# Bioinformatics Tools for Metabolomic Data Processing and Analysis Using Untargeted Liquid Chromatography Coupled With Mass Spectrometry

Andrei G. LAZAR[1], Florina ROMANCIUC[1], Mihai Adrian SOCACIU[2,3] and Carmen SOCACIU[1,2*]

[1]University of Agricultural Sciences and Veterinary Medicine, 3-5 Mănăştur Street, Cluj-Napoca
[2] RTD Center for Applied Biotechnology, CCD-BIODIATECH; 12G Trifoiului Street, 400478, Cluj-Napoca, Romania
[3]University of Medicine and Pharmacy "Iuliu Haţieganu" Cluj-Napoca, 8 Babes Street, 400012, Cluj-Napoca, Romania
*Corresponding author, e-mail: carmen.socaciu@usamvcluj.ro

Bulletin UASVM Animal Science and Biotechnologies 72(2) / 2015
Print ISSN 1843-5262; Electronic ISSN 1843-536X
DOI:10.15835/buasvmcn-asb:11536

**Abstract**

Metabolomics is an important "omics" technology, complementary to genomics and proteomics, as parts of systems biology, giving information (qualitative fingerprints and quantitative profiling) as a mirror of cell and extracellular metabolic activity. A cohort of small metabolites are involved in the control and regulation of cellular functions, as intermediates or final products, their presence or levels being useful for the early diagnosis of different pathologies. Bioinformatics tools are mandatory for a future "computational" metabolomics, needed to manage large number of experimentally acquired data obtained from biological samples (plants, animal or human tissues). This review presents updated information about different high-throughput analytical techniques and data acquisition software (1-2), the pre-processing of data, converted to specific matrices, further processed by specific normalization and alignment procedures (3), then analysed by statistical univariate and multivariate chemometric and /or statistical techniques (4), identifying biomarkers by comparison with databases (5), and finally elucidating the networks and pathways (6). New software is available for data conversion, pre-processing, alignment algorithms, bucketing, normalization, underlying the challenges and comparisons with international data bases. Finally, the accurate identification of individual molecules as biomarkers, either evaluated by untargeted metabolomics techniques (Principal Component Analysis - PCA), Cluster Analysis - CA) or supervised ones (Partial Least Square Discriminant Analysis (PLS-DA) is presented. The accurate identification of metabolites and their involvement in metabolic networks and pathways became possible by well-established databases (HMDB, LIPID MAPS, KEGG, etc.), to validate all experimental data. Bioinformatics is a sine-qua-non tool, to be used and valorised by untargeted or targeted metabolomics, as an integrated technology in systems biology.

**Keywords**: LC/MS *metabolomics, bioinformatics, data processing, chemometrics, databases*

## INTRODUCTION

Metabolomics is an emerging *omics* technology, an important platform of systems biology, providing an integrative overview of the complex living systems. By qualitative and quantitative measurement of metabolites and their dynamic changes in biological systems, metabolomics and metabonomics became widely used in biomedicine, disease diagnostics, and drug pharmacology. Metabolites are small molecules (less than 1500 Da molecular weight), precursors, intermediates or final products of metabolic pathways, involved in cellular reactions. They are involved in the control and regulation of cell functions, very useful to make an appropriate fingerprinting of a physiological *vs* pathological status, providing a functional signature of phenotype, complementary to biochemical in-

formation obtained from genes, transcripts and proteins, as well as being biomarkers for early diagnosis or treatments monitoring (Romero *et al*., 2004; Putri *et al.,* 2013; Johnson *et al*., 2015).

Metabolomics is an emerging discipline with increasing technological advances (Roberts *et al.,* 2012; Putri *et al.,* 2013). Information about different metabolomes is limited, and complications often occur in data analysis, in spite of recent progress in bioinformatics procedures. Recently, scientists reviewed the state of the art regarding analytical methods in untargeted metabolomics (Alonso *et al.,* 2015) applied for biofluids, cell lysate, tissues, etc. (Dunn and Ellis, 2005; Dunn and Hankemeier, 2013). Bioinformatics tools are mandatory for metabolomics, as they can offer solutions for the adequate interpretation of a large number of experimental data, converted in specific matrices and further analysed by statistical methods (Wishart, 2007; Blekherman *et al.,* 2011). Therefore, it is an important goal of metabolomics analysis to assign metabolite identity to facilitate statistical analysis. Metabolomics bioinformatics is currently undergoing development for assigning identity and biological significance of metabolites in specific pathways (Nobeli and Thornton, 2006; Wishart *et al.,* 2007; Johnson *et al.,* 2015).

However, a main challenge in metabolomics is to find adequate bioinformatics tools and setting off parameters in order to obtain reliable and relevant results from biological data sets. The present article presents approaches used in pre-processing, aligning, normalizing and bucketing by providing a basic overview of metabolomics software for metabolomics data obtained with LC-MS method as an example to easily adopt proper bioinformatics strategies for reliable results (Sugimoti*et al.,* 2012; Dunn and Hankemeier, 2013). Several software tools of the multivariate analysis like PCA, CA and PLS are presented, as well as case studies. The identification of molecules is made by MS/MS techniques and comparisons with specific databases such as METLIN, Human metabolome database, LIPID MAPS for lipidomic metabolites, and MassBank (Melamud, 2010). Recently, web-based integrative information was released via Metabolite Atlas and Systems Biology Knowledge base, available online at http://metatlas.nersc.gov and http://kbase.us, respectively. Collaborative research infrastructure for computational metabolomics was also recently created (Workflow4Metabolomics) (Giacomini *et al.,* 2015)

This review underlines the role of bioinformatics for processing and analysis of workflows used nowadays in high-throughput untargeted metabolomic studies. of workflows used nowadays in high-throughput untargeted metabolomic studies. Updated information about different bioinformatics tools is useful to make accurate interpretation of metabolomics data obtained by LC-MS analysis. The large amount of data obtained by this technique requires a flow of data processing procedures (conversion, alignment, bucketing and normalizations, etc.) in order to create matrix data for statistical analysis and interpretation. Finally, the state-of-the-art of the existing databases is presented, as a good step forward for the integration of metabolomics in a general systems biology platform.

## STRATEGIES FOR UNTARGETED METABOLOMICSAND DATA MANAGEMENT

Untargeted metabolomic studies are characterized by simultaneous measurements of metabolites from blood, urine or other biofluids, as a top-down strategy, without considering a particular set of metabolites, but focusing on a global metabolomic profile (fingerprint). Such studies generate "big data", in terms of not only their volume, but also their complexity, implying a need for high performance bioinformatics tools (Putri *et al*., 2013). Conversely, targeted metabolomic studies are hypothesis-driven, characterized by specific metabolites measurements, with high precision and accuracy (Roberts *et al.,* 2012).

