Phytochemistry 70 (2009) 450–456

Contents lists available at ScienceDirect

Phytochemistry

journal homepage: www.elsevier.com/locate/phytochem

Update in Bioinformatics

Web-based resources for mass-spectrometry-based metabolomics: A user's guide

Takayuki Tohge, Alisdair R. Fernie *

Max-Planck-Institut für Molekulare Pflanzenphysiologie, Am Mühlenberg 1, 14476 Potsdam-Golm, Germany

ARTICLE INFO

Article history:
Received 6 February 2009
Accepted 10 February 2009
Available online 13 March 2009

Keywords:
Bioinformatics
Web-resource
Metabolomics
LC–MS
GC–MS
Peak identification
Peak alignment

ABSTRACT

In recent years, a plethora of web-based tools aimed at supporting mass-spectrometry-based metabolite profiling and metabolomics applications have appeared. Given the huge hurdles presented by the chemical diversity and dynamic range of the metabolites present in the plant kingdom, profiling the levels of a broad range of metabolites is highly challenging. Given the scale and costs involved in defining the plant metabolome, it is imperative that data are effectively shared between laboratories pursuing this goal. However, ensuring accurate comparison of samples run on the same machine within the same laboratory, let alone cross-machine and cross-laboratory comparisons, requires both careful experimentation and data interpretation. In this review, we present an overview of currently available software that aids either in peak identification or in the related field of peak alignment as well as those with utility in defining structural information of compounds and metabolic pathways.

© 2009 Elsevier Ltd. All rights reserved.

1. Introduction

In the post-genome era, there has been a dramatic surge in the number of research groups studying metabolite levels by the use of the metabolomics approaches. Two essentially complementary approaches are commonly adopted by those reliant on nuclear magnetic resonance (NMR) and by those reliant on mass spectrometry (MS).

In parallel, bioinformaticians have attempted to compute the size of the metabolome on the basis of genome information (Nobeli et al., 2003). However, such approaches are considerably restrained by the quality of genome annotation, knowledge of precise enzymatic functionality and by the fact that certain reactions can occur within the cell in the absence of catalytic enzymes. Despite these drawbacks, and in spite of the fact that metabolomics is still very much regarded as a developing science, metabolomics has already found great utility in plants both as a diagnostic tool, as a method to discriminate genotypically distinct individuals in a population, as a component of systems biology approaches to understand diurnal metabolic regulation and response to stress, and more recently to perform analysis of natural variance in crop species as part of strategies aimed at crop improvement (see Fiehn (2002), Sumner et al. (2003), Fernie et al. (2004), Oksman-Caldentey and Saito (2005), and Schauer and Fernie (2006) for reviews). However, to correctly address such questions requires both large sample numbers and great care in ensuring data quality. As a result, a recent focus of many groups was the development of computational approaches that allow high-throughput data evaluation in a manner that does not compromise quality (Duran et al., 2003; Kopka et al., 2005; De Vos et al., 2007; Luedemann et al., 2008). Moreover, to gain ultimate value from large-scale experiments the data and associated metadata (detailing exact growth condition etc.) that they produce need to be immortalised in databases. The recently established metabolomics standards initiative provides solid guidelines for reporting standards in metabolomics (Jenkins et al., 2004; Fiehn et al., 2008), which if universally adopted should address this problem. Computational analysis of NMR spectra (often termed chemometrics) is well developed (see for example Holmes and Antti (2002)), and large databases of NMR spectra are already available (see for example Ebbels et al. (2007)). The situation for MS is not quite so advanced; however, recent years have experienced major developments and a large number of web-based tools are currently available that can greatly aid in the quality control and interpretation of gas-chromatography–MS (GC–MS), capillary electrophoresis–MS (CE–MS) and liquid-chromatography–MS (LC–MS). In this review, we detail web-resources that can facilitate MS-based metabolomics either with respect to peak identification and peak alignment or in the structural identification of compounds and/or pathways. Such resources are imperative if the plant metabolism community is to mount a serious

* Corresponding author. Tel.: +49 331 5678211; fax: +49 331 5678408.
E-mail addresses: tohge@mpimp-golm.mpg.de (T. Tohge), fernie@mpimp-golm.mpg.de (A.R. Fernie).

0031-9422/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.phytochem.2009.02.004

attempt towards tackling the grand challenge of metabolomics the identification of the metabolic complement of a plant or even of the entire plant kingdom. At present, even the combination of a wide range of analytical tools only allow us to see a small proportion of the metabolic complement of the cell (Fernie, 2007). That said, a large number of peaks in GC–MS and LC–MS are unidentified and the open sharing of information between laboratories would likely greatly enrich the information content achievable by these approaches. Given that GC–MS and LC–MS technologies are the most widely used MS technologies for metabolomics, we chose to focus our article on these; however, we additionally provide links for other contemporary technologies as well as discuss the prospects and needs of emergent and next-generation technologies.

2. Peak identification

With the exception of diagnostic approaches, which do not empirically require metabolite identifications, the first "informatic" task that metabolite profilers hit is to identify the many, many peaks resulting from the chromatographic separation afforded by gas- or liquid-chromatography. Given the different challenges presented to each technique, GC–MS and LC–MS are best discussed individually.

2.1. GC–MS peak identification

For the identification of GC–MS peaks several approaches are possible. Given that GC–MS often provides information concerning fragments of the derivatised metabolite, it is important to identify the most reliable fragment on which to quantify the metabolite. By and large this is the major fragment, when considered in terms of peak intensity. As a rule the retention time of peaks is defined in comparison to the introduced time standards (Fiehn et al., 2000), whilst the level of a given metabolite is given by reference to the level of a similarly introduced external compound (Huege et al., 2007). In comprehensive metabolic profiling of plant extracts by GC–MS, the majority of compounds measured have not yet been

Table 1
Freely and commercially available web-resources for MS-based metabolomics.

