Metabolomics (2015) 11:1492–1513
DOI 10.1007/s11306-015-0823-6

REVIEW ARTICLE

Computational and statistical analysis of metabolomics data

Sheng Ren[1,2,3] · Anna A. Hinzman[2,4] · Emily L. Kang[3] · Rhonda D. Szczesniak[5,6] · Long Jason Lu[1,2,4,6,7,8]

Received: 17 December 2014 / Accepted: 19 June 2015 / Published online: 28 July 2015

Abstract Metabolomics is the comprehensive study of small molecule metabolites in biological systems. By assaying and analyzing thousands of metabolites in biological samples, it provides a whole picture of metabolic status and biochemical events happening within an organism and has become an increasingly powerful tool in the disease research. In metabolomics, it is common to deal with large amounts of data generated by nuclear magnetic resonance (NMR) and/or mass spectrometry (MS). Moreover, based on different goals and designs of studies, it may be necessary to use a variety of data analysis methods or a combination of them in order to obtain an accurate and comprehensive result. In this review, we intend to provide an overview of computational and statistical methods that are commonly applied to analyze metabolomics data. The review is divided into five sections. The first two sections will introduce the background and the databases and resources available for metabolomics research. The third section will briefly describe the principles of the two main experimental methods that produce metabolomics data: MS and NMR, followed by the fourth section that describes the preprocessing of the data from these two approaches. In the fifth and the most important section, we will review four main types of analysis that can be performed on metabolomics data with examples in metabolomics. These are unsupervised learning methods, supervised learning methods, pathway analysis methods and analysis of time course metabolomics data. We conclude by providing a table summarizing the principles and tools that we discussed in this review.

Keywords Computational · Statistical · Unsupervised learning · Supervised learning · Pathway analysis · Time course data

Electronic supplementary material The online version of this article (doi:10.1007/s11306-015-0823-6) contains supplementary material, which is available to authorized users.

✉ Long Jason Lu
long.lu@cchmc.org;
http://dragon.cchmc.org

[1] Institute for Systems Biology, Jianghan University, Wuhan, Hubei, People's Republic of China

[2] Division of Biomedical Informatics, Cincinnati Children's Hospital Research Foundation, 3333 Burnet Avenue, Cincinnati, OH 45229-3026, USA

[3] Department of Mathematical Sciences, McMicken College of Arts & Sciences, University of Cincinnati, 2815 Commons Way, Cincinnati, OH 45221-0025, USA

[4] Department of Biomedical Engineering, College of Medicine, University of Cincinnati, 231 Albert Sabin Way, Cincinnati, OH 45267-0524, USA

[5] Division of Pulmonary Medicine, Cincinnati Children's Hospital Research Foundation, 3333 Burnet Avenue, Cincinnati, OH 45229-3026, USA

[6] Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital Research Foundation, 3333 Burnet Avenue, Cincinnati, OH 45229-3026, USA

[7] Department of Environmental Health, College of Medicine, University of Cincinnati, 231 Albert Sabin Way, Cincinnati, OH 45267-0524, USA

[8] Department of Computer Science, College of Medicine, University of Cincinnati, 231 Albert Sabin Way, Cincinnati, OH 45267-0524, USA

1 Introduction

Omics is the study of the totality of biomolecules. Just as genomics is the analysis of a complete genome, proteomics is the comprehensive analysis of proteins, and transcriptomics is the comprehensive analysis of gene transcripts, metabolomics is the analysis of the complete set of metabolites, or metabolome, in an organism (Griffin and Shockcor 2004; Oliver 2002). The metabolome represents a large number of compounds including parts of amino acids, lipids, organic acids, or nucleotides. Metabolites are used in or produced by chemical reactions, and their levels can be regarded as the ultimate response of biological systems to genetic or environmental changes. Therefore, it has been suggested that the metabolome is more sensitive to systematic perturbations than the transcriptome and the proteome (Kell et al. 2005).

Cellular processes involve specific metabolites for reactions. Studying and recording these metabolites can lead to the discovery of biomarkers which are measureable biological characteristics that can be used to diagnose, monitor, or predict the risk of diseases (Xia et al. 2012). There are several approaches to studying the metabolome, including target analysis, metabolic profiling and metabolic fingerprinting (Griffin and Shockcor 2004). Target analysis focuses on the quantification of a small number of known metabolites. Metabolic profiling focuses on a larger set of unknown metabolites. Metabolic fingerprinting focuses on the extracellular metabolites. Rather than studying individual metabolites, metabolomics collects quantitative data over a large range of metabolites to obtain an overall understanding of the metabolism associated with a specific condition (Kaddurah-Daouk and Krishnan 2009).

Discovering biomarkers through metabolomics will help diagnose, prevent, and produce drugs for treatment of diseases, including cancer (Griffin and Shockcor 2004), cardiovascular diseases (Griffin et al. 2011), central nervous system diseases (Kaddurah-Daouk and Krishnan 2009), diabetes (Wang-Sattler et al. 2012) and cystic fibrosis (Wetmore et al. 2010). Metabolomics can be a minimally invasive procedure since data can be gathered from plasma, urine, cerebrospinal fluid (CSF), or tissue extracts. It has also been used in studying plants to understand cellular processes and to decode the function of genes, in studying animals to discover biomarkers, in foods research, and in herbal medicines (Putri et al. 2013).

The idea behind metabolomics has been in existence since people have used the sweetness of urine to detect high glucose in diabetes. In the 1960s, chromatographic separation techniques made it possible to detect individual metabolites. Robinson and Pauling's "Quantitative Analysis of Urine Vapor and Breath by Gas–Liquid Partition Chromatography", written in 1971, was the first scientific article about metabolomics (Pauling et al. 1971). The word "metabolome" was coined by Olivier et al. (1998) and defined as the set of metabolites synthesized by an organism. Nicholson et al. first used the word metabonomics in a publication in 1999 to mean "the quantitative measurement of the dynamic multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification" (Nicholson et al. 1999). Griffin, in his paper (Griffin and Shockcor 2004), suggested one of the best definitions of metabolomics given by Oliver is "the complete set of metabolites/low-molecular-weight intermediates, which are context dependent, varying according to the physiology, developmental or pathological state of the cell, tissue, organ or organism".

2 Databases and resources for metabolomics

In 2004, the Metabolomics Society was established to promote the growth, use and understanding of metabolomics in the life sciences. The Metabolomics Society later launched a journal, *Metabolomics,* published by Springer. The Society now has a Twitter feed (@MetabolomicsSoc), which provides news from the Metabolomics Society, its annual international conference, and the Metabolomics journal. METLIN, the first metabolomics database, was also established in 2004. In 2005, the Human Metabolome Project was launched to find and catalogue all of the metabolites in human tissue and biofluids. This metabolite information is kept in the Human Metabolome Database, which produced its first draft in 2007 (Wishart et al. 2013). In recent years the number of papers written about metabolomics has been increasing. More than 800 papers were written in 2009, compared to fewer than 50 in 2002 (Griffiths et al. 2010). As technologies for the quantification and analysis of metabolomics are adapted and improved, the use of metabolomics is expected to continue to grow.

There are many databases containing metabolomics data, and each has different information, ranging from NMR and MS spectra to metabolic pathways. The purpose of metabolic databases is to organize the many metabolites in a way that helps researchers easily identify and analyze metabolomics data. The information found in metabolite databases has continuously been updated in recent years as metabolomics studies have become more widely conducted. Just as metabolomics is a new field and new approaches are still being discovered, metabolomics databases are new and still improving. These databases contain various types of information, including concentration, anatomical location, and related disorders. Among the databases are the Human Metabolome Database (HMDB), MassBank, METLIN, lipid metabolites and pathways strategy (LIPID MAPS), Madison metabolomics consortium database, and Kyoto Encyclopedia of Genes and Genomes (KEGG).

HMDB contains detailed information for 40,444 metabolite entries with chemical, clinical, and molecular biology/biochemistry data (Wishart et al. 2013). Each metabolite in the database includes a "Metabocard" with information including molecular weights, spectra, associated diseases, and biochemical pathways. The purpose of HMDB is to identify all of the metabolites in the human. The 39,293 spectra in MassBank are useful for the chemical identification and structure interpretation of chemical compounds detected by mass spectrometry (MS) (Horai et al. 2010). METLIN is a repository of over 75,000 endogenous and exogenous metabolites from essentially any living creature, including bacteria, plants and animals (Smith et al. 2005). LIPID MAPS is not only the largest database of lipid molecular structures, but the lipid maps resource contains information on the lipid proteome, quantitative estimates of lipids in the human plasma, the first complete map of the macrophage lipidome, and a host of tools for lipid biology, including MS tools, structure tools, and pathway tools (Fahy et al. 2007). The Madison metabolomics consortium database is a resource for metabolomics research based on nuclear magnetic resonance (NMR) spectroscopy and MS (Cui et al. 2008). The current total number of compounds in the Madison metabolomics consortium database is 20,306. Finally, KEGG contains information about metabolic pathways (Kanehisa 2002).

Metabolomics reporting and databases currently suffer from a lack of common language or ontologies (Wishart 2007). The issue is further aggravated by the large number of different types of instruments used in research, each of which has its own language. This makes working with different instruments or other laboratories difficult. A possible solution (Wishart 2007) is standardizing data by entering data into an electronic record-keeping system such as LIMS. The establishment of common reporting standards and data formats would make it much easier to compare and locate metabolomics data.

3 Experimental methods

While tools for transcriptomics and proteomics have made significant improvements in recent years, tools for metabolomics are still emerging. No analytical tool can measure all of the metabolites in an organism, but NMR spectroscopy and MS combined come the closest. That is, using NMR and MS together may result in more complete data than using them individually. NMR spectroscopy and MS are the most common technologies used to collect data from biofluids or tissues. NMR can be used to identify and quantify metabolites from complex mixtures. NMR spectroscopy relies on certain nuclei that possess a magnetic spin and when placed inside a magnetic field can adopt different energy levels that can be observed using radiofrequency waves (Griffin et al. 2011). Proton NMR (^{1}H NMR) is the most commonly used for metabolomics. NMR approaches can typically detect 20–40 metabolites in tissue and 50 in urine samples (Griffin and Shockcor 2004). It is non-destructive because the sample does not come in contact with the detector, usually occurs in a noninvasive manner, requires no chemical derivation, and can be easily reproduced (Armitage and Barbas 2014). A major advantage of NMR is that the signal frequencies observed in an NMR spectrum are directly proportional to the concentration of the nuclei in the sample (Smolinska et al. 2012). However, compared to MS, it has a lower sensitivity, only medium to high abundance metabolites will be detected (Smolinska et al. 2012).

MS-based metabolomics is more commonly used than NMR, judging by the number of publications annually that use each technique (Dettmer et al. 2007). In order to separate the makeup of a mixture, MS is always coupled to other separation techniques. Among all hyphenated MS methods, gas chromatography MS (GC–MS) and liquid chromatography MS (LC–MS) are most popular, as they can be used to detect low-concentration metabolites. GC–MS can be applied to the analysis of low molecular weight metabolites, and it is highly sensitive, quantitative and reproducible (Armitage and Barbas 2014). GC–MS is also preferred in terms of cost and operational issues (Theodoridis et al. 2012). It can typically detect 1000 metabolites (Griffin and Shockcor 2004). LC–MS, which can also be prefixed with high (HPLC) or ultra-high (UPLC) performance, is suitable for the analysis of nonvolatile chemicals, therefore it is complementary to GC–MS (Armitage and Barbas 2014). It has a high sensitivity and is less time consuming than GC–MS, but it can be more expensive (Griffin and Shockcor 2004). One advantage of LC–MS is that it can separate and detect a wide range of molecules and allows for the collection of both quantitative and structural information (Theodoridis et al. 2012).

In addition to the three main stream methods mentioned above, there are also other important spectroscopy and hyphenated methods. Among all spectroscopy methods, vibrational spectroscopy is one of the oldest (Li et al. 2012). There are primarily two vibrational methods utilized: Fourier-transform infrared spectrometry (FT-IR) and Raman spectroscopy (RS). FT-IR is inexpensive and good for high-throughput screening but it is very poor at distinguishing metabolites within a class of compounds

(Griffin and Shockcor 2004) and much less sensitive compared to MS (Patel et al. 2010). Moreover, even combining HPLC with FT-IR, the method HPLC–FT-IR may also have the disadvantage of yielding a low level of detailed molecular identifications (Nin et al. 2012) and the progress in this hyphenated technique is slow (Patel et al. 2010). Raman spectroscopy is an extension of FT-IR and it has been used for the identification of microorganisms of medical relevance (Dunn et al. 2005). However, although there are some advantages to RS over FT-IR, it has similar problems as FT-IR (Griffin and Shockcor 2004). Capillary electrophoresis mass spectrometry (CE-MS) is a powerful separation technique for charged metabolites (Dettmer et al. 2007) and has been predominantly used in targeted metabolomics (Gika et al. 2014). However, since the analytical system stability is not as high as in GC or LC–MS, it is not applied widely in global metabolite profiling (Theodoridis et al. 2012).

