Anal Bioanal Chem (2010) 398:2779–2788
DOI 10.1007/s00216-010-4142-5

REVIEW

Computational mass spectrometry for metabolomics: Identification of metabolites and small molecules

Steffen Neumann · Sebastian Böcker

Received: 14 June 2010 / Revised: 16 August 2010 / Accepted: 18 August 2010 / Published online: 9 October 2010

Abstract The identification of compounds from mass spectrometry (MS) data is still seen as a major bottleneck in the interpretation of MS data. This is particularly the case for the identification of small compounds such as metabolites, where until recently little progress has been made. Here we review the available approaches to annotation and identification of chemical compounds based on electrospray ionization (ESI-MS) data. The methods are not limited to metabolomics applications, but are applicable to any small compounds amenable to MS analysis. Starting with the definition of identification, we focus on the analysis of tandem mass and MS^n spectra, which can provide a wealth of structural information. Searching in libraries of reference spectra provides the most reliable source of identification, especially if measured on comparable instruments. We review several choices for the distance functions. The identification without reference spectra is even more challenging, because it requires approaches to interpret tandem mass spectra with regard to the molecular structure. Both commercial and free tools are capable of mining general-purpose compound libraries, and identifying candidate compounds. The holy grail of computational mass spectrometry is the *de novo* deduction of structure hypotheses for compounds, where method development has only started thus far. In a case study, we apply several of the available methods to the three compounds, kaempferol, reserpine, and verapamil, and investigate whether this results in reliable identifications.

Keywords Mass spectrometry · Metabolomics · Compound identification · Spectral library · Structure elucidation

Electronic supplementary material The online version of this article (doi:10.1007/s00216-010-4142-5) contains supplementary material, which is available to authorized users.

S. Neumann (✉)
Department of Stress and Developmental Biology,
Leibniz Institute of Plant Biochemistry,
06120 Halle, Germany
e-mail: sneumann@ipb-halle.de

S. Böcker
Department of Mathematics and Computer Science,
Friedrich-Schiller-University, Jena,
07743 Jena, Germany
e-mail: sebastian.boecker@uni-jena.de

Introduction

For a long time the established textbooks on MS have included a computer as part of the analytical setup. Initially, the computer replaced the photo platters to display the spectrum, and to save or print the peak lists. Nowadays, with the advent of high-throughput experiments, the complexity of their tasks has grown tremendously, and this is where computational mass spectrometry enters the field.

The signal processing of today's hyphenated MS setups such as GC-MS, LC-MS, or CE-MS requires two-dimensional feature finding and peak picking, which will not be covered here. For a review, see, e.g., [1]. To compare the abundances of the measured compounds in different samples, an *alignment* step is required, because both the chromatographic retention time and (usually to a lesser degree) the mass to charge ratio (m/z) may drift across measurements. Again, this has been covered elsewhere, see, e.g., [2–4].

Often, efficient solutions to the challenges in data analysis exist in the realm of mathematics and algorithm engineering, but researchers in those fields are often not familiar with mass spectrometry, and the problems have to be translated first: an earlier review [5] touched on several aspects of mathematical modeling in mass spectrometry.

At the American Society for Mass Spectrometry (ASMS) conference 2009, a survey among the 600 participants revealed that the identification of compounds was perceived as the bottleneck in interpretation of metabolomics data.[1]

What is identification? At first glance, the question might appear naïve, because there exist a vast number of publications in the field that claim to "identify" compounds using MS data, presenting a list of compounds that were present in some sample. Yet, these "identifications" are often backed by vastly differing levels of evidence.

As an example, consider the work of Iijima et al. [6]: The authors annotated approximately 870 metabolites in tomato (*Solanum lycopersicum*) and calculated molecular formulas based on recalibrated accurate masses obtained by LC-ESI-FTICR-MS. About 500 of these molecular formulas were not found in metabolite databases such as DNP, KNApSAcK, or MotoDB and are claimed to be novel metabolites.

More laborious, but also more informative is the approach taken by, e.g., Böttcher et al. [7]. Here, the calculation of elemental compositions with ultrahigh resolution Fourier transform ion cyclotron resonance (FTICR)-MS is accompanied by extensive tandem mass and pseudo-MS^3 spectra on a CapLC-QTOF instrument. Beyond the purely spectral information, the authors also compare the abundance of the compounds in different *Arabidopsis thaliana* pathway mutants, to make sure the molecular structures are consistent with their role as substrates and products of known enzymes.

A different approach to a highly reliable identification is the combination of both MS and NMR analysis, as done by Glauser et al. [8]. Though a lack of sensitivity has always been touted as a major disadvantage of NMR, modern high-field instruments equipped with cryo-probes or capillary NMR, allow structure elucidation and require only *microgram* amounts of the compound.

