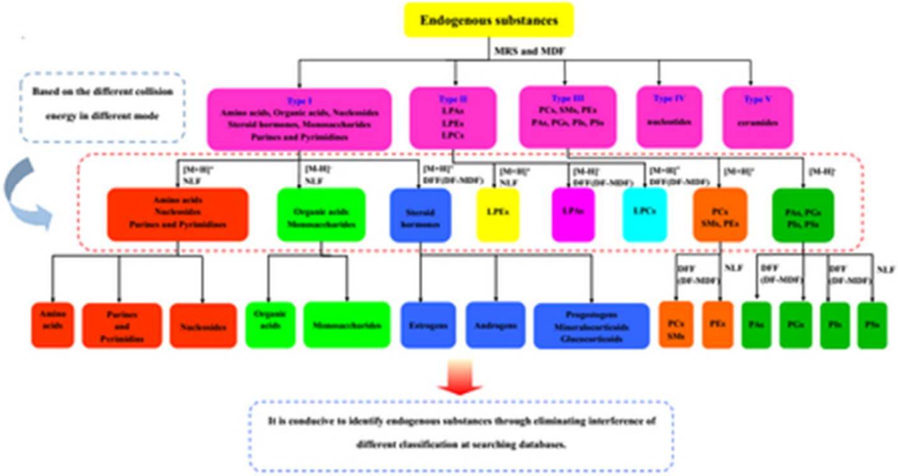
# RSC Advances

Volume 1 | Number 1 | Jan 2013 | Pages 1–100
RSC Advances
www.rsc.org/advances
ROYAL SOCIETY OF CHEMISTRY

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/advances



39x22mm (300 x 300 DPI)

Cite this: DOI: 10.1039/c0xx00000x

www.rsc.org/xxxxxx

ARTICLE TYPE

# A rapid and integrated pyramid screening method to classify and identify complex endogenous substances with UPLC/Q-TOF MS-based metabolomics

**Yubo Li,[a] Zhenzhu Zhang,[a] Zhiguo Hou,[a] Lei Wang,[a] Xin Wu,[a] Liang Ju,[a] Xiuxiu Zhang[a] and Yanjun Zhang*[a]**

## Abstract

Metabolomics plays an important role in disease diagnosis, safety and efficacy of drug evaluation, and microbial research. Liquid chromatography–mass spectrometry (MS) is the main analysis tool in metabolomics studies. Given the existence of many endogenous metabolites with their different categories isomerizing one another, some problems on classification and identification have emerged. These problems result in high false-positive results, as well as a complex and time-consuming substance identification process. Accordingly, this study reviewed literature and retrieved databases to identify endogenous substances in the same category accompanied by a certain mass range (MR) and mass defect range (MDR), as well as an identical or similar fragmentation pattern in the mass spectrum [*i.e.*, characteristic and neutral loss (NL) fragments]. We conducted different MS/MS collision energies to analyze different categories of endogenous substance to discover and summarize their fragmentation patterns. We then used the MR and MDR of the parent ion, diagnostic fragments (together with their MDR), and NL as screening tools to establish a pyramid screening method (PSM) for the rapid classification and identification of metabolites. Finally, we compared the PSM with the conventional identification method through known compounds in literatures. PSM was found to solve the key problem in metabolomics to some extent, namely, the classification and identification of substances. This method also facilitated the further development of metabolomics and provided a new perspective on the screening and identification of target components in other complex samples.

## Introduction

Metabolomics attempts to comprehensively detect and quantitatively analyze low molecular weight metabolites *in vivo* and seeks the physiological and pathological variations of endogenous substances to determine their connection. [1–2] These processes are involved in a series of different types of molecules, such as nucleic acids, hormones, amino acids, organic acids, and lipids. [3–6] Liquid chromatography–mass spectrometry (LC-MS) is more suitable to detect complicated metabolites and identify potential biomarkers in biological samples because of its high sensitivity, wide dynamic range, and simple sample processes, so it has become an important technology for metabolomics data acquisition. [7-8] Considering the complicated LC-MS spectrogram derived from metabolomics that involves a large amount of information, chemometrics and bioinformatics are required to mine and integrate the original data. Despite the complexity of endogenous substances, they are necessary to detect more substances in processing non-targeting metabolic fingerprint analysis and further explore the biological significance of the involved endogenous substances. [9] The physical and chemical properties of different classes of substances in the same complex system are different. Therefore, performing a preliminary classification of metabolites that will benefit the identification of different metabolites is critical. Conventional metabolite identification methods involved in comparing the information of standard substances or referencing data in published literature or matching information in the database are critical. Considering that the standards are not easily available, the mass spectrum analysis is complex and time-consuming, and different classes of metabolites exist as isomers, establishing a rapid and accurate method for the classification and identification of endogenous metabolites based on LC-MS technology is necessary.

Nowadays, mass defect filtering (MDF), diagnostic fragments filtering (DFF), and neutral loss filtering (NLF) show the unique advantages of screening and identification with the development of data processing technology. [10–12] MD is the D-value of the accurate mass and its nearest integer value of the molecular compound. Given that the same class of compounds generally possesses the same core structure, MDF use a range of MD values (MDV) of a certain category to screen compounds. Recently, MDF technology has been used for the detection and characterization of metabolites in biological matrices, [13] designer drugs, [14] analysis of herbal components, [15] distinct proteomics potential of iodine-based reagents, [16] and imaging of lipid species. [17] However, this technology has yet to be reported in the field of metabolomics. When DFF and NLF are applied to the screening and identification of substances, fragment ion abundances are largely influenced by the collision energy. Therefore, studying fragmentation rules of substances under particular collision energy can cause the loss of informational fragments. [18-19] Using MS to analyze fragmentation information, the limitation of the

