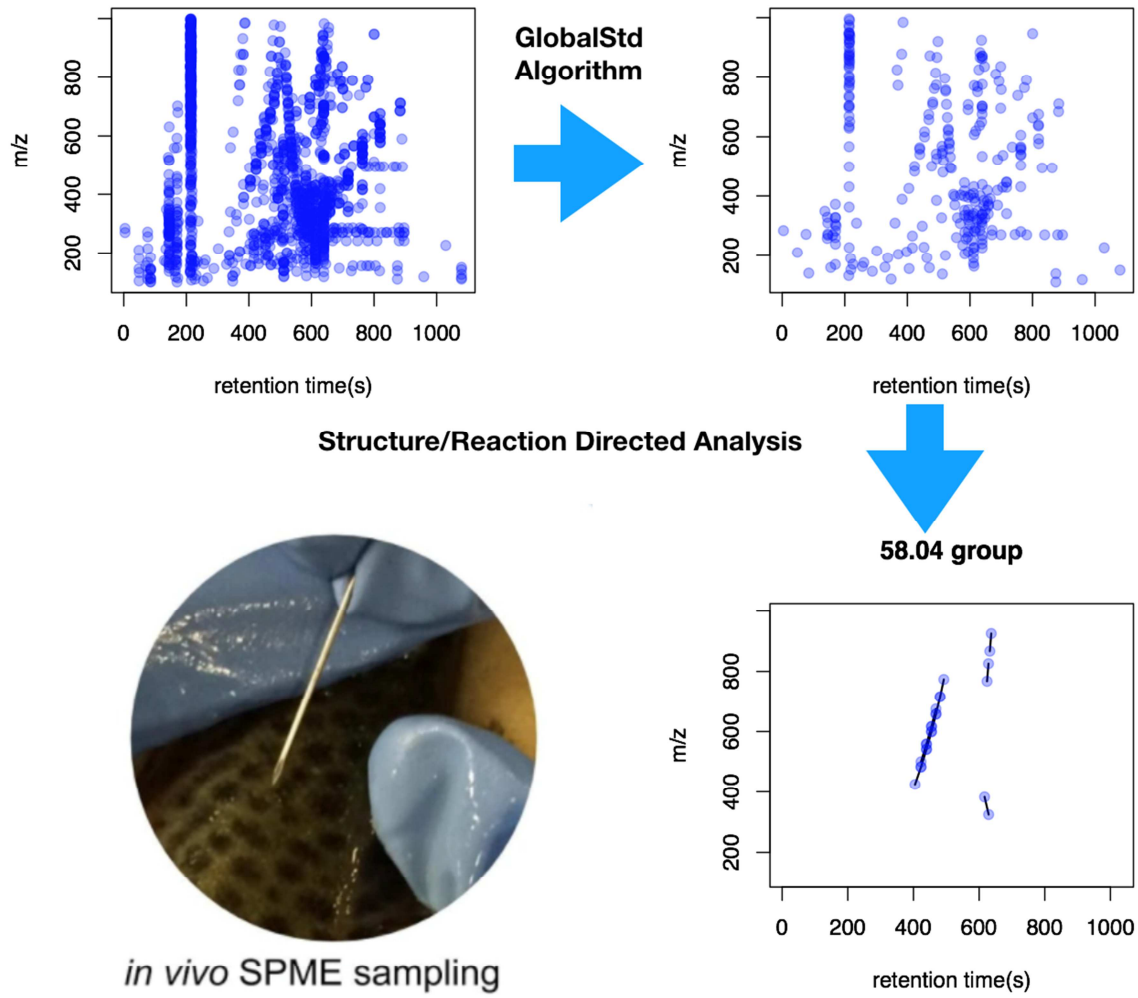
Reference: ACA 236368

To appear in: *Analytica Chimica Acta*

Received Date: 19 August 2018

Accepted Date: 25 October 2018

Please cite this article as: M. Yu, M. Olkowicz, J. Pawliszyn, Structure/reaction directed analysis for LC-MS based untargeted analysis, *Analytica Chimica Acta*, https://doi.org/10.1016/j.aca.2018.10.062.


m/z
200
400
600
800
0
200
400
600
800
1000
retention time(s)
GlobalStd
Algorithm
m/z
200
400
600
800
0
200
400
600
800
1000
retention time(s)
Structure/Reaction Directed Analysis
58.04 group
m/z
200
400
600
800
0
200
400
600
800
1000
retention time(s)
in vivo SPME sampling

Structure/reaction directed analysis for LC-MS based untargeted analysis

Miao Yu[1], Mariola Olkowicz[1] and Janusz Pawliszyn[1*]

[1]Department of Chemistry, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, N2L 3G1, Canada

*Corresponding author: Email: janusz@uwaterloo.ca Phone: +1-519-888-4641. Fax: +1-519-746-0435

Abstracts

In LC-MS based untargeted analysis, data is collected at the peak or ion level, although the investigated biochemistry processes occur at the compound or reaction level. To this end, the presence of redundancy peaks such as co-eluted peaks, multi-chargers, adducts, neutral loss, isotopologues, and fragments ions often muddle subsequent statistical data analysis. In order to fill this gap, between peaks and compounds/reactions, independent components must first be found at the peak level, then evaluated at the compound or reaction levels. Based on paired mass distances (PMD), the algorithm GlobalStd, based on retention time hierarchical cluster analysis and global analysis of PMDs within clusters, is here proposed to extract independent peaks from raw LC-MS data. Following its application, a structure/reaction directed analysis can then be used to evaluate compounds at the structure or biochemistry reaction level, based on similar PMDs among different retention times clusters. As a proof-of-concept, the developed statistical method was applied to data obtained for *in vivo* SPME sampling on fish. In total, 277 independent peaks were demonstrated to stand for most of the variances found for the total 1459 ions detected via LC-MS. Following, both known homologous series or biological reactions along with unknown bio-processes, which may involve oxidation/reduction reactions or homologous series, were analyzed via a structure/reaction directed analysis. The findings of this analysis yielded interesting information regarding the data, for instance denoting the possible occurrence of a biosynthesis process involving L-Carnitine and its precursor 4-Trimethylammoniobutanoic acid. Such PMD relationships could also aid in the screening of annotation results. To this end, semi-quantitative analysis based on structure/reaction directed analysis is also here proposed for further investigation of unknown patterns or for removal of contaminants in metabolomics studies. The developed data-driven algorithm has been included in a PMD package with a GUI interactive document, and is freely available online（https://github.com/yufree/pmd）.

Keywords: metabolomics; LC-MS; *in vivo*; SPME; algorithm

1 Introduction

Untargeted analysis based on Liquid chromatography–mass spectrometry (LC-MS) has been applied in metabolomics,**[1,2]** petroleomics,**[3]** and environmental analysis**[4]** for discoveries of unknown compounds associated with certain biotic or abiotic chemical reactions. In such applications, compounds are usually first analyzed in mass spectrometry as charged ions, while most of the downstream analyses, such as group-wise differences, pathway analysis, and annotations, are performed at the compound level, or between compounds.**[5]** In such cases, charged ions' profiles can be further converted into peaks' profiles via peak detection, using for instance feature detection algorithms such as centWave.**[6]** However, peaks found from untargeted LC-MS analysis at a given retention time are always comprised of a mixture of known and unknown co-eluted peaks, doubly charged ions, adducts, neutral loss, isotopologues, fragments, or molecular ions**[7]**.

