CHAPTER FOURTEEN

An Overview of Metabolomics Data Analysis: Current Tools and Future Perspectives

Santosh Lamichhane*, Partho Sen*, Alex M. Dickens*, Tuulia Hyötyläinen†, Matej Orešič*,‡,1

*Turku Centre for Biotechnology, University of Turku and Åbo Akademi University, Turku, Finland
†Department of Chemistry, Örebro University, Örebro, Sweden
‡School of Medical Sciences, Örebro University, Örebro, Sweden
[1] Corresponding author: e-mail address: matej.oresic@utu.fi

Contents

1. METABOLOMICS: AN OVERVIEW

Metabolites are small molecules (molecular weight <1500 Da) which act as intermediates or end products of cellular metabolism [1]. Metabolomics is a comprehensive analysis of metabolites produced by a biological system or derived from various other external sources such as diet, microbes, or xenobiotic sources [1–3]. Metabolomic analysis is commonly categorized into targeted and untargeted approaches, based on the coverage of the metabolites [4]. Non-targeted analyses are global approaches optimized for covering as many metabolites as possible in a given biological matrix.

Comprehensive Analytical Chemistry, Volume 82
ISSN 0166-526X
https://doi.org/10.1016/bs.coac.2018.07.001

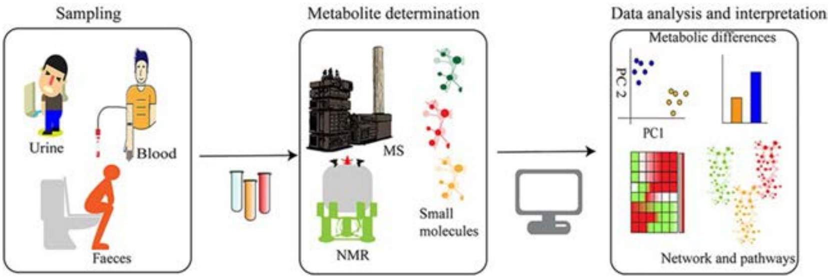

Fig. 1 An illustration of the metabolomics workflow.

In contrast, targeted metabolomics aims at analysing a limited number and/or biologically specific set of metabolites.

Typical metabolomics workflow starts with the collection of biologically relevant samples such as urine, plasma, faeces, and tissue, in order to answer the specified research question(s) [5] (Fig. 1). These biological specimens contain a wide range of metabolites that have the potential to provide insights into fundamental cellular processes. High-throughput analytical platforms such as nuclear magnetic resonance (NMR) spectroscopy, gas chromatography (GC), and/or liquid chromatography (LC) coupled to mass spectrometry (MS) are employed to measure the metabolites within the biological matrix [6, 7]. These complex measurements may result in hundreds of metabolites (non-targeted metabolomics) or a few prespecified metabolites (targeted metabolomics), depending on the objective of the investigation.

Herein, we outline various univariate and multivariate data handling strategies in metabolomics studies. In addition, we elaborate how metabolic modelling together with the integration of metabolomics and other multi-omics datasets can be used as a tool to understand cellular metabolism.

2. METABOLOMICS DATA ANALYSIS

The metabolomics datasets commonly contain hundreds to thousands of variables [6, 8–11]. For example, a single MS measurement typically contains multiple data points. These data points may originate from diverse sources such as variation in instrumental condition, heterogeneity of the sample, or due to reduced signal detection (instrumental error). These may result in an inaccurate identification of metabolites, poor quantitation, and ultimately introduce a statistical error and affect the efficiency of data analysis methods applied. Thus, strict quality control and preprocessing of data generated from the metabolomics analytical platform is necessary to retrieve meaningful information before the actual data analysis.

2.1 Data Preprocessing

The type of data preprocessing depends on the type of experiment performed to acquire the data. The challenge is that metabolomics experiments can generate a large number of data points, which can influence the data modelling; therefore, these data need to be preprocessed reliably. The key to data preprocessing is to ensure that the correct quality controls are present in the experiment. The basic steps involved in the data preprocessing include peak alignment, peak filtering, peak identification, and metabolite identification.

2.1.1 NMR Data Preprocessing

The main advantages of NMR experiments for metabolomics are the reproducibility of the technique due to the quick data acquisition time and the fact that little sample preparation is required. However, the individual spectra can vary from sample to sample due to experimental variation over time. These experimental conditions include: pH variation, temperature deviations, difference in macromolar binding [12]. Furthermore, all spectra also need to be phased correctly, and the baselines have to be adjusted to ensure accurate integration of the peaks. These processes can be performed manually, in NMR acquisition software such as Topspin (Bruker, Germany). However, large numbers of samples are often run in a single NMR experiment, thus making manual processes undesirable. There are a variety of commercial NMR data processing tools available, such as Mestrelab Research NMR tool (Mestrelab Research, S.L, Spain), Amix (Bruker, Germany), and Chenomx NMR Suite (Chenomx, Canada). Additionally, there are packages in both R and MATLAB® which implement published methods for automatic phasing [13, 14], baseline correction [15], and peak alignment [16, 17]. For further reading on NMR preprocessing, please refer to these two reviews which discuss these topics in detail [18, 19]. Following spectral alignment, the peaks are then integrated. NMR peak integration can be performed using established packages in R and MATLAB®. For instance, Hao et al. have developed an R package (BATMAN) for the automated quantification of metabolites from NMR spectra [20]. On the other hand, to simplify the subsequent data analysis and to reduce noise effects the NMR spectra may also be divided into buckets of either set or variable width depending on the algorithm [21, 22]. These buckets can then be used in the subsequent multivariate analysis. However, binning reduces the spectral resolution, increases the risk of peak overlapping and masking of less abundant yet potentially significant metabolites.

2.1.2 Mass Spectrometry Data Preprocessing

MS is a more sensitive technique for measuring metabolites, and as such, the amount of data that is typically generated per instrument run is greatly increased as compared to NMR. However, the basic data preprocessing steps remain similar. The exact steps depend on the type of chromatographic separation technique used and what the type of mass spectrometry data acquired.

2.1.2.1 Targeted Metabolomics Assays

When a small number of metabolites are detected in a targeted fashion, the data preprocessing tends to be straightforward. Generally, the quantitation of the peaks can be performed in the vendor-specific software. Examples include Analyst (AB Sciex), Mass Hunter (Agilent), and MassLynx (Waters). Generally, the quantitation in the acquisition software tends to be somewhat basic and for more advanced projects vendors often sell companion software, such as Multi Quant (AB Sciex), Mass Hunter Quant (Agilent), and TASQ (Bruker). The main aim of all these software packages is to integrate the appropriate peak, normalize to the internal standard and calibrate to a standard curve. Again, this process can also be performed in MATLAB® or R as well as other freeware software; however, the raw MS data often need to be converted to an appropriate format before they can be processed. The peak identification is done based on their mass spectrum and retention times (LC) and retention index (GC).

2.1.2.2 Untargeted Metabolomics Assays

Generally, the processing of the complex untargeted metabolomics data depends on the data acquisition methodology. However, the basic steps remain the same as in targeted approach. Initially, the data have to be aligned; secondly, it is filtered and then normalized. Increasingly, for untargeted metabolomics data are being acquired from high-resolution (HRMS) instruments (such as quadrupole time-of-flight or Orbitrap) for profiling of both polar and nonpolar metabolites. The main reason for this approach is the eventual reproducibility of this type of data in large sample numbers and the speed at which the data can be acquired. Additionally, the inclusion of a chromatographic technique (whether GC or LC) also removes ion suppression issues as observed with direct infusion approaches.