The experimental design is an important aspect to consider before conducting any metabolomics experiments. It is a plan of data-gathering studies, which is constructed to control process variation in the experiments and to ensure potential confounders are not present or are well-characterized (Dunn et al. 2012). The process variation in the experiments can be introduced in the sample collection, storage and preparation steps. For example, sample collection time of day (Slupsky et al. 2007), storage and experiment temperature (Cao et al. 2008; Lauridsen et al. 2007) all have an impact on the metabolites profile determined. These conditions and procedures, if not standardized, may lead to spurious biomarkers being reported and may account for a lack of reproducibility between laboratories (Emwas et al. 2014). Therefore, a standard operating procedure is essential to control variation introduced during the sample preparation process. There are some sample procedures for NMR (Emwas et al. 2014) and MS (Dunn et al. 2011) studies. Controlling for potential confounding factors is also critical and is better addressed in experimental design (Broadhurst and Kell 2006). In metabolomics research, the confounding factors are those variables that correlate with both response variables (e.g. disease status) and metabolites concentrations. Such factors include but are not limited to age, gender (Emwas et al. 2014), diet (Heinzmann et al. 2010), physical activity (Enea et al. 2010) and individual metabolic phenotypes (Assfalg et al. 2008). Those factors, if not properly controlled, could lead to failure of discovering true significance or reporting spurious findings (Dunn et al. 2012). For human studies, we should control for confounders first by defining more specific criteria in selecting subjects, since subjects are heterogeneous with respect to demographic and lifestyle factors. This is especially important in defining healthy control (Scalbert et al. 2009). Then, it is recommended to perform sample randomization in order to reduce the correlation between confounders and sample analysis order and instrument conditions (Dunn et al. 2012). More advanced statistical experimental design methods, for example nested stratified proportional randomization and matched case–control design, can be used when outcomes (e.g. disease status) are imbalance (Dunn et al. 2012; Xia et al. 2012). For animal studies, where many confounding factors can be well controlled, large number of samples are not needed compared to human studies (Emwas et al. 2014). In fact, by using some statistical experimental design methods, such as factorial design and randomized block design, researchers can minimize the number of samples used and control most confounders at the same time (Kilkenny et al. 2009). There are rich literatures discussing experimental design in both statistical methodology (Box et al. 1978; Montgomery 2008) and its applications in high throughput biological assays (Riter et al. 2005; Rocke 2004).

4 Data preprocessing

Data preprocessing plays an important role and can substantially affects subsequent statistical analysis results. It takes place after the raw spectra are collected and serves as the link between raw data and statistical analysis. NMR and MS spectra typically show differences in peak shape, width, and position due to noise, sample differences or instrument factors (Blekherman et al. 2011; Smolinska et al. 2012). The goal of preprocessing is to correct those differences for better quantification of metabolites and improved comparability between different samples. Similar preprocessing considerations and methods can be applied to both MS and NMR (Vettukattil 2015).

Preprocessing for NMR typically includes baseline correction, alignment, binning, normalization, and scaling (Smolinska et al. 2012). Baseline correction is a procedure to correct the distortion in the baseline caused by systematic artifacts. It is very important since signal intensities are calculated with reference to the baseline (Vettukattil 2015). Current automatic baseline correction methods are mostly based on polynomial fitting such as local weighted scatter plot smoothing (Xi and Rocke 2008) and splines (Eilers and Marx 1996). After baseline correction, some of unwanted spectral regions are often removed, such as water and other contaminations (Vettukattil 2015). Due to differences in instrumental factors, salt concentrations, temperature and changes of pH, peak shifts can always been observed between samples. Therefore, alignment must be performed in order to correct those shifts. Since most shifts in NMR are local shifts, it is often insufficient to simply perform global alignment by spectral referencing

(Smolinska et al. 2012). Several automatic methods like icoshift (Savorani et al. 2010) and correlation optimized warping (Tomasi et al. 2004) can be used to perform local alignment. After automatic baseline correction or alignment, it is recommended to visually inspect the processed spectra and one can also choose to manually correct baseline and perform alignment (Vettukattil 2015). Binning (also known as bucketing) is a dimension reduction technique, which divides the spectra into segments and replaces the data values within each bin by a representative value. It is a useful technique when perfect alignment is hard to achieve (Smolinska et al. 2012). Traditional equal sized binning is not recommended since peaks can be split into two bins. Some adaptive binning methods such as Gaussian binning (Anderson et al. 2008) and adaptive binning using wavelet transform (Davis et al. 2007) can overcome this difficulty to some extent. However, binning can reduce spectral resolution, therefore, it may be better to avoid binning when spectral misalignment is not serious or when identification of metabolites is more important (Vettukattil 2015). Normalization can remove or correct for some systematic variations between samples, for example sample dilution factors, which is a key factor in analysis of urinary metabolites (Smolinska et al. 2012), in order to make samples more comparable with each other. Typically, normalization is a multiplication of every row (sample) by a sample specific constant (Craig et al. 2006). One popular normalization technique is total integral normalization, where the total spectral intensity of each sample is the constant. When some of the strong signals change considerably between samples, probabilistic quotient normalization can offer more robust results than total integral normalization (Dieterle et al. 2006). Scaling, in metabolomics data analysis, often refers to the column operations that are performed on each feature (spectral intensity or metabolite concentration) across all samples in order to make the features more comparable. Scaling can affect the results of subsequent statistical analysis and we will briefly discuss this problem in Principal Component Analysis. Commonly used scaling methods include but not limited to autoscaling, Pareto scaling and range scaling. More detailed discussion on these methods can be found in van den Berg et al. (2006) and Timmerman et al. (2015).

Data preprocessing for MS typically includes noise filtering, baseline correction, normalization, peak alignment, peak detection, peak quantification and spectral deconvolution. One should note that not all methods use all of the processing steps listed above, nor do they necessarily perform them in the same order (Coombes et al. 2005). Although many preprocessing steps of MS are similar to NMR, there are still some differences. First, a noise filtering step is often associated with MS data preprocessing to improve peak detection (Blekherman et al. 2011). There are many different noise filters, such as Savitzky–Golay filter, Gaussian filters and wavelet based filters, and wavelet based methods provide the best average performance (Yang et al. 2009), due to their adaptive, multi-scale nature (Coombes et al. 2005). Second, a de-isotoping step, which is specific to MS data, can be used to cluster the isotopic peaks corresponding to the same compounds together to simplify the data matrix (Vettukattil 2015). Third, deconvolution is an important step to separate overlapping peaks in order to improve peak quantification. However, deconvolution also has the potential to introduce errors and extra variability to the process (Coombes et al. 2005). There are many software tools available for NMR and MS data preprocessing, a comprehensive summary of software tools can be found in Vettukattil (2015).

5 Overview of metabolomics data analysis

There are two different approaches to processing metabolomics data: chemometrics and quantitative metabolomics. For the former, we directly perform statistical analysis on spectral patterns and signal intensity data and identify metabolites in the last step if needed. For the latter, we identify all metabolites first and then analyze the metabolites' data directly. Compared to quantitative metabolomics, the key advantage of chemometrics profiling is its ability of automated and nonbiased assessment of metabolites data. But it requires a large number of spectra and strict sample uniformly, which are less concerned in quantitative metabolomics. Therefore, quantitative metabolomics is more amenable to human study or studies that require less day to day monitoring (Matthiesen, SpringerLink (Online service) 2010). However, the data analysis methods behind them are similar. In this section, we will discuss four different types of data analysis methods. Note that these methods are not totally independent; they differ only by serving different research purposes. Within each type of data analysis method, we select the most basic, important and widely used models or methods based on published research and review papers we found in metabolomics. The methods we selected cover most core methods currently in use on metabolomics data analysis platforms, such as MetaboAnalyst (Xia et al. 2009, 2012). We also included methods beyond the scope of such platforms. We gave brief introduction of the background, models and algorithms, important facts and potential limitations for each method that we discussed in detail, together with important references and illustrative examples. For the methods not discussed in great detail, we listed a few key references. At the end, we briefly summarized all methods discussed in Table 2 in order to offer readers a clear overview of these methods.

5.1 Unsupervised learning methods

When we receive the data after pre-processing, we may wish to obtain a general idea of its structure. Unsupervised learning methods allow us to discover the groups or trends in the data. The word “unsupervised” here implies that the data we analyze is unlabeled with class membership. The purpose of unsupervised learning is to summarize, explore and discover. Therefore we may only need a few prior assumptions and a little prior knowledge of the data. Unsupervised learning is usually the first step in data analysis and can help visualize the data or verify any unintended issues with the DOE. Among the many different unsupervised learning methods, we will discuss four of the most commonly used methods in metabolomics data analysis.

5.1.1 Principal component analysis (PCA)

If we have a high-dimensional dataset, e.g., dozens or hundreds of metabolites, peak locations, or spectral bins for each subject, we may wish to find only a few combinations of them that best explain the total variation in the original dataset. PCA is one of the most powerful methods to perform this type of dimension reduction (Jolliffe 2005). The main objective of the PCA algorithm is to replace all correlated variables by a much smaller number of uncorrelated variables, often referred to as principal components (PCs), that still retain most of the information in the original dataset (Jolliffe 2005). Although the number of PCs equals to the number of variables, only a limited number of PCs are interpreted. Moreover, if the first few PCs can explain a large proportion of variation in the data, we can visualize the data using a two-dimensional or three-dimensional plot (sometimes called scores and loadings plots) (Fig. 1).

Before we perform PCA, it is recommended to standardize the variables (Jolliffe 2005). This process consists of centering each of the p vectors and standardizing each of them to have variance equal to 1. By doing so, we actually perform PCA on sample correlation matrix instead of sample covariance matrix. When the variables have widely differing variances or use different units of measurement, the variables with the largest variances may dominate the first few PCs (Jolliffe 2005). If those variances are biologically meaningful, we do not need to standardize all variables; otherwise it is highly recommended to perform standardization before performing PCA (Johnson and Wichern 2007). We can calculate the sample PC scores matrix (denoted by $\boldsymbol{T}_{n\times p}$) of all subjects (denoted by $\boldsymbol{X}_{n\times p}$ which is the original dataset with n subjects and p variables). The process can be expressed by the following formulas $\boldsymbol{T} = \boldsymbol{XP},\ or\ t_{ij} = \boldsymbol{x}_i^T\boldsymbol{p}_j \cdot \boldsymbol{P}_{p\times p}$ is the weight matrix (i.e., loading matrix), $\boldsymbol{x}_i$ is the ith row of $\mathbf{X}$, which represents the values of the ith subject, $\boldsymbol{p}_j$ is the jth column of $\boldsymbol{P}$ which represents the weights of all p original variables on the jth PC, t_{ij} is called the score of the ith subject on the jth PC. All of the p PCs are uncorrelated and their variances are eigenvalues of sample correlation matrix (or sample covariance matrix if not standardized). The largest eigenvalue corresponds to the first PC, the second largest eigenvalue corresponds to the second PC, and so on. In order to measure the contribution of a PC to the total sample variance, we use its corresponding eigenvalue divided by the summation of all eigenvalues as a measure. This is the percentage of variance explained by the corresponding PC. There is no rule for how many PCs to keep; we usually make the decision by checking the “variance explained” measure mentioned above or using a scree plot (Johnson and Wichern 2007).

Here we use a simple example of microbial metabolomics (Hou et al. 2012) to illustrate the basic idea of how to use PCA. Other examples of how PCA was used in metabolomics studies can be found in Heather et al. (2013) and Ramadan et al. (2006). In this study, the authors used PCA to perform strain selection and to discover unique natural products. They assumed bacterial strains producing the same secondary metabolites would group together. The PCA scores and loadings plots of 47 strains are shown in Fig. 1.

In PCA, the scores plot is mainly used to discover groups while the loadings plot is mainly used to find variables that are responsible for separating the groups. In the loadings plot, we mainly check the points that are further from the origin than most other points in the plot. For the scores plot (Fig. 1a), we can see seven identifiable groups; these groups were identified by human eye. In the loadings plot (Fig. 1b), point 1 corresponds to a compound that is responsible for separating group G7 from the other groups. We cannot judge which of the points in the loadings plot are responsible for separating subjects into groups using only these two plots; instead, we should go back to the loading matrix $\boldsymbol{P}_{p\times p}$ to check the weights. Furthermore, the groups shown in the PCA scores plots are not necessarily the biologically meaningful groups. PCA often provides a clue for further investigation. Although the authors mentioned 74 PCs that were generated, which explained 98 % of the variation in the data set, they did not show how much of the variance was explained by the first two PCs. Note that PCA may not be powerful if the first few PCs cannot explain a large proportion of variability of the sample. For example, if the first two PCs in the plot account for only 50 % of the total variation, then the visualization results may be misleading, so we cannot identify the groups of the strains only by graphs.