The resulting ambiguity between the found peaks and their corresponding compounds would thus affect any subsequent statistical analysis, as peak intensities from any given compound would be in proportion to each other and show a strong linear relation.**[8,9]** For example, in an analysis that yields 1000 peaks with statistical significant differences, such peaks may only correspond to 200 compounds. In such a case, then only twenty percent of the found peaks, containing all pertinent information regarding the compounds of interest, would be necessary for subsequent analysis. Also, peaks stemming from the same compound would yield different sensitivities on LC-MS due to different intensities and signal-to-noise ratios and introduce more uncertainties. Such peaks would also necessitate validation with respect to potential false positive annotations of adducts or

fragmental ions.**[9]** Besides, owing to different ionization processes or their elemental composition, some compounds might only yield one peak, while others may present multiple peaks with different kinds of adducts or isotopologues. This redundancy in peaks would result in statistical bias once analysis such as multiple comparisons**[10]** with false discovery rate (FDR) control are carried out. Similar issues have been discussed with respect to genomics research regarding genes with dependence, as such occurrences have been shown to increase the variance of FDR estimators.**[11–13]**

Targeted qualitative and quantitative analyses usually involve the use of standards,**[14]** where for any given compound, qualitative and quantitative ions are selected based on its standard's mass spectrum and/or retention time, with the provision that selected ions for different compounds are distinctive from each other. Following this rationale, in untargeted analysis, if only a few independent peaks were to be selected among all of the identified peaks to stand for compounds prior to further statistical analysis, then any subsequent statistical analyses could then be performed at the compound level, as is performed in targeted analysis, thus significantly simplifying data analysis and reducing uncertainty. Aiming to find such peaks or remove redundancy peaks, previous studies have attempted to screen mass-to-charge ratios (m/z) with the use of predefined adducts, neutral loss, or chemical contaminants.**[9]** However, unknown adducts, neutral loss, or chemical contaminants may also play an important role in the profiles of untargeted peaks.**[7]** As the importance of such peaks cannot be reasonably predicted and must be evaluated on a case-by-case basis, predefined rules based methods may thus miss peaks from unknown background ions or adducts. Peak shapes**[15]** or peak intensities**[16–18]** have been employed to identify pseudospectra of independent compounds as a means to make further annotations or identifications. However, if

the purpose of research is to elucidate overall changes at the structural or reaction level, then it can be reasonably assumed that one peak from each independent compound would provide enough information.

Supposing that an independent peak for each compound is detected and selected, compound identification would nonetheless still require the execution of additional steps, such as tandem mass spectrometry analysis or database queries.**[16,18]** To this end, specific tools have been developed to allow for predictions based on MS/MS database data, such as the Global Natural Products Social Molecular Networking (GNPS)**[19]** and Metlin.**[20]** Mass defect analysis, on the other hand, can be employed to reveal unknown compounds**[14,21]** and compounds with same sub-structures would show similar mass defect values. In petroleomics or environmental analysis, extensions of the concepts of Kendrick mass defect analysis have been employed to find homologous series**[22]** compounds with different base units, such as -CH_2-, -O-, -CH_2O-, from high resolution mass spectrometry data.**[21–23]** Besides, mass defect values could also be used to filter drug metabolites from high resolution mass spectrometry data, since the metabolites of certain parent compounds would show mass defect values within 50mDa of their parents compounds.**[24]**

We could extend the concept of mass defect to paired mass distance (PMD; the distance between two mass-to-charge ratios), since chemical reactions would also involve unique defect values between reactants. For instance, in environmental analysis, a PMD of 33.96102Da is often used to screen for halogenated contaminants, as this PMD corresponds to a dechlorination reaction that involves an exchange between a hydrogen atom and a chlorine atom (H <-> Cl)**[25]**. Thus, PMD-based identification could be used to identify certain homologous series or

substitution reactions in biotic or abiotic processes. As such, further qualitative or quantitative analysis could forego identifications of all detected compounds, only focusing on compounds that present the same PMDs as a group, since such compounds would have similar structures, or participate in the same chemical reactions. Besides, it can be assumed that if a certain compound is involved in multiple common PMDs, such compound would thus play an important role in the untargeted profile of the found peaks.

However, employment of either methods, namely the identification of independent peaks or PMD-based structure/reaction directed analysis, would necessitate a pre-defined PMD for either adducts, neutral loss, isotopologues, or sub-structures as shown in CSPP algorithm**[18]**. To this end, if a heuristic method could be employed to find unknown PMDs based on the statistical properties of the LC-MS peak profile, both known and unknown compounds belonging to adducts, neutral loss, the same homologous series, or biochemistry reactions could thus be identified. Once the LC-MS data is thus 'filtered', then subsequent semi-qualitative or quantitative statistical analyses could be performed for those compounds as a group, thus bypassing the need for identification of each peak found in the raw LC-MS data. Further, such an approach to analysis might reveal unknown novel mechanisms in untargeted studies, such as oxidation processes or substitution reactions.

Selection of PMDs could include PMDs that correspond to certain structures or reactions by element analysis. For instance, reactions involving Oxygen (15.994915 amu), Phosphorus (30.973763 amu) and Sulfur (31.972072 amu) would yield a PMD value corresponding to less than an integer number, such as PMD 13.98 Da and 15.99 Da. On the other hand, reactions that involve Hydrogen

(1.007825 amu) and Nitrogen (14.003074 amu) atoms would always yield reaction-related PMDs that are larger than an integer. For instance, for analysis of data from samples collected from biological sources, PMD-based analysis could be employed to infer which elements are involved in the specific metabolic reaction under study. Such information would help point at the biochemical processes associated with the studied phenomena without necessitating identification of each detected feature.

For short-lived compounds, identification of reactions or dynamic changes occurring among such compounds' structures would reveal important information regarding their biological or environmental profile. While *in vivo* untargeted studies can aid in the identification of previously unreported compounds,[26,27] qualitative analysis of such "unknown" compounds captured by new analytical methods remains a challenge. To this end, solid phase microextraction (SPME) has been successfully applied towards analysis of *in vivo* biological processes to reveal the presence of previously unreported short-lived compounds, which may have gone undetected in analysis employing traditional sampling methods[26,28]. Thus, *in vivo* SPME is presented as a suitable analytical platform to set up and validate a statistical method for identification of unknown compounds.

In the current study, an algorithm, namely GlobalStd, is proposed to remove redundancy peaks in LC-MS based non-targeted analysis, based on peaks' exact mass and retention times. Following application of the algorithm, the resulting independent peaks can then be submitted to a structure/reaction directed analysis at the compound or reaction level. Such a method is designed to detect both known and unknown compounds, as well as reaction relationship among compounds. As a proof-of-concept, the developed method was employed towards untargeted analysis of fish tissue via *in vivo* SPME.

2 Materials and methods

2.1 Chemicals

LC-MS grade acetonitrile, methanol, and water were purchased from Fisher Scientific (Ottawa, ON, Canada). Hexane and acetone were purchased from Sigma-Aldrich (Oakville, ON, Canada). Biocompatible SPME mixed mode probes (45 µm thickness, 15mm length of coating) were provided by Supelco (Bellefonte, PA, USA). Standards, including diazepam, nordiazepam, oxazepam, flunitrazepam, lorazepam, testosterone, and progesterone were used as instrumental QC samples, and purchased from Sigma-Aldrich (Oakville, ON, Canada).

2.2 *In vivo* SPME sampling

All experimental protocols were approved by and carried out in accordance with guidelines established by the University of Waterloo Animal Care Committee (AUPP #10–17). Rainbow trout (*Oncorhynchus mykiss*) (n=3) were purchased from Silver Creek Aquaculture (Erin, ON Canada). Fish were acclimatized to laboratory conditions for two weeks in non-chlorinated water. All fibers used in *in vivo* sampling were preconditioned in methanol/water (50/50, v/v) prior to use. Three fibers were used to sample each fish, and a total of three fish were sampled.

In vivo sampling of fish muscle tissue was carried out by inserting mixed mode SPME fibers into the dorsal-epaxial muscle (near the dorsal fin) of fish after they were anaesthetized with Tricaine mesylate and affixed to a foam bed. After insertion of fibers, fish were allowed to recover in a bucket for a 20 minute period while *in vivo* extractions were carried out. Once the extraction period was concluded, fibers were pulled out, wiped with Kimwipes, and vortexed at 1500 rpm for 5s with ultrapure water to remove any matrix components on the fiber.

Desorption of fibers was performed with 300 µL of acetonitrile/water (80/20, v/v) as solvent for 90 min at 1,000 rpm vortex agitation. Extract solutions were collected for instrumental analysis.

2.3 Instrumental analysis

An ACQUITY ultra performance liquid chromatography (UPLC) M-Class (UPLC) instrument coupled with a mass spectrometer (Xevo G2-S QT of mass spectrometer equipped with ZSpray TM ESI source) was used for instrumental analysis of samples. Chromatography columns (Kinetex 1.7µm PFP, 100A, 100 x 2.1 mm) were eluted with mobile phase A (water with 0.1% Formic acid) and mobile phase B (Acetonitrile with 0.1% Formic acid) at 80 µL/min. The column temperature was set at 30°C, and the samples were kept at 5°C. The injection volume was 10µL. Gradient elution was as follows: 90% A was run for 1 min, reduced to 10% in the following 7 min, then kept for 4 min. Following, mobile phase A was increased back to 90% within 2 min, then maintained for 4 min until the next injection.