The Metabolomics Standards Initiative (MSI) has published several guidelines for the publication of metabolomics experiments. One of these covers the "proposed minimum reporting standards for chemical analysis" [9], which define confidence levels for the identification (or validation) of non-novel chemical compounds, ranging from level 1 for a rigorous identification based on independent measurements of authentic standards, to unidentified signals at level 4, see Table 1 for a summary. The characterization of *novel* compounds usually requires extraction and purification of the substance, and the use of analytical methods beyond mass spectrometry, including 1D and 2D NMR measurements.

Authors who publish their metabolomics experiments should always strive to deliver the most convincing data to support their findings. Although the MSI guidelines allow level 1 or level 2 identifications[2] based on a comparison of "exact mass and isotope pattern", more structural evidence should be presented at these levels: even with the most exact mass and isotope pattern the identification will be limited to the elemental composition, and even for the known metabolites in databases such as KEGG or PubChem, dozens to hundreds of compounds share the same sum formula. The MSI strongly discourages matching an elemental formula to database hits as identification method alone.

For all other MSI levels, the "identification" usually boils down to an annotation with lower levels of confidence. If reference spectra from commercial or public databases are used, authors should consider not only the similarity (score) and the comparability of the analytical setups, but also a possible bias in the database: whether a compound class is represented by a large number of similar compounds, or just by a few examples. It is easy to obtain a wrong, but "unique" identification by chance if a spectrum of a plant metabolite is searched against the Human Metabolite Database (HMDB), which is dedicated to human metabolites [10]. On the other hand, it is as easy to miss the correct compound among many structurally similar members of a large compound class present in the database.

As the elemental composition is the basis of any further identification, tools for their determination have long been a part of most vendor software, and many of today's algorithms are known to perform well in practice. Despite the quickly growing number of elemental compositions for large masses, these methods are usually fast in practice, some of them even for the chemically unrestricted case [11]. Note that the monoisotopic mass of a compound is usually not sufficient to determine its elemental composition, even for mass accuracies that surpass those of any available MS instrument. Therefore the algorithms depend on both the accurate mass and isotope patterns. In their theoretical evaluation, Kind and Fiehn [12] showed that an instrument with 3-ppm mass accuracy and 2% RIA (relative isotope abundance) accuracy allows one to calculate a single sum formula up to 300 Da. In another study [11], this was confirmed experimentally, and for 68 out of 70 compounds measured on an oa-TOF instrument (micrOTOFq, Bruker Daltonik GmbH, Bremen, Germany), the correct solution was ranked first. In KEGG, 81% of all compounds are below 500 Da. The used SIRIUS software is available freely, and details of the statistical analysis have been published [11]. Extensive experimental determination (and optimization) of the mass accuracy and RIA for oa-TOF instruments have

[1] http://metabolomicssurvey.com/

[2] The difference between level 1 and 2 is that the former requires the comparison with authentic standards based on in-house data measured under identical analytical conditions, whereas the latter allows one to use literature values or external databases.

Table 1 MSI levels for validation of non-novel compounds, based on [9]

Level	Name	Minimum requirements
1	Identified compounds	At least two independent and orthogonal data relative to an authentic compound analyzed under identical experimental conditions (e.g., retention time/index and mass spectrum, retention time and NMR spectrum, accurate mass and tandem MS, accurate mass and isotope pattern, full ^{1}H and/or ^{13}C NMR, 2D NMR spectra)
2	Putatively annotated compounds	Similar to level 1, but based on literature values reported for authentic samples by other laboratories
3	Putatively characterized compound classes	Based upon characteristic physicochemical properties of a chemical class of compounds, or by spectral similarity to known compounds of a chemical class
4	Unknown compounds	These metabolites can still be differentiated and quantified based upon spectral data

been performed [13, 14]. For Orbitrap instruments, mass accuracy and dependency upon ion intensities were recently evaluated by Xu et al. [15]. Here, the mass accuracy was characterized as less than 5 ppm with external calibration (although much lower values have been reported elsewhere), and RIA was found to be less than 20%.

The current FT-ICR-MS instruments with superconducting 12 Tesla (or stronger) magnets can easily exceed a resolution of 300,000 in routine measurements, and allow one to resolve the isotopic fine-structure of the individual ^{13}C, ^{15}N for the first isotope peak, or ^{18}O and ^{34}S isotopes. Deriving the elemental composition is then much simplified, and can be achieved by calculating the individual ratios to the monoisotopic peak [16].

Although the correct formula will be among the top ranks for most of these approaches, Matsuda et al. [17] showed that the false discovery rate (FDR) of queries based on exact masses and isotope ratio in databases widely spreads between a few percent up to 100%, depending on mass accuracy, fidelity of the isotopic intensities, and the actual database (KEGG, KNApSAcK, and PubChem).

Structural information for a compound can be obtained in different ways: by exploiting the in-source fragmentation, and/or by performing targeted collision-induced dissociation (CID) MS experiments. They both allow one to measure the mass of molecular fragments, in the latter case after collision in a cell filled with an inert gas such as argon or nitrogen. The compound structure and collision energy determine the degree of fragmentation. It is also possible to continuously increase the energy, e.g., from 5 to 60 eV during a single acquisition, essentially measuring a combined spectrum. These are often termed RAMP spectra. With multiple stages (MS^3 and higher), individual fragments can be analyzed further. Molecular rearrangements during the fragmentation can complicate the interpretation of the spectra.

