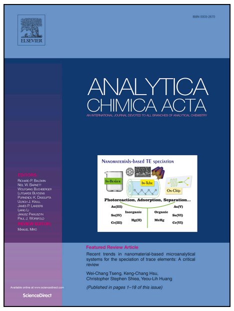
Accepted Manuscript



Chemometric methods in data processing of mass spectrometry-based metabolomics: A review

Lunzhao Yi, Naiping Dong, Yonghuan Yun, Baichuan Deng, Dabing Ren, Shao Liu, Yizeng Liang

PII:	S0003-2670(16)30186-6
DOI:	10.1016/j.aca.2016.02.001
Reference:	ACA 234404
To appear in:	*Analytica Chimica Acta*
Received Date:	25 August 2015
Revised Date:	28 January 2016
Accepted Date:	1 February 2016

Please cite this article as: L. Yi, N. Dong, Y. Yun, B. Deng, D. Ren, S. Liu, Y. Liang, Chemometric methods in data processing of mass spectrometry-based metabolomics: A review, *Analytica Chimica Acta* (2016), doi: 10.1016/j.aca.2016.02.001.

Chemometric methods in data processing of mass spectrometry-based metabolomics: A review

Lunzhao Yi [a]*, Naiping Dong[c], Yonghuan Yun[b], Baichuan Deng[d], Dabing Ren[a], Shao Liu[e], Yizeng Liang[b]

[a]*Yunnan Food Safety Research Institute, Kunming University of Science and Technology, Kunming, 650500,China*
[b]*College of Chemistry and Chemical Engineering, Central South University, Changsha, 410083, China*
[c]*Department of Applied Biology and Chemical Technology, The Hong Kong Polytech nic University, Hong Kong, 999077, China*
[d]*College of Animal Science, South China Agricultural University, Guangzhou, 510642, China*
[e]*Xiangya hospital, Central South University, Changsha, 410008, China*

*Correspondence to: Lunzhao Yi, Yunnan Food safety research institute, Kunming University of Science and Technology, Kunming, 650500, China. Tel.: +86 871 65920302. E-mail address: yilunzhao@kmust.edu.cn.

Abstract

This review focuses on recent and potential advances in chemometric methods in relation to data processing in metabolomics, especially for data generated from mass spectrometric techniques. Metabolomics is gradually being regarded a valuable and promising biotechnology rather than an ambitious advancement. Herein, we outline significant developments in metabolomics, especially in the combination with modern chemical analysis techniques, and dedicated statistical, and chemometric data analytical strategies. Advanced skills in the preprocessing of raw data, identification of metabolites, variable selection, and modeling are illustrated. We believe that insights from these developments will help narrow the gap between the original dataset and current biological knowledge. We also discuss the limitations and perspectives of extracting information from high-throughput datasets.

Keywords: metabolomics; chemometrics; biomarker; identification of metabolites; data preprocessing; modeling

Contents

1. Introduction

Metabolomics refers to the comprehensive and quantitative analysis of metabolites and aims to gather as much metabolic information as possible from a biological system [1]. It is a reproducible and efficient method that can directly reflect biological events. Metabolomics has recently been upgraded from a promising concept to a widespread and valuable biotechnology. Two modern analytical platforms, namely, nuclear magnetic resonance (NMR) and mass spectrometry (MS), have become the methods of choice for metabolic analysis and are used to generate massive amounts of data to answer various biological questions in metabolomics [2-4].

Improved analytical technologies have gradually caused metabolomics datasets to become larger with more intricate inner structures [5]. Thus, the coverage of metabolomics becomes more comprehensive but will consequently demand more advanced chemometric methods [6]. Metabolomics is either targeted or untargeted. In the targeted approach, specific metabolites of known identity are profiled; good quantitative precision is easily obtained. One disadvantage of this approach, however, is its limitation in terms of the breadth of analysis. The dataset of the target approach is simple. Data analysis often focuses on variable selection and modeling. Untargeted metabolomics aims to simultaneously measure of as many metabolites as possible in a biological specimen. Often, the chemical identities of the MS-resolved peaks are not known a priori, and significant chemical/spectral analysis must be performed to identify the metabolites. Deconvolution and normalization of complex spectra in biological samples is therefore critical for this type of datasets.

The raw data from metabolomics presents a gold mine of information [7]. To ensure that the metabolic information is of valuable knowledge, considerable data analysis is required. Chemometrics has become a crucial and dedicated tool for extracting valuable information from data; it presents a complete theory and methodology for every step of metabolomics research, including sampling, experiment design, data pre-processing, metabolite identification, variable selection, and modeling. Chemometrics has thus become one of the cornerstones of metabolomics. However, major changes in the dimensionality and complexity of datasets lead to a significant shift in knowledge discovery. The complexity of metabolomics also presents great challenges on chemometrics to deal with such massive high-dimensional data [6].

Several review papers and guide books on metabolomics have been published [8-10], and these works have provided informative and valuable guidance for researchers. Insights into metabolomics experimental skills, including sample preparation and metabolite analysis, have also been revealed [11]. In this review, we describe recent advances in chemometric methods for data analysis of metabolomics. This review provides a brief but broad overview of the developed methods, the challenges remaining in the data processing of metabolomics, especially those generated by MS, and perspectives on this topic. Various aspects, including raw data pre-processing, metabolite identification, variable selection and modeling, are discussed. The flowchart of data processing in metabolomics is shown in Figure 1.

Insert Figure 1

2. Critique and discussion

2.1 Pre-processing of raw data

Analytical instruments do not provide clean and comparable lists of metabolites. Raw data must be processed to generate a practicable data matrix in a variety of ways [12]. The key step is eliminating the variance and bias in the data analysis to reduce the complexity and enhance metabolically significant signals [13]. Consequently, several algorithms have been developed and multiple open source programs have been applied to process raw MS data acquired on liquid chromatography-mass spectrometry (LC-MS) or gas chromatography-mass spectrometry (GC-MS). Among these, XCMS (https://xcmsonline.scripps.edu/) [13, 14], MZmine (http://sourceforge.net/projects/mzmine/) [15, 16], OpenMS (http://open-ms.sourceforge.net/) [17], and MetAlign (http://www.metalign.nl) [18] have attracted particular attentions for their practicability and effectiveness. Most members of the research community of metabolomics work with these tools, and new programs, such as MetSign [19], MSFACTs [20] and MetaboliteDetector [21] have been steadily developed to increase the quality and efficiency of data preprocessing.

Most of these tools are freely available. Furthermore, through these tools, the exchange of algorithms and data within the community is convenient. In generally, tools for raw data preprocessing include four basic modules, namely, noise filtering and baseline correction, peak detection and deconvolution, alignment, and normalization. In the following sections, we will introduce different chemometric algorithms and strategies for these modules.

2.1.1 Noise filtering and baseline correction

Noise filtering is designed to separate component signals from the background originating from the chemical matrix or instrumental interference, remove measurement noise or baseline distortions [9]. Conventionally, during baseline correction of one-way data (e.g. a chromatogram or mass spectrum), the two ends of a signal peak are manually identified by analysts and piecewise linear approximation is then applied to fit a curve as the baseline [22]. However, this procedure is time-consuming, and its accuracy highly depends on the user’s operating skills. Thus, numerous algorithms have been developed for better estimation of the baseline. Two powerful algorithms, automatic two-side exponential baseline correction algorithm (ATEB) [23]and adaptive iteratively reweighted penalized least squares (airPLS) [22], were recently developed by Liang’s group. These algorithms can automatically and effectively remove the baseline, regardless of whether it is linear or non-linear. Furthermore, unlike methods that require peak detection, these very fast and robust algorithms do not require intervention experience and prior knowledge.

For MS-based datasets, the methods for removing random noise are typically implemented by traditional signal processing techniques in chemometrics. Noise filtering of LC-MS data is more complicated than that of GC-MS data because chemical and random noises are both included in the former. Chemical noise is typically induced by molecules in buffers and solvents and can be especially strong at the beginning and the end of the elution [24]. This type of noise causes a shift in the baseline in the intermediate mass range of LC-MS spectra. To resolve this problem, several filtering methods have been proposed. For example, Haimi *et al*. fitted the baseline by first segmenting a spectrum and performing linear regression through the lowest points of the smoothed spectrum segments [25]. In addition, baseline removal has also been approached by estimating the background from a two-dimensional intensity image and then removing it with two orthogonal (retention time and *m/z*) one-dimensional passes [26].

2.1.2 Peak detection and deconvolution

The purpose of peak detection and deconvolution is to identify and quantify the signals corresponding to the molecules (e.g., the metabolites) in a sample [12]. This step is fundamental for downstream data analysis, such as profile alignment or biomarker identification, and can significantly reduce the complexity of the data [9]. However, given the complexity of the signals and the multiple sources of noise in data, automatic identification of the noise from compound signals is very difficult. The threshold between noise and a signal is difficult to specify, especially when detecting peaks with low-response values.

A peak detection method can identify the true signals correctly and avoid false positives. Unfortunately, high response values do not always guarantee real peaks because some sources of noise can also produce high signals. Conversely, low peaks may correspond to real signals. Therefore, constraints on the peak shapes and criteria of minimal intensity, area or signal-to-noise are widely applied to distinguish real peaks from noise. Several parameters must generally be adjusted to match the characteristics of the MS-based data. Traditionally, peak detection algorithms follow two strategies: derivative techniques or matched filter response.

Derivative-based peak detection methods make use of the fact that the first derivative of a peak will have a positive-to-negative zero-crossing at the local maxima of a peak [27]. Derivative-based methods commonly require increasingly elaborate pre-processing to prevent compounding noise effects [28, 29]. A slope threshold on is often imposed to avoid false positives.

Matched filter methods may become progressively sophisticated as the data complexity increases. One may apply a threshold in the response function to determine the location of chromatographic peaks when applied to chromatographic data by assuming a Gaussian peak shape [30]. A number of popular and open-source software packages, such as XCMS [13] have been developed. XCMS includes three steps: binning, signal determination, and filtering. One weakness of the initially proposed method in XCMS, however, is that the peaks can sometimes be alternatively assigned to two adjacent *m/z* bins. One potential solution to this problem involves combining adjacent extracted ion chromatograms, which represent the analyses of interest. However, this algorithm cannot resolve pairs of co-eluting peaks that fall within half of the *m/z* bin. The developers of XCMS software thus added another algorithm called centWave in later version [31]. The centWave algorithm collects regions containing potentially interesting masses in the raw data and applies continuous wavelet transformation (CWT) and, optionally, Gauss-fitting for chromatographic peak resolution. To circumvent the problems during binning, an alternative fast-computing approach is used in centWave based on the mass accuracy deviation and expected chromatographic peak width. Then, CWT is performed to detect all possible chromatographic peaks. Subsequent filtering is employed to remove candidate peaks in which number of m/z centroids is less then specified threshold. In addition, CWT is also applied to build a robust pattern-matching method for MS peak detection and can be directly used to the raw spectrum. By identifying peaks and assigning a signal-to-noise ratio in the wavelet space according to the two-dimensional CWT coefficient matrix, the pattern matching problem is simplified. Thus, issues surrounding the baseline correction are simultaneously resolved, and the preprocessing steps, such as noise filtering and baseline correction, are not required before peak detection [32].

Selecting an optimal threshold for the above mentioned two strategies is a difficult problem but of essential importance that has been thoroughly discussed in various peak detection approaches [27, 33, 34], whereas no general consensus is reached. Some algorithms have recently been developed based on Bayesian inference [35, 36]. These algorithms make use of chromatographic information (i.e., the expected width of a single peak and the standard deviation of baseline noise), which is regarded as prior information. Finally, the probability of a signal being a peak is estimated, based on some theories or hypotheses, such as the statistical overlap theory [36].

In the high-throughput analysis of metabolites, overlapping peaks are ineluctable. This problem can be resolved by two-dimensional data resolution methods that have been well developed and theorized by the chemometrics community using matrix computation combined with characteristics of spectral data [37-39]. Specifically, multivariate curve resolution-alternating least squares (MCR-ALS) [40] has been extended to processing LC-MS data [41] and shown to be more robust than XCMS [42]. The overlapping peaks can also be resolved by mass spectral deconvolution. Automated mass spectral deconvolution and identification system (AMDIS, NIST) and commercially available tools, such as deconvolution reporting software (DRS, Agilent), AnalyzerPro (SpectralWorks), and ChromaTOF® (LECO), are developed for processing GC-MS data. Most recently, Oliver Fiehn *et al.* [43] proposed an open-source software pipeline, called MS-DIAL, for data-independent acquisition (DIA) - based metabolite identification and quantification by mass spectral deconvolution. MS-DIAL resolves entangled MS/MS spectra by a two-step process: precursor-peak spotting followed by MS/MS-level deconvolution. With this software, DIA can provide high efficacy and accuracy for metabolome coverage.

2.1.3 Alignment

Alignment of detected features in different samples aims to remove shifts among samples for a given signal to guarantee downstream extraction of useful information. Thus far, several alignment techniques have been developed to minimize run-to-run shifts [44]. To make them applicable to chromatographic systems coupled with sophisticated detection instruments, e.g., LC-MS, which have yielded large amounts of two-dimensional data, the dimensionality must be reduced. The reduction could be achieved by generating integrated peak areas or total ion chromatograms (TICs). For one-dimensional data (such as TICs), some kinds of time alignment procedures could be employed as a useful method for tackling this problem of retention time shifts [45]. Examples of these procedures include correlation optimized warping (COW) [46], and dynamic time warping (DTW) [47], recursive alignment by fast Fourier transform (RAFFT) [48]. COW requires large execution times and memory when dealing with huge hyphenated datasets. Artifacts often appear in the fingerprints aligned by DTW because signals are often over-warped when signals are recorded by a mono-channel detector. RAFFT efficiently accelerates the alignment procedure by fast Fourier transform cross-correlation. However, RAFFT may distort the shapes of peaks because it does not consider the peak information when moving segments; this technique only considers the insertion and deletion of data points only at the start and end of segments, which may introduce artifacts and remove peak points. Nonlinear retention time shifts often exist for a real sample; thus, a multi-scale peak alignment (MSPA) approach has been proposed. MSPA involves iteratively dividing a chromatogram into smaller segments to solve the problem of nonlinear retention time shifts in alignment. FFT cross correlation is used to estimate candidate shifts and gradually align peaks step by step. A simple example of the application of MSPA method is demonstrated in Figure 2. The retention time shifts of GC-MS TICs in different samples are successfully removed. Other algorithmic alternatives, such as kernel density [13], component-resolving algorithms [49], and progressive clustering [50], among others, exist. Besides, another alignment methods attempt to integrate peak areas. Although time-consuming and meticulous, this approach is considered as the process of "data cleaning" because the retention time shift, noise pollution, and background shift are cleared simultaneously.

Insert Figure 2

During dimension reduction, loss of information is inevitable. Addressing this issue involves modeling of the high-dimensional data by multi-way analysis methods, which maintain the so-called two dimensional advantages (e.g., mass spectral information of metabolites). For example, the alignment method by Prakash *et al*. [51] and the ChromAlign method [52] both use the raw high-way data. First, these algorithms construct similarity score matrix for similar spectra between two experimental runs. Dynamic programming is applied to find an optimal path through the matrix and define the mapping of paired spectra. In the method proposed by Pierce *et al*. [53], a piecewise single dimension retention time alignment algorithm is applied to align two-dimensional data. In the continuous profile model (CPM), the two-dimensional data is divided into four *m/z* bins as opposed to the alignment of only a single TIC [54]. In addition, some algorithms align the two-way retention time shift more comprehensively, such as the algorithm using a novel indexing scheme [53]. This type of algorithms aligns the fingerprints in different dimensions simultaneously, thereby preserving the separation information in both dimensions.

2.1.4 Normalization

Normalization removes confounding variations attributed to experimental sources, such as analytical noise or experimental bias, and retains relevant variations attributed to biological events [12]. If the signal of majority of metabolites is stable, simple and efficient normalization could be achieved by calculating the relative ratio of the abundance of analytes to all other peaks, such as the unit norm and median intensities normalization [55]. However, the assumption of negligible overall concentration changes is difficult to satisfy; the total concentrations of analytes may be considerably changed because of laboratory system errors and differences among large scale biological experiments. In this case, scaling based on the total chromatogram may seriously distort the data.

Compounds with lower concentrations will be easily altered by analytical noise. To allow the comparison of different metabolites, scaling is required. Autoscaling (1/SD) is the most popular normalization method used in metabolomics; in this method, each variable has equal (unit) variance by multiplying with the inverse of standard deviation (SD). Pareto (1/sqrt(SD)) is softer than autoscaling and can increase the importance of low abundant compounds without significantly amplifying the noise.