The mass spectrometer was run in positive mode with spray voltage 3000V, cone voltage 40V, and source offset 80V. The source temperature was 120°C and the desolvation temperature was 350°C, with desolvation gas flow at 800L/h. Acquisition mode was set as full scan mode, with a mass range of m/z 100-1000. LockMass acquisition was employed to calibrate the mass spectrum, with a scan time of 0.1s, an interval of 120s, and with Leucine enkephalin used as reference material. Pool QC sample, and instrument QC samples were injected before and after nine samples, blank solvent and blank fiber to assess the stability of the mass spectrum throughout analysis. Such quality control showed a stable performance during the analysis.

2.4 Data processing

Following instrumental analysis, raw data was exported from the instrument and converted into mzxml format for further data analysis. Once optimized parameters were attained via employment of the IPO package[29] on pool QC samples, XCMS[30] was used to extract peaks. The GlobalStd algorithm was then employed in a structure/reaction directed analysis to evaluate the obtained profiles from fish *in vivo* sampling. Metlin was used to tentatively annotate the peaks and obtain chemical names via comparisons of chemical formulas, with an accuracy of less than 5 ppm. Annotation was employed to validate the results of the structure/reaction directed analysis.

2.5 GlobalStd algorithm

STEP 1: Retention time cluster analysis

The algorithm GlobalStd was developed to find independent peaks from peak retention time and mass-to-charge ratio profiles. As shown in scheme 1, the first step of GlobalStd encompasses the aggregation of peak groups based on a retention time hierarchical cluster analysis.[31] Such groups include components separated by chromatography that are relatively independent of each other. Once this analysis is concluded, then PMD analysis can be used to screen potential redundancy peaks.

STEP 2: Paired mass distance (PMD) analysis

Redundancy peaks from same compounds should be discarded for further structure/reaction directed analysis. As shown in work by Mahieu et al.,7 unknown adducts or background ions could be revealed by frequent intrinsic relationship analysis. However, the presence of co-eluted compounds can make such

frequency-based methods exclude unknown redundancy peaks from unrelated compounds. On the other hand, as doubly charged ions would show a PMD around 0.5, PMD analysis can enable the exclusion of these mass pairs from further discussion. To avoid the inclusion of common isotopologues, e.g., peaks with 12C and 13C, mass distance pairs around 1 and 2 would be treated as isotopologues groups,[23] and any additional PMD analysis would only include isotopologues with lower mass-to-charge ratios. As such, ions identified via PMD analysis will not have isotopologues or doubly charged ions among the data carried forward for further analysis.

Following the above discussed steps, further PMD analysis can then performed based on the 'global' properties of the PMDs found in each retention time group. If a specific PMD were to appear multiple times in different retention time groups, then such PMD would be assumed to reflect universal paired relationships, such as adducts or neutral loss. At the same time, most of the fragmented ions, co-eluted compounds, or contaminated ions would be removed for further analysis, as their PMDs are unlikely to appear in multiple retention time groups as compared with adducts or neutral loss. Since only PMDs within the same retention time group are addressed in this step, PMDs between independent compounds would thus not be captured.

STEP 3: Selection of independent peaks

The workflow of Step 3 is illustrated in the right part of Scheme 1. Here, within each retention time group, the remaining peaks are grouped into one of two groupings: one that contains singles peak in the retention time group, and another to encompass multiple peaks in the retention time group. Here, single peaks are kept as independent components. The grouping with multiple peaks, on the other

hand, is further grouped into another two categories. One category is comprised of peaks with isotopologue peaks, while the other is comprised of peaks with 'global' PMDs. For retention time groups that contain multiple groups, only the largest mass-to-charge ratios are selected as potential molecular ions or base peaks.

For retention time groups with isotopologues or 'global' PMDs, we could further divide their peaks into three parts: one with isotopologues peaks and no 'global' paired masses, one with 'global' paired masses and no isotopologues, and one last group, containing both isotopologues and 'global' paired masses. For the first group, since 12C containing isotopologues often show higher intensities than isotopologues with 13C, smaller ions are then kept as independent ions. For the second group, all ions with smaller mass-to-charge ratios in the 'global' paired masses are treated as independent peaks. For the third group, all isotopologues with lower mass-to-charge ratios are first extracted. Then, aiming to remove all isotopologues adducts, the mass distances among the remaining isotopologues are calculated, and only the lower mass isotopologues that appear in the 'global' PMDs are kept. Other isotopologue ions can also be kept as potential independent peaks, even if they are not in the 'Global' PMDs.

Once all these steps are concluded, and most if not all repeated peaks, isotopologues, and adduct-related peaks are removed, the peaks from all of the above groups can be combined together as independent peaks, and carried forward for further analytical analysis. While this step is aimed at removing isotopologues and adduct related peaks, the remaining peaks could still contain some adducts ions if these compounds are only shown as adduct ions.

In summary, the goal of the GloabalStd algorithm is to use a minimum amount of peaks to stand for the significant amount of peaks generally found in untargeted analyses by removing redundancy peaks from the same compounds prior to further analytical analysis. To this end, the presented algorithm requires at least two parameters: the cutoff of the retention time hierarchical cluster analysis, and the bottom threshold number of retention time groups for global PMD searches. For example, a threshold of 10 would mean that the selected PMD should appear in at least 10 different retention time groups. Since we employed cluster analysis, the resolution of the chromatography separation could be controlled by the cutoffs of distances between retention time groups. Such a cutoff should reflect the separation capacity of the employed chromatography columns. Selection of an appropriate bottom threshold number for PMD searches, on the other hand, would ensure that retention time groups for PMD analysis can be determined by explorer analysis of the PMDs profiles so as to include all m/z with known PMDs.

2.6 Structure/reaction directed analysis

PMDs can also be used to group compounds in structure/reaction directed analyses. Here PMDs for peaks in different retention time groups are used instead of PMDs of the same retention time groups, as is the case for the GlobalStd algorithm application. To this end, such PMDs would not indicate adducts or background ions, since those peaks are supposedly coming from different compounds. These PMDs may nonetheless be related to certain homologous series or chemistry reactions. To this end, a frequency cutoff could be set to investigate universal homologous series or chemistry reaction related compounds. The presence of isomers would increase the frequencies of certain mass-to-charge ratios, thus ensuring that only one of each isomer remains in the data carried forward for frequency analysis. Such structure/reaction directed analysis could be performed at

the peak level without employing the GlobalStd algorithm. However, peaks stemming from the same compounds would be cumbered by additional noise in the frequency, as shown in the following section.

To make it clear, GlobalStd algorithm is different from published methods like DeltaMS[32] or MSClust[33]. For DeltaMS, mass distances are used to find isotopologues relationship[32]. However, our methods also used such relationship to find adducts, neutral losses, homologous series or chemistry reactions. MSClust[33] use intensity-based cluster analysis to reduce the peaks into compounds while PMD method only use pared mass distances. As we will show in the demonstrated data, our method could show a similar result compared with intensity-based methods. However, since intensity was not used to find independent peaks, our method is robust for the uncertainty in intensity measurement. Another important difference is that our method doesn't use pre-defined neutral losses, adducts lists, homologous series or reaction. All the findings are based on relationship frequency in the data and only the high frequency paired mass distance relationships are kept for further investigation. Current methods such as mass defect, or could not find unknown reactions or adducts while our methods could reveal them if they show a highly frequency in the peaks profile. As for structure/reaction directed analysis, similar but totally different way has been used to find metabolites for known compounds[34]. However, our method directly uses the frequency of paired mass relationships to screen and reveal both known and unknown structures or reactions.

Both the GlobalStd algorithm and the structure/reaction directed analysis workflow have been included in the PMD package, which is freely available online (https://github.com/yufree/pmd). All the documents for this package could be

found online (https://yufree.github.io/pmd/). A graphical user interface (GUI) to perform the presented PMD-based methods was also included in this software package as interactive documents. Experimental data from *in vivo* SPME sampling are also attached in this package for reproducible research purposes.