In the following, we will focus on computational mass spectrometry for the identification of small compounds and metabolites on high-resolution hybrid or multi-stage instruments with electrospray ionization (ESI), such as (Q)TOF, Orbitrap, and FTICR-MS. Because many ideas have been pioneered on electron impact (EI) instruments, we will also enlarge upon these where appropriate.

Acquisition and processing of tandem MS data

With current MS instruments, it is possible to acquire tandem mass spectra in data-dependent-acquisition mode (DDA). Here, the instrument performs an MS^1 survey scan, and selects one or more ions for subsequent MS^2 or even MS^n scans. However, this has the drawbacks that the effective scan rate for tandem MS is reduced by single-stage MS survey scans and secondly the selection of parent peaks only considers the N most intense peaks, possibly including in-source fragments or, e.g., $[M+Na]^+$ adducts. Usually, $[M+H]^+$ ions are preferable for fragmentation, because at common collision energies (10–50 eV) they result in more informative spectra for many compounds.

An advanced scheme to survey the metabolome of *Arabidopsis thaliana* was proposed by Matsuda et al. [18], who performed repeated measurements with DDA in narrow, overlapping (60-Da) windows, and collected results in a "Metabolite Expression Atlas" [19]. Out of almost 1,600 observed metabolite signals, they identified 167 compounds based on the tandem mass spectra.

A different approach is the acquisition of tandem mass spectra, alternating between low and high collision energies without any precursor mass filtering, termed MS^E . Plumb et al. [20] assessed and identified spectra of 10 metabolites from rat urine. A difficulty is the superposition of fragments from all co-eluting compounds, including background ions. The assignment to precursor and product ions based on a statistical test was introduced by Ipsen et al. [21], where ion counts are modeled using a Poisson distribution. The proposed algorithm works well if the acquisition conditions are well controlled, such as low to medium signal intensities, and absence of dynamic gain control in the instrument. A more general but less rigorous method to assign corresponding mass signals is the peak shape correlation using Pearson correlation coefficients (ρ), as described elsewhere [22].

Comparison with reference spectra

Today, one of the most common methods for the identification of compounds using mass spectrometry is the comparison with spectra of authentic standards. Libraries of mass spectra, especially the National Institute of Standards and Technology (NIST) database and the Wiley registry of mass spectral data, have been mentioned as part of a very broad review of chemical signature databases [23]. A summary is given in Table 2. Here, we want to focus on the computational aspects.

Each database or processing tool requires one to score database entries based on a similarity or distance function. The most fundamental scorings are the "peak count" family of measures. They count the number of matching peaks between a query spectrum and each of the database spectra. For this, both spectra can be considered as binary vectors with 0's and 1's for "peak absent" and "peak present", respectively. They can be either fixed-length, with bins of fixed widths such as 1 or 0.1 Da. Alternatively, the binary vectors can result from a matching between the query Q and library spectrum L, with a vector of length $|Q \cup L|$. Common distance functions on binary vectors are the Hamming distance (counting any difference) or the Jaccard coefficient (the fraction of matching peaks). These and other scoring functions differ mostly in how missing or extra peaks are treated. For an overview of distance measures, see [24].

In addition to just counting matches, other measures also include their actual mass and intensities. Stein [25] compared the Euclidean distance, the probability-based matching (PBM), and the normalized dot product (NDP) for the database search of EI spectra, and proposed a modified cosine distance for database retrieval.

Mass and intensity scores can be weighted by using the formula $W = \text{score\_intensity}^m \cdot \text{score\_mass}^n$, where the parameters m=0.6 and n=3 were optimized experimentally on a large training set of EI spectra. The MassBank system [26] uses this measure, but optimized the exponents for ESI spectra of common metabolites with their different mass and intensity distributions to m=0.5 and n=2. MassBank also offers a neutral loss search.

The Human Metabolome Database (HMDB) [10] uses a scoring function based on spectral matching and parameter optimization that was originally developed for peptides [27].

Because peak intensities are inherently variable in CID mass spectra, especially across different acquisition parameters or even instruments, Pavlic et al. [28] and Oberacher et al. [29] proposed and optimized a search function based on a combination of relative and absolute match probabilities, which combine the principle of peak counting and *summed* intensities of matching peaks.

The X-Rank algorithm [30] uses a statistical formulation of the problem, and considers only *ranked* intensities instead of absolute or relative intensities: what is the joint probability of matching the nth peak in one spectrum to the mth peak in another? If this probability can be reliably calculated, the correct library spectrum should be the one with the highest probability. The solution requires a training on a representative dataset.