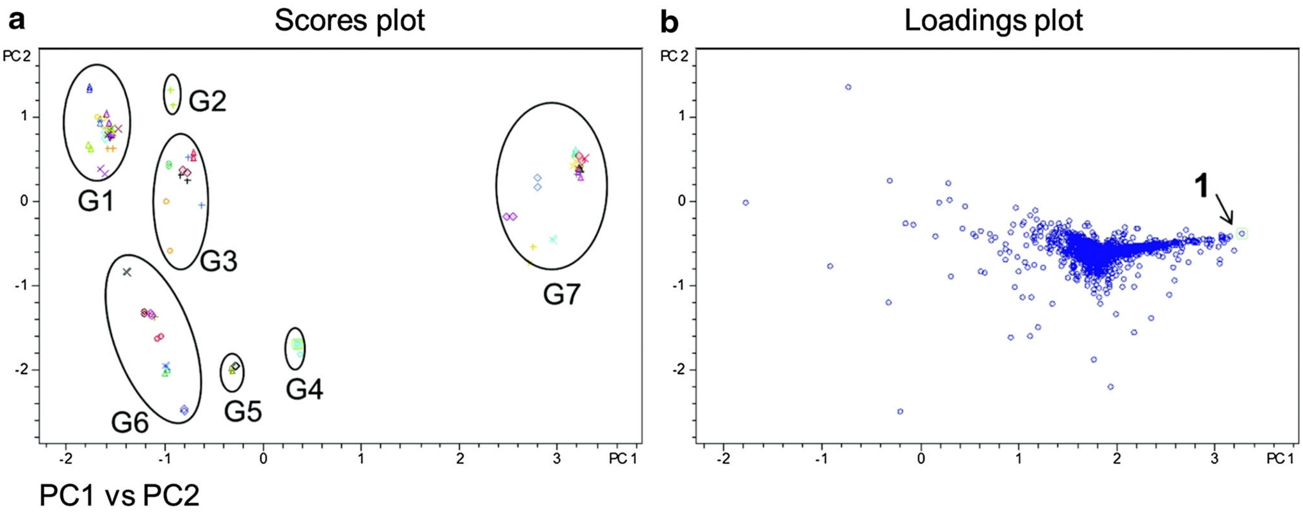

Fig. 1 Using PCA to discover natural products unique to group 7 (Hou et al. 2012)

5.1.2 *Clustering*

Unlike PCA, clustering analysis explicitly aims at identifying groups in the original dataset. All clustering algorithms group the subjects such that the subjects in the same group or cluster are more similar to each other than to subjects in other groups. Different algorithms may use different similarity measures, such as various distances and correlation coefficients. Among the many different clustering methods, we will only introduce two most common methods in metabolomics as well as in many other areas of data analysis.

5.1.2.1 K-means clustering K-means clustering is centroid-based clustering and is a type of partitioned clustering method (Hartigan and Wong 1979). Here centroid-based indicates that each cluster can be represented by a center vector, which may not be an observation in the original dataset, $\boldsymbol{X}_{n \times p}$. Partitioning requires each subject to appear in exactly one cluster. K-means clustering divides subjects into k non-overlapping clusters such that each subject belongs to the nearest mean of the corresponding cluster. If all of the variables are numerical, we generally choose Euclidean distance as the metric that distinguishes between the subject and the center vector. When using Euclidean distance fails to find meaningful clusters, we may consider using other distance metrics, for example Mahalanobis distance, which has the form $\sqrt{(\boldsymbol{x}-\boldsymbol{y})^T \boldsymbol{A}(\boldsymbol{x}-\boldsymbol{y})}$. It is clear that Euclidean distance is a special case of Mahalanobis distance when **A** is identity matrix. In general, **A** is a covariance matrix with unknown form, a general and efficient algorithm (Xing et al. 2002) can be used to learn the parameters in **A** together with performing K-means clustering. Variations of the K-means clustering algorithm include using median instead of mean as the center vector and assigning weights to each variable. There are also some drawbacks to K-means clustering methods. The major problem is that the number of clusters, “K”, is an unknown parameter, and thus we must determine K before we employ the algorithm. Visualization tools such as PCA, multidimensional scaling (MDS) and self-organizing map (SOM) may help to determine K. There are also some statistical methods for estimating K, the most widely used methods are gap statistic (Tibshirani et al. 2001) and weighted gap statistic (Yan and Ye 2007). Another problem is that K-means assumes that subjects in each cluster are distributed spherically around the center (Hamerly and Elkan 2003). This assumption may lead to poor performance on data with outliers or with clusters of various sizes or non-globular shapes (Ertöz et al. 2003). An adaptive Fuzzy c-means clustering (Gunderson 1982, 1983) can be used in these cases. Fuzzy c-means clustering (Bezdek et al. 1981; Dunn 1973) is an extension of K-means where each data point belongs to multiple clusters to a certain degree, which is called membership value. The adaptive Fuzzy c-varieties clustering algorithm (Gunderson 1983) which is based on Gunderson (1982) is a data dependent approach that can seek out cluster shapes and detect a mixture of clusters of different shapes. Therefore, it removes the limitation of imposing non-representative structures in K-means and Fuzzy c-means clustering. An alternative way to solve the arbitrary cluster shapes problem is using kernel K-means (Schölkopf et al. 1998) and it was suggested in Jain (2010). Another limitation of K-means and other clustering methods is that some variables may hardly reflect the underlying clustering structure (Timmerman et al. 2010). One possible way to solve that problem is performing K-means in reduced space (De

Soete and Carroll 1994). Many methods have been proposed to improve the original reduced K-means (De Soete and Carroll 1994), including factorial K-means (Vichi and Kiers 2001) and subspace K-means (Timmerman et al. 2013). An alternative solution is using variable selection (Steinley and Brusco 2008) or variable weighting (Huang et al. 2005). An illustrative example of using K-means clustering on metabolites profiles to explore dietary intake patterns can be found in O'Sullivan et al. (2011). More details of how to use fuzzy c-means in metabolomics were explained in Li et al. (2009).

5.1.2.2 Hierarchical clustering Hierarchical clustering (Johnson 1967) builds a hierarchy and uses a dendrogram to represent the hierarchical structure. Unlike K-means clustering, hierarchical clustering does not provide a single partition of the dataset. It only shows the nested clusters organized as a hierarchical tree and lets the user decide the clusters. In order to form the hierarchical tree, we must choose the similarity metric between pairs of subjects and pairs of clusters. The similarity metric between two subjects is distance. Different clusters will form by using different distance functions. Commonly used distance functions include Euclidean distance, Manhattan distance, Mahalanobis distance and maximum distance. A general discussion of distance functions can be found in Jain et al. (1999). Based on the distance function we choose, we can construct a distance matrix for all subjects before we perform hierarchical clustering. Then we need to select a linkage function, which is the similarity metric for pairs of clusters. Different linkage functions will lead to different clusters. Commonly-used linkage functions include single linkage, complete linkage and average linkage. A general discussion of linkage functions can be found in Hastie et al. (2009). An advantage of hierarchical clustering over K-means is that it does not stop at a special number of clusters found, but will continue to split until every object in the dataset belongs to the same cluster. Therefore, the hierarchical tree may provide some meaningful finding of the real structure of the dataset. However, it also has some drawbacks, for example it may not be robust to outliers.

Hierarchical clustering is often used together with a heat map to visualize the data matrix. Heat maps use different colors to represent different entries in the data matrix. The entries in the data matrix can be either the values of some variables or some statistic, e.g., correlation coefficient or p-value. We can add hierarchical clustering trees on the side or top of the heat map so that we can clearly see the structure of the data. There is a good example of this kind of representation from (Poroyko et al. 2011).

In this paper, the authors studied the effect of different diets on selecting different intestinal microbial communities using a metabolomics approach. They used a heat map to show the significant p-values associated with the relationship between metabolites and bacterial taxa in piglet cecal content (see Supplementary Fig. 1). The dendrogram on the left side shows the hierarchical structure of different genus of bacteria; the one on the top shows the hierarchical structure of different metabolites. This graph helps us visualize the degree to which bacteria were associated with the same or different metabolites. Another similar example of using hierarchical clustering together with heat map representation can be found in Draisma et al. (2013), which used hierarchical clustering to analyze blood plasma lipid profiles of twins.

5.1.3 Self-organizing map (SOM)

SOM is a powerful tool to visualize high-dimensional data (Kohonen 1990); it can thus help us visually discover the clusters in the data. It is an arrangement of nodes in a two-dimensional (may also be 1D or 3D) grid. The nodes are vectors whose dimension is the same as input vectors. Since SOM is a type of artificial neural network (ANN), the nodes are also called neurons. Unlike other types of ANNs, SOM uses a neighborhood function to connect adjacent neurons. The neighborhood function is a monotonically decreasing function of iterated times and the distance between the neighborhood neurons and neuron that matches the input best. It defines the region of influence that the input pattern has on the SOM and the most common choice of the function is Gaussian function. More technical details can be found in Kohonen (1998). In this way, data points located closely in the original data space will be mapped to neurons nearby. Every node can thus be treated as an approximation of a local distribution of the original space, and the resulting map retains its topological structure. Before implementing SOM, one may choose the number of nodes and the shape of grids, either hexagonal or rectangular. Using a hexagonal grid implies that one node will have six bordering nodes. After numerous updating cycles, each subject is finally assigned to a corresponding neuron, and neighboring neurons can be treated as mini clusters. These mini clusters may give hints of metabolic patterns. In order to see the clusters clearly, a unified distance matrix (U-matrix) representation can be constructed on the top of SOM. The U-matrix nodes are located among all neighborhood neurons, and the color codes of each node represent the average Euclidean distance among weight vectors of neighboring neurons (Ultsch 2003). Therefore, the gaps between clusters can be shown by the colors of U-matrix nodes.

There is an example of using SOM with a U-matrix in metabolomics research (Haddad et al. 2009). In their paper, the SOM (see Supplementary Fig. 2) was trained with the metabolome data from three different fermentations of

Corynebacterium glutamicum. The color codes show different Euclidean distances between each output node and its four bordering nodes. White–purple represents short distance, which implies the subjects (black points) have similar metabolic patterns, while green–yellow represents long distance, which implies the subjects have different metabolic patterns. We can see that the clusters were clearly separated by the green–yellow gaps. SOM have also been used to visualize metabolic changes in breast cancer tissue (Beckonert et al. 2003) and to improve clustering of metabolic pathways (Milone et al. 2014).

5.2 Supervised learning methods

The purpose of supervised learning is different from that of unsupervised learning. Supervised learning methods are widely used in discovering biomarkers, classification, and prediction, while unsupervised learning methods cannot complete these tasks. However, these distinctions do not imply that supervised methods are superior to unsupervised methods; rather, each was designed to achieve different objectives of analysis. Supervised learning deals with problems or datasets that have response variables. These variables can be either discrete or continuous. When the variables are discrete, e.g., control group versus diseased group, the problems are called classification problems. When the variables are continuous, e.g., metabolite concentration or gene expression level, the problems are called regression problems. The purpose of supervised learning is to determine the association between the response variable and the predictors (often referred to as covariates) and to make accurate predictions. It is called supervised learning because one or more response variables are used to guide the training of the models. Usually both a training step and a testing step are included. Supervised learning algorithms are applied on the training dataset to fit a model, and then the testing dataset is used to evaluate the predictive power. In these steps, we may encounter the following problems: How to extract or select better predictors? How to evaluate the fitness and predictive power of the model? And what learning methods and algorithms to choose?

For the first problem, the process of choosing relevant predictors is called feature selection or variable selection. There are three main types of feature selection methods: Wrapper, Filter and Embedded (Guyon and Elisseeff 2003). The Wrapper method scores subsets of variables by running every trained model on the test dataset and selecting the model (subset of variables) with the best performance. The Filter method scores subsets of variables by easy-to-compute measures before training the models. The Embedded method, just as its name implies, completes feature selection and model construction at the same time. For the second problem, we first need goodness of fit statistics to measure model fit and predictive power. Commonly used statistics include but are not limited to: root mean square error (RMSE) for regression; sensitivity, specificity and the area under the receiver-operating characteristic (ROC) curve for binary classification. In addition, we need test datasets to assess the predictive power and avoid over-fitting issues. Ideally, model validation should be performed using independent test datasets; however, gathering objective data can be expensive due to limited resources and other pragmatic factors. Therefore, various resampling methods are often used in order to reuse the data efficiently. These methods include cross validation, bootstrapping, jackknifing, randomization of response variables and some others. Among all of them, bootstrapping and cross-validation are used more often in validating supervised learning models (Hastie et al. 2009). Commonly used cross-validation methods include k-fold validation and random sub-sampling validation. Together with resampling methods, we can obtain a set of goodness-of-fit statistics. By averaging them we can obtain a single statistic indicating the fitness and predictive power of the model. For example, if the average of k RMSEs, which can be the result of k-fold validation of model A, is lower than those average RMSEs of other models, then we can conclude that model A is the better one under RMSE criteria. For the third problem, there are many different supervised learning methods to choose from. Here we briefly introduce two of the most widely used methods in metabolomics.