The first step in processing HRMS data is deconvolution, which identifies peaks based on the similarity of the gathered HRMS spectra. These algorithms often contain the ability to deisotope the spectra. This is required due to the relatively high (1.1%) of ^{13}C that occurs naturally. These peak finding algorithms can sometimes be implemented in the vendor-specific software. However, more commonly utilized are open source software packages such as MZmine [23], MS-DIAL [24], and XCMS [25]. These software solutions have been developed by the community and provide powerful tools for the preprocessing of MS data. The open source nature of these software allows for the development of new tools. For example, the ability to process GC–MS-based data was recently added as a package to MZmine. XCMS is both an online tool and executable on your local machine. Recently, Li et al. compared the feasibility of five widely used data analysis software packages (MS-Dial, MZmine 2, XCMS, MarkerView, and Compound Discoverer) for untargeted metabolomics data processing [26]. The authors compared the ability of these packages for feature detection, quantification, and marker selection using a well-defined benchmark sample set. The authors found that, even though these five widely used data analysis software packages gave similar performance in detecting true features from benchmark samples, MZmine 2 outperformed other software in terms of quantification accuracy and was therefore, reported to produce the truest discriminating markers together with the fewest false markers [26]. This could be attributed to the fact that MZmine provides the user with greater flexibility for tuning the various processing parameters and thus offering greater flexibility to optimize them for the specific MS method used.

The main aim of all of the above data processing tools is to extract the peaks from the raw data and then normalize them to internal standards. Increasingly, the idea of quantifying specific metabolites with standard curves and the appropriate internal standards is being used in metabolomics with the remainder of the metabolites quantified relatively based on selected internal standards, such as lipid class-specific standards as used in lipidomics [27, 28]. The outcome of these steps is the integrated peak area, or in rare exceptions peak height data, which can be used in multivariate statistics later on. Prior to these values being used, however, additional filtering steps may be needed. For example, commonly the peaks are filtered based on the number of times they were detected across the samples. The crude and commonly applied way is to set a threshold, e.g., to exclude all peaks that appear in less than 50% of all samples. However, this approach risks eliminating the metabolites that are unique to specific (smaller) subgroups in the sample. Therefore, a better strategy is to exclude peaks at a study group level. Some peak finding algorithms implemented in MZmine [23] and T-Rex (Bruker) will allow second-pass peak detection, i.e., to loosen the parameters used to identify peaks that are closely matched to the original peak. This allows for the drift in retention time for that sample or noise during that specific peak to be taken into account. This second-pass peak detection at the data usually reduces the number of zeros in the dataset. For further processing, however, the remaining zeros often have to be imputed. The simplest method is to use the half-minimum value of that metabolite. However, recent work has shown that using a K-means clustering approach may be a more robust approach [29].

2.1.2.3 Peak Identification

The final step to any data preprocessing is to identify as many of the peaks as possible. This is commonly done by comparing the spectroscopic data with known metabolite libraries. Again, this process can be performed manually for smaller numbers of metabolites, but with larger metabolite screens this becomes unfeasible. Therefore, data processing software includes a library search function and assign a similarity score to each peak. For peaks of interest, these can be followed-up with further experiments. For example, with NMR it is possible to perform 2D spectroscopy to show the coupling patterns of the hydrogens (COSY) or how the hydrogens are connected to the carbons (HMBC) in the molecule. Using MS, fragmentation experiments can be performed to identify and characterize the specific fragments from the molecule. There are a growing number of metabolite libraries which experimental data can be compared to. Some of them are specific to the chromatographic separation technique used. For example, the National Institute of Standards (NIST) GC library is commonly used in vendor-specific software to perform spectral match searches. However, the retention indices are not easily accessible in the NIST libraries and as such the software often does not match based on this critical piece of information. Therefore, other software packages such as Guineu [27] have been developed in order to search the GOLM database (http://gmd.mpimp-golm.mpg.de/). For LC-based experiments, there are several databases such as METLIN (https://metlin.scripps.edu), NIST MS (https://www.nist.gov/), HMDB (http://www.hmdb.ca/), and lipid maps (http://www.lipidmaps.org/). There are also vendor-specific libraries, such as the Bruker library, which can aid with identification as it contains isotope pattern information which on HRMS systems can provide another degree of confidence in the identification. For the identification of unknown metabolites, the GOLM and Japan Mass Bank (http://www.mssj.jp) can be used to predict the likely chemical structures based on the fragmentation pattern or MS/MS spectra, respectively.

2.2 Data Analysis

After data processing, metabolomics measurements usually result in signal intensities across a wide range of metabolites. Univariate and multivariate statistics are used as routine approach to extract relevant information from these complex datasets [30]. Univariate approaches involve analysis of a single variable in question and require prior knowledge of the measured variable [30]. However, biologically specific information may be hidden in the relation between multiple variables (metabolites). Therefore, univariate analysis is limited to providing useful information if the effect occurs in one single variable, which is rare in biological systems. In such cases, untargeted metabolomics analysis in combination with multivariate data analysis (MVDA) is the optimal data handling strategy. MVDA extracts the hidden structure from the complex metabolomics data and determines the pattern of metabolites (if any) that change between various groups, e.g., treatment vs control subjects [31]. Although, the dataset becomes exceedingly complex the unbiased nature of untargeted multivariate approach increases the chances for biological discoveries. Fig. 2 depicts a simplified overview of the metabolomics data analysis workflow.

2.2.1 Univariate Approach

In univariate analysis, only one variable (e.g. metabolite intensity) is used as an input for statistical analysis. In principle, the univariate test compares the statistic value (usually mean or median) between two study groups of samples (e.g. disease vs healthy control). There are a wide variety of univariate statistical tests designed to compare the mean or medians. Based on heteroscedasticity, normality, and independence of the datasets the comparison could be either parametric or nonparametric. Typical univariate statistical tools for metabolomics data include: *t*-test (paired or unpaired), analysis of variance (ANOVA), Wilcoxon rank-sum test, Kruskal–Wallis test or others dependent on experimental design and data distribution [32]. For example, ANOVA, which can be implemented in a study designed

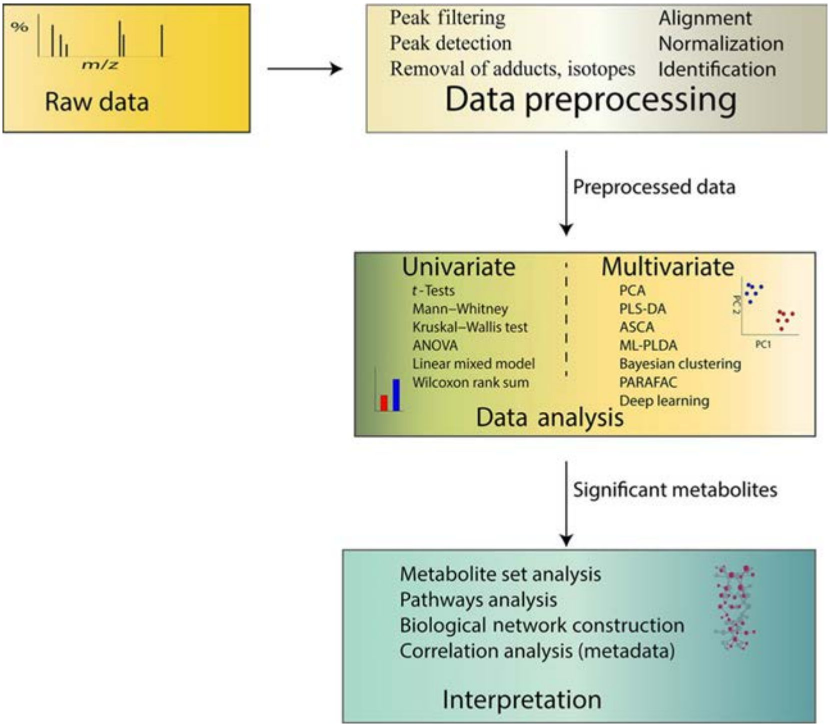

Fig. 2 Metabolomics data handling workflow, from data acquisition, processing, to analysis and interpretation.

to compare the interaction among different factors such as age, gender, treatments (placebo vs test) assuming the data are normally distributed [33]. Similarly, linear mixed models can be used to compute the effects of fixed (groups or class, time, and the interaction between class and time) and random effect (subjects) variables on a response variable (for example, a metabolite) [34]. These comparisons result into a statistical inference (nominal *P*-values), which determines whether the observed differences (mean or median) between the tested study groups are due to randomness or not. However, in the case of metabolomics experiments, the number of hypotheses tests can be large due to multiple metabolites being measured. This increases the possibility of identifying false positive results. Therefore, to reduce the probability of having false positives, the statistical inferences are adjusted by means of multiple testing. Broadhurst and Kell [35] have discussed the statistical strategies for determining the false discoveries in metabolomics and related experiments. The false discovery rate (FDR) estimation for multiple comparisons can be performed using Storey approach [36] or Benjamin and Hochberg approach [37], which are mainly adapted from earlier emerged omics (genomics) technologies. However, it should also be noted that this approach theoretically increases the probability of obtaining false negative results.