During data analysis, researchers tend to assume that the total variations originating from sampling, analytical measurements, and biological events are with equal standard deviations and symmetrically around zero [56]. However, this assumption is not satisfied in many cases. Biological effects related to concentration alterations could vary dramatically for different metabolites. Variations related to certain metabolites are considered heteroscedasticity, which could be detrimental to observations of a particular biological situation [56]. A mathematical transformation, such as log transformation [57] or power transformation [58] is helpful to correct the skewed data before modeling. When the relative standard deviation is constant, a log transformation can perfectly remove heteroscedasticity [57]. However, log transformation presents a serious drawback: the transformation approaches minus infinity when the values are transformed as they approach zero. Power transformation does not have the near-zero artifacts and yields results similar to those of log transformation.

Another sophisticated strategy for normalization is the internal standards (ISs) method, e.g., isotopically labeled internal standards, and quality control (QC) samples in each data acquisition procedure [59]. Comprehensive and representative IS-based normalization is based on a key assumption that the variance exhibited by ISs solely comes from a component with a systematic error. But, a single IS cannot estimate the systematic error of a complex biological matrix. Multiple ISs work better in this case. Further, IS use must aim to decrease the risk of cross-contribution (CC) which can cause serious loss of information, especially when the interfering analytes are related to the factors of interest in metabolomic datasets. If the masses used for quantifying the IS are carefully selected, this problem can be solved easily [60]. However, this attempt is nontrivial in metabolomics research because the biological sample is too complex. Prediction of which ions will produce cross-interference is difficult. Redestig *et al.* presented an effective normalization algorithm that could compensate for systematic CC effects and improve the normalization of mass spectrometry-based metabolomics data [61]. To image the global variability of a measurement system, performing QC before normalization is recommended when visualizing the data by PCA. A QC is a pool of several individuals having similar characteristics. The studied samples are compared with QCs to evaluate their variability. In multivariate statistical analysis, such as PCA, QC samples should appear closely on the scores plot, which indicates that the analytical system has good reproducibility [62].

2.2 Identification of metabolites

Confidently identifying metabolites from MS spectra data has been generally recognized as a significant challenge in the metabolomics community, especially in untargeted analysis, because of the biochemical diversity of metabolites. Given the benefits of advanced computational techniques and methods, advanced mass spectrometry instrumentation, the wealth of knowledge on ion fragmentation, and well-established databases and libraries, especially fruitful works in the past decade, metabolite identification can cover unknowns with reasonable accuracy and could be performed in a high-throughput manner. A variety of overviews have been published on this topic, including basic concepts in compound identification, comprehensive summaries of different identification strategies [63, 64], instructions for practical use [65], and guidelines for beginners of mass spectrometry [66]. Thus, we are going to briefly introduce currently available algorithms and tools valuable for metabolite identification using MS in this section.

2.2.1 Metabolite identification using GC-MS

GC-MS has been routinely used in metabolomics with mature protocols. Great effort has been made to interpret MS spectra from electron impact (EI) ion sources. The most frequently adopted and reliable method for this is library search, where each experimental MS spectrum is compared with the reference MS spectra in the mass spectral library and the similarity score is calculated for each match. The corresponding library compound gaining the highest similarity score is theoretically considered as the one that generates this experimental spectrum. The commonly adopted mass spectral libraries are listed in Table 1. The main factors that influence search results include the quality of the experimental MS spectra, the size of the mass spectral library, and the similarity score calculation algorithms used [67]. From the arithmetic point of view, the method for calculating the similarity score is the most important factor to consider because the quality of the MS spectra significantly depends on the experiment, and the libraries are generally commercially available and thus cannot be freely configured by users and remain relatively small in size. Previous investigations showed that the most robust similarity score calculation method is the dot product using square-rooted mass spectral intensities [68]. However, no comprehensive comparative investigation is performed for high through-put metabolite identification.

Insert Table 1

Given the complexity of metabolites and their EI-MS spectra, such as the existing of isomers and co-eluted components, a target compound does not ideally gain the highest similarity score but is generally located at a higher rank (e.g., second or third rank, or higher) in the hit list. This approach always requires careful manual checking. Therefore, taking other information, such as the retention index (RI, e.g. Kovat's retention index) of a target compound, into consideration will be very helpful [69, 70]. RI is a structurally and physicochemically specific indicator that can effectively differentiate compounds having similar mass spectra. Actually, this indicator and the EI-MS spectrum comprise the widely accepted mass spectral tag (MST) in metabolomics and organize the Golm Metabolome Database(GMD) [71-73] and

BinBase/FiehnLib [74]. The NIST standard reference database includes a large number of RI values. Another improvement, especially in the case of co-elution, can be achieved by mass spectral deconvolution or two-dimensional data resolution methods (see Section 2.1.2). As GC-MS instruments with mass analyzers capable of high resolution and accurate mass measurement are now available; the majority of the false matches can also be filtered by considering the accurate masses of the fragments [75].

The methods independent of a mass spectral library are to learn the structural features of compounds from their experimental mass spectra and then deduce unknown structures from the features of a given spectrum according to previously constructed learning models. This can be achieved in two ways. The first one involves exhaustion of all possible isomers according to the molecular mass extracted from MS spectra by a structure generation module (e.g., MOLGEN [76] and OMG [77]) and retention of the structures that best explain the spectrum according to fragmentation rules. Machine learning algorithms are generally adopted in this procedure to determine whether a substructure is present in the unknown compound. This step can filter out a large number of isomers that do not contain the identified substructures [78]. MOLGEN-MS [79] and MassLib have been developed for this purpose. The web-based algorithm embedded in GMD employs decision trees to predict the 166 most common functional groups in metabolites after training known metabolites in GMD with the corresponding mass spectra data and retention indices [80], thereby providing invaluable information for inferring the structures of unknown metabolites. The second approach is based on library search results under the assumption that similar structures have similar spectra. Possible substructures of unknown compounds can be deduced from library compounds with the top similarity scores [81].

An alternative series of methods directly predict mass spectra for input molecules. Based on the wealth of knowledge on ion fragmentation and aided by advanced computational technologies, accurate prediction of mass spectra has become feasible. Mass Frontier (Thermo Scientific), one of the most commonly adopted software for structure elucidation, uses the HighChem Fragmentation Library, which stores approximately 31,000 fragmentation mechanisms to predict and interpret experimental mass spectra. ACD/MS Fragmenter (ACD/Labs), which is also very powerful for MS spectrum prediction, has gained popularity in the metabolomics community. The freely available tool Mass Spectrum Interpreter, which was released by NIST, uses thermochemical kinetics of general fragmentation reactions summarized from known fragmentation rules to predict mass spectra. Among these powerful methods, a common difficulty is that they cannot effectively extract correct structures from their isomers, as pointed out after comparing different tools [82]. However, improvements can be made by combining different tools [83]. In addition to the above methods, by adopting advantages of high resolution GC-MS, unknown compounds can be putatively identified from accurate *m/z* provided by chemical ionization, *in-silico* predicted retention index and fragmentation patterns without requiring any mass spectral library [84, 85]. This trend is analogous to identifying metabolites in high resolution LC-MS, as will be shown below. A practical guide for small molecule structure elucidation with several strategies that differ from above mentioned computational methods can be found in Ref.[86].

2.2.2 Metabolite identification using LC-MS

For LC-MS, identifying metabolites from MS spectra is not amenable because of the variation of experimental settings, such as chromatographic conditions and mass spectrometry parameters [87]. This step becomes even more serious for discovering unknowns from large and complex metabolite space. Additionally, the fragmentation mechanisms during ionization in the LC-MS platform under various activation energies are still unclear. These factors make the confident interpretation of MS spectra derived from different LC-MS and LC-MS^n platforms a significant challenge. Fortunately, recent active studies have made remarkable advances in metabolite identification and several tools and various databases are publically available (see Table 1 and 2). In general, currently available tools are developed based on two aspects of LC-MS data: accurate mass with other information like isotopic distribution and MS/MS spectra.

Insert Table 2

2.2.2.1 Structure inference by accurate mass combined with other information

The ability to accurately measure *m/z* is one of the most important features of high-resolution mass spectrometry, which has greatly facilitated the whole MS data analysis workflow. The accurate mass calculated from determined *m/z* is generally the first step [66] because it is the simplest and most straight-forward. The formula generation method or the search of a large compound database or metabolism network can be adopted. For formula generation, all combinations of predefined elements with constraints of element number and mass range are exhausted. A number of tools commercially or freely available have been developed to assist this (see Table 2). As expected, very large number of candidate formulas will be generated, especially for a relatively large molecular mass. This phenomenon makes it impracticable to obtain a single assignment of formula to each *m/z* solely based on the accurate mass. Thus, defining the rules to filter out false positives becomes nontrivial.

Among all the developed rules, similarity checking in isotopic distribution is commonly accepted as the most critical criterion. Majority of the spurious formulas could be rejected under this checking [88, 89]. Theoretically, each elemental composition or formula has a unique isotopic distribution because different elements have distinct isotopic abundance distributions in nature. Thus, by comparing the instrument-determined isotopic distribution to the simulated one, the formula candidates can be ranked, with the top ones being the most similar via so called spectral comparison [90] or rejected if the relative isotopic abundances (RIA) between the two distributions are unacceptably different. The exploration to precisely simulate isotopic distribution has been undertaken for decades and several tools are now freely available [91]. If the resolution of an MS instrument is high enough, formulas can be exclusively identified from the RIA of a single element. This strategy is now extended and confirmed with higher-resolution instruments for high-throughput metabolomics analysis [92]. However, high RIA measurement errors can appear in peaks with a low signal-to-noise ratio (S/N), low *m/z*, and the presence of co-eluting species [93-96]. These factors will terribly mislead the identification results [94]. Unfortunately, the systematic evaluation of the influence of RIA measurement error on formulae inference is not performed. A suggestion for eliminating this influence can be setting a larger error tolerance during comparison [95]. Whereas cautions still should be proceeded with when using RIA to identify metabolites and additional information is required.

The second rule is to check whether the generated formulas are reasonable as candidates of metabolites. The famous “Seven Golden Rules” was defined after statistically analyzing formulas extracted from Wiley and NIST02 mass spectral database and the Dictionary of Natural Products [88] and has been demonstrated to be an efficient tool in metabolomics. An updated version of these rules is defined recently after analyzing large scale formulas in the PubChem database [97].

Once formulas are determined or ranked, decoding them to known metabolites in LC-MS feature annotation is subsequently performed, typically by searching large chemical substance databases [98, 99]. The databases frequently adopted in metabolomics are listed in Table 1. Further annotation of ion species can be realized by prior biological knowledge from lists of expected metabolites of the analyzed organism. Metabolites in biological samples are biochemically connected (e.g., chemical transformation) rather than randomly mixed [100]. Thus, the metabolite candidates are mapped onto metabolism networks to gain confident identification [101-103]. For example, MI-Pack maps mass spectral peaks onto the KEGG network database [104] and uses the rigidly defined mass error surface of mass differences between substrate-product pairs derived from the database for metabolite identification [103]. Significant reduction of false negatives and false positives is consequently obtained. This approach is advantageous for metabolite identification and mining related subnetworks, which represent the activity or functions of the metabolites, as demonstrated in recent works [105, 106].

Besides mapping ions to molecular databases, mining relationships between extracted ion features to annotate these ions has also been proven to be a highly effective strategy. This approach can be executed because LC-MS can detect ion series (so called satellite ions) of a metabolite generated by fragmentation reactions during ionization, including neutral losses and ions with different adducts [107, 108]. This process can generate an *in silico* ion network that reveals relationships between metabolites, also known as metabolic biotransformation [100]. CAMERA [109], IDEOM [110], and MAIT [111] *etc*. were developed in this manner.

2.2.2.2 Metabolite identification by MS^n

MS^n is a highly effective technique for structure elucidation. As an indispensable part of the LC-MS system, ionized molecules or molecules in the *m/z* range specified by instruments are gradually dissociated into charged or neutral pieces by hard ionization methods such as the collision-induced dissociation (CID). Recording all the charged fragments and precursor ion forms the MS^n spectrum. This MS^n spectrum generation procedure demonstrates that a molecule's structure can be readily deduced from its MS^n spectrum. Moreover, strategies for interpreting GC-MS spectra (e.g., library search or mass spectrum prediction) can be applied in this deduction. Therefore, several MS^n spectral libraries and computational methods for spectral prediction or structure elucidation are developed (Table1 and 2). The experimental conditions (e.g., collision energy) in MS^n analysis are not as standardized as in GC-MS analysis. Furthermore, the sizes of currently constructed libraries are much smaller compared with the whole metabolism or structure databases and other factors [112, 113]. Thus, metabolite identification via spectral library search is not as popular in MS^n analysis as in GC-MS analysis. Consequently, much more studies are focused on developing computational methods to interpret MS^n spectra without querying spectral libraries.

The algorithms employed in currently developed software for computational MS^n can be categorized into three basic approaches, namely, mass spectrum prediction, *in silico* fragmentation, and *de novo* elucidation [114]. Mass spectrum prediction, which is mainly applied for MS^2, has been well studied in EI spectrum interpretation. This process is also a basic and highly important module in peptide identification under hypothesis-driven proteomics. The enormous diversity of small compounds continues to considerably challenge accurate MS^2 spectral prediction. To predict the MS^2 spectrum for a given structure, Mass Frontier extracts all possible reactions that can occur during the fragmentation of this structure from its own fragmentation reaction library to generate rules for the prediction of fragments and intensities. ACD/MS Fragmenter handles spectrum prediction in a similar way. MetISIS uses a machine-learning algorithm to learn CID kinetics from lipid experimental MS^2 spectra to predict lipid spectra *in silico* [115]. A fragment ion prediction algorithm embedded in MyCompoundID website (http://www.mycompoundid.org/) adopts a “chopping” program to predict the bond cleavage of metabolites to generate theoretical MS^2 spectra for database search [116]. Instead of directly predicting mass spectra, *in silico* fragmentation attempts to elucidate a structure from all candidates that best explains the given MS^2 spectrum. This approach was first employed in EPIC using a bond disconnection algorithm to exhaust all possible substructures of a molecule and compare the substructures to formulas inferred from fragment ions. Then relevant structures were listed for user confirmation [117]. Later, FiD [118] and Mass-MetaSite [119] were developed on the basis of bond dissociation mechanism, and MetFrag extended this procedure [120] by considering rearrangement reactions during molecule fragmentation. An alternative procedure was implemented in FingerID by calculating the likelihood between metabolites in a database and a given experimental MS^2 spectrum in a feature space called fingerprints using an support vector machine (SVM) model [121]. This model was obtained by training fingerprints extracted from the Mass Bank MS/MS (MS^2) spectral library. CFM calculated the likelihood between database metabolites and given MS^2 spectra in accordance with the competitive fragmentation process learned from a spectral library using the expectation maximum algorithm [122].

De novo analysis, however, infers structures from the observed fragments in a given MS^n spectrum. This approach first determines the formulas of fragments according to their high resolution *m/z* and then deduces the structure of a precursor ion using these formulas and the known fragmentation pathways that generate these ions. To date, the most appropriate method employed for this deduction appears to be the construction of a fragmentation tree with nodes being fragment formulas, edges being neutral losses, and the root being the precursor [123, 124]. Therefore, with an appropriate scoring scheme, an experimental MS^n spectrum can be identified by extracting the most optimal fragmentation tree defined by the scores. Even so, the later portion of this procedure has been demonstrated to be extremely computationally intensive, despite already attaining the precursor formulas [125]. This obstacle can be partly solved by heuristic methods [126] and several tools, such as SIRIUS^2 [123, 124] and MAGMa [127].

2.3 Variable selection

Variable selection aims to extracting important metabolites from a mass of metabolites detected by mass spectrometry that can help us to answer biological questions at hand, which plays an essential role in metabolomics. From statistical point of view, this is an optimization approach that discovers an optimal variable combination from the considerable body of variables. However, this process faces a great challenge to address the NP-hard problem called “large p, small n problem” [128]. To date, numerous variable selection methods specific to this problem have been proposed. Some of these suggested strategies are based on statistical features of variables, whereas some are based on the optimization algorithm. Herein, we divide these methods into two kinds of approaches as follows: variable ranking and variable subset selection [129].

2.3.1Variable ranking

Variable ranking is mostly used in revealing informative metabolites or biomarkers. The process of ranking assigns a measure of importance to each variable on the basis of certain criteria. Many PLS-based criteria are frequently employed for variable ranking[130], including PLS loading weights (LW) [131], variable importance on projection (VIP) scores [132], regression coefficient (RC) [133], target projection (TP) [134], and selectivity ratio (SR) [135]. To date, VIP is the most popular one in metabolomics. Yi *et al.* [136] reported that VIP exhibited better efficiency than LW and RC for the metabolomics dataset of nasopharyngeal carcinoma patients. However, for another dataset, the comparison result between different variable ranking methods might be different [137, 138]. Because the efficiency of these methods is data-dependent, it is hard to say which one is the best. We should know that various variable ranking methods are most likely to generate different variable ranking results due to their different principles. Recently, Yun *et al.* use rank aggregation method to emerge all different ranking lists into a final aggregated variable ranking list for biomarker discovery [139]. It is a good attempt to handle this problem. In addition, variable ranking can be conducted based on statistical features between variables and classification label.