DF-MD range (DF-MDR) can narrow the screening range and prevent a false-positive consequence.[16]

We conducted a comprehensive analysis and systemic integration of endogenous metabolites based on the structure commonalities of the same class and specificity between the different classes in the present study using data post-processing techniques (*i.e.*, MDF, DFF, DF-MDF and NLF) to establish a rapid classification and identification method, namely, pyramid screening method (PSM). First, we inducted and collated classified information and primary information of endogenous substances in KEGG and HMDB to build the information database of different categories of endogenous substances. We continuously used the LC-MS technique combined with different collision energies to analyze the MS and two-stage MS of different classes of metabolite standards. Using information on endogenous substance fragments in the literature, Massbank, and Metlin database, we identified and summarized the rule of mass range (MR), MDR, DF, DF-MDR, and NL of endogenous substances belonging to 6 classes (20 subclasses). We then considered the MR and MDR of the parent ion, as well as DF, DF-MDR, and NL as screening tools to build a rapid PSM for the classification and identification of metabolites. Finally, we compared PSM with conventional metabolite identification methods by known substances reported in literature. This study attempts to solve the problem of classification and identification of metabolites. It also provides a feasible method for the classification and identification of potential biomarkers in non-targeted metabolic fingerprint analysis and targeted metabonomic. This study also provides a new way for screening target components in other complex samples.

## Experimental

### Reagents and materials

HPLC-grade methanol was purchased from Oceanpak (Goteborg, Sweden). Distilled water was obtained from Wahaha Company (Hangzhou, China). Adenine and β-thymidine were obtained from Heowns Biochem Technologies LLC (Tianjin, China). Aldosterone and uridine were obtained from Jince Analysis Technology Co. Ltd. (Tianjin, China). LPC18:0 and progesterone were purchased from Sigma-Aldrich (St. Louis, USA). Testosterone and estrone were purchased from Shilan Technology Co. Ltd. (Tianjin, China) and TCI development Co. Ltd. (Shanghai, China), respectively. D-Glucose anhydrous, fructose, succinic acid, dopamine hydrochloride, uracil, 17β-estradiol, L-histidine, L-lysine, and hydrocortisone were all obtained from the National Institute for the Control of Pharmaceutical and Biological Products (Beijing, China). The above mentioned materials were dissolved in methanol/water (1/1, v/v) for the qualitative detection of MS.

### Animal experiment

Plasma and urine samples are from Wistar rats (weighing 200 ± 20 g) raised in SPF level laboratory of Chinese Academy of Medical Sciences & Peking Union Medical College Institute of Biomedical Engineering (Tianjin, China). This study was approved by the Animal Ethics Committee of Tianjin, University of Traditional Chinese Medicine, under permit number TCM-2014-04-A11.

### Sample collection and preparation

Prior to sample collection, food and water were withheld from the animals for 12 hours to eliminate food and water effect in the final result. Plasma was collected from the intraocular angular vein and was placed in a heparinization tube. Then, plasma was separated by centrifugation at 3500 rpm for 15 min and was stored at –80 °C until metabonomic analysis. The plasma was thawed at room temperature before processing. A 300 µL acetonitrile portion was added to 100 µL of the plasma, and the mixture was ultrasonicated in cold water for 10 min, vortexed for 1 min, and then centrifuged at 13000 rpm for 15 min. After centrifugation, the supernatants were collected for direct analysis. Urine samples were collected from Wistar rats and maintained at –80 °C before analysis. Then, the samples were thawed at room temperature and centrifuged at 10000 rpm for 10 min. The supernatant was then collected, diluted with distilled water at a ratio of 1:1, vortexed for 1 min, and then centrifuged at 13000 rpm for 15 min. The supernatant was subjected to UPLC-Q-TOF/MS analysis.

### Chromatographic conditions

Analysis was performed in a Waters UPLC-Q-TOF/MS system (Waters, USA). The 10 and 5 µL plasma samples and urine samples, respectively, were injected into ACQUITY UPLC HSS C18 column (2.1 × 100 mm, 1.7 µm, Waters). The column temperature was set at 40 °C and the flow rate at 0.3 mL/min. The UPLC separation system includes a binary solvent system with mobile phase A (0.1% formic acid in water) and mobile phase B (0.1% formic acid in acetonitrile) in the plasma and urine samples. The gradient profiles for the plasma samples were as follows: 99% A; followed by 0–0.5 min, A: 99%–99%; 0.5–2 min, A: 99%–50%; 2–9 min, A: 50%–1%; 9–10 min, A: 1%–1%; 10–10.5 min, A: 1%–99%; and 10.5–12 min, A: 99%–99%. The gradient profiles of the urine samples started with 99% A, followed by 0–8.5 min, A: 99%–75%; 8.5–11 min, A: 75%–50%; 11–13 min, A: 50%–10%; 13–15 min, A: 10%–1%; 15–17 min, A: 1%–1%; 17–18.5 min, A: 1%–99%; and 18.5–20 min, A: 99 %–99%.

