9

Cheminformatics and Computational Approaches in Metabolomics

Marco Fernandes[1,2] • Bela Sanches[3] • Holger Husi[2,4]

[1]Department of Psychiatry, Warneford Hospital, Translational Neuroscience and Dementia Research, Oxford University, Oxford, UK; [2]Institute of Cardiovascular and Medical Sciences, BHF Glasgow Cardiovascular Research Centre, University of Glasgow, Glasgow, UK; [3]Strathclyde Institute of Pharmacy & Biomedical Sciences (SIPBS), University of Strathclyde, Glasgow, UK; [4]Division of Biomedical Sciences, Centre for Health Science, University of Highlands and Islands, Inverness, UK

Author for correspondence: Holger Husi, Division of Biomedical Sciences, Centre for Health Science, University of Highlands and Islands, Inverness, United Kingdom.
Email: Holger.Husi@uhi.ac.uk

Doi: http://dx.doi.org/10.15586/computationalbiology.2019.ch9

Abstract: Metabolomics can be viewed as an evolved form of chemical analysis, which required an early instrumental revolution in which the technological core of spectroscopy and spectrometry was developed. This was followed by the advent of high-throughput and high-performance liquid chromatography, together with the establishment of compound libraries and database systems. The ease in the use of metabolomics platforms was coupled with an implementation of data mining methods and bioinformatics tools using machine learning approaches. Cheminformatics makes use of software packages and tools to convey workflows and to streamline data analysis. On the other hand, computational biology offers the contextual approach to the functional characterization of metabolite profiles from a dataset, providing ontologies and annotations. In this chapter, we discuss the main technical procedures used in metabolomics data acquisition, data processing and pipelines, followed by data mining and statistical approaches including machine learning, and ultimately how metabolomics data can aid in elucidating aberrant pathways and metabolic dysfunctions in disease.

Keywords: cheminformatics; computational biology; functional annotation; machine learning; metabolomics

In: *Computational Biology.* Holger Husi (Editor), Codon Publications, Brisbane, Australia. ISBN: 978-0-9944381-9-5; Doi: http://dx.doi.org/10.15586/computationalbiology.2019

INTRODUCTION

The metabolome is the genome's final product, which is defined as the total quantitative group of small molecular weight compounds (metabolites) present in a cell or organism that is involved in metabolic reactions (1). Metabolites are small molecules that are chemically transformed during metabolism, providing functional information of the cellular state, which serves as direct signatures of biochemical activity. Therefore, they are easy to correlate with phenotypes when compared to genes and proteins, whose function is subject to epigenetic regulation and post-translational modifications, respectively (2). Metabolomics (Figure 1) is part of the omics strategies (genomics, proteomics and transcriptomics) that aim to describe the metabolome qualitatively and quantitatively by applying various analytical platforms and methods (3).

Metabolomics combines analytical chemistry strategies and is based on several technological platforms such as mass spectrometry (MS) and nuclear magnetic resonance (NMR) spectroscopy with streamline data analysis (4). In metabolomics, the choice of platforms and techniques is less evident, whereas in genomics and proteomics this appears to be more intuitive to implement for a given study, such as the use of next-generation sequencing (NGS) or microarrays and in-gel or in-solution MS, respectively (1). Nevertheless, MS and NMR are usually the preferred choices for metabolome investigations (5). Data generated through these acquisition platforms need to be further processed using different open-source software or commercial software, such as MZmine (6), Mnova, MetAlign (7), MathDAMP (8), MS-DIAL (9), and XCMS (10) (Table 1). The software can be jointly used with other online or commercially available libraries and databases, depending on the purpose of the study, like the Dictionary of Natural Products (DNP), ChemSpider (11), MarinLit (12), or in-house/custom databases, to

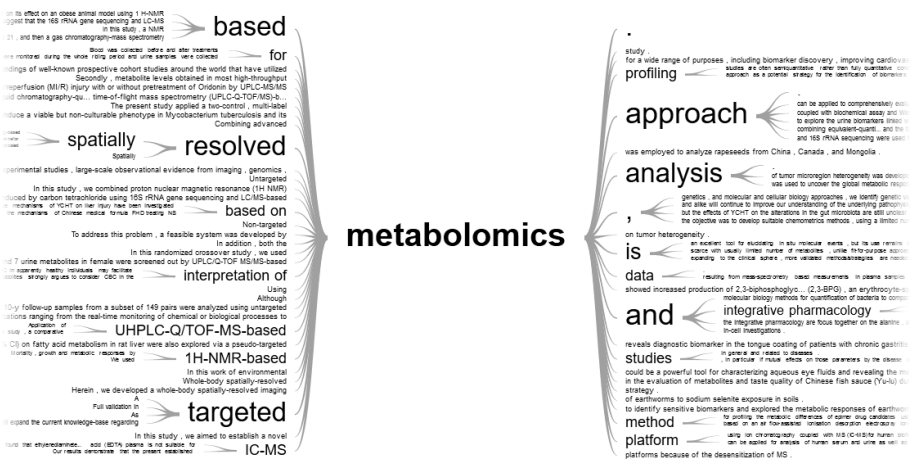

Figure 1 Tree mapping of the most frequent terms in the metabolomics field. Data mining from abstracts indexed in PubMed using as primary key word – "metabolomics" in "Pub-tree" available at https://esperr.github.io/pub-trees/.

TABLE 1 **Software solutions for acquisition and pre-processing data across metabolomics platforms**

Software package	Selected features	Platform	Distribution	Ref.
MS-DIAL	Built-in DIA analysis, annotation and visualization	GC/LC/MS	open-source	(9)
XCMS	User-friendly, retention time correction, statistical analysis	LC/MS		(10)
MZmine2	Batch mode, deconvolution, statistical analysis, visualization	LC/MS		(45)
Mnova	Single suite for processing and visualization	NMR, GC/LC/MS	commercial	—
speaq 2.0	Peak picking and grouping; multivariate statistical functions	1D NMR	open-source	(47)
MetaboAnalyst	Modules for integrative data analysis	NMR, GC/LC/MS		(48)
rDolphin	Enhances ROI by estimation of baseline and signal parameters to maximize fitting of the signals	1H-NMR		(49)
BATMAN	Concentration estimates for known compounds from raw spectra	NMR		(50)
rNMR	Visualisation of NMR signals from multiple spectra concurrently by assigning chemical shift ranges	NMR		(51)

ROI, regions of interest; NMR, nuclear magnetic resonance; IR, infrared Raman; XRF, X-ray fluorescence; DIA, data independent acquisition.

identify secondary metabolites based in the information of the chemical structure of known natural products. Accordingly, the processed data are further subjected to multivariate statistical analysis applying, for example, soft independent modeling by class analogy (SIMCA), which uses unsupervised clustering such as partial component analysis (PCA) or supervised clustering like orthogonal partial least squares discriminant analysis (OPLS-DA), to provide information on the putative bioactive metabolite at the first fractionation step or detect putative biomarkers in a cellular process (13).