Name	Link	Note
"Freely available" software		
Pathway		
AraCyc	http://www.arabidopsis.org/biocyc/index.jsp	*Arabidopsis*
BioCyc	http://biocyc.org/	Organisms
EcoCyc	http://ecocyc.org/	*Escherichia coli*
KaPPA-View3	http://kpv.kazusa.or.jp/kpv3/guestIndex.jsp	*Arabidopsis*, rice, tomato, lotus
KEGG	http://www.genome.jp/kegg/ligand.html	Organisms
MetaCyc	http://metacyc.org/index.shtml	Organisms
MetaCrop	http://metacrop.ipk-gatersleben.de	Crops
PlantCyc	http://www.plantcyc.org/	Plants
Analytical data (samples)		
ARMeC	http://www.armec.org/MetaboliteLibrary/index.html	*Arabidopsis*, potato
KOMICS	http://webs2.kazusa.or.jp/komics/	Tomato
McGill-MD	http://metabolomics.mcgill.ca/	Crops
MotoDB	http://appliedbioinformatics.wur.nl/moto/	Tomato
Analytical data (standard compounds)		
MASSBANK	http://www.massbank.jp/index-e.html	GC–MS, LC–MS, CE–MS, MS^2
METLIN	http://metlin.scripps.edu/	LC–MS, MS^2
MS2T	http://prime.psc.riken.jp/lcms/ms2tview/ms2tview.html	LC–MS, MS^2
MSRI@CSB.DB	http://csbdb.mpimp-golm.mpg.de/csbdb/gmd/msri/gmd_msri.html	GC–MS
NIST Chem WebBook	http://webbook.nist.gov/chemistry/	GC–MS
SDBS	http://riodb01.ibase.aist.go.jp/sdbs/cgi-bin/cre_index.cgi?lang=eng	MS
Publication and literature		
BioMeta	http://biometa.cmbi.ru.nl/	
ChEBI	http://www.ebi.ac.uk/chebi/	
HMDB	http://www.hmdb.ca/	
KNApSAcK	http://kanaya.naist.jp/KNApSAcK/KNApSAcK.php	
MassTRIX	http://mips.gsf.de/proj/metabolomics/	
MMCD	http://mmcd.nmrfam.wisc.edu/	
MMsINC	http://mms.dsfarm.unipd.it/MMsINC.html	
PubChem	http://pubchem.ncbi.nlm.nih.gov/	
UCI ChemDB	http://cdb.ics.uci.edu/CHEM/Web/	
Compound family		
Flavonoid viewer	http://www.metabolome.jp/software/FlavonoidViewer/data/mass	Flavonoids
LIPID MAPS	http://www.lipidmaps.org/index.html	Lipids
LIPIDBANK	http://lipidbank.jp/	Lipids
"For a commercial fee" database		
Adams library	Identification of essential oil components by gas chromatography/quadrupole mass spectroscopy, 4th Edition (2007)	
Dictionary of natural medicines	http://dnp.chemnetbase.com/dictionary-search.do?method=view&id=2714206&si=	
DiscoveryGate	http://www.symyx.com/products/databases/discoverygate/index.jsp	
MASS Frontier	http://www.thermo.com/com/cda/product/detail/1,1055,11442,00.html	
NIST	http://www.nist.gov/srd/onlinelist.htm	
Scifinder	http://www.cas.org/SCIFINDER/SCHOLAR/index.html	
Wiley MS Libraries	http://www.sisweb.com/software/ms/wiley.htm	

formally identified. Over 400 components were detected in *Cucurbita maxima* phloem, using automated mass spectral deconvolution GC–MS, but only 90 of them were tentatively identified (Fiehn, 2003). However, the power of mass spectrometry enables relative measurements of compounds to be reliably made, even when their chemical structures are not known (Halket et al., 2005). Critical to this process is the reliance on both retention time index and spectral databases. However, both of these are publicly available, for example, through a web-based platform of MSRI libraries (www.csbdb.mpimp-golm.mpg.de/gmd.html; Kopka et al., 2005) and downloadable files, which can be imported into the current leading NIST02 mass spectral search program or AMDIS, the automated mass spectral deconvolution and identification system (National Institute of Standards and Technology, Gaithersburg, MD, USA; Ausloos et al., 1999; Stein 1999). Both software systems are publicly available from www.chemdata.nist.gov/mass-spc/amdis/ and www.chemdata.nist.gov/mass-spc/Srch_v1.7/index.html. The Golm libraries, which are also freely available, are classified according to technology and degree of manual mass spectral identification that is required for the library construction. However, a recent study provided the important illustration that irrespective of the technological platform, if the same chromatography column is used then the retention time indices are comparable (Schauer et al., 2005a). A useful introduction to library searching is given in a text by Smith and Busch (1999). In addition, detailed studies of search algorithms have been carried out (Stein and Scott, 1994; McLafferty et al., 1998). Table 1 lists common freely and commercially available libraries, respectively. The mass spectra contained within the NIST library are studied in detail by professional evaluators before inclusion (Ausloos et al., 1999), ensuring high quality. Of particular interest to metabolic profilers may well be the libraries downloadable from the Max-Planck Institute for Plant Physiology in Potsdam-Golm, Germany (http://www.mpimp-golm.mpg.de), which include information concerning derivatives both of known compounds and of a potentially useful library of unassigned spectra with compound listings including retention times.

Several different bioinformatic approaches have recently been developed to aid in GC–MS peak identifications including TagFinder (Luedemann et al., 2008). When taken together with the relative ease of acquisition of authentic standards for primary metabolites, these approaches afford relative high coverage of plant primary metabolism. Furthermore, the availability of information regarding the species studied as well as the peaks identified is readily available within databases such as the Golm Metabolome Database (Kopka et al., 2005; Hummel et al., 2008), housed in the CSB.DB in Golm (Steinhauser et al., 2005) as well as in publications such as Wagner et al. (2003) and Strehmel et al. (2008) and in the website of the Fiehn laboratory (http://fiehnlab.ucdavis.edu/). All of these are highly recommended resources for researchers intent on establishing metabolomics platforms for primary metabolism.

