3

Online Metabolomics Databases and Pipelines

Adam J. Carroll
The Australian National University
Australia

1. Introduction

As metabolomics becomes an increasingly major component of modern biological research, steps must be taken to preserve and make maximal use of the ever increasing torrents of new data entering the public domain. While this task is by no means unique to the field of metabolomics, the complexity, heterogeneity and large sizes of metabolomics datasets make the development of effective metabolomics bioinformatics tools particularly challenging. Despite these challenges, metabolomics specialists have recently been making rapid progress in this area. A wide range of powerful web-based tools designed to facilitate the systematic online storage, processing, dissemination and biological interpretation of technically and biologically diverse metabolomics datasets have now emerged and are rapidly becoming cornerstones of advancement in biological science.

Web-based tools for metabolomics perform a wide variety of functions. These can be divided into several broad categories, including:

1. Storage and dissemination of technical, biological, and physicochemical reference data for metabolites
2. Processing of raw instrument data to generate [metabolite x sample] data matrices suitable for statistical and multivariate data-analysis
3. Database storage and querying of pre-processed relative and/or absolute metabolite level data
4. Statistical and multivariate analysis of pre-processed relative and/or absolute metabolite level data
5. Aiding biological interpretation of metabolomics results by integration of biological knowledge such as known biomarkers or metabolic pathway information.

While some tools are broader in scope than others and some tools can essentially fully service the data-processing requirements of certain metabolomics approaches, it is important to note that no single tool is currently capable of fulfilling every requirement of every metabolomics researcher. This chapter will review the current state of development in the area of web-based informatics tools for metabolomics and explain how currently available tools can be used to accelerate scientific discovery. It will then attempt to predict future developments in the area of metabolomics web-tool development and advise new metabolomics researchers on strategies to maximise their own benefit from these developments.

2. Information about metabolites: Biological cheminformatics

2.1 Background

One of the fundamental questions of metabolomics is "how many metabolites occur in nature, what are their structures, what are their physical, chemical and biological properties and how are they distributed amongst species?". Large-scale efforts to build comprehensive databases of metabolite-related knowledge are beginning to provide at least approximate answers to these questions. Defining "the metabolome" of an organism in qualitative sense, by building well-annotated catalogues of metabolites and their properties, is analogous to sequencing and mapping the genome of an organism. That is, it provides a crucial foundation for the development of analytical approaches and experiments, aids in the interpretation of analytical results and provides an important scaffold upon which to attach new information as it becomes available.

Because metabolites are small molecule chemicals of biological origin, organisation of metabolite information lies at the interface between *bioinformatics* (the management of biological information) and *cheminformatics* (the management of chemical information). While metabolomics researchers will find useful information about metabolites in broad-scoped, general cheminformatic databases, a new generation of biology-focused cheminformatic databases are making it easier for biologists to find cheminformatic data specifically related to biology. This section will guide the reader towards online sources of metabolite information and explain how these information sources can be used to aid metabolomics research.

2.2 Molecular semantics: The metabolite naming issue

One of the challenges associated with finding online information about metabolites can be figuring out what text to enter into search engines. Metabolites can be named in many different ways in many different places online and searching with one name will generally only retrieve resources tagged with that particular name. Moreover, in cutting-edge metabolomics research, it is frequently the case that one is searching for information about a poorly-known or even completely hypothetical metabolite for which one has a structure in mind but for which its common name, if indeed it has one, is unknown. Fortunately, there are ways around these problems, thanks to the thoughtful design of cheminformatic databases. These will be explained below.

For well-known metabolites, finding detailed information is particularly easy. Most metabolite information databases annotate each metabolite entry with a large set of 'synonyms' – a range of different names commonly used to refer a given metabolite. As a result, if one uses a common name to search those databases for information about a well-known metabolite, one will usually find the information they need. In those cases, the key thing is to know which databases to search (these will be outlined shortly).

Finding information on well-known metabolites is relatively easy. However, metabolomics researchers are often interested in discovering new metabolites or learning what little is known about more about poorly-known metabolites. Often, a researcher may know the structure of a theoretical metabolite but have no idea whether it has been observed in nature before let alone what its common name might be. Indeed, such 'theoretical' metabolites

often *have* been observed in nature before and have a common name, but finding this out can be challenging if one does not know where to start. This is where *InChI* codes and comprehensive InChI-enabled cheminformatic databases become *indispensible* (Wohlgemuth et al., 2010).

"InChI" is an abbreviation for "**In**ternational **Ch**emical **I**dentifier", a system of expressing chemical structures as compact strings of text suitable for efficiently and unambiguously conveying chemical structures across text-based systems such as web search engines. The InChI system was developed by the International Union of Pure and Applied Chemistry (IUPAC) and the National Institute of Standards and Technology (NIST). Each unique chemical structure can be converted into its own unique InChI code and vice-versa[1]. There are a range of freely-available software tools that allow one to draw a chemical structure and obtain its InChI code or enter an InChI code and have its structure drawn automatically (see Table 1 for examples). All the major metabolite information databases tag their entries with InChI codes, so if one is uncertain of the name of a target metabolite, the best approach is to generate its InChI code and search with that. Some cheminformatic databases provide web-based structure drawing tools allowing users to effectively generate an InChI code and search with it in a single step. One of the advantages of using an unambiguous structural identifier such as InChI to search a database is that if no hits are obtained, one may fairly safely conclude that the target molecule was not in the database[2]. When a hit is obtained, however, the returned information may include common name(s) for the molecule that can aid in subsequent literature searches. For anyone building metabolite databases or supplying supplementary tables of metabolite data for publication, annotation of these data with InChI codes is highly-recommended (Wohlgemuth et al., 2010). Online tools for generating InChI codes from structures or other identifiers are listed in Table 1. A particularly useful tool for metabolomics researchers is the Chemical Translation Service provided by the lab of Oliver Fiehn (Wohlgemuth et al., 2010) since this tool is capable of batch translations of miscellaneous metabolite identifiers and synonyms to standard InChI codes and other common identifiers.

2.3 Chemical ontologies: Organising metabolites into useful categories

In scientific communication, biologists frequently refer to broad 'classes' of metabolites using terms related to their functional groups (eg. 'alcohols'), their chemical properties (eg. 'organic acids') or biological roles (eg. 'hormones'). Moreover, researchers are often interested in obtaining lists of metabolites that fall in a particular class. For example, a researcher may want to identify metabolites in an organism that contain a particular functional group and will therefore be expected to undergo certain chemical reactions. Potential classes range in scope from very broad (eg. 'organic') to moderately specific (eg. 'alkaloids') to even more specific (eg. 'monoterpenoid indole alkaloids') and so on. While "metabolite classes" like these appear throughout the scientific literature, formalising them

[1] There is one caveat to this statement. The only truly non-ambiguous InChI codes are called "Standard" InChI (often abbreviated to "StdInChI" - these always begin with the string "InChI=1S/"). If building a metabolomics database, it is advisable to use only standard InChI codes.

