Differentiating signals to make biological sense - a guide through databases for MS-based non-targeted metabolomics

Alberto Gil de la Fuente[1,2], Emily Grace Armitage[3,4], Abraham Otero[2], Coral Barbas[1], Joanna Godzien[1*]

1 Centre for Metabolomics and Bioanalysis (CEMBIO), Facultad de Farmacia, Universidad CEU San Pablo, Campus Montepríncipe, Boadilla del Monte, Madrid, Spain
2 Department of Information Technology, Universidad CEU San Pablo, Campus Montepríncipe, Boadilla del Monte, Madrid, Spain
3 Wellcome Trust Centre for Molecular Parasitology, Institute of Infection, Immunity and Inflammation, College of Medical Veterinary and Life Sciences, University of Glasgow, Glasgow, UK
4 Glasgow Polyomics, Wolfson Wohl Cancer Research Centre, College of Medical Veterinary and Life Sciences, University of Glasgow, Glasgow, UK
KEYWORDS: Database, Mediator, Metabolite identification, Metabolomics, Structural elucidation

* Corresponding author
CEMBIO (Centre for Metabolomics and Bioanalysis), Facultad de Farmacia, Universidad CEU San Pablo, Campus Monteprincipe, Boadilla del Monte, 28668 Madrid Spain
Tel: (+34) 913724769, e-mail: joannabarbara.godzien@ceu.es

Number of words: 7434

List of abbreviations:

API: Application Programming Interfaces
BNICE: Biochemical Network Integrated Computational Explorer
CEMBIO: Centre for Metabolomics and Bioanalysis
CFM-ID: Competitive Fragmentation Modelling for Metabolites Identification
CeuMM: CEU Mass Mediator
CSI:FingerID: Compound Structure Identification: FingerID
DRCC: Data Repository and Coordinating Center
EML: Evidence-based Metabolome Library
HMDB: Human Metabolome Database
ID: Identifier
JCBL: Japanese Conference on the Biochemistry of Lipids
KomicMarket : Kazusa Omics Data Market
LipidMaps: LIPID Metabolites and Pathways Strategy
LMPD: LipidMaps Proteome Database
MGP: Metabolome Gene/Protein Database
MINE: Metabolic *In Silico* Network Expansion Databases
MSI: Metabolomics Standards Initiative
MS^n: Multi-Stage Mass Spectrometry
NIMS: Nanostructure Imaging Mass Spectrometry
PEP: Peptides
PIF: Precursor Ion Fingerprinting
ppm: part per million
RefMet: Reference set of Metabolites
REST: REpresentational State Transfer

RT: retention time
SE-CFM: Single Energy CFM
Workbench: UCSD Metabolomics Workbench

ABSTRACT

Metabolite identification is one of the most challenging steps in metabolomics studies and reflects one of the greatest bottlenecks in the entire workflow. The success of this step determines the success of the entire research, therefore the quality at which annotations are given requires special attention. A variety of tools and resources are available to aid metabolite identification or annotation, offering different and often complementary functionalities. In preparation for this article, almost 50 databases were reviewed, from which 17 were selected for discussion, chosen for their on-line ESI-MS functionality. The general characteristics and functions of each database is discussed in turn, considering the advantages and limitations of each along with recommendations for optimal use of each tool, as derived from experiences encountered at the Centre for Metabolomics and Bioanalysis (CEMBIO) in Madrid. These databases were evaluated considering their utility in non-targeted metabolomics, including aspects such as ID assignment, structural assignment and interpretation of results.

INTRODUCTION

The importance of metabolomics and its utility is still increasing, both in terms of the range of applications and their frequency. Amongst the different applications, non-targeted metabolomics plays a vital role, revealing new and unexpected findings that can lead to further research in a particular direction [1-4]. However, the success of this approach highly depends on the possibility to understand and interpret the information hidden within a complex metabolomics dataset. Most metabolomics studies are based on ESI-MS [5-7], usually with a preceding separation step such as liquid chromatography, tending to measure the ratio of mass to charge (*m/z*) and abundance of each ions which originate from chromatographically separated molecules. After data pre-processing and statistical analysis, a list of discriminating signals between sample groups can be obtained [8]. However, to understand the nature of this separation and its cause, masses must be annotated with metabolite identifications, which can be mapped onto biochemical pathways to understand their origins. Metabolite identification is influenced by a range of factors, which should be taken into consideration from the initial experimental design through to the interpretation of results (Figure 1).

To annotate measured masses with metabolite IDs, a data source is needed for comparison. One solution would be to use an in-house library based on the authentic standards analysed under particular conditions. In this way, at least two independent and orthogonal characteristics (e.g. mass and RT) could be used for comparison, providing first, the highest level of identification confidence according to MSI (Metabolomics Standards Initiative) guidelines [9]. This method is rather restrictive though, since only commercially available metabolites can be introduced to the library and used for annotation. New strategies utilising on-line accessible databases that contain a large array of information have emerged to mitigate this shortfall [10-15]. Cross comparison of experimental data

to databases can be performed using only one characteristic (mass) (second level of confidence for MSI) which highlights a limitation compared to using in-house libraries. Nevertheless, the amount of information provided is huge, covering different subclasses and including not only endogenous metabolites but also substances originating from the microbiome, diet, plants or supplementation. Therefore, the coverage of annotations across the data is much more promising. Furthermore, *in silico* predicted compounds are now available, considering biological modifications of known metabolites that may occur under particular conditions [16]. This somehow responds to the clear need to open metabolomics research to consider new or previously unidentified metabolites. Moreover, databases are continuously growing due to the contribution of many researchers.

In 2011 Fiehn and colleagues divided databases into two categories, making a clear distinction between pathway-centric and compound-centric databases [17]. In this review only compound-centric databases are examined, omitting databases such as KEGG (www.genome.jp/kegg), Reactome (www.reactome.org) and Wikipathways (wikipathways.org). Additionally, only on-line, open-access databases are included, omitting commercial resources. Finally, only ESI-MS dedicated resources allowing exact mass searching are assessed. Following these restrictions, 17 data sources were selected for review from a total of 47 considered. For a comprehensive list of those rejected, refer to table 1S (Supplementary Information). Data sources covered in this article are: BioCyc Database Collection (BioCyc) (biocyc.org), Ceu Mass Mediator (CeuMM) (ceumass.eps.uspceu.es), Compound Structure Identification: FingerID (CSI:FingerID) (www.csi-fingerid.org), Human Metabolome Database (HMDB) (www.hmdb.ca), Kazusa Omics Data Market (KomicMarket) (webs2.kazusa.or.jp/komicmarket/index.php), LipidBank (lipidbank.jp), LIPID Metabolites And Pathways Strategy (LipidMaps) (www.lipidmaps.org), MAGMa (www.emetabolomics.org/magma), MassBank (www.massbank.jp), MassTRIX (masstrix3.helmholtz-muenchen.de/masstrix3/), MetFrag (msbi.ipb-halle.de/MetFragBeta), METLIN (metlin.scripps.edu), Metabolic *In Silico* Network Expansion Databases (MINE) (minedatabase.mcs.anl.gov), MycompoundID (www.mycompoundid.org), MzCloud (www.mzcloud.org), MZedDB (maltese.dbs.aber.ac.uk:8888/hrmet/search/addsearch0.php) and UCSD Metabolomics Workbench (Workbench) (www.metabolomicsworkbench.org). The number of compounds contained in each is depicted in Figure 2. Figure 3 illustrates the number of citations of each data source in google scholar, while information on the initial release data and latest updates for each are given in table 2S (supplementary information). All information given on each database is true as of January 15th 2017. It is important to highlight that this review was constructed based not only on literature research but also on usage and revision of databases at the Centre for Metabolomics and Bioanalysis (CEMBIO), Madrid.

Of the data sources reviewed, BioCyc, HMDB, KomicMarket, LipidBank, LipidMaps, MassBank, METLIN, MzCloud and Workbench are considered databases *sensu stricto*. All the other on-line tools reviewed are mediators which use the information provided by databases: CeuMM, CSI:FingerID, MAGMa, MassTRIX, MetFrag, MINE and MZedDB. Detailed information on the sources used by each database and mediator is stated in table 3S (Supplementary Information). Both types of on-line tool are very important for the metabolomics society and both require continued

improvement. Different databases focus on different types of molecules, therefore it is recommended to use a combination of resources for optimal coverage. In this way mediators are advantageous since they perform searches across different sources through a single interface. However, not all mediators offer multi-source usage. For example MAGMa, MetFrag and MINE permit the use of only one source at once. MassTRIX on the other hand searches KEGG, HMDB and LipidMaps together or separately (as defined by the user) and CeuMM permits the search between all combinations of HMDB, KEGG, LipidMaps, Metlin and MINE as required. Within this review, the general characteristics of each of the data sources are detailed, followed by a discussion of functionality to compare and contrast the advantages and limitations of each for different aspects.