3 Results and Discussion

3.1 Retention time groups

A total of 1459 peaks were extracted from *in vivo* SPME samples across 9 samples. As shown in Figure S1, 75 retention time groups were found in *in vivo* SPME datasets. Under the employed chromatography conditions, hydrophilic compounds eluted first, followed by lipophilic compounds. Indeed, some hydrophilic compounds were observed to not retain on the employed column, and to co-elute at the very beginning (see retention time group 6). Some patterns, such as homologous series, could also be observed in the raw data as such compounds eluted sequentially, with an increasing mass-to-charge ratio (m/z). However, the majority of peaks formed what appeared to be a random pattern on the retention time-m/z profile. Peaks within certain retention time groups could be either co-eluted compounds or peaks from same compounds. Hierarchical cluster analysis separates those peaks with a cutoff of 10, which means the complete linkage distances between each retention time group is larger than 10s.

In summary, retention time hierarchical cluster analysis could aid in the search for relatively independent fractions. Following, PMD-based filtering could be applied within each retention time group to further reduce peaks into potential independent peaks.

3.2 PMD analysis

A PMD analysis with cutoff of 10 for the frequency of PMDs between RT cluster for independent peaks of the *in vivo* data indicated 8 retention groups with single peaks. Additionally, 631 isotopologue-related paired mass and 685 multi-charger

related paired mass peaks were found. As shown in figure 1, among the PMDs to appear in more than 10 retention time groups, 10 unique PMDs (which retain 2 digits after the decimal point), involving 431 peaks, were kept out of 443 paired mass peaks. Some PMDs were treated as adducts (such as 21.98Da for adducts between H^+ and Na^+, and 17.03Da for adducts between H^+ and NH_4^+)[35] while some were treated as neutral loss (such as 18.01Da for H_2O).[36] Some polymer-related PMDs, such as PMD 28.03Da (-C_2H_4-) and 44.03Da (-C_2H_4O-) were also found. Unknown adducts/neutral loss such as PMD 45.06Da (-C_2H_7N-) and 66.01Da (-C_4H_2O-or -C_2N_3-) were also identified in the presently discussed data. Conversely, PMD 23.0760Da, shown in Mahieu et. al's work,[7] was not found in this dataset, which means such a PMD may be related to case-by-case unknown background ions. It should be also noted that Mahieu et. al 's analysis directly use global paired mass distances for all mass pairs, while our algorithm only employs the mass distances within each retention time group. Besides, the median Pearson correlation coefficients of the PMD's intensity is 0.88, which implies these paired peaks stem from the same compounds. In summary, PMD analysis within retention time groups could show both known and unknown adducts or background ions from the m/z - retention time profile.

3.3 Independent peaks selection

Application of the GlobalStd algorithm on the data yielded 277 independent ions. As shown in figure S2, ions found by GlobalStd could fit into different scenarios, including groups with lots of co-eluted peaks. Since the developed algorithm only uses m/z and retention times, intensities can thus be further applied to validate the selections. Likewise, Principal component analysis (PCA)[37] can be used to assess changes between score plots of the raw data (containing all peaks) versus that of the selected peaks. As shown in figure 2, PCA similarities would indicate

that the selected peaks, representing around 20% of the original peaks, sufficiently capture variances from all peaks. Considering that this analysis only employs peak mass and retention times, a correlation analysis based on peak intensities could be used as an independent test to further screen peaks.

Further validation can be carried out based on a statistical analysis of peak intensities. As three fish were sampled with three SPME fibers each, no statistical differences should be found among biological replicates. From the raw peaks, 86 peaks out of 1459 peaks showed statistical differences among three fish (F test, p-value cutoff 0.05). With a p-value cutoff of 0.05 for multiple comparisons, 73 (1459*0.05) peaks with significant statistical differences were identified as false positives. From the independent peaks, 17 peaks out of 277 peaks were found to yield statistical differences (F test, p-value cutoff 0.05). Of these, 14 (277*0.05) peaks should be identified as false positives, with p-value cutoff 0.05 for multiple comparisons in independent peaks. After applying a false discovery rate control (q-value cutoff 0.05), no peaks could be identified as true discoveries, in either the raw or selected data. Thus, the statistical analysis would indicate that the algorithm-selected peaks retain information similar to that in the raw data. However, the validation of those peaks are reduced a lot by focused on independent peaks.

In summary, the GlobalStd algorithm can be used to reduce peak numbers with minimal loss of information. As a next step, the peaks selected by the algorithm can then be submitted to structure/reaction directed analysis.

3.4 Structure/reaction directed analysis

277 selected independent peaks were imported for structure directed analysis. Here, only PMDs among different retention time groups were considered for the selected peaks. This setting forced the structure/reaction directed analysis to use peaks which could be separated by chromatography. In total, 19 PMD groups with a frequency larger than 10, as shown in figure 3, were found. All 277 peaks were submitted to Metlin for their chemical formula, with accuracy setting of less than 5 ppm. The settings $[M+H]^+$, $[M+Na]^+$, $[M+NH_4]^+$ and $[M-H_2O+H]^+$ were selected for the database search, according to the PMD analysis above. Potential structure/reactions were then directly investigated via comparisons of chemical formula. As shown in table S1, 119 peaks were involved in those 19 PMD groups. This would indicate tentative identification of a variety of compounds involved in networks of multiple chemical reactions in certain biotic or abiotic systems. For example, $C_{22}H_{41}NO_2$ (m/z 352.3214), which appeared in 7 different PMD groups, was tentatively identified as anandamide, a reported active compound in living systems.[38]

Some of the identified PMD groups highlighted in this analysis have been already associated with known structures or certain bio-processes. For instance, a PMD of 0 indicates isomers, while a PMD of 13.98Da could indicate the exchange of an oxygen atom for two hydrogen atoms, which is associated to an oxidation process followed by H_2O elimination.[35,39] For example, $C_{24}H_{36}O_5$(m/z 405.2616) and $C_{24}H_{38}O_4$(m/z 391.2835)'s ions were tentatively identified with accuracy less than 5 ppm, and a statistically significant intensity correlation (pearson correlation coefficient, 0.8427). Such a relationship might denote the presence of an oxidation process of 3β-Hydroxy-6-oxo-5α-cholan-24-oic acid, according to tentative annotation from Metlin.[20] Likewise, a PMD of 15.99Da might indicate the addition or removal of oxygen atoms, or an oxidation process.[31,35] For example,

$C_7H_{15}NO_2$ (m/z 162.1128) and $C_7H_{15}NO_3$ (m/z 146.1183) could be L-Carnitine and its precursor 4-Trimethylammoniobutanoic acid, as supported by their intensity correlation (pearson correlation coefficient, 0.9519). This biosynthesis process has been reported to occur in humans,[40] and may also similarly occur in fish, since L-Carnitine is also found in fish.[41]

Some of the acquired PMD values could be related to homologous series such as PMD 14.02Da, 28.03Da, and 58.04Da. These in turn could be related to substructures of $-CH_2-$, $-C_2H_4-$ and $-C_3H_6O-$, respectively. Such substructures could be found in fatty acids[31,35] or surfactant[21]. As shown in figure 4, a series of seven compounds, from m/z 425.3120 to m/z 773.5662, and with a PMD of 58.04Da, were identified in the data. The chromatograph also showed a linear elution process, with regular increasing distances. However, a Metlin[20]seach failed to yield corresponding compounds. Previous works[31,35] have treated PMD 58.04Da as acetone condensation, although such a process might not occur between compounds. The identified PMD might also be related to polymers such as Polypropylene Glycol, since their mass spectrum covers the peaks found in our research.[42] While these peaks were not found in fiber control, they may very well be contaminants or unknown compounds.