After raw data acquisition by advanced analytical techniques (liquid or gas chromatography coupled with mass spectrometry, magnetic resonance), data processing is necessary to guarantee the baseline correction, alignment of retention times, the elimination of noise and outliers, then normalization against specific molecules (or internal standards), creating specific matrices (buckets) to be processed statistically.

Fig. 1 shows the untargeted liquid chromatography-mass spectrometry analysis (LC/MS) to establish the metabolomics flow, including raw data acquisition and processing, data analysis,

statistics and interpretation, a real pipeline for an untargeted metabolomic study.

Specific strategies were elaborated to elucidate the physiological and/or pathological pathways in humans, animals or plants. As a recent example, one can mention MetaboLights (Steinbeck *et al.,* 2012; Salek *et al,* 2013;), MetaboAnalyst, a webserver for metabolomic data analysis and interpretation (Xia *et al.,* 2009, 2012), the metabolite atlases (Yao *et al,* 2015), as well as specific databases e.g. Lipidomics Gateway (http://www.lipidmaps.org/) (Yetukuri *et al,* 2007) and LipidXplorer for Consensual Cross-Platform Lipidomics (Herzog *et al.,* 2010).

## 1. RAW DATA ACQUISITION

Qualitative fingerprints and quantitative profiling of metabolites is made through a specialized method of data acquisition done by advanced analytical techniques based on chromatographic separation by Gas-chromatography (GC), Liquid Chromatography (LC) or Capillary Electrophoresis (CE) coupled with detection using photodiode array (DAD) or mass spectrometry (MS). Direct spectrometric identification can be made by MS or MS/MS, Magnetic Resonance (NMR) or imaging techniques (MALDI, MRI).

Recent publications report advances in untargeted ultrahigh performance liquid chromatography/electrospray ionization tandem mass spectrometry based metabolomics, integrated in platforms for the identification and relative quantification of the small-molecule complement of biological systems (Evans *et al,* 2009; Zhang *et al.,* 2012; Yin *et al.,* 2013; Gika *et al.,* 2014).

Mass spectrometry acquires spectral data, expressed as mass-to-charge ratio (m/z) and a relative intensity of the measured compounds. A wide range of instrumental and technical variants ( electrospray ionization, chemical ionization, simple quadrupole, triple quadrupole, time of flight system) are currently available for MS spectrometry, characterized by different ionization and mass selection methods, following advanced separation techniques, as reviewed recently (Tautehahn *et al.,* 2008; Theodoridis *et al.,* 2011; Goodwin *et al.,* 2014). High resolution and accuracy represent the main advantages, although high costs and specialized staff are needed.

## 2. DATA PROCESSING

Spectral data processing is a methodological approach aiming to achieve an accurate identification and quantification of molecular features, arranged in a feature matrix.

Data conversion is the first step in data processing binary format files, using different algorithms produced in house (Castillo *et al.,*

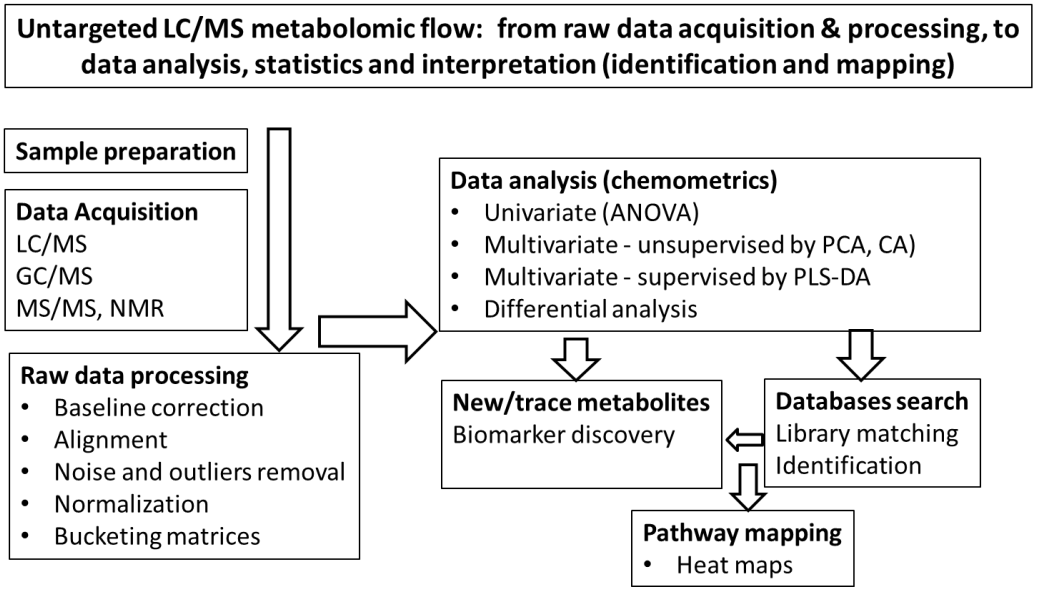


**Fig.1.** The LC/MS untargeted metabolomics flow: raw data acquisition and processing, data analysis, statistics and interpretation, a real pipeline for an untargeted metabolomic study.

2011; Kuhl *et al,* 2011; Hoffmann *et al,* 2012), independent operation systems and software (Sugimoto M. *et al.,* 2012). The most common file formats to store hyphenated-MS data are Net CDF and mzXML. The typical default output files are.RAW file, .RAW directory, .d directory, .d file, peg, .peg file, .raw, .D, Masslynx (.raw files), .wiff file. Data acquisition and processing software, including default output file formats provided by different producers are presented in table 1.

Relevant references were reported about data pre-processing, as well as raw data processing (Liland, 2011), to reduce data complexity for metabolites identification and evaluation (Hansen, 2007; Blekherman *et al.,* 2011). This includes noise filtering, peak detection, alignment, identification and normalization (Podwwojski *et al.,* 2009). There are several softwares which can be utilized for pre-processing, such as Metalign, MZmin, XCMS, Profile Analysis (Table 1).

1. MetAlign combines direct conversion to and from manufacturer formats, as well as netCDF, baseline correction, de-noising, accurate mass calculation, alignment, export of univariate statistical selections to differential MS data file, export to spreadsheets for multivariate statistical analysis and conversion of a multivariate statistical selection to a MS data file, being used and reported by many metabolomics publications, since 1995 to 2013 as the third cited pre-processing tool searched in SciFinder (www.cas.org/products/scifinder) and Web of Science (www.thomsonreuters.com/web-of-science) (Coble *et al.,*2014). MetAlign 3.0 is a new open source version which can be downloaded from the official site (www.metalign.nl) (Lommen *et al.,* 2012).