5.2.1 Partial least squares (PLS)

PLS (Wold 1966) is a method of solving linear models. A general linear model has the form $\boldsymbol{Y} = \boldsymbol{X\beta}+\boldsymbol{\varepsilon}$, where $\boldsymbol{Y}$ is the response variable, it can be a vector (one variable) or a matrix (several variables); $\boldsymbol{X}$ is the design matrix whose columns represent variables and rows represent observations; $\boldsymbol{\beta}$ is the vector (matrix) of parameter coefficients and $\boldsymbol{\varepsilon}$ is the random error vector (matrix) (Martens 1992). Generally, we use the ordinary least square solution of $\boldsymbol{\beta}$, which is $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$. However, in metabolomics analyses, we always have a large number of variables, such as metabolites, peak locations, and spectral bins, but a relatively small number of observations. Moreover, these variables may be linearly dependent, and thus it will be impossible to use the conventional least squares method to solve for $\boldsymbol{\beta}$ in the linear regression model, since it is impossible to invert the singular matrix $\boldsymbol{X}^T\boldsymbol{X}$. At first, principal component regression (PCR) was introduced to solve this problem. Instead of using all original variables, PCR uses the first few PCs from PCA to fit the linear regression model. But it is not clear whether those PCs have high correlation with response variables $\boldsymbol{Y}$ or not. Therefore, PLS was introduced to tackle this problem

(Wold et al. 1984). PLS may also stand for projection to latent structures, which implies how this method works. The underlying model of the PLS method has the form (Wold et al. 2001):

$$\begin{cases} \boldsymbol{Y} = \boldsymbol{UQ}' + \boldsymbol{F} \\ \boldsymbol{X} = \boldsymbol{TP}' + \boldsymbol{E} \end{cases}$$

Similar to PCA, $\boldsymbol{T}$ and $\boldsymbol{U}$ are called $\boldsymbol{X}$ and $\boldsymbol{Y}$ scores, which are matrices formed by latent variables; $\boldsymbol{P}$ and $\boldsymbol{Q}$ are called $\boldsymbol{X}$ and $\boldsymbol{Y}$ loadings, which can be thought of as weight matrices; **E** and **F** are residuals, which are the remaining amounts that cannot be explained by latent variables. The latent variables, which can be thought of as factors, are linear combination of the original $\boldsymbol{X}$ and $\boldsymbol{Y}$ variables, i.e., for each latent variable $\boldsymbol{t}$ and $\boldsymbol{u}$, $\boldsymbol{t} = \boldsymbol{Xw}$ and $\boldsymbol{u} = \boldsymbol{Yc}$; $\boldsymbol{w}$ and $\boldsymbol{c}$ are called weight vectors. These latent variables may have chemical or biological meanings. The PLS method finds a best set of $\boldsymbol{X}$ variables that can explain most of the variation in **Y**. Namely, we should find each latent variable **t** and **u**, such that, under some orthogonal conditions, their covariance reaches its maximum value (Abdi 2010). There are many variants of the PLS and corresponding algorithms, which may have different orthogonal conditions and different methods to estimate scores and loading matrices. It is important to note that PLS is different from PCA and PCR. First, PCA is an unsupervised learning method while PCR and PLS are supervised learning methods. Second, PCR uses the first few PCs in the PCA as predictors to fit a latent variable regression. Thus, predictors in PCR may only explains the variance in $\boldsymbol{X}$ itself while a PLS model tries to find the multi-dimensional direction in the $\boldsymbol{X}$ space that explains the maximum variance direction in the $\boldsymbol{Y}$ space. Therefore, the PLS method may often perform better than PCR.

The only parameter we need to specify in PLS is the number of components to keep. There are two approaches. First, we can use plots to help us decide the components, e.g., the $\boldsymbol{Y}$ and $\boldsymbol{X}$ scores plot or R^2 plot. Another approach is using resampling methods together with a measure of goodness of fit or predictive power. We can select different numbers of components and check their goodness of fit or predictive power. Since the PLS method is a dimension reduction method itself, feature selection is not a required step in PLS. However, in order to improve interpretation, robustness and precision, there are also some feature selection methods that can be used with PLS. For example, we can use a two-sample t test, a filter method, to select variables before running PLS. Sparse PLS (SPLS), which is an embedded method, imposes sparsity when constructing the direction vectors, thereby improves interpretation and achieves good prediction performance simultaneously (Chun and Keleş 2010). Another method called orthogonal projections to latent structures (OPLS) (Trygg and Wold 2002), can be embedded as an integrated part of PLS modeling to remove systematic variation in $\boldsymbol{X}$ that is orthogonal to $\boldsymbol{Y}$, thus also enhancing the interpretation of PLS.

Although PLS was first designed to deal with regression problems, it can also be used in classification problems. One popular method is called PLS-discriminant analysis (DA) (Boulesteix 2004; Nguyen and Rocke 2002). In PLS-DA, $\boldsymbol{Y}$ is a vector whose values represent class memberships. When considering model validation in PLS or PLS-DA, Predicted Residual Sum of Squares (PRESS), Q^2 *and* R^2 can be used in addition to the commonly used diagnostic methods mentioned above. Note that R^2 is a measure of fitness of the model to the training data set while Q^2 and PRESS are used to evaluate the predictive power of the model. For the PLS-DA method, it is recommended to use a double cross-validation procedure (Szymanska et al. 2012) along with the number of misclassifications and the area under the ROC curve as diagnostic statistics. Using similar algorithms as PLS-DA, other variants of PLS, like SPLS and OPLS mentioned above, can also be extended to classification problem, where they called SPLS-DA (Chung and Keles 2010) and OPLS-DA (Bylesjö et al. 2006).

Here we use an example to illustrate some application aspects of the PLS model (Kang et al. 2011) (Fig. 2). The authors used OPLS-DA to classify coronary heart failure (CHF) groups and control groups. Figure 2a is a scores plot of the first two components, which shows the similarities and dissimilarities of the subjects. In this plot, we can see that the diseased and control groups can be clearly separated by the OPLS-DA model. Figure 2b is the corresponding loadings plot. Different from a PCA loadings plot, the metabolites identified are responsible for the classification. The upper section of Fig. 2b shows that metabolites increased in the control group while the lower section shows that metabolites increased in heart failure group. They ran the OPLS-DA model with NMR spectra data (which is the input data matrix **X**) and then identified the metabolites responsible for the separation using results in Fig. 2b.

PLS method has been successfully applied in numerous metabolomics studies for disease classification and biomarker identification (Marzetti et al. 2014; Velagapudi et al. 2010; Zhang et al. 2008). Note that the PLS method sometimes can also be used as a dimension reduction (feature selection) tool rather than as a classification method (Bu et al. 2007).

5.2.2 *Support vector machine (SVM)*

Since metabolomics data is represented in matrix form, every subject is a row vector; thus, each subject can be

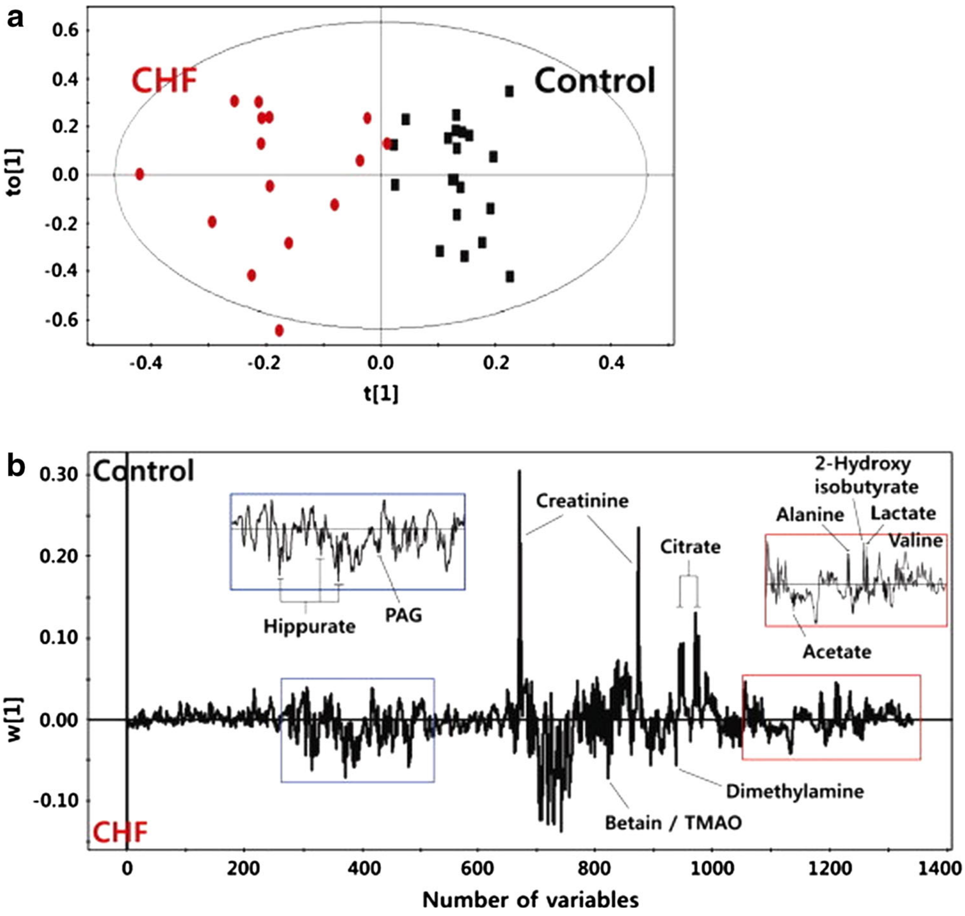

Fig. 2 PLS example (OPLS-DA classification): OPLS-DA scores plot (**a**) and loadings plot (**b**) of urinary NMR spectra obtained from healthy controls and patients with coronary heart failure (Kang et al. 2011)

viewed as a point in a p-dimensional space where p is the number of variables. If we can separate the data into two groups, intuitively we can find a "gap" between these two groups in the p-dimensional space. SVM tries to find such a gap that is as wide as possible (Cortes and Vapnik 1995). The margins for the gap are defined by support vectors, i.e., the points located on the margins. SVM is trained to determine the support vectors. The boundary in the middle of the gap that separates the data is called the separating hyper plane. The prediction is done by deciding to which side of the hyper plane new subjects (observations) belong.

The original SVM algorithm is a linear classifier, which means it can only produce a hyper plane ($p - 1$ dimension plane in p-dimensional space) to classify the data. We aim to find the largest margin, i.e., the largest distance between two groups, which can be solved by quadratic programming. The related mathematical expression of this problem has been documented by Bishop (Bishop 2006). However, it is quite common that the data cannot be linearly separated, i.e., a separating hyper plane does not exist. In this case, we can use kernel trick to map the original data to a higher dimensional space so that it can be linearly separated in that space. Kernel trick or kernel substitution is very useful in extending algorithms. It substitutes the inner product (linear kernel) with other kernels. Commonly used kernels include the polynomial kernel and the Gaussian kernel (Bishop 2006).

Another problem is the stability of the algorithm. If there are outliers or mislabeled data, the original SVM may give an unsatisfactory classification result. In this case, we can use SVM with a soft margin to solve this problem. The soft margin SVM allows some misclassification in the training step by adding slack variables to the original objective function. This modification changes our objective from maximizing the margin between two groups to maximizing the margin as cleanly as possible. Here "clean" means only a few misclassified subjects.

Since SVM is a well regularized method, it does not always require a feature selection step. However, there are some feature selection methods that can be used to enhance the performance of SVM and lower its computational cost. Examples include the recursive feature elimination (RFE) method and L1 norm SVM (Guan et al. 2009). As discussed in Sect. "5.2.1", similar validation methods and diagnostic measures can be applied to the SVM algorithm. Moreover, these validation methods and diagnostic measures can help us select optimum parameters such as which kernel to choose and its parameters.

Compared to the PLS-DA method, one minor disadvantage of SVM is that it is often difficult to visualize and interpret the classification result using plots; especially when the number of variables is large. However, for classification purposes, we still recommended SVM over most other methods (Mahadevan et al. 2008), and SVM has been widely used for classification and prediction in metabolomics research, especially in cancer research (Guan et al. 2009; Henneges et al. 2009; Stretch et al. 2012). Moreover, SVM can also be used in regression problems, where it is called support vector regression (SVR) (Brereton and Lloyd 2010). Detailed discussions on SVR and its applications in metabolomics and chemometrics can be found in Li et al. (2009).

5.3 Pathway analysis methods

Pathway analysis allows us to detect the biological mechanisms in which identified metabolites are involved. Some metabolic pathway analysis methods are directly borrowed from gene pathway analysis, e.g., over-representation analysis (ORA) and enrichment score. Here we provide a brief introduction to ORA, functional class scoring (FCS) and some pathway simulation methods.

5.3.1 Over-representation analysis (ORA)

During the research there may be cases where we have a list of metabolites identified and we only want to know which pathways are involved in the samples being studied. There are many metabolic pathway databases available on the internet. In this case the pathways are already specified and we only need to test which pathway is significantly involved based on the available samples. This kind of pathway analysis is called knowledgebase-driven pathway analysis (Khatri et al. 2012). Among all of the knowledgebase-driven pathway methods, ORA is well-known and the simplest. ORA is used to test whether pathways are significantly different between two study groups. Before performing ORA, we should have a list of metabolites showing significant differences between two groups. This can be done by using two sample tests, e.g., *t* test or nonparametric tests, for all metabolites. Then we select metabolites whose significance reach a predetermined threshold for false discovery rates (FDRs) or p-values. Next, we can perform ORA, which is equivalent to a 2 × 2 contingency table (Table 1) test in statistics. After we obtain all related pathways from knowledgebase, we can count the number of metabolites in or not in both a known pathway and the list, then perform a statistical test for whether this pathway is significantly involved for each known pathway. The most frequently used tests include the Chi-square test, which requires larger sample sizes, and Fisher's exact test, which is more appropriate for smaller cell counts in the table and uses a hypergeometric distribution (Agresti 2014).