Wide varieties of packages for univariate statistical are available for metabolomics data analysis. Vinaixa et al. suggest various resources for univariate analysis to handle the dataset obtained from metabolomics experiments [32]. These univariate methods are based on scripts programmed either in MATLAB® or R [38]. There are also online tools including MetaboAnalyst [39], Metabox [40] for univariate statistical analysis of metabolomics-derived data. In fact, previous publications have demonstrated the applicability of the univariate approach for metabolomics data handling [7, 41, 42]. A large body of literature has shown that univariate analysis leads to the discovery of those metabolites that varies between two study populations (e.g. healthy vs diseased) [7, 42–47]. Researchers often prefer a univariate approach because it is a simple and robust tool for understanding the biological condition of the specified research question(s). However, univariate methods failed to show a significant/meaningful difference between groups if the property of interest lies in the relationship between many variables (also known as covariance) but not in the variance of a specified single variable [30, 33, 48]. Further limitations of univariate include: (i) a priori knowledge about the variables included in the analysis (for hypothesis testing) and (ii) Loss of large-scale information obtained from the complex biological matrix; thereby, reducing the chances of making new discoveries due to restricted analysis of few known variables.

2.2.2 Multivariate Approach

Data generated in untargeted metabolomics studies involve the measurement of more metabolites than the number of samples or subjects [6, 10, 11, 49]. Multivariate approaches simplify the complexity in the high-dimensional data into manageable variables, facilitating a straightforward and meaningful interpretation of the large-scale complex metabolomics datasets. In principle, MDVA reduces the dimensionality of the dataset and is used to find out the concurrent relationship between the variables. Multivariate methods facilitate the determination of specific patterns along with classification and discrimination of variables in the datasets. Table 1 provides an overview of data analysis methods and relevance for metabolomics research.

Table 1 Overview of Data Analysis Methods and Relevance for Metabolomics Research

Data Analysis	Main Methods	Strengths	Weaknesses
Univariate analysis	*t*-Tests Mann–Whitney Kruskal–Wallis test ANOVA Linear mixed model Wilcoxon rank sum	Simple approach in terms of application Powerful method for targeted questions Easy to interpret No masking of information by noisy variables	Misses the variable relationship Loss of collected information Required prior knowledge Difficult to determine the outliers Waste of data FDR correction can increase the chance of false negatives
Multivariate analysis	PCA PLS-DA ASCA ML-PLDA Bayesian clustering PARAFAC Deep learning	Handles tens to hundreds variable simultaneously Lowers the dimensionality into to manage number Independent variable complements other variables because MVVDA focus on the relations between variables Requires no prior knowledge of variables	May be overfitting due to high level of noise Some abundant variable may mask information of other less variables Correlation (covariance) may be challenging to accomplish Biological interpretation, in particular relation to the original variables, may be challenging
Metabolite set analysis	Hypergeometric test Fisher Exact Kolmogorov–Smirnov test	In between uni- and multivariate analysis Biological study design taken in account for comprehensive interpretation	Metabolite sets are difficult to estimate compressively Metabolic pathway databases are incomplete and manually interpreted, which leads to biased interpretations
Data fusion	Multiblock-PCA, -PLS Correlations analysis DISCO JIVE	Global view of the variables under same study system Convenient approach to understand relationships between the datasets	Multivariate correlations might be difficult for interpretation No clear guidelines for metabolomics data fusion exist

Principal component analysis (PCA) [50] and partial least square discriminant analysis (PLS-DA) [51] are two of the most frequently used multivariate methods in metabolomics studies. PCA is a key multivariate tool to identify patterns and outliers in the metabolomics datasets. PCA is an unsupervised approach, which does not require any prior knowledge to identify class differences or groupings in the multivariate datasets. PCA is built on the principle that the redundancy within multivariate datasets can be orthogonally decomposed into few linear weighted variables known as principal components (PCs). These PCs represent a new coordinate system that can be imagined as a mathematical summary of the measured variables in the dataset. A key characteristic of PCA is the PCs being projected in a way that first PC (PC1) explains the greatest variance in the data matrix, followed by the second PC (PC2), the third PC (PC3), and so on. Bro and Smilde have described the mathematical structure of the PCA model in a tutorial review (for mathematical details, see Refs. [50, 52]). PCA is a powerful tool to identify the trends, grouping, and outliers in the complex metabolomics dataset. In addition, PCA could be used to determine confounding factors in metabolomics studies [53]. However, in terms of applicability, if the data are not linearly correlated and within-group variation is less than between-group variation, PCA fails to reveal group structure. Furthermore, valuable information can be lost because in PCA the greatest amount of variance will largely bias the interpretations from analyses [52, 54]. For that reason, PCA can be considered as an observational tool rather than an explicit model for exploring potential biomarkers in a metabolomics study.

PLS-DA is a supervised method where a priori knowledge of the samples or groups is required to build the classification model. The aim in PLS-DA is to find the metabolite pattern that is underlying the discrimination between the two study groups. The basis of the PLS-DA model is PLS [55], which decomposes two matrices (matix1: Metabolite information/intensities, matix2: categorical/classifier variables such as disease vs control state) to find common variability and to construct a model that seeks for covariance between these two matrices [51]. This supervised approach tends to improve the separation between (two or more) groups of samples. For that reason, PLS-DA is widely used for classification purposes and biomarker identification in metabolomics studies [9, 10]. However, the supervised nature of PLS-DA provides the separations between variables (in score space) including variations unrelated to the categorical variable, which implies that the discriminatory information obtained from the scores in PLS-DA analysis may not be directly correlated with the classifying variables [51, 56–58]. Especially, for metabolomics data from human studies where often large variations can be observed between the subjects rather than within the subjects (e.g. treatment effect). It thus complicates the interpretation of PLS-DA information. However, PLS-DA has been extended to related methods such as OPLS-DA [59] and ML-PLSDA [60, 61], in order to discriminate between two or more categorical variables (classes) in the multivariate dataset. Unlike PLS-DA, OPLS-DA uses a single component as a predictor for the class, while ML-PLSDA exploits the paired data structure in the multivariate study design [56, 59]. The advantage of these methods over PLS-DA methods is the possibility to examine the differences within the subjects and between the subjects without being confounded with the actual effect (for instance, the treatment effect) [59]. Alternatively, supervised approach such as PLS-regression, logistic regression modelling could be built on selected variables (derived either PCA or univariate analysis) to assess the feasibility of predicting the categorical information from the dataset [7, 62].

Other than discriminant analysis, experimental design in combination with multiway data analysis can be a powerful approach for the identification of biomarkers in the complex and dynamic metabolomics datasets. Subsequent dimensionality reduction techniques, such as analysis of variance (ANOVA)-simultaneous component analysis [63], model-based clustering [64, 65], parallel factor analysis (PARAFAC)-ASCA [66], and multivariate linear mixed models [67, 68], have been implemented to improve the inference of covariate effects in metabolomics dataset. Based on experimental design, these multiway methods can: (i) increase the interpretability of the complex multivariate model in terms of experimental questions and (ii) reduce the risk of discovery of false relation between experimental covariates including diet, age, gender, etc. and the detected metabolic-markers [69]. Therefore, these methods are becoming increasingly popular for longitudinal metabolomics data analysis as well as for dealing with the covariates in metabolomics experiments. Recently, Xia et al. introduced MetATT, a web-based tool for time-series and two-way metabolomic data analysis, which offers a number of complementary approaches including ASCA, two-way ANOVA through an intuitive web interface [70]. Most of these multiway methods are also readily available in.

R or MATLAB® or within the set of online tools for metabolomic data analysis such as MetaboAnalyst [39] and Metabox [40].