Some unknown PMDs might need further validation analysis. For example, a PMD 16.03 could be related to a mass difference of one carbon atom and four hydrogen atoms. On the other hand, this PMD may correspond to a combination of removing the substructure $-CH_2-$ and a dehydrogenation reaction. In *in vivo* SPME sampling, such a PMD was found between $C_{24}H_{50}NO_7P$ (m/z 496.3410) and $C_{23}H_{46}NO_7P$ (m/z 480.3100) (pearson intensity correlation coefficient of 0.9456), which might indicate a conversion between two kinds of phosphatidylcholine(PC) ($C_{24}H_{50}NO_7P$

512 <-> $C_{23}H_{46}NO_7P$). Another possibility however, is that both of these two
513 compounds stem from the same parent compounds (unknown parent compounds ->
514 $C_{24}H_{50}NO_7P$ + $C_{23}H_{46}NO_7P$). Both types of reactions would show a high
515 correlation coefficient between the two compounds. If this PMD were to appear
516 with high frequency for certain metabolites across the independent peaks, then it
517 could be reasonably concluded that such reactions are not the result of a random
518 combination of two compounds. While all the annotations made in this work need
519 further validation, such as MS/MS analysis or data-based predictions, some
520 preliminary conclusions can be nonetheless drawn at the chemical formula level
521 for the unknown parts of this non-targeted analysis.
522
523 Compounds from homologous series or similar biochemistry reactions, such as
524 lipids, might show response factors with regularity on mass spectrometry
525 analysis.[43] The average responses from certain mass defect groups could be used
526 for a semi-quantitative evaluation of those unknown homologous or reactions in
527 samples. Figure S3 shows the relative standard deviations (RSD%) of compounds
528 in each group among the three fish. The peaks can be further filtered for certain
529 homologous series or similar biochemistry reactions by assigning a threshold based
530 on the attained RSD%. If a given PMD group shows significant average intensity
531 changes among the two conditions, then this change can be directly used to
532 quantitate certain homologous series or biochemistry reactions, which would allow
533 for a circumvention of the use of standards to validate these compounds. Further,
534 an established linear relationship between paired masses could be also used to filter
535 reasonable peaks for subsequent semi-quantitative structure or reaction analysis.
536
537 To this end, such an analysis was also performed on raw peak data, without prior
538 application of the GlobalStd algorithm. As shown in Figure S4, PMD values

yielded a chaotic distribution, with peaks from same compounds, and much noise in the frequency. While known adducts could be used to filter data in each PMD group, unknown adduct ions, such as PMD 66.01, would still be present in the data. Thus, application of the GlobalStd algorithm would be necessary to remove both known and unknown peaks from the same compounds.

4 Conclusions

The current work proposes a data-driven method to evaluate untargeted data at the compound, homologous series, or biochemistry reaction levels without the use of standards or intensity data. The presented methods could be used to remove redundancy peaks from data profile and to select independent peaks for further structure/reaction directed analysis. As this process is software automated and based on a heuristic search, it enables the unveiling of both known and unknown relationships between the peaks. PMD values can be used to elucidate bio-processes at the reaction level, as well as to aid in more accurate peak annotations based on their oxidation-reduction properties. To this end, the establishment of a database of PMDs and their corresponding homologous and reactions might aid in much easier exploration of “unknown unknown” compounds.

Acknowledgments

We gratefully acknowledge Professor Mark R. Servos and Ms. Leslie Bragg for kindly providing us with fish and the use of a facility for *in vivo* sampling as well as for their assistance in the sampling. We are truly grateful for Ms. Sofia Lendor' help on instrumental analysis of the samples.

Funding

This research was financially supported by the Industrial Research Chair of the National Sciences and Engineering Research Council of Canada (NSERC-IRC). We are also very grateful to Waters Corporation for kindly providing the seed instrument Xevo G2-S QTof to our laboratory. The authors thankfully acknowledge Supelco (Millipore Sigma) for providing the mixed-mode SPME fibers used in this study.

References

[1] A.C. Schrimpe-Rutledge, S.G. Codreanu, S.D. Sherrod, J.A. McLean, Untargeted Metabolomics Strategies-Challenges and Emerging Directions, J. Am. Soc. Mass Spectrom. 27 (2016) 1897–1905.

[2] A. Scalbert, L. Brennan, O. Fiehn, T. Hankemeier, B.S. Kristal, B. van Ommen, E. Pujos-Guillot, E. Verheij, D. Wishart, S. Wopereis, Mass-spectrometry-based metabolomics: limitations and recommendations for future progress with particular focus on nutrition research, Metabolomics. 5 (2009) 435–458.

[3] N.E. Oro, C.A. Lucy, Analysis of the Nitrogen Content of Distillate Cut Gas Oils and Treated Heavy Gas Oils Using Normal Phase HPLC, Fraction Collection and Petroleomic FT-ICR MS Data, Energy Fuels. 27 (2013) 35–45.

[4] S. Samanipour, M. Reid, K. Bæk, K.V. Thomas, Combining a deconvolution and a universal library search algorithm for the non-target analysis of data independent LC-HRMS spectra, Environ. Sci. Technol. (2018). doi:10.1021/acs.est.8b00259.

[5] A. Alonso, S. Marsal, A. Julià, Analytical methods in untargeted metabolomics: state of the art in 2015, Front Bioeng Biotechnol. 3 (2015) 23.

[6] R. Tautenhahn, C. Böttcher, S. Neumann, Highly sensitive feature detection for high resolution LC/MS, BMC Bioinformatics. 9 (2008) 504.

[7] N.G. Mahieu, G.J. Patti, Systems-Level Annotation of a Metabolomics Data Set Reduces 25 000 Features to Fewer than 1000 Unique Metabolites, Anal. Chem. 89 (2017) 10397–10406.

[8] John Wiley & Sons, Inc., ed., Mass Spectrometry, in: Kirk-Othmer Encyclopedia of Chemical Technology, John Wiley & Sons, Inc., Hoboken, NJ, USA, 2000: p. 542.

[9] X. Domingo-Almenara, J.R. Montenegro-Burke, H.P. Benton, G. Siuzdak, Annotation: A Computational Solution for Streamlining Metabolomics Analysis, Anal. Chem. 90 (2018) 480–489.

[10] J. Xia, D.I. Broadhurst, M. Wilson, D.S. Wishart, Translational biomarker discovery in clinical metabolomics: an introductory tutorial, Metabolomics. 9 (2013) 280–299.

[11] A.B. Owen, Variance of the Number of False Discoveries, J. R. Stat. Soc. Series B Stat. Methodol. 67 (2005) 411–426.

[12] J.T. Leek, J.D. Storey, Capturing heterogeneity in gene expression studies by surrogate variable analysis, PLoS Genet. 3 (2007) 1724–1735.

[13] Y. Benjamini, D. Yekutieli, The Control of the False Discovery Rate in Multiple Testing under Dependency, Ann. Stat. 29 (2001) 1165–1188.

[14] L.D. Roberts, A.L. Souza, R.E. Gerszten, C.B. Clish, Targeted metabolomics, Curr. Protoc. Mol. Biol. Chapter 30 (2012) Unit 30.2.1–24.

[15] B. Zhou, J.F. Xiao, L. Tuli, H.W. Ressom, LC-MS-based metabolomics, Mol. Biosyst. 8 (2012) 470–481.

[16] K. Uppal, D.I. Walker, D.P. Jones, xMSannotator: An R Package for Network-Based Annotation of High-Resolution Metabolomics Data, Anal. Chem. 89 (2017) 1063–1067.

[17] C.D. Broeckling, F.A. Afsar, S. Neumann, A. Ben-Hur, J.E. Prenni, RAMClust: a novel feature clustering method enables spectral-matching-based annotation for metabolomics data, Anal. Chem. 86 (2014) 6812–6817.