### Mass spectrometer conditions

We employed the Xevo G2 Q-TOF mass spectrometer (MS) from Waters Corporation (Milford, MA, USA) to detect precursor ions and product ions of endogenous substances via an electrospray ionization interface. There are instrument configuration parameters listed as follows. Resolution can reach 22,000 when we detected endogenous small molecules of molecular mass less than 1000. The value of Lteff and Veff is 1800 and 6316.84, respectively. And it can detect 2 minimum points in peak. The Xevo G2 Q-TOF MS collect data through the WatersADC as acquisition device, which maximum acquisition rate (TOF mode) is 30 scans/sec (MS or MS/MS) and average single ion intensity is 33. Lockspray configuration was obtained under following conditions: reference scan frequency of 10.00 sec; reference cone voltage of 40.00 V; reference collision energy of 6.000; reference DRE Setting of 99.900. And the experimental MS parameters were as follows: drying gas temperature of 325 °C; drying gas flow of 10 ml/min; desolvation gas flow of 600 L/h; capillary voltage of 3.5 kV; fragmentor voltage of 6 kV; collision energies

Cite this: DOI: 10.1039/c0xx00000x

www.rsc.org/xxxxxx

of 10 eV to 20 eV, 20 eV to 30 eV, and 30 eV to 40 eV; nebulizer pressure of 350 psi; evaporative and auxiliary gases with high purity nitrogen; and reference ions ($[M+H]^+ = 556.2771$, $[M-H]^- = 554.26$) were employed to ensure accuracy during spectral acquisition. The range of data acquisition was 50 Da to 1000 Da.

### Data analysis

Based on the classification of small molecules in the KEGG database (http://www.genome.jp/kegg/), the study combined the KEGG database with the HMDB database (http://www.hmdb.ca/) to comprehensively search endogenous substances. We recorded detailed information of endogenous substances and built a in-house database using Excel spreadsheets using the Y-axis as the compound name and X-axis as the compound information. The X-axis included monoisotopic molecular weight, molecular formula, and structural formula. The study also included MS fragments of endogenous small molecules in MassBank (http://www.massbank.jp), Medline (http://www.ncbi.nlm.nih.gov/pubmed/), and other reported literature. Finally, we summarized the rule of fragmentation pattern in the same category substances.

We introduced the process of building an endogenous substances database in the case of amino acids. First, we identified a clear classification of the major endogenous small molecules through the KEGG database, and then searched amino acids among its classes. Approximately 700 acquired substances were then exported. Excluding the exogenous substances, derivatives replaced the functional group, compounds containing special elements (*i.e.*, Se, F, Cl, and Br), and substances of non-amino acid monomers. The self-built database recorded 113 amino acids that included 23 isomer groups. Finally, we obtained fragmentation information of amino acids provided by the HMDB, Massbank, Metlin, literature, and reference substance experiment.

## Results and discussion

### Mass range screening and mass defect filtering

MD is the value of the difference between the accurate mass and integer mass of compound. Considering that the same general class of compounds has the same or similar core structures, a relatively narrow MDR is observed within a certain MR. We attempted to limit the MR and MDR of the parent ion in the initial classification of endogenous substances based on the compound characteristics. Using information on endogenous substances in the self-built database, we calculated the MDV of each class based on the MD formula and summarized the MR and MDR in different classes of compounds. Notably, the corresponding MDV of the MR's maximum (or minimum) values of the same class can be same with the MDR's maximum (or minimum) values, which is due to the MDV being associated with the number of atoms and atomic composition of compounds. In purine and pyrimidine bases, for example, Cytosine has a minimum mass of 111.0432, whereas 8-Hydroxy-7-methyl-guanine has the largest mass of 181.0599. However, their corresponding MDVs are not extreme values of bases. We eventually set the MR and MDR of purine and pyrimidine bases at 111.0 to 181.1 and 0.0272 to 0.0702, respectively.

The scatter plot of the accurate mass of compounds with the corresponding MDV is given in Figure 1. This figure shows that the different classes of substances have a certain MR and MDR, and mass range screening (MRS) is conducive for classifying and screening different classes or different subclasses combined with MDF. LPCs and PCs have similar physical and chemical properties. We generally have some difficulty in extraction and separation using physical and chemical methods. Considering that the MRs of LPCs and PCs are 412.2 to 607.5 and 673.4 to 957.9, respectively, we can use the different MRs of these two substances to rapidly classify LC-MS. Thus, MRs have a certain role in filtering different classes of substances. However, the mere use of MR would hinder the further filtering of substances and increase the degree of their confusion. The MR of nucleotides and steroids were 304.0 to 524.0 and 270.1 to 378.3, respectively. Thus, we cannot separate them by simply using accurate MR. However, they have different MDRs (*i.e.*, the former is -0.0314 to 0.0682, whereas the latter is 0.1412 to 0.2403). These two classes of substances can achieve further screening by MDF. Thus, the double limit of MR and MDR can help to accurately screen endogenous substances.

Therefore, we divided endogenous substances into five types from the initial six classes (20 subclasses) by MR and MDR. The first type (I) includes amino acids, organic acids, monosaccharides, nucleic acids (containing bases and nucleosides), and steroid hormones (containing estrogens, androgens, progestagens, glucocorticoids, and mineralocorticois), which had MRs and MDRs of 60.0 to 378.3 and -0.0081 to 0.2403, respectively. The MRs and MDRs of type (II) phospholipids (containing LPAs, LPCs, and LPEs) are 410.2 to 607.5 and 0.2385 to 0.4577, respectively. Type (III) phospholipids (containing PCs, SMs, PEs, PAs, PGs, PIs, and PSs) had 631.4 to 957.9 and 0.4111 to 0.8126, respectively. Type (V) had nucleic acids (containing one to three phosphate groups) with MRs and MDRs of 304.0 to 524.0 and -0.0314 to 0.0682, respectively. Type (IV) has phospholipids (including ceramides) with MRs and MDFRs of 481.4 to 679.7 and 0.4494 to 0.6843, respectively. Given the complex composition of endogenous substances, the MR and MDR overlap and intersect in different categories of substances. Although different categories of substances can be distinguished, these substances cannot be accurately classified. We need to use other data processing techniques in conjunction with screen complex endogenous substances to classify and analyze the endogenous substances rapidly and accurately.