Finally, it depends on the tandem MS library, whether the search can be constrained by the parent ion mass. MassBank will search all spectra, whereas METLIN [31] has an option to filter only the correct precursor masses. HMDB searches entries within a user-defined precursor mass window.

Computational analysis of tandem mass spectra

The simplest information which can be extracted from accurate mass MS data is the elemental composition, see, e.g., [32], and subsequent lookup in compound libraries as implemented in MZSearcher [33]. Recently, improved algorithms have been created which exploit the additional information present in tandem mass and MS^n spectra for sum formula calculations.

The commercial SmartFormula3D software (Bruker Daltonics) is an extension of an elemental composition calculator. The software predicts elemental compositions of both precursor and product ions, and filters all elemental compositions of the precursor that are incompatible with

Table 2 Overview of several spectral libraries (only ESI spectra)

Library	Compounds	Spectra	Accuracy	Comment
NIST '08	5,308	14,802	Nominal	Commercial license
METLIN	2,658	13,896	Accurate	Web interface, SOAP Web service planned
HMDB	921	2,565	Nominal	Web interface, download free for noncommercial purposes
MassBank	2,189	9,218	Accurate/Nominal	Web interface, SOAP Web service. Free subset for download planned

Data presented concern content size, accuracies, license, and availability. For NIST and MassBank the number of compounds is an upper bound, ignoring possible redundancy

the (smaller mass, thus likely more accurate) compositions of the product ions.

SIRIUS Starburst [34] aims to calculate the correct sum formula from MS and MS/MS data, but in addition it will propose a tree representation that is very often close to the actual fragmentation tree of the compound. In any case, the organization as a tree simplifies the subsequent manual interpretation and structure elucidation.

Because the fragmentation is a gas-phase reaction, the cleavage sites can be approximated and described with rules of possible fragmentation reactions. The commercial software ACD/MS Fragmenter [35] is such a tool that uses a database of fragmentation rules. Pelander et al. used both SmartFormula3D and the ACD Fragmenter to differentiate structural isomers in a comparatively low number of phase I metabolites of quetiapine from LC/TOF MS spectra [36], and recently generalized the survey to 111 compounds in 48 isomer groups [37].

A similar problem is approached by Mass Frontier [38], which was originally targeted at electron impact (EI) spectra as obtained from GC-MS. Support for ESI spectra has been added later. The fragmentation schemes have been extracted from the literature and in-house spectral libraries. Horai et al. [39] annotated the spectra of 453 metabolites in MassBank, and used both ACD Fragmenter and MassFrontier in a manual process to verify the annotation. From an overall 120,000 peaks, only 3% could be annotated with confidence. Some of the fragmentation rules used by MassFrontier also cover the negative ionization mode, but Heinonen et al. [40] report that for some compounds MassFrontier (version 5) is not able to identify any fragments in negative mode. For both ACD/MS Fragmenter and MassFrontier further algorithmic details are not published.

Another class of algorithms strives to interpret the tandem MS spectra, and to assign fragment structures to observed peaks. The systematic bond disconnection method is independent of any rule sets. A member of this class is the EPIC tool [41]. It matches resulting product ions from a single precursor structure against the peaks measured on a high-resolution mass spectrometer. The application provides a Web front end, and allows one to generate a report including the user-approved fragment structures.

Heinonen et al. [40, 42] proposed an algorithm that, given both a metabolite's molecular structure and its tandem mass spectrum, tries to predict the fragmentation tree of the metabolite by interpreting the tandem mass spectrum. Among sets of equivalent possible fragmentations, their tool can select the solution which minimizes the bond dissociation energy. Even this seemingly simple problem turns out to be computationally hard. The authors show that in general, it is NP complete to decide whether a given molecular structure will generate a fragment of a certain mass.[3] The authors present heuristics that work well for small molecules, but require hours of running time for molecules above 350 Da. But even if the correct molecular formula is known for the parent peak and all fragment peaks in the spectrum, the problem remains NP hard [43]. None of the above approaches is capable of handling non-trivial rearrangements during fragmentation.

The number of compounds in spectral databases is low compared to the estimated number of metabolites for a given organism. Computational mass spectrometry approaches help to identify the "known" unknowns, i.e., those metabolites which are in a compound library, but without any reference spectra. Once a tool is available that allows one to assign (sub-)structures to peaks in reasonable time, it is also possible to screen comparatively large general-purpose molecular databases and calculate a score of the agreement between the spectrum and candidate compounds.

This approach was first used by Hill et al. [44] who measured the spectra of 102 test compounds on a Micromass Q-TOF II in positive mode at different collision energies. For each compound they retrieved on average 270 candidates from the PubChem database, and used Mass Frontier (version 4) to predict the tandem mass spectra. The agreement score was a simple peak count between predicted and measured spectra. The median rank of the correct compounds in an evaluation on the PubChem database was four if no manual expert knowledge was used.