[2] Some metabolite databases were built prior to the release of Standard InChI and have been annotated using non-standard InChI codes (always beginning with "InChI=1/"). It is always a good idea to check which InChI type a database uses before searching it with an InChI code.

into accurately and systematically defined and 'chemical ontologies' that can be used in practically useful ways is a non-trivial task. Despite this, a number of metabolite-related databases have begun developing and/or employing hierarchical systems of compound classification, allowing users to browse lists of metabolites via classification trees (ontologies). Examples of databases employing compound ontologies or hierarchical compound taxonomies for annotation of metabolite information include PubChem, ChEBI (Degtyarenko et al., 2008), the BioCyc family of metabolic pathway databases (Caspi et al., 2010), the Human Metabolome Database (HMDB) (Wishart et al., 2007) and MetabolomeExpress (Carroll et al., 2010). The ChEBI compound ontology is by far the most advanced and comprehensive ontology for biological small molecules and is downloadable in open formats from the ChEBI website. Its adoption is recommended in the development of new metabolomics databases.

Tool (URL)	**Features**
ChEBI Advanced Search (http://www.ebi.ac.uk/chebi/advancedSearchForward.do#)	**Input:** Structure drawing tool **Output:** StdInChI, SMILES **Search capabilities:** substructure, similarity, identical structure
PubChem Structure Search (http://pubchem.ncbi.nlm.nih.gov/search/search.cgi)	**Input:** Structure drawing tool **Output:** StdInChI, StdInChIKey, SMILES, SMARTS, Formula **Search capabilities:** substructure, superstructure, similarity, identical structure
Fiehn Lab Chemical Translation Service (http://uranus.fiehnlab.ucdavis.edu:8080/cts/homePage)	**Input:** Any major database ID, common synonym or structure identifier **Output:** Any major database ID, common synonym or structure identifier **Search capabilities:** simple search by any major database ID, common synonym or structure identifier

Table 1. Recommended online tools for generating unambiguous InChI structural identifier strings from structures, names or other identifiers

2.4 Physicochemical information

Physicochemical information about metabolites includes information about their physical and chemical properties such as their structures, molecular formulas, molecular weights, melting and boiling points, solubilities in different solvents at different temperatures, polarities, pKa, light absorbance and fluorescence properties, energy contents, refractive indices and other similar types of basic empirical information. This kind of information can be extremely useful when designing extraction, sample clean-up or analyte-enrichment protocols, for example.

2.5 Recommended sources of general metabolite information

Many online databases offer information about metabolites. These have varying scales and scopes of content, query tools and modes of access. In these aspects, several databases stand out from all others and these are described below.

2.5.1 ChemSpider

Description: A freely-accessible collection of compound data from across the web with a very versatile search engine.

Scope: all chemicals – not just metabolites

Semantic content: *Many* synonyms, identifiers and external database IDs and link-outs

Physicochemical content: Masses, formula, experimental melting point, physical state, appearance, stability, storage compatibility, safety. A substantial amount of additional predicted data.

Biological content: Links to MeSH

Analytical content: Some compounds have spectra

Noteworthy tools: Search by physicochemical properties

Modes of access: search (by synonym, InChI, SMILES, CAS, structure), API. Limited to 5000 structures per day.

Strengths: Enormous index of chemicals that is widely linked to external online resources – a good starting point if looking for information on a particular chemical.

Limitations: Broad focus means extracting desired subsets of information can be difficult. Cannot be downloaded. Results only returned in HTML - not spread sheet format. Limited to 5000 structures per day.

URL: http://www.chemspider.com/

2.5.2 Human Metabolome Database (HMDB)

Description: A comprehensive, freely-available knowledgebase of human metabolite information.

Scope: Human (*Homo sapiens*) metabolites

Semantic content: metabolite names, formulas, masses, structures, InChI, SMILES, external database IDs and link-outs; Chemical taxonomy

Physicochemical content: Masses, formula, water solubility, hydrophobicity, melting point, physiological charge, physical state

Biological content: Presence in human cellular compartments and biofluids; Measured concentrations in biofluids; reactions; enzymes; enzyme genes; disease associations; descriptions of biological roles

Analytical content: many compounds have LC/MS, GC/MS and/or NMR spectra obtained under standardised conditions

Noteworthy tools: Versatile 'Data Extractor'. Searching based on spectral properties

Modes of access: browse, search, complex query with data extractor, download

Strengths: Comprehensive, may be freely downloaded in entirety. Human focus is good for human metabolomics researchers.

Limitations: Important fields are empty for some very common metabolites. Being limited to human metabolites limits utility for other research areas. Downloadable flat-file format requires parsing in order to be usable in spread sheets or local databases.

Reference: (Wishart et al., 2007)

URL: http://www.hmdb.ca/

2.5.3 Chemical Entities of Biological Interest (ChEBI)

Description: A freely-available dictionary of small molecule chemicals of interest to biologists.

Scope: Small molecules of biological interest (endogenous biochemicals and exogenous bioactive compounds)

Semantic content: metabolite names, formulas, masses, structures, InChI, SMILES, external database IDs and link-outs; ChEBI Chemical Ontology

Physicochemical content: Formal charge

Biological content: None

Analytical content: None

Noteworthy tools: Structure based search, Versatile advanced query, Chemical and functional ontology-based browsing

Modes of access: browse, search, complex query, FTP download

Strengths: May be downloaded in bulk. Versatile advanced query tool. Query results downloadable in useful formats. Well-designed ontology.

Limitations: Human-centric. Far from comprehensive, particularly for non-human-related information. References to supporting literature are not provided with biofunctional ontology assignments. No species occurrence information.

Reference: (Degtyarenko et al., 2008)

URL: http://www.ebi.ac.uk/chebi

2.5.4 PubChem

Description: A freely available general dictionary of chemicals.

Scope: Any small molecules

Semantic content: metabolite names, formulas, masses, structures, InChI, SMILES, external database IDs and link-outs; MeSH chemical classification

Physicochemical content: Formal charge, partition coefficient, H-bonding donor and acceptor counts

Biological content: Bioactivity, bioassay results, safety and toxicology, associations with metabolic pathways in KEGG

Analytical content: None

Noteworthy tools: Structure and structural similarity-based search, versatile advanced query, chemical and biomedical ontology-based browsing, chemical structure clustering

Modes of access: browse, search, complex query, FTP download

Strengths: Huge number of compounds. Highly annotated. Some information is available for bulk download. Versatile advanced query tool. Extensive link-outs.

Limitations: Much compound information displayed on the website is not provided for bulk download. No species occurrence information.

URL: http://pubchem.ncbi.nlm.nih.gov

2.6 Metabolic pathway databases

A wide range of biological information about metabolites is available online. Utilising this information can aid in the development of hypotheses, the design of experiments and the biological interpretation of metabolomics results. For this purpose, among the most useful types of database are metabolic pathway databases. These play a crucial role in metabolomics research by systematically capturing and providing a close representation of current knowledge about: a) which metabolites occur in particular biological systems; b) the enzymatic and non-enzymatic reactions that link different metabolites together into metabolic pathways; c) the enzymes that carry out these reactions and the genes that encode them; and d) the allosteric interactions and signalling networks that regulate these genes and gene products. Another highly useful function that some metabolic pathway databases carry out is to visually overlay metabolomic datasets over pathway diagrams to provide biological contexts aiding the biological interpretation of results. Some of most useful metabolic pathway databases are described below.

2.6.1 Kyoto Encyclopedia of Genes and Genomes (KEGG)

Description: A knowledgebase of genomes, genes, gene-products their properties and the metabolic and regulatory pathways they form.