### Diagnostic fragments filtering, diagnostic fragments-mass defect filtering and neutral loss filtering

Endogenous substances exist in a wide variety and complex chemical composition *in vivo*, but the same class of compounds has similar or identical core structures. Thus, the same fragmentation behavior of these compounds is observed during an induced collision. Therefore, we use this feature to search for fragmentation patterns in MS. Notably, fragment ions are impacted in a larger extent by the collision energy. When the collision energy is too low, the fragment ion abundance can increase. In contrast, the fragment ions can fracture to other fragment ions. Thus, the single collision energy can result in missing information on the fragment ions. There are BPI chromatograms of plasma and urine of rat sample at the positive and negative mode combined with different energy collision showing Figure 2.

Diagnostic fragments are defined as the characteristic fractured fragments that the same class of compounds can form in the fracturing process. Therefore, DFF can be used to screen different classes of substances, as well as to classify and identify known compounds or speculate unknown compound structures. However, other similar mass fragment ions are generated by interference when the DF ions are extracted because of the measurement error of the instrument. Thus, we can calculate the MDF value of the DF in the error range of ±10 ppm (instrument permissible error range), then built a DF-MDR that eliminates false positive results. NL is the difference mass level (m) between the MS and two-stage MS, and is utilized to roughly classify substances by a certain type of feature to replace groups. For example, amino acids easily lose neutral fragments of 17 Da ($NH_3$) and 46 Da (HCOOH) in the positive mode. However, a majority of small molecules of endogenous substances have carboxy, amino, or other functional groups. Thus, the feature of NL fragments is weaker so that it cannot accurately be screened for substances.

According to the different home-made classes of endogenous substances database, we selected representative substances from each class to MS<2> scans in different collision energies (*i.e.*, 10–20, 20–30, and 30–40 eV) (Table 1) and finally summarized rules of six categories (18 subclass) of endogenous substances using DF, DF-MDR, and NL (Table S1, S2). The bases and nucleosides of nucleic acids are separated from each other by different NL fragmentation [*i.e.*, the former is $NH_3$ or HNCO, whereas the latter is ($C_5H_8O_3$) or ($C_5H_8O_4$)]. Steroids and PCs were filtered because of different DFs (*i.e.*, the former is 97.0653 $[C_6H_9O]^+$ and 109.0653 $[C_7H_9O]^+$, whereas the latter is 184.0739 $[C_5H_{15}NO_4P]^+$). We found that the same class and different subclasses can be screened by DF and NL fragments based on the above methods. However, we also discovered that similar DFs or NLs exist in different classes of compounds, which results in selected confusion within the substances. Therefore, the mere use of DFF or NLF cannot accurately classify and screen different categories of substances. By combining with MDR and MDS established at “3.1” term, we used DFF, NLF, and DF-MDR to screen based on broad classification. This method better distinguishes different classes of substances and significantly improves the screening accuracy.

**Pyramid screening method**

We used MDF technology combined with MDS, DFF, DF-MDF, and NLF to build a PSM to rapidly classify and identify endogenous metabolites. This method is similar to a pyramid-type structure and can convert rough classification into detailed classification. PSM is effective for the rapid, accurate classification and identification of different classes of substances and subclasses of the congeneric substances. PSM provides a feasible method for the accurate screening substances based on mass and MDV. DF, DF-MDR, and NL serve as a further continuation and supplementary methods. When the PSM is used to rapidly and accurately classify and screen endogenous substances, we first have to confirm the m/z value of the parent ion to clarify the scope of mass and MD. Considering the different fragmentation behavior in different classes in MS, we then rapidly classify and filter endogenous substances. Finally, we further confirm the compound structure by querying the database based on confirming the classification of compounds and information of fragment ions. There is the schematic of the PSM showing Figure 3.

When the mass and MDV of the compound are in the ranges of 60.0 to 378.3 and -0.0081 to 0.2403, respectively, the compound is classified as type I substance (*i.e.*, amino acid, base and nucleoside, organic acid, monosaccharide and steroid hormone). Subsequently, we further screened fragment ions under the different induced collisions. If a compound loss of NL fragments involves $NH_3$ and HCOOH in the positive mode, the substance is identified as an amino acid. If a compound loses $NH_3$ or HNCO, it is identified as a base of nucleic acid. If a compound loses $C_5H_8O_3$ and $C_5H_8O_4$, it is identified as a nucleoside. If the compound contains m/z 97.06 $[C_6H_9O]^+$, m/z 109.06 $[C_7H_9O]^+$, and 121.06 $[C_8H_9O]^+$ of DFs and corresponded to the DF-MDR of 0.0643 to 0.0663, 0.0642 to 0.0664, and 0.0640 to 0.0666 Da, respectively, the compound can be an androgen, progesterone, glucocorticoid, or mineralocorticoid. If the compound contains m/z 145.06 $[C_{10}H_9O]^-$ of DFs and correspond to DF-MDR of 0.0638 to 0.0667 Da in the negative mode, the compound is identified as an estrogen. If a compound that lost $CO_2$ is identified as an organic acid and also loses $H_2O$ and $C_2H_4O_2$, it is identified as a monosaccharide.

When the mass and MDV of the compound are in the ranges of 410.2 to 607.5 and 0.2385 to 0.4577, respectively, the compound is classified as a type II substance, which includes an LPE, LPA, and LPC. If the compound contains m/z 184.07 $[C_5H_{15}NO_4P]^+$ of DFs and correspond to the DF-MDR of 0.0720 to 0.0758 Da in the positive mode, the compound is an LPC. The compound that loses ($C_2H_8NO_4P$) is identified as an LPE. A compound with an m/z 152.99 $[C_3H_6PO_5]^-$ DFs and corresponding DF-MDR of 0.9937 to 0.9968 Da is a PA in the negative mode.

