Bioinformatics Tools for Mass Spectroscopy-Based Metabolomic Data Processing and Analysis

Masahiro Sugimoto¹,²,³,⁴, Masato Kawakami¹,⁴, Martin Robert¹,², Tomoyoshi Soga¹,⁴ and Masaru Tomita¹,⁴

¹Institute for Advanced Biosciences, Keio University, Tsuruoka, Yamagata 997-0017, Japan
²Systems Biology Program, Graduate School of Media and Governance, Keio University, Fujisawa, Kanagawa 252-8520, Japan
³Graduate School of Medicine and Faculty of Medicine Kyoto University, Yoshida-Konoe-cho, Sakyo-ku, Kyoto 606-8501, Japan
⁴Department of Environment and Information Studies, Keio University, Fujisawa, Kanagawa 252-8520, Japan

Abstract: Biological systems are increasingly being studied in a holistic manner, using omics approaches, to provide quantitative and qualitative descriptions of the diverse collection of cellular components. Among the omics approaches, metabolomics, which deals with the quantitative global profiling of small molecules or metabolites, is being used extensively to explore the dynamic response of living systems, such as organelles, cells, tissues, organs and whole organisms, under diverse physiological and pathological conditions. This technology is now used routinely in a number of applications, including basic and clinical research, agriculture, microbiology, food science, nutrition, pharmaceutical research, environmental science and the development of biofuels. Of the multiple analytical platforms available to perform such analyses, nuclear magnetic resonance and mass spectrometry have come to dominate, owing to the high resolution and large datasets that can be generated with these techniques. The large multidimensional datasets that result from such studies must be processed and analyzed to render this data meaningful. Thus, bioinformatics tools are essential for the efficient processing of huge datasets, the characterization of the detected signals, and to align multiple datasets and their features. This paper provides a state-of-the-art overview of the data processing tools available, and reviews a collection of recent reports on the topic. Data conversion, pre-processing, alignment, normalization and statistical analysis are introduced, with their advantages and disadvantages, and comparisons are made to guide the reader.

Keywords: Bioinformatics, mass spectrometry, metabolome, metabolomics, software development, statistical analysis, systems biology.

1. INTRODUCTION

Metabolomics or metabolome analysis aims to conduct the simultaneous determination and quantitative analysis of intracellular metabolites. Since metabolomics is concerned with small molecules that are the substrates and products, of cellular activity, it allows to explore in a direct and immediate way the biological system/environment interface. This can be appreciated by the great sensitivity of metabolite levels to subtle pharmacological and toxicological intervention [1-6]. As a consequence, metabolomics is playing an increasingly important role in systems biology, a field that aims to integrate information collected at multiple biological levels. It is now used widely in many applications including microbiology, diagnostic biomarker discovery, toxicological testing, food and beverage analysis, plant and animal phenotyping, and drug discovery and development [7-12].

Nuclear magnetic resonance (NMR) is one of the most commonly used analytical techniques in metabolomics studies [13]. To date, a number of large-scale, studies using NMR have been reported, including blood urine and serum metabolome profiling [14-15]. This technique has been popular in metabolomic studies because of its quantitative nature and high reproducibility. In addition, NMR spectra provide a wealth of biochemical information not available by other means [16-20]. It also has definitive advantage that it can be used in non-destructive ways to enable metabolomic profiling in vivo [21-22] and even allow metabolite imaging in biological samples [23-24]. However, the relatively low sensitivity of NMR, and the spectral overlap that often occurs, limits the number and variety of metabolites that can be simultaneously observed. Hyphenated mass spectrometry (MS) methods, such as GC-MS [25], LC-MS [26] and CE-MS [27], currently provide higher sensitivity, and are the leading analytical platform for metabolite profiling [28-31]. Because of the diverse physical and chemical properties (for example, molecular weight, polarity and solubility) of the metabolites contained in typical samples, no single analytical methodology can profile datasets comprehensively. Thus, metabolomics, in the strictest sense, is very challenging, and the term is used broadly to cover approaches concerned with investigating subsets of the metabolome [32]. GC-MS, LC-MS and CE-MS are generally capable of profiling volatile, singly or multiply charged metabolites. Hyphenated
The analytical usually flow starts from data conversion, detecting signal peaks, normalization and comparison of multiple datasets to generate a data matrix that includes the detected peaks of all given samples (alignment). The differentiation of signals from noise by interpretation of the mass spectrum and the identification of detected features using, for example, alignments with standard compound data, are also important. Finally, processed data are analyzed using statistical methods and data mining. A recent addition to this straightforward analytical process is the quality control (QC) of data processing. This process does not simply involve the use of QC methods after data processing [58], but rather is used as part of an iterative feedback loop between data processing and QC [59] (Fig. 1).

The following section introduces recent literature related to 1) data conversion, 2) feature detection, 3) alignment, 4) scaling/normalization, 5) identification, and 6) QC. See also the following references: [58, 60-61].

2.1. Data Conversion

Data processing starts with file format conversion from the MS-vendor dependent binary format to more common formats, to allow subsequent processing to be carried out on independent operation systems and software. A common and open framework and data description is important if data are to be shared among laboratories [62-64]. NetCDF and mzXML are the most commonly used file formats to store hyphenated-MS data [65]. Owing to recent rapid improvements in the throughput and resolution of MS, individual data files have become large, which compounds problems associated with the large numbers of datasets handled in metabolomics projects. Although these common file formats simplify data sharing between laboratories, the problem of handling a large number of large datasets remains. While removing small intensity peaks and data compression using irreversible filtering, as can be implemented in mzMine [56] and mzMine2 [52], is the simplest way to diminish data size, they risk distorting subsequent data analysis. Although Mass++ allows the direct import of various binary files provided by MS vendors into standard software [66], it merely accesses the binary data through a vendor-provided application programmable interface (API). This dramatically reduces throughput and does not solve the problem of MS-vendor software dependency. Although it cannot be shown directly without access to the source code, most vendor-provided hyphenated-MS instrument binary formats (for example, .wiff files and .D formats provided by Applied Biosystems and Agilent Technologies Inc.) can be estimated to contain a series of mass spectra data, since mass spectra are usually collected in this way. This data structure results in much longer data access times to output a chromatograph or an electropherogram if the data points included in the mass spectra are not unique over the chromatograph or electropherogram. To solve these size and structure problems, we developed a compact binary file format that facilitates rapid access to chromatographs or electropherograms and mass spectra [67]. Although there is currently a trade-off between facilitating quick data access and the availability of a generic file format, the development and standardization of file formats that fulfill the requirement for rapid access should be a priority.
2.2. Feature Detection

In the typical analytical flow, three-dimensional data incorporating retention or migration times, \( m/z \) and intensity data are first converted to piles of two-dimensional chromatography/electropherograms, by integrating data points within a specific range along the \( m/z \) axis (ion extraction or data binning). Second, background reduction (or baseline removal) and smoothing of the data are conducted to reduce false positive detection. Third, local maxima are found as peak top candidates, or a mathematical model is fitted to find peak-like shapes within the chromatographs or electropherograms. These are used to identify peaks over a user-specified threshold, which may be in the form of a peak height, peak area or signal-to-noise (S/N) ratio [52-54, 67-68]. Although wavelet transformation and Gaussian-curve fitting (or matched filter) is a commonly used means to distinguish signal from noise [53, 68], fully automatic processing remains difficult owing to the complex peak shapes often observed in LC-MS and, in particular, CE-MS. Interactive tuning of the algorithms is therefore often required [52, 67]. Other options are to identify peaks at matched locations (\( m/z \) and time), even under the initially-defined threshold after the alignment process [69]. Such feedback procedures and QC will be discussed further in section 2.7.