Table 1 ORA analysis table

	Metabolites on list	Metabolites not on list	Subtotal
Metabolites in the pathway	a	b	a + b
Metabolites not in the pathway	c	d	c + d
Subtotal	a + c	b + d	

5.3.2 Functional class scoring (FCS)

ORA is simple to perform, but it has several drawbacks. First, much information is lost since only the most significant metabolites are used and the rest are ignored, and only the number of identified metabolites is considered. Second, the optimal threshold is unclear. Third, it assumes improper independence. For example, it assumes that each metabolite and pathway is independent of others, but in reality, this assumption may not be valid. Therefore, another class of methods called FCS was proposed to address some of the limitations in ORA. A general framework of univariate FCS methods works as follows: First, obtain single-metabolite statistics (e.g., t-statistic and z-statistic) by computing differential expression of individual metabolites. Second, aggregate those single-metabolite statistics to compute a pathway level statistic. This pathway level statistic can be univariate or multivariate. Commonly used univariate pathway level statistics include mean, median and enrichment score (Holmans 2010). For multivariate statistics, a widely used statistic is Hotelling's T^2 statistic, which has an F distribution under the null hypothesis (Johnson and Wichern 2007). The final step is hypothesis testing. There are two kinds of null hypothesis: competitive and self-contained. A competitive test considers metabolites both within and outside of the pathway, while a self-contained test ignores metabolites that are not in the pathway. In other words, for a competitive test, the null hypothesis being tested is that the association between the specific pathway and disease is average; while for a self-contained test, the null hypothesis is that there is no association between the specific pathway and disease (Holmans 2010). For the multivariate statistics, the null hypothesis is self-contained since the null hypothesis is that there is no association between metabolites in the pathway and the phenotype (Holmans 2010). Although multivariate statistics take the correlation of different metabolites into account, they may not be necessarily more powerful than univariate statistics (Khatri et al. 2012). There are also some drawbacks of the FCS method. If two pathways have the same metabolites, FCS will give the same result. That is, FCS does not take the reactions among metabolites (topology structure) into account. One way to address this problem is to use a correlation measure, such as the Pearson correlation coefficient, to help us choose the most suitable pathway; another method is to use pathway reconstruction.

5.3.3 Metabolic pathway reconstruction and simulation

Metabolic pathway/network reconstruction and simulation are a batch of methods used to refine or construct metabolic networks. A reconstruction collects all of the relevant metabolic information of an organism and compiles it in a mathematical model. The relevant metabolic information includes all related known chemical reactions, previously constructed networks, experimental data, and related research results. After compiling the data into a model, we can obtain the output of the system and then use it to refine our model and perform the simulation iteratively. If we have the knowledge of all involved metabolites, then we can enhance the predictive capacity of the reconstructed models by connecting the metabolites within the pathways. The pathway models can be roughly classified into one of two categories: static (stoichiometric network models) and kinetic models. We will first discuss static models and then give a brief introduction on kinetic modeling.

The mathematical model behind static models is a linear system. If we treat the metabolic network as a system, then, based on mass conservation of internal metabolites within a system, we can express the reaction network by a stoichiometric matrix ($\boldsymbol{S}$). Each element of $\boldsymbol{S}(s_{ij})$ represents the coefficient of metabolite i involved in reaction j. At a steady state, we have $\boldsymbol{S}\boldsymbol{v} = \boldsymbol{0}$, since there is no accumulation of internal metabolites in the system (Schilling et al. 1999). This linear system is called the flux-balance equation. Here $\boldsymbol{v}$ represents fluxes through the associated reactions in $\boldsymbol{S}$. These linear equations define the entire reaction network; all of the solutions to this linear system are valid steady-state flux distributions. In general, $\boldsymbol{S}$ is an $m \times n$ matrix where the number of columns is larger than the number of rows ($n > m$), which means there are more reactions than metabolites (Schilling et al. 1999). Moreover, $\boldsymbol{S}$ is a full rank matrix, which means $rank(\boldsymbol{S}) = m$. Therefore, there are multiple solutions to this linear system. Different solutions define different pathways. Based on different research purposes, we may impose different constraints on the linear system and obtain different types of solutions.

Here we introduce three kinds of solutions that are the most widely used in metabolic pathway analysis. The first type of solution is called elementary modes. In addition to the flux-balance equation, we add the constraints that all fluxes are greater than 0. Then, by applying the convex analysis method, we can find a solution set called elementary modes (EM) if the following properties are satisfied:

(i) Uniqueness: The solution set is unique for a given network.
(ii) Non-decomposability: Each solution in the solution set consists of the minimum number of reactions that it needs to exist as a functional unit. If any reaction in a solution set were removed, the whole solution set could not operate as a functional unit.
(iii) The solution set is the set of all routes through a metabolic network consistent with the second property (Papin et al. 2004).

The second type of solution is called extreme pathways (EP). By convex analysis, the solution set is called the extreme pathways if it is under the same constraints of EM and follows properties below:

(i) Uniqueness: The solution set is unique for a given network.
(ii) Non-decomposability: Each solution in the solution set consists of the minimum number of reactions that it needs to exist as a functional unit.
(iii) This solution set is the systemically independent subset of the elementary modes; that is, no solution in the set can be represented as a nonnegative linear combination of any other solutions in the solution set, namely, they are convex basis vectors (Papin et al. 2004).

By using these two kinds of solutions, we can analyze or construct metabolic pathways and networks. Note that we may have a finite number of solutions for EM and EP, which means that we may obtain several different pathways. A numerical example of the calculation and use of the EM and EP can be found in Förster et al. (2002). The key difference between EM and EP is that they treat internal reversible and irreversible reactions differently. EP analysis decouples all internal reversible reactions into forward and reverse directions while EM analysis accounts for reaction directionality through a series of rules in the corresponding calculations of the modes. Moreover, EPs are subsets of EMs, i.e., the numbers of extreme pathways are smaller (potentially much smaller) than or equal to the number of elementary modes.

Another important solution corresponds to an independent analysis method called flux balance analysis (FBA). FBA differs from EM or EP by imposing more constraints and an objective function. Depending on the purpose of the research being performed, we have different problems of interest pertaining to the pathway. For example, we may want to maximize or minimize the flux of certain reactions; or want to limit some flux to a certain interval and see how the pathway changes. Therefore we need to impose an objective function on the linear system. The problem is thus transformed into an optimization problem. We can use linear programming to solve this problem; it does not matter how many linear constraints we have to impose on the flux vector. Note that FBA generally gives only one solution. This is in contrast to EM and EP, which give several solutions. Figure 3 shows how to perform an FBA (Raman and Chandra 2009). First, we specify the reaction network that contains all metabolites and detailed information on all possible reactions (Fig. 3, 1st step). Internal fluxes are denoted by $v_i(i = 1, 2)$ and exchange fluxes are denoted by $b_j(j = 1, 2, 3, 4, 5)$. Then, after building the linear system based on the network structure (Fig. 3, 2nd and 3rd steps), we can add a biologically relevant objective function and relevant constraints (Fig. 3, 4th and 5th steps). The remainder of the task is linear programming (Fig. 3, last step), which can be accomplished using software packages, such as MATLAB COBRA toolbox (Becker et al. 2007).

The pathway constructed using the methods above cannot show the dynamic state such as regulatory effects, how the enzymes work, or whether the pathway is in stable steady state. Therefore, we may use a kinetic model to simulate the metabolomics network (Tomar and De 2013). A kinetic reaction network model can be described by ordinary or partial differential equations (ODE or PDE, respectively). For example, we can simulate the network based on the following simple ODE (Steuer 2007).

$$\frac{d}{dt}\boldsymbol{c}(t) = \boldsymbol{S} \times \boldsymbol{v}(\boldsymbol{c}, \boldsymbol{k});$$

$\boldsymbol{c}(t)$ is the time-dependent concentration vector of m internal metabolites, $\boldsymbol{k}$ represents the Michaelis–Menten kinetics parameters, and S is the stoichiometric matrix. $\boldsymbol{v}(\boldsymbol{c}, \boldsymbol{k})$ is a vector of enzyme-kinetic rate equations that consists of nonlinear functions of $\boldsymbol{c}$ and $\boldsymbol{k}$. We can see that if we let the left-hand-side term equal zero (indicating that it is at a steady state) and let $\boldsymbol{v}$ be a flux vector, then the equation is exactly the flux balance equation. Given an initial condition of $\boldsymbol{c}(0)$, the value of the kinetic parameters and the rate equations $\boldsymbol{v}$, we can simulate the data through the ODE given above. A common choice for the rate equations (for every v in $\boldsymbol{v}$) is $v = \frac{v_{\max} c}{c+k}$, where $v_{\max}$ is the maximal reaction velocity. However, sometimes we do not know the explicit form of the rate equations or it is difficult to estimate the kinetic parameters k. In these cases, we can use the structural kinetic modeling method (SKM) (Steuer 2007). The SKM method uses the Jacobian matrix as a local linear approximation of the rate equations. The Jacobian matrix consists of all first order partial derivatives of the rate equations and it can be rewritten and estimated using the SKM method (Wiechert 2002; Steuer 2007).

5.4 Analysis methods for time course data

Variables, for example the concentration of metabolites, may change with time, thereby creating a time dimension in the dataset. Unsupervised learning and data visualization tools are still initially useful for giving us a general idea of the data structure. We can also use visualization tools such as PCA, SOM, and heat maps with a hierarchical clustering structure to detect patterns and groups/clusters of the data. The only difference is that we should include a time dimension. In addition, by drawing profile graphs we can check the profiles of metabolites or subjects for different clusters. However, if we want to compare the temporal profiles (similar or different patterns of change) of metabolites between different subjects or groups of subjects, we need to introduce statistical methods different from those described above. Among different statistical methods that can be used to analyze time course data (Smilde et al. 2010), we only introduce analysis of variance (ANOVA) based methods.

If we are analyzing one variable over time, e.g., the expression level of a protein or concentration of a metabolite, and we want to test whether the temporal profiles of this variable are significantly different under different experimental conditions, a natural choice is to use two-way ANOVA, which is often used when studying the effects of different treatments in chemical or biological experiments. Here we show the basics of this ANOVA model. In metabolomics research, the experimental condition (α) and time effect (τ) can be treated as two fixed effects. The general linear model for a two-way ANOVA is:

$$Y_{ijk} = \mu + \alpha_i + \tau_j + (\alpha\tau)_{ij} + \varepsilon_{ijk};$$

where Y_{ijk} refers to the measurement obtained from the kth subject at the jth time point under the ith condition $(i = 1, 2, \ldots, a;\ j = 1, 2, \ldots, b;\ k = 1, 2, \ldots n)$; α and τ are fixed effects corresponding to condition and time; a and b are the total number of levels for each effect; n is the number of replicates (often corresponding to subjects) for each combination of condition and time effects; μ is the overall (grand) mean. In many applications involving the traditional two-way ANOVA, ε_{ijk} are assumed to be independent random errors following a normal distribution, denoted by $N(0, \sigma^2)$ (Kutner 2005).

Although it seems reasonable to use a two-way ANOVA to analyze time course data, the data may be better described by a repeated measures (RM) model. The main difference between the two-way ANOVA and the RM

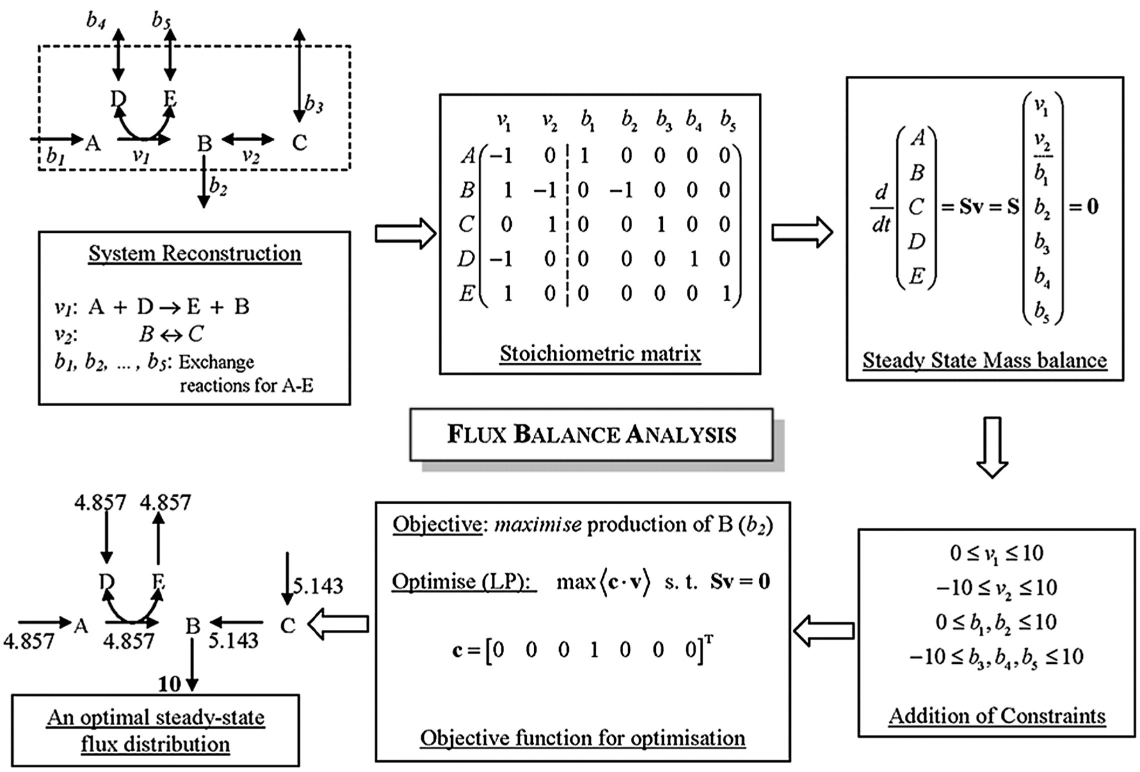

Fig. 3 Flow chart of FBA (Raman and Chandra 2009)

model is that the RM model has a subject error term (δ_{ik}), which takes variation within each group or subject into consideration. The following is the model of RM:

$$Y_{ijk} = \mu + \alpha_i + \delta_{ik} + \tau_j + (\alpha\tau)_{ij} + \varepsilon_{ijk};$$

where $i = 1, 2, \ldots, \mathrm{a}$; $\mathrm{j} = 1, 2, \ldots, \mathrm{b}$; $\mathrm{k} = 1, 2, \ldots, \mathrm{n}$. δ_{ik} are subject effects (within group error) which follow a normal distribution $N(0, \sigma_\delta^2)$ and are independent of random error ε_{ijk}. Other notations are the same as the two-way ANOVA model mentioned above. In metabolomics research, there are often some differences between subjects even within the same groups. If the subjects vary a great deal within each group, then the RM model will be more powerful than the simple two-way ANOVA analysis (Milliken and Johnson 2009). Since each subject is repeatedly assessed in most time course studies, we strongly recommend to use the RM model instead of two-way ANOVA. With the RM model, the user can specify a variety of correlation structures for the measurement error. Compound-symmetric correlation is often assumed in repeated measures analyses. This covariance structure for the ε_{ijk} allows for within-subject correlation that is common over time. If the correlation is expected to decay over time, it is advantageous to consider autoregressive covariance structures (Brockwell and Davis 2002) or exponential covariance functions (Szczesniak et al. 2013).