[18] K. Morreel, Y. Saeys, O. Dima, F. Lu, Y. Van de Peer, R. Vanholme, J. Ralph, B. Vanholme, W. Boerjan, Systematic structural characterization of

627 metabolites in Arabidopsis via candidate substrate-product pair networks, Plant
628 Cell. 26 (2014) 929–945.
629 [19] M. Wang, J.J. Carver, V.V. Phelan, L.M. Sanchez, N. Garg, Y. Peng, D.D.
630 Nguyen, J. Watrous, C.A. Kapono, T. Luzzatto-Knaan, C. Porto, A. Bouslimani,
631 A.V. Melnik, M.J. Meehan, W.-T. Liu, M. Crüsemann, P.D. Boudreau, E.
632 Esquenazi, M. Sandoval-Calderón, R.D. Kersten, L.A. Pace, R.A. Quinn, K.R.
633 Duncan, C.-C. Hsu, D.J. Floros, R.G. Gavilan, K. Kleigrewe, T. Northen, R.J.
634 Dutton, D. Parrot, E.E. Carlson, B. Aigle, C.F. Michelsen, L. Jelsbak, C.
635 Sohlenkamp, P. Pevzner, A. Edlund, J. McLean, J. Piel, B.T. Murphy, L. Gerwick,
636 C.-C. Liaw, Y.-L. Yang, H.-U. Humpf, M. Maansson, R.A. Keyzers, A.C. Sims,
637 A.R. Johnson, A.M. Sidebottom, B.E. Sedio, A. Klitgaard, C.B. Larson, C.A.B. P,
638 D. Torres-Mendoza, D.J. Gonzalez, D.B. Silva, L.M. Marques, D.P. Demarque, E.
639 Pociute, E.C. O'Neill, E. Briand, E.J.N. Helfrich, E.A. Granatosky, E. Glukhov, F.
640 Ryffel, H. Houson, H. Mohimani, J.J. Kharbush, Y. Zeng, J.A. Vorholt, K.L.
641 Kurita, P. Charusanti, K.L. McPhail, K.F. Nielsen, L. Vuong, M. Elfeki, M.F.
642 Traxler, N. Engene, N. Koyama, O.B. Vining, R. Baric, R.R. Silva, S.J. Mascuch,
643 S. Tomasi, S. Jenkins, V. Macherla, T. Hoffman, V. Agarwal, P.G. Williams, J.
644 Dai, R. Neupane, J. Gurr, A.M.C. Rodríguez, A. Lamsa, C. Zhang, K. Dorrestein,
645 B.M. Duggan, J. Almaliti, P.-M. Allard, P. Phapale, L.-F. Nothias, T. Alexandrov,
646 M. Litaudon, J.-L. Wolfender, J.E. Kyle, T.O. Metz, T. Peryea, D.-T. Nguyen, D.
647 VanLeer, P. Shinn, A. Jadhav, R. Müller, K.M. Waters, W. Shi, X. Liu, L. Zhang,
648 R. Knight, P.R. Jensen, B.O. Palsson, K. Pogliano, R.G. Linington, M. Gutiérrez,
649 N.P. Lopes, W.H. Gerwick, B.S. Moore, P.C. Dorrestein, N. Bandeira, Sharing and
650 community curation of mass spectrometry data with Global Natural Products
651 Social Molecular Networking, Nat. Biotechnol. 34 (2016) 828–837.
652 [20] C. Guijas, J.R. Montenegro-Burke, X. Domingo-Almenara, A. Palermo, B.
653 Warth, G. Hermann, G. Koellensperger, T. Huan, W. Uritboonthai, A.E. Aisporna,

654 D.W. Wolan, M.E. Spilker, H.P. Benton, G. Siuzdak, METLIN: A Technology
655 Platform for Identifying Knowns and Unknowns, Anal. Chem. 90 (2018) 3156–
656 3164.
657 [21] M. Loos, H. Singer, Nontargeted homologue series extraction from
658 hyphenated high resolution mass spectrometry data, J. Cheminform. 9 (2017) 12.
659 [22] L. Sleno, The use of mass defect in modern mass spectrometry, J. Mass
660 Spectrom. 47 (2012) 226–236.
661 [23] E.M. Thurman, I. Ferrer, The isotopic mass defect: a tool for limiting
662 molecular formulas by accurate mass, Anal. Bioanal. Chem. 397 (2010) 2807–
663 2816.
664 [24] H. Zhang, D. Zhang, K. Ray, M. Zhu, Mass defect filter technique and its
665 applications to drug metabolite identification by high-resolution mass spectrometry,
666 J. Mass Spectrom. 44 (2009) 999–1016.
667 [25] K.J. Jobst, L. Shen, E.J. Reiner, V.Y. Taguchi, P.A. Helm, R. McCrindle, S.
668 Backus, The use of mass defect plots for the identification of (novel) halogenated
669 contaminants in the environment, Anal. Bioanal. Chem. 405 (2013) 3289–3297.
670 [26] D. Vuckovic, I. de Lannoy, B. Gien, R.E. Shirey, L.M. Sidisky, S. Dutta, J.
671 Pawliszyn, In vivo solid-phase microextraction: capturing the elusive portion of
672 metabolome, Angew. Chem. Int. Ed Engl. 50 (2011) 5344–5348.
673 [27] O. Teahan, S. Gamble, E. Holmes, J. Waxman, J.K. Nicholson, C. Bevan,
674 H.C. Keun, Impact of analytical bias in metabonomic studies of human blood
675 serum and plasma, Anal. Chem. 78 (2006) 4307–4318.
676 [28] V. Bessonneau, J. Ings, M. McMaster, R. Smith, L. Bragg, M. Servos, J.
677 Pawliszyn, In vivo microsampling to capture the elusive exposome, Sci. Rep. 7
678 (2017) 44038.

679 [29] G. Libiseller, M. Dvorzak, U. Kleb, E. Gander, T. Eisenberg, F. Madeo, S.
680 Neumann, G. Trausinger, F. Sinner, T. Pieber, C. Magnes, IPO: a tool for
681 automated optimization of XCMS parameters, BMC Bioinformatics. 16 (2015) 118.
682 [30] C.A. Smith, E.J. Want, G. O'Maille, R. Abagyan, G. Siuzdak, XCMS:
683 processing mass spectrometry data for metabolite profiling using nonlinear peak
684 alignment, matching, and identification, Anal. Chem. 78 (2006) 779–787.
685 [31] C.C. Bridges Jr., Hierarchical Cluster Analysis, Psychol. Rep. 18 (1966)
686 851–854.
687 [32] T.U.H. Baumeister, N. Ueberschaar, W. Schmidt-Heck, J.F. Mohr, M.
688 Deicke, T. Wichard, R. Guthke, G. Pohnert, DeltaMS: a tool to track isotopologues
689 in GC- and LC-MS data, Metabolomics. 14 (2018) 41.
690 [33] Y.M. Tikunov, S. Laptenok, R.D. Hall, A. Bovy, R.C.H. de Vos, MSClust: a
691 tool for unsupervised mass spectra extraction of chromatography-mass
692 spectrometry ion-wise aligned data, Metabolomics. 8 (2012) 714–718.
693 [34] E.A.P. Ekanayaka, M.D. Celiz, A.D. Jones, Relative mass defect filtering of
694 mass spectra: a path to discovery of plant specialized metabolites, Plant Physiol.
695 167 (2015) 1221–1232.
696 [35] B.O. Keller, J. Sui, A.B. Young, R.M. Whittal, Interferences and
697 contaminants encountered in modern mass spectrometry, Anal. Chim. Acta. 627
698 (2008) 71–81.
699 [36] R.M. Smith, ed., Neutral Losses and Ion Series, in: Understanding Mass
700 Spectra: A Basic Approach, John Wiley & Sons, Inc., Hoboken, NJ, USA, 2004:
701 pp. 121–149.
702 [37] G.H. Dunteman, Principal Components Analysis, SAGE, 1989.
703 [38] D.L. Lynch, P.H. Reggio, Cannabinoid CB1 receptor recognition of
704 endocannabinoids via the lipid bilayer: molecular dynamics simulations of CB1

transmembrane helix 6 and anandamide in a phospholipid bilayer, J. Comput. Aided Mol. Des. 20 (2006) 495–509.