2.3. Alignment of Multiple Data Sets

The alignment of multiple datasets, i.e. the elimination of retention or migration times shifts between datasets, is a central topic of data processing in the metabolomics field, and is associated with specific technical difficulties. Therefore, many alignment techniques have been developed [70]. The retention time variance of GC-MS and LC-MS datasets is non-linear [71], and thus multiple sophisticated time correction methods have been developed. The alignment of CE-MS data is especially difficult because of the low reproducibility of migration times [54], and robust and versatile alignment procedures are therefore required. Here, we review the three major alignment algorithms used for the temporal dimension. In addition, the normalization of mass/charge ratio (\( m/z \)) calculated by MS is also introduced.

Time Correlation Optimized Warping

Time correlation optimized warping (COW) divides chromatograms into small segments and shifts individual segments to maximize the correlation coefficient between a reference and test chromatograph. The algorithm itself has inherent problems; a larger number of segments leads to greater accuracy, but raises the risk of dividing the targeted metabolite peaks. To optimize the degree of segmentation, the use of heuristic and global optimization processes, such as genetic algorithms, has been proposed [72]. To date, benchmark tests with only small numbers of peaks have been performed [72], and the method should be evaluated using data with a large number of peaks, observed by high resolution MS.

Parametric Time Warping

The parametric time warping method aligns a given chromatogram with a reference chromatogram using second degree polynomial functions, called warping functions [73]. Coefficients in warping functions are optimized to minimize the time difference between selected matched peaks in reference and aligned chromatograms. Thus, the method relies on the presence of a number of known matched peaks among the samples to be aligned. Although the addition of internal standards (IS) is the most simple way to achieve this, it has several disadvantages: (i) suitable IS compounds must be carefully selected, for example the IS compounds must not normally be present in the samples; (ii) additional sample preparation is required; and most significantly (iii) the added IS may cause ion suppression effects and degrade the quantitative reliability of the observed profiles. Despite these problems, rapid computation time is an important advantage of this method. Lower flexibility and accuracy has been reported for this method in comparison with COW and dynamic time warping (DTW) [70].

Dynamic Time Warping

DTW finds the matched peaks among multiple datasets automatically to produce warping functions. Dynamic programming (DP) has historically been used in homology searching of genes or genomes, and has been used for matching peaks [74]. The parameters that characterize DP results, such as gap penalty, make this method parametric. Thus, empirical iterative multi-step optimization of these parameters has been used in CE-MS data processing software [54] and interactive graphical user interfacing [67]. In contrast, recent modifications to DTW using multiple chromatograms with different \( m/z \), instead of one-dimensional information available from total ion chromatography, reduced the impact of the parametric problems embedded in the original DTW algorithm [75].

Calibration of Mass Values (\( m/z \) Alignment)

Exact masses (mass-to-charge ratio (\( m/z \)) values), produced by detectors in time-of-flight (TOF)-MS instruments are usually calculated based on online calibration with one or more reference substances that are co-injected with the sample. This is known as the mass lock system [76]. The \( m/z \) values detected for individual peaks fluctuate depending on several factors, including temperature, the abundance of ions simultaneously entering the MS, and the processing ability, type and specifications of the MS detector [77]. Thus, the data acquired should be further calibrated. Typically a calibration curve generated using the peaks of known \( m/z \) is applied to correct \( m/z \) values of other peaks of interest (offline or software calibration) [78-80]. The \( m/z \) values are intrinsically calibrated for the whole chromatograph or electropherogram time axis, since the factors influencing \( m/z \) shifts can change even during the course of a single run [81]. In addition, \( m/z \) value correction can be carried out using peak intensities relative to the intensities of internal standards [82], using the location of background noise observed throughout the measurement [83], and using statistical approaches with multiple datasets [69]. Ideally, these methods should be integrated to optimize \( m/z \) normalization.

2.4. Scaling and Normalization

The elimination of unwanted systematic bias, while maintaining genuine biological differences in the observed datasets, is essential for subsequent analyses to identify significant metabolites. The systematic bias derived from
variation in sample concentration, especially when handling biofluids such as urine, blood and saliva samples, must be removed. Deviation in signal intensities due to measurement errors, for example poor MS sensitivity, must also be removed. To address the former problem, metabolomic analyses typically use endogenous metabolites, for example creatinine, to normalize overall urine metabolite concentrations [84]. However, this method is not always sufficient to eliminate systematic bias, and a recent mouse metabolomic study revealed a correlation between overall urinary metabolites and several physical parameters, such as age and weight [85]. The latter bias is generally removed using two approaches. Despite the increased technical complexity of sample preparation, the use of internal standard compounds added to the sample before or after extraction is the most common approach. The use of multiple internal standards to normalize closely eluting peaks with similar m/z values has also been reported [86]. Otherwise, normalization methods based on several statistical models (unit norm [87] median [88] and quantile [58]), scaling methods (auto scaling, range scaling, Pareto scaling, vast scaling and level scaling) [61], and data transformation (log and power) have been widely used. These methods are, however, inferior to the internal standard-based methods [58].

2.5. Identification of Metabolites

Global metabolic profiles or fingerprints that do not necessarily assign observed features to particular metabolites can be very powerful means of classifying and directly comparing samples. They highlight metabolomics as providing a global molecular signature allowing us to discriminate groups of samples in contrast to more conventional comparisons based on single metabolite. However, metabolite identification from spectral data remains indispensable for providing mechanistic insight into specific cellular or disease processes and in quality control/assurance industry, for example. The accurate identification of a compound usually requires the ability to match candidate spectra with standard compounds run under the same conditions. Matching to either externally or internally applied standards has been commonly used, the latter making use of isotopically labeled standards or samples. However, the lack of readily available standard compounds remains a major obstacle to confirming the identity of observed compounds. The purification of compounds from complex samples allows access to standards; however, this can be an expensive and time-consuming process. Several tools that estimate compound composition using isotope distribution or fragmentation patterns in the mass spectrum have been developed [89-92]. Databases that include a large number candidate compounds are also indispensable (see review [43]). A theoretical study estimated that the mass spectral information available from mass spectrometers with accuracy approaching 1 ppm, such as TOFMS, is not sufficient to identify peaks without a matched standard compound, as multiple candidate compounds are often retrieved from the large public databases [93]. The Human Metabolome Project has already identified more than 4,000 putative endogenous metabolites from human serum using GC-MS, LC-MS and NMR profiles with computer-aided literature mining [12]. Many studies thus use tandem MS, which generates more informative spectra including many fragment peaks, for compound identification [94-95]. Efforts have also been made to use retention time information to reduce the number of possible candidates. These efforts are based on reverse engineering techniques [96-99] or theoretical simulation [100], which predict the retention/migration times from the metabolite structure. The quantification of observed peaks in the absence of matched standard compounds is also difficult, but computational prediction techniques have been developed [101]. The combined use of such computational methods can greatly reduce the number of candidates and aid metabolite identification.