The MetFrag suite [45] can directly query an upstream database (currently KEGG, PubChem, and ChemSpider are supported, custom SDF structure files can be uploaded). MetFrag performs an in silico fragmentation and ranks the candidates based on the number of molecular fragments which explain the measured peaks. On the same test set that was used by Hill et al. [44], MetFrag performs better than the commercial MassFrontier, in particular the average and standard deviation of the correct ranks are lower. MetFrag has an Open Source license, and both a Web front end and Web service interface are available.

In their paper Levsen et al. [46] measured and interpreted a large number of compounds from various classes, and also used *ab initio* calculations to explain some of the fragmentation mechanisms. The use of computational chemistry methods such as density functional theory (DFT) promises to predict the fragmentation sites based on the simulation of bond elongation upon protonation [47, 48]. The approach showed good accuracy for fluconazole, voriconazole, and maraviroc with two of its breakdown products, but has rather high computational demands.

[3] If a problem is NP complete, then there cannot exist an algorithm with running time polynomial in the input size unless P=NP. It is widely believed among computer scientists that the latter is not the case.

Identification case studies

Finally, we wanted to apply some of the methods mentioned above to arbitrarily selected compounds: kaempferol is a plant secondary metabolite found in the flavonoid pathway; reserpine and verapamil are two well-known drugs, see Fig. 1 for their structures and PubChem identifiers. The chemical and spectroscopic details are provided in the "Electronic supplementary material". This case study is not meant to be an exhaustive evaluation, but provides qualitative observations. The selection of tools and databases was purely determined by their availability to us.

The first spectral data we investigated were from kaempferol. To determine the elemental composition from accurate mass and the isotope distribution, we used MS^1 data obtained on an ESI-QTOF instrument (micrOTOFQ-II, Bruker Daltonics). We evaluated Bruker SmartFormula and SIRIUS [11] calculators. Both tools returned the correct formula $C_{15}H_{10}O_6$ ranked first. This is no surprise, given the rather low molecular mass (286.048 Da) of the compound, and the good accuracy of the instrument for both mass and relative isotope intensities. However, knowing the correct molecular formula will not get us very far: PubChem listed 159 compounds with this molecular formula, in KEGG we found 12.

Beyond the elemental composition, tandem mass spectra provide the next level of evidence for the identification. As mentioned before, the comparison with reference spectra in an identical analytical setup would be preferred. In reality, spectral libraries such as the NIST database or MassBank contain spectra from a variety of instruments. HMDB and METLIN were each measured (so far) on a single instrument model, which should be beneficial if the local instrument is comparable.

We used the MassBank record PB000166 (acquired on a QStar Pulsar i, Applied Biosystems; 40 eV; removing peaks of 5% intensity or less) to search in MassBank, METLIN, and HMDB. We always tried to search without matching the precursor mass, to include not only exact matches, but also similar compounds (in terms of spectral similarity). This strategy was used in MassBank and METLIN, and we mimicked that by specifying the maximum allowed precursor mass window in HMDB, which was not a fixed value, but depends on the remaining input.

METLIN is focused on human metabolites, and does not contain any kaempferol spectra. Therefore, it is also of interest whether the results might help a manual interpretation, if the compound is not contained in the spectral library. With the QStar spectrum the first hits were several flavonoid-like structures, including fisetin, genistein, and koparin, in the first three positions. In HMDB we searched with the maximum allowed precursor tolerance of 2 Da, any collision energy was allowed. The database returned 24 results, with vitamin A, kaempferol, and 2-hydroxyestrone ranked in the top three. Together, both databases suggested that our substance is a flavonoid, but gave little confidence to decide which one. This result list should be a strong motivation to query libraries which have a better coverage for flavonoids.

Therefore, we used MassBank, which has several plant research institutes among the consortium members. Searching MassBank returned a list of mostly flavonoids among the top 20 results, with two kaempferol measurements from a QTOF Premier ranked first (30 eV, score 0.91) and third (RAMP 5–60 eV, score 0.75).

We cross-checked the results with kaempferol spectra from a Waters QTOF instrument (PR040029), and also the medium energy HMDB nominal mass spectrum measured on a Waters Quattro triple quadrupole. The full results are shown in Table 3.

If the metabolite (or compound class) was likely not contained in the spectral library, we would have to consult computational mass spectrometry approaches to interpret the spectrum. We used MetFrag to obtain structure hypotheses based on the elemental composition search in PubChem, and subsequent scoring of the experimental and in silico peaks. The first seven hits were flavonoid-like structures, with kaempferol ranked third. If only the molecular formula and the MetFrag results were available, it would be necessary to purchase and measure a number of flavonoids for a convincing identification. But because of the excellent fit in MassBank, and the rather sharp decline of scores for other candidates, an MSI level 1 identification could concentrate on the verification of kaempferol.

We tested the FiD software [40] with all KEGG elemental composition candidates individually, but the scores provided by the algorithm did not result in a viable ranking (results are shown in the "Electronic supplementary material"). We expect that the algorithm can easily be extended to provide scores suitable for database searches.