GENERAL CHARACTERISATION

This section contains a short description of each database/mediator. Functionalities, advantages and limitations of each database are detailed in table 1.

BioCyc [18], developed by SRI International (Menlo Park, California), is a collection of curated databases for different organisms. Databases are organised according to the level of manual updates they have received. Tier-1 databases such as EcoCyc (for *E. coli*) and HumanCyc are highly curated, while most BioCyc databases (Tier 2 and 3) have been computationally derived. These databases are particularly applicable to organism specific metabolite identification and metabolic reconstructions using the pathway search.

CeuMM (ceumass.eps.uspceu.es), a collaborative development from the CEMBIO and the Bioengineering Laboratory of Polytechnic Faculty at Universidad CEU San Pablo Spain,is a tool which performs an automated search across external data sources (HMDB, KEGG, LipidMaps, METLIN and MINE) and provides possible identifications for a given mass (unifying similar hits given from more than one database into a single hit).

CSI:FingerID [19] is a database specific for MS^n identification. It supports further research on peaks unidentified at the MS level. It is a collaborative development between Friedrich Schiller University, Germany and Helsinki Institute for Information Technology at Aalto University, Finland, that combines fragmentation tree computation and machine learning to improve both the total percentage of identified molecules and the precision of identification.

HMDB [20] is a database devoted to human metabolism developed with support from the Canadian Institutes of Health Research, Alberta Innovates - Health Solutions and The Metabolomics Innovation Centre. For each data entry, information is given on the chemical, biological and clinical characteristics as well as references to the literature including reported disease associations, related enzymes and transporters in addition to links to external databases such as KEGG.

Komic Market (Kazusa Omics Data Market) is a database of metabolite annotations from MS peaks detected in metabolomics studies. It comes from the project “Development of Fundamental Technologies for Controlling the Material Production Process of Plants”, supported by the New Energy and Industrial Technology Development Organisation, Japan.

LipidBank [21] is the official database of the Japanese Conference on the Biochemistry of Lipids (JCBL). This database is devoted to neutral lipids. It covers several different classes and all

molecular information is manually curated and approved by experts in lipid research. Each entry includes a lipid name, molecular structure, spectral information, and literature references.

LipidMaps [22] is funded by a large-scale collaborative research grant ("Glue Grant") from the NIH National Institute of General Medical Sciences. Its aim is to provide identification and quantitation of mammalian lipids including the quantification of changes in response to perturbation. LipidMaps Proteome Database (LMPD) is also included in this resource.

MAGMa [23] is an annotation tool developed within the eMetabolomics project, funded by the Netherlands eScience Center at Wageningen University in collaboration with the Netherlands Metabolomics Centre. MS^n data can be uploaded as a hierarchical tree of fragment peaks, based on *m/z* or chemical formulae and candidate molecules are automatically retrieved from PubChem, KEGG or HMDB. A matching score is calculated based on the quality of explanation of the fragment peaks.

MassBank [24] is a public repository of mass spectral data based on sharing identifications and structure elucidations of chemical compounds detected by mass spectrometry. MassBank is accessible through two domains: Japanese (http://massbank.jp) and European (http://massbank.eu) (NORMAN MassBank). The tool is deployed in both domains, but some functions are only provided in the Japanese one.

MassTRIX [25, 26] is an on-line tool for the annotation of high precision mass spectrometry data. Results are displayed on organism specific KEGG pathway maps and any additional genomic or transcriptomic information can be added. The tool was developed at the Helmholtz Zentrum München in a collaboration between Philippe Schmitt-Kopplin and Karsten Suhre.

MetFrag [27, 28] is a tool designed for *in silico* fragmentation data for computer assisted identification of metabolite mass spectra using general chemical rules based on standard reactions. Its development is concentrated around Leibniz Institute of Plant Biochemistry and Eawag: Swiss Federal Institute for Aquatic Science and Technology. It is currently available through two web pages: MetFrag Web 2010 and the updated MetFrag Web beta. A search can be performed against the listed databases or from a fully customised file, allowing the use of the *in silico* fragmentation function on users own compounds. It provides a score based on the algorithms implemented.

METLIN [29] is a trademark of the Scripps Research Institute, which develops and applies mass spectrometry-based technologies for understanding metabolism. It includes cloud-based data processing informatics (XCMS), and nanostructure imaging mass spectrometry (NIMS). With almost 1,000,000 real compound entries (not from prediction), this is one of the largest databases available. Entries in METLIN include metabolites, lipids, steroids, plant and bacterial metabolites, small peptides and exogenous drug metabolites and toxicants. IsoMETLIN - A module for isotope-based metabolomics is also included.

MINE [16] taps into data sources such as KEGG, EcoCyc, YMDB and Chemical Damage, generating theoretically possible metabolites based on known entities. It does this using an algorithm called the Biochemical Network Integrated Computational Explorer (BNICE) and expert curated reaction rules based on the Enzyme Commission classification system. The tool comes from collaboration between several research centres including Northwestern University, Argonne

National Laboratory, West Coast Metabolomics Center, University of California and Davis and King Abdulaziz University

MycompoundID [30, 31] is a web-based resource developed at the University of Alberta for identification of compounds based on chemical properties including accurate mass. Different searches are possible including MS, MS^2, PEP searches of unlabelled and dimethyl labelled peptides, and chemical isotope labelled MS data. Searches are performed across an evidence-based metabolome library (EML) which consists of 8,021 known human endogenous metabolites and their predicted metabolic products including 375,809 compounds from one metabolic reaction and 10,583,901 from two reactions. *In silico* predicted compounds are generated from HMDB entries.

MzCloud (www.mzcloud.org) is a trademark of HighChem LLC from Slovakia. It is an advanced database of high-resolution MS^n spectra acquired under different conditions that are filtered, recalibrated and arranged into spectral trees. Identification is possible through the Precursor Ion Fingerprinting (PIF) tool that can expand on compounds that are already listed in the database to new metabolites, identified based on substructure information through the comparison of product ion spectra of structurally related compounds. It is also a repository for databases of contributors.

MZedDB [32] is a database for metabolite signal annotation developed by the Aberystwyth University High Resolution Mass Spectrometry Laboratory. It is largely derived from established repositories (aracyc, dico, HMDB, KEGG, lmdb, mammal, metacyc, plant, ricecyc) and performs automated, high throughput analysis of data derived from soft ionisation. It is possible to apply rules about adduct formation and neutral losses to prove or discard certain hits. Also, a molecular formula generator is available for identifying molecules based on chemical formulae.

Workbench [33], developed within the Metabolomics Program's Data Repository and Coordinating Center (DRCC) and sponsored by the Common Fund of the National Institutes of Health, serves as a national and international repository for metabolomics data and metadata, providing analysis tools and access to metabolite standards, protocols, tutorials and training material. MS search for ID assignment is possible using three types of database: a virtual database of lipid classes, a reference set of metabolites (RefMet) and the Metabolomics Workbench Metabolite database (combination of compounds from LipidMaps, ChEBI, HMDB, BMRB, PubChem and KEGG). The Human Metabolome Gene/Protein Database (MGP) is also available.

FUNCTIONALITY

There are some key considerations which determine the applicability of different data sources in the metabolomics workflow. One key consideration is whether or not the resource is freely available (which can differ between academia and industry). Table 2 presents information on the licence and data usage policies for different databases. Additionally, only some tools offer the possibility to save searches or export results that is particularly useful in large-scale or multi-platform studies with a huge number of masses requiring annotation. A summary of these characteristics including the exact information that can be exported using different tools is given in

table 4S (supplementary information). Some on-line tools provide Application Programming Interfaces (APIs). An API is a common language for communication between different computer systems. APIs enable search automation and integration into workflows of third-party metabolomics tools. Galaxy Workflow4metabolomics is an example of a tool where many other external metabolomics tools can be integrated through their APIs [34]. Some databases do not provide APIs (Table 2) and others are now out of service (for example METLIN's API has been out of service since 2011 due to security issues). APIs may be developed in different paradigms and Representational State Transfer (REST) is one example for constructing web services [35, 36]. REST architecture leads to a stateless model where resources can be accessed through primitive methods such as GET or POST. On-line tools which implement this API usually provide resources independently and are not used as methods for performing queries based on experimental masses. APIs can be developed for a specific programming language as shown in table 2.

Another consideration is how user-friendly each resource is. Of course each resource can be more or less useful for a particular purpose and the assessment of each can be highly subjective, however to provide a guide of the main practical aspects of each data source, table 3 summarises design features, asynchronicity (lack of need for a full page reload every time the user performs an action), login requirements and ease of familiarisation for each source.

Due to the range of tools available, global characterisation is challenging without separating them by functionality. Functionality will therefore be discussed under the following classifications: *i)* ID assignment, *ii)* structural assignment and *iii)* data interpretation. ID assignment involves annotation of peaks with known metabolites. Structural assignment includes MS^n information used for structural confirmation or elucidation by matching structural similarity to known compounds on the MS or MS^n level. Data interpretation covers any information useful to understand and interpret results including pathway analysis, literature search, depiction of metabolites and their classification.