Screening for new compounds of pharmacological interest for a specific disease or a disease class has a long history of success cases. For instance, the use of high-throughput screening (HTS) methods for early-stage drug discovery directly yielded cyclosporin A (14), a fungal-derived immunosuppressant medication, and mevastatin, a mold-derived agent, used to normalize cholesterol levels (15). Likewise, drug discovery using structure-based drug design (SBDD) led to the development of new drug candidates such asdorzolamide (16), which is a topical ophthalmic agent applied in the treatment of glaucoma. This method was also

used to develop imatinib, a cancer chemotherapy agent for specific treatment of many leukemia subtypes. Other examples include vemurafenib, a BRAF inhibitor used as chemotherapeutic agent in late stages of melanoma (17). Although it becomes apparent that an ideal workflow for earlier drug discovery should rely on a whole range of tools, from detection and analytical platforms, used either coupled or in parallel, through to computational and statistical steps (Figure 2). This will not only assist in the investigation of novel compounds, but accelerate the discovery stage or even to boost drug repurposing programs (18). This becomes even more apparent when costs are factored in, since the development of a new drug, from target identification to the availability of a final product including approval for prescription to the general public by a governmental or local state authority, involves a multi-step procedure, which can easily take around 12 to 15 years, and is associated with extremely high costs for companies (19). This process starts with basic research that includes lead identification, synthesis scale-up and in vitro pharmacology. This is followed by preclinical development which includes assessing of in vitro toxicity and measuring specific activities by conducting studies of absorption, distribution, excretion and metabolism, and activity of relevant enzymes (20). Alternatively, if the lead target such as a protein is known and its 3D-structure has been elucidated, in silico approaches to predict drug–enzyme interactions can be pursued, using docking algorithms and other well-established computational structure-based approaches (21).

METABOLOMICS DATA ACQUISITION AND PRE-PROCESSING

This section gives an overview of common detection methodologies in metabolomics, conversion of machine data to spectral files and mapping to both known and putative libraries, and ultimately construction of discovery matrices allowing peak-metabolite pairing and quantitative measures. Acquiring raw data from metabolomics analytical platforms and their conversion to extracted data, such as peak lists and spectral bins, requires specific software packages that in many cases need proprietary licenses that are often tied to the platform manufacturer (Table 1).

Mass spectrometry analytical platform

Mass spectrometry (MS) is the analytical technique of choice in metabolomics for identification and/or quantification of varied classes of metabolites, consisting in the production of gas-phase ions that are then detected and characterized by their mass and charge (22). Basically, a mass spectrometer consists of a sample inlet, an ion source, a mass analyzer and a detector and, in that order, functions by introducing the sample into the mass spectrometer, generates gas-phase ions via an ionization technique, separates the ions according to their mass-to-charge ratio (m/z) and generates an electric current from the incident ions that is proportional to their abundances (22). Moreover, the combination of separation techniques such as gas chromatography (GC), high performance liquid chromatography (HPLC), and capillary electrophoresis (CE) allows improved metabolite identification and quantification by MS, which is particularly beneficial when dealing with complex biological samples (5). The recent introduction of a reengineered

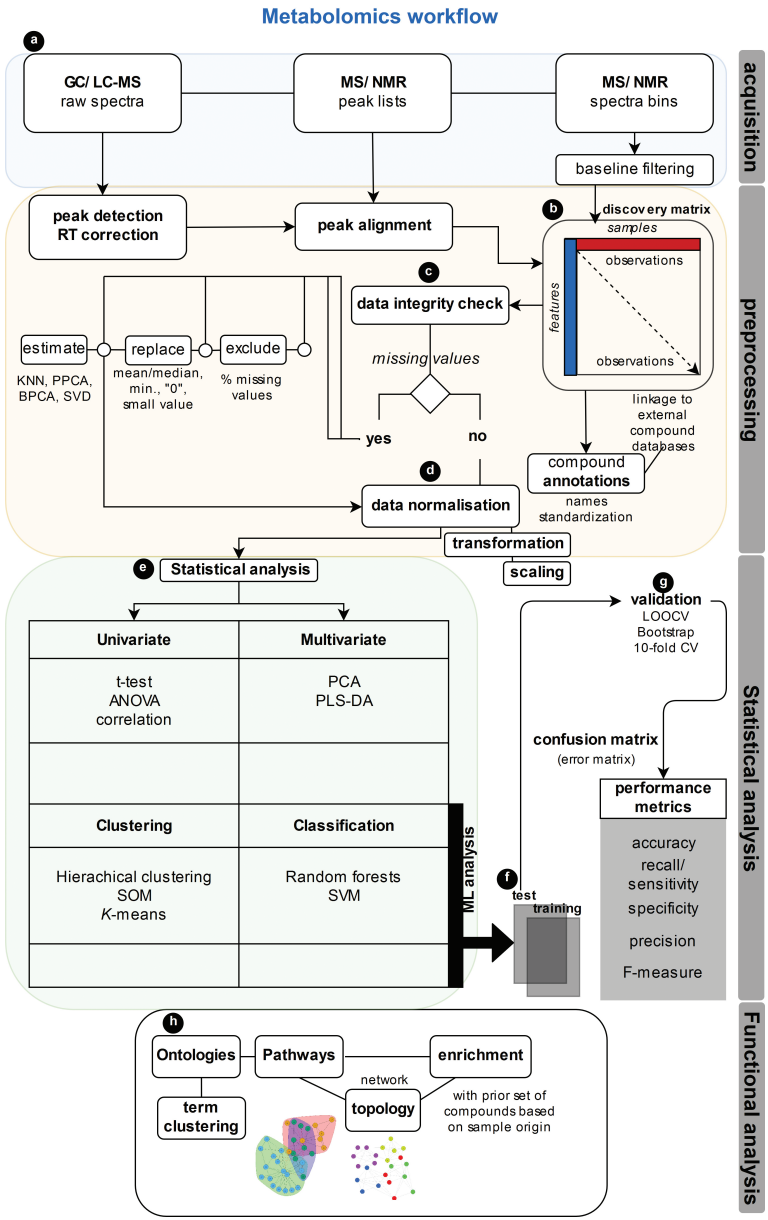

Figure 2 Metabolomics workflows. Data acquisition (a), pre-processing steps including discovery matrix generation (b), data integrity check (c) and data normalisation (d). Followed by statistical analysis (e), machine learning (ML) approaches (f) and validation based of randomisation (g), and functional analysis (h). **Abbreviations**: Gas Chromatography (GC), Liquid Chromatography (LC), Mass Spectrometry (MS), Nuclear Magnetic Resonance (NMR), Leave-One-Out Cross Validation (LOOCV), *n*-times/fold Cross Validation (CV), k-nearest neighbours (KNN), probabilistic principal component analysis (PPCA), Bayesian principal component analysis (BPCA), singular value decomposition (SVD), analysis of variance (ANOVA) , Principal Component Analysis (PCA), Partial Least Squares (PLS).