Species: many species from many different classes

Metabolic pathway content: metabolite names, formulas, masses, structures and external database IDs; reactions; reactant-product atom mappings; pathways; enzymes; enzyme

genes; orthologies; bioactivities; allosteric interactions / regulatory pathways; pathway, compound, taxonomy and biological process ontologies

Noteworthy features: Structural similarity search

Modes of access: browse, search, API, FTP download (requires subscription)

Strengths: Enormous amount of information. The largest source of atom-mapped reactions available.

Limitations: Broad focus means extracting desired subsets of information can be challenging. Query tools are limited.

Reference: (Ogata et al., 1999)

URL: http://www.genome.jp/kegg/

2.6.2 BioCyc and the "Cyc" family of metabolic pathway databases

Description: Similar to KEGG. A collection of Pathway / Genome Databases (PGDBs) built using software that predicts metabolic pathways from genome sequences and subsequently refined by varying degrees of expert curation.

Species: BioCyc itself includes highly-curated PGDBs for 3 organisms: *Escherichia coli* (EcoCyc), *Arabidopsis thaliana* (AraCyc), *Saccharomyces cerevisiae* (YeastCyc). Another highly-curated PGDB called MetaCyc compiles pathway and enzyme information from >1900 organisms (mainly single-cell organisms) into a single reference database. See also the separate HumanCyc, PlantCyc and many other "Cyc" databases.

Metabolic pathway content: metabolite names, formulas, masses, structures, InChI, SMILES, Gibbs free energies and external database IDs; reactions; pathways; enzymes; enzyme genes; allosteric interactions / regulatory pathways; compound, pathway, gene and enzyme ontologies; links to literature supporting pathways

Noteworthy features: Interactive cellular overview pathway display, Regulatory overview, Genome browser, Advanced query tool, Powerful API, Omics viewer

Modes of access: browse, search, API, FTP

Strengths: Enormous amount of information. Powerful and intuitive query tools and API make extraction of data subsets easy. Returns results in simple tables and open XML pathway exchange formats. Omics viewer allows overlaying of omics (including metabolomics) data onto pathway views.

Limitations: Some useful and easily-fillable fields are empty for some metabolites. The Cyc databases often refer to generic entities such as 'a fatty acid' – this can limit their utility when researchers are interested in modelling connections between certain specific entities.

Reference: (Caspi et al., 2010)

URLs: http://biocyc.org, http://humancyc.org, http://www.plantcyc.org and others

2.6.3 Reactome

Description: An interactive collection of curated, peer-reviewed metabolic pathways with cross-referencing of reactions and pathways between organisms. Pathways are displayed via an intuitive GUI but may be downloaded in a variety of open formats.

Species: A variety of species. Most comprehensive for human.

Metabolic pathway content: hierarchically organised curated and peer-reviewed metabolic pathways; reactions; reaction-gene associations

Noteworthy features: Interactive pathway viewer

Modes of access: browse, search and download

Strengths: Peer-reviewed, user-friendly, different subcellular metabolite pools are treated as separate entities

Limitations: Reaction-centric. Not much information about metabolites and does not provide any tools for overlaying metabolite expression data.

Reference: (Croft et al., 2011)

URL: http://www.reactome.org

2.6.4 KappaView

Description: A web-based tool allowing users to overlay metabolite- and gene-expression responses and correlations onto custom pathway diagrams or onto a collection of neat, simple and interactive metabolic pathway diagrams.

Species: A variety of species.

Metabolism-related content: hierarchically organised curated metabolic pathways; reactions; reaction-gene associations

Noteworthy features: Gene and metabolite expression overlay

Modes of access: browse, search and download

Strengths: User-friendly; neat/simple diagrams; may be integrated into third party websites using a flexible API; can also overlay metabolite-metabolite, gene-gene and metabolite-gene correlations

Limitations: Does not support InChI

Reference: (Sakurai et al., 2010)

URL: http://kpv.kazusa.or.jp/kpv4/

2.6.5 Human Metabolome Database (HMDB) Pathways

Description: A comprehensive, freely-available knowledgebase of human metabolite information.

Species: Human (*Homo sapiens*)

Metabolism-related content: A set of 'textbook' style metabolic pathway diagrams with metabolites hyperlinked to HMDB metabolite information pages and enzymes hyperlinked to UniProt database.

Noteworthy features: None

Modes of access: browse, search

Strengths: Easy to understand

Limitations: Not downloadable. No documented API.

Reference: (Wishart et al., 2007)

URL: http://www.hmdb.ca

2.6.6 KNApSAcK

Description: A comprehensive species-metabolite relationship database for plants. Although not strictly a metabolic pathway database, this database is useful for identifying plant species that contain a certain chemical or identifying chemicals that have been reported in a particular plant species or higher level taxon.

Species: Plants

Metabolism-related content: References to literature reporting the presence of compounds in different plant species. Chemical structures. Masses.

Noteworthy features: References to literature.

Modes of access: browse, search

Strengths: Contains information on many plant-specific specialised metabolites.

Limitations: Data itself is not downloadable.

Reference: (Shinbo et al., 2006)

URL: http://kanaya.naist.jp/KNApSAcK/

3. Online analytical reference spectra for metabolomics

3.1 The roles of analytical reference libraries in metabolomics research

The first online metabolomics databases to store and disseminate actual instrument data for metabolites generally provided spectral reference libraries. These spectral libraries provide reference signals for authentic standard compounds and sometimes also for 'unknown' metabolites obtained through the analysis of standards and biological materials under controlled conditions. The de-novo construction of large analytical reference libraries requires expertise in chemistry, is time consuming and expensive. Centralization of spectral reference data in expert-curated public repositories helps the metabolomics community by: 1) making it easier and cheaper for new labs to build their

own data processing pipelines; 2) reducing the probability of metabolite misidentification by non-specialists; and 3) promoting efficient communication about 'unknown' metabolites that are recognisable on the basis of their analytical properties but for which no structural information is available.

3.2 Types of analytical reference spectra available online

Reference spectra are available from a number of online sources. Types of reference data available include downloadable mass-spectral and retention-index (MSRI) libraries for gas chromatography / mass spectrometry (GC/MS) (Kopka et al., 2005; Schauer et al., 2005; Carroll et al., 2010), searchable but not-downloadable MSRI data (Skogerson et al., 2011), NMR spectra collected under standardized conditions (Wishart et al., 2007; Cui et al., 2008; Ulrich et al., 2008) and MS and MS/MS spectra from a wide range of platforms including accurate mass instruments (Smith et al., 2005; Horai et al., 2010). In addition, most cheminformatic and metabolic pathway databases provide accurate monoisotopic mass information for metabolites which can help provide candidate identities for accurate-mass LC/MS and direct-infusion (DI)/MS peaks. These data sources are described in detail later.

3.3 Reference data for Nuclear Magnetic Resonance (NMR)

One of the great advantages that NMR has over mass-spectrometry is that chemical shifts and coupling constants – unlike mass-spectral fragmentation patterns - are, under readily controllable conditions, absolute physical constants that may be readily and accurately reproduced between different makes and model of instrument. Reference libraries of NMR spectra of metabolites, acquired under standardized conditions, are therefore of broad utility by the metabolomics research community. The major sources of standardized NMR spectra for metabolomics are the Madison Metabolomics Consortium Database (Cui et al., 2008), the Biological Magnetic Resonance Bank (Ulrich et al., 2008) and the HMDB (Wishart et al., 2007). These are detailed shortly.