2.2. LC–MS peak identification

By contrast, to the situation for primary metabolism, the identification of secondary metabolites is massively hindered by the lack of available standards. It is impossible to purchase standard substances comprehensively because diversity of chemical structure is too large, and complex compounds are almost unavailable commercially and are expensive. Moreover, LC–MS studies are complicated by the fact that the levels of secondary metabolites are highly divergent between different organs, growth conditions and species (Petersen et al., 2002; Yonekura-Sakakibara et al., 2008; Hanhineva et al., 2008). For these reasons, peak identification is generally performed by the use of combinatorial strategies whereby the literature information is taken alongside available compounds in an attempt to identify specific peaks (see for example Tohge et al. (2005)). For this reason in order to predict the chemical structure of the detected peaks often requires reference to multiple databases of chemical compounds and spectra.

2.3. Metabolite databases for peak prediction

Publicly accessible databases, which are useful for the peak annotation, are shown in Table 1, here we detail the most commonly useful databases. Researchers should select the database on the basis of their purposes and type of samples analysed. Detected peaks are searched within the database by using either the *m/z* value or the compound molecular weight (MW) predicted from *m/z* value and adduct ions. For peak annotation, the most useful way is to search with information on each biosynthetic pathway, since similar metabolic species have similar structural properties and hence similar elution times and mass spectra. Preknown information concerning biosynthetic origins of compounds can thus render peak annotation relatively facile and also to minimise the chances of false annotations. Several databases are organized on the bases of biosynthetic pathways using a literature-based curation strategy. BioCyc (Karp et al., 2005), KEGG (Kanehisa and Goto, 2000; Kanehisa et al., 2006) and MetaCyc (Zhang et al., 2005; Caspi et al., 2008) contain generalised pathways for all organisms, whilst PlantCyc (http://www.plantcyc.org/) and MetaCrop (Crops; Grafahrend-Belau et al., 2008) contain a generalised pathway for all plant species. By contrast, AraCyc (*Arabidopsis*; Zhang et al., 2005), EcoCyc (*Escherichia coli*; Karp et al., 2007) and KaPPA-View3 (*Arabidopsis*, Rice, Tomato, Lotus; Tokimatsu et al., 2005) contain species-specific pathways. However, these should be used with caution since these databases are prone to contain incomplete pathways, since the databases are built solely on the basis of current literature information. Other effective databases are ARMec (*Arabidopsis* and potato; http://www.armec.org/MetaboliteLibrary/index.html), KOMICS (Tomato; Iijima et al., 2008), McGill-MD (Crops; http://metabolomics.mcgill.ca/) and MotoDB (Tomato; Moco et al., 2006), which provide information of all detected peaks obtained from analysing particular plant samples. Whilst such metabolite profiling catalogues are highly informative, it must be borne in mind that similar analytic conditions such as separation column and analytical instrument must be used. When this is done, it is relatively straightforward to share and compare the obtained results to archival data comprising both known and unknown peaks with their underlying retention time and spectra. In this vein, the databases, MASSBANK (Horai et al., 2008), METLIN (Smith et al., 2005), MS2T (Matsuda et al., 2008), MSRI@CSB.DB (Kopka et al., 2005), NIST Chem WebBook (Linstrom and Mallard, 2001) and SDBS provide highly useful resources of MS of standard substances. Since the searching function in these databases is well established, the user can either input *m/z* values, predict molecular weight, fragment ion peaks or predict compositional formula for searching. In MASSBANK, the user can also input the spectrum data as search query. HMDB (Human; Wishart et al., 2007) and KNApSAcK (Plants; Shinbo et al., 2006) are species-specific metabolite databases compiled by the literature survey. In instances in which the researcher has very little *a priori* knowledge concerning detected compounds or unknown peaks, BioMeta (Ott and Vriend, 2006), CHeBI (Degtyarenko et al., 2008), MassTRIX (Suhre and Schmitt-Kopplin, 2008), MMCD (Cui et al., 2008), MMsINC (Masciocchi et al., 2009), PubChem (Rosania et al., 2007) and UCI ChemDB (Chen et al., 2005, 2007), which list the MW of compounds represented in the literature, are highly useful for peak prediction. However, due to the immense search scale afforded by such databases, caution is required in their interpretation, and predictions on this basis alone must be regarded as "putatives". In case of targeted metabolite profiling, Flavonoid Viewer (Arita, 2008; Arita and

Suwa, 2008), LIPIDBANK (Watanabe et al., 2000) and LIPID MAPS (Fahy et al., 2005) provide wide coverage of compound families of specific pathways, and are therefore highly useful for obtaining information concerning closely related compounds. The databases Adams library, the Dictionary of Natural Medicines, DiscoveryGate, MASS Frontier, NIST, SciFinder and Wiley MS libraries are available for a commercial fee. Details concerning the size and coverage of these databases, which illustrate their potential use, are provided in Table 1. Whilst it is likely that the needs of most academic researchers will be satisfied by freely available websites, those with specific interest in more obscure compounds or in the identification of unknown structures may well want to investigate these sites.

3. Peak alignment

3.1. GC–MS peak alignment

As hinted, the above peak alignment for GC–MS evaluation is relatively well established with libraries for retention time indices being freely available in Golm (Wagner et al., 2003; Kopka et al., 2005; Schauer et al., 2005a; Hummel et al., 2008; Strehmel et al., 2008). The utility of these libraries was recently demonstrated in a multi-laboratory study analysing extracts from widely divergent tissue samples including *Aradiposis thaliana* (Schauer et al., 2005b), *Lotus japonica* (Desbrosses et al., 2005) and the *Solanum esculentum* complex (Schauer et al., 2005b). This report illustrated that mass spectral retention indices initially defined in Golm could be used to define metabolite content using a machine of similar performance in Wales, Sweden and Australia (Schauer et al., 2005a). Furthermore, it illustrated that if the same chromatography column was used, these indices were even independent of the machine type. A similar, freely available application – MS-FACTs – was developed, principally for *Medicago truncatula*, which also allows the alignment of GC–MS data, as well as other aspects of data processing (Duran et al., 2003). Furthermore, empirical computer scripts for correctly assigning peaks are described in Lisec et al. (2006), these scripts could prove highly useful especially in situations wherein a limited amount of metabolite information is required from a very large number of biological samples. The final tool that is worthy of mention here is MetAlign (Vorst et al., 2005), which was although developed for LC–MS analysis of secondary metabolites (see the following section), can also be used in GC–MS applications. However, the utility of this application is defined in detail in the following section.