chromatographic technology such as ultra-high-pressure liquid chromatography (UHPLC) has led to enhanced resolution, higher throughput, lower running times and better cost-effectiveness than traditional HPLC. The use of MS in metabolomics has important advantages such as requiring small sample volumes and provides highly sensitive detection and metabolite identification via interpretation of the spectra and comparison of molecular formula determination via precise mass measurements (23). Additionally, MS is also destructive, and therefore an analyzed sample is not recoverable, and is a relatively slow detection methodology, unlike NMR spectroscopy (23).

Nuclear magnetic resonance analytical platform

NMR spectroscopy is a widely used technique for metabolomics studies with many benefits, such as being specific and at the same time non-selective and non-destructive, and requires no separation or derivatization, is fast and offers highly reproducible and quantitative analyses (1). ANMR spectrum is specific and unique to each compound and provides valuable structural information about the components of the analyzed sample. It combines the information of chemical shift (the nature of the chemical environment), signal multiplicities (neighboring signals), homonuclear and heteronuclear coupling constants, integrals of the signals (number of protons), spin–spin coupling (number and nature of neighbors and connectivity information), and relaxation or diffusion (size of molecule and large-scale environment of location) (24). Although one-dimensional (1D) proton (H) and carbon (C) NMR is one of the most used modes, currently alternative techniques are available, offering additional chemical and structural information, since, in some cases, 1H and 13C NMR are insufficient to provide enough information to entirely characterize metabolites (5) and resolve their identity. To complement the 1D experiments, it is possible to perform two-dimensional (2D) correlation spectroscopy such as 1H-1H COSY, 1H-13C HMBC, 1H-13C HMQC, 1H-13C HSQC,1H-1H ROESY, and 1H-1H NOESY, which enables the elucidation of complex structures. Additionally, samples can be reused, since this technique is non-destructive and does not require pre-selection of analysis conditions like ion source, which is a pre-requisite of MS, or chromatographic operating conditions such as stationary phase, mobile phase, and temperature (1).

Metabolite identification strategies, libraries and algorithms

The metabolomics field has been evolving according to the need for chemical characterization of the composition of biological matrices and extracts from a diverse range of organisms. A fundamental task and simultaneously one of the major bottlenecks in many research areas that use metabolomics workflows is to accurately identify unknown small molecules from the MS and NMR spectra data. Therefore, libraries containing reference spectra with peak assignment to metabolites from previous experiments are being collated and maintained in spectral and compound databases. NMR-based spectral databases are SDBS (13C-NMR, ESR and Raman spectra) (25) (13C-NMR, ESR and Raman spectra), BioMagResBank (26), NMRShiftDB2 (27) and The Birmingham Metabolite Library Nuclear Magnetic Resonance database (BML-NMR) (28). On the other hand, MS-based spectral databases consist of METLIN (29), NIST (30), GMD (31) and MassBank (32).

The Madison Metabolomics Consortium Database (MMCD) (33), The Human Metabolome Database (HMDB) (34) and MetaboLights (35) cover both MS and NMR spectra. Splitting by analytical platform and type of content, either selecting only by spectral data or selecting only by compound annotations, is rather conceptual, since many "modern" metabolomics databases aim to implement both contents in an integrative way. Despite the steady increase in the number of metabolite identities across databases, many cannot be detected through this strategy of database matching due to the absence of their spectral information. Conventional approaches for the identification of these unknowns require reduction of sample complexity by successive steps of fractionation, in order to isolate the target metabolite or compound from the complex mixture, which poses several technical challenges and is highly time-consuming. However, it often does not guarantee identification of low-abundance metabolites via NMR or other spectroscopic techniques (36). Instead, either using the raw or crude sample mixture or even partial sample fractionations can achieve elucidation of the metabolite structure. Then software implementations such as MetFrag2 (37) and CSI:FingerID (38) are available, where MS2 (MS/MS) LC-MS/MS spectra of an unknown experimental metabolite is compared with the in silico generated MS2 fragmentation spectra of putative metabolite structures to find a best match. Other approaches include the use of NMR chemical shifts, in a straight analogy with the previously mentioned strategy, comparing in this case the deconvoluted experimental chemical shifts of unknown metabolites with predictions to yield a best match, where deconvolution is a process to remove instrument-specific signal distortions (39). Recently, the possibility to perform joint analysis with complementary platforms such as NMR and MS was suggested to solve the current paradigm of identification of unknown metabolites (40). Hybrid strategies, such as the SUMMIT MS/NMR (41), primarily resolves all the chemical formulas of the sample detected in the MS1 spectra and then generates all the possible structure permutations. This follows a prediction of NMR chemical shifts for each structural rearrangement and comparison with experimental records acquired to consistently identify molecular structures from both platforms. Other groups used oversimplistic approaches by correlating signal intensities from peak lists from NMR and LC-MS data as proof of principle for the identification of individual metabolites in a sample (42).

Data pre-processing

This step aims to generate a matrix that typically comprehends features (rows) and samples (columns) with each pair coding for an observation from primary raw data. Here, the analysis cascade usually is performed in a stepwise manner and also involves other pre-processing workflows for quality control (QC) dependent on the nature of the acquisition platform, for instance, deconvolution of overlapping peaks, peak picking, integration and alignment (43).