Inference procedure for RM model is quite similar to ANOVA. We can calculate the statistics and p-values by decomposing the total sum of squares (SST) into different parts as shown in the following formula:

$$\mathrm{SST} = \mathrm{SS}(\alpha) + \mathrm{SS}(\tau) + \mathrm{SS}(\alpha\tau) + \mathrm{SS}(\delta) + \mathrm{SSE}$$

After validating all assumptions (normality, independence and homogeneity of the variances), we turn to look at the ANOVA table (Kutner 2005; Milliken and Johnson 2009). The F statistics each follow an F distribution under the null hypothesis (corresponding effects are all zero) with corresponding numerator degrees of freedom. Therefore, we have three effects to test: two main effects (α and τ) and an interaction effect. Note that we must always test the interaction between α and τ first, since a significant interaction may mask the significance of main effects and influence the explanation of data. The null hypothesis for the interaction effect is: $(\alpha\tau)_{ij} = 0$ for all i and j. If the interaction effect is significant, it implies that the temporal profiles for different groups (experimental conditions) are different. Usually, the estimated treatment means plots (and many other plots) will give us a straightforward

Table 2 Summary of commonly used metabolomics data analysis methods

Type of analysis	Basic methods		Goal	Application	Input	Output	Software[a]
Basic statistical testing	Fold change		Biomarker discovery; feature selection for supervised learning	Fold change of metabolite concentrations between two groups	Data tables with class memberships: each row represents one subject and each column represents concentration of a metabolite/MS and NMR peak list or spectral bin	Lists of selected metabolites with p-values, volcano plot (for two groups)	MetaboAnalyst[b], R and MATLAB
	Statistical testing	Two sample tests (e.g. t test)		Identify significantly different expressed metabolites between two groups			
		ANOVA with post hoc analysis		Identify significantly different expressed metabolites for multiple groups			
Unsupervised learning	Principal component analysis (PCA)		Data grouping and visualization	Reduce dimensionality and check clusters visually	Same as above, but no need for class memberships	Scores and loadings plots for visualization	MetaboAnalyst, R and MATLAB
	Clustering	K-means, fuzzy c-means, K-means in reduced space		Group the subjects into k different clusters		Cluster labels for each subject	
		Hierarchical		Show how subjects form different clusters		Heatmap, dendrogram	
	Self-organizing map (SOM)			Reduce dimensionality and check clusters visually		SOM	
Supervised learning—classification	Support vector machine (SVM) Partial least square-discriminant analysis (PLS-DA), OPLS-DA and SPLS-DA		Predict class memberships (disease diagnostic), biomarker discovery	All of them are classification methods suitable for metabolomics research; need to compare their performance in real data analysis	Same as above, but class memberships are needed	A prediction model with selected metabolites as predictors	MetaboAnalyst, R,and MATLAB
Supervised learning—regression	Support vector regression (SVR) Partial least square (PLS), OPLS and SPLS		Predict continuous variables, calibration, biomarker discovery	All of them are regression methods suitable for metabolomics research; need to compare their performance in real data analysis	The response variables should be continuous vector or matrix; for calibration problem, the response variables are concentrations and covariates are spectral intensities	A prediction model with selected variables as predictors	
Pathway analysis	Over-representation analysis (ORA) Functional class scoring (FCS)/enrichment analysis		Find biologically meaningful metabolite sets or pathways that are associated with certain diseases or characters	Find related pathways or metabolite sets using statistical testing	Significant metabolites list and reference pathways Metabolites list with concentrations and reference pathways or metabolite sets	A list of selected pathways with their p-values or FDRs	MetaboAnalyst, bioconductor[c]

Table 2 continued

Type of analysis	Basic methods		Goal	Application	Input	Output	Software[a]
	Elementary modes/ Extreme pathways		Simulation and pathway reconstruction	Analyze or reconstruct metabolic pathways	Pathway, stoichiometric matrix	Several different possible flux distributions	MATLAB
	Flux Balance Analysis			Find the flux distribution by maximizing/ minimizing a reaction based on some constraints	Pathway, stoichiometric matrix, constraints, and an objective function	One unique flux distribution	
	Kinetic reaction network model			Simulate the dynamic behavior of metabolites reaction networks	Pathway, stoichiometric matrix, Kinetic parameter and rate equation	Simulation results of network kinetics	
Time course data analysis	Analysis of variance (ANOVA)	Fixed effects ANOVA	Test time dependent effects; compare time profiles for metabolites from different subjects	Two or more given factors	Data tables with labels (class membership indicators): each row represents the observation value of one subject at one time point and columns represent metabolites concentrations/MS and NMR peak lists or spectral bins	Testing results of time dependent effects for each metabolite, time profile plots	MetaboAnalyst, R
		Repeated measures		ANOVA method when we have multiple within subjects measurements (measured at different time points)			
	ANOVA-simultaneous component analysis (ASCA)			ANOVA for multivariate response		Scores and loadings plots of sub models, time profiles plots	
	Functional based methods (e.g. smoothing splines mixed effects model)			Estimate sets of curves (metabolic profiles) and quantify their differences		Testing results of the differences in time profiles between groups for each metabolite	
	Time series models (e.g. ARMA model)		Model the dynamic properties (e.g. biorhythm) of metabolites data	Long time series data modeling		A model that best describes the dynamic properties of underlying biological process	

[a] We only picked some commonly used software packages or web tools in metabolomics

[b] MetaboAnalyst is an easy-to-use web based tool that covers most basic computational and statistical methods for metabolomics (Xia et al. 2009, 2012)

[c] Bioconductor was built on R and provided many useful packages for bioinformatics (Gentleman et al. 2004), including packages for metabolic pathway analysis (Luo and Brouwer 2013; Zhang and Wiemann 2009)

explanation of main and interaction effects (Milliken and Johnson 2009).

If the response variable **Y** is a matrix, then we will need to analyze multiple metabolites simultaneously, while taking their correlation structure into consideration. In this case, we need another method called ANOVA-simultaneous component analysis (ASCA). ASCA is a generalization of ANOVA from the univariate case to the multivariate case. Statistically, traditional MANOVA is a generalization of ANOVA. However, MANOVA will not work if the covariance matrices are singular or the assumption of multivariate normality is violated. The idea behind ASCA comes from principal component analysis, which decomposes the original data matrix into a component score and a loading matrix plus an error term. The following is the full model (Smilde et al. 2005):

$$\boldsymbol{X}_{hi_h} = \boldsymbol{lm}^T + \boldsymbol{T}_K\boldsymbol{P}_l^T + \boldsymbol{T}_{Kh}\boldsymbol{P}_2^T + \boldsymbol{T}_{Khi_h}\boldsymbol{P}_3^T + \boldsymbol{E}_{hi_h};$$

With the following constraints:

(1) $\boldsymbol{l}^T\boldsymbol{T}_K = 0^T$

(2) $\sum_{h=1}^{H} \boldsymbol{T}_{Kh} = 0$

(3) $\sum_{i_h=1}^{I_h} \boldsymbol{T}_{Khi_h} = 0$

$h = 1, \ldots, H$; $i_h = 1, \ldots, I_h$. H denotes the number of groups and I_h denotes the number of replicates in group h. $\boldsymbol{X}_{hi_h}$ is a $K * J$ matrix where K denotes the number of available time points and J denotes the number of variables. $\boldsymbol{l}$ is a K dimensional vector of ones and $\boldsymbol{m}$ is a J dimensional vector of overall means of variables, so each row in the matrix $\boldsymbol{lm}^T$ represents the overall mean of the variables. $\boldsymbol{T}_K$ is the matrix of "time" effect; $\boldsymbol{T}_{\mathrm{Kh}}$ represents the "time" and treatment interaction effect; $\boldsymbol{T}_{Khi_h}$ is the interaction of treatment, time, and subject; $\boldsymbol{E}_{hi_h}$ is the matrix of residuals. The corresponding $\boldsymbol{P}$ matrices are loading matrices. Unlike in ANOVA, we will use the scores plots of the first few components for each effect ($\boldsymbol{T}$ matrices, a.k.a., sub-model) to detect the time main effect or interaction effect; and we can use the corresponding loadings plots to detect which variables are responsible for the variation. There are some examples of ASCA scores plots from (Nueda et al. 2007).

In their paper, they mainly discussed the application of ASCA on time course microarray data. As an example, in the scores plots of their simulation study (see Supplementary Fig. 3), sub-model a represents the time main effect and sub-model b + ab represents the treatment effect (treatment main effect and its interaction with time effect). The percentages on the left show how much the components in the sub-model (principal components kept in this $\boldsymbol{T}$ matrix) explain the variation in the corresponding effect (sub-model). The first plot shows the positive time main effect exists; the following two show differences between subjects on the same treatment. Note that in this example, the time-treatment effect was not modeled independently, which is different from that in the original paper of ASCA (Smilde et al. 2005).

There are many other methods for analyzing time course data that we did not discuss in this paper, such as the time-series data analysis (ARMA model) (Smilde et al. 2010). The ARMA model makes sense only when we have many more than just two or three time points. If our dataset has many time points and we wish to find and compare the profile curves, then a functional based method (Berk et al. 2011) may be a good choice. The paper proposed a smoothing splines mixed effects model that treats each longitudinal measurement as a smooth function of time and uses a functional t-type test statistic to quantify the difference between two sets of curves. See (VanDyke et al. 2012) for a biomedical application related to this approach. Furthermore, since metabolomics is well suited to longitudinal studies, if we have many time points and experimental conditions, we can use the Hierarchical Linear Model to fit the data. It treats the time profile as a function and the parameters of the time profile function as random variables (Jansen et al. 2004).

6 Conclusions

Metabolomics is a rapidly growing field that has greatly improved our understanding of the metabolic mechanisms behind biological processes as well as human diseases. Its broad goal is to understand how the overall metabolism of an organism has been changed under different conditions. Metabolomics has been used to study diseases such as cystic fibrosis, central nervous system diseases, cancer, diabetes, and cardiac disease. Using metabolomics could lead to the discovery of more accurate biomarkers that will help diagnose, prevent, and monitor the risk of disease. This review briefly introduced the background of metabolomics, NMR and MS strategies, and data pre-processing. We then placed our main focus on the data analysis of metabolomics and described mainstream data analysis methods in current metabolomics research. These include unsupervised learning methods, supervised learning methods, pathway analysis methods and time course data analysis. Finally, in Table 2, we summarized the key points of the methods discussed, as well as some basic methods such as fold change and two sample t test that were not included in this review. We hope our review will be a useful reference for researchers without this type of background in data analysis.

Acknowledgments We would like to express our great appreciation to Dr. Lilliam Ambroggio and Dr. Lindsey Romick-Rosendale for their valuable and constructive suggestions to our review. Their willingness to give their time so generously has been very much appreciated. This study is funded by the NIH Grant R01 HL116226 to RDS and LJL.

Compliance with Ethical Standards

Conflict of interest Sheng Ren, Anna A. Hinzman, Emily L. Kang, Rhonda D. Szczesniak and L. Jason Lu declare that we have no conflict of interest and we have included separately signed conflict of interest forms in this manuscript.

References

Abdi, H. (2010). Partial least squares regression and projection on latent structure regression (PLS regression). *Wiley Interdisciplinary Reviews: Computational Statistics, 2*, 97–106.

Agresti, A. (2014). *Categorical data analysis*. New York: Wiley.

Anderson, P. E., Reo, N. V., DelRaso, N. J., Doom, T. E., & Raymer, M. L. (2008). Gaussian binning: A new kernel-based method for processing NMR spectroscopic data for metabolomics. *Metabolomics, 4*, 261–272.

Armitage, E. G., & Barbas, C. (2014). Metabolomics in cancer biomarker discovery: Current trends and future perspectives. *Journal of Pharmaceutical and Biomedical Analysis, 87*, 1–11.

Assfalg, M., et al. (2008). Evidence of different metabolic phenotypes in humans. *Proceedings of the National Academy of Sciences, 105*, 1420–1424.