***i)* ID assignment**

ID assignment relates the exact mass of a compound detected to the exact mass of a known metabolite in a database (with a given tolerance suitable for the instrument used in data acquisition). It is the only option when there is no more than MS level data available and therefore no structural elucidation can be performed [37]. Of the data sources discussed in this review, the following are suitable for ID assignment: BioCyc, CeuMM, HMDB, LipidMaps, MassBank, MassTRIX, METLIN, MINE, MycompoundID, MZedDB and Metabolomics Workbench. The remainder of this section discusses the features that are deemed as relevant for the ID assignment task; all these features are summarised in Table 2.

Tolerance

In non-targeted metabolomics, identification power is determined by the mass accuracy of the data; databases can provide high precision when masses are recorded to four or more decimals. Databases offer the possibility to set a tolerance either in absolute (Da or mDa) or relative (ppm) terms (table 2). The majority of databases give absolute freedom to establish the tolerance, while MassTRIX, LipidMaps and Workbench define set ranges of tolerance. Each measurement, regardless

of the power of the instrumentation, comes with some inaccuracy. For this reason, it is necessary to establish an appropriate tolerance for each dataset. A good way to decide the tolerance is to assess the error on an internal standard or well-known compound. Choosing whether the tolerance should be absolute or relative is also important. For example, a relative error of 10 ppm on a low molecular weight compound such as choline (MW=104.1075Da) would be in the range ±0.0020Da, while for PC(21:0/22:6) (MW=875.6404Da), 10 ppm would be in the range ±0.0176Da.

Search mode

An important aspect to evaluate databases is whether searches can be performed by batch (multiple masses can be submitted simultaneously) or only single searches are permitted. Manually querying hits mass by mass can be tedious and repetitive if not impractical.

Adducts

During the process of ionisation using ESI, adducts that alter the detected mass of the metabolite can be formed [38]. Working in positive mode, the most common adduct formations are: $[M+H]^+$, $[M+Na]^+$, $[M+NH_4]^+$ and $[M+H-H_2O]^+$ and in negative mode: $[M-H]^-$, $[M+HCOO]^-$, $[M+Cl]^-$ and $[M-H-H_2O]^-$ [39]. A great deal of time can be saved with the option of searching multiple adducts and multimers simultaneously [32]. This is of particular importance for datasets obtained using high sensitivity equipment, where different adducts are detected, even those with very low abundance. This plays an even more relevant role when multi-signals originating from a single molecule are not combined into single values during data reprocessing. On inspection of the data sources, three types of search can be distinguished: neutral mass search only, *m/z* search for a single adduct and *m/z* for multi-adducts. Information on the search mode for each database is presented in table 2 and a detailed list of possible adducts is given in table 5S (Supplementary Information). Lipids are best identified by their *m/z* and applying knowledge about possible ionisation and adduct formations in order to select adequate hits. By ordering these possible hits by RT, different adducts corresponding to the same molecule can easily be identified. Moreover, this method allows the identification of mis-assignments considering the chemical properties and elution order. It is important though, when selecting possible adducts for ID assignment, only to allow those expected to minimise the risk of mis-assignment. Small molecules and acids should be also searched considering possible in-source fragmentation with the most common neutral loss of water [40, 41]. Some databases, for example MZedDB, offer the option to select multi-adducts following a list of defined rules regarding adduct formation [39] (Putative ionisation product tab). These rules were established considering aspects such as the number of particular elements or chemical groups in a molecule (-OH, -COOH, $-NH_2$ etc.), the number of electrons or charges and information on non-covalently bound products and solvents.

Although there are no on-line tools that can combine metabolic features split by multi-adducts, some tools (e.g. METLIN) do offer the option to calculate the mass of different adducts, multimers and charges for any given compound. Similar options are also offered in LipidBank, LipidMaps and Metabolomics Workbench where *m/z* value is given for single adduct. In MZedDB, even when there is no compound listed for an exact mass in the database, the generated chemical formula can be used to predict *m/z* values for different adducts (adduct manipulation tab).

The possibility for batch searching and searching considering multi-adducts are of vital importance when considering the usefulness of a resource. Figure 4 depicts these functions for the different data sources considered in this review.

Exporting options

The purpose of ID assignment can be to provide a quick putative hit for detected masses, or to generate a longer list of options that can be later used in ID confirmation by MS^n analysis. Regardless of the purpose, the list of hits should be easily exportable. Most of the databases offer the possibility to save search results in a chosen format, e.g. csv, xls or sdf. KomicMarket, LipidBank, MassBank, MassTRIX, METLIN, MzCloud, MZedDB and Workbench do not offer automatic data download options for MS searches, thus results must be manually copied from the webpage. Workbench offers the option to save results but only one compound at a time which can render it ineffective for larger datasets.

Filters

The number of hits for any given search mass can be quite high. Careful filtration of this list to reduce the number of plausible hits is required. This filtration is generally performed manually, however CeuMM, CSI:FingerID and MZedDB offer functions to aid this process by restricting hits based on chemical alphabet (a list of elements selected based on expectation in given samples) or by restricting or including halogens and metals in the hits based on expectation. LipidMaps, BioCyc and Workbench offer the alternative option of allowing selection of expected compound classes, (e.g. lipids, carnitines, amino acids) and excluding all other hits in order to filter the number of matches. LipidMaps, by definition, searches only lipids and related compounds, however it is possible to restrict the search to a particular class, category or chemical composition in the ontology section (e.g. considering number of carbons, double bonds, rings or particular functional groups). MzCloud offers a useful list of filter categories (see list 1 in Supplementary Information) to aid both MS and MS^n searches. One relevant possibility is to exclude some compounds from it, an option also present in METLIN. Since most databases were constructed considering utility in human studies [15], the option to restrict certain types of compound can be particularly useful when using databases for different (model) organisms with a more controlled metabolome [42]. Such options are possible in BioCyc, LipidBank and MassTRIX, where the former two use different data sources based on the restrictions and the latter highlights more plausible hits by organism selection in the output.

***In silico* compounds**

Since ID assignment is restricted to available database entries, many experimental masses can be left unannotated after a search. As a solution to this, some databases now include the option to predict compounds *in silico* with the aid of chemical rules or restrictions. Expansion of the known metabolome can be performed using as an example the BNICE framework (Computational framework for predictive biodegradation) with hand-curated reaction rules generalised from chemical theory and literature [16]. LipidMaps and Workbench include a virtual database of lipids created by combining head groups with acyl/alkyl chains, including glycerophospholipids, glycerolipids, sphingolipids, acyl carnitines, acyl CoAs, cholesteryl esters and wax esters. Also a list of virtual fatty acids (OH:hydroxyl, Ke:keto(oxo), Ep:epoxy, cyclo:ring) and cardiolipins is available. Two

mediators: MINE and MycompoundID are open for all types of metabolites, not only lipids, and consider some biotransformation reactions that are known to commonly occur. MycompoundID takes the approach of searching one or two chemical transformations over compounds from HMDB. For example alanine - methylalanine (positively changed in mass), or sphinganine and dehydrosphinganine (negatively changed in mass). The list of possible biotransformations includes 76 positions and is based on literature revision [31]. A similar function is present in MINE, however in contrast to MycompoundID, the search cannot be restricted to just real or predicted compounds and therefore the list of hits is longer and mixed. CeuMM searches the MINE database, restricting the hits to generated compounds only. This is based on API services provided by MINE, but not accessible from MINE's on-line service itself.

***ii)* Structural assignment**

While for some purposes putative identification is sufficient, the majority of researchers require a more defined approach to metabolite identification, especially where potential biomarkers are being proposed. MS^n data is required for this purpose to confirm hits by comparison of a compounds fragmentation pattern relative to MS^n (usually MS^2) spectra in databases, or better still to the fragmentation pattern of the authentic standard analysed under the same experimental conditions. Amongst the databases discussed in this review, ten offer functions related to the use of MS^2 spectra: CSI:FingerID, HMDB, KomicMarket, LipidMaps, MAGMa, MassBank, MetFrag, METLIN, MycompoundID and MzCloud.

MS^2

When comparing experimental fragmentation to spectral resources in databases, it is vital to consider the instrumentation and parameters used in data acquisition, since fragmentation can be highly dependent on both these aspects. For this reason, HMDB, LipidMaps, MassBank and MzCloud are particularly useful given the amount of information available with spectra. The type of mass analyser, tolerance for precursor and product ions, collision energy and ion mode are particularly relevant. A list of experimental *m/z* values (product ions with or without precursor) and corresponding abundances are used to search and compare against relevant spectra in the databases. Depending on the database, the upload of this information can vary, but once uploaded the matching process is similar. Careful experimental design considering the options available in databases can significantly improve the efficiency of metabolite annotation using fragmentation comparison. For example, data are usually acquired using fixed collision energies of 10, 20 and 40 eV; therefore it is sensible to collect data on an unknown compound using one of these thresholds. When data are acquired using a slope for collision energy determination (particularly relevant for very fragile compounds) several different spectra available in the databases should be checked to improve the likelihood of a good match.