3.4 Reference data for Gas-Chromatography / Mass-Spectrometry (GC/MS)

The most useful reference data for GC/MS are downloadable MSRI libraries. These are libraries of mass-spectra and retention indices for peaks observed in GC/MS chromatograms obtained by the GC/MS analysis of pure compounds and biological samples under standardised conditions (Kopka et al., 2005; Schauer et al., 2005). When the same standardized conditions are employed for GC/MS analysis in different laboratories, a single common MSRI library can be used for the high-confidence identification of common metabolite signals in those different labs (Schauer et al., 2005). Researchers setting up new GC/MS metabolomics platforms are advised to consider adopting a standardised GC/MS protocol already supported by a publicly-available MSRI library such as those available from the Golm Metabolome Database (Kopka et al., 2005) or MetabolomeExpress (Carroll et al., 2010) since this will enable them to share MSRI libraries with those labs and benefit from ongoing efforts to extend those libraries and annotate the large number of 'unknown' metabolites detected in GC/MS chromatograms of biological samples.

3.5 Reference data for liquid chromatography-MS, MS/MS and MS^n

While the low-cost and operational simplicity of GC/MS has led it to become the most widely employed analytical platform in metabolomics, an increasing number of laboratories are adopting complementary techniques based on liquid chromatography (LC)- and direct infusion (DI)/MS methods that employ different ionisation techniques and more advanced mass-spectrometers capable of MS, MS/MS, MS^3 and MS^n modes of analysis together with much higher mass accuracy and resolution than is provided by most standard GC-MS systems. In the paragraphs below, the various types of non GC/MS, MS-based metabolomics techniques such as LC/MS, DI/MS and capillary electrophoresis (CE)/MS including tandem MS and MS^n methods will be referred to collectively as "LC/MS" techniques.

While GC/MS metabolomics is dominated almost entirely by electron impact ionisation (EI) methods using the industry-standardised ionisation energy of 70eV, yielding highly-reproducible fragmentation spectra between different GC/MS instruments, such broad standardisation has not occurred for LC/MS. For LC/MS, the enormous diversity of mass-spectrometer types, combined with a lack of highly-developed LC 'retention-index' systems present significant challenges towards the creation of standardized MSRI reference libraries, analogous to those available for GC/MS, capable of unambiguous cross-laboratory peak identification for LC/MS.

The simplest type of online reference data for LC/MS metabolomics are the accurate, monoisotopic masses and molecular formulas of metabolites and, in some cases, their stable-isotope-labelled isotopomers. The data-processing packages provided with MS instruments capable of high-accuracy mass measurements generally allow users to create custom libraries of accurate masses and/or molecular formulas (for improved match scoring based on the shapes of isotopic envelopes) for target analytes to assist with peak identification. Although accurate masses or molecular formulas alone are not sufficient to unambiguously identify metabolite signals (due to the high frequency of structural isomers across nature), using these data in a rational manner can often provide valuable clues about the possible identities of peaks.

A good way of reducing (*but not eliminating*) ambiguity in accurate mass-based assignments is to build a separate accurate mass library for each biological system under investigation and to include in each library only those metabolites for which literature evidence exists to support their presence in that organism. An easy way of doing this is to use the advanced query tool provided with each of the BioCyc family of metabolic pathway databases (of which there are many). While the metabolite sets thus obtained may not be complete, this is a fast way of obtaining a good quality starting set.

Another approach for reducing ambiguity in LC/MS peak identifications is to use MS/MS spectral similarity as a scoring parameter to complement accurate-mass MS based assignments (see (Matsuda et al., 2009; Matsuda et al., 2010) for good examples). The major online sources of MS/MS spectra for metabolites are MassBank (Horai et al., 2010), METLIN (Smith et al., 2005), ReSpect for Phytochemicals (http://spectra.psc.riken.jp/menta.cgi/index) and the HMDB (Wishart et al., 2007). These databases each have different strengths and limitations which will be outlined shortly. With the notable exception of ReSpect for Phytochemicals, a drawback that these databases share is a lack of support for bulk

downloading of spectra. That said, MassBank does provide a powerful API to partially overcome the need for bulk download while the METLIN website currently reports that an API is in development.

3.6 The need for chromatographic retention data in LC/MS reference databases

It is important to note that, for high-confidence peak identifications that meet minimum reporting standards outlined by the Metabolomics Standards Initiative (MSI) (Sansone et al., 2007), it is necessary to support peak identifications with an additional, orthogonal identification parameter. In the case of LC/MS, where chromatography is used, this parameter is generally retention time or relative retention time agreement with an authentic standard. Unfortunately, there appear to be few if any LC/MS reference databases that provide retention time or relative retention time information. Absolute retention times vary from instrument to instrument and from column to column (even between columns of the same make and model), and are therefore considered to be of limited use for high-confidence inter-laboratory peak identification. However, *relative* retention times (or *retention indices*), where the retention time of each peak is expressed relative to one or two other peaks in the same chromatogram, are far more stable (Tarasova et al., 2009) and may provide an avenue to the compilation of LC-MS reference libraries capable of providing MSI-compliant peak identifications by combining accurate mass MS or MS/MS spectra with meaningful and highly reproducible retention index (RI) properties. Complementary to this approach would be the further development of RI-prediction models that can accurately predict the LC retention indices of metabolites based on their structures (Hagiwara et al., 2010).

It is important to note that sufficient RI reproducibility may only be achievable with certain simple types of stationary and mobile phase combinations whereby a single stationary phase interaction mechanism (eg. hydrophobic interactions in C18 reversed-phase chromatography or hydrogen-bonding interactions in silanol based normal phase chromatography) applies to all analytes. In separations over mixed-mode stationary phases where multiple interaction mechanisms occur, there is more potential for variations in chromatographic conditions to differentially affect different peaks, thus changing their relative retention times. Public databases of "Accurate Mass / retention Time (AMT) tags" are playing increasingly important roles in peptide identification in LC-MS proteomics (Hagiwara et al., 2010). A similar trend is to be expected in metabolomics.

3.7 Major online sources of analytical reference spectra for metabolomics

3.7.1 Madison Metabolomics Consortium Database (MMCD)

Description: An analytical reference database and signal-matching tool for metabolomics.

Species: Not species-constrained

Reference data: Standardized NMR spectra for 791 different metabolites (^{1}H, ^{13}C, DEPT90, DEPT135, [^{1}H, ^{1}H] TOCSY and [^{1}H, ^{13}C] HSQC). General information on >20 000 metabolites.

Noteworthy features: NMR spectrum-based search. Batch search capability.

Modes of access: browse, search and download individual spectra via web interface. Bulk FTP download of raw spectra via BMRB FTP site.

Strengths: Enormous resource for NMR metabolomics. Includes a wide range of metabolites including those that don't occur in humans (eg. plant-specific metabolites). Spectral matching tools provide batch-processing capability.

Limitations: No support for bulk download of metabolite information based on complex query

Reference: (Cui et al., 2008)

URL: http://mmcd.nmrfam.wisc.edu

3.7.2 Human Metabolome Database (HMDB)

Description: A comprehensive, freely-available knowledgebase of human metabolite information.