Becker, S. A., Feist, A. M., Mo, M. L., Hannum, G., Palsson, B. Ø., & Herrgard, M. J. (2007). Quantitative prediction of cellular metabolism with constraint-based models: The COBRA Toolbox. *Nature Protocols, 2*, 727–738.

Beckonert, O., Monnerjahn, J., Bonk, U., & Leibfritz, D. (2003). Visualizing metabolic changes in breast-cancer tissue using 1H-NMR spectroscopy and self-organizing maps. *NMR in Biomedicine, 16*, 1–11.

Berk, M., Ebbels, T., & Montana, G. (2011). A statistical framework for biomarker discovery in metabolomic time course data. *Bioinformatics, 27*, 1979–1985. doi:10.1093/bioinformatics/btr289.

Bezdek, J. C., Coray, C., Gunderson, R., & Watson, J. (1981). Detection and characterization of cluster substructure i. linear structure: Fuzzy c-lines. *SIAM Journal on Applied Mathematics, 40*, 339–357.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.

Blekherman, G., et al. (2011). Bioinformatics tools for cancer metabolomics. *Metabolomics, 7*, 329–343. doi:10.1007/s11306-010-0270-3.

Boulesteix, A.-L. (2004). PLS dimension reduction for classification with microarray data. *Statistical Applications in Genetics and Molecular Biology, 3*, 1–30.

Box, G. E., Hunter, W. G., & Hunter, J. S. (1978). *Statistics for experimenters*. New York: Wiley.

Brereton, R. G., & Lloyd, G. R. (2010). Support vector machines for classification and regression. *Analyst, 135*, 230–267.

Broadhurst, D. I., & Kell, D. B. (2006). Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics, 2*, 171–196.

Brockwell, P. J., & Davis, R. A. (2002). *Introduction to time series and forecasting* (Vol. 1). Boca Raton: Taylor & Francis.

Bu, H.-L., Li, G.-Z., Zeng, X.-Q., Yang, J. Y., & Yang, M. Q. (2007). Feature selection and partial least squares based dimension reduction for tumor classification. In *Proceedings of the 7th IEEE international conference on bioinformatics and bioengineering, 2007 (BIBE 2007)* (pp. 967–973). New York: IEEE.

Bylesjö, M., Rantalainen, M., Cloarec, O., Nicholson, J. K., Holmes, E., & Trygg, J. (2006). OPLS discriminant analysis: Combining the strengths of PLS-DA and SIMCA classification. *Journal of Chemometrics, 20*, 341–351.

Cao, H., Dong, J., Cai, C., & Chen, Z. (2008). Investigations on the effects of NMR experimental conditions in human urine and serum metabolic profiles. In *The 2nd international conference on bioinformatics and biomedical engineering, 2008 (ICBBE 2008)* (pp. 2236–2239). New York: IEEE.

Chun, H., & Keleş, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72*, 3–25.

Chung, D., & Keles, S. (2010). Sparse partial least squares classification for high dimensional data. *Statistical Applications in Genetics and Molecular Biology*. doi:10.2202/1544-6115.1492.

Coombes, K. R., Tsavachidis, S., Morris, J. S., Baggerly, K. A., Hung, M. C., & Kuerer, H. M. (2005). Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. *Proteomics, 5*, 4107–4117.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*, 273–297.

Craig, A., Cloarec, O., Holmes, E., Nicholson, J. K., & Lindon, J. C. (2006). Scaling and normalization effects in NMR spectroscopic metabonomic data sets. *Analytical Chemistry, 78*, 2262–2267.

Cui, Q., et al. (2008). Metabolite identification via the Madison metabolomics consortium database. *Nature Biotechnology, 26*, 162–164.

Davis, R. A., Charlton, A. J., Godward, J., Jones, S. A., Harrison, M., & Wilson, J. C. (2007). Adaptive binning: An improved binning method for metabolomics data using the undecimated wavelet transform. *Chemometrics and Intelligent Laboratory Systems, 85*, 144–154.

De Soete, G., & Carroll, J. D. (1994). K-means clustering in a low-dimensional Euclidean space. In E. Diday, et al. (Eds.), *New approaches in classification and data analysis* (pp. 212–219). Heidelberg: Springer.

Dettmer, K., Aronov, P. A., & Hammock, B. D. (2007). Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews, 26*, 51–78. doi:10.1002/mas.20108.

Dieterle, F., Ross, A., Schlotterbeck, G., & Senn, H. (2006). Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. *Analytical Chemistry, 78*, 4281–4290.

Draisma, H. H., Reijmers, T. H., Meulman, J. J., van der Greef, J., Hankemeier, T., & Boomsma, D. I. (2013). Hierarchical clustering analysis of blood plasma lipidomics profiles from mono-and dizygotic twin families. *European Journal of Human Genetics, 21*, 95–101.

Dunn, J. C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics, 3*, 32–57.

Dunn, W. B., Bailey, N. J., & Johnson, H. E. (2005). Measuring the metabolome: Current analytical technologies. *Analyst, 130*, 606–625.

Dunn, W. B., Wilson, I. D., Nicholls, A. W., & Broadhurst, D. (2012). The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomic studies of humans. *Bioanalysis, 4*, 2249–2264.

Dunn, W. B., Broadhurst, D., Begley, P., Zelena, E., Francis-McIntyre, S., Anderson, N., et al. (2011). Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nature Protocols, 6*, 1060–1083.

Eilers, P. H., & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science, 11*, 89–102.

Emwas, A.-H., Luchinat, C., Turano, P., Tenori, L., Roy, R., Salek, R. M., et al. (2014). Standardizing the experimental conditions for using urine in NMR-based metabolomic studies with a particular focus on diagnostic studies: A review. *Metabolomics, 11*(4), 872–894.

Enea, C., et al. (2010). ^{1}H NMR-based metabolomics approach for exploring urinary metabolome modifications after acute and chronic physical exercise. *Analytical and Bioanalytical Chemistry, 396*, 1167–1176.

Ertöz, L., Steinbach, M., & Kumar, V. (2003). Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. In: *SDM 2003*, SIAM (pp. 47–58).

Fahy, E., Sud, M., Cotter, D., & Subramaniam, S. (2007). LIPID MAPS online tools for lipid research. *Nucleic Acids Research, 35*, W606–W612.

Förster, J., Gombert, A. K., & Nielsen, J. (2002). A functional genomics approach using metabolomics and in silico pathway analysis. *Biotechnology and Bioengineering, 79*, 703–712.

Gentleman, R. C., et al. (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology, 5*, R80.

Gika, H. G., Theodoridis, G. A., Plumb, R. S., & Wilson, I. D. (2014). Current practice of liquid chromatography–mass spectrometry in metabolomics and metabonomics. *Journal of Pharmaceutical and Biomedical Analysis, 87*, 12–25.

Griffin, J. L., Atherton, H., Shockcor, J., & Atzori, L. (2011). Metabolomics as a tool for cardiac research. *Nature, 8*, 630–643.

Griffin, J. L., & Shockcor, J. P. (2004). Metabolic profiles of cancer cells. *Nature Reviews Cancer, 4*, 551–561. doi:10.1038/nrc1390.

Griffiths, W. J., Koal, T., Wang, Y., Kohl, M., Enot, D. P., & Deigner, H. P. (2010). Targeted metabolomics for biomarker discovery. *Angewandte Chemie, 49*, 5426–5445. doi:10.1002/anie.200905579.

Guan, W., Zhou, M., Hampton, C. Y., Benigno, B. B., Walker, L. D., Gray, A., et al. (2009). Ovarian cancer detection from metabolomic liquid chromatography/mass spectrometry data by support vector machines. *BMC Bioinformatics, 10*, 259. doi:10.1186/1471-2105-10-259.

Gunderson, R. W. (1982). Choosing the r-dimension for the FCV family of clustering algorithms. *BIT Numerical Mathematics, 22*, 140–149.

Gunderson, R. W. (1983). An adaptive FCV clustering algorithm. *International Journal of Man-Machine Studies, 19*, 97–104.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research, 3*, 1157–1182.

Haddad, I., Hiller, K., Frimmersdorf, E., Benkert, B., Schomburg, D., & Jahn, D. (2009). An emergent self-organizing map based analysis pipeline for comparative metabolome studies. *In Silico Biology, 9*, 163–178.

Hamerly, G., & Elkan, C. (2003). Learning the k in k-means. *Advances in Neural Information Processing Systems, 16*, 281–288.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 28*, 100–108.

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., & Tibshirani, R. (2009). *The elements of statistical learning* (Vol. 2). Berlin: Springer.

Heather, L. C., Wang, X., West, J. A., & Griffin, J. L. (2013). A practical guide to metabolomic profiling as a discovery tool for human heart disease. *Journal of Molecular and Cellular Cardiology, 55*, 2–11.

Heinzmann, S. S., Brown, I. J., Chan, Q., Bictash, M., Dumas, M. E., Kochhar, S., et al. (2010). Metabolic profiling strategy for discovery of nutritional biomarkers: Proline betaine as a marker of citrus consumption. *The American Journal of Clinical Nutrition, 92*, 436–443.

Henneges, C., Bullinger, D., Fux, R., Friese, N., Seeger, H., Neubauer, H., et al. (2009). Prediction of breast cancer by profiling of urinary RNA metabolites using Support Vector Machine-based feature selection. *BMC Cancer, 9*, 104.

Holmans, P. (2010). Statistical methods for pathway analysis of genome-wide data for association with complex genetic traits. *Advances in Genetics, 72*, 141–179. doi:10.1016/B978-0-12-380862-2.00007-2.

Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., et al. (2010). MassBank: A public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry, 45*, 703–714.

Hou, Y., et al. (2012). Microbial strain prioritization using metabolomics tools for the discovery of natural products. *Analytical Chemistry, 84*, 4277–4283. doi:10.1021/ac202623g.

Huang, J. Z., Ng, M. K., Rong, H., & Li, Z. (2005). Automated variable weighting in k-means type clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*, 657–668.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters, 31*, 651–666.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys (CSUR), 31*, 264–323.

Jansen, J. J., Hoefsloot, H. C., Boelens, H. F., van der Greef, J., & Smilde, A. K. (2004). Analysis of longitudinal metabolomics data. *Bioinformatics, 20*, 2438–2446. doi:10.1093/bioinformatics/bth268.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika, 32*, 241–254.

Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis* (6th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Jolliffe, I. (2005). *Principal component analysis*. New YorK: Wiley Online Library.

Kaddurah-Daouk, R., & Krishnan, K. R. (2009). Metabolomics: A global biochemical approach to the study of central nervous system diseases. *Neuropsychopharmacology, 34*, 173–186. doi:10.1038/npp.2008.174.

Kanehisa, M. (2002). The KEGG database. *Novartis Foundation Symposium, 247*, 91–101 **; discussion 101–3, 119–28, 244–52**.

Kang, S. M., Park, J. C., Shin, M. J., Lee, H., Oh, J., Hwang, G. S., et al. (2011). (1)H nuclear magnetic resonance based metabolic urinary profiling of patients with ischemic heart failure. *Clinical Biochemistry, 44*, 293–299. doi:10.1016/j.clinbiochem.2010.11.010.

Kell, D. B., Brown, M., Davey, H. M., Dunn, W. B., Spasic, I., & Oliver, S. G. (2005). Metabolic footprinting and systems biology: The medium is the message. *Nature Reviews Microbiology, 3*, 557–565. doi:10.1038/nrmicro1177.

Khatri, P., Sirota, M., & Butte, A. J. (2012). Ten years of pathway analysis: Current approaches and outstanding challenges. *PLoS Computational Biology, 8*, e1002375. doi:10.1371/journal.pcbi.1002375.

Kilkenny, C., Parsons, N., Kadyszewski, E., Festing, M. F., Cuthill, I. C., Fry, D., et al. (2009). Survey of the quality of experimental design, statistical analysis and reporting of research using animals. *PLoS ONE, 4*, e7824.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE, 78*, 1464–1480.

Kohonen, T. (1998). The self-organizing map. *Neurocomputing, 21*, 1–6.

Kutner, M. H. (2005). *Applied linear statistical models* (5th ed.). McGraw-Hill/Irwin: Boston.

Lauridsen, M., Hansen, S. H., Jaroszewski, J. W., & Cornett, C. (2007). Human urine as test material in 1H NMR-based metabonomics: Recommendations for sample preparation and storage. *Analytical Chemistry, 79*, 1181–1186.

Li, H., Liang, Y., & Xu, Q. (2009a). Support vector machines and its applications in chemistry. *Chemometrics and Intelligent Laboratory Systems, 95*, 188–198.

Li, X., Lu, X., Tian, J., Gao, P., Kong, H., & Xu, G. (2009b). Application of fuzzy c-means clustering in data analysis of metabolomics. *Analytical Chemistry, 81*, 4468–4475.

Li, F., Wang, J., Nie, L., & Zhang, W. (2012). *Computational methods to interpret and integrate metabolomic data*. New York: INTECH Open Access Publisher.

Luo, W., & Brouwer, C. (2013). Pathview: An R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics, 29*, 1830–1831.

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

Martens, H. (1992). *Multivariate calibration*. New York: Wiley.

Marzetti, E., Landi, F., Marini, F., Cesari, M., Buford, T. W., Manini, T. M., et al. (2014). Patterns of circulating inflammatory biomarkers in older persons with varying levels of physical performance: A partial least squares-discriminant analysis approach. *Frontiers in Medicine*, *1*, 27. doi:10.3389/fmed.2014.00027.