3.2. LC–MS peak alignment

In contrast to the well-established processing of NMR, or even to the relatively well-established data processing for GC–MS analysis, the non-targeted LC–MS analysis is in its infancy. Several factors hamper the development of uniform processing for LC–MS. Firstly, there is a great diversity in the machine and column type under use in the different laboratories engaged in metabolomics research worldwide. Secondly, given the enormous chemical diversity apparent in secondary metabolism (De Luca and St Pierre, 2000), there are often quite diverse needs of researches at this level. However, all researchers encounter problems concerning limit of detection due to high background noise and non-uniformity of compound elution time. Thirdly, LC–MS is particularly susceptible to ion-suppression, due to its common reliance on electrospray ionisation which complicates comparison between samples of widely different compositions (Fernie et al., 2004). Whilst very recent developments of LC–MS technology based on TOF-MS (time of flight mass spectroscopy) and UPLC (ultra performance liquid-chromatography) facilitate the generation of highly sensitive analysis without delay of elution time, a number of other "solutions" exist to cope with data obtained from older instrumentation. Thus, although processing software for unbiased mass peak extraction and alignment of LC–MS data remain unresolved, freely available useful softwares such as MetAlign (Vorst et al., 2005; De Vos et al., 2007), MSFact (Duran et al. 2003), XCMS2 (Smith et al., 2006; Benton et al., 2008), BINBASE (Fiehn et al., 2005) and MZmine2 (Katajamaa and Oresic, 2005) can all facilitate comprehensive data analyses for non-targeted metabolomics approaches. These approaches all follow the same general principle for alignment, as shown in Fig. 1, which uses relatively simple chromatograms to illustrate the problem of shifting retention time and to how this can be addressed. Raw binary files are exchanged by mass spectrometry software to ANDI-NetCDF files (the standard Mass-Spectrometry Data Interchange format), for peak alignment. Subsequently, differences in elution time between each analysis are adjusted with respect to internal standards and/or to major detected peaks. During the process of peak alignment, each parameter such as noise tolerance and peak thresholds should be adjusted on the basis of judgements regarding signal intensity and peak shape of internal standard and identified compound, since the concentration of each compound and number of detected peaks per chromatogram can be highly variable between samples. It is recommended that a certain number of abundant peaks of the chromatogram are annotated prior to alignment in order to further

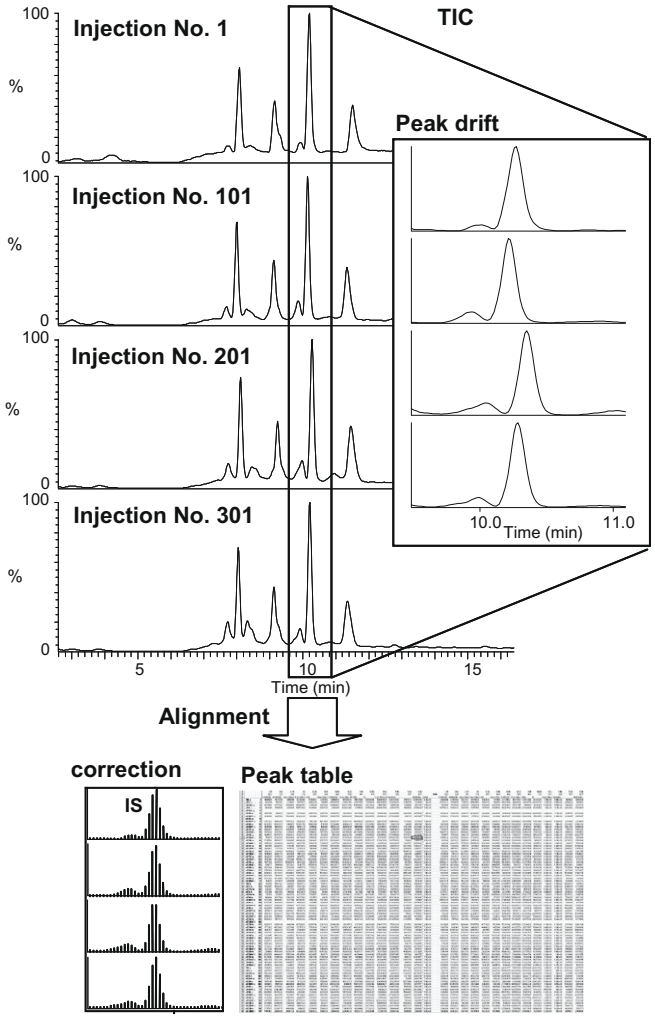

Fig. 1. Schematic overview of peak drift and correction in LC–MS.

improve this process. As an alternative approach, the comparison of samples with and without isotope labels on feeding experiment additionally allows the removal of false positive peaks and also greatly reduces the influence of ion-suppression effects (Giavalisco et al., 2008).

4. Data interpretation

The collection of information concerning metabolite abundance is a challenging task, that said, the interpretation of the results is often, at least, equally daunting. This is somewhat easier by utilizing one of the several metabolic pathway databases available such as AraCyc (Mueller et al., 2003), Kappa-Viewer (Tokimatsu et al., 2005) and MapMan (Thimm et al., 2004). These softwares afford integrated overviews of metabolic diagrams and render hypothesis/conclusion drawing far more facile. Furthermore, these databases include information on enzymatic reaction underlying these metabolic pathways and on cross-referencing with the information of chemical structure of annotated peaks aiding in gene annotation, or even in helping to clarify pathway structures (see for example Hirai et al. (2004), Tohge et al. (2005), Morikawa et al. (2006), Arita and Suwa (2008)). MapMan and Kappa-Viewer also render cross-entity comparison facile and greatly aid in systems-wide analyses of metabolic pathways or other biological processes (Thimm et al., 2004; Tokimatsu et al., 2005).