2. MZmine (Katajamaa *et al.*, 2006, 2007) allows easy comparisons of data across multiple samples, linear normalization with/without internal standards. For peak detection, it creates unique masses for connecting data points of successive scans, using 1-4 algorithms to build deconvoluted chromatograms (Pluskal *et al.,* 2010). MZmine2, which was released recently, contains identification methods to be used online in databases such as PubChem, KEGG, HMDB and METLIN.

3. XCMS package is the first cited software useful for pre-processing metabolic data, which incorporates matched filtration, peak detection, retention time alignment, and peak matching (Coble *et al.,* 2014). LC-MS feature detection can be done with an original implemented algorithm Filter (Katajamaa *et al*., 2007) and a new developed algorithm named CentWave (Tautenhahn *et al.,* 2008). Both algorithms are integrated in the Bioconductor R-package XCMS (http:// www.bioconductor.org/). Nonlinear retention time alignment, which is the most important for LC-MS, can be done using endogenous metabolites (Smith, 2006). This type of realignment is done by matching peaks representing the same analytes from different samples, very useful for untargeted analysis. An extension of this package, XCMS$^2$ can search automatically high quality MS-MS data in METLIN database (Benton *et al.,* 2008).

4. Profile analysis 2.0, provided by BrukerDaltonics, is applied for LC/MS pre-processing data and statistical analysis, including peak picking, filtering and normalization. Find Molecular Features (FMF) is a peak finding algorithm for quantitatively pinpointing relevant information; it eliminates noise and outliers,

**Tab.1.** Data acquisition and processing software, including default output file formats provided by producers*

| Provider of LC/MS (QTOF) | Manufacturer software | Default output format | Conversion software | Conversion output formats |
|---|---|---|---|---|
| Waters | Masslynx V 4.0 | Masslynx .raw file | MetAlign | .net; .cdf |
| Applied BioSystems | Analyst QS | .wiff file | File Translator | *.cdf |
| Bruker | Hystar Compass | .d, YEP, BAF, FID | Data Analysis Metabolic profiler | *.cdf, *.txt |
| AB SCIEX | Analyst | .wiff | Marker View | .mzXML |
| Agilent | MassHunter | .d | Mass Profiler MS Convert | |
| ThermoFisher | Xcalibur | .raw | Xcalibur | |

*Data sources of software inputs/outputs are available at: https://wiki.nbic.nl/index.php/Metabolomics_Data_formats_/_data_sources
https://bix-lab.ucsd.edu/display/Public/Data+Conversion+to+.mzXML

finally creating bucket tables of data. There are two options for creating such buckets: rectangular and advance bucketing. This software has various filtering and normalization options (www.bruker.de).

**Features and peak detection.** The objective of this step is to identify and quantify the features present in the spectra. Peak-based methods are the most common algorithmic choice for feature-detection in MS-based studies (Gika *et al*., 2014; Rafiei and Sleno, 2015), to detect the peaks across the spectrum and integrate their areas to provide a quantification of the underlying metabolite. Peak detection applies algorithms which analyse each sample spectrum independently (Tautenhahn *et al,* 2008, 2012; Pluskal *et al.,* 2010). In the first step, the spectra are smoothed using the Wavelet transform-based filter's superior performance, while in the second step, the different metabolite peaks are identified using one or multiple detection thresholds applied to different parameters such as the signal-to-noise ratio, intensity, or the area of each peak from the resulting filtered spectra.

Martens *et al.* (2010) report the new standard for mass Spectrometry data analysis mzML. Tools and databases of the KOMICS webportal for pre-processing, mining, and dissemination of metabolomics data were recently reported (Sakurai *et al*., 2014)

**Spectral alignment** is one of the main processing steps in metabolomic studies involving multiple samples. When analysing multiple spectra, the position of the peaks corresponding to the same metabolic feature may be affected by non-linear shifts. In MS-based analysis, peak shifts are observed across the retention time axis, and are generally associated with changes in the stationary phase of the chromatographic column. Spectral alignment methods are applied to correct this undesired variability in the samples that can profoundly affect the quality of the study (Podwwojski *et al.,* 2009; Jiang *et al.,* 2013). The spectral alignment algorithms can be divided in two main groups: methods where the spectral data are aligned before peak detection (a) and peak-based methods, where spectral peaks are aligned after the peak detection, using their coordinates (b), e.g. m/z and retention time in LC/MS). Spectral alignment methods (a) are also classified into warping and segmenting methods, the first ones applying a non-linear transformation to the retention time axis in order to maximize the correlation between the spectra. The alignment is then performed by either stretching or shrinking spectral segments to reach this correlation maximization. This is done by splitting the original sample and reference spectra into small segments, and by separately aligning each pair of segments. Alignment is performed through dynamic programming in such a way that limited changes in segment lengths are allowed. This way, the overall correlation between both spectra is effectively maximized. That is also based on dynamic programming, where a warping path is computed to which the connected data points of each spectrum are equivalent. (Alonso *et al.,* 2015).

A good methodological report (Hoffmann *et al*., 2012) introduces two algorithms for retention time alignment of GC-MS datasets by two methods: BIPACE (bidirectional peak assignment and cluster extension) and CEMAPP-DTW (Center-star multiple alignment by pairwise partitioned dynamic time warping). Recently released, the Icoshift algorithm (Savorani *et al.,* 2010) is one of the most commonly used segmentation methods which is based on the convergence toward a reference signal.

Compared to the warping and segmentation alignment methods, peak-based methods (b) are applied after peak detection. This type of method is implemented in the XCMS software (Tautenhahn *et al.,* 2012). Since the shifts along the m/z axis are minimal and the m/z axis has a high resolution, the data can be safely binned in m/z intervals, and peak alignment can be performed on each bin, along the chromatographic time. Another common alignment method used in MS is the RANSAC algorithm (Pluskal *et al.,* 2010), which allows the corresponding peaks across samples to be identified by LOESS regression on different retention times and m/z windows.

## 3. DATA ANALYSIS BY CHEMOMETRICS (STATISTICAL EVALUATION)

Chemometrics became an interdisciplinary science of extracting and analysing data cohorts, solving descriptive and predictive problems from experimental data. Once the metabolite features are identified, there are multiple univariate and multivariate statistical methods that can be used to perform the desired study analysis. These

groups of techniques are commonly known as chemometric methods (Van der Greef, 2005; Lavine and Workman, 2006; Madsen *et al.,* 2010),

Each metabolomic feature is intrinsically related to the concentration of a particular metabolite. Depending on the analytical technique and the spectral processing workflow that have been used, different metabolomic features are used as input for data analysis. Progress in using bioinformatics tool MetScape2, for the analysis and visualization of metabolomics and gene expression data was reported recently (Karnovschi *et al.,* 2012).

Spectral peak areas are the most commonly used features in high-throughput MS-based metabolomics data. They are computed through the integration of the peaks identified and aligned using the methods described in the previous section.