2.3.2 Variable subset selection

Subset selection refers to the search for an optimal subset from all variables that satisfy an optimality criterion. Any variable ranking method can be transformed into a variable subset selection algorithm by introducing a threshold on the variable importance values. The assignment of this threshold can be subjective or achieved by statistical method [129]. Usually, a trade-off between model prediction accuracy and the number of selected variables is considered. The most straightforward proposal for this purpose is to use cross validation (CV) procedure to determine the threshold. This approach estimates the generalization error using different number of variables and chooses the number that minimizes the prediction error (CV error). That is, after ranking variables from the most important to the least by some criteria (e.g., VIP), models are built by adding these variables sequentially until all are included, and CV error obtained by each model is recorded. The best variable subset can then be determined to be the first *n* variables if minimum CV error is achieved after adding *n*th variable. In addition, some criteria related to the classification algorithm can be employed for subset selection. The objective function is a pattern classifier, which evaluates variable subsets according to their predictive accuracy by statistical re-sampling (e.g., bootstrapping) or CV. Usually, optimization algorithm is combined with the classification algorithm. And, variable subset selection seeks the optimal or near-optimal subset with respect to an objective function. For example, genetic algorithm - Bayesian network （GA-BayesN） approach [140] combines the optimization algorithm GA with a classifier. Compared with the variable ranking method, subset selection generally achieves better prediction accuracy because the latter considers the specific interactions between the classifier and dataset. In the process, subset selection utilizes a mechanism to avoid overfitting through re-sampling or CV measures of prediction accuracy. However, the approach entails training of a classifier for each variable subset, leading to low execution and high computation. Moreover, the solution lacks generality because subset selection combines the bias of the classifier with the fitness evaluation function.

2.3.3 Variable selection considering the interaction effect among variables

In fact, finding an optimal subset or variable ranking is not always preferred unless the interaction among multiple variables is considered. The collective effect of variables should be considered because the joint performance of a set of variables is better than the additive independent contributions of its individuals [141]. To address this problem, Zhao and Liu introduced a variable subset selection method, called INTERACT [142]. This approach is based on inconsistency and symmetrical uncertainty measurements for finding interacting features. The group proposed variable interactions can be implicitly managed with a carefully designed variable evaluation metric and a search strategy with a specially designed data structure. The metric and strategy together take the combination effects of variables into considerations when performing variable selection. The method proposed in Breiman’s work [143] somewhat considers the combination effects of variables on the basis of the random forest (RF) and permutation test. The variable importance is assessed by the percent increase of misclassification error when the variable is randomly permuted in a RF. However, all variables are involved in the RF model, thus, providing a good reflection of the synergetic effect among multiple variables is difficult to accomplish.

Recently, Liang’s group proposed a new strategy for variable selection, called model population analysis (MPA) [144]. This method provides a general framework for the development of data analysis methods. Figure 3 illustrates the outline of the MPA. MPA involves three steps. Firstly, (1) sampling method (e.g., Monte Carlo sampling (MCS)) is employed to randomly produce N sub-datasets (e.g., 10,000). Then, (2) a sub-model is built on each sub-dataset. And finally, (3) statistical analysis is employed to evaluate outcomes of interest (e.g., prediction errors) for all established N sub-models. With this approach, the variables are identified as informative, uninformative, or interfering variables according to the differences between the cases and control samples. Figure 4 illustrates the prediction error distributions of the three kinds of variables after permutation. Uninformative and interfering variables are useless because of their potential undesirable influence on the modeling. Thus, discovering the optimal variable subset or ranking in the informative variables can produce compelling results.

Insert Figure 3

Insert Figure 4

Subwindow permutation analysis (SPA) [145] combines the above-mentioned concepts on the MCS method and MPA. SPA assesses each variable's importance on the basis of the sub-models obtained by MCS technique. Informative variables are identified and ranked by *p* values obtained by the Mann-Whitney U test on two distributions of prediction errors. Another method, margin influence analysis (MIA) [146] is also based on the concepts of MCS and MPA. Although designed to operate with SVM in identifying informative variables, MIA also offers a measure for each variable on the basis of the differences between the prediction errors from the inclusion and exclusion of this variable. However, the chance of each variable to be sampled by MCS is not the same. Some variables are selected more frequently than others; hence, assessing the importance of each variable using the above introduced strategy does not appear appropriate. So, a new sampling method in the variable space, called binary matrix sampling (BMS) [147], was proposed. This method not only considers the synergetic effect among multiple variables, but also guarantees that each variable is selected with equal probability and a population of different variable combinations is concurrently generated. With this population of variable subset, Yun *et al.* introduced a method called variable importance analysis, which is based on random variable combination (VIAVC) [137]. VIAVC employs the MPA strategy and finds the optimal subset of variables by observing the differences between the prediction errors of inclusion and exclusion of each variable. Meanwhile, Deng *et al*. developed an optimization algorithm called variable iterative space shrinkage approach (VISSA) to determine optimal variable combinations [148]. Each variable is assigned a weight according to its importance during modeling in VISSA. The weight of each variable accumulates through an iterative procedure and the variables are selected when their weights reach “1”. Two rules are highlighted in the VISSA algorithm. First, the variable space shrinks smoothly in each step. Second, the variable space is optimized in each step.

Although the above mentioned methods considered the synergetic effect among multiple variables, these approaches rarely investigate the complementary information between variables. By contrast, the variable complementary network (VCN) is an overall method that visualizes the complementary processes among biological variables [149]. VCN accumulates the information from several classification models obtained by MCS in variable space, quantitatively computes the complementary information between variables. Thus, it can effectively discover biomarkers with the aid of mutual associations among metabolites.   For comparison, Table 3 lists several variable selection methods.

Insert Table 3

2.4 Modeling of the data

To explore the high-dimensional metabolomics datasets and discover valuable information on biological events, a number of machine-learning methods have been applied. Main characteristics of the machine learning methods which will be described below are summarized in Table 4. It contains the category, advantages and disadvantages of each method, and also some applications in metabolomics.

Insert Table 4

2.4.1 Unsupervised methods

Unsupervised methods are usually used to explore the overall structure of a dataset, finding trends and groupings within the dataset. These methods contribute an unbiased view of the data. Several unsupervised methods are available, during which principal component analysis (PCA), hierarchical cluster analysis (HCA), and self-organization mapping (SOM) are the most frequently used examples in metabolomics.

PCA transforms the high-dimensional variables into a small number of orthogonal factors, called PCs, containing the largest variance [150, 151]. PCA provides the projection of samples into low dimensional (usually two- or three-dimensional) PC space, enabling the visualization of the sample distribution. HCA aims to group relatively similar samples in one cluster and relatively dissimilar objects in another [152, 153].

SOM is a neural-network algorithm [154]. For high-dimensional data, SOM can form a non-linear projection on a regular, low-dimensional grid. The clustering in the data space and the metric-topological relations of the data items is clearly visible. SOM is a useful tool to characterize metabolic patterns and interrelationship between samples [155-157]. For example, similar responses on primary and secondary metabolites were characterized in microorganisms across stimuli using MS-based metabolomics and SOM [155].

PARAFAC2 [158] is an extension of PARAFAC [159] which can be used to model three-way data with a trilinear structure. PARAFAC2 can be considered as the generalization of PCA to a higher order of data. It allows the simultaneous processing of all samples, deconvolution of metabolites, elimination of chromatographic baseline, and alignment of retention time shifts. PARAFAC2 can provide simple and robust models upon the application of some constraints. The advantage of PARAFAC2 is its ability of finding and modeling the shifted peaks of the same chemical compounds, with the disadvantage of being sensitive to noise [160, 161]. Goodacre *et al.* [162] employed PARAFAC2 to model the metabolic profiles of meat and characterise the hygiene status of pork chops which undergo a spoilage process.

2.4.2 Supervised methods

Supervised techniques support a priori known data structures to train patterns and rules to predict new data, which can be classified as linear methods, such as PLS-DA, linear discriminant analysis (LDA), orthogonal projections to latent structures discriminant analysis (OPLS-DA), and non-linear methods, including RF and SVM *etc.*

LDA attempts to find a linear function on the basis of original variables, which maximizes the ratio of between-class variance and minimizes the ratio of within-class variance [152]. LDA is a fast and powerful tool for discriminant analysis, in which parameter optimization is not necessary. The number of samples must be larger than that of the variables, ensuring that the inverse of the covariance matrix can be obtained [163].

The most widely used supervised method for classification in chemometrics is PLS-DA [164], which is a combination of PLS regression and LDA. One advantage of PLS-DA is its ability to handle highly collinear data. Moreover, PLS-DA can provide excellent insights into the cause of discrimination by checking the behavior of variables (e.g. variable importance, see Section 2.3.1). As such, PLS-DA is also a useful tool in biomarker discovery. The recent modification of PLS-DA is the OPLS-DA [165]. The systematic variations in data matrix X can be split into two parts through the orthogonal signal correction (OSC) technique [166]: one part exhibits linear responsiveness, whereas another is linearly orthogonal to the response. OPLS-DA supposes that only the variance related to the response is useful for modeling [167]. It gives better visualization and interpretation than PLS-DA [168] and has been widely applied in modeling and biomarker discovery in metabolomics [169-171].

2.4.3 Non-linear methods

Complex interactions occur in different levels of biological organizations; hence, biological processes commonly follow a non-linear response. In these cases, non-linear pattern recognition methods are required to characterize metabolomics data. Many non-linear techniques have been proposed in pattern recognition and machine learning research fields. Among these methods, kernel-PLS, RF and SVMs are three popular methods used in metabolomics.

Kernel-based models transform data using some specific functions called kernels. By using the kernel transformation, researchers can transform the non-linear problem of the original data into a higher-dimensional feature space. Afterward, the non-linear problem becomes linear and can be solved easily. The kernel functions appear in various types, and users can choose appropriate kernel transformation for a certain dataset. Positive semi-definite is one requirement of the kernel matrix [172]. Meanwhile, the dot product is the simplest kernel function for the data matrix. The radial basic function is another frequently used kernel function that requires tuning of parameters relating to the width of the Gaussian. Kernel-based classification methods, such as kernel Fisher discriminant analysis (K-FDA) [173], kernel PLS (K-PLS) [174], and kernel OPLS (KO-PLS) [175] have been developed and all exhibit obvious advantages in solving non-linear problems.

SVM is another powerful kernel-based classifier that utilizes a set of objects called support vectors to define decision boundaries and separate binary class[176]. SVM focuses on finding a hyper-plane that splits two classes perfectly, whereas the thickness of the margins is maximized. Hence, for each class, the distance of the plane to the data point is the closest [177, 178]. If a point is situated on the wrong side of the margin, the margin is maximized by penalizing the point. The step can split the overlapping classes. Support vectors are the points on the boundary or on the wrong side of the margin supporting the split. When classes are separated by a non-linear boundary, the kernel method is used to find the boundary. SVM is particularly suitable for the data of small sample sizes. The scheme is also capable of handling both linear and non-linear problems of classification by applying linear and non-linear kernels.

The major disadvantage of SVM is that the model is lack of transparency and variable importance is difficult to obtain. Another disadvantage is that it does not provide a universal means of solving non-linear problems. Hence, kernel functions should be selected discreetly [179]. It has been applied in toxicology research [180, 181], food research [182], and *etc*.

RF [143] is an ensemble-learning method that consists of a large number of classification and regression trees (CART). It is highly powerful classifier for high-dimensional data. A random resampling method with replacement called bootstrapping [183] is used to select training samples from the original samples (bootstrap samples) to construct a classification tree. Bootstrapping is carried out many times to build a large group of simple CARTs. Model accuracy is improved with the help of bootstrapping using resample means to estimate sample means [184]. Two powerful and efficient machine-learning techniques, bagging and random feature selection are employed in RF. For bagging, each CART is trained on the bootstrap samples of the training dataset. Predictions are obtained from the majority of votes of the CARTs. During RF model construction, only about two-thirds of training samples are used due to the intrinsic property of bootstrap sampling. Thus the remaining samples can serve as an internal testing set to monitor the prediction error termed out-of-bag error (OOB error). Besides, variable importance can also be obtained by comparing OOB error difference between normal variable and its random permutation, as has been introduced in previous section. RF has shown better performance than many of the classifiers such as PLS-DA and OPLS-DA with external validation [185].

Other examples also showed that RF and other approaches could be the alternatives to PLS-DA [186]. It has been applied to metabolomics research of hepatocellular carcinoma [187], breast cancer [188], metabolic syndrome [189], and *etc*.

2.4.4 Model tuning and model validation

The tuning of parameters is of great importance when building a model. CV [190] is the most commonly used model tuning method because it selects a model on the basis of prediction ability. Leave-one-out CV, K-fold CV [191] and Monte Carlo CV [192] are important branches of CV. Recently, CV has faced up some criticisms. For example, it may provide exceedingly optimistic results of the model prediction ability [193]. An alternative is to use double CV (DCV) which involves two loops: the inner loop is used for model tuning, and outer loop is adopted for model validation [194].

Model validation is a process on deciding whether results quantify hypothesized relationships between variables and responses and provide accurate estimation of the model prediction ability. Supervised machine-learning methods, such as PLS-DA, hold a high tendency for over-fitting, especially in high dimensional data [195, 196]. Thus, a careful model validation is desired.

Several criteria can evaluate the prediction ability of a model including sensitivity, specificity, accuracy, the receiver operating characteristic (ROC) curve, and the cross-validated coefficient of determination (Q^2). For a perfect classification, the value of specificity should be close to 1, and 1- specificity should be preferably close to 0. When the area under the ROC curve (AUC) is closer to 1, the method performs better. Recently, a criterion was developed by combing Q^2 and model stability (S)

[197]. The results show that, when a clear maximum of Q^2 is not obtained, S can provide additional information of over-fitting and it helps in finding the optimal nLVs. We believe that the criterion will be efficient for model selection of metabolomics.

The most common strategies and recommended for model validation are independent test set, CV and permutation test. Ideally, model validation employs an independent test set assumed to be representative and independent from the training data. A number of algorithms can be adopted to divide samples into training and test sets, including the Duplex algorithm [198], Kennard-and-Stone algorithm [199], and SPXY algorithm [200]. However, the ideal situation is usually unsatisfied in actual settings, often resulting in bias findings.

In CV, model tuning and model validation processes are carried out simultaneously. When the optimal model parameter is determined, the characteristics of prediction ability, such as Q^2, are obtained by tuning parameters. However, in DCV the model tuning and model validation processes are carried out separately using inner loop and outer loop, respectively. DCV has shown more accurate estimations of error rates than six-fold CV [194].

Permutation test is another powerful approach for model validation. The class labels of samples are permutated randomly in a permutation test. By repeating the permutation test numerous times, a group of “wrong” models are built, and the distribution for accuracy, Q^2, and AUC can be obtained. For a validated model, the difference between the “right” models and the “wrong” models should be significant. This difference can be characterized by statistical hypothesis testing. The permutation test also offers many applications in metabolomics studies [201, 202].

The modeling of metabolomics data is a kind of systematic work. For exploratory studies, unsupervised methods, such as PCA, provide an informative first look at the dataset structures and relationships between groups. Then, supervised methods, such as PLS-DA and OPLS-DA, are applied to classify the samples as well as discover biomarkers. When these classifiers fail to work properly, non-linear models SVM and RF are applied to further explore the non-linear relationship within the data. In addition, the parameters of each model should be well tuned and the model should be validated with caution to ensure its prediction ability for future samples.

2.5. One eye on the future

To date, numerous authors have demonstrated that data processing based on an individual datasets limit the complete understanding of the chemical complexity of the metabolome. Substantial data and information is generated from numerous experimental platforms (e.g., NMR, GC-MS or LC-MS). Consequently, the combination of information becomes increasingly necessary and important in extending metabolite coverage and characterizing biological systems [5]. The greatest future challenge is on how to efficiently integrate massive information from various sources (i.e., data fusion problem). Merging information from multiple datasets with different structural characteristics and extracting the common or distinctive features will unquestionably form a crucial element for the more comprehensive prospect of metabolomics.