2.6. Quality Control of Data Processing

A number of algorithms have been developed for data processing, especially for peak detection and alignment, and various parameters can be used to characterize the quality of data processing [59]. The selection of the best algorithm, and the best parameters, to analyze the datasets obtained is not an easy task. Thus, QC evaluation based on various benchmark tests is important to understand the features of each algorithm and their parameters [102].

A comparison of peak detection algorithms of LC-MS data using centWave [68], matched filter implemented in XCMS [53] and MZmine [56] showed that there was only a partial overlap in the results obtained with these methods, and a number of peaks were only detected by one software (not overlapped) [68]. Even with the same algorithm, the use of different parameters strongly affected peak detection performance [58]. Evaluation of the alignment of LC-MS data using six freely available software packages, including XCMS [53], MZmine [56], msInspect [103] and OpenMS [55], concluded that no single software perfectly aligned the datasets [104]. The annotation of metabolite identities using fixed confidence thresholds has been recommended for data reporting, as has quantitative assessment of the annotation quality using the false discovery rate (FDR) [105]. Another approach is to provide a sophisticated graphical interface that enables specific steps of data processing to be rerun using different parameters [52]. Scripting tools may also be used to accelerate the optimization process and to minimize the need for user interactivity. Another possible means to improve performance entails the development of an iterative analytical framework with machine learning methods that allow the program to be trained to tune parameters using the difference between automated and manual data processing [59]. It is evident that subsequent statistical analysis will benefit if care is taken at the processing stage, and that automatic data processing for peak detection, alignment and annotation remain far from perfect.

3. DATA ANALYSIS IN METABOLOMICS

Once a data matrix has been produced from raw data, subsequent steps usually involve different forms of statistical analysis and data mining to allow the identification of samples or variables (metabolites) that capture the bulk of variation between datasets and that may represent candidates for biologically meaningful variables. Typical analyses of metabolomic data consist of two phases; initially an overview of the given datasets is generated using
Hierarchical clustering (HCL) (Fig. 2A) aligns datasets by generating dendrograms using the following procedure: 1) calculate the similarity of the two samples using a specific metric, such as Pearson correlation, Euclidean, mutual information and covariance values; 2) align the most similar samples as neighbors or pair them as a single cluster; and 3) reiterate step 1 and 2 until all samples are aligned [33]. Non-hierarchical clustering (non-HCL) also divides data into clusters but without any hierarchical organization. The K-means and fuzzy c-means methods are typical examples of non-HCL [33]. In the K-means method, k data points are initially randomly chosen to be close to the mean of each cluster, a new mean is then calculated for each cluster and the patterns are reassigned to the new means. This process is repeated until the cluster means are such that no pattern moves from one cluster to another [109]. The K-means method assigns each datapoint into only one cluster while the fuzzy c-means method allows data to be assigned to multiple clusters [110]. Fuzzy c-means also calculates the probability of a datapoint belonging to each cluster [111]. These analyses are widely used when the number of clusters for the samples is unknown, and can be used for one-time snapshot profiling as well as time-course data.

3.3. Partial Least Squares Analysis

Partial least squares (PLS) (Fig. 2C), a regression-based method, builds a low-dimensional sub-space based on linear combinations of the original X variables. It makes use of additional Y information by adjusting the model to capture the (Y)-related variation into the original X variables [37]. PLS is particularly useful when fewer observations (samples) are available than measured variables (metabolites). In metabolomics, PLS-based classification and PLS-discriminant analysis (PLS-DA) have been widely used to sharpen the separation between groups or observations. This is achieved by rotating PCs to maximize the separation between known classes, and to elucidate the variables that carry the class separating information [33,112-113]. Similarly to loading plots in PCA, S-plots visualize both the covariance and the correlation between metabolites and the modeled class designation. The S-plot therefore helps to identify statistically significant and potentially biochemically significant metabolites, based both on contributions to the model and their reliability [114]. Despite its powerful ability to separate classes, care must be taken during fitting of PLS-DA to the training detaining datasets, which exaggerate...
generalization ability. Usually cross-validation or permutation tests are used to assess the ability of the trained PLS-DA model [115]. Orthogonal projections to latent structures (OPLS)-DA, an extension of PLS-DA featuring an integrated orthogonal signal correction filter to remove variability not relevant to class separation, has been used increasingly owing to its robustness against noise [21,116].

3.4. Random Forests

Random forests (RF) is a relatively new machine learning method typically used to discriminate two groups (Fig. 2D). The fundamental concept of RF is to allow data structures to be understood without dimensional reduction, and this method is therefore different from conventional methods such as PCA and PLS-DA. This classification algorithm was developed by Leo Breiman [117] and uses an ensemble of classification trees. Each of the classification trees is built using a bootstrap sample of the data, and at each split, the candidate set of variables is a random subset of the variables. Thus, RF uses both bagging (bootstrap aggregation), a successful approach for combining unstable learners, and random variable selection for tree building. Each tree is unpruned (grown fully) so as to obtain low-bias trees. At the same time, bagging and random variable selection result in low correlation of the individual trees [118]. The algorithm yields an ensemble that can achieve both low bias and low variance (by averaging over a large ensemble of low-bias, high-variance but low-correlation trees) [119].

3.5. Conventional Statistical Analysis

Because metabolomics generates data on multiple (dozens or hundreds) different metabolites, global overview methods that take into account the possible correlations between variables are the main tools used. However, when used appropriately, monovariate methods can also provide useful insight and remain widely used, especially for secondary biomarker analyses.

Although multivariate classification methods are often used to identify biomarkers, the discrimination of individual metabolites is usually assessed by conventional univariate statistical tests, such as Student’s t-test and the Mann-Whitney test for two classes, or ANOVA and Kruskal-Wallis for multiple classes (≥3). Dependency or correlations between metabolites, inadequate sample size, and large FDR due to multiple hypothesis testing must be taken into account when applying these methods [120]. Corrections of the p-value and/or calculation of false discovery rates must be carried out to limit the number of false positives that increase linearly with the number of variables [120]. Multivariate analysis has the advantage of considering the general patterns in the whole dataset, but it introduces additional challenges and sources of variability owing to the necessary data pre-treatment and scaling used to analyze all variables at once [61]. Thus, biomarkers should be rigorously evaluated by a combination of these statistical analyses and several validation methods, such as cross-validation and bootstrap analysis [121]. Recently, the FDR and receiver operating characteristics (ROC) methods have been frequently used to identify significantly different metabolites in the given classes.

The FDR method [122], is commonly used in gene expression analyses, and is now also used in metabolomic studies, [11], where a large number of variables are analyzed simultaneously, and thus multiple comparisons are conducted. In practice, FDR establishes a threshold for the significance level (q-value) that can be expected to represent false positives among all significant hypotheses to reject optimistic significance. To account for multiple comparisons, each FDR is estimated by the product of the significance level (Type I error rate) and the number of null hypotheses tested, divided by the number of null hypotheses rejected [123].