When the mass and MDV of the compound are in the ranges of 631.4 to 957.9 and 0.4111 to 0.8126, respectively, the compound is classified as type III, namely, a PC, SM, PE, PS, PI, PA, and PG. A compound that contains m/z 184.07 $[C_5H_{15}NO_4P]^+$ of DFs and corresponds to a DF-MDR of 0.0720 to 0.0758 Da in the positive mode is either a PC or SM, whereas a compound that loses ($C_2H_8NO_4P$) is a PE. A compound that contains m/z 241.01 $[C_6H_{10}PO_8]^-$ of DFs and has a corresponding DF-MDR of 0.0088 to 0.0137 Da in the negative mode is a PI. A compound with m/z 152.99 $[C_3H_6PO_5]^-$ DFs and corresponding DF-MDR of 0.9937 to 0.9968 Da is either a PA or a PG. A compound that loses ($C_3H_5NO_2$) is a PS.

Cite this: DOI: 10.1039/c0xx00000x

www.rsc.org/xxxxxx

When the mass and MDV of the compound are in the ranges of 304.0 to 524.0 and -0.0314 to 0.0682, respectively, the compound is included in type IV, namely, a nucleoside. The mass and MDV of a nucleoside containing one phosphate group are 304.0 to 368.1 and 0.0253 to 0.0682, respectively. The mass and MDV of a nucleoside containing two phosphate groups are 387.0 to 443.1 and 0.0021 to 0.0346, respectively. The mass and MDV of a nucleoside containing three phosphate groups are 466.9 to 524.0 and -0.0314 to 0.0009, respectively.

When the mass and MDV of the compound are in the ranges of 481.4 to 679.7 and 0.4494 to 0.6843, respectively, the compound is identified as type V, namely, a ceramide. Finally, we performed a query in the HMDB bases to screen the different classes of compounds to quickly and accurately screen the target compound.

### PSM for the classification and identification of endogenous substances in biological samples

The chromatographic peak data in positive ion mode were extracted by MassHunter software to import an Excel spreadsheet before working.[20-21] After a multivariate statistical analysis (*i.e.*, principal component analysis and partial least-squares-discriminant analysis) of the data, we found that the ions that are the furthest ones from the origin contribute significantly to the clustering of different groups. Final, we discovered 22 potential bio-markers correlated with renal toxicity through S-plot and variable importance plot. We compared PSM with conventional identification methods for the classification and identification of known compounds in this study, in which LPCs are selected for their significant effect on renal toxicity.[22-23] The ion at Rt = 12.66 min and $[M+H]^+$ = 482.3233 can contain an odd number of nitrogen atoms in the positive mode because its precise molecular weight is 481.3168. When we use PSM, the MR and MDR of substances are confirmed first, so this potential biomarker is identified as Type II (containing LPAs, LPCs, LPEs). The substances can easily produce two clear fragments of m/z 184.07 and 104.01. Finally, we judged the substance as LPCs based on different diagnosis fragmentation and NL fragments in Type II. We utilized m/z 509.3481 to retrieve the HMDB database to clarify the structure of the substance and found two categories of substances (*i.e.*, LPEs and LPCs). We thus determine that the substance is LysoPC (17:0). The conventional identification process is as follows. First, we predicted the elemental composition of the parent ion by software. We then searched the HMDB and KEGG using mass or adduct mass to acquire possible endogenous substances. However, many compounds exist after excluding exogenous substances. Next, we compared the secondary spectrum information of all provided endogenous substances or match information in the database or reference data in literature, which is beneficial to confirm the substance structures. However, acquiring the standards of endogenous substances is difficult and the MS analysis process is complicated and time-consuming. Different categories of substances with the same mass interfere with the identification process of one another. PSM is established through many studies, standard substances experiments, and MS information in databases. Breaking rules of substances combined with mass properties of the same category of substances are determined, which digs and optimizes the process of classifying and identifying substances. This method can effectively decrease the interference of different categories in the identification process. It can also save time for material identification. Unlike the conventional identification process, PSM can save at least a third of the total time. However, this method lacks specificity to different substances in the same category.

We constructed a PSM of endogenous metabolites using MDF, DFF, DF-MDF, and NLF in this study. Unlike the conventional identification method, PSM can quickly and accurately classify and identify endogenous metabolites. Given the limitations of human knowledge and modern instruments, the conventional identification method cannot classify and confirm all potential endogenous markers. PSM attempts to use mass and MDV to predict the unknown substances. We can then extract and separate unknown substances based on the category attribution of unknown metabolites and the principle of similar physical and chemical properties in the same class substances. This approach is also combined with modern separation techniques, which can facilitate the further study of the unknown. PSM can provide a new clue for the separation and understanding of unknown compounds, as well as significantly promote the development of metabolomics.

Meanwhile, the PSM has some effect on the classification of unknown compounds. Because of the diversity of endogenous substances and the limit of existing technologies, the conventional identification method has some disadvantages on classification and identification of potential endogenous substances. While the PSM utilized the detected mass and the MDV combined with fragmentation rules of substances under different collision energies to assign classification of unknown substances. We analyzed unknown substances involved in the literatures, such as unknown compounds with m/z 294.1555[24] and m/z 324.0724[25]. There are not detailed information on mass spectrometry provided in the literatures, they can only be rough classification that they belong to Type I. The PSM requires a lot of work on structure prediction of unknown substances. Through assigning classification of unknown compounds and utilizing the same class of substances with similar chromatographic behavior, the method can further understand unknown substances and contribute to separate substances in future work.

## Conclusion

PSM is established based on multiple databases, spectrum information of different categories of representative compounds, and many studies. This method can rapidly and accurately screen and exclude different categories of isomeric substances that narrow the retrieval range of substances in the database, so it contributes to identifying the target compounds. This scenario shows that it can solve some existing key problems of metabolomics studies, such as the tedious and poor accuracy identification process of endogenous substances. Meanwhile, this method can contribute to affiliated categories of unknown compounds, which provides a new clue to the discovery and separation of unknowns. Apart from importantly promoting metabolomics study, PSM also significantly guides in the classification, identification, discovery, and separation of lipidomics, petrochemicals, and traditional Chinese medicine analysis.