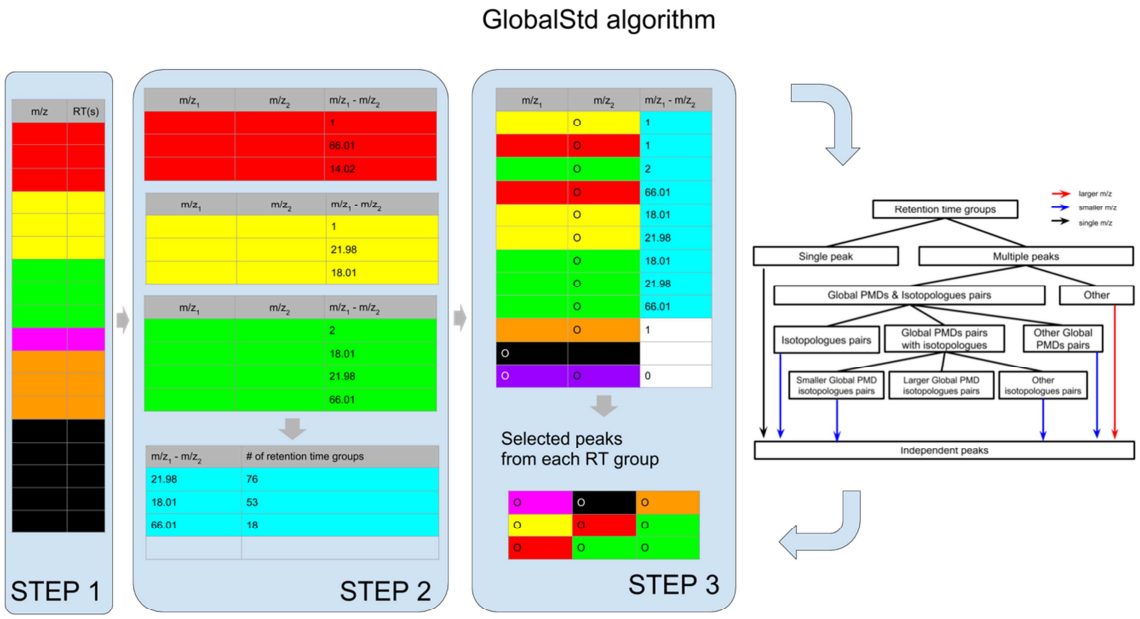
[39] H. Tong, D. Bell, K. Tabei, M.M. Siegel, Automated data massaging, interpretation, and e-mailing modules for high throughput open access mass spectrometry, J. Am. Soc. Mass Spectrom. 10 (1999) 1174–1187.

[40] C.J. Rebouche, A.G. Engel, Tissue distribution of carnitine biosynthetic enzymes in man, Biochimica et Biophysica Acta (BBA) - General Subjects. 630 (1980) 22–29.

[41] C. Rigault, F. Mazué, A. Bernard, J. Demarquoy, F. Le Borgne, Changes in l-carnitine content of fish and meat during domestic cooking, Meat Sci. 78 (2008) 331–335.

[42] G.M. Neumann, P.G. Cullis, P.J. Derrick, Mass Spectrometry of Polymers: Polypropylene Glycol, Zeitschrift Für Naturforschung A. 35 (1980). doi:10.1515/zna-1980-1015.

[43] K. Yang, X. Han, Accurate quantification of lipid species by electrospray ionization mass spectrometry - Meet a key challenge in lipidomics, Metabolites. 1 (2011) 21–40.



Scheme 1. Demonstration of GlobalStd algorithm. Different colors stand for peaks from different retention time groups. Blue stands for mass pairs with high frequency paired mass distances (PMD). Step 1 indicates the retention time cluster analysis to find Pseudospectra for potential compounds. Step 2 indicates the PMD-based global search. Step 3 indicates selection of independent peaks and detailed process is shown on the right flowchart.

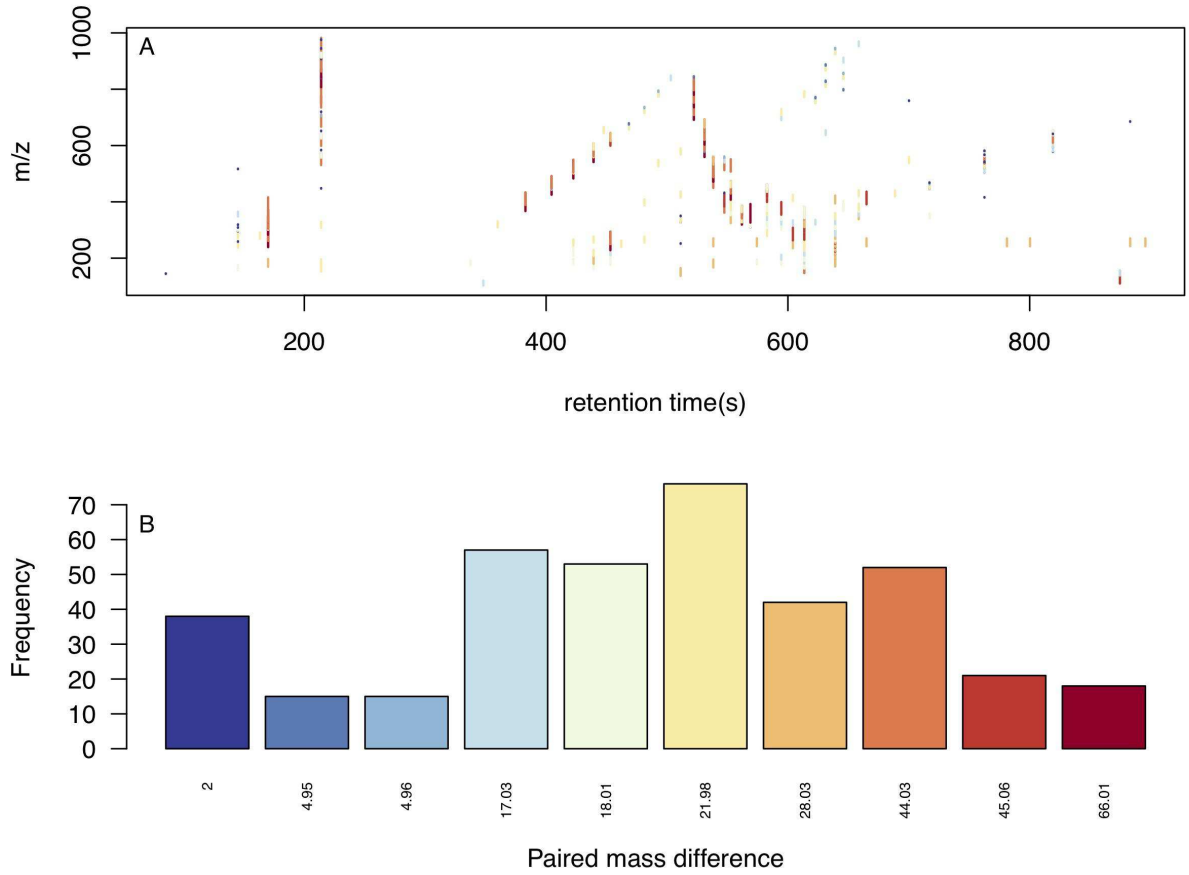


Figure 1. Paired mass distance (PMD) analysis for *in vivo* SPME sampling data. The plot on the top of the figure(A) illustrates the PMD relationship across retention time – m/z profile, while the bottom plot(B) shows the corresponding PMD frequency. The paired relationships are reflected by the lines between peaks. The colors of the segments in the top plot correspond to the colors in the bar plot below, indicating the PMD groups.

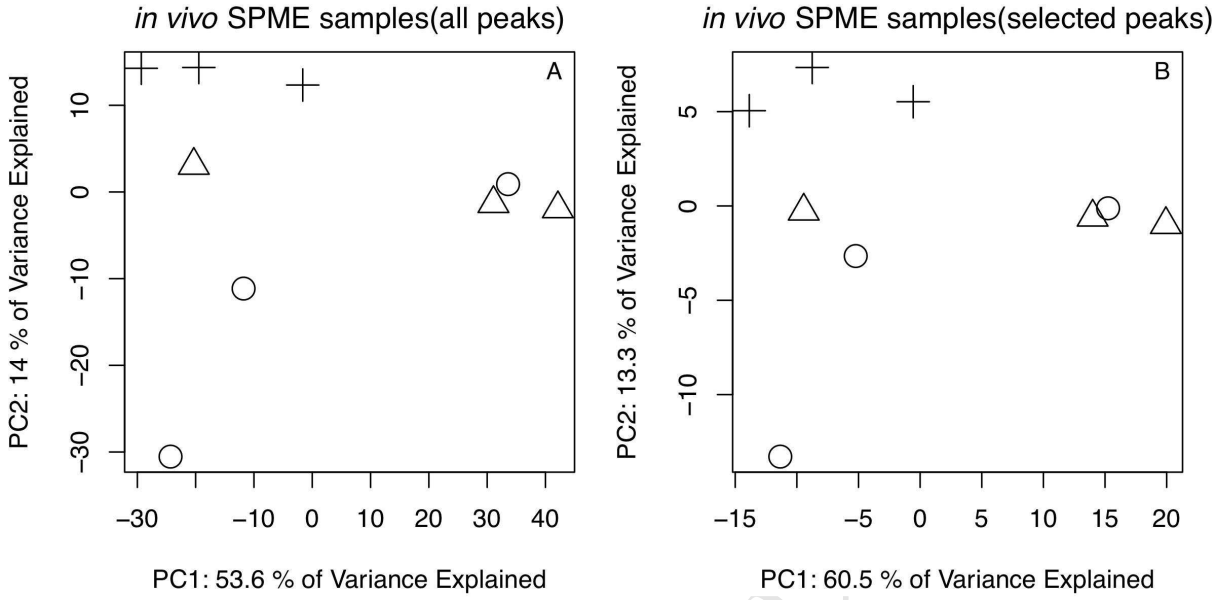


Figure 2. Score plot from principal components analysis (PCA) of raw peaks (A) and Std peaks selected by GlobalStd algorithm (B). Use of the same symbols in plots indicates technique replicates, while different symbols indicate biological replicates.