One of the initial steps in the analysis of mass spectrometry data is to convert the vendor-specific binary files to an open or universal format. Thus, LC-MS raw data can be split by ionization mode (positive and negative) using, for instance, the ProteoWizardmsConvert tool (44), and then imported and processed using the open-source MZmine2 (45) toolbox or other software solutions displayed in Table

1. MZmine2 can carry out peak detection, alignment, deconvolution (decomposing overlapping peaks), peak picking and deisotoping, filtering (e.g., removing low-intensity peaks) and gap-filling when, for instance, peaks were detected in some runs or scans but not in others. Additionally, this allows the prediction of putative molecular formulas for each feature by minimizing mis-assignment of features by stepwise removing adducts and complexes (45). This is followed by verifying how novel the "new" compound is by applying dereplication methods, which are particularly relevant for the discovery of new compounds derived from natural product metabolomic data, since it filters from the analysis all the known ones (46). Similar approaches can be found in subtractive and differential genome analysis. DEREPLICATOR+is such an improved algorithm for the dereplication task of core importance in natural products discovery (46).This algorithm assembles theoretical spectra of peptides from non-ribosomal peptide synthetases and ribosomally synthesized post-translationally modified peptide synthetases by first generating a decoy database of peptidic natural products. It then builds predicted spectra for all peptidic products within the database, thereby generating and attributing a score for each peptide and associated spectrum matches, calculating P-values and correction for multiple testing using false discovery rate for the former pairs matching and infer the initial seed of peptidic products by spectral network approaches. Customized libraries with relevant peptidic products can be created by applying dereplication algorithms and further explored or reused by coupling with state-of-the-art software toolboxes such as MZmine2.

On the other hand, the acquired NMR data can be processed with the commercial MestReNova (Mnova) software to confirm and elucidate chemical structures. The 1D and 1H spectrums are processed using the following steps: The baseline is corrected by manual phasing and by using the Whittaker Smoother, and Gaussian is set to 1 Hz for apodization. The chemical shifts are given in ppm and the coupling constants are given in Hz. Chemical shifts in ppm are used to generate the unique primary ID while there are no other secondary IDs considered. It is possible to add the integral number that gives information about the number of hydrogens present in the structure and the multiplicity indicating the neighboring number of hydrogens, thereby allowing a positive assignment of measured data and structure information.

DATA MINING APPROACHES, STATISTICAL ANALYSIS AND ML METHODS

This section will give a brief description of some ML algorithms and performance metrics with examples from the literature of their implementation in the analysis workflow of metabolomics datasets. This includes the initial use of dimensionality reduction methods for visual inspection or data summarization tasks, additional feature selection through filtering metabolites that show higher variability across samples and further computational downstream analysis (Table 2). The popularity and choice of ML algorithms is highly dependent on the domain of science, availability, computational cost, model complexity and interpretability. The eternal model trade-off between "too simple," yet highly biased, and "too complex," yet

TABLE 2 Machine learning methods and algorithms

Class	Description	Implementation/toolbox	Weka	KNIME	TensorFlow	Caret
Association rule learning algorithms	Rules extraction to explain variables association	Apriori and Eclat algorithms	+++	++	+	+++
Artificial neural network algorithms including deep learning	Neural networks construction	Perceptron, back-propagation, Hopfield network, [a]RBFN, CNN, stacked auto-encoders	++	++	+++	++
Bayesian algorithms	Bayes' theorem for classification and regression problems	Naive Bayes, Gaussian Naive Bayes, Bayesian Network, McMC	+++	++	+++	+++
Dimensionality reduction	Unsupervised and supervised approaches to resolve multidimensional data structures	[b]PCA, CCA, PLS, OPLS, MDS, LDA, MDA, QDA, FDA	+++	+++	+++	+++
Ensemble algorithms	Composite of multiple models trained independently in which their individual predictions are fused to yield enhanced overall predictions	Boosting, bootstrapped aggregation (bagging), AdaBoost, stacked generalization (blending), [c]GBM, GBRT, random forests (RF)	+++	++	+++	+++
Decision tree	Trained on the data for classification and regression problems providing a flowchart-like structure model where nodes denote tests on an attribute with each branch represents outcome of a test and each leaf node holds a class label	Classification and regression tree, C4.5 and C5.0, decision stump, regression tree	+++	++	+	++
Regularization	Penalization measures to convey simple models	[d]LASSO, ridge, elastic net	+++	++	++	
Instance based	Comparison of test samples with train samples	[e]kNN, SOM, SVM	+++	++	+++	
Regression	Model relationship between features and sample, error as measure	[f]OLSR, LOESS, linear regression	+++	+++	+++	

Standalone software or analysis framework solutions are available (Weka (W), KNIME (K), TensorFlow (T) library and Caret R package) and can perform most of algorithmic tasks described. Natively supports (+++), supports with add-ons/plugins or extensions (++), or not available or poorly described (+).

[a]Radial Basis Function Network (RBFN), Deep Boltzmann Machine (DBM), Deep Belief Networks (DBN), Convolutional Neural Network (CNN).

[b]Principal Component Analysis (PCA), Canonical Correspondence Analysis (CCA),Partial Least Squares (PLS),Orthogonal PLS (OPLS), Multidimensional Scaling (MDS), Linear Discriminant Analysis (LDA),Mixture Discriminant Analysis (MDA), Quadratic Discriminant Analysis (QDA),Flexible Discriminant Analysis (FDA).

[c]Gradient Boosting Machines (GBM), Regression Trees (GBRT), random forests (RF).

[d]Least Absolute Shrinkage and Selection Operator; (LASSO).

[e]k-Nearest Neighbors (kNN), Self-Organizing Map (SOM), Support Vector Machine (SVM).

[f]Ordinary Least Squares Regression (OLSR), Locally Estimated Scatterplot Smoothing (LOESS).

highly variable, is a core concept in statistics and ML. Standard ML performance metrics such as area under the curve (AUC) are derived from receiver operating characteristic curves (ROC), R2/Q2 ratios, and *k*-fold cross-validation. This also includes concepts like sensitivity, the ratio of the proportion of true positives and the sum of the proportion of false negatives and true positives, which in medical sciences could be interpreted as the proportion of individuals with disease whose test is positive. This is in contrast to specificity, the ratio of the proportion of true negatives and the sum of the proportion of true negatives and false positives is the proportion of individuals without disease whose test returned negative.