## Acknowledgements

This project was supported by the National basis Research Program of China (973 Program) (2011CB505300, 2011CB505302) and the National Natural Science foundation of China (No. 81273998).

## Notes and references

a *Tianjin State Key Laboratory of Modern Chinese Medicine, School of Traditional Chinese Materia Medica, Tianjin University of Traditional Chinese Medicine, 312 Anshan west Road, Tianjin 300193, China.*

* *Author for correspondence: Tel and Fax number: +86-22-59596223; E-mail: tianjin_tcm001@sina.com.*

## References

1. A. H. Zhang, H. Sun, G. L. Yan, P. Wang, Y. Han and X. J. Wang. *Cancer Lett.*, 2014, **345**, 17-20.
2. I. Tzoulaki, T. M. D. Ebbels, A. Valdes, P. Elliott and J. P. A. Ioannidis. *Am J Epidemiol.*, 2014, **180**, 129-139.
3. J. H. Wang, T. T. Christison, K. Misuno, L. Lopez, A. F. Huhmer, Y. Y. Huang and S. Hu. *Anal. Chem.*, 2014, Doi: 10.1021/ac500951v.
4. N. Szoboszlaia, X. H. Guo, O. Ozohanics, J. Oláh, A. Gömöry, V. G. Mihucz, A. Jeney and K. Vékey. *Anal. Chim. Acta.*, 2014, **819**, 108-115.
5. P. Liu, J. N. Duan, P. J. Wang, D. W. Qian, J. M. Guo, E. X. Shang, S. L. Su and Y. P. Tang. *Mol. BioSyst.*, 2013, **9**, 77-87.
6. T. Sun, S. Pawlowski and M. E. Johnson. *Anal. Chem.*, 2011, **83**, 6628-6634.
7. W. Yang, Y. H. Chen, C. Xi, R. P. Zhang, Y. M. Song, Q. M. Zhan, X. F. Bi and Z. Abliz. *Anal. Chem.*, 2013, **85**, 2606-2610.
8. H. Y. Li, L. V. DeSouza, S. Ghanny, W. Li, A. D. Romaschin, T. J. Colgan and K. W. Michael Siu. *J. Proteome Res.*, 2007, **6**, pp 2615-2622.
9. R. Amathieu, M. N. Triba, P. Nahon, N. Bouchemal, W. Kamoun, H. Haouache, J. C. Trinchet, P. Savarin, L. L. Moyec and G. Dhonneur. *Plos One*. 2014, **9**, e89230.
10. J. Guo, M. L. Zhang, C. S. Elmore and K. Vishwanathan. *Anal. Chim. Acta.*, 2013, **780**, 55-64.
11. L. W. Qi, H. Y. Wang, H. Zhang, C. Z. Wang, P. Li and C. S. Yuan. *J Chromatogr A.*, 2012, **1230**, 93-99.
12. S. Qiao, X. W. Shi, R. Shi, M. Liu, T. Liu, K. R. Zhang, Q. Wang, M. C. Yao and L. T. Zhang. *Anal Bioanal Chem.*, 2013, **405**, 6721-6738.
13. M. S. Zhu, L. Ma, D. L. Zhang, K. Ray and W. P. Zhao. *Drug Metab Dispos.*, 2006, **34**, 1722-1733.
14. M. Grabenauer, W. L. Krol, J. L. Wiley and B. F. Thomas. *Anal Chem.*, 2012, **84**, 557-5581.
15. T. Xie, Y. Liang, H. P. Hao, J. Y. A, L. Xie, P. Gong, C. Dai, L. S. Liu, A. Kang, X. Zheng and G. J. Wang. *J Chromatogr A.*, 2012, **1227**, 234-244.
16. Y. Shi, B. Bajrami and X. D. Yao. *Anal Chem.*, 2009, **81**, 6438-6448.
17. R. C. Murphy, J. A. Hankin and R. M. Barkley. *J. Lipid. Res.*, 2009, **50**, S317-S322.
18. D. W. Hill, T. Z. Kertesz, D. Fontaine, R. Friedman and D. F. Grant. *Anal. Chem.*, 2008, **80**, 5574-5582.
19. J. Qu, Q. L. Liang, G. A. Luo and Y. M. Wang. *Anal. Chem.*, 2004, **76**, 2239-2247.
20. Y. B. Li, X. X. Zhang, H. F. Zhou, Y. M. Wang, S. M. Fan, L. Zhang, L. Ju, X. Wu, H. Y. Wu and Y. J. Zhang. *Rsc. Adv.*, 2014, **4**, 8260-8270.
21. X. X. Zhang, Y. B. Li, H. F. Zhou, S. M. Fan, L. Zhang, Y. M. Wang, Z. Z. Zhang, L. Wang and Y. J. Zhang, *J. Pharm. Biomed. Anal.*, 2014, **97**, 151-156.
22. Y. Y. Zhao, X. L. Cheng, J. H. Cui, X. R. Yan, F. Wei, X. Bai and R. C. Lin, *Clinica. Chimica. Acta.*, 2012, **413**, 1438-1445.
23. Y. Y. Zhao, L. Zhang, F. Y. Long, X. L. Cheng, X. Bai, F. Wei and R. C. Lin. *Chem-Biol Interact.*, 2013, **201**, 31-38.
24. F. X. Zhang, Z. H. Jia, P. Gao, H. W. Kong, X. Li, J. Chen, Q. Yang, P. Y. Yin, J. S. Wang, X. Lu, F. M. Li, Y. L. Wu, G. W. Xu. *Talanta.*, 2009 ; **79**: 836-844.
25. W. D. Dai, C. Wei, H. W. Kong, Z. H. Jia, J. K. Han, F. X. Zhang, Z. M.Wu, Ya. Gu, S. L. Chen, Q. Gu, X. Lu, Y. L. Wu, G. W. Xu. *J Pharm Biomed Anal.*, 2011; 56: 86-92.