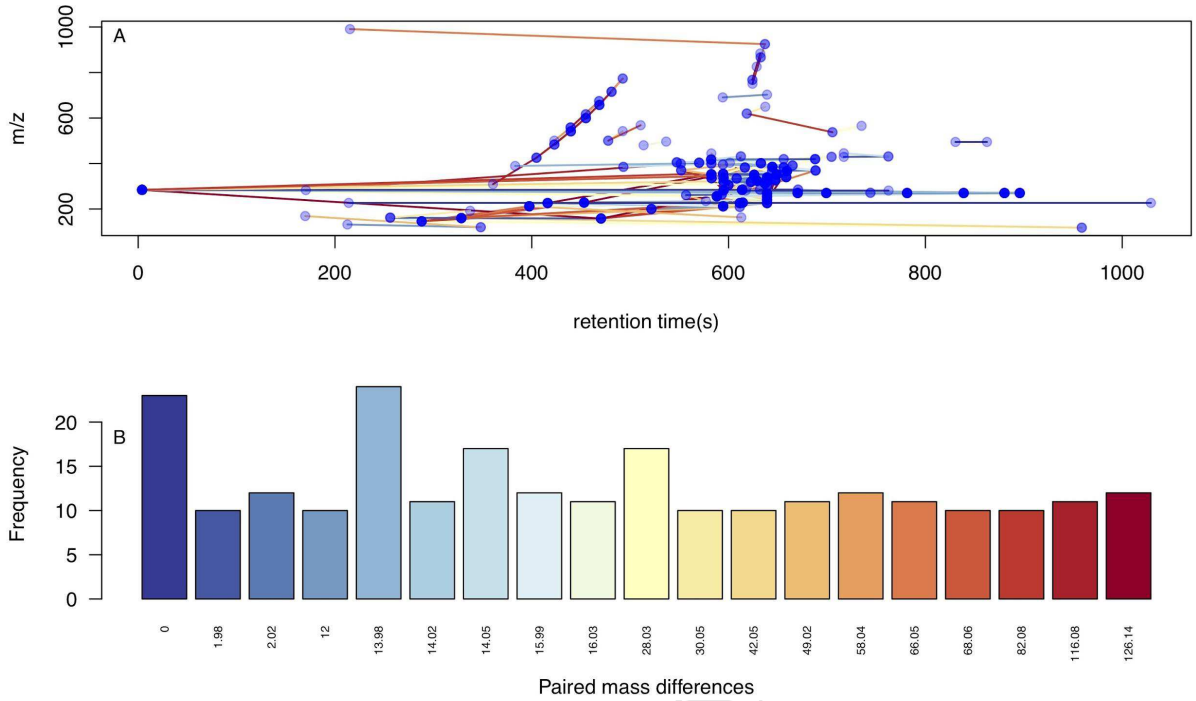


Figure 3. Structure/reaction directed analysis in *in vivo* SPME sampling data for peaks selected via application of the GlobalStd algorithm. The plot above (A) illustrates the PMD relationship across retention time and m/z profile, while the bottom plot (B) indicates the corresponding PMD frequency. The paired relationships are reflected by the line between peaks. The colors of the segments in figure 3A correspond the colors in the bar plot below, which categorize the PMD groups.

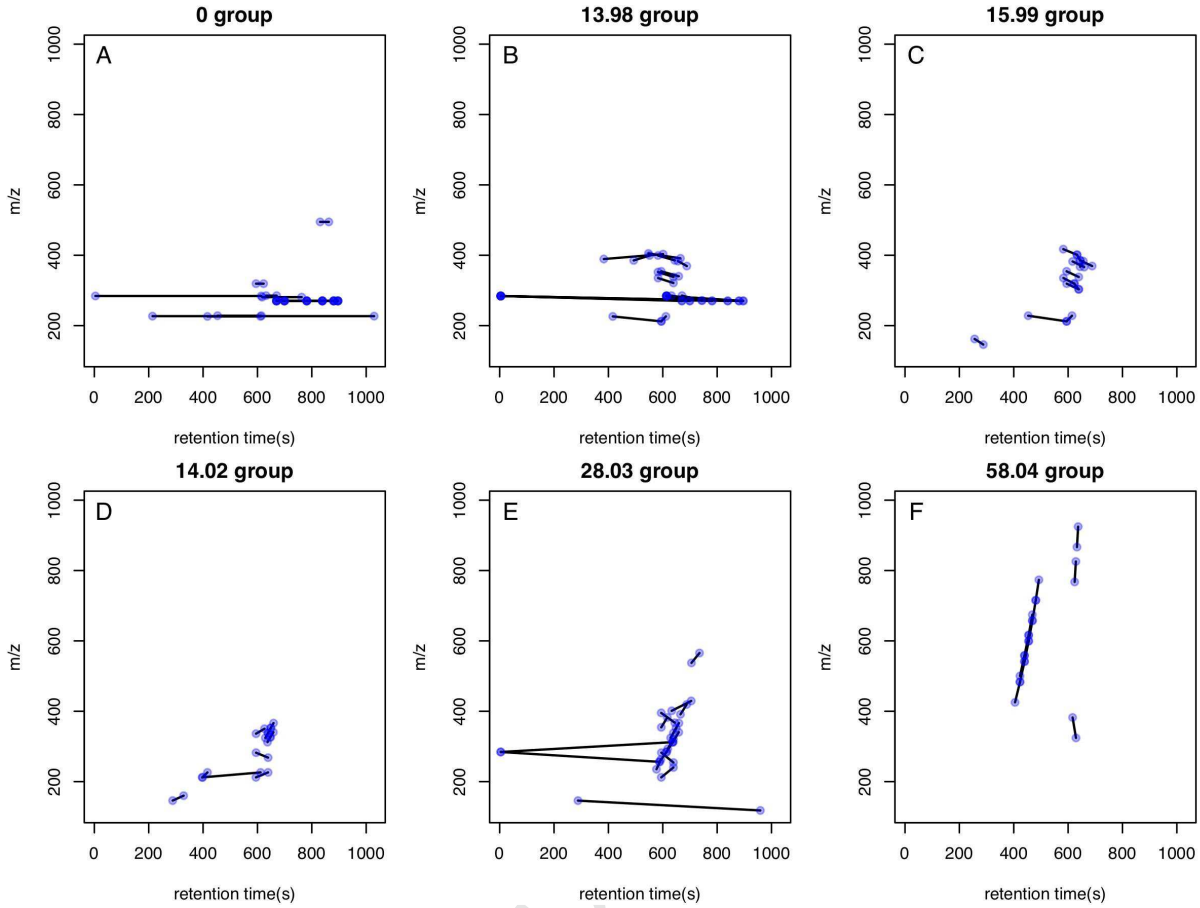


Figure 4. m/z - retention time peak profiles obtained via GlobalStd algorithm analysis for PMD 0Da (A, isomers), PMD 13.98Da (B, replacement of oxygen atom and two hydrogen atoms), PMD 15.99Da (C,oxidation), PMD 14.02Da (D, homologous series with -CH_2-), PMD 28.03Da (E, homologous series with -C_2H_4-), and PMD 58.04Da (F, homologous series with -C_3H_6O-). The paired relationships are reflected by the lines between peaks.

Highlights

- Algorithms were developed to reduce redundant peaks in metabolomics data profile
- 20% of the original peaks could stand for the major variances for data
- Quantitative analysis could be performed at structure/reaction level
- Unknown structure/reaction relationships could be revealed by *in vivo* SPME sampling