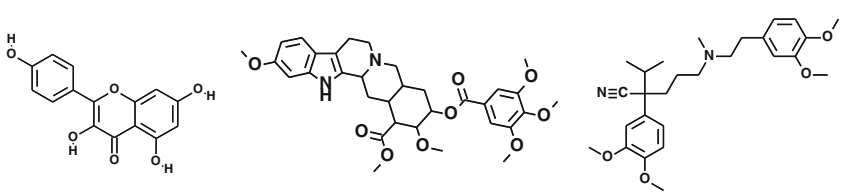

Fig. 1 Molecular structures of (*left* to *right*): kaempferol (**1**) $C_{15}H_{10}O_6$, monoisotopic mass 286.048 Da, PubChem CID 5280863; reserpine (**2**) $C_{33}H_{40}O_9N_2$, 608.273 Da, CID 5770; and verapamil (**3**) $C_{27}H_{38}N_2O_4$, 454.283 Da, CID 2520

Table 3 Identification of kaempferol by searching a query tandem mass spectrum from one library in the other two libraries, and with the MetFrag search in PubChem

Searching in...	Query spectrum, DB and entry		
	MassBank PB000166	MassBank PR040029	HMDB 05801
MassBank	1, 3	1, 19	1, 19
METLIN	NA	NA	NA
HMDB	2, 9, 23	3, 6, 11	1, 2, 5
MetFrag	3	5	11

Numbers are positions in the output list where kaempferol was reported by the search. MassBank and HMDB report all spectra (from different instruments or collision energy settings). Smaller numbers imply that we have to consider less hits to identify the correct compound. METLIN does not contain kaempferol spectra, but returned similar structures at the top of the result list

NA not applicable

Second, we investigated data from reserpine. The drug has a higher molecular mass of 608.273 Da, and could be expected to be a challenge to identify. Still, SmartFormula found the correct elemental composition ranked second. Together, the first two suggestions produced 120 hits in PubChem, whereas in KEGG we had only a single hit for the correct solution. SIRIUS applies less strict chemical filtering by default, and found the correct solution at rank 14 of its output. On the other hand, $C_{33}H_{40}O_9N_2$ was the *only* molecular formula among all 168 SIRIUS candidate formulae which was found in the compound libraries. PubChem listed 55 compounds with this molecular formula, in KEGG we found only a single one.

For reserpine we used the 40-eV spectrum from METLIN entry 2253, measured on an Agilent 6510 ESI Q-TOF. MassBank returned reserpine spectra on the first two ranks, measured on a single-quadrupole instrument[4] (ZQ, Waters). This time, HMDB had not one record within the allowed 2-Da window of the parent ion, and with the nominal mass WA002661 spectrum METLIN was also unable to return reserpine at all. All results are show Table 4.

As additional evidence, we analyzed all 55 PubChem candidates with the calculated $C_{33}H_{40}O_9N_2$ composition in MetFrag, which found the solution well based on the QTOF spectrum, but less so with the nominal masses from the ZQ instrument. For further confirmation it would be reasonable to purchase 5–10 authentic standards, or to validate the structure candidates in an NMR experiment.

Finally, we investigated the verapamil data, focusing on the tandem MS analysis. We used mass spectral data from the METLIN entry 3009 (40-eV spectrum, measured on an Agilent 6510 ESI Q-TOF), the MassBank record KOX00895 (10–50 eV; 5 spectra merged; measured on a QStar, 1% relative intensity threshold), and the HMDB01850 record.

Table 4 Identification of reserpine by searching with a QTOF spectrum from METLIN, or the single-quadrupole Waters ZQ spectrum in all three libraries. METLIN returns the compound ID, whereas MassBank and HMDB report all spectra

Searching in...	Query spectrum, DB and entry	
	METLIN 2253	MassBank WA002661
MassBank	1, 2	1–4
METLIN	1	Not found
HMDB	NA	NA
MetFrag	1	15

The search in MassBank returned the query spectrum and five other spectra, measured on a single-quadrupole instrument (ZQ, Waters) and an ion trap (LC/MSD Trap XCT, Agilent Technologies) in the first six positions. Then, METLIN also returned verapamil ranked first, and all runners-up had a much lower score. In HMDB we searched with the maximum allowed parent ion tolerance of (in this case) 27 Da: HMDB returned 30 results for the KOX00895 peaks, with two structurally quite different compounds (loperamide and deoxycholic acid glycine conjugate) at the top of the list, and the best matching verapamil spectrum ranked third. The same searches with peak data from the METLIN entry 3009 produced comparable results. The full results are shown in Table 5.