Dimensionality reduction and multivariate analysis

Today, an extensive variety of statistical methods is available, ranging from unsupervised methods, such as principal components analysis (PCA), or hierarchical clustering (HCA) to supervised methodologies like partial least squares (PLS), partial least squares discriminant analysis (PLS-DA) and orthogonal partial least squares discriminant analysis (OPLS-DA) (52). Processed MS and NMR data usually are in the form of a matrix of signal intensities signal origins, and, since both are in the same format, it is possible to apply standard analysis techniques to both (53). The first step in metabolomics data analysis is using PCA as an initial exploratory and visualization method that gives an overview of the variability of the dataset as the samples are grouped based on similarity or differences within the group of samples. This enables the detection of trends, groups and outliers, and it is possible to visualize the data as a score plot and a loading plot. In the score plot, each point represents an individual sample, while the loading plot gives information about which variables have the greatest contribution to the positioning of the samples on the scores plot and are responsible for the clustering of samples (24). PCA analysis is followed by supervised pattern recognition techniques, which applies class information of the samples to maximize the separation between different groups of samples and detect the metabolic signatures that contribute to the classifications (24). OPLS-DA is the most used supervised methodology, which has the same predictive power as PLS but gives better interpretation of the relevant variables. This methodology provides information about the causes for class separation (54). In metabolomics, most of the analysis workflows are bespoke procedures, thus requiring implementation of individual software solutions for a given task. For instance, MS-derived MZmine Ids can be combined with ionization mode (positive and negative) to generate a unique primary ID, while other variables like retention time (RT), m/z and molecular weight (MW) should be considered as secondary IDs. Then, using OPLS-DA to compare among groups, it is possible to discriminate and rank metabolites according to their variable importance in projection (VIP) value, ranging from 0 to 1. This is achieved by applying Pareto scaling, which is similar to autoscaling (55), and models can be validated based on multiple correlation (R^2) and cross-validation (Q^2) coefficients as well as by permutation tests for the supervised method.

Kernel methods

Support vector machine (SVM) is the best well-known classification algorithm within machine learning kernel methods, which is the gathering of kernel functions able to map any two points in the initial space representation based on the distances between them into the new space representation, avoiding the computational burden to compute all data point coordinates into the new space. SVM is broadly applied to many classification problems, and a boost in its use was observed with the rise of omics high-throughput data since in most setups it performs well with multidimensional and noisy data. Conceptually, it aims at solving classification problems by finding optimal decision margins between two sets of points belonging to two distinct categories. A decision margin can be described as a line on a surface separating training data into two spaces corresponding to two categories. The classification of new data points is to verify which side of the decision margin they fall on. The data are mapped to a new high-dimensional representation where the decision margin can be expressed as a hyperplane, which is a straight line in any case of dealing with only two dimensions. An optimal decision margin is computed by trying to maximize the distance between the hyperplane and the nearby data points from each class, a procedure named margin maximization, which allows generalization to new samples outside of the training dataset (56). Thereby, data points nearby the maximum margin hyperplane that sit on the margin are so-called support vectors. SVM is a good generalization classifier and has shown good performance using metabolomics data. For instance, Mahadevan et al. (57) did show that SVM gives better predictive models for diagnosis of pneumonia among individuals based on NMR spectral data measured in urine, yielding a classification accuracy greater than 99% using only 30 features selected via recursive feature elimination (RFE). On the other hand, traditional PLS-DA achieved >98% accuracy using 50 features ranked by VIP score. Others built classifiers using SVM with LOO cross-validation for the diagnostic purpose of ovarian cancer with an accuracy superior to 90% using LC/TOF-MS metabolic data detected in serum samples (58). Similarly, using ultra performance (UP) liquid chromatography (LC) with tandem MS for the detection of serum metabolites in early-stage ovarian cancer, the authors claim that using only 16 features selected by SVM-RFE, they are able to discriminate early ovarian cancer (N=46) from healthy controls (N=49) with perfect performance metrics in accuracy, sensitivity and specificity (59).

Ensemble algorithms, decision trees and random forests

Popular ensemble algorithms are bagging and boosting. The first trains each unconstrained model in parallel and the latter trains constrained models in series, learning from the previous ones, and thus evolving overtime. In ML, random forests (RF) (60) is a widely used ensemble algorithm that combines the output of multiple randomly generated decision trees into a composite averaged tree model. RF is applied in many domains of science in classification and regression tasks since it is easy to train and does not require complex tuning adjustments. Additionally, RF yields accurate and robust predictions and is recognized to be less prone to over fitting, a term used to describe the generation of a statistical

model that fits too well to the test or investigation data and fails subsequently in fitting subsequent data, since the rise in the number of each independent randomized tree in an ensemble model would be less likely to increase the generalization error (60). In metabolomics, RF has proven its value in many classification tasks, for instance, by building classifiers to distinguish colorectal cancer (CRC) patients and healthy individuals, as well as pre-surgical against post-surgical CRC patients based on the GC-MS measured urinary metabolome (61). After evaluation of the classification performance, RF, compared with LDA, SVM and PLS via AUC, R2/Q2 and 10-fold cross-validation, outperformed in all of those metrics. Ranking each metabolite through the RF Gini score, and further selection of those with a score >50, yielded, among others, homovanillate and lysine, which are able to discriminate healthy and CRC cases in an early-stage discovery study. Other examples of applicability of this ML algorithm are the development of classifiers able to discriminate among a large set of individuals infected with Zika virus with a specificity and sensitivity over 95% through the use of previously built RF classifiers containing 42 spectral signatures measured in blood using high-resolution mass spectrometry (62).

Functional annotation and biological interpretation of metabolomic data

At this stage, it is expected that a set of compounds or metabolites are identified in at least one chemical database. This simplifies further analysis since most of the available functional and enrichment analysis tools require different database identifiers. Thereby, once identified in any database, it becomes relatively trivial to cross-map compounds to other databases. Additionally, if information of sample concentration or expression is known and allows comparison across sample groups, for example, case versus control, this should be incorporated in the analysis. After having generated a list or matrix with annotated metabolites or compounds and their expression, concentration or ratio metric quantitative values, one can perform enrichment analysis, over-representation analysis, topology-based pathway analysis and activity profiling within pathways. This can be accomplished by using KEGG mapper web server functionality (https://www.genome.jp/kegg/mapper.html). However, this requires that the input is KEGG accession IDs that can be converted from chemical names using web solutions such as the CTS (https://cts.fiehnlab.ucdavis.edu) or MetaboAnalyst (48) ID converter functionality. The final output of the analysis however is only a list of pathways with the number of "hits" found. For a more formal statistical determination of pathway importance modules, Metabo Analyst can be used for enrichment or topological pathway analysis (48). Network-based analysis can be performed using Cytoscape (63), a standalone Java application, which provides multidimensional representations of large-scale networks. This platform supports directed, undirected and weighted graphs, filtering functionalities, merging and extensions for searching active sub-networks and pathway modules, and also incorporates a built-in statistical analysis of the network parameters. Several plug-ins are available for specific tasks, such as metabolomics integration with genomics and proteomics which is implemented in the

MetScape app (64). Additionally, Cytoscape allows interfacing with R and Python, which is useful for scaling and automation of tasks. For pathway editing and mapping metabolites or joint integrative analysis with genomics and proteomics, PathVisio (65) enables visualization and pathway statistical inference using firstly BridgeDb (66) to cross map molecular identifiers and then relies on curated collections of pathways from Wiki Pathways (67) and Reactome (68). In this tool, estimation of over-representation of pathways is based on a Z-score statistical procedure under the hypergeometric distribution and a P-value ranking based on a permutation procedure, which compares actual and permuted Z-scores.