Species: Human (*Homo sapiens*)

Reference data: Standardized MS/MS and NMR spectra (^{1}H, ^{13}C, ^{13}C HSQC, TOCSY) for >780 metabolites. GC/MS MSRI reference data on ~300 metabolites.

Noteworthy features: NMR, MS/MS and GC/MS spectrum-based search

Modes of access: browse, search and bulk download (bulk download of MS/MS spectra only provides *images* of spectra).

Strengths: A large set of standardized NMR and GC/MS spectra help new labs to quickly set up metabolite profiling platforms.

Limitations: No support for bulk download of metabolite information based on complex query. No batch-processing capabilities for spectral matching. No API for integration with other web tools.

Reference: (Wishart et al., 2007)

URL: http://www.hmdb.ca

3.7.3 METLIN

Description: A repository for metabolite information and tandem mass spectrometry data.

Species: Not formally species-constrained but is fairly human-centric

Reference data: Accurate masses of >44000 metabolites. >28000 high-resolution Quadrupole/Time-Of-Flight (Q/TOF) MS/MS spectra for ~5000 metabolites. Multiple collision energies.

Noteworthy features: Batch searching of mzXML MS/MS files against the database. Integration with XCMS LC/MS data-processing pipeline. Neutral loss search.

Modes of access: Search only. API in development

Strengths: A large set of standardized NMR and GC/MS spectra help new labs to quickly set up metabolite profiling platforms.

Limitations: No bulk-download (must be purchased from instrument manufacturer).

Reference: (Smith et al., 2005)

URL: http://metlin.scripps.edu

3.7.4 MassBank

Description: A repository for mass-spectra of pure compounds. Features a unique design involving a centralised interface but a distributed network of data servers providing the mass-spectra.

Species: Not species constrained. Not limited to biological metabolites.

Reference data: >29000 mass spectra from a wide range of instrument types including, but not limited to, GC/MS, LC/MS and LC-MS/MS.

Noteworthy features: Batch searching of MS/MS files against the database. Neutral loss search. Most sophisticated and powerful spectral search and visualisation capabilities of all available mass-spectral repositories.

Modes of access: Search, browse and API.

Strengths: Many spectra, powerful search capabilities.

Limitations: No bulk-download. However, individual spectra may be downloaded in text format.

Reference: (Horai et al., 2010)

URL: http://www.massbank.jp/

3.7.5 ReSpect for Phytochemicals

Description: An interactive collection of MS^n spectra of plant metabolites, collected by the LC/MS metabolomics group of the RIKEN Plant Science Center.

Species: Plant species.

Reference data: A total of >8500 MS/MS spectra including >3000 spectra from the literature, >4000 triple quadrupole MS/MS spectra corresponding to >861 standard compounds and >1000 Q/TOF spectra corresponding to >550 standard compounds. Includes both +ve and –ve ionization modes.

Noteworthy features: Spectral search online using cosine method

Modes of access: Search, browse and complete download.

Strengths: Contains many plant-specific spectra not available elsewhere. Free for bulk download.

Limitations: No API. No batch search capability.

URL: http://spectra.psc.riken.jp/menta.cgi/index

3.7.6 MS-MS Fragment Viewer 1.0

Description: A database of Liquid Chromatography Fourier Transfer Ion Cyclotron Resonance Mass Spectroscopy (LC-FT/ICR-MS), Ion-Trap Tandem Mass Spectroscopy (IT-MS/MS), Fourier Transform Tandem Mass Spectroscopy (FT-MS/MS) and photodiode array (PDA) spectra with predicted structures of fragment ions observed in LC-FT/ICR-MS.

Species: Plant species.

Reference data: Spectral data for 116 different flavonoids.

Noteworthy features: Predicted structures of fragment ions observed in LC-FT/ICR-MS.

Modes of access: Search only.

Strengths: Ultra-high mass accuracy of FT/ICR-MS.

Limitations: Limited range of spectra. Search only. No browse capability means it is impossible to know what to search for. No spectral-based searching.

URL: http://webs2.kazusa.or.jp/msmsfragmentviewer/

3.7.7 MoTo DB

Description: A liquid chromatography-mass spectrometry-based metabolome database for tomato

Species: Tomato (*Solanum lycopersicum)*

Reference data: Masses, retention times, UV/Vis properties and MS/MS fragment information for a range of metabolites reported to occur in tomato plants.

Noteworthy features: Includes retention times.

Modes of access: Search only.

Strengths: Provides literature references to support peak annotations.

Limitations: Very limited search capability. No browse capability. No download.

Reference: (Moco et al., 2006)

URL: http://appliedbioinformatics.wur.nl/moto/

3.7.8 The Golm Metabolome Database (GMD)

Description: An interactive and downloadable database of electron impact (EI) ionization mass-spectra and associated retention indices of metabolite peaks detected by GC-EI-Quadrupole (GC-EI-Q-MS) and GC-EI-Time Of Flight (GC-EI-TOF-MS) instruments operated under standardized conditions.

Species: Not formally species-constrained but is plant-centric.

Reference data: Contains MSRI data for ~4500 analytes (different chemical derivatives) corresponding to ~1500 metabolites

Noteworthy features: Decision-tree based substructure prediction.

Modes of access: Search, browse and API

Strengths: Very comprehensive. Free for download. Well curated and supported.

Limitations: Does not provide innate support for sharing of MSRI libraries by arbitrary users.

Reference: (Kopka et al., 2005)

URL: http://gmd.mpimp-golm.mpg.de/Default.aspx

3.7.9 MetabolomeExpress

Description: An interactive database of downloadable MSRI libraries, raw and processed GC/MS metabolite profiling datasets and a database of metabolic phenotypes observed in any organism using any analytical technique. Includes a complete GC/MS data processing pipeline and cross-study data mining tools.

Species: Not formally species-constrained but current content is plant-centric.

Reference data: A number of GC/MS MSRI libraries are downloadable from the website. Golm Metabolome Database MSRI libraries are provided for use within the data processing pipeline.

Noteworthy features: Members may independently upload their own MSRI libraries for interactive dissemination and use within the GC/MS data-processing pipeline.

Modes of access: browse and FTP

Strengths: Libraries free for download. Provides a built-in GC/MS data processing pipeline.

Limitations: No API. No search.

Reference: (Carroll et al., 2010)

URL: http://www.metabolome-express.org

4. Web-based data processing pipelines for metabolomics

4.1 Background

Less than a decade ago, software packages enabling processing and analysis of metabolomics datasets were restricted to a limited range of desktop software programs. Would-be metabolomics researchers would have to download or purchase and install software on local computers, set up local reference libraries for peak identification and sometimes develop custom in-house computer scripts to adapt the outputs of various programs into the formats required by programs used for downstream analysis. These challenges were compounded by the fact that available programs often lacked the kinds of specialised, biology-related features desirable for metabolomics research. However, the understandable widespread dissatisfaction of metabolomics researchers with this situation has, over the last decade, driven rapid development of powerful online, platform-independent data processing pipelines tailored

towards the needs of metabolomics research. Thanks to the availability of these packages and the availability of standardised analytical reference libraries, it is now quite feasible for researchers with limited experience to conduct detailed processing and analysis of their instrumental datasets with little more than a fast internet connection, an up-to-date web-browser and, in some cases, an FTP-client program for uploading data. This section will provide an overview of the types of data processing pipelines that are currently accessible online and compare the most powerful examples in more detail.