Although we had convincing evidence for verapamil in MassBank and METLIN, we nevertheless analyzed the 436 PubChem entries with formula $C_{27}H_{38}N_2O_4$ in MetFrag. MetFrag was able to explain 6 out of the 11 peaks we also used in the spectral library tests from KOX00895, ranking the correct compound (and several highly similar structures) in second position, but all in all 16 similar compounds obtained the same score as verapamil. In a next step a user would try to

Table 5 Identification of verapamil with two QTOF and one triple-quadrupole spectrum

Searching in...	Query spectrum, DB and entry		
	METLIN 3009	MassBank KOX00895	HMDB HMDB01850
MassBank	1–7, 9, 18	1–6, 10, 13, 15	Not found
METLIN	1	1	Not found
HMDB	9, 19, 28	3, 14, 28	1, 12, 29
MetFrag	8	2	Not found

[4] In single-quadrupole instruments, a higher voltage can be applied to the sampling cone to induce in-source fragmentation.

manually assign fragment peaks to likely structures, possibly also using a substructure search in a spectral library, to verify the predicted fragments. Taking both MassBank and MetFrag results into account, it would be sufficient to validate a small number of authentic standards to obtain an MSI level 1 identification.

Conclusion

With modern high resolution mass spectrometers, the determination of the elemental composition for low to medium weight metabolites from accurate measurements is clearly feasible. The number of results (and false positives) in a subsequent search in the compound libraries depends on the database size and content. But clearly, this is only the first step of compound identification.

The different distance measures for spectra comparison in reference libraries have come a long way. The results from MassBank show that spectral libraries can not only retrieve the correct compound—even from different instruments—among the top hits, but also related structures (HMDB also returned some quite different structures, though). Not surprisingly, the queries with QTOF spectra achieved much better rankings than the nominal mass quadrupole data.

But the results of the case study also clearly show that an MSI level 2 identification based on tandem mass spectra is difficult to achieve because spectral libraries have a low (and divergent) coverage of the chemical space, and the ranked result lists are just that; simply using the best hit as "truth" can be very misleading. On the other hand, the library results often allow one to obtain the correct compound class, i.e., a level 3 identification. Consequently, unless authors make explicit statements how the spectral match was (structurally) interpreted, simply reporting the best hits can at best be considered a highly putative identification.

Without reference spectra, a compound identification based on current in silico methods is currently not possible. However, both our case study and the more extensive results published for ACD Fragmenter and MetFrag make it clear that they are valuable tools to augment the rather sparse reference libraries and to direct selection of authentic compounds for in-house comparison.

An open question remains, whether the spectral libraries should contain (1) individual collision energies, (2) merged spectra from several spectra, or (3) RAMP spectra, where the instrument ramped up the collision energy during the acquisition. If a single eV spectrum is compared to a library of merged (or RAMP) spectra, not only the number of peaks in the library will be higher, but the intensities in the library are unlikely to be anywhere close to what can be observed for a single energy. This should be true for MassBank, which modifies the intensity with an exponent ($score\_intensity^{0.5}$) as a term of the scoring function, but even more so for METLIN, which recently adopted the X-Rank algorithm and currently contains spectra measured at a set of collision energies. In the future, we expect a rapid growth of the public spectral libraries. Analogous to the wealth of algorithms for the analysis of sequence data available with the large sequence databases, we also anticipate new and more robust algorithms in the area of computational mass spectrometry.

All spectral libraries would benefit from opening up their data. Such a step allows one to use the experimental spectra for the training of computational mass spectrometry algorithms, which in turn can be used for quality control, e.g., to calculate mass accuracies, detect incorrect metadata both within and across libraries.

But unlike the sequence analysis area, identification is still missing a possibility to search for those compounds that have not been recorded in any library or structure database. The existing spectra interpretation algorithms return some kind of score, but none offers a reliable p value, and methods for these tasks are highly sought.

Acknowledgements We greatly appreciate the intense discussions with our colleagues in several mass spectrometry labs, such as with Edda von Roepenack-Lahaye, Christoph Böttcher, Stephan Schmidt, and Jürgen Schmidt (all at IPB Halle), and Ales Svatos (MPICE, Jena). Thanks also to all participants of our "Computational Mass Spectrometry" workshop at the DGMS 2010 in Halle.

References

1. Zhang J, Gonzalez E, Hestilow T, Haskins W, Huang Y (2009) Review of peak detection algorithms in liquid-chromatography-mass spectrometry. Curr Genomics 10(6):388–401
2. America AHP, Cordewener JHG (2008) Comparative LC-MS: a landscape of peaks and valleys. Proteomics 8(4):731–749
3. Lange E, Tautenhahn R, Neumann S, Gröpl C (2008) Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements. BMC Bioinformatics 9:375+
4. Vandenbogaert M, Li-Thiao-Té S, Kaltenbach H-M, Zhang R, Aittokallio T, Schwikowski B (2008) Alignment of LC-MS images, with applications to biomarker discovery and protein identification. Proteomics 8(4):650–672
5. Meija J (2006) Mathematical tools in analytical mass spectrometry. Anal Bioanal Chem 385(3):486–499
6. 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(5):949–962
7. Böttcher C, von Roepenack-Lahaye E, Schmidt J, Schmotz C, Neumann S, Scheel D, Clemens S (2008) Metabolome analysis of biosynthetic mutants reveals a diversity of metabolic changes and