A ROC [124] curve is a statistical representation that simultaneously expresses both sensitivity and specificity to separate binary class datasets, for example to discriminate healthy control and patient datasets. The curve is plotted by fractions of sensitivity as the Y-axis vs. fractions of false positive rate (1- specificity) as the X-axis (Fig. 2E)). The test is used to differentiate performance of one or a combination of biomarkers; an area under the curve (AUC) of 1.0 indicates perfect separation without any false negatives or false positives, while an AUC of 0.5 is equivalent to random separation only.

AUC evaluates only the rank of the metabolites associated with the given classes, and therefore it does not count fold-change or the concentration itself. Meanwhile, FDR evaluates the relative significance of the metabolites in a large group of metabolites. Thus, the use of a combination of different methods, along with multivariate analyses, can achieve more efficient screening than any single method.

3.6. Data Mining Analysis

In addition to classification methods, other data mining methods have also been used in metabolomic data analyses to discriminate two classes, for example support vector machine (SVM) [125-126], artificial neural networks (ANN) [127] and decision tree [128]. ANN has been particularly widely used for various applications in MS-based studies, including in metabolite identification [97], classification [129], optimization of separation parameters [130] and QC of data processing [59] (see review [131]). In comparative study, a class of LC/MS peaks was predicted by four data mining techniques, k-NN, SVM, PLS-DA and Naïve Bayes, and revealed that the former two methods performed better than the latter two [132]. However, it is usually difficult to select the best method for the analysis of a given metabolomic dataset a priori, and the development of a pipeline with multiple analytical tools is therefore necessary. Visualization of metabolomic data in a pathway form also requires several data mining techniques. Small relevance and conditioned metabolic pathways have been predicted and then merged to generate pruned networks [133]. Small sub-pathways were estimated with only relevant nodes, for example metabolite and enzymes, to reduce complexity and to enhance interpretability [134]. Both of these method attempts to find new relevant connections, rather than to assign the observed data to known maps.

4. VISUALIZATION AND DATA SHARING

Here we discuss data visualization to facilitate the interpretation of large metabolomic profiles. Data
standardization is also discussed to realize open and shared access to metabolomics technologies.

4.1. Visualization of Metabolomics Data

Data visualization using a heatmap or a pathway form facilitates comprehension of the metabolomic change/response to the experimental setting. MetaboAnalyst visualizes experimental metabolomic data using heatmap visualization and offers common statistical analyses, such as PCA, PLS-DA, and HCL [135-136]. Pathway Project [137] visualizes data in the form of several graph types, such as bar graphs, time-courses and simple circles corresponding to metabolite concentration at the metabolite node on the KEGG pathway [138]. Similar web-based network visualization tools for BioCyc [139] are also available [140]. Both tools take advantage of Google Map API zoom and search functions, which can be helpful when looking for interesting details in large metabolomic datasets. The editable pathway tool is also useful when new molecular interactions that are not available in public database are to be explored [141].

4.2. Standardization of Metabolomics Data Reporting

In addition to the standardization of raw file format and data processing tools, the standardization of the reporting of metabolomic data information has also received attention. This would facilitate experimental replication, interrogation and comparison over multiple investigators and laboratories. The Metabolomics Society has formed five working groups, biological context metadata, chemical analysis, data processing, ontology and data exchange, to establish guidelines for reporting standards [142]. The Chemical Analysis Working Group, part of the Metabolomics Standards Initiative, proposed a set of minimum information that should be provided when reporting chemical analyses, and these included metadata from MS and NMR data, sample processing protocol, data processing, metabolite identification, and even unknown metabolites in the obtained dataset [143]. Attempts to define standards for data reporting have been made but unfortunately are still not widely used [35, 142-143]. To maximize the value of metabolomic datasets, it is important that data is made publicly available in formats, and with metadata, that are widely accepted as standard. In this sense, the field of metabolomics lags behind genomics and proteomics. Some of the reasons for this slow adoption of standards include the heterogeneity of analytical platforms and vendors, and the complexity of sample processing, which remains the focus of ongoing investigation. A Metabolomics Standard Initiative was recently initiated by the Metabolomics Society, and aims to develop standards for data exchange, ontology and guidelines for data reporting to solve some of the current issues (http://msi-workgroups.sourceforge.net/).

5. SOFTWARE TOOLS

A number of free software packages are already available for the processing and analysis of metabolomic data, and Table 1 gives a sample directory of these. Both web services and desktop applications are available. The table is not necessarily exhaustive, but should help to identify commonly used solutions. Several statistical tools listed were designed for NMR data analyses but might be also useful for MS data analyses. Here, we focused only on tools used specifically in metabolomics studies, and did not review free or commercial generic software for multivariate analysis or other standard statistical analysis. We emphasize mainly tools for preprocessing and data visualization. Moreover, details of these packages are not reviewed here, and the reader is referred to the original publication or project web site for more information.

6. APPLICATIONS

Here, the use of statistical methods in several applications is discussed. Note that several of the statistical analysis applications introduced here used NMR data. The same multivariable techniques can technically be used for MS data analyses, but it should be noted that MS-data includes a larger number of variables (metabolites) and therefore more redundant variables. However, appropriate statistical analyses and MS data may provide more powerful insight into biological context.

PCA and PLS-DA have been the most popular and widely used analyses in metabolomic studies. Although PCA can visualize the similarities and differences in the observed data with unknown classes, it is generally used as a weaker classification tool for class known problems. It is therefore generally used as a first screening method for classification problems, prior to PLS-DA. For example, while PCA was able to give adequate separation resolution of various conditions, for example smokers and non-smokers in a salivary metabolite profile, PLS-DA was subsequently used to maximize resolution [144]. A similar approach was adopted for the discrimination of lung cancer sufferers using urine metabolomic profiles [116] and pancreatic cancer using serum metabolomic profiles [145]. HCL has also been used to assess data structure by aligning datasets based on their profile's similarities [146-148], and this method is often used to classify samples with known classes, similarly to PCA. It has been applied to biomarker discovery, to classify control and patient groups, with key branches in its dendrogram indicating biomarker candidates [149]. Although this example was not a metabolomics application, a particularly successful example of HCL involved the clustering of gene expression in breast cancer, which suggested the existence of a new subtype of breast cancer in addition to the known classes [150]. The assessment of the analytical results of these methods can only be performed with known classes, and new findings should be analyzed further once consistency between results and known classes has been confirmed.

The over-fitting of a developed model to a given dataset should be carefully avoided, especially when using MS data, since it usually involves a large number of variables and small sample numbers. RF is expected to be a useful classification method when we use such datasets. Because the algorithm itself does not limit the application, RF has been used for biomarker discovery in urine metabolomic profiles from breast cancer patients [125] and in plant applications to explore genotype-dependent variables in metabolomic profiles in Arabidopsis and potato [151-152]. When RF and margin-based classifiers, such as SVM and PLS-DA, were compared, RF and SVM were found to have similar accuracy and both were slightly better than PLS-DA.
Table 1. Software List for Metabolomic Analysis