An increasing number of papers have been published to discuss the problem of data fusion since 2005 [5, 56, 203]. Data fusions often focus on handling multiple datasets generated by several analytical platforms and analyzing longitudinal metabolomic data with time-resolved models. Boccard and Rudaz proposed the four main approaches to data fusion: low-level, mid-level, high-level and kernel-based data fusion [5]. Low-level fusion simply merges data matrices from different platforms into a single matrix for regression analysis or discriminant analysis. Mid-level fusion first extracts relevant features from each data source and then concatenates these features into a single matrix. In high-level fusion, separate models are obtained from each data source and the results of each model are combined to obtain the final decision [204]. Kernel-based methods employ kernel functions to transform data into high-dimensional feature spaces and generate kernel matrices. The kernel matrices are then merged to construct a single matrix for modeling [205]. The selection of data fusion methods depends on the difference between data sources. Data sources with larger differences often entail higher levels of data fusion. So far, methods for data fusion mainly focus on the low and middle levels [203, 206, 207]. Further fusion includes the integration with various “omics” fields, such as genomics, transcriptomics, and proteomics. These methods are all effective strategies for describing a whole biological system. However, we should be careful to avoid the network discordance when metabolomics are integrated with other “omics” [208].

3. Conclusions

In summary, metabolomics plays an essential role in basic research for elucidating environmental effects, gene functions, and defining cellular processes. To date, research on this field entails much exercise of caution with regard to data acquisition, processing, and information interpretation because of the numerous limitations related to data processing in metabolomics. We herein emphasize four issues, which are of great importance for data processing in metabolomics as follows. 1) Automatic and effective data preprocessing remains a difficult task, especially for the detection, alignment, and deconvolution of peaks with low responses. 2) The confident identification of unknown metabolites from complex MS spectra data remains as a great challenge. 3) NP-hard problems in variable selection must be addressed but barely solved by all researchers. 4) New efficient model validation methods and indices are urgently desired. Furthermore, these methods must be carefully selected in practice to guarantee that the objective models are fully validated and with good prediction ability for future actual samples. All of these problems, along with the high-dimensional characteristics of metabolomic datasets pose numerous fundamental questions in chemometrics. Chemometrics is facing enormous challenges to develop robust and efficient methods to answer various biological questions derived from metabolomics. We believe that this review can guide practitioners of metabolomics, and provide insights into its present uses as well as new data processing applications.

Conflicts of interest statement

The author declares no conflicts of interest.

Acknowledgements

This work was supported financially by National Nature Foundation Committee of P.R. China (No.21465016and No.21473257), Science and Technological Program for Dongguan's Higher Education, Science and Research, and Health Care Institutions (2012108102032). In addition, the authors would like to thank the referees and editor for their useful comments and suggestions, which have helped us to improve our manuscript.

References

[1] R. Goodacre, S. Vaidyanathan, W.B. Dunn, G.G. Harrigan, D.B. Kell, Metabolomics by numbers: acquiring and understanding global metabolite data, Trends in biotechnology, 22 (2004) 245-252.

[2] H.K. Kim, Y.H. Choi, R. Verpoorte, NMR-based plant metabolomics: where do we stand, where do we go?, Trends in biotechnology, 29 (2011) 267-275.

[3] J.W. Allwood, R. Goodacre, An introduction to liquid chromatography–mass spectrometry instrumentation applied in plant metabolomic analyses, Phytochemical analysis, 21 (2010) 33-47.

[4] L. Yi, C. Song, Z. Hu, L. Yang, L. Xiao, B. Yi, W. Jiang, Y. Cao, L. Sun, A metabolic discrimination model for nasopharyngeal carcinoma and its potential role in the therapeutic evaluation of radiotherapy, Metabolomics, 10 (2014) 697-708.

[5] J. Boccard, S. Rudaz, Harnessing the complexity of metabolomic data with chemometrics, Journal of Chemometrics, 28 (2014) 1-9.

[6] J. van der Greef, A.K. Smilde, Symbiosis of chemometrics and metabolomics: past, present, and future, Journal of Chemometrics, 19 (2005) 376-386.

[7] R. Goodacre, Making sense of the metabolome using evolutionary computation: seeing the wood with the trees, Journal of experimental botany, 56 (2005) 245-254.

[8] A.H. BaniMustafa, N.W. Hardy, A Strategy for Selecting Data Mining Techniques in Metabolomics, Plant Metabolomics, Springer2012, pp. 317-333.

[9] M. Katajamaa, M. Orešič, Data processing for mass spectrometry-based metabolomics, Journal of Chromatography A, 1158 (2007) 318-328.

[10] A.M. De Liyera, M. Sysi-Aho, L. Jacob, J.A. Gagnon-Bartsch, S. Castillo, J.A. Simpson, T.P. Speed, Statistical Methods for Handling Unwanted Variation in Metabolomics Data, Analytical chemistry, 87 (2015) 3606-3615.

[11] M. Ernst, D.B. Silva, R.R. Silva, R.Z. Vêncio, N.P. Lopes, Mass spectrometry in plant metabolomics strategies: from analytical platforms to data acquisition and processing, Natural product reports, (2014).

[12] S. Castillo, P. Gopalacharyulu, L. Yetukuri, M. Orešič, Algorithms and tools for the preprocessing of LC–MS metabolomics data, Chemometrics and Intelligent Laboratory Systems, 108 (2011) 23-32.

[13] C.A. Smith, E.J. Want, G. O'Maille, R. Abagyan, G. Siuzdak, XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification, Analytical chemistry, 78 (2006) 779-787.

[14] H. Benton, D. Wong, S. Trauger, G. Siuzdak, XCMS2: processing tandem mass spectrometry data

for metabolite identification and structural characterization, Analytical chemistry, 80 (2008) 6382-6389.

[15] M. Katajamaa, J. Miettinen, M. Orešič, MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data, Bioinformatics, 22 (2006) 634-636.

[16] T. Pluskal, S. Castillo, A. Villar-Briones, M. Orešič, MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data, BMC bioinformatics, 11 (2010) 395.

[17] M. Sturm, A. Bertsch, C. Gropl, A. Hildebrandt, R. Hussong, E. Lange, N. Pfeifer, O. Schulz-Trieglaff, A. Zerck, K. Reinert, O. Kohlbacher, OpenMS-An open-source software framework for mass spectrometry, BMC Bioinformatics, 9 (2008).

[18] R.C. De Vos, S. Moco, A. Lommen, J.J. Keurentjes, R.J. Bino, R.D. Hall, Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry, Nature protocols, 2 (2007) 778-791.

[19] X. Wei, W. Sun, X. Shi, I. Koo, B. Wang, J. Zhang, X. Yin, Y. Tang, B. Bogdanov, S. Kim, MetSign: A computational platform for high-resolution mass spectrometry-based metabolomics, Analytical chemistry, 83 (2011) 7668-7675.

[20] A.L. Duran, J. Yang, L. Wang, L.W. Sumner, Metabolomics spectral formatting, alignment and conversion tools (MSFACTs), Bioinformatics, 19 (2003) 2283-2293.

[21] K. Hiller, J. Hangebrauk, C. Jäger, J. Spura, K. Schreiber, D. Schomburg, MetaboliteDetector: comprehensive analysis tool for targeted and nontargeted GC/MS based metabolome analysis, Analytical chemistry, 81 (2009) 3429-3439.

[22] Z.-M. Zhang, S. Chen, Y.-Z. Liang, Baseline correction using adaptive iteratively reweighted penalized least squares, Analyst, 135 (2010) 1138-1146.

[23] X. Liu, Z. Zhang, Y. Liang, P.F. Sousa, Y. Yun, L. Yu, Baseline correction of high resolution spectral profile data based on exponential smoothing, Chemometrics and Intelligent Laboratory Systems, 139 (2014) 97-108.

[24] M. Hilario, A. Kalousis, C. Pellegrini, M. Mueller, Processing and classification of protein mass spectra, Mass spectrometry reviews, 25 (2006) 409-449.

[25] P. Haimi, A. Uphoff, M. Hermansson, P. Somerharju, Software tools for analysis of mass spectrometric lipidome data, Analytical chemistry, 78 (2006) 8324-8331.

[26] M. Bellew, M. Coram, M. Fitzgibbon, M. Igra, T. Randolph, P. Wang, D. May, J. Eng, R. Fang, C. Lin, A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS, Bioinformatics, 22 (2006) 1902-1909.

[27] G. Vivó-Truyols, J. Torres-Lapasió, A. Van Nederkassel, Y. Vander Heyden, D. Massart, Automatic program for peak detection and deconvolution of multi-overlapped chromatographic signals: Part I: Peak detection, Journal of Chromatography A, 1096 (2005) 133-145.

[28] K.M. Pierce, R.E. Mohler, A Review of chemometrics applied to comprehensive two-dimensional separations from 2008–2010, Separation & Purification Reviews, 41 (2012) 143-168.

[29] S. Krishnan, J.T. Vogels, L. Coulier, R.C. Bas, M.W. Hendriks, T. Hankemeier, U. Thissen, Instrument and process independent binning and baseline correction methods for liquid chromatography–high resolution-mass spectrometry deconvolution, Analytica Chimica Acta, 740 (2012) 12-19.

[30] R. Danielsson, D. Bylund, K.E. Markides, Matched filtering with background suppression for improved quality of base peak chromatograms and mass spectra in liquid chromatography–mass spectrometry, Analytica Chimica Acta, 454 (2002) 167-184.

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

[32] P. Du, W.A. Kibbe, S.M. Lin, Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching, Bioinformatics, 22 (2006) 2059-2065.

[33] K.C. Leptos, D.A. Sarracino, J.D. Jaffe, B. Krastins, G.M. Church, MapQuant: Open‐source software for large‐scale protein quantification, Proteomics, 6 (2006) 1770-1782.

[34] C.A. Hastings, S.M. Norton, S. Roy, New algorithms for processing and peak detection in liquid chromatography/mass spectrometry data, Rapid communications in mass spectrometry, 16 (2002) 462-467.

[35] G. Vivó-Truyols, Bayesian approach for peak detection in two-dimensional chromatography, Analytical chemistry, 84 (2012) 2622-2630.

[36] M. Lopatka, G. Vivó-Truyols, M. Sjerps, Probabilistic peak detection for first-order chromatographic data, Analytica Chimica Acta, 817 (2014) 9-16.

[37] Y.Z. Liang, O.M. Kvalheim, Resolution of two-way data: theoretical background and practical problem-solving - Part 1: Theoretical background and methodology, Fresen J Anal Chem, 370 (2001) 694-704.

[38] L.W. Hantao, H.G. Aleme, M.P. Pedroso, G.P. Sabin, R.J. Poppi, F. Augusto, Multivariate curve resolution combined with gas chromatography to enhance analytical separation in complex samples: A review, Analytica Chimica Acta, 731 (2012) 11-23.

[39] C. Ruckebusch, L. Blanchet, Multivariate curve resolution: A review of advanced and tailored applications and challenges, Analytica Chimica Acta, 765 (2013) 28-36.

[40] R. Tauler, Multivariate curve resolution applied to second order data, Chemometrics and Intelligent Laboratory Systems, 30 (1995) 133-146.

[41] E. Gorrochategui, J. Jaumot, R. Tauler, A protocol for LC-MS metabolomic data processing using chemometric tools, Protocol Exchange, (2015).

[42] M. Navarro-Reig, J. Jaumot, A. Garcia-Reiriz, R. Tauler, Evaluation of changes induced in rice metabolome by Cd and Cu exposure using LC-MS with XCMS and MCR-ALS data analysis strategies, Analytical and Bioanalytical Chemistry, 407 (2015) 8835-8847.

[43] H. Tsugawa, T. Cajka, T. Kind, Y. Ma, B. Higgins, K. Ikeda, M. Kanazawa, J. VanderGheynst, O. Fiehn, M. Arita, MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis, Nature methods, 12 (2015) 523-526.

[44] R. Smith, D. Ventura, J.T. Prince, LC-MS alignment in theory and practice: a comprehensive algorithmic review, Briefings in bioinformatics, 16 (2015) 104-117.

[45] K.J. Johnson, B.W. Wright, K.H. Jarman, R.E. Synovec, High-speed peak matching algorithm for retention time alignment of gas chromatographic data for chemometric analysis, Journal of Chromatography A, 996 (2003) 141-155.

[46] N.-P.V. Nielsen, J.M. Carstensen, J. Smedsgaard, Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping, Journal of Chromatography A, 805 (1998) 17-35.

[47] V. Pravdova, B. Walczak, D. Massart, A comparison of two algorithms for warping of analytical signals, Analytica Chimica Acta, 456 (2002) 77-92.

[48] J.W. Wong, C. Durante, H.M. Cartwright, Application of fast Fourier transform cross-correlation for the alignment of large chromatographic and spectral datasets, Analytical chemistry, 77 (2005) 5655-5661.

[49] V.P. Andreev, T. Rejtar, H.-S. Chen, E.V. Moskovets, A.R. Ivanov, B.L. Karger, A universal denoising and peak picking algorithm for LC-MS based on matched filtration in the chromatographic time domain, Analytical chemistry, 75 (2003) 6314-6326.

[50] D.P. De Souza, E.C. Saunders, M.J. McConville, V.A. Likić, Progressive peak clustering in GC-MS Metabolomic experiments applied to Leishmania parasites, Bioinformatics, 22 (2006) 1391-1396.

[51] A. Prakash, P. Mallick, J. Whiteaker, H. Zhang, A. Paulovich, M. Flory, H. Lee, R. Aebersold, B. Schwikowski, Signal maps for mass spectrometry-based comparative proteomics, Molecular & cellular proteomics, 5 (2006) 423-432.

[52] R.G. Sadygov, F. Martin Maroto, A.F. Hühmer, ChromAlign: a two-step algorithmic procedure for time alignment of three-dimensional LC-MS chromatographic surfaces, Analytical chemistry, 78 (2006) 8207-8217.

[53] K.M. Pierce, L.F. Wood, B.W. Wright, R.E. Synovec, A comprehensive two-dimensional retention time alignment algorithm to enhance chemometric analysis of comprehensive two-dimensional separation data, Analytical chemistry, 77 (2005) 7735-7743.

[54] J. Listgarten, R.M. Neal, S.T. Roweis, P. Wong, A. Emili, Difference detection in LC-MS data for protein biomarker discovery, Bioinformatics, 23 (2007) e198-e204.

[55] W. Wang, H. Zhou, H. Lin, S. Roy, T.A. Shaler, L.R. Hill, S. Norton, P. Kumar, M. Anderle, C.H. Becker, Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards, Analytical chemistry, 75 (2003) 4818-4826.

[56] R.A. van den Berg, H.C. Hoefsloot, J.A. Westerhuis, A.K. Smilde, M.J. van der Werf, Centering, scaling, and transformations: improving the biological information content of metabolomics data, BMC genomics, 7 (2006) 142.

[57] O.M. Kvalheim, F. Brakstad, Y. Liang, Preprocessing of analytical profiles in the presence of homoscedastic or heteroscedastic noise, Analytical chemistry, 66 (1994) 43-51.

[58] R. Sokal, F. Rohlf, Assumptions of analysis of variance, Biometry: The Principles and Practice of Statistics in Biological Research. 3rd ed. New York: WH Freeman, (1995) 396-406.

[59] H.G. Gika, G. Theodoridis, J. Extance, A.M. Edge, I.D. Wilson, High temperature-ultra performance liquid chromatography–mass spectrometry for the metabonomic analysis of Zucker rat urine, Journal of Chromatography B, 871 (2008) 279-287.

[60] R. Liu, D. Lin, W. Chang, C. Liu, W. Tsay, J. Li, T. Kuo, Issues to address when isotopically labeled analogues of analytes are used as internal standards, Anal. Chem, 74 (2002) 618AJ626A.

[61] H. Redestig, A. Fukushima, H. Stenlund, T. Moritz, M. Arita, K. Saito, M. Kusano, Compensation for Systematic Cross-Contribution Improves Normalization of Mass Spectrometry Based Metabolomics Data, Analytical chemistry, 81 (2009) 7974-7980.

[62] H.G. Gika, E. Macpherson, G.A. Theodoridis, I.D. Wilson, Evaluation of the repeatability of ultra-performance liquid chromatography–TOF-MS for global metabolic profiling of human urine samples, Journal of Chromatography B, 871 (2008) 299-305.

[63] D.S. Wishart, Computational strategies for metabolite identification in metabolomics, Bioanalysis, 1 (2009) 1579-1596.

[64] T. Kind, O. Fiehn, Advances in structure elucidation of small molecules using mass spectrometry, Bioanal Rev, 2 (2010) 23-60.

[65] D.G. Watson, A rough guide to metabolite identification using high resolution liquid chromatography mass spectrometry in metabolomic profiling in metazoans, Comput Struct Biotechnol J, 4 (2013) e201301005.

[66] M. Holcapek, R. Jirasko, M. Lisa, Basic rules for the interpretation of atmospheric pressure ionization mass spectra of small molecules, J Chromatogr A, 1217 (2010) 3908-3921.

[67] I. Koo, S. Kim, X. Zhang, Comparative analysis of mass spectral matching-based compound identification in gas chromatography-mass spectrometry, J Chromatogr A, 1298 (2013) 132-138.