Cite this: DOI: 10.1039/c0xx00000x

www.rsc.org/xxxxxx

**Table 1.** Main class of endogenous substances having fragmentation patterns in the mass spectrum in different scan modes.

| NO. | Analyte | Classification | Subclass | Monoisotopic | Scan | Precursor | Fragment ion |
|---|---|---|---|---|---|---|---|
| | | | | Mass | mode | ion | |
| 1 | Adenine | Nucleic acids | Bases | 135.0545 | + | 136.06 | 119.03 |
| 2 | Uracil | | | 112.0273 | + | 113.03 | 70.02 |
| 3 | β-thymidine | | Nucleosides | 242.0903 | + | 243.09 | 127.05 |
| 4 | Uridine | | | 244.0695 | + | 245.07 | 113.03 |
| 5 | Estrone | Steroid hormones | Estrogens | 270.162 | - | 269.15 | 145.04 |
| 6 | 17β-estradiol | | | 272.1776 | - | 271.16 | 145.04 |
| 7 | Testosterone | | Androgens | 288.2089 | + | 289.21 | 109.06 |
| 8 | Progesterone | | Progestagens | 314.2246 | + | 315.23 | 109.06 |
| 9 | Hydrocortisone | | Glucocorticoids | 362.2093 | + | 363.21 | 109.06 |
| 10 | Aldosterone | | Mineralocorticoid | 360.1937 | + | 361.20 | 109.06 |
| 11 | L-Histidine | Common amino acids | | 155.0695 | + | 156.07 | 110.07 |
| 12 | L-Lysine | | | 146.1055 | + | 147.11 | 84.08 |
| 13 | Glucose | Monosaccharides | | 180.0634 | - | 179.05 | 161.04 |
| 14 | D-Fructose | | | 180.0634 | - | 179.05 | 161.04 |
| 15 | Succinic acid | Organic acids | | 118.0266 | - | 117.01 | 73.02 |
| 16 | LPC(18:0) | Phospholipids | | 523.3637 | + | 524.36 | 184.07 |

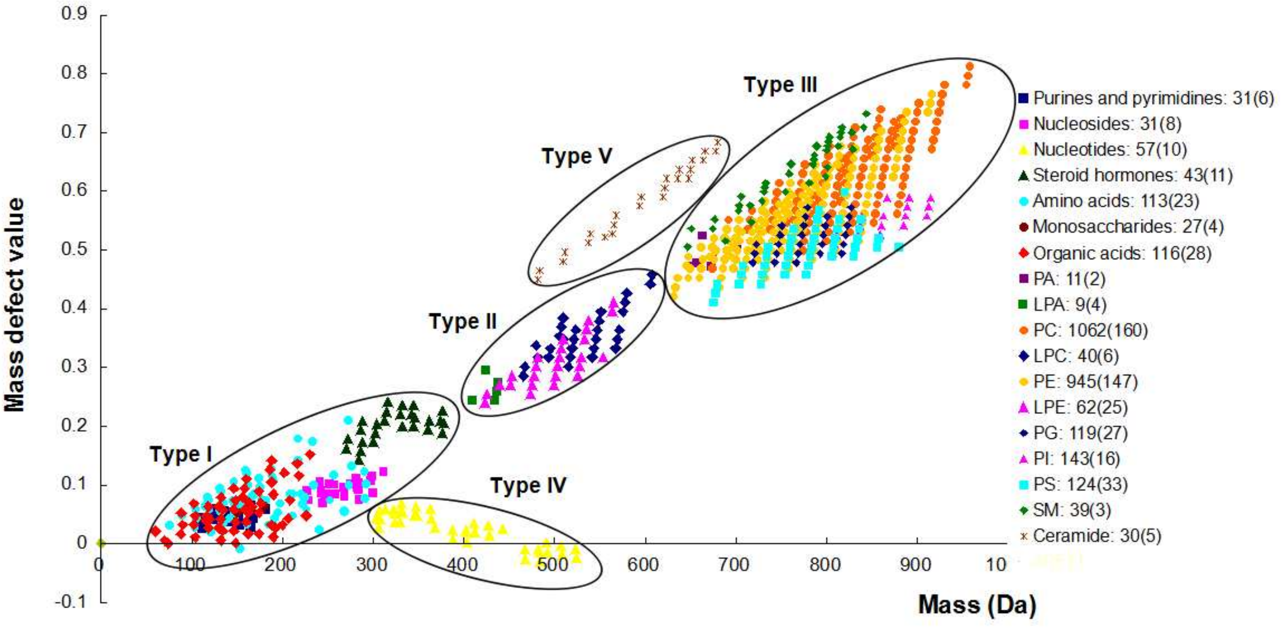


**Fig. 1.** Mass defect filtering plot of different classes of endogenous substances though HMDB and KEGG databases. Type I is amino acids, organic acids, monosaccharides, bases and nucleosides of nucleic acids, and steroid hormone; Type II is LPCs, LPEs, and LPAs; Type III is PCs, SMs, PEs, PAs, PGs, PIs, and PSs; Type IV is nucleotides; and Type V is ceramides. For example, PC:1062(160) in the MDF plot requires the search 1062 PCs, which includes 160 groups of isomerides.