CONCLUSION

The metabolomics field is rapidly evolving and appears to be catching up with genomics and proteomics approaches, which are more established in the research for disease biomarkers. Nevertheless, to establish foundations, protocols and standard operating procedures (SOPs), a more detailed evaluation of how to handle missing data is required through the assessment of the effects of imputation of missing values by means of statistical analysis across analytical metabolomics platforms and by the type of biological matrix. Inclusion and integration of other contextual biological counterparts such as genomics and proteomics will support a global overview of the system in study. Currently, matching experimental spectral data requires the query of many individual database resources to enable the best coverage and maximize compound identification. Ideally, those resources should cover both spectral and compound chemical characteristics, along with biological activities aggregated from many sources, and records would preferentially be manual-annotated and corrected to ensure the highest quality. Structural elucidation of new compounds is a complex, challenging and time-consuming task, but computational-assisted tools and algorithms will reduce such burden and potentiate in-line joint analysis of higher dimensional NMR experiments with high-resolution MS to achieve accurate identifications (24). In the years to come, we will undoubtedly see advances in the development of comprehensive metabolite spectral libraries, algorithms and bioinformatics tools for functional characterization and biological interpretation of metabolite profiles, thereby not only improving our understanding of biology and etiology of disease but also having an impact on drug discovery and personalized medical therapies.

Conflict of interest: The authors declare that they have no conflicts of interest with respect to research, authorship and/or publication of this chapter.

REFERENCES

1. Dunn WB, Bailey NJ, Johnson HE. Measuring the metabolome: Current analytical technologies. Analyst. 2005;130(5):606–25. http://dx.doi.org/10.1039/b418288j
2. Patti GJ, Yanes O, Siuzdak G. Innovation: Metabolomics: The apogee of the omics trilogy. Nat Rev Mol Cell Biol. 2012;13(4):263–9. http://dx.doi.org/10.1038/nrm3314
3. Lopes AS, Cruz EC, Sussulini A, Klassen A. Metabolomic strategies involving mass spectrometry combined with liquid and gas chromatography. Adv Exp Med Biol. 2017;965:77–98. http://dx.doi.org/10.1007/978-3-319-47656-8_4
4. Griffiths WJ, Koal T, Wang Y, Kohl M, Enot DP, Deigner HP. Targeted metabolomics for biomarker discovery. Angew Chem Int Ed Engl. 2010;49(32):5426–45. http://dx.doi.org/10.1002/anie.200905579
5. Villas-Boas SG, Mas S, Akesson M, Smedsgaard J, Nielsen J. Mass spectrometry in metabolome analysis. Mass Spectrom Rev. 2005;24(5):613–46. http://dx.doi.org/10.1002/mas.20032
6. Katajamaa M, Oresic M. Processing methods for differential analysis of LC/MS profile data. BMC Bioinformatics. 2005;6:179. http://dx.doi.org/10.1186/1471-2105-6-179
7. Lommen A. MetAlign: Interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing. Anal Chem. 2009;81(8):3079–86. http://dx.doi.org/10.1021/ac900036d
8. Baran R, Kochi H, Saito N, Suematsu M, Soga T, Nishioka T, et al. MathDAMP: A package for differential analysis of metabolite profiles. BMC Bioinformatics. 2006;7:530. http://dx.doi.org/10.1186/1471-2105-7-530
9. Tsugawa H, Cajka T, Kind T, Ma Y, Higgins B, Ikeda K, et al. MS-DIAL: Data-independent MS/MS deconvolution for comprehensive metabolome analysis. Nat Methods. 2015;12(6):523–6. http://dx.doi.org/10.1038/nmeth.3393
10. 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(3):779–87. http://dx.doi.org/10.1021/ac051437y
11. Williams AJ, Tkachenko V, Golotvin S, Kidd R, McCann G. ChemSpider – Building a foundation for the semantic web by hosting a crowd sourced databasing platform for chemistry. J Cheminform. 2010;2(Suppl 1):O16. http://dx.doi.org/10.1186/1758-2946-2-S1-O16
12. Blunt J, Munro M. MarinLit database. Canterbury: University of Canterbury; 2012.
13. Robotti E, Marengo E. Chemometric multivariate tools for candidate biomarker identification: LDA, PLS-DA, SIMCA, ranking-PCA. Methods Mol Biol. 2016;1384:237–67. http://dx.doi.org/10.1007/978-1-4939-3255-9_14
14. Kreis W, Soricelli A. Cyclosporins: Immunosuppressive agents with antitumor activity. Experientia. 1979;35(11):1506–8. http://dx.doi.org/10.1007/BF01962813
15. Brown AG, Smale TC, King TJ, Hasenkamp R, Thompson RH. Crystal and molecular structure of compactin, a new antifungal metabolite from Penicillium brevicompactum. J Chem Soc Perkin 1. 1976;(11):1165–70. http://dx.doi.org/10.1039/p19760001165
16. Sugrue MF, Mallorga P, Schwam H, Baldwin JJ, Ponticello GS. Preclinical studies on L-671,152, a topically effective ocular hypotensive carbonic anhydrase inhibitor. Br J Pharmacol. 1989;98(Suppl):820P. http://dx.doi.org/10.3109/02713689008999600
17. Halaban R, Zhang W, Bacchiocchi A, Cheng E, Parisi F, Ariyan S, et al. PLX4032, a selective BRAF(V600E) kinase inhibitor, activates the ERK pathway and enhances cell migration and proliferation of BRAF melanoma cells. Pigment Cell Melanoma Res. 2010;23(2):190–200. http://dx.doi.org/10.1111/j.1755-148X.2010.00685.x
18. Roy A. Early probe and drug discovery in academia: A minireview. High Throughput. 2018;7(1):pii: E4. http://dx.doi.org/10.3390/ht7010004
19. Mohs RC, Greig NH. Drug discovery and development: Role of basic biological research. Alzheimers Dement (NY). 2017;3(4):651–7. http://dx.doi.org/10.1016/j.trci.2017.10.005
20. Pereira F, Aires-de-Sousa J. Computational methodologies in the exploration of marine natural product leads. Mar Drugs. 2018;16(7):236. http://dx.doi.org/10.3390/md16070236