LipidMaps and KomicMarket are the only two databases covered that do not contain the option to search against MS^2 spectra. Furthermore, the MS^2 spectra that are present in these databases are often limited by single ion mode or collision energy. However, these databases do

offer alternative useful information. LipidMaps has valuable information on possible ionisation and fragmentation, while KomicMarket contains a huge number of unannotated compounds with information on extraction, measurement and detection including example MS^2 spectra for many entries. HMDB and METLIN in contrast to other MS^2 databases allow determination of collision energy in the search parameters. Of the databases with MS^2 search-match functionality, all except HMDB, MAGMa, MassBank and MzCloud, offer the possibility to determine adducts. Most databases use mirror graphs (HMDB, METLIN, MassBank) to display experimental and database spectral matches, or present the query and library spectra together with the difference spectrum showing exactly which peaks do not match (MzCloud). MassBank offers the very useful option of visualising and comparing several spectra at once, with options to change various display settings.

Another way to evaluate the MS^2 match efficiency is using a score (particularly advantageous when considering multiple hits). HMDB, MAGMa, MassBank, MetFrag, METLIN, MycompoundID and MzCloud all generate scores for this purpose. HMDB presents three scores: Fit, RFit and purity [43]. Fit is calculated comparing the library spectrum to the acquired one and RFit is the opposite. MycompoundID generates scores for fit in explaining product ions. MzCloud generates three scores that correspond to different algorithms useful for structure explanation (HighChem HighRes, Opt.Data Product and NIST(modified)).

MS^n

MS^n (n>2) data can be particularly useful to determine the exact identification of a metabolite that has strong structural similarities with other compounds, often encompassing vastly different biological function. Differences can be as small as a position of a double bond or functional group. Specific analysers are required to generate such data (ion trap, Fourier transform ion cyclotron resonance or orbitrap) and data must later be organised into structural trees illustrating the fragmentation patterns. CSI:FingerID, MassBank MzCloud and MAGMa contain the relevant information to identify molecules in this way. MzCloud supplies a wide variety of filters and options for MS^n searching. Identification can be performed in compound mode through tree search or in substructure mode for subtree search. In MzCloud, spectral comparison at any MS level can be performed on filtered or recalibrated spectra, where results can be additionally filtered based on compound or spectrum (ionisation mode, mass analyser, ion activation, collision energy etc.). The possibility to assign substructures or explain neutral losses is most useful, making MzCloud highly valuable for use with MS^n data. CSI:FingerID and MAGMa follow a different strategy for identification. Fragmentation trees are computed and used to predict the molecular structure fingerprint using a machine learning approach, which can later be searched against structures in PubChem (CSI:FingerID) and/or KEGG or HMDB (MAGMa).

Predicted MS^2

Although new entries are continually made to MS^2 spectral libraries, the number of available standards is restricted and therefore the databases will never be complete. To overcome this, fragmentation prediction can be especially useful. Predicted MS^2 spectra are available in HMDB, MetFrag, METLIN, MycompoundID and MzCloud. Differences in the algorithms used in each do lead to (often relevant) differences in the result and therefore careful analysis is required while using

these functions. HMDB and METLIN predict spectra using Competitive Fragmentation Modelling for Metabolites Identification (CFM-ID), a method that learns and generates models of collision-induced dissociation (CID) fragmentation from data (cfmid.wishartlab.com/). In single energy CFM (SE-CFM) [44], ESI- MS^2 fragmentation is modelled as a stochastic, homogeneous, Markov process involving state transitions between charged fragments. MetFrag obtains a candidate list from compound libraries based on the precursor mass, subsequently ranked by the agreement between measured and *in silico* predicted fragments [28]. It is a combinatorial fragmentor using the bond disconnection, top-down approach, starting with an entire molecular graph and removing each bond successively. MzCloud, in contrast to other databases, uses Mass Frontier (Thermo Scientific™) for the prediction of fragments, applying general fragmentation rules for more than a hundred thousand mechanisms, published in peer-reviewed journals.

Amongst other databases offering spectral prediction, CSI:FingerID, MAGMa and MetFrag do not contain real spectra. In MetFrag, searches are performed in two steps: first a database search is employed to find possible candidates corresponding to a particular parent ion and second product ions are explained. MetFusion (msbi.ipb-halle.de/MetFusion/), an extension of MetFrag, combines information from GPD, MassBank or METLIN with candidates generated in MetFrag [11]. CSI:FingerID combines fragmentation tree computation and machine learning to increase the number of MS^2 spectra available [14]. Support vector machines are employed for directly predicting a chemical fingerprint that is used to search for the metabolite with the closest match. MAGMa annotates hierarchical spectral trees obtained from multistage MS^n experiments. It performs queries using a selected source to explain fragments and score and rank candidate substructure matches.

Structure search

Structure searches using MS^2 data can be used in three modes: similarity, substructure and exact, whereby parts of the structure can be matched to find candidates with similar structures or candidates containing the observed structures as a substructure. Structure search options are available in HMDB, LipidMaps, MzCloud, MassBank, BioCyc and MINE (details given in table 4). The method for structure search is similar for most, except BioCyc where queries are performed through four different input options (chemical formula, SMILE InChI key or InChI string) rather than through uploading or drawing the structure. HMBD and MINE database compute a similarity threshold which can be used to filter out non-relevant candidates. It is also possible in HMDB to make a search from a pre-selected compound. In this way structures need not be drawn, instead particular metabolites can be selected and their structures used in the search. MzCloud offers the widest selection of filters, where a search can be restricted to certain compounds or precursors and several aspects of the structure can be ignored including charges, radicals, adducts and isotopes.

Additional functions

METLIN contains a very useful function for identification of unknowns: it allows searching by a list of fragments or neutral losses ignoring the precursor ion. This is particularly applicable when in-source fragmentation is high and the precursor ion is not present in the dataset. A similar assessment of fragments and neutral losses can be made in MassBank through the option “prediction” when working in the Japanese domain, although the precursor ion must also be

present. MyCopmoundID contains a useful feature called “deisitope”. This can be used to perform a search using only the first isotope, excluding all other natural isotopic peaks to avoid false matching. Moreover, this data source has the option to restrict candidate matches by filters including min/max precursor mass, intensity or score of fit. Useful tools are available within some of the data sources to explain unidentified fragments by predicting formulae from *m/z*. MassBank performs this based on data from Keio and Riken ESI-QTOF-MS^2, generating a list of possible formulae from the database given a suitable tolerance, that can be restricted to particular elements.

iii) Data interpretation

Metabolite annotation, performed on either MS or MS^2 levels can lead to a long list of possible candidates. If there is no possibility to obtain additional information about the structure, other mechanisms must be employed to exclude certain hits. Physical and chemical properties, origin, or biological role can be useful considerations for this. Some data sources offer clear advantages over others to assist the user in this regard.

Pathways

As already stated, pathway-centric databases are excluded from this review, however some of the databases considered do contain pathway related functions worth mentioning. Pathway information is available in BioCyc, CeuMM, HMDB, MassTRIX, MINE and Workbench. HMDB pathway information is based on its sister platform Small Molecule Pathway Database SMPDB (smpdb.ca/). All SMPDB pathways include information on the relevant organs, subcellular compartments, protein complex cofactors, protein complex locations, metabolite locations, chemical structures and protein complex quaternary structures, which might be particularly important for multi-omics studies. BioCyc also uses its own pathways, which are built and curated based on evidence from the literature. CeuMM, MassTRIX, MINE and Workbench use KEGG (http://www.genome.jp/kegg/) pathway information. In addition to KEGG, workbench uses HMDB/SMPDB information. CeuMM has the option to upload a list of metabolite KEGG identifiers and identify involved pathways ordered by number of hits.

Description and classification

HMDB provides a great deal of information about each metabolite entry. This information is stored in a “metabocard” which details the taxonomy, ontology, physical, chemical and biological properties, spectra, expected physiological concentrations, literature references and appropriate links. LipidBank also contains very useful information for data interpretation, including genetic, bioactivity and metabolic data in addition to literature references and Workbench provides literature references too. Table 6S (Supplementary information) details the information provided in each data source.

Classification approaches can be used to help filter or interpret hits given in databases using a forest or tree approach, for which taxonomy and ontology can be useful [45]. This data is available in BioCyc and HMDB for all metabolites and in LipidMaps and Workbench for lipids only, calling on LipidMaps whose nomenclature is the recognised standard for lipid classification. BioCyc includes

additional useful information including metabolic reactions in which metabolites are involved or information on their presence or abundance in culture medium, for example. This is particularly useful when considering the plausibility of a metabolite as a statistically significant feature of a study and can also be useful in the experimental design stage to choose certain experimental conditions if there are particular metabolites of interest that may be affected by that. Similarly, MINE provides information about enzymes and products of reactions in which metabolites are involved.