**Univariate methods** analyse metabolomic features independently. They are common, easy to use statistical analysis approaches, which get fast interpretation. However, their main disadvantage is that they do not take into account the presence of interactions between the different metabolic features, which is usually the case in biological samples, with very complex composition and many correlations between features from the same metabolite and inter-metabolites of the same pathway. Also, the effect of potential confounding variables (age, gender, diet, body mass index) is not taken into account by these methods, increasing the probability of obtaining false positive or false negative results (Saccenti *et al.,* 2014; Alonso *et al,* 2015)

In contrast to univariate methods, **multivariate analysis** methods take into account all the metabolomic features simultaneously and, therefore, can identify relationship patterns between them. These pattern-recognition methods can be classified into supervised and unsupervised methods (Worley and Powers, 2013).

In **unsupervised methods**, the similarity patterns within the data are identified without taking into account the type or class of the metabolite. These are often applied to summarize the complex metabolomic data. They provide an effective way to detect data patterns to be correlated with experimental and/or biological variables. Principal component analysis (PCA) is the most commonly used unsupervised method in metabolomic studies (Alonso *et al*., 2015). PCA is based on the linear transformation of the metabolic features into a set of linearly uncorrelated (i.e., orthogonal) variables known as principal components. This decomposition method maximizes the variance explained by the first component, while the subsequent components explain increasingly reduced amounts of variance. At the same time, PCA minimizes the covariance between these components. After applying the PCA, a set of loading and score vectors are obtained: loading vectors represent the principal components and each vector coefficient corresponds to the contribution of each variable to the principal component (i) while the score vectors represent the projection of each sample onto the new orthogonal basis. Plotting the scores over the first principal components is a convenient way of summarizing the global dataset, since first principal components "capture" most of the dataset variability. PCA is also used in metabolomics studies to assess data quality, since it can identify sample outliers or reveal hidden biases in the study, as well as the impact of technical variation in the analysis of metabolic profiles (Gika *et al.,* 2014; Yin *et al.,* 2013). Other unsupervised methods, such as hierarchical clustering analysis (HCA) and self-organizing maps (SOMs), have also been applied to metabolomic data (Tsugawa, 2011; Goodwin *et al.,* 2014), being suitable to detect non-linear parameters that are not covered by PCA and also to visualize metabolic phenotypes and feature patterns based on their similarity (Alonso *et al.,* 2015)

**Supervised methods** are used to identify metabolic patterns that are correlated with the phenotypic variable of interest while suppressing other sources of variance (Xia *et al*., 2013). Partial least squares (PLS) is one of the most widely used supervised method in metabolomics, being used either as a regression analysis (i.e., quantitative variable of interest) or as a binary classifier (PLS-DA). Unlike PCA, PLS components do not maximize the explained dataset variance but the covariance between the variable of interest and the metabolomic data. Therefore, the loadings of PLS components represent a measure of how much a feature contributes to the discrimination of the different sample groups. However, one weakness of PLS is that some metabolic features that are not correlated with the variable of interest

can influence the results. To surpass this problem, orthogonal PLS (O-PLS) was developed, as a model which factorizes the data variance into two components: a first component correlated with the variable of interest, and a second uncorrelated component (i.e., orthogonal). Classification of metabolomics samples is commonly performed by fitting the discriminant analysis versions of PLS and O-PLS models, named PLS-DA, O-PLS-DA. In the last years, a progressive move from the use of PLS models to O-PLS has been observed in the metabolomics field.

In metabolomics, many classical statistical tools (t-test, ANOVA) such as correlation analysis and simple linear regression also utilize pairs of variables, and these methods can be extended to manage hundreds or thousands of variables simultaneously. Scales of relationship between variables are reflected by covariance or correlation, which are used in multivariate analysis apart from the mean (Saccenti *et al.,* 2014). Several software tools are or can be used for multivariate analysis in metabolomics; however the article present only a few used in multivariate analysis such are: SIMCA (Tsugawa *et al*, 2011), Unscrambler 10.X, ProfileAnalysis 2.0 andMultiBase. Table 4 presents types of data analysis.

**Unscrambler10.X** software includes many functions, e.g. advanced regression, classification and predictive modelling tools, Cluster Analysis, PCA, PLS-DA and many more. This software provides great visualization option for PCA plots such as 2-D and 3-D graphics, as well as common rotation on the 3-D results that helps to identify and understand patterns in large or complex data sets. PCA is useful to see differences between samples, what variables are involved in most of these differences and whether these variables are correlated or independent from one another. PLS-DA sharpens the difference between groups by rotating PCA components in order to obtain maximum separation between classes, and it helps to understand which variables are responsible for the molecular separation. Cluster Analysis is done by K-Means methodology, based on specific distance measurements. More information about this software can be found on the producer's website (http://www.camo.com/)

**Profile Analysis** includes PCA, Hierarchical Clustering Analysis (HCA) and PLS-DA. It extracts the relevant statistics on the distribution of samples and group characteristics using PCA. Profile Analysis links statistical results with the original LC-MS data provided by Hystar (Bruker), enabling fast on-the-fly identification of potential biomarkers by SmartFormula. PCA and HCA quickly reveal sample clustering helping to focus on the relevant information in a data set. Visualization of PCA plots are available only in 2-D form on this software.

**Multibase** provides a set of powerful, graphical statistical tools, an Excel Add-In program, which can directly process Excel data without any troublesome conversions. It includes PCA, Partial Least Squares (PLS), Cluster Analysis and others. With PCA, the sample distribution is simplified and the underlying factors explain the pattern of variable and sample correlations. The sample group distributions of different categories are shown with ellipses. All information about

**Tab.2.** Tools available for LC/MS or MS metabolomics spectral processing and data analysis

| Tool | Type | Website | Reference |
|---|---|---|---|
| MetaboAnalyst2 | web | http://www.metaboanalyst.ca/ | Xia *et al.,*2012 |
| XCMS | R | http://metlin.scripps.edu/xcms/ | Smith *et al.,* 2006 |
| MetSign | MatLab | http://metaopen.sourceforge.net/ | Lommen and Kools, 2012 |
| XCMS online | Web | https://xcmsonline.scripps.edu/ | Tautenhahn *et al.,*2012 |
| MAVEN | Application | http://genomics-pubs.princeton.edu/mzroll | Melamud *et al.,*2010 |
| mzMine2 | Application | http://mzmine.sourceforge.net/ | Pluskal *et al.,*2010 |
| AStream | R | http://www.urr.cat/AStream/AStream.html | Alonso *et al.,*2015 |
| Camera | R | http://metlin.scripps.edu/xcms/ | Kuhl *et al.,*2011 |
| MetDAT | Web | http://smbl.nus.edu.sg/METDAT2/ | Xia e*t al, 2009* |

Multibase can be found on the website of the developer (http://www.numericaldynamics.com/).