5. Emergent and next-generation technologies

Thus far, we have concentrated our discussion on GC– and LC–MS technologies. However, improvement in the coverage of metabolomics techniques will likely best be served by alternative approaches such as fourier transform-ion cyclotron–MS (FT-ICR–MS), which has already been used in plant metabolomics (Aharoni et al., 2003; Hirai et al., 2004; Tohge et al., 2005; Giavalisco et al., 2008; Iijima et al., 2008), or even in field asymmetric waveform ion mobility MS–MS (FAIMS–MS) which has yet been postulated as a tool with an impressive potential (Fernie and Schauer, 2009). These techniques provide very high accuracy mass data of each analyte. However, despite this fact, whilst the information they provide is enough to give exact elemental composition of each analyte, this alone cannot provide information concerning the exact chemical structure (Kind and Fiehn, 2006, 2007). Several recent studies have concentrated on ways to improve structural information. For this purpose two alternate strategies exist. Firstly, multi-dimensional MS analysis can be performed and the structure of the analyte can be constructed retrospectively. This approach has recently been taken for FT-ICR–MS (Ohta et al., 2007; Iijima et al., 2008; Takahashi et al., 2008), and is used in conjunction with an associated database (KNApSaCK; Shinbo et al., 2006) and its application of data processing (DrDMASS; Oikawa et al., 2006). The second approach, that of isotope labelling, has a more global application in that it allows unambiguous identification of metabolites on the basis of the incorporation of a signature mass (Giavalisco et al., 2008). Following this approach, the authors were able to provide quantitative information on the levels of some 1024 metabolites in *Arabidopsis*. In a similar approach, the natural abundance of heavy isotope was used as a reference for the identification of some 869 metabolites in tomato (IIjima et al., 2008). Whilst these two examples are by no means the first to utilize FT-ICR–MS technology, being preceded by the studies of both strawberry and *Arabidopsis* (Aharoni et al., 2003; Hirai et al., 2004; Tohge et al., 2005), they do provide a clear strategy for the unambiguous assignment of chemical structure. The recent paper of Giavalisco et al. (2008) defines a clear strategy towards improving our coverage of the metabolome. If this strategy is followed, using several independent metabolic precursors, it would appear likely that a multitude of important information concerning the regulation of primary metabolism would become accessible.

6. Conclusions

In this review, we have documented the impressive number of public and commercially available resources available, which can aid in the initial analysis of metabolomics data sets by MS-based approaches. These web-resources should provide indispensible data for the researcher's intent for setting up either GC–MS or LC–MS platforms of there own. It is worth noting that a wide range of excellent web-accessible statistical tools also exist for the interpretation of the resultant data sets (see Nobeli and Thornton (2006), Steinfath et al. (2008), Wiklund et al. (2008) for details). Despite the immense importance of such tools, we were unfortunately not able to describe them in detail due to space limitations. As a final comment to reiterate that stated already in the introduction, it is through the communal sharing of information such as that deposited in the databases cited within this article that the science of metabolomics is likely to expand and succeed. As such we would be greatly heartened to see further similar web-resources being opened in the near future.

Acknowledgements

Discussions with Joachim Kopka, Lothar Willmitzer, Patrick Giavalisco, Nicolas Schauer, Bjoern Usadel and Masanori Arita were highly useful during the writing of this review.

References

Adams, R.P., 2007. Identification of Essential Oil Components by Gas Chromatography-Quadrupole Mass Spectroscopy, 4th ed. Allured Pub Corp, Carol Stream, IL, USA.

Aharoni, A., Ric de Vos, C.H., Verhoeven, H.A., Maliepaard, C.A., Kruppa, G., Bino, R., Goodenowe, D.B., 2003. Nontargeted metabolome analysis by use of Fourier transform ion cyclotron mass spectrometry. OMICS 6, 217–234.

Arita, M., 2008. A pitfall of wiki solution for biological databases, Brief. Bioinform. (Epub ahead of print).

Arita, M., Suwa, K., 2008. Search extension transforms Wiki into a relational system: a case for flavonoid metabolite database. BioData Min. 17, 1–8.

Ausloos, P., Clifton, C.L., Lias, S.G., Mikaya, A.I., Stein, S.E., Tchekhovskoi, D.V., Sparkman, O.D., Zaikin, V., Zhu, D., 1999. The critical evaluation of a comprehensive mass spectral library. J. Am. Soc. Mass Spectrom. 10, 287–299.

Benton, H.P., Wong, D.M., Trauger, S.A., Siuzdak, G., 2008. XCMS2: processing tandem mass spectrometry data for metabolite identification and structural characterization. Anal. Chem. 80, 6382–6389.

Caspi, R., Foerster, H., Fulcher, C.A., Kaipa, P., Krummenacker, M., Latendresse, M., Paley, S., Rhee, S.Y., Shearer, A.G., Tissier, C., Walk, T.C., Zhang, P., Karp, P.D., 2008. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res. 36, D623–D631.

Chen, J., Swamidass, S.J., Dou, Y., Bruand, J., Baldi, P., 2005. ChemDB: a public database of small molecules and related chemoinformatics resources. Bioinformatics 21, 4133–4139.

Chen, J.H., Linstead, E., Swamidass, S.J., Wang, D., Baldi, P., 2007. ChemDB update – full-text search and virtual chemical space. Bioinformatics 23, 2348–2351.

Cui, Q., Lewis, I.A., Hegeman, A.D., Anderson, M.E., Li, J., Schulte, C.F., Westler, W.M., Eghbalnia, H.R., Sussman, M.R., Markley, J.L., 2008. Metabolite identification via the Madison Metabolomics Consortium Database. Nat. Biotechnol. 26, 162–164.

De Luca, V., St Pierre, B., 2000. The cell and developmental biology of alkaloid biosynthesis. Trends Plant Sci. 5, 168–173.