Workbench contains information about previous projects and research where particular metabolites were already found. The highly detailed data includes a further description of project, samples used, conditions applied and treatments and analytical conditions employed. Even measured abundances for particular masses across all the samples are stated.

CONCLUSIONS

Data analysis is a critical, but often an under-considered aspect of metabolomics research. In general, close to 50% of features detected in a non-targeted metabolomics study are unidentified compounds, leading to an important loss of information. Moreover, if features are mis-identified, data is wrongly interpreted and false conclusions are drawn onto which new experiments can be proposed. It is therefore vital to get this step right and be aware of the advantages and limitations of the tools at our disposal. As discussed, there is a range of different open access resources, with different characteristics that have been critically reviewed here. On-line tools will benefit from the input of a broad spectrum of scientists interested in metabolomics. However, the community as a whole should contribute to establish rules about data collected using different extraction protocols and analytical methods.

elps201700070-sup-0001-tableS1-S6.doc

Supplementary Information

ACKNOWLEDGEMENTS

This work was supported by grants from the Spanish Ministerio de Economía y Competitividad (grant CTQ2014-55279-R) and USPCEU PCON10/2016. AGF acknowledges Fundación Universitaria San Pablo CEU for his PhD fellowship.

All authors declare neither financial nor commercial conflicts of interest.

REFERENCES:

[1] Dunn, W. B., Lin, W., Broadhurst, D., Begley, P., Brown, M., Zelena, E., Vaughan, A. A., Halsall, A., Harding, N., Knowles, J. D., Francis-McIntyre, S., Tseng, A., Ellis, D. I., O'Hagan, S., Aarons, G.,

Benjamin, B., Chew-Graham, S., Moseley, C., Potter, P., Winder, C. L., Potts, C., Thornton, P., McWhirter, C., Zubair, M., Pan, M., Burns, A., Cruickshank, J. K., Jayson, G. C., Purandare, N., Wu, F. C. W., Finn, J. D., Haselden, J. N., Nicholls, A. W., Wilson, I. D., Goodacre, R., Kell, D. B., *Metabolomics* 2015, *11*, 9-26.
[2] Peng, B., Li, H., Peng, X.-X., *Protein Cell* 2015, *6*, 628-637.
[3] Zhang, A., Sun, H., Yan, G., Wang, P., Wang, X., *Biomed Research International* 2015, *2015*, 1-6.
[4] Johnson, C. H., Ivanisevic, J., Siuzdak, G., *Nature Reviews Molecular Cell Biology* 2016, *17*, 451-459.
[5] Dunn, W. B., Ellis, D. I., *TrAC Trends in Analytical Chemistry* 2005, *24*, 285-294.
[6] Cao, M., Fraser, K., Rasmussen, S., *Metabolites* 2013, *3*, 1036-1050.
[7] Werner, E., Heilier, J.-F., Ducruix, C., Ezan, E., Junot, C., Tabet, J.-C., *Journal of Chromatography B* 2008, *871*, 143-163.
[8] Godzien, J., Ciborowski, M., Angulo, S., Barbas, C., *Electrophoresis* 2013, *34*, 2812-2826.
[9] Sumner, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, C. A., Fan, T. W., Fiehn, O., Goodacre, R., Griffin, J. L., Hankemeier, T., Hardy, N., Harnly, J., Higashi, R., Kopka, J., Lane, A. N., Lindon, J. C., Marriott, P., Nicholls, A. W., Reily, M. D., Thaden, J. J., Viant, M. R., *Metabolomics* 2007, *3*, 211-221.
[10] Bingol, K., Bruschweiler-Li, L., Li, D., Zhang, B., Xie, M., Brüschweiler, R., *Bioanalysis* 2016, *8*, 557-573.
[11] Gerlich, M., Neumann, S., *Journal of Mass Spectrometry* 2013, *48*, 291-298.
[12] Hufsky, F., Rempt, M., Rasche, F., Pohnert, G., Böcker, S., *Analytica Chimica Acta* 2012, *739*, 67-76.
[13] Rojas-Cherto, M., Peironcely, J. E., Kasper, P. T., Van Der Hooft, J. J. J., De Vos, R. C. H., Vreeken, R., Hankemeier, T., Reijmers, T., *Analytical Chemistry* 2012, *84*, 5524-5534.
[14] Shen, H. B., Duhrkop, K., Bocker, S., Rousu, J., *Bioinformatics* 2014, *30*, 157-164.
[15] Vinaixa, M., Schymanski, E., Neumann, S., Navarro, M., Salek, R., Yanes, O., *TrAC Trends in Analytical Chemistry* 2016, *78*, 23-35.
[16] Jeffryes, J. G., Colastani, R. L., Elbadawi-Sidhu, M., Kind, T., Niehaus, T. D., Broadbelt, L. J., Hanson, A. D., Fiehn, O., Tyo, K. E. J., Henry, C. S., *Journal of Cheminformatics* 2015, *7*, 1-8.
[17] Fiehn, O., Barupal, D. K., Kind, T., *Journal of Biological Chemistry* 2011, *286*, 23637-23643.
[18] Caspi, R., Billington, R., Ferrer, L., Foerster, H., Fulcher, C. A., Keseler, I. M., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L. A., Ong, Q., Paley, S., Subhraveti, P., Weaver, D. S., Karp, P. D., *Nucleic Acids Research* 2016, *44*, D471-D480.
[19] Dührkop, K., Shen, H., Meusel, M., Rousu, J., Böcker, S., *Proceedings Of The National Academy Of Sciences Of The United States Of America* 2015, *112*, 12580-12585.
[20] Wishart, D. S., Jewison, T., Guo, A. C., Wilson, M., Knox, C., Liu, Y., Djoumbou, Y., Mandal, R., Aziat, F., Dong, E., Bouatra, S., Sinelnikov, I., Arndt, D., Xia, J., Liu, P., Yallou, F., Bjorndahl, T., Perez-Pineiro, R., Eisner, R., Allen, F., Neveu, V., Greiner, R., Scalbert, A., *Nucleic Acids Research* 2013, *41*, D801-D807.
[21] Yasugi, E., Watanabe, K., *Tanpakushitsu Kakusan Koso. Protein, Nucleic Acid, Enzyme* 2002, *47*, 837-841.
[22] Cotter, D., Maer, A., Guda, C., Saunders, B., Subramaniam, S., *Nucleic acids research* 2006, D507-D510.
[23] Ridder, L., van der Hooft, J. J. J., Verhoeven, S., de Vos, R. C. H., van Schaik, R., Vervoort, J., *Rapid Commun Mass Spectrom* 2012, *26*, 2461-2471.