| Name               | Main Application                  | Specific Features                                                                 | Ref.    | License                        | User Interface |
|--------------------|-----------------------------------|-----------------------------------------------------------------------------------|---------|--------------------------------|----------------|
| OpenMS             | Raw data processing               | C++ libraries for MS data processing, including feature detection and protein/peptide identification | [57]    | Lesser GNU Public License (LGPL) | C++ library    |
| CDK-Taverna        | Workflow                          | A workflow based data processing library for cheminformatics                       | [156]   | LGPL                            | Plug-in of Java |
| Metabonomic Package | Statistical analysis of NMR data | Multivariate analysis, such as PCA, PLS, k-nearest neighbor classification, neural networks. | [127]   | GPL                            | R language<sup>2</sup> |
| metaXCMS           | Importing XCMS output             | Post processing of XCMS for comparison of multiple (≥3) classes and visualizing statistical analyses. | [157]   | Free                           | R language<sup>3</sup> and GTK |
| XCMS               | Processing LC-MS raw data         | R module for data processing, including feature detection and peak alignment        | [53]    | Free                           | R language<sup>3</sup> |
| XCMS2              | Importing tandem mass spectrometry (MS/MS) raw data | Processing of tandem mass spectrometry data for metabolite identification and structural characterization | [158]   | Free                           | Plug-in of R language<sup>3</sup> |
| MetDDL             | Data processing of LC-MS and GC/MS data | A Matlab script for data processing and visualizing multiple datasets.            | [159]   | Free                           | Matlab script  |
| MetaScape          | Pathway visualization / statistical analysis | A Cytoscape plug-in for visualizing and interpreting metabolomic data in the context of human metabolic networks | [160]   | Free                           | Plug-in of Cytoscape |
| MetaboliteDetector | Importing NetCDF and FastFlight GC-MS data | Comprehensive analysis, including chromatogram compression, feature detection, alignment and compound identification. | [161]   | GNU public license (GPL)       | Local application (GUI) |
| MetAlign           | Importing many common formats, including Masslynx, Xcalibur, netCDF, and the old-style HP/Agilent format of GC-MS / LC-MS data | Interface-driven data processing program. Includes baseline correction, smoothing, feature detection and alignment | [162]   | Free                           | Local application (GUI) |
| MAVEN              | Data processing of LC-MS and pathway visualization | Tools for all aspects of data analysis, from feature extraction to pathway-based graphical data display | [59]    | Free                           | Local application (GUI) |
| LIMSA              | Data processing / mass spectrometric lipidome data | Tool finds and integrates peaks in a mass spectrum and matches the peaks with a user-supplied list of expected lipids. | [163]   | Free                           | Local application (GUI) |
| centWave           | Data processing of LC-MS data     | Detection of close and partially overlapping features; also has the highest overall recall | [68]    | Free                           | Local application (GUI) |
| mzMine2            | Data processing of MS data        | Modular framework for processing, visualizing and analyzing mass spectrometry-based molecular profile data | [52]    | Free                           | Local application (GUI) |
| JDAMP              | Data processing of CE-MS data     | Data processing, alignment, differential display                                   | [67]    | Free for academic users         | Local application (GUI) |
| CytoScape          | Pathway visualization / statistical analysis | Software for the visualisation and analysis of biological networks                  | [164]   | Free                           | Local application (GUI) |
| metaP-server       | Statistical analysis, database searching, pathway visualization | A web-based metabolomics data analysis tool                                         | [165]   | Free                           | Web            |
| MetDAT             | Statistical analysis, database searching, pathway visualization | A modular and workflow-based free online pipeline for mass spectrometry data processing, analysis and interpretation | [166]   | Free                           | Web            |
| ChromaA            | Alignment, chromatography-mass spectrometry | Signal-based retention time alignment for chromatography-mass spectrometry data | [167]   | Free                           | Web            |
| MZedDB             | Data processing                    | Interactive m/z annotation tool                                                    | [92]    | Free                           | Web            |
| Pathway projector  | Pathway visualization              | A Web-based zoomable pathway browser that uses KEGG atlas and Google Maps API     | [137]   | Free                           | Web            |
| MetPA              | Pathway visualization / statistical analysis | A web-based metabolomics tool for pathway analysis and visualization               | [168]   | Free                           | Web            |
| MetExplore         | Pathway visualization              | A web server to link metabolomic experiments and genome-scale metabolic networks   | [169]   | Free                           | Web            |
and unfortunately the profusion of new and interesting tools, greater interchangeability and interoperability between tools, are likely to produce further progress. There is a need for metabolomics is still rapidly evolving, and ongoing efforts methods for the user. The field of data analysis for other things. This review gives an overview of the options that can be chosen from, and highlights recent efforts to depend on the analytical platform, the experimental design a data analysis solution is not straightforward and will specialized for certain steps or visualization. The selection of all steps from raw data to statistical analysis, while others are processing, some of which are capable of performing most or possible when only a few metabolite sets are used, rather than all available data, which is used in PCA and PLS-DA. Approaches using all available metabolites are appropriate when studying overall variation, but are not useful for clinical usage, for example in the development of diagnosis techniques using a single or a few markers. Thus, integrative analyses using multivariate analysis, feature selection, and assessment of individual or a few markers are standard techniques that are useful for general purposes.

As should be apparent, multiple solutions exist for data processing, some of which are capable of performing most or all steps from raw data to statistical analysis, while others are specialized for certain steps or visualization. The selection of a data analysis solution is not straightforward and will depend on the analytical platform, the experimental design and data type, and on computational infrastructure, among other things. This review gives an overview of the options that can be chosen from, and highlights recent efforts to integrate these solutions to generate simple, yet powerful methods for the user. The field of data analysis for metabolomics is still rapidly evolving, and ongoing efforts are likely to produce further progress. There is a need for greater interchangeability and interoperability between tools, and unfortunately the profusion of new and interesting tools originating from numerous small groups often tends to limit this goal. Developers should consider these factors when promoting particular solutions. This will stimulate data sharing and exchange, and therefore improve adoption by a community of users who are often overwhelmed by a range of possibilities, and who may therefore tend to stick to tools that emphasize usability rather than quality or performance.

In this review article, we reviewed multiple tools for processing and analysis of MS data. Multiple metabolomics platforms together with the appropriate data processing and analysis tools can allow us to identify discriminating features in a set of samples, with multiple applications in research, diagnosis, etc. However, beyond class discrimination, understanding the biological mechanisms responsible for the variance in observed profiles remains an important issue. For this, the constant development and improvement of computational techniques for metabolite identification, accurate quantification, data integration, and pathway visualization is important and will continue to be the focus of bioinformatics efforts in the coming years.

**CONCLUSION**

Remarkable improvements in analytical instruments, including MS, have enabled the profiling of metabolites with increasingly high throughput and high precision. Bioinformatics, which facilitates the interpretation of the output of these instruments, is essential to the successful analysis of large dataset metabolomic applications. Tool development must keep up with the improvements in analytical instruments and thus represents an important challenge, but has great potential to add value to metabolomic datasets.

**CONFLICT OF INTEREST**

Declared none.

**ACKNOWLEDGEMENTS**

This work was supported by research funds from the Yamagata Prefectural Government and the City of Tsuruoka. We thank Akira Okawa at the Riken Plant Science Center and Fumio Matsuta at Kobe University for fruitful discussions.

**REFERENCES**

[1] Lindon JC, Holmes E, Nicholson JK. Metabonomics in pharmaceutical R&D. FEBS J 2007; 274: 1140-51.