De Vos, R.C., Moco, S., Lommen, A., Keurentjes, J.J., Bino, R.J., Hall, R.D., 2007. Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry. Nat. Protoc. 2, 778–791.

Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M., Ashburner, M., 2008. ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Res. 36, D344–D350.

Desbrosses, G.G., Kopka, J., Udvardi, M.K., 2005. *Lotus japonicus* metabolic profiling. Development of gas chromatography–mass spectrometry resources for the study of plant–microbe interactions. Plant Physiol. 137, 1302–1318.

Duran, A.L., Yang, J., Wang, L.J., Sumner, L.W., 2003. Metabolomics spectral formatting, alignment and conversion tools (MSFACTs). Bioinformatics 19, 2283–2293.

Ebbels, T.M., Keun, H.C., Beckonert, O.P., Bollard, M.E., Lindon, J.C., Holmes, E., Nicholson, J.K., 2007. Prediction and classification of drug toxicity using probabilistic modeling of temporal metabolic data: the consortium on metabonomic toxicology screening approach. J. Proteome Res. 6, 4407–4422.

Fahy, E., Subramaniam, S., Brown, H.A., Glass, C.K., Merrill, A.H., Murphy, R.C., Raetz, C.R., Russell, D.W., Seyama, Y., Shaw, W., Shimizu, T., Spener, F., van Meer, G., VanNieuwenhze, M.S., White, S.H., Witztum, J.L., Dennis, E.A., 2005. A comprehensive classification system for lipids. J. Lipid Res. 46, 839–861.

Fernie, A.R., 2007. The future of metabolic phytochemistry: larger numbers of metabolites, higher resolution, greater understanding. Phytochemistry 68, 2861–2880.

Fernie, A.R., Schauer, N., 2009. Metabolomics-assisted breeding: a viable option for crop improvement? Trends Genet. 25, 39–48.

Fernie, A.R., Trethewey, R.N., Krotzky, A.J., Willmitzer, L., 2004. Metabolite profiling: from diagnostics to systems biology. Nat. Rev. Mol. Cell. Biol. 5, 763–769.

Fiehn, O., 2002. Metabolomics – the link between genotype and phenotype. Plant Mol. Biol. 48, 155–171.

Fiehn, O., 2003. Metabolic networks of *Cucurbita maxima* phloem. Phytochemistry 62, 875–886.

Fiehn, O., Kopka, J., Dörmann, P., Altmann, T., Trethewey, R.N., Willmitzer, L., 2000. Metabolite profiling for plant functional genomics. Nat. Biotechnol. 18, 1157–1161.

Fiehn, O., Wohlgemuth, G., Scholz, M., 2005. Setup and annotation of metabolomic experiments by integrating biological and mass spectrometric metadata. Proc. Lect. Note Bioinform. 3615, 224–239.

Fiehn, O., Wohlgemuth, G., Scholz, M., Kind, T., Lee do, Y., Lu, Y., Moon, S., Nikolau, B., 2008. Quality control for plant metabolomics: reporting MSI-compliant studies. Plant J. 53, 691–704.

Giavalisco, P., Hummel, J., Lisec, J., Inostroza, A.C., Catchpole, G., Willmitzer, L., 2008. High-resolution direct infusion-based mass spectrometry in combination with whole (13)c metabolome isotope labeling allows unambiguous assignment of chemical sum formulas. Anal. Chem. 15, 9417–9425.

Grafahrend-Belau, E., Weise, S., Koschützki, D., Scholz, U., Junker, B.H., Schreiber, F., 2008. MetaCrop: a detailed database of crop plant metabolism. Nucleic Acids Res. 36, D954–D958.

Halket, J.M., Waterman, D., Przyborowska, A.M., Patel, R.K., Fraser, P.D., Bramley, P.M., 2005. Chemical derivatization and mass spectral libraries in metabolic profiling by GC/MS and LC/MS/MS. J. Exp. Bot. 56, 219–243.

Hanhineva, K., Rogachev, I., Kokko, H., Mintz-Oron, S., Venger, I., Kärenlampi, S., Aharoni, A., 2008. Non-targeted analysis of spatial metabolite composition in strawberry (Fragariaxananassa) flowers. Phytochemistry 69, 2463–2481.

Hirai, M.Y., Yano, M., Goodenowe, D.B., Kanaya, S., Kimura, T., Awazuhara, M., Arita, M., Fujiwara, T., Saito, K., 2004. Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. USA 101, 10205–10210.

Holmes, E., Antti, H., 2002. Chemometric contributions to the evolution of metabonomics: mathematical solutions to characterising and interpreting complex biological NMR spectra. Analyst 127, 1549–1557.

Horai, H., Arita, M., Nishioka, T., 2008. Comparison of ESI-MS spectra in MassBank database. Proc. Int. Conf. Biomed. Eng. Inform. 2, 853–857.

Huege, J., Sulpice, R., Gibon, Y., Lisec, J., Koehl, K., Kopka, J., 2007. GC–EI-TOF-MS analysis of in vivo carbon-partitioning into soluble metabolite pools of higher plants by monitoring isotope dilution after 13CO2 labelling. Phytochemistry 68, 2258–2272.

Hummel, J., Selbig, J., Walther, D., Kopka, J., 2008. The Golm Metabolome Database: a database for GC–MS based metabolite profiling. In: Hohmann, S. (Ed.) Topics in Current Genetics: Nielsen, J., Jewett, M. (Eds.) Metabolomics. Springer-Verlag, Berlin, Heidelberg, New York.

Iijima, Y., Nakamura, Y., Ogata, Y., Tanaka, K., 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. Plant J. 54, 949–962.

Jenkins, H., Hardy, N., Beckmann, M., Draper, J., Smith, A.R., Taylor, J., Fiehn, O., Goodacre, R., Bino, R.J., Hall, R., Kopka, J., Lane, G.A., Lange, B.M., Liu, J.R., Mendes, P., Nikolau, B.J., Oliver, S.G., Paton, N.W., Rhee, S., Roessner-Tunali, U., Saito, K., Smedsgaard, J., Sumner, L.W., Wang, T., Walsh, S., Wurtele, E.S., Kell, D.B., 2004. A proposed framework for the description of plant metabolomics experiments and their results. Nat. Biotechnol. 22, 1601–1606.