[24] Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., Ojima, Y., Tanaka, K., Tanaka, S., Aoshima, K., Oda, Y., Kakazu, Y., Kusano, M., Tohge, T., Matsuda, F., Sawada, Y., Hirai, M. Y., Nakanishi, H., Ikeda, K., Akimoto, N., Maoka, T., Takahashi, H., Ara, T., Sakurai, N., Suzuki, H., Shibata, D., Neumann, S., Soga, T., Nishioka, T., Saito, K., Oda, Y., Taguchi, R., Iida, T., Funatsu, K., Matsuura, F., *Journal of Mass Spectrometry* 2010, *45*, 703-714.
[25] Suhre, K., Schmitt-Kopplin, P., *Nucleic Acids Research* 2008, *36*, W481-W484.
[26] Wagele, B., Witting, M., Schmitt-Kopplin, P., Suhre, K., *PLOS ONE* 2012, *7*, 1-5.
[27] Wolf, S., Schmidt, S., Muller-Hannemann, M., Neumann, S., *BMC Bioinformatics* 2010, *11*, 1-12.
[28] Ruttkies, C., Schymanski, E. L., Wolf, S., Hollender, J., Neumann, S., *Journal of Cheminformatics* 2016, *8*, 1-16.
[29] Smith, C. A., O'Maille, G., Want, E. J., Qin, C., Trauger, S. A., Brandon, T. R., Custodio, D. E., Abagyan, R., Siuzdak, G., *Therapeutic Drug Monitoring* 2005, *27*, 747-751.
[30] Huan, T., Tang, C., Li, R., Shi, Y., Lin, G., Li, L., *Analytical Chemistry* 2015, *87*, 10619–10626.
[31] Li, L., Li, R., Zhou, J., Zuniga, A., Stanislaus, A. E., Wu, Y., Huan, T., Zheng, J., Shi, Y., Wishart, D. S., Lin, G., *Analytical Chemistry* 2013, *85*, 3401-3408.
[32] Draper, J., Enot, D. P., Parker, D., Beckmann, M., Snowdon, S., Lin, W., Zubair, H., *BMC Bioinformatics* 2009, *10*, 1-16.
[33] Sud, M., Fahy, E., Cotter, D., Azam, K., Vadivelu, I., Burant, C., Edison, A., Fiehn, O., Higashi, R., Nair, S. K., Sumner, S., Subramaniam, S., *Nucleic Acids Research* 2016, *44*, D463-D470.
[34] Pétéra, M., Le Corguille, G., Landi, M., Monsoor, M., Tremblay Franco, M., Duperier, C., Martin, J.-F., Jacob, D., Guitton, Y., Lefebvre, M., Pujos-Guillot, E., Giacomoni, F., Thévenot, E., Caron, C., *Bioinformatics* 2015, *31*, 1493–1495.
[35] Severance, C., *Computer* 2015, 7-9.
[36] Fielding, R. T., Taylor, R. N., *ACM Transactions on Internet Technology* 2002, *2*, 115-150.
[37] Brown, M., Dobson, P., Patel, Y., Francis-Mcintyre, S., Begley, P., Broadhurst, D., Tseng, A., Kell, D. B., Dunn, W. B., Winder, C. L., Carroll, K., Swainston, N., Spasic, I., Goodacre, R., *Analyst* 2009, *134*, 1322-1332.
[38] Bowen, B. P., Northen, T. R., *Journal of the American Society for Mass Spectrometry* 2010, *21*, 1471-1476.
[39] Godzien, J., Ciborowski, M., Martínez-Alcázar, M. P., Samczuk, P., Kretowski, A., Barbas, C., *Journal of Proteome Research* 2015, *14*, 3204-3216.
[40] Godzien, J., Armitage, E. G., Angulo, S., Martinez-Alcazar, M. P., Alonso-Herranz, V., Otero, A., Lopez-Gonzalvez, A., Barbas, C., *Electrophoresis* 2015, *36*, 2188–2195.
[41] Xu, Y.-F., Lu, W., Rabinowitz, J. D., *Analytical Chemistry* 2015, *87*, 2273-2281.
[42] Dhanasekaran, A. R., Pearson, J. L., Ganesan, B., Weimer, B. C., *BMC Bioinformatics* 2015, *16*, 1-13.
[43] Wishart, D. S., *Bioanalysis* 2009, *1*, 1579-1596.
[44] Allen, F., Pon, A., Wilson, M., Greiner, R., Wishart, D., *Nucleic Acids Research* 2014, *42*, W94-99.
[45] Godzien, J., Ciborowski, M., Armitage, E. G., Jorge, I., Camafeita, E., Burillo, E., Martín-Ventura, J. L., Rupérez, F. J., Vázquez, J., Barbas, C., *Journal of Proteome Research* 2016, *15*, 1762-1775.

FIGURES

Figure 1. Different aspects of metabolite identification in the metabolomics workflow.

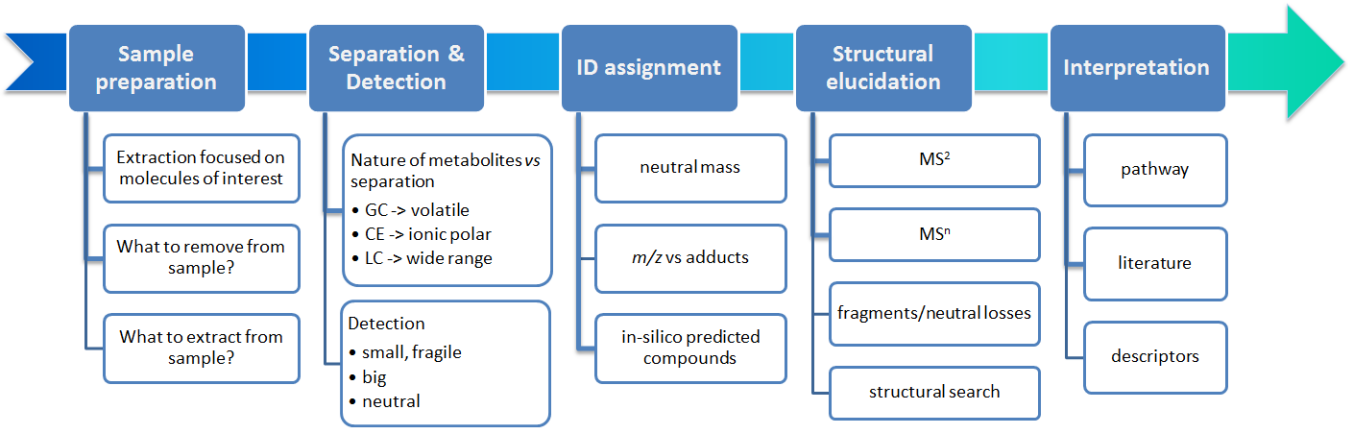


Figure 2. Number of compounds available in different data sources. Those containing only previously detected compounds are depicted in blue and those that include *in silico* generated compounds are depicted in red. CeuMM is the only mediator which gives information on the total number of compounds and is therefore the only mediator represented here.

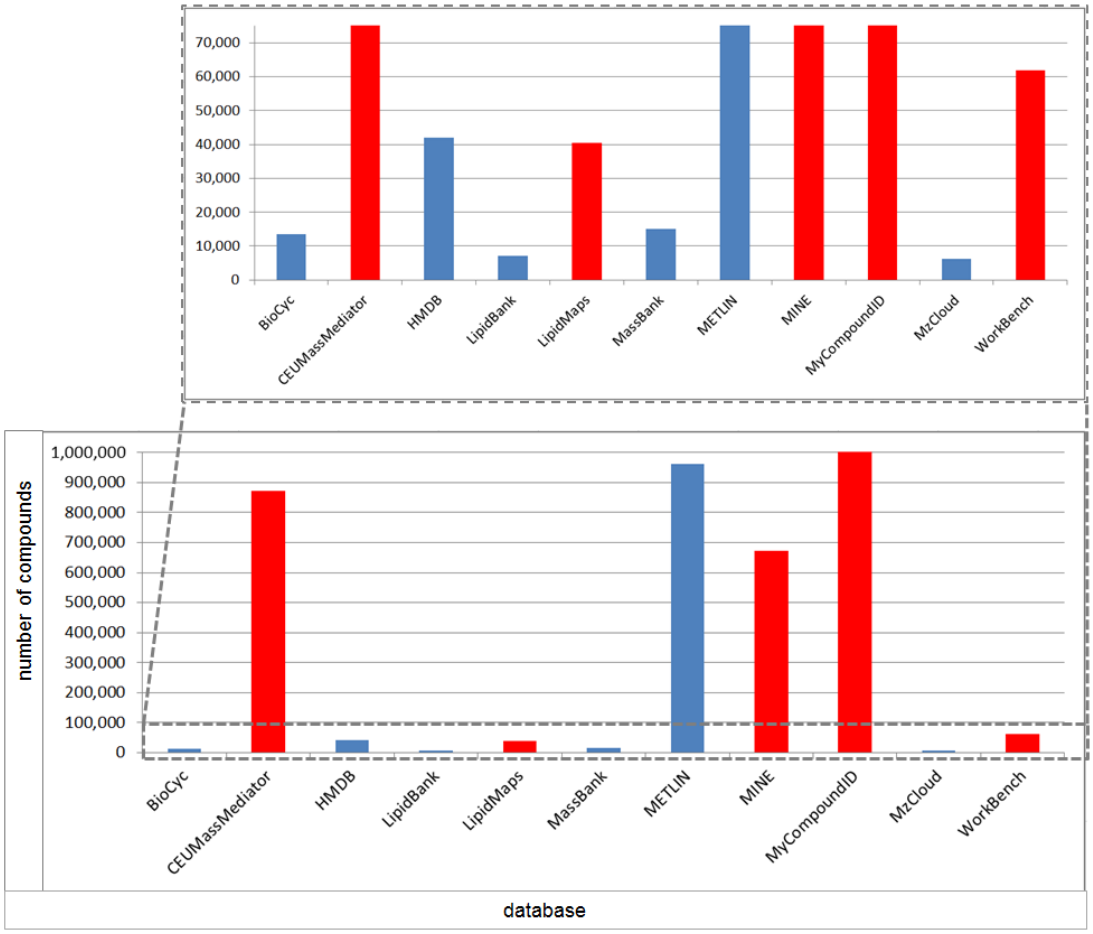


Figure 3. Number of citations of each data source by name in google scholar (as of 15th January 2017).