21. Issa NT, Wathieu H, Ojo A, Byers SW, Dakshanamurthy S. Drug metabolism in preclinical drug development: A survey of the discovery process, toxicology, and computational tools. Curr Drug Metab. 2017;18(6):556–65. http://dx.doi.org/10.2174/1389200218666170316093301
22. Becker S, Kortz L, Helmschrodt C, Thiery J, Ceglarek U. LC-MS-based metabolomics in the clinical laboratory. J Chromatogr B Analyt Technol Biomed Life Sci. 2012;883–884:68–75. http://dx.doi.org/10.1016/j.jchromb.2011.10.018
23. Lindon JC, Nicholson JK, Wilson ID. Directly coupled HPLC-NMR and HPLC-NMR-MS in pharmaceutical research and development. J Chromatogr B Biomed Sci Appl. 2000;748(1):233–58. http://dx.doi.org/10.1016/S0378-4347(00)00320-0
24. Dona AC, Kyriakides M, Scott F, Shephard EA, Varshavi D, Veselkov K, et al. A guide to the identification of metabolites in NMR-based metabonomics/metabolomics experiments. Comput Struct Biotechnol J. 2016;14:135–53. http://dx.doi.org/10.1016/j.csbj.2016.02.005
25. Yamamoto O, Someno K, Wasada N, Hiraishi J, Hayamizu K, Tanabe K, et al. An integrated spectral data base system including IR, MS, ^{1}H-NMR, ^{13}C-NMR, ESR and Raman Spectra. Anal Sci. 1988;4(3):233–9. http://dx.doi.org/10.2116/analsci.4.233
26. Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, et al. BioMagResBank. Nucleic Acids Res. 2008;36(Database issue):D402–8. http://dx.doi.org/10.1093/nar/gkm957
27. Kuhn S, Schlorer NE. Facilitating quality control for spectra assignments of small organic molecules: Nmrshiftdb2 – A free in-house NMR database with integrated LIMS for academic service laboratories. Magn Reson Chem. 2015;53(8):582–9. http://dx.doi.org/10.1002/mrc.4263
28. Ludwig C, Easton JM, Lodi A, Tiziani S, Manzoor SE, Southam AD, et al. Birmingham metabolite library: A publicly accessible database of 1-D 1 H and 2-D 1 H J-resolved NMR spectra of authentic metabolite standards (BML-NMR). Metabolomics. 2012;8(1):8–18. http://dx.doi.org/10.1007/s11306-011-0347-7
29. Smith CA, O'Maille G, Want EJ, Qin C, Trauger SA, Brandon TR, et al. METLIN: A metabolite mass spectral database. Ther Drug Monit. 2005;27(6):747–51. http://dx.doi.org/10.1097/01.ftd.0000179845.53213.39
30. Linstrom PJ, Mallard WG. The NIST Chemistry WebBook: A chemical data resource on the internet. J Chem Eng Data. 2001;46(5):1059–63. http://dx.doi.org/10.1021/je000236i
31. Kopka J, Schauer N, Krueger S, Birkemeyer C, Usadel B, Bergmuller E, et al. GMD@CSB.DB: The golm metabolome database. Bioinformatics. 2005;21(8):1635–8. http://dx.doi.org/10.1093/bioinformatics/bti236
32. Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, Suwa K, et al. MassBank: A public repository for sharing mass spectral data for life sciences. J Mass Spectrom. 2010;45(7):703–14. http://dx.doi.org/10.1002/jms.1777
33. Cui Q, Lewis IA, Hegeman AD, Anderson ME, Li J, Schulte CF, et al. Metabolite identification via the Madison Metabolomics Consortium Database. Nat Biotechnol. 2008;26(2):162–4. http://dx.doi.org/10.1038/nbt0208-162
34. Wishart DS, Feunang YD, Marcu A, Guo AC, Liang K, Vazquez-Fresno R, et al. HMDB 4.0: The human metabolome database for 2018. Nucleic Acids Res. 2018;46(D1):D608–d17. http://dx.doi.org/10.1093/nar/gkx1089
35. Kale NS, Haug K, Conesa P, Jayseelan K, Moreno P, Rocca-Serra P, et al. MetaboLights: An open-access database repository for metabolomics data. Curr Protoc Bioinformatics. 2016;53:14.3.1–8. http://dx.doi.org/10.1002/0471250953.bi1413s53
36. Koehn FE, Carter GT. The evolving role of natural products in drug discovery. Nat Rev Drug Discov. 2005;4(3):206–20. http://dx.doi.org/10.1038/nrd1657
37. Ruttkies C, Schymanski EL, Wolf S, Hollender J, Neumann S. MetFrag relaunched: Incorporating strategies beyond in silico fragmentation. J Cheminform. 2016;8:3. http://dx.doi.org/10.1186/s13321-016-0115-9
38. Duhrkop K, Shen H, Meusel M, Rousu J, Bocker S. Searching molecular structure databases with tandem mass spectra using CSI: FingerID. Proc Natl Acad Sci U S A. 2015;112(41):12580–5. http://dx.doi.org/10.1073/pnas.1509788112
39. Komatsu T, Ohishi R, Shino A, Kikuchi J. Structure and metabolic-flow analysis of molecular complexity in a 13C-labeled tree by 2D and 3D NMR. Angew Chem Int Ed. 2016;55(20):6000–3. http://dx.doi.org/10.1002/anie.201600334