[68] S.E. Stein, D.R. Scott, Optimization and testing of mass spectral library search algorithms for compound identification, J Am Soc Mass Spectrom, 5 (1994) 859-866.

[69] W.B. Dunn, D. Broadhurst, P. Begley, E. Zelena, S. Francis-McIntyre, N. Anderson, M. Brown, J.D. Knowles, A. Halsall, J.N. Haselden, A.W. Nicholls, I.D. Wilson, D.B. Kell, R. Goodacre, Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry, Nat Protoc, 6 (2011) 1060-1083.

[70] J. Kopka, Current challenges and developments in GC-MS based metabolite profiling technology, Journal of Biotechnology, 124 (2006) 312-322.

[71] J. Kopka, N. Schauer, S. Krueger, C. Birkemeyer, B. Usadel, E. Bergmuller, P. Dormann, W. Weckwerth, Y. Gibon, M. Stitt, L. Willmitzer, A.R. Fernie, D. Steinhauser, GMD@CSB.DB: the Golm Metabolome Database, Bioinformatics, 21 (2005) 1635-1638.

[72] C. Wagner, M. Sefkow, J. Kopka, Construction and application of a mass spectral and retention time index database generated from plant GC/EI-TOF-MS metabolite profiles, Phytochemistry, 62 (2003) 887-900.

[73] N. Schauer, D. Steinhauser, S. Strelkov, D. Schomburg, G. Allison, T. Moritz, K. Lundgren, U. Roessner-Tunali, M.G. Forbes, L. Willmitzer, A.R. Fernie, J. Kopka, GC-MS libraries for the rapid identification of metabolites in complex biological samples, Febs Letters, 579 (2005) 1332-1337.

[74] T. Kind, G. Wohlgemuth, D.Y. Lee, Y. Lu, M. Palazoglu, S. Shahbaz, O. Fiehn, FiehnLib: Mass Spectral and Retention Index Libraries for Metabolomics Based on Quadrupole and Time-of-Flight Gas Chromatography/Mass Spectrometry, Anal Chem, 81 (2009) 10038-10048.

[75] N.W. Kwiecien, D.J. Bailey, M.J.P. Rushp, J.S. Cole, A. Ulbrich, A.S. Hebert, M.S. Westphall, J.J. Coon, High-Resolution Filtering for Improved Small Molecule Identification via GC/MS, Anal Chem, 87 (2015) 8328-8335.

[76] C. Benecke, R. Grund, R. Hohberger, A. Kerber, R. Laue, T. Wieland, Molgen(+), a Generator of Connectivity Isomers and Stereoisomers for Molecular-Structure Elucidation, Analytica Chimica Acta, 314 (1995) 141-147.

[77] J.E. Peironcely, M. Rojas-Cherto, D. Fichera, T. Reijmers, L. Coulier, J.L. Faulon, T. Hankemeier, OMG: Open Molecule Generator, J Cheminform, 4 (2012) 21.

[78] E.L. Schymanski, C. Meinert, M. Meringer, W. Brack, The use of MS classifiers and structure generation to assist in the identification of unknowns in effect-directed analysis, Analytica Chimica Acta, 615 (2008) 136-147.

[79] A. Kerber, R. Laue, M. Meringer, K. Varmuza, MOLGEN-MS: Evaluation of low resolution electron impact mass spectra with MS classification and exhaustive structure generation, in: E. Gelpi (Ed.) Advances in Mass Spectrometry 15, Wiley, 2001, pp. 939-940.

[80] J. Hummel, N. Strehmel, J. Selbig, D. Walther, J. Kopka, Decision tree supported substructure prediction of metabolites from GC-MS profiles, Metabolomics, 6 (2010) 322-333.

[81] S.E. Stein, Chemical substructure identification by mass spectral library searching, J Am Soc Mass Spectrom, 6 (1995) 644-655.

[82] E.L. Schymanski, M. Meringer, W. Brack, Matching Structures to Mass Spectra Using Fragmentation Patterns: Are the Results As Good As They Look?, Anal Chem, 81 (2009) 3608-3617.

[83] E.L. Schymanski, C.M.J. Gallampois, M. Krauss, M. Meringer, S. Neumann, T. Schulze, S. Wolf, W. Brack, Consensus Structure Elucidation Combining GC/EI-MS, Structure Generation, and Calculated Properties, Anal Chem, 84 (2012) 3287-3295.

[84] S. Kumari, D. Stevens, T. Kind, C. Denkert, O. Fiehn, Applying In-Silico Retention Index and Mass Spectra Matching for Identification of Unknown Metabolites in Accurate Mass GC-TOF Mass Spectrometry, Anal Chem, 83 (2011) 5895-5902.

[85] O. Fiehn, J. Kopka, R.N. Trethewey, L. Willmitzer, Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry, Anal Chem, 72 (2000) 3573-3580.

[86] L.X. Zhang, C.L. Tang, D.S. Cao, Y.X. Zeng, B.B. Tan, M.M. Zeng, W. Fan, H.B. Xiao, Y.Z. Liang, Strategies for structure elucidation of small molecules using gas chromatography-mass spectrometric data, Trac-Trend Anal Chem, 47 (2013) 37-46.

[87] J.M. Halket, D. Waterman, A.M. Przyborowska, R.K.P. Patel, P.D. Fraser, P.M. Bramley, Chemical derivatization and mass spectral libraries in metabolic profiling by GC/MS and LC/MS/MS, Journal of Experimental Botany, 56 (2005) 219-243.

[88] T. Kind, O. Fiehn, Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry, BMC Bioinformatics, 8 (2007) 105.

[89] J.C.L. Erve, M. Gu, Y.D. Wang, W. DeMaio, R.E. Talaat, Spectral Accuracy of Molecular Ions in an LTQ/Orbitrap Mass Spectrometer and Implications for Elemental Composition Determination, J Am Soc Mass Spectr, 20 (2009) 2058-2069.

[90] Y.D. Wang, M. Cu, The Concept of Spectral Accuracy for MS, Anal Chem, 82 (2010) 7055-7062.

[91] D. Valkenborg, I. Mertens, F. Lemiere, E. Witters, T. Burzykowski, The isotopic distribution conundrum, Mass Spectrometry Reviews, 31 (2012) 96-109.

[92] T. Nagao, D. Yukihira, Y. Fujimura, K. Saito, K. Takahashi, D. Miura, H. Wariishi, Power of isotopic fine structure for unambiguous determination of metabolite elemental compositions: in silico evaluation and metabolomic application, Anal Chim Acta, 813 (2014) 70-76.

[93] Y. Xu, J.F. Heilier, G. Madalinski, E. Genin, E. Ezan, J.C. Tabet, C. Junot, Evaluation of Accurate Mass and Relative Isotopic Abundance Measurements in the LTQ-Orbitrap Mass Spectrometer for Further Metabolomics Database Building, Anal Chem, 82 (2010) 5490-5501.

[94] B.P. Koch, T. Dittmar, M. Witt, G. Kattner, Fundamentals of molecular formula assignment to ultrahigh resolution mass data of natural organic matter, Anal Chem, 79 (2007) 1758-1763.

[95] R.J.M. Weber, A.D. Southam, U. Sommer, M.R. Viant, Characterization of Isotopic Abundance Measurements in High Resolution FT-ICR and Orbitrap Mass Spectra for Improved Confidence of Metabolite Identification, Anal Chem, 83 (2011) 3737-3743.

[96] A. Knolhoff, J. Callahan, T. Croley, Mass Accuracy and Isotopic Abundance Measurements for HR-MS Instrumentation: Capabilities for Non-Targeted Analyses, J Am Soc Mass Spectr, 25 (2014) 1285-1294.

[97] A. Lommen, Ultrafast PubChem Searching Combined with Improved Filtering Rules for Elemental Composition Analysis, Anal Chem, 86 (2014) 5463-5469.

[98] Z.J. Zhu, A.W. Schultz, J.H. Wang, C.H. Johnson, S.M. Yannone, G.J. Patti, G. Siuzdak, Liquid chromatography quadrupole time-of-flight mass spectrometry characterization of metabolites guided by the METLIN database, Nature Protocols, 8 (2013) 451-460.

[99] J. Little, A. Williams, A. Pshenichnov, V. Tkachenko, Identification of “Known Unknowns” Utilizing Accurate Mass Data and ChemSpider, J Am Soc Mass Spectr, 23 (2012) 179-185.

[100] R. Breitling, S. Ritchie, D. Goodenowe, M.L. Stewart, M.P. Barrett, Ab initio prediction of metabolic networks using Fourier transform mass spectrometry data, Metabolomics, 2 (2006) 155-164.

[101] G.T. Gipson, K.S. Tatsuoka, B.A. Sokhansanj, R.J. Ball, S.C. Connor, Assignment of MS-based metabolomic datasets via compound interaction pair mapping, Metabolomics, 4 (2008) 94-103.

[102] S. Rogers, R.A. Scheltema, M. Girolami, R. Breitling, Probabilistic assignment of formulas to mass peaks in metabolomics experiments, Bioinformatics, 25 (2009) 512-518.

[103] R.J.M. Weber, M.R. Viant, MI-Pack: Increased confidence of metabolite identification in mass spectra by integrating accurate masses and metabolic pathways, Chemometr Intell Lab, 104 (2010) 75-82.

[104] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, M. Kanehisa, KEGG: Kyoto Encyclopedia of Genes and Genomes, Nucleic Acids Res, 27 (1999) 29-34.

[105] H. Doerfler, X. Sun, L. Wang, D. Engelmeier, D. Lyon, W. Weckwerth, mzGroupAnalyzer--predicting pathways and novel chemical structures from untargeted high-throughput metabolomics data, PLoS One, 9 (2014) e96188.

[106] S. Li, Y. Park, S. Duraisingham, F.H. Strobel, N. Khan, Q.A. Soltow, D.P. Jones, B. Pulendran, Predicting Network Activity from High Throughput Metabolomics, PLoS Comput Biol, 9 (2013) e1003123.

[107] N. Huang, M.M. Siegel, G.H. Kruppa, F.H. Laukien, Automation of a Fourier transform ion cyclotron resonance mass spectrometer for acquisition, analysis, and E-mailing of high-resolution exact-mass electrospray ionization mass spectral data, J Am Soc Mass Spectr, 10 (1999) 1166-1173.

[108] M. Brown, W.B. Dunn, P. Dobson, Y. Patel, C.L. Winder, S. Francis-McIntyre, P. Begley, K. Carroll, D. Broadhurst, A. Tseng, N. Swainston, I. Spasic, R. Goodacre, D.B. Kell, Mass spectrometry tools and metabolite-specific databases for molecular identification in metabolomics, Analyst, 134 (2009) 1322-1332.

[109] C. Kuhl, R. Tautenhahn, C. Bottcher, T.R. Larson, S. Neumann, CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets, Anal Chem, 84 (2012) 283-289.

[110] D.J. Creek, A. Jankevics, K.E. Burgess, R. Breitling, M.P. Barrett, IDEOM: an Excel interface for analysis of LC-MS-based metabolomics data, Bioinformatics, 28 (2012) 1048-1049.

[111] F. Fernandez-Albert, R. Llorach, C. Andres-Lacueva, A. Perera, An R package to analyse LC/MS metabolomic data: MAIT (Metabolite Automatic Identification Toolkit), Bioinformatics, 30 (2014) 1937-1939.

[112] S. Stein, Mass Spectral Reference Libraries: An Ever-Expanding Resource for Chemical Identification, Anal Chem, 84 (2012) 7274-7282.

[113] E. Werner, J.F. Heilier, C. Ducruix, E. Ezan, C. Junot, J.C. Tabet, Mass spectrometry for the identification of the discriminating signals from metabolomics: Current status and future trends, J Chromatogr B, 871 (2008) 143-163.

[114] F. Hufsky, K. Scheubert, S. Bocker, Computational mass spectrometry for small-molecule fragmentation, Trac-Trend Anal Chem, 53 (2014) 41-48.

[115] L.J. Kangas, T.O. Metz, G. Isaac, B.T. Schrom, B. Ginovska-Pangovska, L. Wang, L. Tan, R.R. Lewis, J.H. Miller, In silico identification software (ISIS): a machine learning approach to tandem mass spectral identification of lipids, Bioinformatics, 28 (2012) 1705-1713.

[116] T. Huan, C.Q. Tang, R.H. Li, Y. Shi, G.H. Lin, L. Li, MyCompoundID MS/MS Search: Metabolite

Identification Using a Library of Predicted Fragment-Ion-Spectra of 383,830 Possible Human Metabolites, Analytical Chemistry, 87 (2015) 10619-10626.

[117] A.W. Hill, R.J. Mortishire-Smith, Automated assignment of high-resolution collisionally activated dissociation mass spectra using a systematic bond disconnection approach, Rapid Commun Mass Sp, 19 (2005) 3111-3118.

[118] M. Heinonen, A. Rantanen, T. Mielikainen, J. Kokkonen, J. Kiuru, R.A. Ketola, J. Rousu, FiD: a software for ab initio structural identification of product ions from tandem mass spectrometric data, Rapid Commun Mass Spectrom, 22 (2008) 3043-3052.

[119] B. Bonn, C. Leandersson, F. Fontaine, I. Zamora, Enhanced metabolite identification with MS(E) and a semi-automated software for structural elucidation, Rapid Commun Mass Spectrom, 24 (2010) 3127-3138.

[120] S. Wolf, S. Schmidt, M. Muller-Hannemann, S. Neumann, In silico fragmentation for computer assisted identification of metabolite mass spectra, BMC Bioinformatics, 11 (2010) 148.

[121] M. Heinonen, H. Shen, N. Zamboni, J. Rousu, Metabolite identification and molecular fingerprint prediction through machine learning, Bioinformatics, 28 (2012) 2333-2341.

[122] F. Allen, R. Greiner, D. Wishart, Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification, Metabolomics, (2014) 1-13.

[123] S. Bocker, F. Rasche, Towards de novo identification of metabolites by analyzing tandem mass spectra, Bioinformatics, 24 (2008) i49-i55.

[124] F. Rasche, A. Svatos, R.K. Maddula, C. Bottcher, S. Bocker, Computing fragmentation trees from tandem mass spectrometry data, Anal Chem, 83 (2011) 1243-1251.

[125] F. Hufsky, M. Rempt, F. Rasche, G. Pohnert, S. Böcker, De novo analysis of electron impact mass spectra using fragmentation trees, Analytica Chimica Acta, 739 (2012) 67-76.

[126] I. Rauf, F. Rasche, F. Nicolas, S. Böcker, Finding Maximum Colorful Subtrees in Practice, in: B. Chor (Ed.) Lect N Bioinformat, Springer Berlin Heidelberg2012, pp. 213-223.

[127] L. Ridder, J.J.J. van der Hooft, S. Verhoeven, R.C.H. de Vos, R. van Schaik, J. Vervoort, Substructure-based annotation of high-resolution multistage MSn spectral trees, Rapid Communications in Mass Spectrometry, 26 (2012) 2461-2471.

[128] J. Boccard, J.L. Veuthey, S. Rudaz, Knowledge discovery in metabolomics: an overview of MS data handling, Journal of separation science, 33 (2010) 290-304.

[129] I. Narsky, F.C. Porter, Methods for Variable Ranking and Selection, Statistical Analysis Techniques in Particle Physics, Wiley-VCH Verlag GmbH & Co. KGaA2013, pp. 385-415.

[130] T. Mehmood, K.H. Liland, L. Snipen, S. Sæbø, A review of variable selection methods in partial least squares regression, Chemometrics and Intelligent Laboratory Systems, 118 (2012) 62-69.

[131] S. Wold, M. Sjöström, L. Eriksson, Partial Least Squares Projections to Latent Structures (PLS) in Chemistry, Encyclopedia of Computational Chemistry, John Wiley & Sons, Ltd2002.

[132] S. Favilla, C. Durante, M.L. Vigni, M. Cocchi, Assessing feature relevance in NPLS models by VIP, Chemom. Intell. Lab. Syst, 129 (2013) 76-86.

[133] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemom. Intell. Lab. Syst, 58 (2001) 109-130.

[134] T. Rajalahti, R. Arneberg, A.C. Kroksveen, M. Berle, K.-M. Myhr, O.M. Kvalheim, Discriminating Variable Test and Selectivity Ratio Plot: Quantitative Tools for Interpretation and Variable (Biomarker) Selection in Complex Spectral or Chromatographic Profiles, Analytical Chemistry, 81 (2009) 2581-2590.

[135] O.M. Kvalheim, Interpretation of partial least squares regression models by means of target

projection and selectivity ratio plots, J. Chemometr, 24 (2010) 496-504.

[136] L.Z. Yi, N.P. Dong, S.T. Shi, B.C. Deng, Y.H. Yun, Z.B. Yi, Y. Zhang, Metabolomic identification of novel biomarkers of nasopharyngeal carcinoma, Rsc Advances, 4 (2014) 59094-59101.