4.2 Functions carried out by online data-processing tools

Any ideal metabolomics data-processing pipeline, whether online or offline, should be able to: a) identify and quantify biologically-relevant signals from raw instrument files and distinguish them from biologically irrelevant signals; b) identify non-redundant metabolite signals and, where possible, annotate them with their molecular identities; c) assemble a [metabolite x sample] data matrix appropriately normalised to sample volumes, internal standards and/or other useful normalisation factors; d) facilitate determination and statistical analysis of relative metabolite levels between sample classes; e) carry out multivariate analyses such as principal components analysis (PCA), hierarchical clustering analysis (HCA) and partial least squares discriminant analysis (PLS-DA); and f) provide facilities to assist biological interpretation of results (eg. mapping of detected metabolite responses onto metabolic pathways, over-representation analysis and biomarker detection). While the vast majority of online metabolomics data-processing tools carry out only one or a few of these functions, there are systems capable of carrying out all of these functions. The functionalities of a variety of web-based data processing tools for metabolomics are summarised in Table 2.

5. Online metabolomics data repositories

5.1 Background

The long-standing scientific tradition of openly disclosing supporting primary data whenever scientific claims are made has been a fundamental factor underlying the credibility of science. However, in more recent years, the scale and complexity of primary datasets has risen dramatically, presenting ever-new challenges to this tradition with the widespread emergence of high-throughput metabolomics technologies in bioscience being a good example.

In this author's view, it is absolutely crucial that the culture of open primary data disclosure is maintained, and that "challenges" should not become "excuses". Even in the extreme case of next-generation DNA sequencing where the sizes of typical primary datasets (after parsing of raw image data) are typically measured in the 10's of gigabytes (at least 10 times larger than typical metabolomics datasets), scientists have risen to the challenge by providing online storage space and developing specialised data repository systems capable of systematically archiving and effectively disseminating these data (Kaminuma et al., 2010; Cochrane et al., 2011; Leinonen et al., 2011).

Given the relatively small sizes of metabolomics datasets and the fact that metabolomics techniques predate next-generation sequencing by a considerable number of years, it is difficult to think of a satisfactory justification for the number of scientific claims that have been made on the basis of metabolomics datasets that have not, at the very least, been made freely available for download from a publicly accessible web site. That said, recent years

Feature / Tool Name:	Metabolome Express	XCMS Online	Metabo Analyst	MeltDB	metDAT 2.0
Reference:	(Carroll et al., 2010)	(Smith et al., 2006)	(Xia et al., 2009)	(Neuweger et al., 2008)	(Biswas et al., 2010)
Raw data processing	GC/MS	LC/MS	GC/MS, LC/MS, NMR	GC/MS, LC/MS	DI/MS
Raw data visualization	Interactive	Static		Static	Static
Peak detection	+	+	+	+	binning
Peak ID method	MSRI	accurate mass		MS	accurate mass
Peak ID ambiguity	Unambiguous	Ambiguous		Ambiguous	Ambiguous
Peak alignment	identification-based	COW	COW (XCMS)	COW (XCMS)	+
Processed data handling					
Data matrix construction	+	+	+	+	+
Normalisation	+		+	+	+
Fold change calculation	+	+	+	+	+
Univariate statistics	t-test	t-test	t-test, ANOVA	+	+
Correlation analysis	+		+	+	+
Multivariate analysis					
PCA	+		+	+	+
PLS-DA			+	+	+
Cluster analysis	+		+	+	+
Classification and feature analysis			+	+	
Biological interpretation					
Pathway mapping	In development			+	+
Phenotype recognition	+				
ORA	In development		+		

Table 2. Comparison of features of major web-based data processing pipelines for metabolomics. Only tools capable of some level of raw data processing have been included. A '+' indicates the presence of a feature. COW = Correlation Optimized Warping (the algorithm used by XCMS); MSRI = Mass Spectral and Retention Index; PCA = Principal Components Analysis; PLS-DA=Partial Least Squares-Discriminant Analysis; ORA=Over-representation Analysis; DI/MS = Direct Infusion / Mass Spectrometry

have seen a strong increase in the number of metabolomics labs sharing primary datasets from their own websites and even the emergence of centralized metabolomics data repositories allowing arbitrary labs to share their datasets publicly without even having to set up their own website. These groups that have been voluntarily driving the free and open dissemination of primary metabolomics data should be commended! The following sections will highlight the data sharing efforts that have been made by individual groups within the metabolomics community and describe the centralized metabolomics data repositories that are currently in operation and/or development.

5.2 Online databases sharing raw and/or partially-processed experimental datasets

5.2.1 DROP met: Data resources of plant metabolomics

Description: A part of the PRIMe (Platform for RIKEN Metabolomics) website.

Species: Plant species

Reference data: Provides a simple download page allowing free download of raw and/or processed LC/MS and GC/MS datasets and metadata from 8 different peer-reviewed publications emerging from the RIKEN Plant Science Center.

Noteworthy features: Metadata for each raw data file is provided in a systematic, MSI-compliant format.

Modes of access: browse

Strengths: Data are easy to find and well annotated.

Limitations: Metabolic phenotypes are not stored in a database. There is no way of querying the data without downloading, extracting biological information and importing into a local database.

URL: http://prime.psc.riken.jp/?action=drop_index

5.2.2 KomicMarket (Kazusa omics data market)

Description: A freely accessible database of annotations of metabolite peaks from FT-ICR-MS analysis of standard compounds and plant samples.

Species: Plant species

Reference data: Metabolites detected in tomato fruits by FT-ICR-MS. 215 standard compounds detected by FT-ICR-MS.

Noteworthy features: None

Modes of access: Search, browse and API.

Strengths: Good collection of high mass-accuracy flavonoid spectra. API makes download of spectra and associated annotations relatively easy.

Limitations: No bulk download of spectra needlessly makes access to spectra more challenging.

Reference: (Iijima et al., 2008)

URL: http://webs2.kazusa.or.jp/komics/index.php

5.2.3 MassBase 1.0

Description: A mass-spectral tag archive for metabolomics.

Species: Plant species

Reference data: Provides raw and processed GC/MS, LC/MS and CE/MS data for download.

Noteworthy features: None

Modes of access: Search, browse and download.

Strengths: One of very few sites to archive and disseminate raw chromatograms.

Limitations: No bulk-download of data – data files must be downloaded one at a time via the web interface. Chromatograms provided in proprietary binary formats. Limited metadata provided. Peak annotations are not provided.

URL: http://webs2.kazusa.or.jp/massbase/index.php/

5.2.4 SetupX

Description: A study design database for GC/MS metabolomics experiments.

Species: Plant species

Reference data: Provides raw and processed GC/MS data for download together with metadata.

Noteworthy features: Metabolite detections are searchable by species and species are searchable by metabolite detections.

Modes of access: Search, browse and download.

Strengths: One of very few sites to archive and disseminate raw chromatograms. Experimental datasets may be downloaded as single zipped files.

Limitations: Enormous sizes of zipped experimental dataset files means that download errors frequently occur during long downloads. No quantitative information is provided with metabolite detections and there is no way to compare the results of different experiments.