Kanehisa, M., Goto, S., 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27–30.

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., Hirakawa, M., 2006. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. 34, D354–D357.

Karp, P.D., Ouzounis, C.A., Moore-Kochlacs, C., Goldovsky, L., Kaipa, P., Ahrén, D., Tsoka, S., Darzentas, N., Kunin, V., López-Bigas, N., 2005. Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. Nucleic Acids Res. 33, 6083–6089.

Karp, P.D., Keseler, I.M., Shearer, A., Latendresse, M., Krummenacker, M., Paley, S.M., Paulsen, I., Collado-Vides, J., Gama-Castro, S., Peralta-Gil, M., Santos-Zavaleta, A., Penaloza-Spinola, M.I., Bonavides-Martinez, C., Ingraham, J., 2007. Multidimensional annotation of the *Escherichia coli* K-12 genome. Nucleic Acids Res. 35, 7577–7590.

Katajamaa, M., Oresic, M., 2005. Processing software for differential analysis of LC/MS profile data. BMC Bioinform. 6, 179.1–179.12.

Kind, T., Fiehn, O., 2006. Metabolomic database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm. BMC Bioinform. 7, 234.

Kind, T., Fiehn, O., 2007. Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. BMC Bioinform. 8, 115.

Kopka, J., Schauer, N., Krueger, S., Birkemeyer, C., Usadel, B., Bergmüller, E., Dörmann, P., Weckwerth, W., Gibon, Y., Stitt, M., Willmitzer, L., Fernie, A.R., Steinhauser, D., 2005. GMD@CSB.DB: the Golm Metabolome Database. Bioinformatics 21, 1635–1638.

Linstrom, P.J., Mallard, W.G., 2001. The NIST Chemistry WebBook: a chemical data resource on the internet. J. Chem. Eng. Data 46, 1059–1063.

Lisec, J., Schauer, N., Kopka, J., Willmitzer, L., Fernie, A.R., 2006. Gas chromatography mass spectrometry-based metabolite profiling in plants. Nat. Protoc. 1, 387–396.

Luedemann, A., Strassburg, K., Erban, A., Kopka, J., 2008. TagFinder for the quantitative analysis of gas chromatography–mass spectrometry (GC–MS)-based metabolite profiling experiments. Bioinformatics 24, 732–737.

Masciocchi, J., Frau, G., Fanton, M., Sturlese, M., Floris, M., Pireddu, L., Palla, P., Cedrati, F., Rodriguez-Tomé, P., Moro, S., 2009. MMsINC: a large-scale chemoinformatics database. Nucleic Acids Res. 37, D284–D290.

Matsuda, F., Yonekura-Sakakibara, K., Niida, R., Kuromori, T., Shinozaki, K., Saito, K., 2008. MS/MS spectral tag (MS2T)-based annotation of non-targeted profile of plant secondary metabolites. Plant J. (Epub ahead of print).

McLafferty, F.W., Zhang, M.Y., Stauffer, D.B., Loh, S.Y., 1998. Comparison of algorithms and databases for matching unknown mass spectra. J. Am. Soc. Mass Spectrom. 1, 92–95.

Moco, S., Bino, R.J., Vorst, O., Verhoeven, H.A., de Groot, J., van Beek, T.A., Vervoort, J., de Vos, C.H., 2006. A liquid chromatography mass spectrometry based Metabolome Database for Tomato. Plant Physiol. 141, 1205–1218.

Morikawa, T., Mizutani, M., Aoki, N., Watanabe, B., Saga, H., Saito, S., Oikawa, A., Suzuki, H., Sakurai, N., Shibata, D., Wadano, A., Sakata, K., Ohta, D., 2006. Cytochrome P450 CYP710A encodes the sterol C-22 desaturase in *Arabidopsis* and tomato. Plant Cell 18, 1008–1022.

Mueller, L.A., Zhang, P., Rhee, S.Y., 2003. AraCyc: a biochemical pathway database for *Arabidopsis*. Plant Physiol. 132, 453–460.

Nobeli, I., Thornton, J.M., 2006. A bioinformatician's view of the metabolome. Bioessays 28, 534–545.

Nobeli, I., Ponstingl, H., Krissinel, E.B., Thornton, J.M., 2003. A structure-based anatomy of the *E. coli* metabolome. J. Mol. Biol. 334, 697–719.

Ohta, D., Shibata, D., Kanaya, S., 2007. Metabolic profiling using Fourier-transform ion-cyclotron-resonance mass spectrometry. Anal. Bioanal. Chem. 389, 1469–1475.

Oikawa, A., Nakamura, Y., Ogura, T., Kimura, A., Suzuki, H., Sakurai, N., Shinbo, Y., Shibata, D., Kanaya, S., Ohta, D., 2006. Clarification of pathway-specific inhibition by Fourier transform ion cyclotron resonance/mass spectrometry-based metabolic phenotyping studies. Plant Physiol. 142, 398–413.

Oksman-Caldentey, K.M., Saito, K., 2005. Integrating genomics and metabolomics for engineering plant metabolic pathways. Curr. Opin. Biotechnol. 16, 174–179.

Ott, M.A., Vriend, G., 2006. Correcting ligands, metabolites, and pathways. BMC Bioinform. 7, 517.

Petersen, B.L., Chen, S., Hansen, C.H., Olsen, C.E., Halkier, B.A., 2002. Composition and content of glucosinolates in developing *Arabidopsis thaliana*. Planta 214, 562–571.

Rosania, G.R., Crippen, G., Woolf, P., States, D., Shedden, K., 2007. A cheminformatic toolkit for mining biomedical knowledge. Pharm. Res. 24, 1791–1802.

Schauer, N., Fernie, A.R., 2006. Plant metabolomics: towards biological function and mechanism. Trends Plant Sci. 11, 508–516.