[137] Y.-H. Yun, F. Liang, B.-C. Deng, G.-B. Lai, C.M.V. Goncalves, H.-M. Lu, J. Yan, X. Huang, L.-Z. Yi, Y.-Z. Liang, Informative metabolites identification by variable importance analysis based on random variable combination, Metabolomics, 11 (2015) 1539-1551.

[138] M. Farrés, S. Platikanov, S. Tsakovski, R. Tauler, Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation, Journal of Chemometrics, 29 (2015) 528-536.

[139] Y.-H. Yun, B.-C. Deng, D.-S. Cao, W.-T. Wang, Y.-Z. Liang, Variable importance analysis based on rank aggregation with applications in metabolomics for biomarker discovery, Analytica Chimica Acta, (2016).

[140] E. Correa, R. Goodacre, A genetic algorithm-Bayesian network approach for the analysis of metabolomics and spectroscopic data: application to the rapid identification of Bacillus spores and classification of Bacillus species, BMC bioinformatics, 12 (2011) 33.

[141] D. Anastassiou, Computational analysis of the synergy among multiple interacting genes, Molecular Systems Biology, 3 (2007) n/a-n/a.

[142] Z. Zhao, H. Liu, Searching for interacting features in subset selection, Intelligent Data Analysis, 13 (2009) 207-228.

[143] L. Breiman, Random forests, Machine Learning, 45 (2001) 5-32.

[144] H.-D. Li, Y.-Z. Liang, Q.-S. Xu, D.-S. Cao, Model population analysis for variable selection, J. Chemometr, 24 (2010) 418-423.

[145] H.-D. Li, M.-M. Zeng, B.-B. Tan, Y.-Z. Liang, Q.-S. Xu, D.-S. Cao, Recipe for revealing informative metabolites based on model population analysis, Metabolomics, 6 (2010) 353-361.

[146] H.-D. Li, Y.-Z. Liang, Q.-S. Xu, D.-S. Cao, Recipe for Uncovering Predictive Genes Using Support Vector Machines Based on Model Population Analysis, IEEE. ACM. T. Comput. Bi, 8 (2011) 1633-1641.

[147] H. Zhang, H. Wang, Z. Dai, M.-s. Chen, Z. Yuan, Improving accuracy for cancer classification with a new algorithm for genes selection, BMC Bioinformatics, 13 (2012) 1-20.

[148] B.-c. Deng, Y.-h. Yun, Y.-z. Liang, L.-z. Yi, A novel variable selection approach that iteratively optimizes variable space using weighted binary matrix sampling, Analyst, 139 (2014) 4836-4845.

[149] H.-D. Li, Q.-S. Xu, W. Zhang, Y.-Z. Liang, Variable complementary network: a novel approach for identifying biomarkers and their mutual associations, Metabolomics, 8 (2012) 1218-1226.

[150] J.E. Jackson, A User's Guide to Principal Components, Wiley, New York, 1991.

[151] J. Xu, F.L. Hu, W. Wang, X.C. Wan, G.H. Bao, Investigation on biochemical compositional changes during the microbial fermentation process of Fu brick tea by LC-MS based metabolomics, Food Chem, 186 (2015) 176-184.

[152] A.R. Webb, Statistical pattern recognition, John Wiley & Sons2003.

[153] L. Jing, Z.T. Lei, G.W. Zhang, A.C. Pilon, D.V. Huhman, R.J. Xie, W.P. Xi, Z.Q. Zhou, L.W. Sumner, Metabolite profiles of essential oils in citrus peels and their taxonomic implications, Metabolomics, 11 (2015) 952-963.

[154] T. Kohonen, S.-O. Maps, Springer series in information sciences, Self-organizing maps, 30 (1995).

[155] C.R. Goodwin, B.C. Covington, D.K. Derewacz, C.R. McNees, J.P. Wikswo, J.A. McLean, B.O. Bachmann, Structuring Microbial Metabolic Responses to Multiplexed Stimuli via Self-Organizing Metabolomics Maps, Chemistry & biology, (2015).

[156] J.K. Kim, M.R. Cho, H.J. Baek, T.H. Ryu, C.Y. Yu, M.J. Kim, E. Fukusaki, A. Kobayashi, Analysis of metabolite profile data using batch-learning self-organizing maps, Journal of Plant Biology, 50 (2007) 517-521.

[157] A.D. Patterson, H. Li, G.S. Eichler, K.W. Krausz, J.N. Weinstein, A.J. Fornace, F.J. Gonzalez, J.R. Idle, UPLC-ESI-TOFMS-Based Metabolomics and Gene Expression Dynamics Inspector Self-Organizing Metabolomic Maps as Tools for Understanding the Cellular Response to Ionizing Radiation, Analytical Chemistry, 80 (2008) 665-674.

[158] J.M. Amigo, T. Skov, R. Bro, J. Coello, S. Maspoch, Solving GC-MS problems with parafac2, TrAC Trends in Analytical Chemistry, 27 (2008) 714-725.

[159] R. Bro, PARAFAC. Tutorial and applications, Chemometr Intell Lab, 38 (1997) 149-171.

[160] B. Khakimov, J.M. Amigo, S. Bak, S.B. Engelsen, Plant metabolomics: Resolution and quantification of elusive peaks in liquid chromatography-mass spectrometry profiles of complex plant extracts using multi-way decomposition methods, J Chromatogr A, 1266 (2012) 84-94.

[161] J.M. Amigo, M.J. Popielarz, R.M. Callejon, M.L. Morales, A.M. Troncoso, M.A. Petersen, T.B. Toldam-Andersen, Comprehensive analysis of chromatographic data by using PARAFAC2 and principal components analysis, J Chromatogr A, 1217 (2010) 4422-4429.

[162] Y. Xu, W. Cheung, C.L. Winder, W.B. Dunn, R. Goodacre, Metabolic profiling of meat: assessment of pork hygiene and contamination with Salmonella typhimurium, Analyst, 136 (2011) 508-514.

[163] C.M. Bishop, Pattern recognition and machine learning, springer New York2006.

[164] M. Barker, W. Rayens, Partial least squares for discrimination, J Chemometr, 17 (2003) 166-173.

[165] J. Trygg, S. Wold, Orthogonal projections to latent structures (O-PLS), J Chemometr, 16 (2002) 119-128.

[166] R. Madsen, T. Lundstedt, J. Trygg, Chemometrics in metabolomics-A review in human disease diagnosis, Anal Chim Acta, 659 (2010) 23-33.

[167] A. Kiss, C. Bordes, C. Buisson, F. Lasne, P. Lanteri, C. Cren-Olive, Data-handling strategies for metabonomic studies: example of the UHPLC-ESI/ToF urinary signature of tetrahydrocannabinol in humans, Anal Bioanal Chem, 406 (2014) 1209-1219.

[168] T. Verron, R. Sabatier, R. Joffre, Some theoretical properties of the O-PLS method, J Chemometr, 18 (2004) 62-68.

[169] A.H. Zhang, H. Sun, Y. Han, G.L. Yan, Y. Yuan, G.C. Song, X.X. Yuan, N. Xie, X.J. Wang, Ultraperformance Liquid Chromatography-Mass Spectrometry Based Comprehensive Metabolomics Combined with Pattern Recognition and Network Analysis Methods for Characterization of Metabolites and Metabolic Pathways from Biological Data Sets, Analytical Chemistry, 85 (2013) 7606-7612.

[170] B. Dieme, S. Mavel, H. Blasco, G. Tripi, F. Bonnet-Brilhault, J. Malvy, C. Bocca, C.R. Andres, L. Nada-Desbarats, P. Emond, Metabolomics Study of Urine in Autism Spectrum Disorders Using a Multiplatform Analytical Methodology, J Proteome Res, 14 (2015) 5273-5282.

[171] J. Hadrevi, M. Bjorklund, E. Kosek, S. Hallgren, H. Antti, M. Fahlstrom, F. Hellstrom, Systemic differences in serum metabolome: a cross sectional comparison of women with localised and widespread pain and controls, Scientific Reports, 5 (2015).

[172] J. Shawe-Taylor, N. Cristianini, Kernel methods for pattern analysis, Cambridge university press2004.

[173] D.S. Cao, M.M. Zeng, L.Z. Yi, B. Wang, Q.S. Xu, Q.N. Hu, L.X. Zhang, H.M. Lu, Y.Z. Liang, A novel kernel Fisher discriminant analysis: constructing informative kernel by decision tree ensemble for

metabolomics data analysis, Anal Chim Acta, 706 (2011) 97-104.

[174] B. Walczak, D. Massart, The radial basis functions—partial least squares approach as a flexible non-linear regression technique, Anal Chim Acta, 331 (1996) 177-185.

[175] M. Bylesjo, M. Rantalainen, J. Nicholson, E. Holmes, J. Trygg, K-OPLS package: Kernel-based orthogonal projections to latent structures for prediction and interpretation in feature space, Bmc Bioinformatics, 9 (2008) 106.

[176] V. Vapnik, Statistical Learning Theory, John Willey & Sons, New York, 1998.

[177] H.D. Li, Y.Z. Liang, Q.S. Xu, Support vector machines and its applications in chemistry, Chemometr Intell Lab, 95 (2009) 188-198.

[178] J. Luts, F. Ojeda, R. Van de Plas, B. De Moor, S. Van Huffel, J.A.K. Suykens, A tutorial on support vector machine-based methods for classification problems in chemometrics, Anal Chim Acta, 665 (2010) 129-145.

[179] C.J. Burges, A tutorial on support vector machines for pattern recognition, Data mining and knowledge discovery, 2 (1998) 121-167.

[180] Y.B. Li, L. Ju, Z.G. Hou, H.Y. Deng, Z.Z. Zhang, L. Wang, Z. Yang, J. Yin, Y.J. Zhang, Screening, Verification, and Optimization of Biomarkers for Early Prediction of Cardiotoxicity Based on Metabolomics, J Proteome Res, 14 (2015) 2437-2445.

[181] Y.B. Li, H.Y. Deng, L. Ju, X.X. Zhang, Z.Z. Zhang, Z. Yang, L. Wang, Z.G. Hou, Y.J. Zhang, Screening and validation for plasma biomarkers of nephrotoxicity based on metabolomics in male rats, Toxicology Research, 5 (2016) 259-267.

[182] V.G. Uarrota, R. Moresco, B. Coelho, E.d.C. Nunes, L.A.M. Peruch, E.d.O. Neubert, M. Rocha, M. Maraschin, Metabolomics combined with chemometric tools (PCA, HCA, PLS-DA and SVM) for screening cassava (Manihot esculenta Crantz) roots during postharvest physiological deterioration, Food Chem, 161 (2014) 67-78.

[183] B. Efron, Bootstrap methods: another look at the jackknife, The annals of statistics, (1979) 1-26.

[184] B.F. Manly, Randomization, bootstrap and Monte Carlo methods in biology, CRC Press2006.

[185] I.M. Scott, W. Lin, M. Liakata, J.E. Wood, C.P. Vermeer, D. Allaway, J.L. Ward, J. Draper, M.H. Beale, D.I. Corol, J.M. Baker, R.D. King, Merits of random forests emerge in evaluation of chemometric classifiers by external validation, Analytica Chimica Acta, 801 (2013) 22-33.

[186] P.S. Gromski, H. Muhamadali, D.I. Ellis, Y. Xu, E. Correa, M.L. Turner, R. Goodacre, A tutorial review: Metabolomics and partial least squares-discriminant analysis - a marriage of convenience or a shotgun wedding, Anal Chim Acta, 879 (2015) 10-23.

[187] R. Gao, J.H. Cheng, C.L. Fan, X.F. Shi, Y. Cao, B. Sun, H.G. Ding, C.J. Hu, F.T. Dong, X.Z. Yan, Serum Metabolomics to Identify the Liver Disease-Specific Biomarkers for the Progression of Hepatitis to Hepatocellular Carcinoma, Scientific Reports, 5 (2015).

[188] J.H. Huang, L. Fu, B. Li, H.L. Xie, X.J. Zhang, Y.J. Chen, Y.H. Qin, Y.H. Wang, S.H. Zhang, H.Y. Huang, D.F. Liao, W. Wang, Distinguishing the serum metabolite profiles differences in breast cancer by gas chromatography mass spectrometry and random forest method, RSC Advances, 5 (2015) 58952-58958.

[189] Z. Lin, C.M.V. Goncalves, L. Dai, H.M. Lu, J.H. Huang, H.C. Ji, D.S. Wang, L.Z. Yi, Y.Z. Liang, Exploring metabolic syndrome serum profiling based on gas chromatography mass spectrometry and random forest models, Anal Chim Acta, 827 (2014) 22-27.

[190] M. Stone, Cross-validatory choice and assessment of statistical predictions, Journal of the Royal Statistical Society. Series B (Methodological), (1974) 111-147.

[191] S. Geisser, The predictive sample reuse method with applications, J Am Stat Assoc, 70 (1975) 320-328.

[192] J. Shao, Linear Model Selection by Cross-validation, J Am Stat Assoc, 88 (1993) 486-494.

[193] D. Krstajic, L. Buturovic, D. Leahy, S. Thomas, Cross-validation pitfalls when selecting and assessing regression and classification models, J Cheminformatics, 6 (2014) 10.

[194] J.A. Westerhuis, H.C.J. Hoefsloot, S. Smit, D.J. Vis, A.K. Smilde, E.J.J. van Velzen, J.P.M. van Duijnhoven, F.A. van Dorsten, Assessment of PLSDA cross validation, Metabolomics, 4 (2008) 81-89.

[195] R.G. Brereton, Consequences of sample size, variable selection, and model validation and optimisation, for predicting classification ability from analytical data, Trac-Trend Anal Chem, 25 (2006) 1103-1111.

[196] H.D. Li, Y.Z. Liang, Q.S. Xu, D.S. Cao, Model population analysis for variable selection, J Chemometr, 24 (2010) 418-423.

[197] B.C. Deng, Y.H. Yun, Y.Z. Liang, D.S. Cao, Q.S. Xu, L.Z. Yi, X. Huang, A new strategy to prevent over-fitting in partial least squares models based on model population analysis, Anal Chim Acta, 10.1016/j.aca.2015.04.045 (2015).

[198] R.D. Snee, Validation of regression models: methods and examples, Technometrics, 19 (1977) 415-428.

[199] R.W. Kennard, L.A. Stone, Computer Aided Design of Experiments, Technometrics, 11 (1969) 137-148.

[200] R.K.H. Galvao, M.C.U. Araujo, G.E. Jose, M.J.C. Pontes, E.C. Silva, T.C.B. Saldanha, A method for calibration and validation subset partitioning, Talanta, 67 (2005) 736-740.

[201] Z.Z. Huang, Y.J. Chen, W. Hang, Y. Gao, L. Lin, D.Y. Li, J.C. Xing, X.M. Yan, Holistic metabonomic profiling of urine affords potential early diagnosis for bladder and kidney cancers, Metabolomics, 9 (2013) 119-129.

[202] S. Bovo, G. Mazzoni, D.G. Calo, G. Galimberti, F. Fanelli, M. Mezzullo, G. Schiavo, E. Scotti, A. Manisi, A.B. Samore, F. Bertolini, P. Trevisi, P. Bosi, S. Dall'Olio, U. Pagotto, L. Fontanesi, Deconstructing the pig sex metabolome: Targeted metabolomics in heavy pigs revealed sexual dimorphisms in plasma biomarkers and metabolic pathways, Journal of Animal Science, 93 (2015) 5681-5693.

[203] J. Forshed, H. Idborg, S.P. Jacobsson, Evaluation of different techniques for data fusion of LC/MS and 1 H-NMR, Chemometrics and Intelligent Laboratory Systems, 85 (2007) 102-109.

[204] T. Doeswijk, A. Smilde, J. Hageman, J. Westerhuis, F. van Eeuwijk, On the increase of predictive performance with high-level data fusion, Anal Chim Acta, 705 (2011) 41-47.

[205] A. Smolinska, L. Blanchet, L. Coulier, K.A. Ampt, T. Luider, R.Q. Hintzen, S.S. Wijmenga, L.M. Buydens, Interpretation and visualization of non-linear data fusion in kernel space: study on metabolomic characterization of progression of multiple sclerosis, Plos One, 7 (2012).

[206] R. Bro, H.J. Nielsen, F. Savorani, K. Kjeldahl, I.J. Christensen, N. Brünner, A.J. Lawaetz, Data fusion in metabolomic cancer diagnostics, Metabolomics, 9 (2013) 3-8.

[207] L. Blanchet, A. Smolinska, A. Attali, M.P. Stoop, K.A. Ampt, H. van Aken, E. Suidgeest, T. Tuinstra, S.S. Wijmenga, T. Luider, Fusion of metabolomics and proteomics data for biomarkers discovery: case study on the experimental autoimmune encephalomyelitis, BMC bioinformatics, 12 (2011) 254.