Furthermore, new machine learning techniques and artificial intelligence methods such as deep neural networks, t-distributed stochastic neighbour embedding (t-SNE) [71] can be adapted for the analysis of metabolomics datasets [72, 73]. These techniques are becoming increasingly popular and are effective in classification [73] as well as visualization of the multivariate dataset as compared to other conventional techniques [74]. However, implementation of deep learning methods to metabolomics datasets is still evolving and extensive statistical evaluation is required to benchmark the performances of these methods, and also to infer its sensitivity and applicability.

2.2.3 Assessment of Reliability of Supervised Multivariate Models

One major challenge in metabolomics data analysis is to validate the resultant multivariate models due to the high level of noise and the small number of samples in the dataset [35]. Especially, supervised multivariate models are prone to overfitting of the data, which results in misleading or irreproducible metabolic signatures [51]. Therefore, to determine the genuine relationships in the dataset rather than the representation of noise, it is vital to properly validate and statistically evaluate the supervised multivariate models. There are two steps that might be considered critical when generating supervised multivariate models (i) the selection of the optimal number of latent variables and (ii) the assessment of the overall quality of the model.

The optimal number of latent variables will provide proper classification between samples from different predefined classes [57]. The appropriate number of latent variables can be determined by cross-validation before establishing the actual supervised model. Cross-validation can be performed by fragmenting metabolomic dataset into training, test, and/or validation sets. In general, a supervised model is developed and optimized using the validation and training sets. The test set is then used to evaluate whether they are classified correctly or not, i.e., testing the model performance in the classification of future observations [51, 56, 57].

For the supervised multivariate models, the statistical significance of the generated model can be estimated by comparing the diagnostic statistics such as sensitivity, specificity, the number of misclassifications (NMC), the area under the receiver operating characteristic (AUROC), Q2 and/or discriminant Q2 [51, 56, 57]. The Q2 has been commonly used as diagnostic statistics for supervised multivariate models, which was initially introduced for classification model PLS to evaluate the class prediction ability [55]. Q2 is calculated as $Q2 = 1 -$ prediction error sum of squares (PRESS)/total sum of squares (TSS).

Metabolomics statistical packages such as SIMCA, PLS-toolbox, and other packages written in R or MATLAB® have used Q2 as a default diagnostic statistic to evaluate the PLS-DA models. Comparison of various diagnostic statistics showed that NMC and AUROC are more efficient and reliable than Q2 or DQ2 [51, 56]. Mathematically, NMC can be calculated as the sum of false positive and false negative class prediction. In other words, NMC can be understood as the misclassification error referred to the number of tests that we know that belong to a category, group, or class [51]. Zero NMC denotes perfect discrimination (i.e. 100% correct classification). On the other hand, AUROC is the plot of the sensitivity (true positives found as a percentage of all positives) vs 1-specificity (false positives as a percentage of all negatives). In general, AUROC values range from zero to one. Zero means no discrimination and one means there is perfect discrimination between classes.

Even though NMC and AUROC measure the statistical relevance, permutation testing is recommended to determine the statistical significance of the diagnostic statistics (*P*-value) [51, 57]. Permutation tests assume that there is no difference between the two groups that have been randomized. Permutation testing for supervised multivariate analysis is usually performed by randomly assigning the class labels to the samples, and then evaluates the statistical significance. The rationale is that the random class labels should not be able to predict the classes very well [51, 56]. Other than *P*-values, Variable Importance for Projection (VIP) and selectivity ratio (SR) can be used for selection of the most relevant and predictable variables in the supervised multivariate models including PLS-DA [75]. VIP is probably the most commonly used multivariate method implemented for variable selection in metabolomics [76, 77]. Thereafter, the SR method is increasingly being used for the selection of chemical signatures in omics research [78–80]. Farrés et al. have discussed the possible advantages and disadvantages of applying these variable selection methods in metabolomics dataset, however, the choice of the method may be dependent on experimental design and the aim of the study [75].

2.2.4 Conjugation of Metabolomics With Other Omics

Multivariate data-fusion approaches such as joint and individual variation explained (JIVE) [81], canonical correlation analysis (CCA), DIStinct COmmon SCA (DISCO-SCA) [82], multiblock (PLS/PCA) methods [83] may also open new avenues for enhancing the biological interpretation of the metabolomics datasets [62, 83, 84]. Li et al., exemplified by fusing multidimensional datasets originating from epigenetics, transcriptional, and post-transcriptional analysis using multiblock PLS analysis [85]. This approach has enabled the authors to identify multilayers of gene regulatory modules that were vital for ovarian tumour progression. In another study, Kostic et al. analysed the association between gut microbes and faecal metabolites using penalized CCA analysis [43]. Here, authors were able to identify a statistical association between gut microbes (*Ruminococcus* and *Veillonella*) and circulating metabolites (bile acids). Besides, multivariate data fusion, other statistical associations between the metabolome data, metadata, and/or other omics data could provide a mechanistic overview of how metabolites are associated with the biological processes [84]. For instance, Pedersen et al. performed correlation analysis between operational taxonomic units (OTU) pairs of gut microbiome and metabolites that provided an explanation for the effect of insulin resistance on the metabolic activity of gut [86]. However, most of these integrative approaches especially multivariate fusion tools are not well established, as these approaches failed to explain the comprehensive functional biology. Therefore, it is essential to develop other large-scale computational techniques that can guide us to understand the biological aspects of metabolomics datasets [84].

3. METABOLITE SET ANALYSIS AND PATHWAY OVERREPRESENTATIONS

Several pathway databases that store diverse molecular information are available. These resources were designed with different objective(s). Kyoto encyclopaedia of genes and genomes (KEGG) [87] (http://www.genome.jp/kegg/), a widely used pathway database was built to capture cellular metabolism and processes of an organism. Recently, KEGG has embedded 16 different databases that contain information about genes, proteins, reactions, and pathways for various organism including human [88]. In addition, KEGG datasets and tools have been used for pathway predictions, metabolic network reconstructions, and overrepresentation analysis. Moreover, KEGG provided a platform for development of web-based tools [89] such as KeggArray (http://www.kegg.jp/kegg/download/kegtools.html) that allows visualization and interpretation of metabolomics alongside other omics datasets.

MetaCyc [90] (https://metacyc.org/) is a knowledge base of experimentally validated metabolic pathways. The database is curated based on bibliographic references. It contains reference pathways of most of the organism listed in the tree of life. The pathway tool allows the reconstruction of the metabolic network of an organism from its sequenced genome. MetaCyc also aided in the development of other tools such as HUMAnN2 [91]. In addition, PlantCyc (https://www.plantcyc.org/) is a metabolic reference database that includes 900 metabolic pathways, genes, enzymes, and compounds from 350 plant species [92].

The small molecule pathway database (SMPDB) [93, 94] (http://www.smpdb.ca) was designed for pathway elucidation and analysis. It contains ~30,000 small molecule pathways found in humans. Each metabolite is cross-linked to HMDB [95] (http://www.hmdb.ca/) and DrugBank [96] (https://www.drugbank.ca/). HMDB is an open source knowledge base that lists information about metabolites found in the human. The database also stores disease-specific metabolomics datasets which can be downloaded for additional analysis. For instance, marking the enrichment of serine in human cancer cells among the abundant cellular metabolites set as a background.

REACTOME [97] (https://reactome.org/) database provides an ordered and layered network of molecular transformations in human. It also helps in visualization, interpretation, and analysis of pathways. Furthermore, Ingenuity® pathway analysis (IPA) [98], another commercial tool allows the integration and interpretation of multiomics including metabolomics datasets for pathway analysis and identification. In addition to casual network analysis, IPA-BioProfiler guides in the identification of therapeutic or toxicity targets and markers.

Furthermore, ChEBI (https://www.ebi.ac.uk/chebi/), HumanCyc (https://humancyc.org/), Lipid MAPS (http://www.lipidmaps.org/), ChemSpider (http://www.chemspider.com/), METLIN (https://metlin.scripps.edu/), Recon2 [99], virtual metabolic human (https://vmh.uni.lu/), and human metabolic atlas (HMA) [100] (http://www.metabolicatlas.org/) provide structural and functional information of metabolites, their associated reactions, and/or biochemical pathways.