As a summary for data processing and analysis, Table 2 includes a list of the most important tools available for LC/MS metabolomics spectral processing and data analysis

To illustrate the diverse numerical interactions among multiple, high-complexity data, novel Venn Diagram programs were created to compare, superpose and visualize datasets with differentially regulated data sets. Recently, Cai *et al*. (2013) reported updated information about the utility of Venn diagram generators (www.pangloss.com/seidel/protocols/venn4.cgi) and the web application for comparison of biological dataset lists.

## 4. DATABASES AND METABOLITE IDENTIFICATION

Metabolite identification is one of the major challenges of high throughput metabolomic analysis, an indispensable step which confers the biological meaning of the associated features in a metabolomic study. In MS-based studies, the common metabolite identification approach is based on querying metabolomic databases for the neutral molecular mass values of the identified peaks using a tolerance window. Recently, an automatic identification toolkit (MAIT) included in an R package to analyse LC/MS metabolomic data was created (Fernández-Albert *et al.,* 2014).

**Biomarker discovery and identification** is one of the most promising applications of metabolomics, especially in the medical sciences. They are usually determined using supervised analysis models, capable to aggregate the evidence of multiple metabolites. The usefulness of the resulting classification models is proved by their use in clinical investigations. Recently, Gao *et al*. (2010) reported the use of MetScape, a Cytoscape plug-in system useful for visualizing and interpreting metabolomic data in the context of human metabolic networks.

**Identification tools: metabolomic databases**

Several free electronic databases are available, such as Human Metabolomic Data Base (Wishart *et al.,* 2009), LIPID MAPS (Sud *et al.,* 2007), KEGG (Kanehisa *et al,* 2012), MetaCyc (Caspi *et al.,* 2014), HumanCyc (Romero *et al.,* 2004), Reactome (Croft D. *et al.,* 2011), Metlin (Smith *et al.,* 2005), which offer the possibility to identify unknown molecules based on their m/z, visualized pathway maps and detailed information about metabolites. Some of these databases (HMDB, METLIN) contain human biomedical information. Metabolic pathway databases (KEGG, MetaCyc, HumanCyc, Reactome) contain reference to different organisms, reaction and pathways, and facilitate the exploration of metabolism between species (Wishart *et al.,* 2007).

The existing metabolome databases (more than 30) are based on mass spectrum, and they are compound and pathways oriented, as described recently by Fukushima and Kusano (2013).

Table 3 presents the most widely known databases using LC/MS data, their characteristics, website links and references.

To simplify data analysis and improve accessibility, a free, accessible, easy-to-use web server for metabolomic data analysis called MetaboAnalyst is functional. It includes options for data processing, normalization, multivariate statistical analysis (t-tests, PCA, PLS-DA) metabolite identification and pathway mapping. It includes also a large library of reference spectra to facilitate compound identification from many input spectra. The new version, MetaboAnalyst 2.0 is hosted on a powerful server with substantially modified code to take advantage the server's multi-core CPUs for computationally intensive tasks. A downloadable version of MetaboAnalyst 2.0, along with detailed instructions for local installation is now available (Xia *et al.,* 2012). HORA suite (Human blOod Range vAlidator) consists of a Java application used to validate the metabolomic analysis of human blood against a database that stores the normal plasma and serum range concentrations of metabolites. The goal of HORA is to find the metabolites that are outside the normal range and to show those not present in the list provided by the user, for different thresholds of concentration (Bruschi *et al.,* 2008)

**The pathway and network analysis** amplifies the information generated by metabolomic studies useful for systems biology. Both approaches exploit the relational properties found in metabolomic data. Pathway analysis uses the preliminary biological knowledge to analyse metabolite patterns from an integrative point of view. Alternatively, network analysis uses

**Tab. 3**. List of Metabolomics databases and their main characteristics and website links and references

| Database short name | Main Characteristics | Web address link and relevant references |
|---|---|---|
| GMD | Database based on GC/MS data | www.gmd.mpimp-golm.mpg.de (Kopka, 2005) |
| HMDB | MS search data on metabolites (MZ format), clinical data, molecular biology and biochemistry; Metabolite data in biofluids (blood, urine, saliva, etc.); Searching of compounds is easy, advanced search tool allows users to select or search over several combinations of subfields. | http://www.hmdb.ca (Wishart *et al*., 2013) |
| HORA suite | Java application, open-source software, including Database and software for human metabolomics. | http://www.paternostrolab.org (Bruschi *et al*, 2008) |
| Human Cyc | Similar to MetaCyc but adapted to human metabolism. Tools for querying, visualizing and analysing the underlying database and for analysing omics data. Still in progress. | http://humancyc.org/ (Romero *et al*., 2004; Wishart *et al.,* 2007). |
| KEGG | Full name: Kyoto Encyclopedia of Genes and Genomes<br>Includes 17 databases, including biological systems' information, genomic information, chemical information and health information.<br>The most known and complete pathway database, for many organisms. | http://www.genome.jp/kegg/<br>http://www.kegg.jp/<br>(Kanehisa *et al.,* 2012) |
| LIPID MAPS LipidBank (for natural lipids) | A lipidomic database which offers: LIPID MAPS, detailed biochemical pathways, protocols, analytical tools for lipid quantitation, structure drawing tools, mass spectrometric standards and experimental data.<br>It includes two databases: LIPID MAPS Structure Database (LMSD), LIPID MAPS Proteome Database (LMPD) (8500 genes and 12 000 proteins from several species and humans). | http://www.lipidmaps.org/ (Melanund, 2010)<br>http://lipidbank.jp (Fukushima and Kusano, 2013) |
| LSMD | 37500 lipid structures with MS/MS data | http://www.lipidmaps.org (Sud *et al.,* 2007) |
| Mapman | A stand-alone tool for analysis of omics data | http://mapman.gabipd.org |
| MassBank and Bio-Massbank | 2337 metabolites and more than 41000 spectra LC/MS and GC/MS | http://www.massbank.jp; http://www.bio.massbank.jp; (Horai *et al*., 2010) |
| MassBase | Mass Spectral archive for LC-MS data | http://webs2.kazusa.or.jp/massbase (Fukushima &Kusano, 2013) |
| MeltDB | A web-based system for data analysis and management of metabolomics | http://meltdb.cebitec.uni-bielefeld.de (Kessler *et al*, 2013) |
| MetaboAnalyst | A web-based metabolomic data processing tool which includes NMR spectra, MS peak lists, compound/concentration data in a wide variety of formats | http://www.metaboanalyst.ca (Xia *et al*., 2012) |
| MetaboLights | An open-access repository considered a new "cosmos" database for metabolomics and metadata. | www.ebi.ac.uk/metabolight (Steinbeck *et al.,* 2012; Salek *et al,* 2013; Haug *et al.,* 2013;) |
| Metabolomics. jp | A wiki database for metabolomics | http://metabolomics.jp (Fukushima and Kusano, 2013) |
| MetaboSearch | Mass-Based Metabolite Identification Using Multiple Databases | http://omics.georgetown.edu:443/ MetaboSearch.jsp (Zhou *et al.,* 2012) |
| MetaCyc | Experimental database, which may elucidate metabolic pathways from all domains of life. It contains reactions, chemical compounds and genes. It is an *online* encyclopedia of metabolism, stores and predicts metabolic pathways involved in primary/ secondary metabolism, sequenced genomes, supports metabolic engineering *via* enzyme database and metabolite database helping metabolomics research. | http://metacyc.org/ (Caspi *et al.,* 2014) |
| MetiTree | Database for small molecules | http://www.MetiTree.nl |
| METLIN | Contains metabolite information and tandem mass spectrometry data. Facilitates metabolite identification, provides MS/MS comprehensive metabolite data. Each metabolite is linked to outside resources. | http://metlin.scripps.edu/ (Tautenhahn *et al.,* 2012) |
| MetScape | A cytoscape plugin for visualizing metabolomic data | http://metscape.ncibi.org |