[2] Swann J, Wang Y, Abecia L, et al. Gut microbiome modulates the toxicity of hydrazine: a metabonomic study. Mol Biosyst 2009; 5:

| Table 1. cont… |  |
|-----------------|-----------------|
| **MSEA** | Pathway visualization | A web-based tool to identify biologically meaningful patterns in quantitative metabolomic data | [170] Free | Web |
| **MetabolomeXpress** | Pipeline for data processing and statistical analysis of GC/MS data | Data processing, statistical analysis (e.g. HCL), metabolite identification and heat map visualization | [171] Free access for non-commercial and academic users | Web |
| **Chromaligner** | Alignment of LC-MS data | Alignment of LC-MS chromatographs using the COW algorithm | [172] Free access | Web |

*) R language (http://www.r-project.org/).
Quantitative metabolite analysis using capillary electrophoresis mass spectrometry. J Proteome Res 2003; 2: 488-94.

Aharoni A, Ric de Vos CH, Verhoeven HA, et al. Nontargeted metabolome analysis by use of Fourier Transform Ion Cyclotron Mass Spectrometry. OMICS 2002; 6: 217-34.

Castillo JJ, Hayes A, Mohammed S, Gaskell SJ, Oliver SG. An optimized protocol for metabolome analysis in yeast using direct infusion electrospray mass spectrometry. Phytochemistry 2003; 62: 929-37.

Theodoridis G, Gika HG, Wilson ID. Mass-spectrometry-based holistic analytical approaches for metabolite profiling in systems biology studies. Mass Spectrom Rev 2011; 30: 884-906.

Bajaj S, Shulvay VC. LC-MS-based metabolomics. Methods Mol Biol 2011; 708: 213-28.

Morton MR, Soga T. Metabolome analysis by capillary electrophoresis-mass spectrometry. J Chromatogr A 2007; 1168: 237-46.

Blekerman G, Laubenbacher R, Cortes DF, et al. Bioinformatics tools for cancer metabolomics. Metabonomics 2011; 7: 329-343.

Katagajama M, Oresic M. Data processing for mass spectrometry-based metabolomics. J Chromatogr A 2007; 1158: 318-28.

Arita M. What can metabolomics learn from genomics and proteomics? Curr Opin Biotechnol 2009; 20: 610-5.

Holmes C, McDonald F, Jones M, Ozdemir V, Graham JE, et al. Metabolomics: a genome-wide association study of metabolite profiles in human serum. PLoS Genet 2008; 4: e1000282.

Shihi N, Nakahigashi K, Baba T, et al. Multiple high-throughput analyses monitor the response of E. coli to perturbations. Science 2007; 316: 593-7.

Fiehn O, Kopka J, Dorrman P, Altmann T, Trethewey RN, Willmitzer L. Metabolite profiling for plant functional genomics. Nat Biotechnol 2000; 18: 1157-61.

Hirayama A, Kami K, Sugimoto M, et al. Quantitative metabolome profiling of colon and stomach cancer microenvironment by capillary electrophoresis time-of-flight mass spectrometry. Cancer Res 2009; 69: 4198-215.

Mukharji S, Poisson LM, Rajendiran TM, et al. Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. Nature 2009; 457: 910-4.

Katajamaa M, Orešič M. Informatics and computational strategies for the study of metabolic networks. Mol Biosyst 2008; 4: 121-7.

Bowen BP, Northen TR. Dealing with the unknown: metabolomics and gender differences. Analyst 2003; 128: 819-23.

50. Li Y, Djurić-Ilić PM, Maletić-Savatić M. Systems biology approach to imaging of neural stem cells. Methods Mol Biol 2011; 711: 421-34.

51. Soga T, Ohashi Y, Ueno Y, Naraoaka H, Tomita M, Nishioka T. Metabolomics in human prostate cancer detection. Sci Transl Med 2010; 2: 16ra8.

52. Pluskal T, Castillo S, Villar-Broines A, Oresic M. mzmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinformatics 2010; 11: 395.
Sturm M, Bertsch A, Gröpl C,
Field D, Sansone SA, Collis A,
Melamud E, Vastag L, Rabinowitz JD. Metabolomic analysis and
Sugimoto M, Hirayama A, Ishikawa T,
Tanaka S, Aoshima K, Miura Y, YO. 57th ASMS Conference on
Current Bioinformatics, 2012
[53] Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. XCMS:
processing mass spectrometry data for metabolite profiling using
nonlinear peak alignment, matching, and identification. Anal Chem 2006; 78: 779-87.
[54] Baran R, Kochi H, Saito N, et al. MathDAMP: a package for
differential analysis of metabolite profiles. BMC Bioinformatics 2006; 7: 530.
Bertsch A, Gröpl C, Reinert K, Kohlbacher O. OpenMS and TOPP:
open source software for LC-MS data analysis. Methods Mol Biol 2011; 696: 353-67.
[56] Katajamaa M, Miettinen J, Drešič M. MZmine: toolbox for
processing and visualization of mass spectrometry based molecular profile data. Bioinformatics 2006; 22: 634-6.
[57] Sturm M, Bertsch A, Gröpl C, et al. Open MS - an open-source
software framework for mass spectrometry. BMC Bioinformatics 2008; 9: 163.
[58] Brodsky L, Moussaieff A, Shafah N, Aharoni A, Rogachev I.
Evaluation of peak picking quality in LC-MS metabolomics data. Anal Chem 2010; 82: 9177-87.
[59] Melamud E, Vastag L, Rabinowitz JD. Metabolomic analysis and
visualization engine for LC-MS data. Anal Chem 2010; 82: 9818-26.
[60] Garcia-Pérez, Vallejo M, Garcia A, Legido-Quigley C, Barbas C. Metabolic fingerprinting with capillary electrophoresis. J Chromatogr A 2008; 1204: 150-9.
[61] van den Berg RA, Hoefsloot HC, Westerhuis JA, Smilde AK, van
der Werf MJ. Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics 2006; 7: 142.
[62] Field D, Sansone SA, Collis A, et al. Megascience. Omics data sharing. Science 2009; 326: 234-6.
[63] Jenkins H, Hardy N, Beckmann M, et al. A proposed framework
for the description of plant metabolomics experiments and their results. Nat Biotechnol 2004; 22: 1601-6.
[64] Hardy NW, Taylor CF. A roadmap for the establishment of
standard data exchange structures for metabolomics. Metabolomics 2007; 3: 1573-890.
[65] Pedroni PG, Eng JK, Hubley R, et al. A common open
representation of mass spectrometry data and its application to proteomics research. Nat Biotechnol 2004; 22: 1459-66.
[66] Tanaka S, Aoshima K, Miura Y, YO. 57th ASMS Conference on
Mass Spectrometry and Allied Topics (American Society for Mass Spectrometry). 2009; 31 May to 04 June; PA.
[67] Sugimoto M, Hirayama A, Ishikawa T, et al. Differential
metabolomics software for capillary electrophoresis-mass spectrometry data analysis. Metabolomics 2010; 6: 27-41.
Tautenhahn R, Böttcher C, Neumann S. Highly sensitive feature
alignment for UPLC-MS and HPLC-MS based metabolomics:
comparison of three algorithms for chromatograms alignment. J
Chromatogr A 2008; 1204: 150-9.
[68] Robinson MD, De Souza DP, Keen WW, et al. A dynamic
programming approach for the alignment of signal peaks in
multiple gas chromatography-mass spectrometry experiments. BMC Bioinformatics 2007; 8: 419.
Christin C, Hoefsloot HC, Smilde AK, Suits F, Bischoff R, Horvatovich PL. Time alignment algorithms based on selected
mass traces for complex LC-MS data. J Proteome Res 2010; 9: 1483-95.
[70] Diaz R, Ibáñez M, Sancho JV, Hernández F. Building an empirical
mass spectra library for screening of organic pollutants by ultra-
high-pressure liquid chromatography/hybrid quadrupole time-of-
flight mass spectrometry. Rapid Commun Mass Spectrom 2011;
25: 355-69.
[71] Nordström A, O'Maille G, Qin C, Siuzdak G. Nonlinear data
sharing. Science 2009; 326: 234-6.
[72] Draper J, Enot DP, Parker D, et al. Metabolite signal identification
in accurate mass metabolomics data with MZeDB, an interactive
m/z annotation tool utilising predicted ionisation behaviour 'rules'. BMC Bioinformatics 2009; 10: 227.
[73] Wolf S, Schmidt S, Müller-Hannemann M, Neumann S. In silico
detection for high resolution LC/MS. BMC Bioinformatics 2008; 9:
501.
[74] Wang W, Zhou H, Lin H, et al. Quantification of proteins and
metabolites by mass spectrometry without isotopic labeling or
spiked standards. Anal Chem 2003; 75: 4818-26.
[75] Brown M, Wedge DC, Goodacre R, et al. Automated workflows for
accurate mass-based putative metabolite identification in
LC/MS-derived metabolomic datasets. Bioinformatics 2011; 27: 1108-12.
[76] Hillmer JK, Bothern B. Physical signal modulation of time-of-flight
mass analyzers increases precision and decreases noise. Rapid
Commun Mass Spectrom 2009; 21: 7985-8.
[77] Shinoda K, Sugimoto M, Tomita M, Ishihama Y. Informatics for
differential metabolic analysis in metabolomics data. BMC Bioinformatics 2010; 11: 148.
time prediction and isotopic pattern information in CE-TOFMS data. Electrophoresis 2010; 31: 2311-8.