Cite this: DOI: 10.1039/c0xx00000x

www.rsc.org/xxxxxx

ARTICLE TYPE

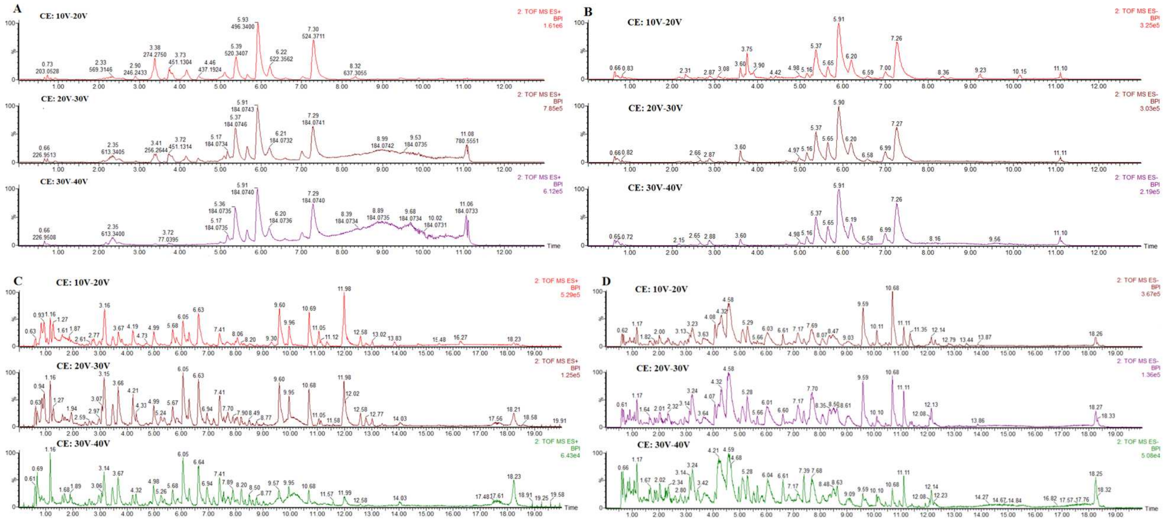


**Fig. 2.** A and B are BPI chromatograms of rat plasma samples in the positive and negative mode combined different collision energy (10v-20, 20v-30, 30v-40eV), respectively. C and D are BPI chromatograms of rat urine samples in the positive and negative mode combined different collision energy (10v-20, 20v-30, 30v-40eV), respectively.

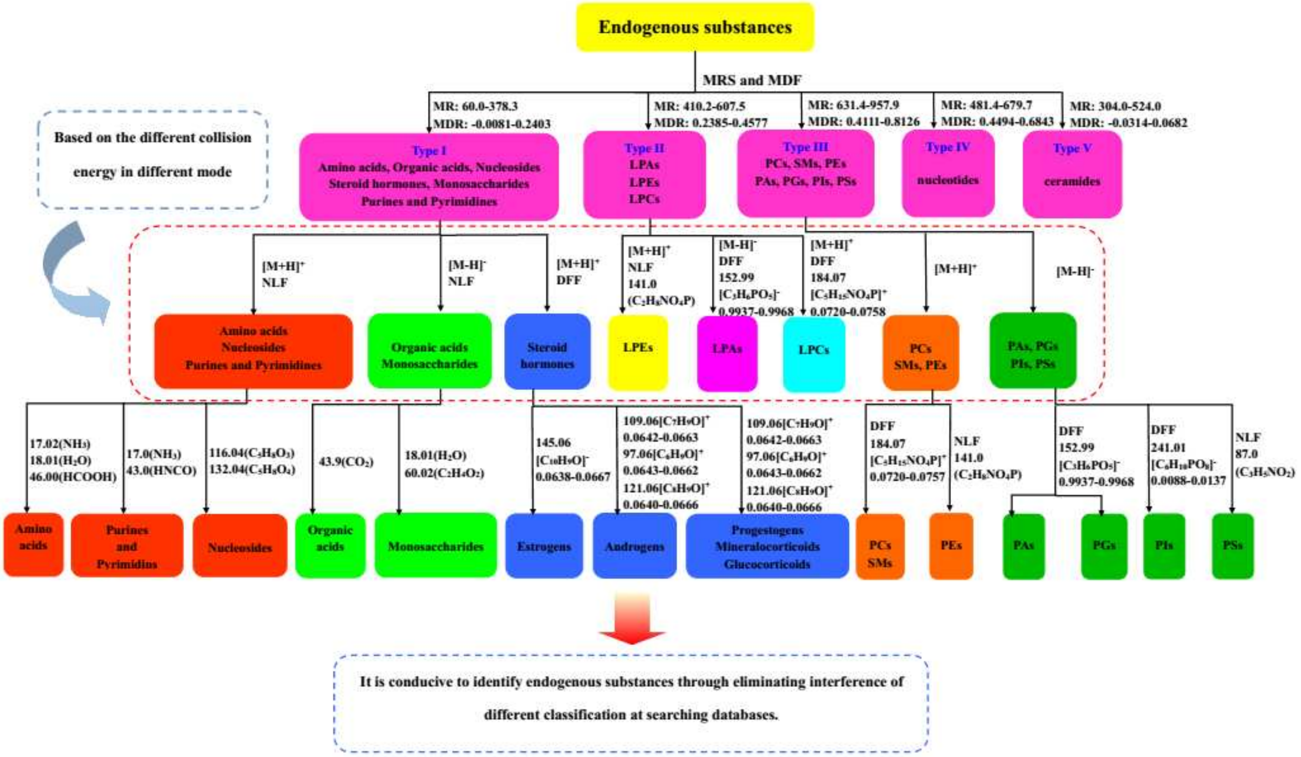


**Fig. 3.** Schematic of the PSM.

5