| | | |
|---|---|---|
| MMCD | Madison Metabolomics Consortium Database - metabolite identification | http://mmcd.nmrfam.wisc.edu/ (Cui *et al,* 2008) |
| NIST | Mass spectra database based on LC/MS and GC/MS, $MS^2$ analytical data | http://www.sisweb.com/software/ms/nist (Fukushima and Kusano, 2013) |
| Pathvisio | A tool to visualize biological pathways | http://www.pathvisio.org |
| PubChem ChEBI | | http://pubchem.ncbi.nlm.nih.gov http://www.ebi.ac.uk/chebi |
| Reactome | It is a manually curated database, which provides intuitive bioinformatics tools for the visualization, interpretation and analysis of pathway knowledge. | http://www.reactome.org/ (Croft *et al*., 2011) |
| SMPDB 2.0 | More than 1500 metabolites mapping >730 small molecules pathways found in humans. More than 350 hand-drawn pathways of small molecule metabolism involved in disease pathways and signalling, 280 being unique molecules. | http://www.smpdb.ca (Frolkis *et al,* 2010; Jewison *et al.,* 2014) |
| VANTED | A stand-alone tool for mapping omics data into metabolic pathways | http://vanted.ipk-gatersleben.de |
| WikiPathways | 1910 pathways | http://wikipathways.org/ (Kelder *et al,* 2012) |

the high degree of correlation existing inside the metabolomics data, to build metabolic networks that characterize the complex relationships existing in the set of measured metabolites.

Until now, when analysing metabolomic data, no prior knowledge regarding metabolite relationships was assumed, but in the last years, biological knowledge available for metabolomics studies has been constantly increasing. Metabolic pathways link together metabolites related to the same biological process, directly or indirectly connected by one or multiple enzymatic reactions.

Biological databases such as Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa *et al.,* 2012), Small Molecule Pathway DataBase (SMPDB) (Jewison *et al*., 2014), EHMN (Ma *et al.,* 2007), WikiPathways (Kelder *et al*., 2012), MetaCyc (Caspi *et al*., 2008) provide exhaustive information about a large number of metabolic pathways. The availability of these databases therefore enables the use of pathway-based approaches in metabolomics. These methods are referred to as a metabolite set enrichment analysis (MSEA), and are methodologically based on the gene set enrichment analysis (GSEA) approach, designed for pathway analysis of gene-expression data (Khatri *et al.,* 2012). Three different approaches have been developed to perform MSEA (Xia and Wishart, 2010) as presented by Alfonso *et al.* (2015).

## 5. INTEGRATION OF OMICS DATA

Systems biology integrates methods which improve our understanding of biological processes, including metabolomics. New platforms, such as http://kbase.us (Knowledge based platform on systems biology) are functional, as well as integrative metabolite atlases (Bowen and Northen, 2010; Yao *et al.,* 2015) which provide the framework and interface, web-based access to raw mass spectrometry data related to chemicals detected. The Bioportal also offers an enhanced web service for medical ontology to access and to use in software applications (Whetzel *et al,* 2011). Xia *et al* (2013) reports another web-based tool INMEX, for integrative meta-analysis of expression data.

Metabolomics is a technology which offers a comprehensive detection of small metabolites but the information became difficult to harmonize considering the complicated analytical experiments (metadata), the big number of databases, as shown before, and the low reusability of the published data. To tackle these issues, Ara *et al. (*2015) report the creation of Metabolonote: a Wiki-based database for managing hierarchical metadata of metabolome analyses, using a well-defined metadata and description format called TogoMD. A total of 808 metadata from 35 species are yet available at http://metabolonote.kazusa.or.jp

To summarize, during the last years, high-throughput technologies improved the analysis of the biologic variability at multiple molecular levels, from the genome, epigenome, transcriptome, proteome to metabolome level, using systems biology approaches. In spite of this progress, major improvements are expected in the next years,

3Omics being a good example of this new type of systems biology integrative analysis including metabolomics (Kuo *et al.,* 2013). Collaborative research infrastructures for computational metabolomics were also recently created, such as Workflow4Metabolomics (Giacomoni *et al.,* 2015).

## CONCLUSION

The review summarizes updated knowledge related to the role and impact of bioinformatics tools in LC/MS based metabolomics, presenting the flow of data management, from data acquisition, to data processing, statistical analysis and interpretation, biomarker discovery and detection, in agreement with existing databases and integrative platforms.

Considering the large variety of analytical methods and protocols, and the large number of data acquisition methods with pre-processing algorithms (made *in-house*), it is difficult to assume clear conclusions related to standardized procedures.

Concerning the unsupervised or supervised analysis (PCA, Cluster Analysis, PLS-DA, etc.), commercial software is available and easy to apply.

Finally, the identification of biomarkers is accessible by comparison with free databases (HMDB, LIPID MAPS, KEGG, etc.), which help the elucidation of metabolites/biomarkers and specific pathways.

To conclude, specific strategies to use bioinformatics tools are needed, to provide accurate evaluations integrated in the 'omics' technology, applied in systems biology.