[99] Horai H, Arita M, Kanaya S, et al. MassBank: a public repository for sharing mass spectral data for life sciences. J Mass Spectrom 2010; 45: 703-14.

[100] Lee R, Ploemen AS, Niewczas L, Britz-McKibbin P. Integrative metabolomics for characterizing unknown low-abundance metabolites by capillary electrophoresis-mass spectrometry with computer simulations. Anal Chem 2007; 79: 403-15.

[101] Chalkauf KR, Lee R, Mills C, Britz-McKibbin P. Virtual quantification of metabolites by capillary electrophoresis-electrospray ionization-mass spectrometry: predicting ionization efficiency without chemical standards. Anal Chem 2009; 81: 2506-15.

[102] Parsons HM, Ekman DR, Collette TW, Viant MR. Spectral relative standard deviation: a practical benchmark in metabolomics. Analyst 2009; 134: 478-85.

[103] Bellmew M, Coram M, Fitzgibbon M, et al. A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS. Bioinformatics 2006; 22: 1902-9.

[104] Lange E, Tautenhahn R, Neumann S, Grop C. Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements. BMC Bioinformatics 2008; 9: 375.

[105] Matsuda F, Shino Y, Okawa A, et al. Assessment of metabolome annotation quality: a method for evaluating the false discovery rate of elemental composition searches. PLoS One 2009; 4: e7490.

[106] Fisher JD, Holliday GL, Rahman SA, Thornton JM. The MetaCyc Database of metabolic pathways. Nucleic Acids Res 2000; 28: 279-88.

[107] Čuplović-Culf M, Belacel N, Culf AS, et al. NMR metabolic analysis of samples using fuzzy K-means clustering. Magn Reson Chem 2009; 47 Suppl 1: S96-104.

[108] Dougherty ER, Barrera J, Brun M, et al. Inference from clustering with application to gene-expression microarrays. J Comput Biol 2002; 9: 105-26.

[109] Pears MR, Cooper JD, Mitchell HM, Mortishire-Smith RJF, Pearce DA, Griffin JL. High resolution 1H NMR-based metabolomics indicates a neurotransmitter cycling deficit in cerebral tissue from a mouse model of Batten disease. J Biol Chem 2005; 280: 42508-14.

[110] Faria AV, Macedo FC Jr., Ferreira MM, Cendes F. Classification of brain tumor extracts by high resolution 1H NMR using partial least squares discriminant analysis. Braz J Med Biol Res 2011; 44: 149-64.

[111] Žiketić L, Takăc J, Holmén I, Orešič M. Functional prediction of unidentified lipids using supervised classifiers. Metabolomics 2010; 6: 18-26.

[112]suit of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS. Bioinformatics 2006; 22: 1902-9.

[113] Seyfried EL, Fiehn O, Bezaire V, et al. Long-chain fatty acid combustion rate is associated with unique metabolite profiles in skeletal muscle mitochondria. PLoS One 2010; 5: e9834.

[114] Fawcett T. An introduction to ROC analysis. Pattern recognition letters 2006; 27: 861-74.

[115] Do JH, Choi DK. Clustering approaches to identifying gene expression patterns from DNA microarray data. Mol Cells 2008; 27: 279-88.

[116] Fawcett T. An introduction to ROC analysis. Pattern recognition letters 2006; 27: 861-74.

[117] Parson et al. 2009. 10: 363.

[118] Seifert EL, Fiehn O, Bezaire V, et al. Long-chain fatty acid combustion rate is associated with unique metabolite profiles in skeletal muscle mitochondria. PLoS One 2010; 5: e9834.

[119] Fawcett T. An introduction to ROC analysis. Pattern recognition letters 2006; 27: 861-74.

[120] Nam H, Chung BC, Kim Y, Lee K, Lee D. Combining tissue transcriptomics and urine metabolomics for breast cancer biomarker identification. Bioinformatics 2009; 25: 3151-7.

[121] Hennegues C, Bullinger D, Fux R, et al. Prediction of breast cancer by profiling of urinary RNA metabolites using Support Vector Machine-based feature selection. BMC Cancer 2009; 9: 104.

[122] Izquierdo-García JL, Rodríguez I, Kyriazis A, et al. A novel R-package graphic user interface for the analysis of metabolic profiles. BMC Bioinformatics 2009; 10: 363.

[123] Brougham DF, Ivanova G, Gottschalk M, Collins DM, Eustace AJ, O’Connor R, Havel J. Artificial neural networks for classification in metabolic studies of whole cells using 1H nuclear magnetic resonance. J Biomed Biotechnol 2011; 2011: 158094.

[124] Jansen MA, Kiwata J, Arceo F, Faull KF, Hanrathan P, Porter E. Evolving neural network optimization of cholesteryl ester separation by reversed-phase HPLC. Anal Bioanal Chem 2010; 397: 2376-87.

[125] Hanrathan P. Computational neural networks driving complex analytical problem solving. Anal Chem 2010; 82: 4307-13.

[126] Takeda I, Stretch C, Barnaby P, et al. Proposed minimum information standards for life sciences. J Mass Spectrom 2010; 45: 277-84.

[127] Sugimoto M, Goto H, Otomo K, et al. Metabolomic profiles and novel diagnostic approach for pancreatic cancer. Metabolomics 2010; 6: 211-21.