Reference: (Scholz and Fiehn, 2007)

URL: http://fiehnlab.ucdavis.edu:8080/m1/

5.2.5 PlantMetabolomics.org

Description: A database of processed, large-scale metabolic phenotype information obtained from an array of different *Arabidopsis thaliana* T-DNA insertion mutants.

Species: Arabidopsis thaliana

Reference data: Provides relative metabolite levels of a large number of metabolites in a large number of *Arabidosis thaliana* mutants.

Noteworthy features: None.

Modes of access: Search, browse and download.

Strengths: Data on a very wide range of metabolites. Incorporates phenotypic notes on mutants.

Limitations: Important metadata fields are frequently left empty. Raw data files are not provided. Origins of processed results are not transparent. There is no way to align and compare global phenotypes of mutants.

Reference: (Bais et al., 2010)

URL: http://plantmetabolomics.vrac.iastate.edu/ver2/index.php

5.2.6 Mery-B

Description: A repository for plant metabolomics datasets including experimental metadata processed data and raw data for NMR experiments.

Species: Plants.

Reference data: Provides NMR-based metabolite quantification data for a variety of tissues from a variety of species grown under a variety of conditions. Based on ~1000 spectra. Chemical shift peak assignment information is provided.

Noteworthy features: Interactive raw data viewers for 1D NMR and GC/MS data.

Modes of access: Search and browse.

Strengths: Contains data from a range of peer-reviewed publications and references to literature are clearly presented. Raw NMR spectra and GC chromatograms are available for visualisation. All experimental protocols are provided.

Limitations: Tools for statistical analysis are not yet functional. Data are not downloadable for offline analysis. Analytical reference libraries are not provided. Peak assignments are not seamlessly integrated into the raw data viewer. No direct links between statistical results and raw data vizualisation. Interface is not very intuitive.

Reference: (Ferry-Dumazet et al., 2011)

URL: http://www.cbib.u-bordeaux2.fr/MERYB/home/home.php

5.2.7 MetabolomeExpress

Description: An interactive, centralized metabolomics data repository for metabolomics data from all organisms and all analytical platforms that provides a variety of cross-study data-mining tools for analysis of metabolic phenotypes. Processed data may be uploaded in a simple tab-delimited format. Alternatively, raw GC/MS data may be uploaded and

processing online using the integrated data-processing pipeline before being imported into the data repository.

Species: Not formally species-constrained but current content is plant-centric. Data from other systems is currently being gathered from the literature.

Reference data: MSRI libraries, GC/MS chromatograms, processed results, metadata in systematic formats. Database currently includes >12000 publicly available metabolite response statistics representing >100 metabolic phenotypes from 8 species under 22 different experiments in 16 different peer-reviewed publications.

Noteworthy features: Members may independently upload their own MSRI libraries for interactive dissemination and use within the GC/MS data-processing pipeline. Provides tools for cross-study meta-analysis and database-driven phenotype recognition by pattern matching.

Modes of access: browse and FTP

Strengths: All public data free for download. Provides a built-in GC/MS data processing pipeline. Allows cross-study analysis. Processed metabolite response statistics are transparently linked to underlying raw data in an interactive raw data viewer.

Limitations: No API. No search. Raw data processing pipeline needs to be extended to support analytical platforms other than GC/MS. Does not provide as many multivariate analysis and classification tools as other web-based metabolomics data-processing systems.

Reference: (Carroll et al., 2010)

URL: http://www.metabolome-express.org

6. Conclusion

The field of metabolomics informatics development is moving very rapidly. New data-processing tools and new data repositories will continue to emerge. As they do, an increasingly important area to make progress in will be in the standardization of universal data exchange formats that allow free flow of data between compliant databases. Similarly important will be the development of user-friendly metadata capture tools that make systematic annotation of their datasets as painless as possible for biologists. These developments will require the development of new ontologies and/or the extension of existing ontologies that do not cover all of the terms required to describe metabolomics experiments. The efficient sharing and mining of well-annotated and well-quality-controlled metabolomics data across the internet will undoubtedly lead to many important discoveries in the future.

7. References

Bais P, Moon SM, He K, Leitao R, Dreher K, Walk T, Sucaet Y, Barkan L, Wohlgemuth G, Roth MR, Wurtele ES, Dixon P, Fiehn O, Lange BM, Shulaev V, Sumner LW, Welti R, Nikolau BJ, Rhee SY, Dickerson JA (2010) PlantMetabolomics.org: a web portal for plant metabolomics experiments. Plant physiology 152: 1807-1816

Biswas A, Mynampati KC, Umashankar S, Reuben S, Parab G, Rao R, Kannan VS, Swarup S (2010) MetDAT: a modular and workflow-based free online pipeline for mass spectrometry data processing, analysis and interpretation. Bioinformatics 26: 2639-2640

Carroll AJ, Badger MR, Harvey Millar A (2010) The MetabolomeExpress Project: enabling web-based processing, analysis and transparent dissemination of GC/MS metabolomics datasets. BMC Bioinformatics 11: 376

Caspi R, Altman T, Dale JM, Dreher K, Fulcher CA, Gilham F, Kaipa P, Karthikeyan AS, Kothari A, Krummenacker M, Latendresse M, Mueller LA, Paley S, Popescu L, Pujar A, Shearer AG, Zhang P, Karp PD (2010) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic acids research 38: D473-479

Cochrane G, Karsch-Mizrachi I, Nakamura Y (2011) The International Nucleotide Sequence Database Collaboration. Nucleic acids research 39: D15-D18

Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, Caudy M, Garapati P, Gopinath G, Jassal B, Jupe S, Kalatskaya I, Mahajan S, May B, Ndegwa N, Schmidt E, Shamovsky V, Yung C, Birney E, Hermjakob H, D'Eustachio P, Stein L (2011) Reactome: a database of reactions, pathways and biological processes. Nucleic acids research 39: D691-697

Cui Q, Lewis IA, Hegeman AD, Anderson ME, Li J, Schulte CF, Westler WM, Eghbalnia HR, Sussman MR, Markley JL (2008) Metabolite identification via the Madison Metabolomics Consortium Database. Nature biotechnology 26: 162-164

Degtyarenko K, de Matos P, Ennis M, Hastings J, Zbinden M, McNaught A, Alcantara R, Darsow M, Guedj M, Ashburner M (2008) ChEBI: a database and ontology for chemical entities of biological interest. Nucleic acids research 36: D344-350

Ferry-Dumazet H, Gil L, Deborde C, Moing A, Bernillon S, Rolin D, Nikolski M, de Daruvar A, Jacob D (2011) MeRy-B: a web knowledgebase for the storage, visualization, analysis and annotation of plant NMR metabolomic profiles. BMC plant biology 11: 104

Hagiwara T, Saito S, Ujiie Y, Imai K, Kakuta M, Kadota K, Terada T, Sumikoshi K, Shimizu K, Nishi T (2010) HPLC Retention time prediction for metabolome analysi. Bioinformation 5: 255-258

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 MY, Nakanishi H, Ikeda K, Akimoto N, Maoka T, Takahashi H, Ara T, Sakurai N, Suzuki H, Shibata D, Neumann S, Iida T, Funatsu K, Matsuura F, Soga T, Taguchi R, Saito K, Nishioka T (2010) MassBank: a public repository for sharing mass spectral data for life sciences. Journal of mass spectrometry : JMS 45: 703-714