Matthiesen, R., & SpringerLink (Online Service). (2010). Bioinformatics methods in clinical research. In S. Krawetz & S. Misener (Eds.), *Methods in molecular biology, methods and protocols*. Totowa: Humana Press.

Milliken, G. A., & Johnson, D. E. (2009). *Analysis of messy data* (2nd ed.). Boca Raton: CRC Press.

Milone, D. H., Stegmayer, G., López, M., Kamenetzky, L., & Carrari, F. (2014). Improving clustering with metabolic pathway data. *BMC Bioinformatics, 15*, 101.

Montgomery, D. C. (2008). *Design and analysis of experiments*. New York: Wiley.

Nguyen, D. V., & Rocke, D. M. (2002). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics, 18*, 39–50.

Nicholson, J. K., Lindon, J. C., & Holmes, E. (1999). 'Metabonomics': Understanding the metabolic responses of living systems to pathphysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica, 29*, 1181–1189.

Nin, N., Izquierdo-García, J., & Lorente, J. (2012). The metabolomic approach to the diagnosis of critical illness. In *Annual update in intensive care and emergency medicine* (pp. 43–52). Berlin: Springer.

Nueda, M. J., Conesa, A., Westerhuis, J. A., Hoefsloot, H. C., Smilde, A. K., Talón, M., et al. (2007). Discovering gene expression patterns in time course microarray experiments by ANOVA-SCA. *Bioinformatics, 23*, 1792–1800. doi:10.1093/bioinformatics/btm251.

Oliver, S. G. (2002). Functional genomics: Lessons from yeast. *Philosophical Transactions of the Royal Society of London. Series B, Biological sciences, 357*, 17–23. doi:10.1098/rstb.2001.1049.

Oliver, S. G., Winson, M. K., Kell, D. B., & Baganz, F. (1998). Systematic functional analysis of the yeast genome. *Trends in Biotechnology, 16*, 373–378

O'Sullivan, A., Gibney, M. J., & Brennan, L. (2011). Dietary intake patterns are reflected in metabolomic profiles: Potential role in dietary assessment studies. *The American Journal of Clinical Nutrition, 93*, 314–321.

Papin, J. A., Stelling, J., Price, N. D., Klamt, S., Schuster, S., & Palsson, B. O. (2004). Comparison of network-based pathway analysis methods. *Trends in Biotechnology, 22*, 400–405. doi:10.1016/j.tibtech.2004.06.010.

Patel, K. N., Patel, J. K., Patel, M. P., Rajput, G. C., & Patel, H. A. (2010). Introduction to hyphenated techniques and their applications in pharmacy. *Pharmaceutical Methods, 1*, 2–13.

Pauling, L., Robinson, A. B., Teranishi, R., & Cary, P. (1971). Quantitative analysis of urine vapor and breath by gas-liquid partition chromatography. *Proceedings of the National Academy of Sciences of the United States of America, 68*, 2374–2376.

Poroyko, V., Morowitz, M., Bell, T., Ulanov, A., Wang, M., Donovan, S., et al. (2011). Diet creates metabolic niches in the "immature gut" that shape microbial communities. *Nutricion Hospitalaria, 26*, 1283–1295. doi:10.1590/S0212-16112011000600015.

Putri, S. P., Nakayama, Y., Matsuda, F., Uchikata, T., Kobayashi, S., Matsubara, A., et al. (2013). Current metabolomics: Practical applications. *Journal of Bioscience and Bioengineering, 115*, 579–589. doi:10.1016/j.jbiosc.2012.12.007.

Ramadan, Z., Jacobs, D., Grigorov, M., & Kochhar, S. (2006). Metabolic profiling using principal component analysis, discriminant partial least squares, and genetic algorithms. *Talanta, 68*, 1683–1691.

Raman, K., & Chandra, N. (2009). Flux balance analysis of biological systems: Applications and challenges. *Briefings in Bioinformatics, 10*, 435–449. doi:10.1093/bib/bbp011.

Riter, L. S., Vitek, O., Gooding, K. M., Hodge, B. D., & Julian, R. K. (2005). Statistical design of experiments as a tool in mass spectrometry. *Journal of Mass Spectrometry, 40*, 565–579.

Rocke, D. M. (2004). Design and analysis of experiments with high throughput biological assay data. *Seminars in Cell & Developmental Biology, 15*, 703–713.

Savorani, F., Tomasi, G., & Engelsen, S. B. (2010). icoshift: A versatile tool for the rapid alignment of 1D NMR spectra. *Journal of Magnetic Resonance, 202*, 190–202.

Scalbert, A., Brennan, L., Fiehn, O., Hankemeier, T., Kristal, B. S., van Ommen, B., et al. (2009). Mass-spectrometry-based metabolomics: Limitations and recommendations for future progress with particular focus on nutrition research. *Metabolomics, 5*, 435–458.

Schilling, C. H., Schuster, S., Palsson, B. O., & Heinrich, R. (1999). Metabolic pathway analysis: Basic concepts and scientific applications in the post-genomic era. *Biotechnology Progress, 15*, 296–303.

Schölkopf, B., Smola, A., & Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation, 10*, 1299–1319.

Slupsky, C. M., Rankin, K. N., Wagner, J., Fu, H., Chang, D., Weljie, A. M., et al. (2007). Investigations of the effects of gender, diurnal variation, and age in human urinary metabolomic profiles. *Analytical Chemistry, 79*, 6995–7004.

Smilde, A. K., Jansen, J. J., Hoefsloot, H. C., Lamers, R. J., van der Greef, J., & Timmerman, M. E. (2005). ANOVA-simultaneous component analysis (ASCA): A new tool for analyzing designed metabolomics data. *Bioinformatics, 21*, 3043–3048. doi:10.1093/bioinformatics/bti476.

Smilde, A. K., Westerhuis, J. A., Hoefsloot, H. C. J., Bijlsma, S., Rubingh, C. M., Vis, D. J., et al. (2010). Dynamic metabolomic data analysis: A tutorial review. *Metabolomics, 6*, 3–17. doi:10.1007/s11306-009-0191-1.

Smith, C. A., O'Maille, G., Want, E. J., Qin, C., Trauger, S. A., Brandon, T. R., et al. (2005). METLIN: A metabolite mass spectral database. *Therapeutic Drug Monitoring, 27*, 747–751.

Smolinska, A., Blanchet, L., Buydens, L. M., & Wijmenga, S. S. (2012). NMR and pattern recognition methods in metabolomics: From data acquisition to biomarker discovery: A review. *Analytica Chimica Acta, 750*, 82–97. doi:10.1016/j.aca.2012.05.049.

Steinley, D., & Brusco, M. J. (2008). Selection of variables in cluster analysis: An empirical comparison of eight procedures. *Psychometrika, 73*, 125–144.

Steuer, R. (2007). Computational approaches to the topology, stability and dynamics of metabolic networks. *Phytochemistry, 68*, 2139–2151. doi:10.1016/j.phytochem.2007.04.041.

Stretch, C., Eastman, T., Mandal, R., Eisner, R., Wishart, D. S., Mourtzakis, M., et al. (2012). Prediction of skeletal muscle and fat mass in patients with advanced cancer using a metabolomic approach. *The Journal of Nutrition, 142*, 14–21.

Szczesniak, R. D., McPhail, G. L., Duan, L. L., Macaluso, M., Amin, R. S., & Clancy, J. P. (2013). A semiparametric approach to estimate rapid lung function decline in cystic fibrosis. *Annals of Epidemiology, 23*, 771–777.

Szymanska, E., Saccenti, E., Smilde, A. K., & Westerhuis, J. A. (2012). Double-check: Validation of diagnostic statistics for PLS-DA models in metabolomics studies. *Metabolomics, 8*, 3–16. doi:10.1007/s11306-011-0330-3.

Theodoridis, G. A., Gika, H. G., Want, E. J., & Wilson, I. D. (2012). Liquid chromatography–mass spectrometry based global metabolite profiling: A review. *Analytica Chimica Acta, 711*, 7–16.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63*, 411–423.

Timmerman, M. E., Ceulemans, E., De Roover, K., & Van Leeuwen, K. (2013). Subspace K-means clustering. *Behavior Research Methods, 45*, 1011–1023.

Timmerman, M. E., Ceulemans, E., Kiers, H. A., & Vichi, M. (2010). Factorial and reduced K-means reconsidered. *Computational Statistics & Data Analysis, 54*, 1858–1871.

Timmerman, M. E., Hoefsloot, H. C., Smilde, A. K., & Ceulemans, E. (2015). Scaling in ANOVA-simultaneous component analysis. *Metabolomics,*. doi:10.1007/s11306-015-0785-8.

Tomar, N., & De, R. K. (2013). Comparing methods for metabolic network analysis and an application to Metabolic Engineering. *Gene, 521*, 1–14.

Tomasi, G., van den Berg, F., & Andersson, C. (2004). Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *Journal of Chemometrics, 18*, 231–241.

Trygg, J., & Wold, S. (2002). Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics, 16*, 119–128.

Ultsch, A. (2003). *U*-matrix: A tool to visualize clusters in high dimensional data*. Marburg: Fachbereich Mathematik und Informatik.

van den Berg, R. A., Hoefsloot, H. C., Westerhuis, J. A., Smilde, A. K., & van der Werf, M. J. (2006). Centering, scaling, and transformations: Improving the biological information content of metabolomics data. *BMC Genomics, 7*, 142.

VanDyke, R., Ren, Y., Sucharew, H. J., Miodovnik, M., Rosenn, B., & Khoury, J. C. (2012). Characterizing maternal glycemic control: A more informative approach using semiparametric regression. *Journal of Maternal-Fetal and Neonatal Medicine, 25*, 15–19.

Velagapudi, V. R., et al. (2010). The gut microbiota modulates host energy and lipid metabolism in mice. *Journal of Lipid Research, 51*, 1101–1112.

Vettukattil, R. (2015). Preprocessing of raw metabonomic data. *Metabonomics: Methods and Protocols, 1*, 123–136.

Vichi, M., & Kiers, H. A. (2001). Factorial k-means analysis for two-way data. *Computational Statistics & Data Analysis, 37*, 49–64.

Wang-Sattler, R., Yu, Z., Herder, C., Messias, A. C., Floegel, A., He, Y., et al. (2012). Novel biomarkers for pre-diabetes identified by metabolomics. *Molecular Systems Biology,*. doi:10.1038/msb.2012.43.

Wetmore, D. R., Joseloff, E., Pilewski, J., Lee, D. P., Lawton, K. A., Mitchell, M. W., et al. (2010). Metabolomic profiling reveals biochemical pathways and biomarkers associated with pathogenesis in cystic fibrosis cells. *Journal of Biological Chemistry, 285*, 30516–30522. doi:10.1074/jbc.M110.140806.

Wiechert, W. (2002). Modeling and simulation: Tools for metabolic engineering. *Journal of Biotechnology, 94*, 37–63.

Wishart, D. S. (2007). Current progress in computational metabolomics. *Briefings in Bioinformatics, 8*, 279–293.

Wishart, D. S., Jewison, T., Guo, A. C., Wilson, M., Knox, C., Liu, Y., et al. (2013). HMDB 3.0—The human metabolome database in 2013. *Nucleic Acids Research, 41*, D801–D807. doi:10.1093/nar/gks1065.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. *Multivariate Analysis, 1*, 391–420.

Wold, S., Ruhe, A., Wold, H., & Dunn, W. J. (1984). The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing, 5*, 735–743.

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems, 58*, 109–130.

Xi, Y., & Rocke, D. M. (2008). Baseline correction for NMR spectroscopic metabolomics data analysis. *BMC Bioinformatics, 9*, 324.

Xia, J., Broadhurst, D. I., Wilson, M., & Wishart, D. S. (2012a). Translational biomarker discovery in clinical metabolomics: An introductory tutorial. *Metabolomics, 9*, 280–299. doi:10.1007/s11306-012-0482-9.

Xia, J., Mandal, R., Sinelnikov, I. V., Broadhurst, D., & Wishart, D. S. (2012b). MetaboAnalyst 2.0—A comprehensive server for metabolomic data analysis. *Nucleic Acids Research, 40*, W127–W133.

Xia, J., Psychogios, N., Young, N., & Wishart, D. S. (2009). MetaboAnalyst: A web server for metabolomic data analysis and interpretation. *Nucleic Acids Research, 37*, W652–W660.

Xing, E. P., Jordan, M. I., Russell, S., & Ng, A. Y. (2002). Distance metric learning with application to clustering with side-information. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems* (pp. 505–512). Cambridge, MA: MIT Press.

Yan, M., & Ye, K. (2007). Determining the number of clusters using the weighted gap statistic. *Biometrics, 63*, 1031–1037.

Yang, C., He, Z., & Yu, W. (2009). Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis. *BMC Bioinformatics, 10*, 4.

Zhang, S., Gowda, G. N., Asiago, V., Shanaiah, N., Barbas, C., & Raftery, D. (2008). Correlative and quantitative 1 H NMR-based metabolomics reveals specific metabolic pathway disturbances in diabetic rats. *Analytical Biochemistry, 383*, 76–84.

Zhang, J. D., & Wiemann, S. (2009). KEGGgraph: A graph approach to KEGG PATHWAY in R and bioconductor. *Bioinformatics, 25*, 1470–1471.