***Acknowledgement*:** This paper was partly supported from the European DISCO project FP7-KBBE-2013-613513.

## REFERENCES

1. Reviews on Metabolomics and bioinformatics
2. Alonso A, Marsal S, Julia A (2015). Analytical Methods in untargeted metabolomics: state of the art in 2015, Frontiers in BioengBiotechnol 3:1-20
3. Blekherman G, Laubenbacher R, Cortes DF, *et al* (2011). Bioinformatics tools for cancer metabolomics. Metabolomics7(3): 329–343.
4. Dunn WB, Ellis DI (2005). Metabolomics: Current analytical platforms and methodologies. Trac-Trends Anal Chem 24 (4): 285-294.
5. Dunn WB, Hankemeier T (2013). Mass spectrometry and metabolomics: past, present and future. Metabolomics 9: 1–3.
6. Johnson CH, Ivanisevic J, Benton HP, Siuzdak G (2015). Bioinformatics: The Next Frontier of Metabolomics. Anal Chem87(1): 147–156.
7. Metabolite Atlas. Available online: http://metatlas.nersc.gov
8. Nobeli I, Thornton JM (2006). A bioinformatician's view of the metabolome. BioEssays 28:534–545.
9. Putri SP, Yamamoto S, Tsugawa H, Fukusaki E (2013). Current metabolomics: technological advances. *J BiosciBioeng*116: 9–16.
10. Roberts LD, Souza AL, Gerszten RE, Clish CB (2012). Targeted metabolomics. CurrProtocMolBiol Chapter 30: 1–24.
11. Romero PJ, Wagg ML, Green D *et al* (2004). Computational prediction of human metabolic pathways from the complete human genome. Genome Biol 6: 1-17.
12. Sugimoto M, Kawakami M, Robert M, Soga T, Tomita M (2012). Bioinformatics Tools for Mass Spectroscopy-Based Metabolomic Data Processing and Analysis. Current Bioinformatics7(1): 96–108.
13. The Systems Biology Knowledgebase. Available online: http://kbase.us
14. Wishart DS (2007). Current progress in computational metabolomics. Briefings in Bioinformatics8(5): 279–293.
15. Yetukuri L, Katajamaa M, Gema Medina-Gomez G, *et al* (2007**).** Bioinformatics strategies for lipidomics analysis: characterization of obesity related hepatic steatosis. BMC Systems Biology 1:12-27.
16. Data Acquisition
17. Evans AM, DeHaven CD, Barrett T, *et al* (2009). Integrated, nontargeted ultrahigh performance liquid chromatography/electrospray ionization tandem mass spectrometry platform for the identification and relative quantification of the small-molecule complement of biological systems. Anal Chem 81: 6656–6667.
18. Gika, HG, Theodoridis GA, Plumb RS, Wilson ID (2014). Current practice of liquid chromatography–mass spectrometry in metabolomics and metabonomics. J Pharm Biomed Anal 87: 12–25.
19. Goodwin CR, Sherrod SD, Marasco F *et al* (2014). Phenotypic mapping of metabolic profiles using self-organizing maps of high-dimensional mass spectrometry data. Anal Chem 86: 6563–6571.
20. Herzog R, Schuhmann K, Schwudke D *et al* (2010) LipidXplorer: A Software for Consensual Cross-Platform Lipidomics. *PLoS ONE 7:* e29851.
21. Tautenhahn R, Bottcher C, Neumann S (2008). Highly sensitive feature detection for high resolution LC/MS. BMC Bioinformatics 9:504-512.
22. Theodoridis G, Gika HG, Wilson ID (2011). Mass spectrometry-based holistic analytical approaches for metabolite profiling in systems biology studies. Mass Spectrom Rev 30: 884–906.

23. Yin P, Peter A, Franken H *et al* (2013). Preanalytical aspects and sample quality assessment in metabolomics studies of human blood. ClinChem 59(5):833-45.
24. Zhang A, Sun H, Wang P *et al* (2012). Modern analytical techniques in metabolomics analysis. Analyst 137: 293–300.

## Data processing

25. Benton HP, Wong DM, Trauger SA, Siuzdak G (2008). XCMS(2): Processing tandem mass spectrometry data for metabolite identification and structural characterization. Anal Chem 80C:6382–6389.
26. Castillo S, Gopalacharyulu P, Yetukuri L, Orešič M (2011). Algorithms and tools for the preprocessing of LC-MS metabolomics data. Chem and Intelligent Lab Systems108(1):23–32.
27. Coble JB, Fraga CG (2014). Comparative evaluation of preprocessing freeware on chromatography/mass spectrometry data for signature discovery. J Chromatogr A1358:155–164.
28. Hoffmann N, Keck M, Neuweger H, *et al* (2012). Combining peak- and chromatogram-based retention time alignment algorithms for multiple chromatography-mass spectrometry datasets. BMC Bioinformatics 13: 214-234.
29. Jiang W, Zhang ZM, Yun Y *et al* (2013). Comparisons of five algorithms for chromatogram alignment. Chromatographia 7: 1067–1078.
30. Katajamaa M, Miettinen J, Oresic M (2006) MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. Bioinformatics 22:634-636.
31. Katajamaa M, Oresic M (2007). Data processing for mass spectrometry-based metabolomics. J Chromatogr 1158: 318-328.
32. Kuhl C, Tautenhahn R, Böttcher C *et al* (2011). CAM-ERA: an integrated strategy forcompound spectra extraction and annotation of liquid chromatography/mass spectrometry datasets. Anal Chem 84: 283–289.
33. Lommen A, Kools HJ (2012). MetAlign 3.0: Performance enhancement by efficient use of advances in computer hardware. Metabolomics8(4): 719–726.
34. Martens L, Chambers M, Sturm M *et al* (2010) mzML—a Community Standard for Mass Spectrometry Data. Mol. Cell. Proteomics 10(1): R110.000133.
35. MelamudE,Vastag L, Rabinowitz JD (2010). Metabolomic analysis and visualization engine for LC-MS data. Anal Chem 82: 9818–9826.
36. Pluskal T, Castillo S,Villar-Briones A, Oresic M (2010). MZmine2: modular framework for processing, visualizing and analyzing mass spectrometry - based molecular profile data. BMC Bioinformatics 11:395-405.
37. Podwojski K, Frisch A, Chamrad, DC *et al* (2009). Retention time alignment algorithms for LC/MS data must consider non-linear shifts. Bioinformatics 25(6): 758–764.
38. Rafiei A, Sleno R (2015).Comparison of peak-picking workflows for untargeted liquid chromatography/ high-resolution mass spectrometry metabolomics data analysis. Rapid Comm Mass Spectrom29:119-127
39. Sakurai N, Ara T, Enomoto M *et al* (2014). Tools and databases of the KOMICS webportal for preprocessing, mining, and dissemination of metabolomics data. Biomed Res Int 194812.
40. Smith CA, Want EJ, O'Maille G *et al* (2006). XCMS: processing massspectrometry data for metabolite profiling using non- linear peak alignment, matching, and identification. Anal Chem 78: 779–787.