Over the recent years, with the advancement of novel techniques and availability of high-throughput datasets, these databases have been expanded with an aim to fill the voids in cellular metabolism of different organisms, yet the system remains underdetermined abided by several challenges such as validation of predicted pathways. Combining different resources could help to fill this void by helping to develop a consensus reference pathway for each organism with improved metabolite-pathway coverage. This strategy has been limited by inappropriate standardization, multiple nomenclatures, and database redundancy. However, the knowledge base when standardized can be used as a background for metabolite set enrichment analysis (MSEA) and identifying the overrepresented disease-specific pathways in individuals [101–104]. Furthermore, it can be used for organism-specific metabolic reconstructions.

Metabolite set analysis, MSA or MSEA, is a group-based approach that can be used to test whether a set of metabolites is associated with a given phenotype, ultimately linking chemical data to biological information [103, 105]. MSEA follows the concept of gene set enrichment analysis (GSEA) [106]. The differentially abundant metabolites (case vs control) along with their statistical significance derived from univariate statistics (*P*-value, *t*-stats, or *z*-score) can be used as an input. MSEA applies these significances to the biologically connected metabolite groups [107]. For each set of metabolites, an overall enrichment score is computed to evaluate whether the metabolites involved are significantly altered in the given experimental condition [107]. In other words, MSEA usually takes a list of significantly altered metabolites from a biological matrix and use this information to compare it with a background metabolite set to indicate a biochemical pathway at a certain condition that can be investigated [103, 105].

In practice, MSEA depends on background metabolite library and statistical significance. The accuracy of MSEA predictions relies on (i) appropriate selection of statistical tests, which could be count based (hypergeometric test or Fisher Exact) or distribution based (Kolmogorov–Smirnov test) [101, 105, 108] and (ii) an extensively curated and biologically meaningful metabolite library [101, 103] to accurately bridge the chemical information to metabolic processes. There are various packages for MSEA such as MPEA [101], MSEA [103], PAPi [109], MBRole [104, 110], and ChemRICH [105], MetaboAnalyst [39] and Metabox [40] are web-based pipelines. These tools generate a comparative visualization (network maps) highlighting up- and downregulated metabolic pathway. Evaluation of several enrichments tools applied to metabolomics datasets is reviewed in Ref. [102].

4. MATHEMATICAL MODELLING OF METABOLISM

Metabolic models serve as a guide to postulate new hypotheses that can be tested and validated experimentally [111–113]. These models can be simulated under physiological conditions to understand biological mechanisms. Metabolic models provide scaffolds on which biological datasets can be overlaid. These models generate information about the regulation of metabolic pathways, metabolite uptake or secretion rates of the cells/tissue, growth and metabolic capacities of the organism under various conditions. Different strategies of metabolic modelling are discussed.

4.1 Constraint-Based Modelling

The availability of DNA sequences of a cell, tissue, or organism, combined with fluxomics and metabolomic data gives an opportunity for genome-wide metabolic reconstructions. These reconstructions can be subjected to predict phenotypes such as growth, flux distribution, metabolic capacities. Flux balance analysis (FBA) is a constrained-based approach (CBA) that can predict an organism's phenotypes under various conditions given the biomass composition [114]. FBA is static and based on steady-state assumption. It does not consider enzyme kinetics. However, this approach is suitable for un- or underdetermined metabolic system with limited knowledge of reaction mechanism(s) and enzyme kinetics. FBA has been widely used for discovery of drug targets [115, 116].

Genome-scale metabolic modelling (GSMM) is a constrained-based modelling where the models are stoichiometrically bounded with biochemical, genetic, and genomic information within a computational framework [114, 117–120]. The layered and ordered structure of genome-scale networks allows the integration of different types of omics data such as transcriptomics, proteomics, metabolomics, and fluxomics [121]. GSMM bridges and infers the metabolic genotype–phenotype relationship of an organism. Several algorithms were designed to integrate and contextualize omics data into the GSMM platform, but only handful of methods are known for integration of metabolomics datasets. One such method is INIT (Integrative Network Inference for Tissues) or tINIT algorithm which uses cell-specific protein and metabolic abundances to model and contextualize metabolic networks [122]. This draft metabolic network is curated with bibliographic evidence and expert's knowledge of metabolism. Metabolomic datasets overlaid on genome-scale networks have spotted reporter reaction(s); the reactions marked by significant and coordinated changes in the surrounding metabolites following environmental/genetic perturbations. On combining transcriptome data, it is possible to infer whether the reactions are hierarchically or metabolically regulated [123].

Moreover, GSMM has been extensively used to study metabolism in microbes [124, 125], lower eukaryotes, and human [99, 120]. Recon 1 was built with a vision to integrate and analyse biological datasets [126].

Thereafter, Edinburgh human metabolic network (EHMN), Recon 2 [99], Recon 2.2 [127], and human metabolic reaction (HMR) [100, 128] were designed to study human metabolism.

GSMMs together with metabolomics [2] and metagenomics [129] have been used to study gut microbial metabolism [84, 125, 130, 131] and their interaction with the host. GSMM as a tool was used to predict changes in the amino acids levels in the gut community. The predictions were validated by faecal and blood metabolomics data [130]. Furthermore, GSMM as an integrative tool has been used to model diet-tissue [120] and multitissue interactions in humans [132]. GSMM was studied in the context of various diseases such as cancer [133], nonalcoholic fatty liver disease (NAFLD) [100, 134], and diabetes [135]. The future scope of GSMMs to understand metabolism of immune cells in context to various diseases such as type 1 diabetes, rheumatic arthritis is reviewed in Ref. [111].

Elementary mode analysis (EMA) is another constraint-based metabolic pathway analysis tool that links the cellular phenotype to their corresponding genotype [136]. Unlike GSMM, EMA fragments the complex metabolic network into a minimal and unique set of enzymatic reactions called elementary modes (EM) [137]. EMs are required for maintenance of cellular functions at steady state. On the one hand, elementary flux modes (EFMs) or simply EMs depend on homogeneous linear constraints sourced from reaction irreversibilities and assumption of steady state. On the other hand, FBA depends on heterogenous linear constraints such as minimal and maximal reaction rates. Thus, it was not straightforward to integrate EM into FBA directly unless elementary flux vectors (EFVs) were introduced. EFV is a unifying framework that was designed to overcome such limitations [137]. EMA has a broad range of applications such as characterization of cellular phenotypes [138], analysis of metabolic network, and design of strains for metabolic engineering [139].

4.2 Kinetic Modelling of Metabolic Pathways

Kinetic modelling (KM) is an alternative to static approaches [112, 140, 141]. Kinetic models are mechanistically represented by metabolic reactions, and stoichiometry of reactants, products, and cofactors. These models can incorporate metabolite concentration, reaction rates, and several other biological parameters making it more complex in computational time and costs. KM is also based on steady-state assumptions where the reaction rates are functions of concentrations of the reactants, enzymes, and cofactors present in the metabolic network of interest. A parameterized kinetic models can be used to simulate dynamics of metabolic processes, time-dependent distribution of metabolic fluxes and concentrations of metabolites under physiological conditions, and effect of specific gene knock-outs on cellular metabolism [112, 113]. As KMs take into account metabolite concentrations and enzyme kinetics, they are considered to be more precise than the static models in identifying critical variables and parameters. The main setback for the development of KMs is lack of kinetic data and stringency in parameter estimation. Sensitivity analysis should be performed to test the robustness of model parameters [113]. Recently, hybrid algorithms combining local optimization for continuous rate parameters and nonlocal optimization for discretized flux variables have been introduced to study the kinetic networks [113]. In spite of its limitations, kinetic models are precious tools for pharmacokinetics study, biomarker discovery, and drug design.

5. CONCLUSIONS

Overall, this chapter provided an overview of metabolomics data analysis strategies. As indicated in this chapter, the primary objective of untargeted metabolomics studies is to identify metabolic signatures specific to the given biological condition (e.g. disease vs healthy). Given that, metabolomics experiments are large-scale and complex measurements, the discovery of metabolite signature mainly depends on the proper data processing and analysis. In this chapter, we provided an objective description of metabolomics data preprocessing, univariate data analysis, and multivariate analysis, highlighting their strong and weak aspects. It is also very important to note, while metabolomics experiments links genotype to phenotype, cross-validation of results obtained during metabolomics experiments is vital to avoid any false discovery and bias interpretation. For metabolomics data interpretation, metabolite set analysis, pathways analysis may assist the practitioner in biological interpretation of metabolomics dataset. Advance computational strategy and knowledge-based approach such as genome-scale metabolic modelling could be integrated within metabolomics study design to understand these cellular processes better. We believe GSMMs could be an excellent tool for deconvolution of complex biological processes to generate hypothesis and for novel metabolic marker discovery in metabolomics studies.