40. Marshall DD, Powers R. Beyond the paradigm: Combining mass spectrometry and nuclear magnetic resonance for metabolomics. Prog Nucl Magn Reson Spectrosc. 2017;100:1–16. http://dx.doi.org/10.1016/j.pnmrs.2017.01.001
41. Bingol K, Bruschweiler-Li L, Yu C, Somogyi A, Zhang F, Bruschweiler R. Metabolomics beyond spectroscopic databases: A combined MS/NMR strategy for the rapid identification of new metabolites in complex mixtures. Anal Chem. 2015;87(7):3864–70. http://dx.doi.org/10.1021/ac504633z
42. Li X, Luo H, Huang T, Xu L, Shi X, Hu K. Statistically correlating NMR spectra and LC-MS data to facilitate the identification of individual metabolites in metabolomics mixtures. Anal Bioanal Chem. 2019;411(7):1301–9. http://dx.doi.org/10.1007/s00216-019-01600-z
43. Bijlsma S, Bobeldijk I, Verheij ER, Ramaker R, Kochhar S, Macdonald IA, et al. Large-scale human metabolomics studies: A strategy for data (pre-) processing and validation. Anal Chem. 2006;78(2):567–74. http://dx.doi.org/10.1021/ac051495j
44. Kessner D, Chambers M, Burke R, Agus D, Mallick P. ProteoWizard: Open source software for rapid proteomics tools development. Bioinformatics. 2008;24(21):2534–6. http://dx.doi.org/10.1093/bioinformatics/btn323
45. Pluskal T, Castillo S, Villar-Briones A, Oresic M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinformatics. 2010;11:395. http://dx.doi.org/10.1186/1471-2105-11-395
46. Mohimani H, Gurevich A, Shlemov A, Mikheenko A, Korobeynikov A, Cao L, et al. Dereplication of microbial metabolites through database search of mass spectra. Nat Commun. 2018;9(1):4035. http://dx.doi.org/10.1038/s41467-018-06082-8
47. Beirnaert C, Meysman P, Vu TN, Hermans N, Apers S, Pieters L, et al. speaq 2.0: A complete workflow for high-throughput 1D NMR spectra processing and quantification. PLoS Comput Biol. 2018;14(3):e1006018. http://dx.doi.org/10.1371/journal.pcbi.1006018
48. Chong J, Soufan O, Li C, Caraus I, Li S, Bourque G, et al. MetaboAnalyst 4.0: Towards more transparent and integrative metabolomics analysis. Nucleic Acids Res. 2018;46(W1):W486–w94. http://dx.doi.org/10.1093/nar/gky310
49. Canueto D, Gomez J, Salek RM, Correig X, Canellas N. rDolphin: A GUI R package for proficient automatic profiling of 1D (1)H-NMR spectra of study datasets. Metabolomics. 2018;14(3):24. http://dx.doi.org/10.1007/s11306-018-1319-y
50. Hao J, Astle W, De Iorio M, Ebbels TM. BATMAN – An R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a Bayesian model. Bioinformatics. 2012;28(15):2088–90. http://dx.doi.org/10.1093/bioinformatics/bts308
51. Lewis IA, Schommer SC, Markley JL. rNMR: Open source software for identifying and quantifying metabolites in NMR spectra. Magn Reson Chem. 2009;47(Suppl 1):S123–6. http://dx.doi.org/10.1002/mrc.2526
52. Shi L, Westerhuis JA, Rosen J, Landberg R, Brunius C. Variable selection and validation in multivariate modelling. Bioinformatics. 2019;35(6):972–80. http://dx.doi.org/10.1093/bioinformatics/bty710
53. Spicer R, Salek RM, Moreno P, Canueto D, Steinbeck C. Navigating freely-available software tools for metabolomics analysis. Metabolomics. 2017;13(9):106. http://dx.doi.org/10.1007/s11306-017-1242-7
54. Westerhuis JA, van Velzen EJ, Hoefsloot HC, Smilde AK. Multivariate paired data analysis: Multilevel PLSDA versus OPLSDA. Metabolomics. 2010;6(1):119–28. http://dx.doi.org/10.1007/s11306-009-0185-z
55. 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. http://dx.doi.org/10.1186/1471-2164-7-142
56. Duan K, Keerthi SS, Poo AN. Evaluation of simple performance measures for tuning SVM hyperparameters. Neurocomputing. 2003;51:41–59. http://dx.doi.org/10.1016/S0925-2312(02)00601-X
57. Mahadevan S, Shah SL, Marrie TJ, Slupsky CM. Analysis of metabolomic data using support vector machines. Anal Chem. 2008;80(19):7562–70. http://dx.doi.org/10.1021/ac800954c
58. Guan W, Zhou M, Hampton CY, Benigno BB, Walker LD, Gray A, et al. Ovarian cancer detection from metabolomic liquid chromatography/mass spectrometry data by support vector machines. BMC Bioinformatics. 2009;10:259. http://dx.doi.org/10.1186/1471-2105-10-259

59. Gaul DA, Mezencev R, Long TQ, Jones CM, Benigno BB, Gray A, et al. Highly-accurate metabolomic detection of early-stage ovarian cancer. Sci Rep. 2015;5:16351. http://dx.doi.org/10.1038/srep16351
60. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32. http://dx.doi.org/10.1023/A:1010933404324
61. Chen T, Cao Y, Zhang Y, Liu J, Bao Y, Wang C, et al. Random forest in clinical metabolomics for phenotypic discrimination and biomarker selection. Evid Based Complement Alternat Med. 2013;2013:298183. http://dx.doi.org/10.1155/2013/298183
62. Melo C, Navarro LC, de Oliveira DN, Guerreiro TM, Lima EO, Delafiori J, et al. A machine learning application based in random forest for integrating mass spectrometry-based metabolomic data: A simple screening method for patients with zika virus. Front Bioeng Biotechnol. 2018;6:31. http://dx.doi.org/10.3389/fbioe.2018.00031
63. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Res. 2003; 13(11):2498–504. http://dx.doi.org/10.1101/gr.1239303
64. Gao J, Tarcea VG, Karnovsky A, Mirel BR, Weymouth TE, Beecher CW, et al. Metscape: A cytoscape plug-in for visualizing and interpreting metabolomic data in the context of human metabolic networks. Bioinformatics. 2010;26(7):971–3. http://dx.doi.org/10.1093/bioinformatics/btq048
65. Fried JY, van Iersel MP, Aladjem MI, Kohn KW, Luna A. PathVisio-faceted search: An exploration tool for multi-dimensional navigation of large pathways. Bioinformatics. 2013;29(11):1465–6. http://dx.doi.org/10.1093/bioinformatics/btt146
66. van Iersel MP, Pico AR, Kelder T, Gao J, Ho I, Hanspers K, et al. The BridgeDb framework: Standardized access to gene, protein and metabolite identifier mapping services. BMC Bioinformatics. 2010;11:5. http://dx.doi.org/10.1186/1471-2105-11-5
67. Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N, et al. WikiPathways: A multifaceted pathway database bridging metabolomics to other omics research. Nucleic Acids Res. 2018;46(D1):D661–d7. http://dx.doi.org/10.1093/nar/gkx1064
68. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al. The reactome pathway knowledgebase. Nucleic Acids Res. 2018;46(D1):D649–d55. http://dx.doi.org/10.1093/nar/gkx1132