[128] Singhvi M, Koseki T, Hirayama A, et al. Metabolomic profiles and novel diagnostic approach for pancreatic cancer. Metabolomics 2010; 6: 211-21.

[129] Fawcett T. An introduction to ROC analysis. Pattern recognition letters 2006; 27: 861-74.

[130] Byrnes RW, Cotter D, Maer A, Li J, Nadeau D, Subramaniam S. Transcriptomics and urine metabolomics for breast cancer participation. OMICS 2006; 10: 158-63.

[131] Jourdan F, Cottret L, Hu C, et al. Use of reconstituted metabolic networks to assist in metabolic data visualization and mining. Metabolomics 2010; 6: 312-21.

[132] Hanrathan P. Computational neural networks driving complex analytical problem solving. Anal Chem 2010; 82: 4307-13.

[133] Jourdan F, Cottret L, Hu C, et al. Use of reconstituted metabolic networks to assist in metabolic data visualization and mining. Metabolomics 2010; 6: 312-21.

[134] Yetukuri L, Tilkka J, Holmén I, Orešič M. Functional prediction of unidentified lipids using supervised classifiers. Metabolomics 2010; 6: 18-26.

[135] Xie J, Wishart DS. Web-based inference of biological patterns, functions and pathways from metabolomic data using MetaboAnalyst. Nat Protoc 2011; 6: 743-60.

[136] Xie J, Psychogios N, Young N, Wishart DS. MetaMeth: a web server for automated metabolite data analysis and interpretation. Nucleic Acids Res 2009; 37: W652-60.

[137] Kono N, Arakawa K, Ogawa R, et al. Pathway projector: web-based zoomable pathway browser using KEGG atlas and Google Maps API. PLoS One 2009; 4: e7710.

[138] Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res 2010; 38: D355-60.

[139] Caspi R, Foerster H, Fulcher CA, et al. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res 2008; 36: D623-31.

[140] Frédéric M, Karp PD. Web-based metabolite network visualization with a zooming user interface. BMC Bioinformatics 2011; 12: 176.

[141] Byrnes RW, Cotter D, Maer A, Li J, Nadeau D, Subramaniam S. An editor for pathway drawing and data visualization in the Biopathways Workbench. BMC Syst Biol 2009; 3: 99.

[142] Fiehn O, Kristal B, van Ommen B, et al. Metabolomics 2005; 1: 107.

[143] Fawcett T. An introduction to ROC analysis. Pattern recognition letters 2006; 27: 861-74.

[144] Fiehn O, Kristal B, van Ommen B, et al. Metabolomics 2005; 1: 107.

[145] Fawcett T. An introduction to ROC analysis. Pattern recognition letters 2006; 27: 861-74.
sensory evaluation scores of Japanese sake and metabolome profiles. J Agric Food Chem 2010; 58: 374-83.

[148] Iino K, Sugimoto M, Soga T, Tomita M. Profiling of the charged metabolites of traditional herbal medicines using capillary electrophoresis time-of-flight mass spectrometry. Metabolomics; DOI:10.1007/s11306-011-0290-7.

[149] Kotowska A, Sworzczak K, Stepnowski P. Urine metabolomics analysis for adrenal incidentaloma activity detection and biomarker discovery. J Chromatogr B Analyt Technol Biomed Life Sci 2011; 879: 359-63.

[150] Sorlie T, Perou CM, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A 2001; 98: 10869-74.

[151] Enot DP, Beckmann M, Overy D, Draper J. Predicting interpretability of metabolome models based on behavior, putative identity, and biological relevance of explanatory signals. Proc Natl Acad Sci U S A 2006; 103: 14865-70.

[152] Beckmann M, Enot DP, Overy D, Draper J. Representation, comparison, and interpretation of metabolome fingerprint data for total composition analysis and quality trait investigation in potato cultivars. J Agric Food Chem 2007; 55: 3444-51.

[153] Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. Proc Natl Acad Sci U S A 2002; 99: 6562-6.

[154] Soga T, Sugimoto M, Honma M, et al. Serum metabolomics reveals gamma-glutamyl dipeptides as biomarkers for discrimination among different forms of liver disease. J Hepatol 2011; 55: 896-905.

[155] Sugimoto M, Wong DT, Hirayama A, Soga T, Tomita M. Capillary electrophoresis mass spectrometry-based saliva metabolomics identified oral, breast and pancreatic cancer-specific profiles. Metabolomics 2010; 6: 78-95.

[156] Kastenmüller G, Römisch-Margl W, Wägele B, Altmäier E, Suhre K. metaP-server: a web-based metabolomics data analysis tool. J Biomed Biotechnol 2011; 2011.

[157] Tautenhahn R, Patti GJ, Kalisiak E, et al. metaXCMS: second-order analysis of untargeted metabolomics data. Anal Chem 2011; 83: 696-700.

[158] Benton HF, Wong DM, Trauger SA, Siuzdak G. XCMS2: processing tandem mass spectrometry data for metabolite identification and structural characterization. Anal Chem 2008; 80: 6382-9.

[159] Grigsby CC, Rizki MM, Tamburino LA, Petsch RL, Shiyano PA, Cool DR. Metabolite differentiation and discovery lab (MeDDL): a new tool for biomarker discovery and mass spectral visualization. Anal Chem 2010; 82: 4386-95.

[160] Gao J, Tarcea VG, Karnovsky A, et al. Metscape: a Cytoscape plug-in for visualizing and interpreting metabolomic data in the context of human metabolic networks. Bioinformatics 2010; 26: 971-3.

[161] Hiller K, Hangebrauk J, Jager C, Spura J, Schreiber K, Schomburg D. MetaboliteDetector: comprehensive analysis tool for targeted and nontargeted GC/MS based metabolome analysis. Anal Chem 2009; 81: 3429-39.

[162] Lommen A. MetAlign: interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing. Anal Chem 2009; 81: 3079-86.

[163] Haimi P, Uphoff A, Hermansson M, Somerharju P. Software tools for analysis of mass spectrometric lipidome data. Anal Chem 2006; 78: 8324-31.

[164] Kohl M, Wiese S, Warscheid B. Cytoscape: software for visualization and analysis of biological networks. Methods Mol Biol 2011; 696: 291-303.

[165] Kuhn T, Willighagen EL, Zielesny A, Steinbeck C. CDK-Taverna: an open workflow environment for cheminformatics. BMC Bioinformatics 2010; 11: 159.

[166] Benton HP, Wong DM, Trauger SA, Siuzdak G. XCMS2: processing tandem mass spectrometry data for metabolite identification and structural characterization. Anal Chem 2008; 80: 6382-9.

[167] Cottret L, Wildridge D, Vinson F, et al. MetExplore: a web server to link metabolomic experiments and genome-scale metabolic networks. Nucleic Acids Res 2010; 38: W132-7.

[168] Carroll AJ, Badger MR, Harvey MA. The MetabolomeExpress Project: enabling web-based processing, analysis and transparent dissemination of GC/MS metabolomics datasets. BMC Bioinformatics 2010; 11: 376.

[169] Wang SY, Ho TJ, Kuo CH, Tseng YJ. Chromaligner: a web server for chromatogram alignment. Bioinformatics 2010; 26: 2338-9.