Iijima Y, Nakamura Y, Ogata Y, Tanaka Ki, Sakurai N, Suda K, Suzuki T, Suzuki H, Okazaki K, Kitayama M, Kanaya S, Aoki K, Shibata D (2008) Metabolite annotations based on the integration of mass spectral information. The Plant Journal 54: 949-962

Kaminuma E, Mashima J, Kodama Y, Gojobori T, Ogasawara O, Okubo K, Takagi T, Nakamura Y (2010) DDBJ launches a new archive database with analytical tools for next-generation sequence data. Nucleic acids research 38: D33-D38

Kopka J, Schauer N, Krueger S, Birkemeyer C, Usadel B, Bergmuller E, Dormann P, Weckwerth W, Gibon Y, Stitt M, Willmitzer L, Fernie AR, Steinhauser D (2005) GMD@CSB.DB: the Golm Metabolome Database. Bioinformatics 21: 1635-1638

Leinonen R, Sugawara H, Shumway M (2011) The Sequence Read Archive. Nucleic acids research 39: D19-D21

Matsuda F, Hirai MY, Sasaki E, Akiyama K, Yonekura-Sakakibara K, Provart NJ, Sakurai T, Shimada Y, Saito K (2010) AtMetExpress Development: A Phytochemical Atlas of Arabidopsis Development. Plant Physiology 152: 566-578

Matsuda F, Yonekura-Sakakibara K, Niida R, Kuromori T, Shinozaki K, Saito K (2009) MS/MS spectral tag-based annotation of non-targeted profile of plant secondary metabolites. Plant J 57: 555-577

Moco S, Bino RJ, Vorst O, Verhoeven HA, de Groot J, van Beek TA, Vervoort J, de Vos CH (2006) A liquid chromatography-mass spectrometry-based metabolome database for tomato. Plant Physiology 141: 1205-1218

Neuweger H, Albaum SP, Dondrup M, Persicke M, Watt T, Niehaus K, Stoye J, Goesmann A (2008) MeltDB: a software platform for the analysis and integration of metabolomics experiment data. Bioinformatics 24: 2726-2732

Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic acids research 27: 29-34

Sakurai N, Ara T, Ogata Y, Sano R, Ohno T, Sugiyama K, Hiruta A, Yamazaki K, Yano K, Aoki K, Aharoni A, Hamada K, Yokoyama K, Kawamura S, Otsuka H, Tokimatsu T, Kanehisa M, Suzuki H, Saito K, Shibata D (2010) KaPPA-View4: a metabolic pathway database for representation and analysis of correlation networks of gene co-expression and metabolite co-accumulation and omics data. Nucleic acids research

Sansone SA, Fan T, Goodacre R, Griffin JL, Hardy NW, Kaddurah-Daouk R, Kristal BS, Lindon J, Mendes P, Morrison N, Nikolau B, Robertson D, Sumner LW, Taylor C, van der Werf M, van Ommen B, Fiehn O (2007) The metabolomics standards initiative. Nature biotechnology 25: 846-848

Schauer N, Steinhauser D, Strelkov S, Schomburg D, Allison G, Moritz T, Lundgren K, Roessner-Tunali U, Forbes MG, Willmitzer L, Fernie AR, Kopka J (2005) GC-MS libraries for the rapid identification of metabolites in complex biological samples. FEBS Lett 579: 1332-1337

Scholz M, Fiehn O (2007) SetupX--a public study design database for metabolomic projects. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing: 169-180

Shinbo Y, Nakamura Y, Altaf-Ul-Amin M, Asahi H, Kurokawa K, Arita M, Saito K, Ohta D, Shibata D, Kanaya S (2006) KNApSAcK: A Comprehensive Species-Metabolite Relationship Database. *In* K Saito, RA Dixon, L Willmitzer, eds, Plant Metabolomics, Vol 57. Springer Berlin Heidelberg, pp 165-181

Skogerson K, Wohlgemuth G, Barupal DK, Fiehn O (2011) The volatile compound BinBase mass spectral database. BMC Bioinformatics 12: 321

Smith CA, O'Maille G, Want EJ, Qin C, Trauger SA, Brandon TR, Custodio DE, Abagyan R, Siuzdak G (2005) METLIN: a metabolite mass spectral database. Therapeutic drug monitoring 27: 747-751

Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Analytical chemistry 78: 779-787

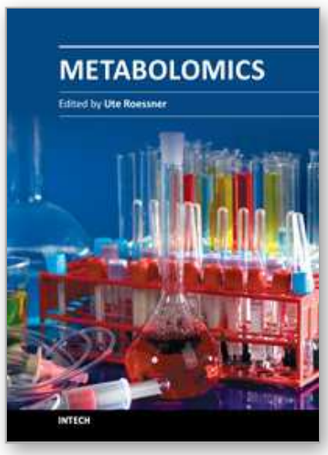
Tarasova IA, Guryca V, Pridatchenko ML, Gorshkov AV, Kieffer-Jaquinod S, Evreinov VV, Masselon CD, Gorshkov MV (2009) Standardization of retention time data for AMT tag proteomics database generation. Journal of Chromatography B 877: 433-440

Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z, Nakatani E, Schulte CF, Tolmie DE, Kent Wenger R, Yao H, Markley JL (2008) BioMagResBank. Nucleic acids research 36: D402-408

Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, Cheng D, Jewell K, Arndt D, Sawhney S, Fung C, Nikolai L, Lewis M, Coutouly MA, Forsythe I, Tang P, Shrivastava S, Jeroncic K, Stothard P, Amegbey G, Block D, Hau DD, Wagner J, Miniaci J, Clements M, Gebremedhin M, Guo N, Zhang Y, Duggan GE, Macinnis GD, Weljie AM, Dowlatabadi R, Bamforth F, Clive D, Greiner R, Li L, Marrie T, Sykes BD, Vogel HJ, Querengesser L (2007) HMDB: the Human Metabolome Database. Nucleic acids research 35: D521-526

Wohlgemuth G, Haldiya PK, Willighagen E, Kind T, Fiehn O (2010) The Chemical Translation Service--a web-based tool to improve standardization of metabolomic reports. Bioinformatics 26: 2647-2648

Xia J, Psychogios N, Young N, Wishart DS (2009) MetaboAnalyst: a web server for metabolomic data analysis and interpretation. Nucleic acids research 37: W652-660



Metabolomics

Edited by Dr Ute Roessner

ISBN 978-953-51-0046-1
Hard cover, 364 pages
Publisher InTech
Published online 10, February, 2012
Published in print edition February, 2012

Metabolomics is a rapidly emerging field in life sciences, which aims to identify and quantify metabolites in a biological system. Analytical chemistry is combined with sophisticated informatics and statistics tools to determine and understand metabolic changes upon genetic or environmental perturbations. Together with other 'omics analyses, such as genomics and proteomics, metabolomics plays an important role in functional genomics and systems biology studies in any biological science. This book will provide the reader with summaries of the state-of-the-art of technologies and methodologies, especially in the data analysis and interpretation approaches, as well as give insights into exciting applications of metabolomics in human health studies, safety assessments, and plant and microbial research.

IntechOpen