[208] A.R. Fernie, M. Stitt, On the discordance of metabolomics with proteomics and transcriptomics: coping with increasing complexity in logic, chemistry, and network interactions scientific correspondence, Plant Physiology, 158 (2012) 1139-1145.

[209] S. Bocker, M.C. Letzel, Z. Liptak, A. Pervukhin, SIRIUS: decomposing isotope patterns for

metabolite identification, Bioinformatics, 25 (2009) 218-224.

[210] B. Zhou, J. Wang, H.W. Ressom, MetaboSearch: tool for mass-based metabolite identification using multiple databases, PLoS One, 7 (2012) e40096.

[211] M. Gerlich, S. Neumann, MetFusion: integration of compound identification strategies, Journal of Mass Spectrometry, 48 (2013) 291-298.

[212] F. Allen, A. Pon, M. Wilson, R. Greiner, D. Wishart, CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra, Nucleic Acids Res, 42 (2014) W94-99.

[213] J. Draper, D.P. Enot, D. Parker, M. Beckmann, S. Snowdon, W. Lin, H. Zubair, Metabolite signal identification in accurate mass metabolomics data with MZedDB, an interactive m/z annotation tool utilising predicted ionisation behaviour 'rules', BMC Bioinformatics, 10 (2009) 227.

[214] S.E. Stein, An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data, J Am Soc Mass Spectr, 10 (1999) 770-781.

[215] C. Steinbeck, Y.Q. Han, S. Kuhn, O. Horlacher, E. Luttmann, E. Willighagen, The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics, J Chem Inf Comp Sci, 43 (2003) 493-500.

[216] M.A. Hall, Correlation-based feature selection for machine learning, The University of Waikato, 1999.

[217] M. Ben-Bassat, Pattern recognition and reduction of dimensionality, Handbook of Statistics, 2 (1982) 773-910.

[218] J. Liang, S. Yang, A. Winstanley, Invariant optimal feature selection: A distance discriminant and feature ranking based solution, Pattern Recognition, 41 (2008) 1429-1439.

[219] L. Yu, H. Liu, Efficient feature selection via analysis of relevance and redundancy, The Journal of Machine Learning Research, 5 (2004) 1205-1224.

[220] H.-D. Li, Y.-Z. Liang, Q.-S. Xu, D.-S. Cao, Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration, Ana. Chim. Acta, 648 (2009) 77-84.

[221] M.D. Cao, B. Sitter, T.F. Bathen, A. Bofin, P.E. Lønning, S. Lundgren, I.S. Gribbestad, Predicting long-term survival and treatment response in breast cancer patients receiving neoadjuvant chemotherapy by MR metabolic profiling, NMR in Biomedicine, 25 (2012) 369-378.

[222] E. Alba, J. Garcia-Nieto, L. Jourdan, E. Talbi, Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms, Evolutionary Computation, 2007. CEC 2007. IEEE Congress on, 2007, pp. 284-290.

[223] Y.-H. Yun, W.-T. Wang, M.-L. Tan, Y.-Z. Liang, H.-D. Li, D.-S. Cao, H.-M. Lu, Q.-S. Xu, A strategy that iteratively retains informative variables for selecting optimal variable subset in multivariate calibration, Anal. Chim. Acta, 807 (2014) 36-43.

[224] Q. Mao, M. Bai, J.D. Xu, M. Kong, L.Y. Zhu, H. Zhu, Q. Wang, S.L. Li, Discrimination of leaves of Panax ginseng and P. quinquefolius by ultra high performance liquid chromatography quadrupole/time-of-flight mass spectrometry based metabolomics approach, Journal of Pharmaceutical and Biomedical Analysis, 97 (2014) 129-140.

[225] J.S. Wang, T. Reijmers, L.J. Chen, R. Van der Heijden, M. Wang, S.Q. Peng, T. Hankemeier, G.W. Xu, J. Van der Greef, Systems toxicology study of doxorubicin on rats using ultra performance liquid chromatography coupled with mass spectrometry based metabolomics, Metabolomics, 5 (2009) 407-418.

[226] H.H.M. Draisma, T.H. Reijmers, J.J. Meulman, J. van der Greef, T. Hankemeier, D.I. Boomsma,

Hierarchical clustering analysis of blood plasma lipidomics profiles from mono- and dizygotic twin families, Eur J Hum Genet, 21 (2013) 95-101.

[227] H.-P. Kriegel, P. Kröger, A. Zimek, Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering, ACM Transactions on Knowledge Discovery from Data (TKDD), 3 (2009) 1.

[228] L. Vaclavik, A. Schreiber, O. Lacina, T. Cajka, J. Hajslova, Liquid chromatography–mass spectrometry-based metabolomics for authenticity assessment of fruit juices, Metabolomics, 8 (2012) 793-803.

[229] M.L. Ouyang, Z.M. Zhang, C. Chen, X.B. Liu, Y.Z. Liang, Application of sparse linear discriminant analysis for metabolomics data, Anal Methods-Uk, 6 (2014) 9037-9044.

[230] L.C. Phua, C.H. Wilder-Smith, Y.M. Tan, T. Gopalakrishnan, R.K. Wong, X.H. Li, M.E. Kan, J. Lu, A. Keshavarzian, E.C.Y. Chan, Gastrointestinal Symptoms and Altered Intestinal Permeability Induced by Combat Training Are Associated with Distinct Metabotypic Changes, J Proteome Res, 14 (2015) 4734-4742.

[231] T. Rajalahti, R. Arneberg, A.C. Kroksveen, M. Berle, K.M. Myhr, O.M. Kvalheim, Discriminating Variable Test and Selectivity Ratio Plot: Quantitative Tools for Interpretation and Variable (Biomarker) Selection in Complex Spectral or Chromatographic Profiles, Analytical Chemistry, 81 (2009) 2581-2590.

[232] E.C.Y. Chan, P.K. Koh, M. Mal, P.Y. Cheah, K.W. Eu, A. Backshall, R. Cavill, J.K. Nicholson, H.C. Keun, Metabolic Profiling of Human Colorectal Cancer Using High-Resolution Magic Angle Spinning Nuclear Magnetic Resonance (HR-MAS NMR) Spectroscopy and Gas Chromatography Mass Spectrometry (GC/MS), J Proteome Res, 8 (2009) 352-361.

[233] X. Lin, Q. Wang, P. Yin, L. Tang, Y. Tan, H. Li, K. Yan, G. Xu, A method for handling metabonomics data from liquid chromatography/mass spectrometry: combinational use of support vector machine recursive feature elimination, genetic algorithm and random forest for feature selection, Metabolomics, 7 (2011) 549-558.

[234] S. Mahadevan, S.L. Shah, T.J. Marrie, C.M. Slupsky, Analysis of metabolomic data using support vector machines, Analytical Chemistry, 80 (2008) 7562-7570.

[235] Y. Liu, Z. Hong, G. Tan, X. Dong, G. Yang, L. Zhao, X. Chen, Z. Zhu, Z. Lou, B. Qian, G. Zhang, Y. Chai, NMR and LC/MS-based global metabolomics to identify serum biomarkers differentiating hepatocellular carcinoma from liver cirrhosis, International Journal of Cancer, 135 (2014) 658-668.

Figure legend

Fig.1. Flowchart of data analysis in metabolomics.

Fig.2. GC-MS total ion chromatograms (TICs) of tangerine peels, before (A) and after peak alignment (B). Retention time shifts in different samples were removed successfully by the multi-scale peak alignment (MSPA) approach.

Fig.3. Concept and outline of model population analysis (MPA).

Fig.4. Prediction error distribution of an informative, uninformative, or interfering variable before (white) and after permutation (gray) of 1000 times. Random sampling is employed. (A). Informative variable; prediction error increases after permutation. (B). Uninformative variable; prediction error shows no significant difference before and after permutation. (C). Interfering variable; prediction error may decrease after permutation.

Table 1. Available databases and libraries for metabolite identification

Name	Access[a]	Current Size[b]	Website
MS Spectral Library			
NIST 14	c	276,248(242,466)	http://www.nist.gov/srd/nist1a.cfm
Wiley Registry of Mass Spectral Data	c	670,000(570,000)	http://onlinelibrary.wiley.com/book/10.1002/9780470175217
GolmMetabolome Database^{RT}	d	26,587	http://gmd.mpimp-golm.mpg.de/
FiehnLib	d	1000	http://fiehnlab.ucdavis.edu/projects/FiehnLib/index_html
MassBank	d	40,889	http://www.massbank.jp/
NIST MS/MS Library	c	234,284(9390)	http://www.nist.gov/srd/nist1a.cfm
ReSpect	d	9017	http://spectra.psc.riken.jp/
METLIN	w	71,808	http://metlin.scripps.edu
Chemical Substance Database			
PubChem Compound Dabatase	d	>53 million	http://www.ncbi.nlm.nih.gov/pccompound
ChemSpider	w	>21 million	http://www.chemspider.com/
Manchester Metabolomics Database	d	42,553	http://dbkgroup.org/MMD/
BiGG Database	w	2835	http://bigg.ucsd.edu/bigg
BioCyc (MetaCyc)		UNKNOWN	http://biocyc.org/
CAS Registry	c	>89 million	http://www.cas.org/
CSLS	w	UNKNOWN	http://cactus.nci.nih.gov/
GDB databases	d	~166 billion	http://www.gdb.unibe.ch/gdb/
Dictionary of Natural	c	240,007	http://dnp.chemnetbase.com/dictionary-search.do?method=view&id=1079994

Products			5&struct=start&props=&&si=
Beilstein database	c	>500 million	http://www.elsevier.com/online-tools/reaxys
KEGG ligand database	d	17,282	http://www.genome.jp/kegg/ligand.html
ChEBI	d	40,211	http://www.ebi.ac.uk/chebi/
HMDB	d	41,806	http://www.hmdb.ca/
KNApSAcK	d	50,899	http://kanaya.naist.jp/KNApSAcK/
LIPID MAPS	d	37,566	http://www.lipidmaps.org/
LipidBank	w	7,009	http://www.lipidbank.jp/
METLIN	w	240,501	http://metlin.scripps.edu
SDBS	w	34,000	http://sdbs.db.aist.go.jp/sdbs/cgi-bin/cre_index.cgi

[a] Access right to the database, c, d and w denote commercial, downloadable and online access, respectively.

[b] Number of unique compounds for corresponding library are provided in the bracket.

[RT] Retention indices are included.

Table 2. Available metabolite identification tools and related tools assisting metabolite identification

Name	Reference	Website
GC-MS Spectrum Identification		
MassLib		http://www.masslib.com/[c]
MOLGEN-MS	[79]	http://molgen.de/?src=documents/molgenms.html[d,w]
Mass Spectrum Interpreter	[81]	http://chemdata.nist.gov/mass-spc/interpreter/[d]
Accurate Mass		
MetWorks		http://www.thermoscientific.com[c]
MetabolitePilot		http://www.absciex.com[c]
Seven Golden Rules	[88]	http://fiehnlab.ucdavis.edu/projects/Seven_Golden_Rules/[d]
SIRIUS	[209]	http://bio.informatik.uni-jena.de/sirius2/[d]
MI-Pack	[103]	http://www.biosciences-labs.bham.ac.uk/viant/mipack[d]
MetaboSearch	[210]	http://omics.georgetown.edu/metabosearch.html[d]
MS/MS Spectrum Prediction		
Mass Frontier		http://www.thermoscientific.com[c,g]
ACD/MS Fragmenter		http://www.acdlabs.com[c,g]
MetISIS	[115]	http://omics.pnl.gov/software/[d]
MyCompoundID	[116]	www.mycompoundid.org[w]
***In silico* Fragmentation**		
FiD	[118]	http://www.cs.helsinki.fi/group/sysfys/software/fragid/[d]
Mass-MetaSite	[119]	http://www.moldiscovery.com/software/massmetasite[c]
MetFrag	[120]	http://c-ruttkies.github.io/MetFrag/[d,w]
FingerID	[121]	https://github.com/icdishb/fingerid[d]
MetFusion	[211]	http://msbi.ipb-halle.de/MetFusion/[w]

CFM-ID	[122, 212]	http://cfmid.wishartlab.com/[d,w]
***De Novo* Analysis**		
SIRIUS^{2}	[123, 124]	http://bio.informatik.uni-jena.de/sirius2/[d]
MAGMa	[177]	http://www.emetabolomics.org/
Molecule Ion Annotation		
PUTMEDID-LCMS	[108]	http://www.mcisb.org/resources/putmedid.html[d]
CAMERA	[109]	http://metlin.scripps.edu/xcms/useful_links.php[d]
IDEOM	[110]	http://mzmatch.sourceforge.net/ideom.php[d]
MZedDB	[213]	http://maltese.dbs.aber.ac.uk:8888/hrmet/index.html[w]
MAIT		
Mass Spectra Deconvolution		
AMDIS	[214]	http://chemdata.nist.gov/dokuwiki/doku.php?id=chemdata:amdis[d]
DeconvolutionReporting Software		http://www.chem.agilent.com/en-US/products-services/Software-Informatics/Deconvolution-Reporting-Software-%28DRS%29/Pages/default.aspx[c]
AnalyzerPro		http://www.spectralworks.com/analyzerpro.html[c]
ChromaTOF®		http://www.leco.com/products/separation-science/software-accessories/chromatof-software[c]
Formula Generation		
OMG	[77]	http://sourceforge.net/projects/openmg/[d]
The Chemistry Development Kit	[215]	http://sourceforge.net/projects/cdk/[d]
Formula To Mass To Formula		http://www.ch.ic.ac.uk/java/applets/f2m2f/[w]
Molecular Formula finder		http://www.chemcalc.org/mf_finder[w]
HiRes		http://hires.sourceforge.net/[w,d]

[c]Commercially available. [d]Freely downloadable to the local site. [w]Freely accessed via web interface. [g]Also suitable for GC-MS spectrum.

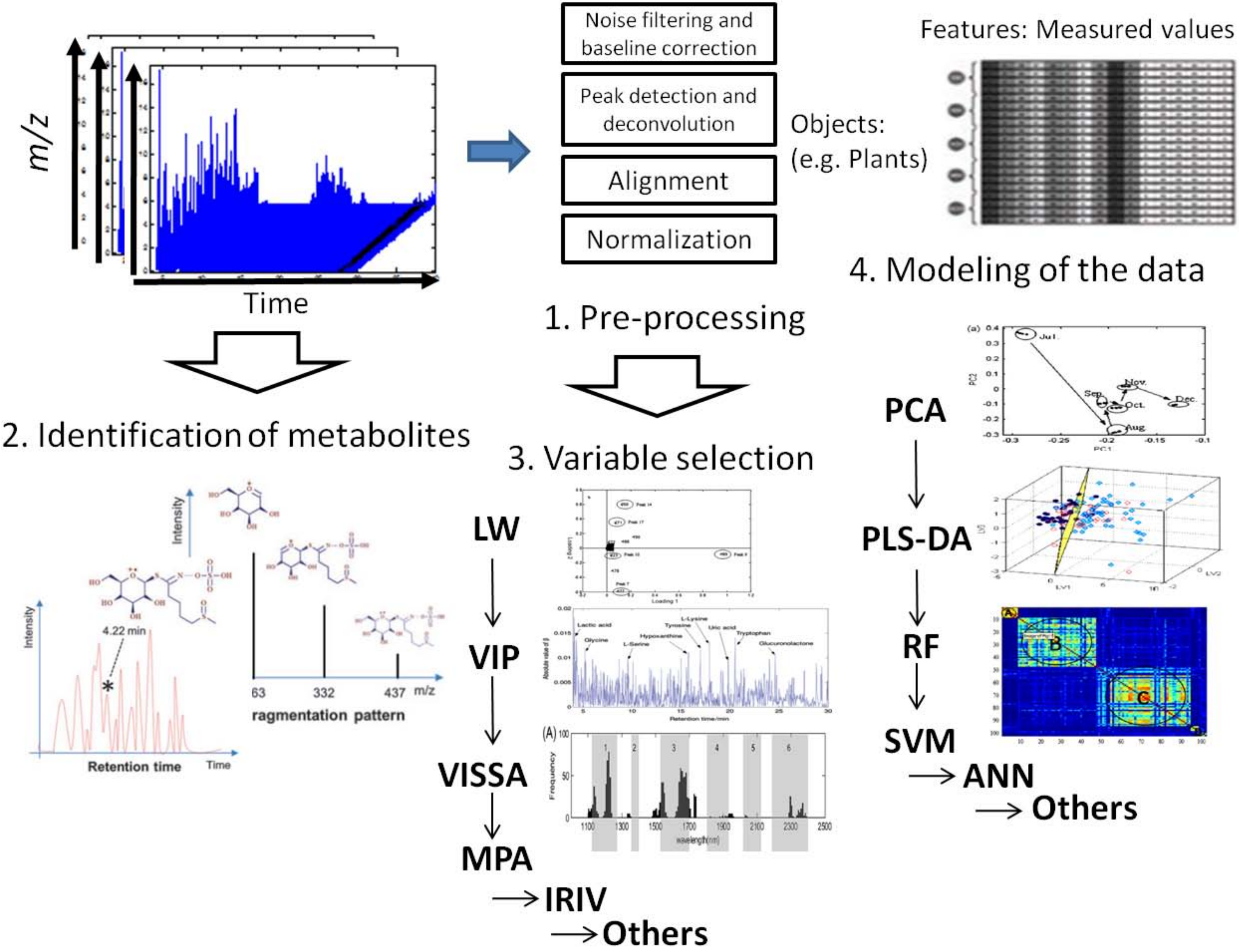
Table 3. A taxonomy of variable selection techniques with the mentioned methods