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

Schauer, N., Zamir, D., Fernie, A.R., 2005b. Metabolic profiling of leaves and fruit of wild species tomato: a survey of the *Solanum lycopersicum* complex. J. Exp. Bot. 56, 297–307.

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. Biotechnol. Agr. Forest. 57, 166–181.

Smith, R.M., Busch, K.L., 1999. Understanding Mass Spectra – A Basic Approach. Wiley.

Smith, C.A., O'Maille, G., Want, E.J., Qin, C., Trauger, S.A., Brandon, T.R., Custodio, D.E., Abagyan, R., Siuzdak, G., 2005. METLIN: a metabolite mass spectral database. Ther. Drug Monit. 27, 747–751.

Smith, C.A., Want, E.J., O'Maille, G., Abagyan, R., Siuzdak, G., 2006. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Anal. Chem. 78, 779–787.

Stein, S.E., 1999. An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. J. Am. Soc. Mass Spectrom. 10, 770–781.

Stein, S.E., Scott, D.R., 1994. Optimization and testing of mass spectral library search algorithms for compound identification. J. Am. Soc. Mass Spectrom. 5, 859–866.

Steinfath, M., Groth, D., Lisec, J., Selbig, J., 2008. Metabolite profile analysis: from raw data to regression and classification. Physiol. Plant. 132, 150–161.

Steinhauser, D., Usadel, B., Luedemann, A., Thimm, O., Kopka, J., 2005. GMD@CSB.DB: the Golm Metabolome Database. Bioinformatics 21, 1635–1638.

Strehmel, N., Hummel, J., Erban, A., Strassburg, K., Kopka, J., 2008. Retention index thresholds for compound matching in GC–MS metabolite profiling. J. Chromatogr. B Analyt. Technol. Biomed. Life Sci. 871, 182–190.

Suhre, K., Schmitt-Kopplin, P., 2008. MassTRIX: Mass TRanslator Into Pathways. Nucleic Acids Res. 36, W481–W484.

Sumner, L.W., Mendes, P., Dixon, R.A., 2003. Plant metabolomics: large-scale phytochemistry in the functional genomics era. Phytochemistry 62, 817–836.

Takahashi, H., Kai, K., Shinbo, Y., Tanaka, K., Ohta, D., Oshima, T., Altaf-Ul-Amin, M., Kurokawa, K., Ogasawara, N., Kanaya, S., 2008. Metabolomics approach for determining growth-specific metabolites based on Fourier transform ion cyclotron resonance mass spectrometry. Anal. Bioanal. Chem. 391, 2769–2782.

Thimm, O., Bläsing, O., Gibon, Y., Nagel, A., Meyer, S., Krüger, P., Selbig, J., Müller, L.A., Rhee, S.Y., Stitt, M., 2004. MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J. 37, 914–939.

Tohge, T., Nishiyama, Y., Yokota Hirai, M., Yano, M., Nakajima, J., Awazuhara, M., Inoue, E., Takahashi, H., Goodenowe, D.B., Kitayama, M., Noji, M., Yamazaki, M., Saito, K., 2005. Functional genomics by integrated analysis of metabolome and transcriptome of *Arabidopsis* plants over-expressing an MYB transcription factor. Plant J. 42, 218–235.

Tokimatsu, T., Sakurai, N., Suzuki, H., Ohta, H., Nishitani, K., Koyama, T., Umezawa, T., Misawa, N., Saito, K., Shibata, D., 2005. KaPPA-view: a web-based analysis tool for integration of transcript and metabolite data on plant metabolic pathway maps. Plant Physiol. 138, 1289–1300.

Vorst, O.F.J., Vos, C.H. de, Lommen, A., Staps, R.V., Visser, R.G.F., Bino, R.J., Hall, R.D., 2005. A non-directed approach to the differential analysis of multiple LC–MS-derived metabolic profiles. Metabolomics 1, 169–180.

Wagner, C., Sefkow, M., Kopka, J., 2003. Construction and application of a mass spectral and retention time index database generated from plant GC/EI-TOF-MS metabolite profiles. Phytochemistry 62, 887–900.

Watanabe, K., Yasug, E., Oshima, M., 2000. How to search the glycolipid data in LIPIDBANK for Web: the newly developed lipid database. Trend Glycosci. Glycotechnol. 12, 175–184.

Wiklund, S., Johansson, E., Sjostrom, L., Mellerowicz, E.J., Edlund, U., Shockcor, J.P., Gottfries, J., Moritz, T., Trygg, J., 2008. Visualization of GC/TOF-MS-based metabolomics data for identification of biochemically interesting compounds using OPLS class models. Anal. Chem. 80, 115–122.

Wishart, D.S., Tzur, D., Knox, C., Eisner, R., Guo, A.C., Young, N., Cheng, D., Jewell, K., Arndt, D., Sawhney, S., Fung, C., Nikolai, L., Lewis, M., Coutouly, M.A., Forsythe, I., Tang, P., Shrivastava, S., Jeroncic, K., Stothard, P., Amegbey, G., Block, D., Hau, D.D., Wagner, J., Miniaci, J., Clements, M., Gebremedhin, M., Guo, N., Zhang, Y., Duggan, G.E., Macinnis, G.D., Weljie, A.M., Dowlatabadi, R., Bamforth, F., Clive, D., Greiner, R., Li, L., Marrie, T., Sykes, B.D., Vogel, H.J., Querengesser, L., 2007. HMDB: the Human Metabolome Database. Nucleic Acids Res. 35, D521–D526.

Yonekura-Sakakibara, K., Tohge, T., Matsuda, F., Nakabayashi, R., Takayama, H., Niida, R., Watanabe-Takahashi, A., Inoue, E., Saito, K., 2008. Comprehensive flavonol profiling and transcriptome coexpression analysis leading to decoding gene–metabolite correlations in *Arabidopsis*. Plant Cell 20, 2160–2176.

Zhang, P., Foerster, H., Tissier, C.P., Mueller, L., Paley, S., Karp, P.D., Rhee, S.Y., 2005. MetaCyc and AraCyc. Metabolic pathway databases for plant research. Plant Physiol. 138, 27–37.