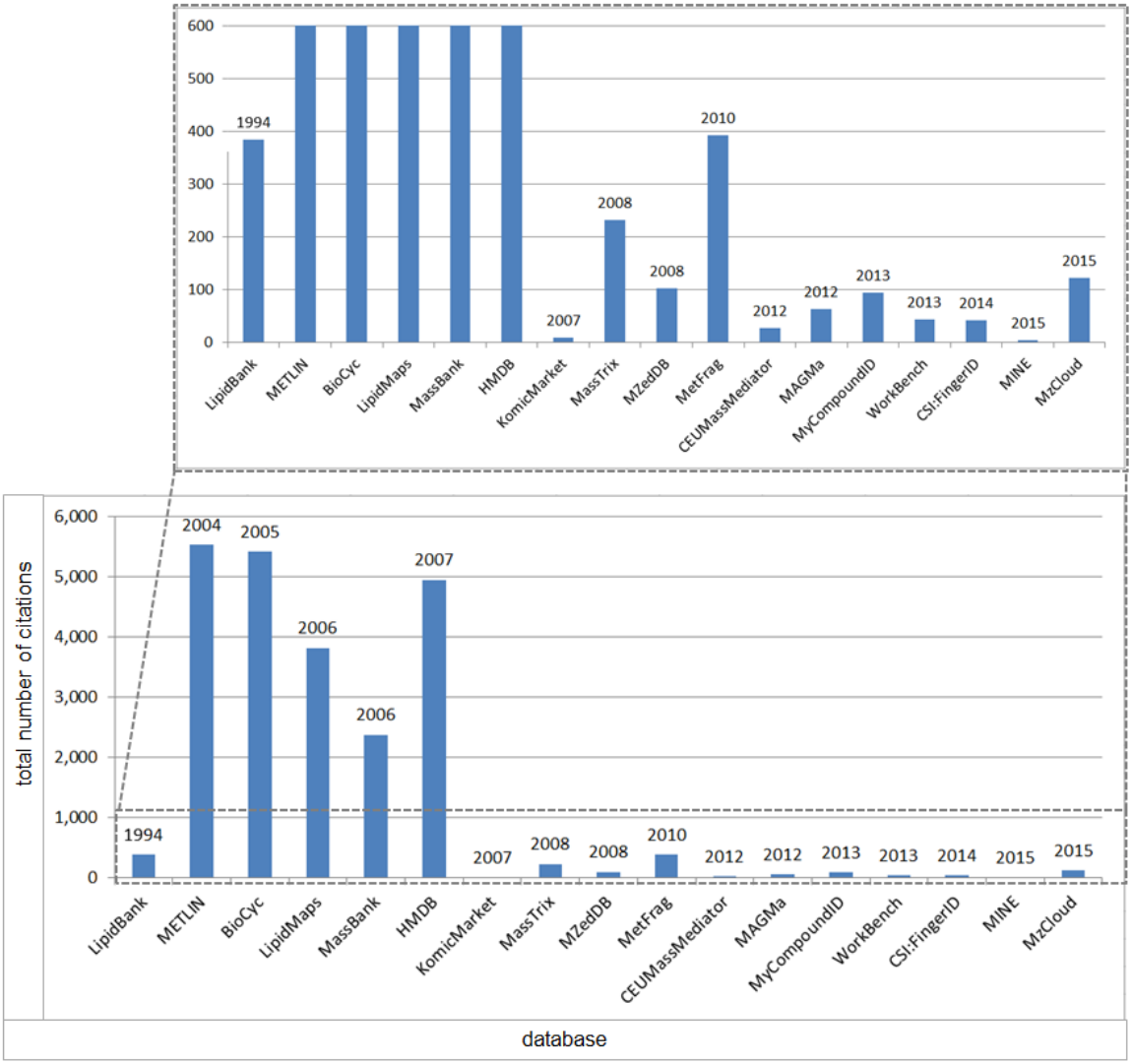
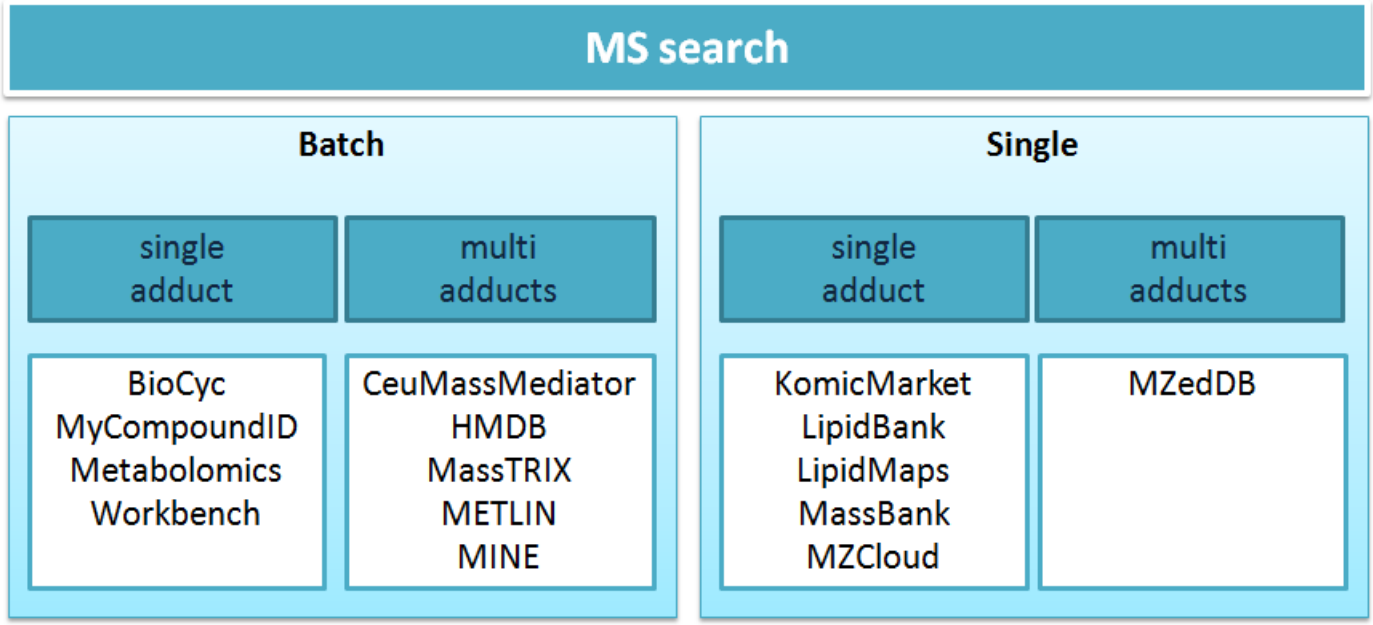


Figure 4. Classification of on-line tools for performing MS searches based on their features.



TABLES

Table 1. On-line tool characteristics.

On-line tool	Functionalities	Strong points	Weak points
BioCyc	- ID assignment - Structure search - Data interpretation[a]	- Organism selection - Information about possible reactions of compound - Literature references - Ontology search - Multi-conditions search - Customisable results	- Limit for batch searches based on URL length - Only neutral mass search - Subscription model (not freely available) - Limited information in exported file by default[b]
CeuMM	- ID assignment	- Unlimited search in batch mode - Multi-adduct - Chemical alphabet - Possibility to choose data source	- No structure available - No API
CSI:FingerId	- MS^n spectral search (fragmentation tree based on molecular formula prediction)	- Chemical alphabet	- Fixed relative error - Limited number of adducts - Positive ionisation mode only - No mass of compounds in results - No exporting option - No API
HMDB	- ID assignment - MS^2 search - Structure search - Data interpretation[c]	- Batch mode (700 masses at once) - Comprehensive characterisation of metabolites - High quality real and predicted spectra - Multi-adduct - Multi-conditions search - Spectra comparison - Very user-friendly	- No compound name in exported results - No exporting option for MS^2 search - No API

KomicMarket	- ID assignment	- Easy comparison with other studies - Filter by species (only 3) - Filter by analytical method - Filter by sample type - RT information for some compounds	- Single search - Single adduct - Limited number of adducts - No name or formula assigned for most compounds - No exporting option - No API
LipidBank	- ID assignment - Data interpretation[c]	- Hierarchical organisation - Biological activity, physical properties, spectral data, organism and references available	- Single search - No monoisotopic mass - Query only from average neutral mass - Out-of-date front-end technology and design - No exporting option - No API
LipidMaps	- ID assignment - Structural assignment - Data interpretation[c]	- Hierarchical organisation - Physicochemical properties, spectral data and references available - Ontology search - MS^2 library for standards	- Single search - Neutral mass - Fixed absolute error - MS^2 spectra only for single collision energy
MAGMa	- MS^n spectral search (fragmentation tree based on substructure prediction)	- Substructure search - Tolerance in Da + ppm	- No adduct search - No API
MassBank	- ID assignment - MS^n search - Structural assignment	- Filter by analytical method - Molecular formula generator - Repository for contributors - Package view for multi-hits comparison in MS^n search	- Batch mode only under request for MS^1 - Neutral mass - No unification about experimental conditions - No exporting option - No API
MassTRIX	- ID assignment - Data interpretation[d]	- Unlimited search in batch mode - Organism selection	- Fixed relative or absolute error - Limited list of adducts