ACKNOWLEDGEMENT

The Juvenile Diabetes Research Foundation (2-SRA-2014-159-Q-R to M.O.) and Academy of Finland (292568 to M.O.) supported this work.

REFERENCES

[1] O. Fiehn, Metabolomics—the link between genotypes and phenotypes, Plant Mol. Biol. 48 (1–2) (2002) 155–171.

[2] J.K. Nicholson, J.C. Lindon, E. Holmes, 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data, Xenobiotica 29 (11) (1999) 1181–1189.

[3] R. Goodacre, et al., Metabolomics by numbers: acquiring and understanding global metabolite data, Trends Biotechnol. 22 (5) (2004) 245–252.

[4] G.J. Patti, O. Yanes, G. Siuzdak, Innovation: metabolomics: the apogee of the omics trilogy, Nat. Rev. Mol. Cell Biol. 13 (4) (2012) 263–269.

[5] S. Lamichhane, et al., Strategy for nuclear-magnetic-resonance-based metabolomics of human feces, Anal. Chem. 87 (12) (2015) 5930–5937.

[6] S. Lamichhane, et al., Optimizing sampling strategies for NMR-based metabolomics of human feces: pooled vs. unpooled analyses, Anal. Methods 9 (30) (2017) 4476–4480.

[7] M. Orešič, et al., Cord serum lipidome in prediction of islet autoimmunity and type 1 diabetes, Diabetes 62 (9) (2013) 3268–3274.

[8] J.P. Posti, et al., Metabolomics profiling as a diagnostic tool in severe traumatic brain injury, Front. Neurol. 8 (2017) 398.

[9] M.D. Cao, et al., Metabolic characterization of triple negative breast cancer, BMC Cancer 14 (1) (2014) 941.

[10] S. Lamichhane, et al., Gut microbial activity as influenced by fiber digestion: dynamic metabolomics in an in vitro colon simulator, Metabolomics 12 (2) (2016) 25.

[11] A.B. García-García, et al., 1H HR-MAS NMR-based metabolomics analysis for dry-fermented sausage characterization, Food Chem. 240 (2018) 514–523.

[12] A.C. Dona, et al., A guide to the identification of metabolites in NMR-based metabonomics/metabolomics experiments, Comput. Struct. Biotechnol. J. 14 (2016) 135–153.

[13] Q. Bao, et al., A robust automatic phase correction method for signal dense spectra, J. Magn. Reson. 234 (2013) 82–89.

[14] L. Chen, et al., An efficient algorithm for automatic phase correction of NMR spectra based on entropy minimization, J. Magn. Reson. 158 (1–2) (2002) 164–168.

[15] W. Dietrich, C.H. Rüdel, M. Neumann, Fast and precise automatic baseline correction of one- and two-dimensional NMR spectra, J. Magn. Reson. (1969) 91 (1) (1991) 1–11.

[16] T. Skov, et al., Automated alignment of chromatographic data, J. Chemom. 20 (11 – 12) (2006) 484–497.

[17] F. Savorani, G. Tomasi, S.B. Engelsen, Icoshift: a versatile tool for the rapid alignment of 1D NMR spectra, J. Magn. Reson. 202 (2) (2010) 190–202.

[18] D.S. Wishart, Quantitative metabolomics using NMR, TrAC Trends Anal. Chem. 27 (3) (2008) 228–237.

[19] L.R. Euceda, G.F. Giskeødegård, T.F. Bathen, Preprocessing of NMR metabolomics data, Scand. J. Clin. Lab. Invest. 75 (3) (2015) 193–203.

[20] J. Hao, et al., BATMAN—an R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a Bayesian model, Bioinformatics 28 (15) (2012) 2088–2090.

[21] J.J. Ellinger, et al., Databases and software for NMR-based metabolomics, Curr. Metabolomics 1 (1) (2013) 28–40, https://doi.org/10.2174/2213235X11301010028.

[22] T. De Meyer, et al., NMR-based characterization of metabolic alterations in hypertension using an adaptive, intelligent binning algorithm, Anal. Chem. 80 (10) (2008) 3783–3790.

[23] T. Pluskal, et al., MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data, BMC Bioinf. 11 (2010) 395.

[24] H. Tsugawa, et al., MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis, Nat. Methods 12 (6) (2015) 523–526.

[25] C.A. Smith, et al., XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification, Anal. Chem. 78 (3) (2006) 779–787.

[26] Z. Li, et al., Comprehensive evaluation of untargeted metabolomics data processing software in feature detection, quantification and discriminating marker selection, Anal. Chim. Acta 1029 (2018) 50–57.

[27] S. Castillo, et al., Data analysis tool for comprehensive two-dimensional gas chromatography/time-of-flight mass spectrometry, Anal. Chem. 83 (8) (2011) 3058–3067.

[28] M. Sysi-Aho, et al., Normalization method for metabolomics data using optimal selection of multiple internal standards, BMC Bioinf. 8 (1) (2007) 93.

[29] K.T. Do, et al., Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies, bioRxiv (2018).

[30] E. Saccenti, et al., Reflections on univariate and multivariate analysis of metabolomics data, Metabolomics 10 (3) (2014) 361–374.

[31] T. Hyötyläinen, M. Orešič, Systems biology strategies to study lipidomes in health and disease, Prog. Lipid Res. 55 (2014) 43–60.

[32] M. Vinaixa, et al., A guideline to univariate statistical analysis for LC/MS-based untargeted metabolomics-derived data, Metabolites 2 (4) (2012) 775–795.

[33] J. Bartel, J. Krumsiek, F.J. Theis, Statistical methods for the analysis of high-throughput metabolomics data, Comput. Struct. Biotechnol. J. 4 (2013), e201301009.

[34] T. Suvitaival, et al., Lipidome as a predictive tool in progression to type 2 diabetes in Finnish men, Metabolism 78 (2018) 1–12.

[35] D.I. Broadhurst, D.B. Kell, Statistical strategies for avoiding false discoveries in metabolomics and related experiments, Metabolomics 2 (4) (2006) 171–196.

[36] J.D. Storey, A direct approach to false discovery rates, J. R. Stat. Soc. Ser. B Stat Methodol. 64 (3) (2002) 479–498.

[37] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. R. Stat. Soc. Ser. B Methodol. 57 (1995) 289–300.

[38] R.C. Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, 2014.

[39] J. Xia, et al., MetaboAnalyst 3.0—making metabolomics more meaningful, Nucleic Acids Res. 43 (W1) (2015) W251–W257.

[40] K. Wanichthanarak, et al., Metabox: a toolbox for metabolomic data analysis, interpretation and integrative exploration, PLoS One 12 (1) (2017), e0171046.

[41] D. La Torre, et al., Decreased cord-blood phospholipids in young age-at-onset type 1 diabetes, Diabetes 62 (11) (2013) 3951–3956.

[42] T.H. Haukaas, et al., Impact of freezing delay time on tissue samples for metabolomic studies, Front. Oncol. 6 (2016) 17.

[43] A.D. Kostic, et al., The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes, Cell Host Microbe 17 (2) (2015) 260–273.

[44] S. Lamichhane, et al., Dynamics of plasma lipidome in progression to islet autoimmunity and type 1 diabetes: type 1 diabetes prediction and prevention study (DIPP), Sci. Rep. 8 (1) (2018) 10635, https://doi.org/10.1038/s41598-018-28907-8.

[45] M. Oresic, et al., Dysregulation of lipid and amino acid metabolism precedes islet autoimmunity in children who later progress to type 1 diabetes, J. Exp. Med. 205 (13) (2008) 2975–2984.

[46] A. O'Gorman, et al., Identification of a plasma signature of psychotic disorder in children and adolescents from the Avon Longitudinal Study of Parents and Children (ALSPAC) cohort, Transl. Psychiatry 7 (2017) e1240.