Methods	Classifier	Interpretability	Consider the interaction effect among variables or not	Variable ranking or subset selection	Computa tion speedy	Reference
PLS-weights	PLS	Based on loading weight matrices of PLS modeling	NO	Ranking	High	[131]
PLS-VIP	PLS	Accumulate the importance of each variable being reflected by loading weights from each latent variable of PLS	NO	Ranking	High	[132]
PLS-regression coefficient	PLS	A single measure of association between each variable and the response.	NO	Ranking	High	[133]
Correlation	No classifier	Calculate simply between variables and classification label.	NO	Ranking	High	[216]
Information gain	No classifier		NO	Ranking	High	[217]
Euclidean distance	No classifier		NO	Ranking	High	[218]
Mutual information	No classifier		NO	Ranking	High	[219]
CARS	PLS	Realize a competitive feature selection based on the absolute regression coefficients.	NO	Subset selection	High	[220]
GA-PLS-DA	PLS-DA	GA is used as an optimal algorithm to find the optimal subset with PLS-DA classifier.	NO	Subset selection	Low	[221]
PSO-SVM	SVM	PSO is used as an optimal algorithm to find the optimal subset with SVM classification method.	NO	Subset selection	Medium	[222]
Random Forest	Decision Tree	Rank the variables by the percent increase of misclassification error when the	YES	Ranking	Medium	[143]

		variable is permuted randomly.				
SPA	PLS-DA	Identify and rank the informative variable based on the difference between the prediction errors of normal and permutated subwindow for each variable.	YES	Ranking	Medium	[145]
MIA	SVM	Give a measure based on the difference between the prediction errors of inclusion and exclusion for each variable with the margin of SVM	YES	Ranking	Medium	[146]
INTERACT	No classifier	Based on inconsistency and symmetrical uncertainty measurements for finding interacting features	YES	Subset selection	High	[142]
VCN	PLS-DA	Compute the complementary information between variables and then effectively discover biomarker with the help of mutual associations of metabolites.	YES	Ranking	Medium	[149]
IRIV	PLS	Find the optimal subset of variables through observing the difference between the prediction errors of inclusion and exclusion for each variable.	YES	Subset selection	Medium	[223]
VISSA	PLS	Search for the optimal variable combinations through shrinking the variable space smoothly	YES	Subset selection	Medium	[148]

Table 4. An overview of multivariate analysis methods for modeling

Method	Category	Advantage	Disadvantage	Applications in metabolomics
PCA	unsupervised	Suit to provide an overview of a large dataset.	Class information is not considered.	[224, 225]
HCA	unsupervised	Suit to provide an overview of the clusters of samples.	Class information is not considered. Variable importance is not obtained.	[226, 227]
SOM	unsupervised	Account for non-linear in the data	Class information is not considered.	[155-157]
PARAFAC2	unsupervised	Can handle shifted data with baseline	Can be more sensitive to noise	[160, 162]
LDA	supervised	Easy and fast. Suit to linear and low dimensional data.	Not suit to high dimensional data	[228, 229]
PLS-DA	supervised	Particularly suit to linear and co-linear data.	Not suit to unbalanced data.	[4, 230, 231]
OPLS-DA	supervised	Particularly suit to linear and co-linear data. Good visualization ability and interpretation ability.	Not suit to unbalanced data.	[169-171, 232]
SVM	supervised	Suit to linear and nonlinear problem. High flexibility in modeling non-linear data.	Lack of transparency of the results. Model tuning is complex	[180-182, 233, 234]
RF	supervised	Suit to linear and nonlinear problem. Resistance to outliers.	Relatively low computation speed	[187-189, 235]



Noise filtering and baseline correction

Peak detection and deconvolution

Alignment

Normalization

1. Pre-processing

Features: Measured values

Objects: (e.g. Plants)

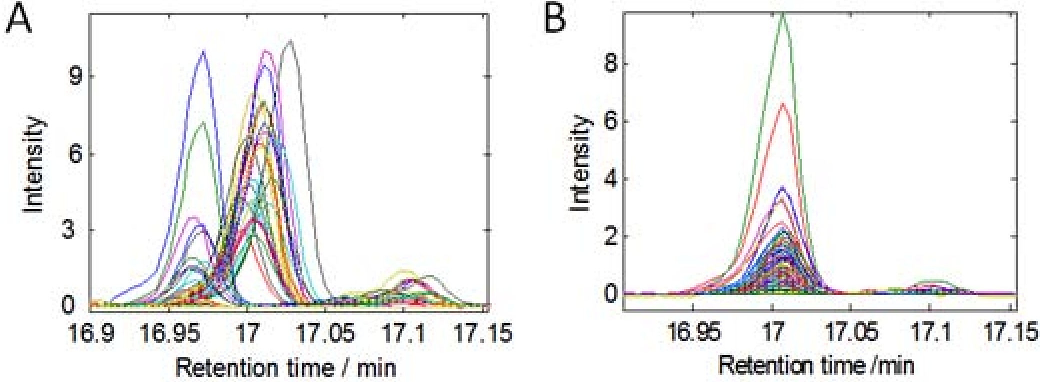
4. Modeling of the data

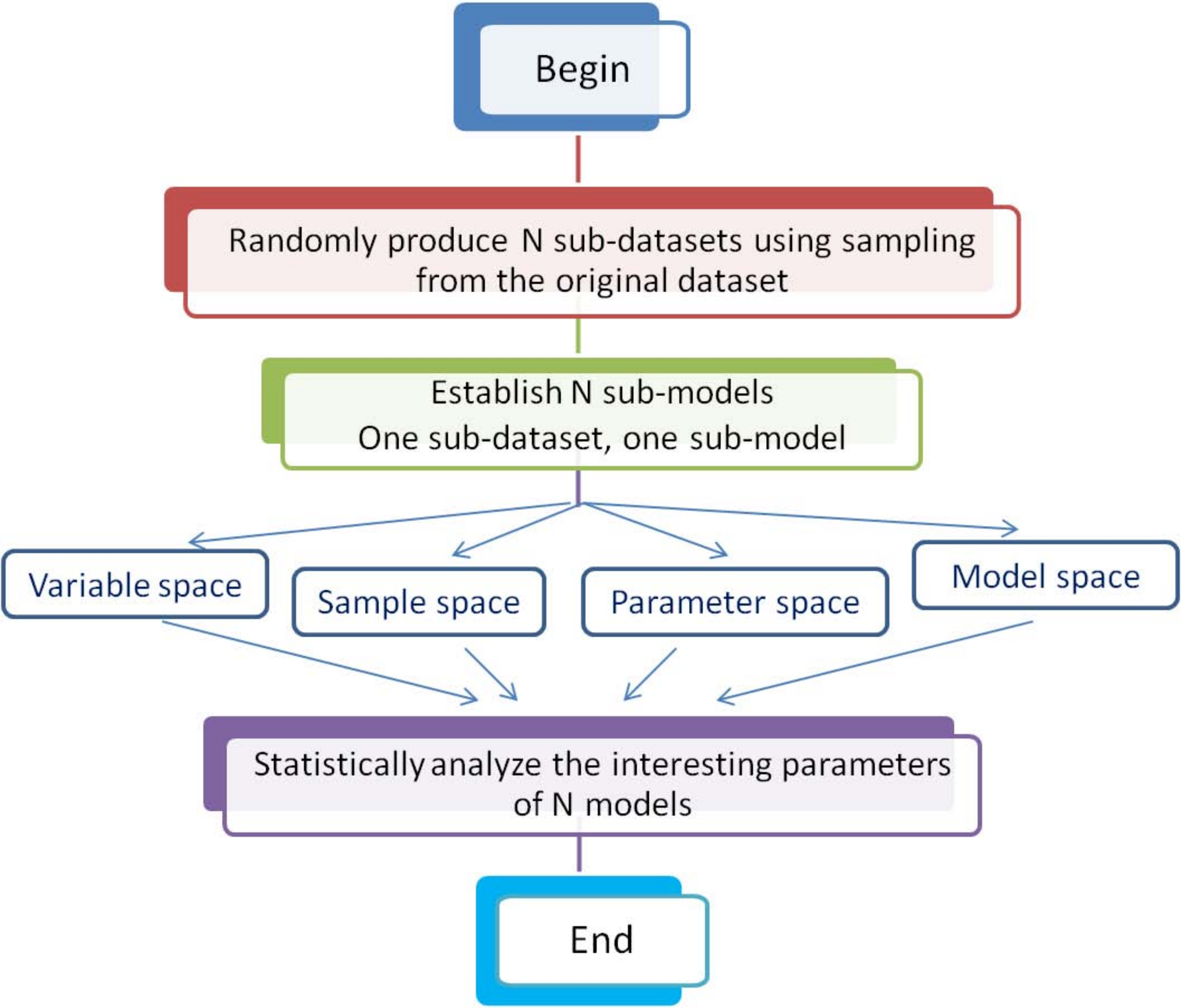
2. Identification of metabolites

3. Variable selection

LW → VIP → VISSA → MPA → IRIV → Others

PCA → PLS-DA → RF → SVM → ANN → Others


A
Intensity
0
3
6
9
16.9
16.95
17
17.05
17.1
17.15
Retention time / min
B
Intensity
0
2
4
6
8
16.95
17
17.05
17.1
17.15
Retention time /min



Begin

Randomly produce N sub-datasets using sampling from the original dataset

Establish N sub-models
One sub-dataset, one sub-model

Variable space

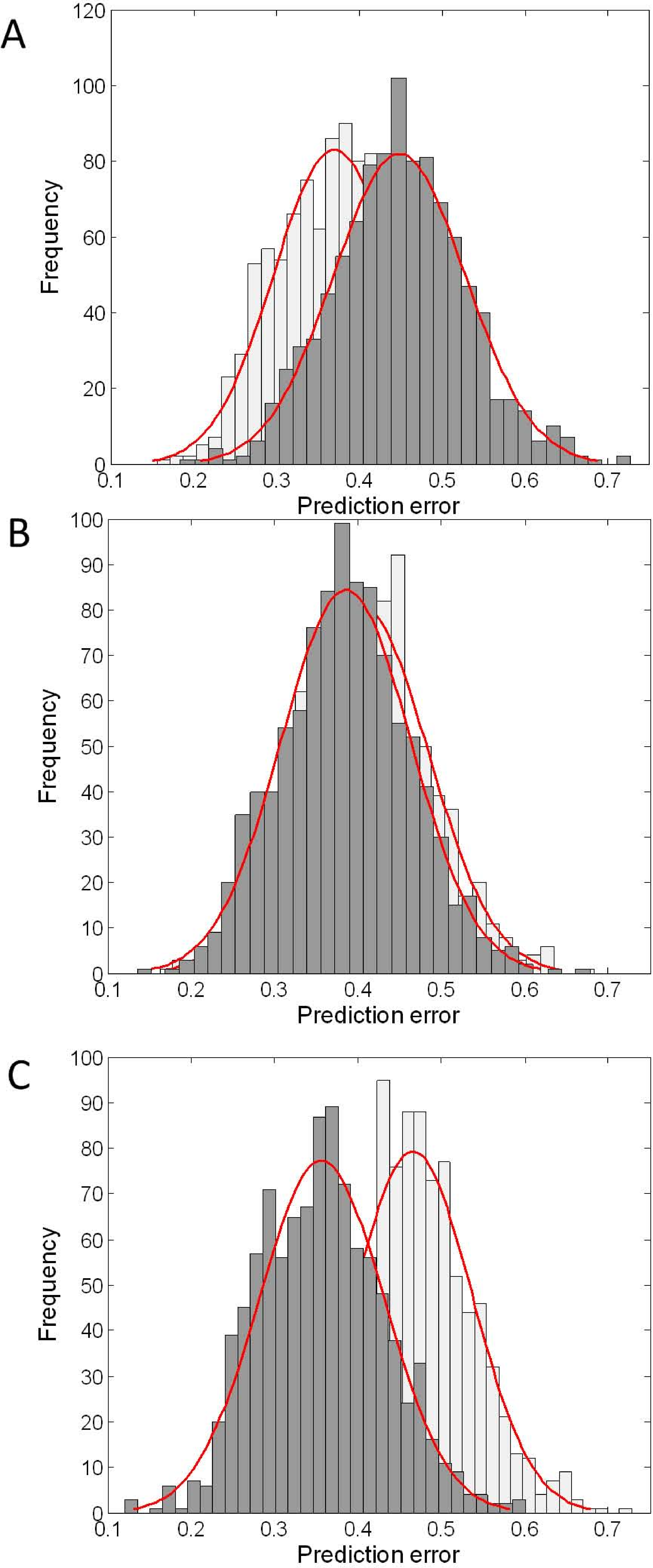
Sample space

Parameter space

Model space

Statistically analyze the interesting parameters of N models

End


A
120
100
80
60
40
20
0
Frequency
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Prediction error
B
100
90
80
70
60
50
40
30
20
10
0
Frequency
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Prediction error
C
100
90
80
70
60
50
40
30
20
10
0
Frequency
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Prediction error

Baichuan Deng received his M.S. and Ph.D. in chemometrics from the Department of Chemistry at the University of Bergen (Norway), in 2012 and 2015, respectively. He is currently a Distinguished Associate Professor in South China Agriculture University. His main research interests include nutritional metabolomics, chemometrics and bioinformatics.

Naiping Dong obtained both a B.S. and a Ph.D. from Central South University (China). He is now research associate in Department of Applied Biology and Chemical Technology, the Hong Kong Polytechnic University. In 2013, He obtained his Ph.D. in analysis of high throughput tandem mass spectrometry in proteomics with Prof. Yizeng Liang as supervisor. Now, his research mainly focuses on applying chemometric and statistical methods in processing chromatographic and mass spectrometric data generated from biological samples.

Shao Liu is the Professor of Xiangya Hospital, Central South University. He has published more than 100 research papers on leading scientific journals. Major research interests include: （1）Natural medicine resources and drug discovery；（2）Quality control of traditional Chinese medicine and metabolomics based on chemometrics.

Yizeng Liang is a professor of chemometrics and analytical chemistry of College of Chemistry and Chemical Engineering, Central South University, China. Editor of “Chemometrics and Intelligent Laboratory Systems” (since 2007). Research interests include analytical chemistry and chemometrics; quality control of traditional Chinese medicines; metabolomics and proteomics; Data mining in chemistry and Chinese medicines. Professor Liang has published more than 420 scientific research papers since 1989 in SCI source journals. Besides, he has published 10 books (8 in Chinese and 2 in English).

Dabing Ren obtained both a B.S. and a Ph.D. from Central South University (China). In 2012, he started his Ph.D. in Professor Yizeng Liang’s group with a focus on countercurrent chromatography (CCC). He mainly used thermodynamic models to study the partition behavior of solutes involved in the CCC separation process. And he developed some useful methods used to correlate and predict the solute partition coefficient in biphasic solvent systems. In 2015, he started career in Kunming University of Science and Technology (China). Now, his research interest mainly covers metabolomics, chromatographic analysis and mass spectroscopy.

Lunzhao Yi is a professor of analytical chemistry and food science of Yunnan Food Safety Research Institute, Kunming University of Science and Technology, China. In 2004, she started her Ph.D. in Professor Yizeng Liang's group with a focus on chemometrics and metabolomics. In 2007, she started career in Central South University (China) until 2014. Her research interests include analytical chemistry, chemometrics, metabolomics and food chemistry. Professor Yi has published more than 50 scientific research papers since 2006 in SCI source journals.

Yonghuan Yun received his B.S. in Pharmaceutical Engineering from Central South University (China) in 2011. He is currently pursuing his Ph.D. in Analytical Chemistry at Central South University under the supervision of Professor Yizeng Liang. His research is focused on developing new algorithm of chemometrics and bioinformatics in the field of near infrared and Raman spectroscopy, metabolomics and genomics.