			- No direct data query, queue jobs system - No exporting option - No API
MetFrag	- MS^n *in silico* explanation (based on structure fragmentation)	- Well-structured downloaded files for explanation of fragments	- Single adduct
METLIN	- ID assignment - MS^2 search - ID assignment for isotope labelling - Fragment search - Neutral loss search	- Batch mode (500 masses at once) - Multi-adduct - Option to include/remove drugs, peptides and toxicants - Information on where compounds can be purchased as standards - Spectra comparison	- Confusing MS^2 spectra (differences in real/predicted and/or energy collision) - No possibility to exclude predicted spectra for MS^2 search - No exporting option - No API - Problems with access (often banned)
MINE	- ID assignment - Structural assignment - Data interpretation[e]	- Unlimited search in batch mode - Multi-adduct - Multi-conditions search - Information about possible reactions of compounds - Possibility to choose data source	- No clear indication and distinction between real and predicted compounds - No possibility to limit search to only real or predicted results in the on-line version
MyCompoundID	- ID assignment - MS^2 search - ID assignment for isotope labelling	- Unlimited search in batch mode for MS^1 - Batch search for MS^2 search (100 spectra at once) - Detailed information about MS^2 peaks explained from library - Deisotope function for MS^2	- Single adduct - Limited list of adducts - Exporting option only available for one mass at a time - No API
MzCloud	- ID assignment	- Compound filter (See list 1 in	- No adduct search (Only $[M+H]^+$, $[M-H]^-$)

	- MS^n search - Structural assignment - Fragment search - Data interpretation[a]	Supplementary Information) - Contributor repository	- No exporting option - No API - Built in Microsoft Silverlight (technology deprecated by Microsoft)
MZedDB	- ID assignment	- Multi-adduct - Chemical alphabet - Molecular formula generator - Possibility to choose data source - Adduct/neutral loss rules	- Single search - No exporting option - No API
WorkBench	- ID assignment - Structural assignment - Data interpretation[a]	- Unlimited search in batch mode - Ontology search - Repository for contributors - Possibility to choose data source[f]	- Single adduct - Only absolute error - No exporting option

a) organism selection, information about reactions and pathways information
b) only information about mass, compound name and chemical formula, without any link for further researching
c) detailed description and references
d) pathway analysis
e) information about possible reactions
f) three options: virtual database of lipids, a reference set of metabolites and Metabolomics Workbench Metabolite Database (database collected from multiple repositories: LIPID MAPS, ChEBI, HMDB, BMRB, PubChem, and KEGG)

Table 2. Features available in each database.

Feature Description		Databases
Source	Database	BC, HM, KM, LB, LM, MB[a], ME, MZC, WB
	Mediator	CMM, CF, MG, MT, MF, MI, MY, MZD, WB
MS^n	MS^2	CF, HM, KM, LM, MG, MB, MF, ME, MY, MZC
	MS^n	CF, MG, MB, MZC,
	Real spectra	HM, KM, LM, MB, ME, MY, MZC
	Predicted spectra	CF, HM, MG, MF,ME, MY, MZC
Search mode for MS	Single	KM, LB, LM, MG, MB[b], MF, MZC, MZD
	Batch	BC[c], CMM, HM (700), MT, ME (500), MI, MY, WB
Search mode for MS^2	Single	CF, HM, MG, MF, ME
	Batch	MB, MY (100), MZC
Adducts*	Single	KM, MY, WB
	Multi	CMM, CF, HM, MT, MF, ME, MI, MZD
	Neutral	BC, LB[d], LM, MG, MB, MZC
Last update*	[0-1 years]	BC, CMM, CF, HM, LM, MG, MB, MF, ME, MI, MZC, WB
	[1-3 years]	MY
	[3-more years]	KM, LB, MT, MZD
Licensing	Open	CMM, LM (BSD), MG (Apache), MF (GNU), MI (CC 4.0)
	Proprietary	BC, KM, LB, MT, ME, MZC, MZD, WB
	Not Specified/Depends on contributor	CF, HM, MB, MY, MZ
Usage of data	Free (Non-commercial)	CMM, CF, HM, KM, LM, MG, MF, ME, MI, MZC, MZD, WB
	Free (All purposes)	MG
	Fee	BC (except EcoCyc and MetaCyc)
	Not Specified/Depends on contributor	LB, MB, MY
Export formats*	csv, xls, tsv	BC, CMM, HM, LM, MG, MF, MI, MY[e]

	sdf	LM, MG, MF,
	html(only)	CF, KM, LB, MB, MT, ME, MZC, MZD, WB
API	REST	BC, LM, WB
	WebService	BC, KM, MI
	Other programming languages	BC (Python, Perl, Java, and Lisp), LM (PHP), MF (R), MI (Python, JavaScript, Perl)
	None	CMM, CF, HM, LB, MG, MB, MT, ME, MY, MZC, MZD
Search Options	Mass	BC, CMM, CF, HM, KM, LB[d], LM, MG, MB, MT, MF, ME, MI, MY, MZC, MZD, WB
	Formula	BC, CF, HM, KM, LB, LM, MG, MB, MF, ME, MI, MZD, WB
	Name	BC, HM, KM, LB, LM, MB, ME, MI, MZC, MZD, WB
	ID	BC, HM, KM, LB, LM, MF, ME, MI
	Ontology	BC, LM, WB
	Substructure/Sub-formula	BC, HM, LM, MB, MI, MzC loud
	Origin of compound[f]	BC, LB, MT
	Chemical Alphabet	CMM, CF, MZD
	Nature of compound[g]	HM[h], ME[i]
	Join several conditions	BC, CMM, CF, HM[h], LB, LM, MB, MT, MF, ME[h], MI[h], MZD, WB
Tolerance	ppm	BC, CMM, CF (2.5-15), MF, ME, MZD
	Da	LM (0.01-100), MB, WB (0.0005-1)
	Both	HM, KM (0-1Da, 0-100ppm), MG, MT (0.001-1Da, 0.1-3ppm), MI (0-15mDa,0-15ppm), MY, MZC
* Details available in Supplementary Information (Table S1, S2 and S3) a) Data repository. Data comes from contributors b) Batch mode available only by mail request c) Number of input masses limited by URL length d) Only average mass e) Peak by peak f) Distinguished by organism e.g. human, mice, *E.Coli* etc.		BC = BioCyc Database Collection (BioCyc) CMM = Ceu Mass Mediator CF = Compound Structure Identification: FingerID (CSI:FingerID) HM = Human Metabolome Database (HMDB) KM = Kazusa Omics Data Market (KomicMarket) LB = LipidBank LM = LIPID Metabolites And Pathways Strategy (LipidMaps) MG = MAGMa

g) The type of compound e.g. toxins, drug, exogenous etc. h) Only available for single search i) Distinction for drugs, peptides and toxicant	MB = MassBank MT = MassTRIX MF = MetFrag ME = METLIN MI = Metabolic *In Silico* Network Expansion Databases (MINE) MY = MycompoundID MZC = MzCloud MZD = MZedDB WB = UCSD Metabolomics Workbench (Workbench)

Table 3. User-friendliness of the tools.

	Design	Asynchronous techniques	Login mandatory	Easiness for familiarisation
BioCyc	★★★★☆	✓	✓	★★★☆☆
CeuMM	★★★★☆	✓	✗	★★★★☆
CSI:FingerID	★★★★★	✓	✗	★★★★☆
HMDB	★★★★★	✓	✗	★★★★★
KomicMarket	★☆☆☆☆	✗	✗	★★☆☆☆
LipidBank	★★☆☆☆	✗	✗	★★★★★
LipidMaps	★★★★★	✓	✗	★★★★★
MAGMa	★★★★☆	✓	✗	★★★☆☆
MassBank	★★★★☆	✓	✗	★★★☆☆
MassTRIX	★★☆☆☆	✗	✗	★★★★★
MetFrag	★★★★☆	✓	✗	★★★★☆
METLIN	★★★☆☆	✓	✓[a]	★★★★★
MINE	★★★☆☆	✓	✗	★★★★☆
MycompoundID	★★★☆☆	✗	✗	★★★★★
MzCloud	★★★★☆	✓	✗	★☆☆☆☆
MZedDB	★★☆☆☆	✗	✗	★★☆☆☆
Workbench	★★★★☆	✓	✗	★★★★☆

a) Locking users out when multiple consecutive searches are performed

Table 4. Features for structure search

Data source	Software	Search mode			Filter
BioCyc	-		substructure	exact	
HMDB	MarvinJS, ChemAxon	similarity	substructure	exact	- similarity threshold - molecular weight (range)
LipidMaps	GGA Ketcher		substructure	exact	- all - curated records only - computationally generated records only
MassBank	not stated		substructure		- search in MassBank - search in KNApSAck
MINE	MarvinJS, ChemAxon	similarity	substructure	exact	- similarity threshold
MzCloud	not stated		substructure	identity	- filter compounds - search in compound - search in precursor - ignore charges - ignore radicals - ignore adducts - ignore isotopes