[47] S. Lamichhane, et al., Metabolic fate of ^{13}C-labelled polydextrose and impact on the gut microbiome: a triple-phase study in a colon simulator, J. Proteome Res. 17 (2018) 1041–1053.

[48] P. Ebrahimi, et al., Chemometric analysis of NMR spectra, in: G.A. Webb (Ed.), Modern Magnetic Resonance, Springer International Publishing, Cham, 2017, pp. 1–20.

[49] S. Lamichhane, et al., Impact of dietary polydextrose fiber on the human gut metabolome, J. Agric. Food Chem. 62 (40) (2014) 9944–9951.

[50] R. Bro, A.K. Smilde, Principal component analysis, Anal. Methods 6 (9) (2014) 2812–2831.

[51] J.A. Westerhuis, et al., Assessment of PLSDA cross validation, Metabolomics 4 (1) (2008) 81–89.

[52] J. Lever, M. Krzywinski, N. Altman, Principal component analysis, Nat. Methods 14 (2017) 641.

[53] M. Oresic, et al., Human serum metabolites associate with severity and patient outcomes in traumatic brain injury, EBioMedicine 12 (2016) 118–126.

[54] K.J. Parsons, W.J. Cooper, R.C. Albertson, Limits of principal components analysis for producing a common trait space: implications for inferring selection, contingency, and chance in evolution, PLoS One 4 (11) (2009), e7957.

[55] L. Ståhle, S. Wold, Partial least squares analysis with cross-validation for the two-class problem: a Monte Carlo study, J. Chemom. 1 (3) (1987) 185–196.

[56] J.A. Westerhuis, et al., Multivariate paired data analysis: multilevel PLSDA versus OPLSDA, Metabolomics 6 (1) (2010) 119–128.

[57] E. Szymańska, et al., Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies, Metabolomics 8 (Suppl. 1) (2012) 3–16.

[58] B. Worley, R. Powers, Multivariate analysis in metabolomics, Curr. Metabolomics 1 (1) (2013) 92–107.

[59] M. Bylesjö, et al., OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification, J. Chemom. 20 (8–10) (2006) 341–351.

[60] E.J.J. van Velzen, et al., Multilevel data analysis of a crossover designed human nutritional intervention study, J. Proteome Res. 7 (10) (2008) 4483–4491.

[61] O.E.D. Noord, E.H. Theobald, Multilevel component analysis and multilevel PLS of chemical process data, J. Chemom. 19 (5–7) (2005) 301–307.

[62] A.K. Smilde, et al., Common and distinct components in data fusion, J. Chemom. 31 (7) (2017).

[63] A.K. Smilde, et al., ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data, Bioinformatics 21 (13) (2005) 3043–3048.

[64] T. Suvitaival, S. Rogers, S. Kaski, Stronger findings for metabolomics through Bayesian modeling of multiple peaks and compound correlations, Bioinformatics 30 (17) (2014) i461–i467.

[65] I. Huopaniemi, et al., Two-way analysis of high-dimensional collinear data, Data Min. Knowl. Disc. 19 (2) (2009) 261–276.

[66] J.J. Jansen, et al., PARAFASCA: ASCA combined with PARAFAC for the analysis of metabolic fingerprinting data, J. Chemom. 22 (2) (2008) 114–121.

[67] Y. Mei, S.B. Kim, K.-L. Tsui, Linear-mixed effects models for feature selection in high-dimensional NMR spectra, Expert Syst. Appl. 36 (3/Pt. 1) (2009) 4703–4708.

[68] B. Ernest, et al., MetabR: an R script for linear model analysis of quantitative metabolomic data, BMC. Res. Notes 5 (2012) 596.

[69] A.K. Smilde, et al., Dynamic metabolomic data analysis: a tutorial review, Metabolomics 6 (1) (2010) 3–17.

[70] J. Xia, I.V. Sinelnikov, D.S. Wishart, MetATT: a web-based metabolomics tool for analyzing time-series and two-factor datasets, Bioinformatics 27 (17) (2011) 2455–2456.

[71] N. Kessler, et al., Learning to classify organic and conventional wheat—a machine learning driven approach using the MeltDB 2.0 metabolomics analysis platform, Front. Bioeng. Biotechnol. 3 (2015) 35.

[72] Y. Date, J. Kikuchi, Application of a deep neural network to metabolomics studies and its performance in determining important variables, Anal. Chem. 90 (3) (2018) 1805–1810.

[73] F.M. Alakwaa, K. Chaudhary, L.X. Garmire, Deep learning accurately predicts estrogen receptor status in breast cancer metabolomics data, J. Proteome Res. 17 (1) (2018) 337–347.

[74] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2605.

[75] M. Farrés, et al., Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation, J. Chemom. 29 (10) (2015) 528–536.

[76] B. Xi, et al., Statistical analysis and modeling of mass spectrometry-based metabolomics data, Methods Mol. Biol. 1198 (2014) 333–353 (Clifton, N.J.).

[77] R. Weng, et al., Metabolomics approach reveals integrated metabolic network associated with serotonin deficiency, Sci. Rep. 5 (2015) 11864.

[78] T. Rajalahti, et al., Biomarker discovery in mass spectral profiles by means of selectivity ratio plot, Chemom. Intell. Lab. Syst. 95 (1) (2009) 35–48.

[79] M.P. Gomez-Carracedo, et al., Objective chemical fingerprinting of oil spills by partial least-squares discriminant analysis, Anal. Bioanal. Chem. 403 (7) (2012) 2027–2037.

[80] S. Karimi, B. Hemmateenejad, Identification of discriminatory variables in proteomics data analysis by clustering of variables, Anal. Chim. Acta 767 (2013) 35–43.

[81] E.F. Lock, et al., Joint and Individual variation explained (JIVE) for integrated analysis of multiple data types, Ann. Appl. Stat. 7 (1) (2013) 523–542.

[82] M. Schouteden, et al., SCA with rotation to distinguish common and distinctive information in linked data, Behav. Res. Methods 45 (3) (2013) 822–833.

[83] A.K. Smilde, J.A. Westerhuis, S. de Jong, A framework for sequential multiblock component methods, J. Chemom. 17 (6) (2003) 323–337.

[84] S. Lamichhane, et al., Gut metabolome meets microbiome: a methodological perspective to understand the relationship between host and microbe, Methods (2018) (in press).

[85] W. Li, et al., Identifying multi-layer gene regulatory modules from multi-dimensional genomic data, Bioinformatics 28 (19) (2012) 2458–2466.

[86] H.K. Pedersen, et al., Human gut microbes impact host serum metabolome and insulin sensitivity, Nature 535 (7612) (2016) 376–381.

[87] M. Kanehisa, et al., Data, information, knowledge and principle: back to metabolism in KEGG, Nucleic Acids Res. 42 (D1) (2013) D199–D205.

[88] M. Kanehisa, et al., KEGG as a reference resource for gene and protein annotation, Nucleic Acids Res. 44 (D1) (2015) D457–D462.

[89] M. Kotera, et al., The KEGG databases and tools facilitating omics analysis: latest developments involving human diseases and pharmaceuticals, in: Next Generation Microarray Bioinformatics, Springer, 2012, pp. 19–39.

[90] R. Caspi, et al., The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases, Nucleic Acids Res. 36 (Suppl. 1) (2007) D623–D631.

[91] S. Abubucker, et al., Metabolic reconstruction for metagenomic data and its application to the human microbiome, PLoS Comput. Biol. 8 (6) (2012), e1002358.

[92] K. Dreher, Putting the plant metabolic network pathway databases to work: going offline to gain new capabilities, in: Plant Metabolism, Springer, 2014, pp. 151–171.

[93] A. Frolkis, et al., SMPDB: the Small Molecule Pathway Database, Nucleic Acids Res. 38 (Database issue) (2010) D480–D487.

[94] T. Jewison, et al., SMPDB 2.0: big improvements to the Small Molecule Pathway Database, Nucleic Acids Res. 42 (Database issue) (2014) D478–D484.

[95] D.S. Wishart, et al., HMDB 4.0: the human metabolome database for 2018, Nucleic Acids Res. 46 (D1) (2018) D608–D617.