## Data Analysis

41. Cai H, Chen H, Yi1 T, *et al* (2013). VennPlex–A Novel Venn Diagram Program for Comparing and Visualizing Datasets with Differentially Regulated Datapoints. PlosOne 8(1): e53388.
42. Fernández-Albert F, Llorach R, Andrés-Lacueva C, Perera A (2014). An R package to analyse LC/MS metabolomic data: MAIT (metabolite automatic identification toolkit). Bioinformatics 30:1937–1939.
43. Hansen MAE (2007). Data analysis. In:S. G. Villas-Boas SG, Roessner E (Eds.) Metabolome analysis: An introduction (pp. 146–187). Hoboken NJ: Wiley .
44. Karnovsky A, Weymouth T, Hull T, *et al* (2012). MetScape2 bioinformatics tool for the analysis and visualization of metabolomics and gene expression data. Bioinformatics 28: 373–380.
45. Lavine B, Workman J (2006). Chemometrics. Anal Chem 78:4137-4145
46. Liland KH (2011). Multivariate methods in metabolomics - from pre-processing to dimension reduction and statistical analysis. Trends Anal Chem 30(6): 827–841.
47. Madsen R, Lundstedt T, Trygg J (2010). Chemometrics in metabolomics–A review in human disease diagnosis. Anal ChimActa 659: 23–33.
48. Saccenti E, Hoefsloot HC, Smilde AK *et al* (2014). Reflections on univariate and multivariate analysis of metabolomics data. Metabolomics10(3): 361–374.
49. Tsugawa H, Tsujimoto Y, Arita M,. Bamba T, Fukusaki E (2011). GC/MS based metabolomics: Development of a data mining system for metabolite identification by using soft independent modeling of class analogy (SIMCA). BMC Bioinformatics 12:131–143.
50. Van der Greef J, Smilde AK (2005). Symbiosis of chemometrics and metabolomics: past, present, and future, J Chemometrics 19: 376–386.
51. Worley B, Powers R (2013). Multivariate Analysis in Metabolomics. Curr Metabolomics1: 92–107.
52. Xia J, Mandal R, Sinelnikov IV *et al* (2012). MetaboAnalyst2.0–a comprehensive server for metabolomic data analysis. Nucl Acids Res 40: W127–W133.
53. Xia J, Psychogios N, Young N, Wishart DS (2009). MetaboAnalyst: a web server for metabolomic data analysis and interpretationNucl Acids Res.37 (suppl 2): W652-W660.

## Data bases

54. Ara T, Enomoto M, Arita M *et al* (2015). Metabolonote: a wikibased database for managing hierarchical metadata of metabolome analyses. Frontiers Bioeng Biotechnol 3:1-9

55. Bowen BP, Northen TR (2010). Dealing with the unknown: Metabolomics and metabolite atlases. J Am Soc Mass Spectrom 21: 1471–1476.
56. Bruschi S, Calzolari D, Coquin L, Paternostro G (2008) HORA suite: a database and software for human metabolomics. Metabolomics 4:90–93.
57. Caspi R, Altman T, Billington R, *et al* (2014). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases Nucleic Acids Research 42: D459-D471
58. Croft D, O'Kelly G, Wuetal G (2011). Reactome: a database of reactions pathways and biological processes. Nucl Acids Res 39: D691–D697.
59. Frolkis A, Knox C, Lim E *et al* (2010). SMPDB: The Snall Molecule Pathway Database. Nucl Acids Res 38: D480-D487.
60. Fukushima A, Kusano M (2013). Recent progress in the development of metabolome databases for plant systems biology. Frontiers Plant Sci 4:1-11.
61. Gao J, Tarcea VG, Karnovsky A, *et al* (2010). MetScape: a Cytoscape plug-in for visualizing and interpreting metabolomic data in the context of human metabolic networks. Bioinformatics 26: 971–973.
62. Giacomoni, F, Le Corguillé G, Monsoor M *et al* (2015). Workflow4Metabolomics: A collaborative research infrastructure for computational metabolomics. Bioinformatics 31: 1493–1495.
63. Haug K, Salek RM, Conesa P *et al* (2013). MetaboLights-an open-access general-purpose repository for metabolomics studies and associated metadata. Nucl Acids Res 41: D781–D786.
64. Horai H, Arita M, Kanaya S, *et al* (2010). Mass-Bank: a public repository for sharing mass spectral data for life sciences. J Mass Spectrom 45: 703–714.
65. Jewison T, Su Y, Disfany FM, *et al* (2014). SMPDB 2.0:big improvements to the small molecule pathway database. Nucl Acids Res 42: D478–D484.
66. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M (2012). KEGG for integration and interpretation of large-scale molecular datasets. Nucl Acids Res 40: D109–D114.
67. Kopka J, Schauer N, Krueger S *et al* (2005). GMD@CSB.DB: the Golm metabolome database. Bioinformatics 21:1635–1638.
68. Kuo TC, Tian TF, Tseng YJ (2013). 3Omics: a web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data. BMC Systems Biology 7:64-78.
69. Salek RM, Haug K, Conesa P *et al* (2013). The MetaboLights repository: curation challenges in metabolomics. Database (Oxford) 2013: bat029
70. Smith CA, O'Maille G, Want EJ *et al*. (2005). METLIN: a metabolite mass spectral database. Ther Drug Monitoring 27:747–751.
71. Steinbeck C, Conesa P, Haug K *et al* (2012). MetaboLights: towards a new COSMOS of metabolomics data management. Metabolomics 8:757–760.
72. Sud ME, Fahy D, Cotter A, Brown EA *et al* (2007). LMSD: LIPID MAPS structure database Nucleic Acids Res 35: D527–D532
73. Tautenhahn R, Cho K, Uritboonthai W (2012). An accelerated workflow for untargeted metabolomics using the METLIN database. Nature Biotechnology 30:826-828
74. Whetzel PL, Noy NF, Shah NH, etal. (2011). BioPortal:enhanced functionality via new Web services from the national center for biomedical ontology to access and use ontologies in software applications. Nucl Acids Res. 39: W541–W545.
75. Wishart DS, JewisonT, Guo AC *et al* (2013). HMDB 3.0– the human metabolome database in 2013. Nucl Acids Res 41: D801–D807.
76. Xia J, Fjell CD, Mayer ML, Pena OM *et al* (2013). INMEX--a web-based tool for integrative meta-analysis of expression data. Nucleic Acids Res 41:W63-W70.
77. Yao Y, Sun T, Wang T (2015). Analysis of Metabolomics Datasets with High-Performance Computing and Metabolite Atlases. Metabolites *5*: 431-442
78. Zhou B, Wang J, Ressom HW (2012). MetaboSearch: Tool for Mass-based metabolite Identification Using Multiple Databases. PLoS ONE 7(6):e40096
79. Kessler N, Neuweger H, Bonte A *et al* (2013). MeltDB 2.0-advances of the metabolomics software system. Bioinformatics. 29(19):2452-9
80. Cui Q, Lewis IA, Hegeman AD, *et al* (2008). Metabolite identification via the Madison Metabolomics Consortium Database, *Nature Biotechnology*, 26:162-170.