[96] D.S. Wishart, et al., DrugBank: a comprehensive resource for in silico drug discovery and exploration, Nucleic Acids Res. 34 (Database issue) (2006) D668–D672.

[97] S. Jupe, et al., Reactome—a curated knowledgebase of biological pathways: megakaryocytes and platelets, J. Thromb. Haemost. 10 (11) (2012) 2399–2402.

[98] A. Kramer, et al., Causal analysis approaches in Ingenuity Pathway Analysis, Bioinformatics 30 (4) (2014) 523–530.

[99] I. Thiele, et al., A community-driven global reconstruction of human metabolism, Nat. Biotechnol. 31 (5) (2013) 419–425.

[100] A. Mardinoglu, et al., Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease, Nat. Commun. 5 (2014) 3083.

[101] M. Kankainen, et al., MPEA—metabolite pathway enrichment analysis, Bioinformatics 27 (13) (2011) 1878–1879.

[102] A. Marco-Ramell, et al., Evaluation and comparison of bioinformatic tools for the enrichment analysis of metabolomics data, BMC Bioinf. 19 (1) (2018) 1.

[103] J. Xia, D.S. Wishart, MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data, Nucleic Acids Res. 38 (Web Server issue) (2010) W71–W77.

[104] M. Chagoyen, F. Pazos, MBRole: enrichment analysis of metabolomic data, Bioinformatics 27 (5) (2011) 730–731.

[105] D.K. Barupal, O. Fiehn, Chemical Similarity Enrichment Analysis (ChemRICH) as alternative to biochemical pathway mapping for metabolomic datasets, Sci. Rep. 7 (1) (2017) 14567.

[106] A. Subramanian, et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, Proc. Natl. Acad. Sci. U. S. A. 102 (43) (2005) 15545–15550.

[107] D.K. Barupal, S. Fan, O. Fiehn, Integrating bioinformatics approaches for a comprehensive interpretation of metabolomics datasets, Curr. Opin. Biotechnol. 54 (2018) 1–9.

[108] C.A. de Leeuw, et al., The statistical properties of gene-set analysis, Nat. Rev. Genet. 17 (2016) 353.

[109] R.B. Aggio, K. Ruggiero, S.G. Villas-Boas, Pathway Activity Profiling (PAPi): from the metabolite profile to the metabolic pathway activity, Bioinformatics 26 (23) (2010) 2969–2976.

[110] J. López-Ibáñez, F. Pazos, M. Chagoyen, MBROLE 2.0—functional enrichment of chemical compounds, Nucleic Acids Res. 44 (Web Server issue) (2016) W201–W204.

[111] P. Sen, E. Kemppainen, M. Orešič, Perspectives on systems modelling of human peripheral blood mononuclear cells, Front. Mol. Biosci. 4 (2017) 96.

[112] P. Sen, H.J. Vial, O. Radulescu, Mathematical modeling and omic data integration to understand dynamic adaptation of apicomplexan parasites and identify pharmaceutical targets, in: Comprehensive Analysis of Parasite Biology: From Metabolism to Drug Discovery, 7, John Wiley & Sons, 2016, p. 457.

[113] P. Sen, H.J. Vial, O. Radulescu, Kinetic modelling of phospholipid synthesis in *Plasmodium knowlesi* unravels crucial steps and relative importance of multiple pathways, BMC Syst. Biol. 7 (1) (2013) 123.

[114] J.D. Orth, I. Thiele, B.Ø. Palsson, What is flux balance analysis? Nat. Biotechnol. 28 (3) (2010) 245–248.

[115] G. Plata, et al., Reconstruction and flux-balance analysis of the Plasmodium falciparum metabolic network, Mol. Syst. Biol. 6 (1) (2010) 408.

[116] S. Fatumo, et al., Estimating novel potential drug targets of Plasmodium falciparum by analysing the metabolic network of knock-out strains in silico, Infect. Genet. Evol. 9 (3) (2009) 351–358.

[117] N.D. Price, J.L. Reed, B.Ø. Palsson, Genome-scale models of microbial cells: evaluating the consequences of constraints, Nat. Rev. Microbiol. 2 (11) (2004) 886–897.

[118] E.J. O'Brien, J.M. Monk, B.O. Palsson, Using genome-scale models to predict biological capabilities, Cell 161 (5) (2015) 971–987.

[119] A. Bordbar, et al., Constraint-based models predict metabolic and associated cellular functions, Nat. Rev. Genet. 15 (2) (2014) 107–120.

[120] P. Sen, A. Mardinogulu, J. Nielsen, Selection of complementary foods based on optimal nutritional values, Sci. Rep. 7 (1) (2017) 5413.

[121] A.S. Blazier, J.A. Papin, Integration of expression data in genome-scale metabolic network reconstructions, Front. Physiol. 3 (2012) 299.

[122] R. Agren, et al., Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT, PLoS Comput. Biol. 8 (5) (2012), e1002518.

[123] T. Cakir, et al., Integration of metabolome data with metabolic networks reveals reporter reactions, Mol. Syst. Biol. 2 (2006) 50.

[124] J. Forster, et al., Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network, Genome Res. 13 (2) (2003) 244–253.

[125] S. Magnúsdóttir, et al., Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota, Nat. Biotechnol. 35 (1) (2017) 81–89.

[126] N.C. Duarte, et al., Global reconstruction of the human metabolic network based on genomic and bibliomic data, Proc. Natl. Acad. Sci. U. S. A. 104 (6) (2007) 1777–1782.

[127] N. Swainston, et al., Recon 2.2: from reconstruction to model of human metabolism, Metabolomics 12 (7) (2016) 1–7.

[128] A. Mardinoglu, et al., Integration of clinical data with a genome-scale metabolic model of the human adipocyte, Mol. Syst. Biol. 9 (2013) 649.

[129] C. Simon, R. Daniel, Metagenomic analyses: past and future trends, Appl. Environ. Microbiol. 77 (4) (2011) 1153–1161.

[130] S. Shoaie, et al., Quantifying diet-induced metabolic changes of the human gut microbiome, Cell Metab. 22 (2) (2015) 320–331.

[131] S. Shoaie, J. Nielsen, Elucidating the interactions between the human gut microbiota and its host through metabolic modeling, Front. Genet. 5 (2014) 86.

[132] A. Bordbar, et al., A multi-tissue type genome-scale metabolic network for analysis of whole-body systems physiology, BMC Syst. Biol. 5 (1) (2011) 180.

[133] K. Yizhak, et al., Modeling cancer metabolism on a genome scale, Mol. Syst. Biol. 11 (6) (2015) 817.

[134] T. Hyötyläinen, et al., Genome-scale study reveals reduced metabolic adaptability in patients with non-alcoholic fatty liver disease, Nat. Commun. 7 (2016).

[135] L. Väremo, et al., Proteome-and transcriptome-driven reconstruction of the human myocyte metabolic network and its use for identification of markers for diabetes, Cell Rep. 14 (6) (2016) 1567.

[136] C.T. Trinh, A. Wlaschin, F. Srienc, Elementary mode analysis: a useful metabolic pathway analysis tool for characterizing cellular metabolism, Appl. Microbiol. Biotechnol. 81 (5) (2009) 813–826.

[137] S. Klamt, et al., From elementary flux modes to elementary flux vectors: metabolic pathway analysis with arbitrary linear flux constraints, PLoS Comput. Biol. 13 (4) (2017), e1005409.

[138] Z. Chen, et al., Elementary mode analysis for the rational design of efficient succinate conversion from glycerol by *Escherichia coli*, J Biomed Biotechnol 2010 (2010) 518743.

[139] M. Kumar, S. Saini, K. Gayen, Elementary mode analysis reveals that Clostridium acetobutylicum modulates its metabolic strategy under external stress, Mol. BioSyst. 10 (8) (2014) 2090–2105.

[140] A. Khodayari, et al., A kinetic model of Escherichia coli core metabolism satisfying multiple sets of mutant flux data, Metab. Eng. 25 (2014) 50–62.

[141] W.W. Adams, Glycerol Production in *Plasmodium falciparum*: Towards a Detailed Kinetic Model, Stellenbosch University, Stellenbosch, 2015.