allows identification of a large number of new compounds in arabidopsis. Plant Physiol 147(4):2107–2120
8. Glauser G, Guillarme D, Grata E, Boccard J, Thiocone A, Carrupt P-A, Veuthey J-L, Rudaz S, Wolfender JL (2008) Optimized liquid chromatography-mass spectrometry approach for the isolation of minor stress biomarkers in plant extracts and their identification by capillary nuclear magnetic resonance. J Chromatogr A 1180(1–2):90–98
9. Sumner LW, Amberg A, Barrett D, Beale M, Beger R, Daykin C, Fan T, Fiehn O, Goodacre R, Griffin JL, Hankemeier T, Hardy N, Harnly J, Higashi R, Kopka J, Lane A, Lindon JC, Marriott P, Nicholls A, Reily M, Thaden J, Viant MR (2007) Proposed minimum reporting standards for chemical analysis. Metabolomics 3 (3):211–221
10. 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 M-A, 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 Res 35(suppl 1):D521–D526
11. Böcker S, Letzel M, Lipták ZS, Pervukhin A (2009) STRTUS: decomposing isotope patterns for metabolite identification. Bioinformatics 25(2):218–224
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 Bioinformatics 7(1):234
13. Bristow T, Constantine J, Harrison M, Cavoit F (2008) Performance optimisation of a new-generation orthogonal-acceleration quadrupole-time-of-flight mass spectrometer. Rapid Commun Mass Spectrom 22(8):1213–1222
14. Laures AM-F, Wolff J-C, Eckers C, Borman PJ, Chatfield MJ (2007) Investigation into the factors affecting accuracy of mass measurements on a time-of-flight mass spectrometer using Design of Experiment. Rapid Commun Mass Spectrom 21(4):529–535
15. Xu Y, Heilier J-F, Madalinski G, Genin E, Ezan E, Tabet J-C, Junot C (2010) Evaluation of accurate mass and relative isotopic abundance measurements in the LTQ-Orbitrap mass spectrometer for further metabolomics database building. Anal Chem 82 (13):5490–5501. doi:10.1021/ac100271j
16. Miura D, Tsuji Y, Takahashi K, Wariishi H, Saito K (2010) A strategy for the determination of the elemental composition by Fourier transform ion cyclotron resonance mass spectrometry based on isotopic peak ratios. Anal Chem 82(13):5887–5891
17. Matsuda F, Shinbo Y, Oikawa A, Hirai MY, Fiehn O, Kanaya S, Saito K (2009) Assessment of metabolome annotation quality: a method for evaluating the false discovery rate of elemental composition searches. PLoS ONE 4(10):e7490
18. 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(3):555–577
19. 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 Physiol 152(2):566–578
20. Plumb RS, Johnson KA, Rainville P, Smith BW, Wilson ID, Castro-Perez JM, Nicholson JK (2006) UPLC/MS(E); a new approach for generating molecular fragment information for biomarker structure elucidation. Rapid Commun Mass Spectrom 20(13):1989–1994
21. Ipsen A, Want EJ, Lindon JC, Ebbels TMD (2010) A statistically rigorous test for the identification of parent-fragment pairs in LC-MS datasets. Anal Chem 82(5):1766–1778
22. Tautenhahn R, Böttcher C, Neumann S (2007) Annotation of LC/ESI-MS mass signals. In: Hochreichter S, Wagner R (eds) Bioinformatics research and development (BIRD 2007). Lecture notes in computer science, vol 4414. Springer, Heidelberg, pp 371–380
23. Borland L, Brickhouse M, Thomas T, Fountain AW (2010) Review of chemical signature databases. Anal Bioanal Chem 397 (3):1019–1028
24. Gower JC, Legendre P (1986) Metric and Euclidean properties of dissimilarity coefficients. J Classif 3(1):5–48
25. Stein SE (1994) Estimating probabilities of correct identification from results of mass spectral library searches. J Am Soc Mass Spectrom 5(4):316–323
26. 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, Nakanishi H, Ikeda K, Akimoto N, Maoka T, Takahashi H, Ara T, Shibata D, Neumann S, Iida T, Tanaka K, 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. J Mass Spectrom 45:703–714
27. Dworzanski JP, Snyder AP, Chen R, Zhang H, Wishart D, Li L (2004) Identification of bacteria using tandem mass spectrometry combined with a proteome database and statistical scoring. Anal Chem 76(8):2355–2366
28. Pavlic M, Libiseller K, Oberacher H (2006) Combined use of ESI-QqTOF-MS and ESI-QqTOF-MS/MS with mass-spectral library search for qualitative analysis of drugs. Anal Bioanal Chem 386(1):69–82
29. Oberacher H, Pavlic M, Libiseller K, Schubert B, Sulyok M, Schuhmacher R, Csaszar E, Köfeler HC (2009) On the inter-instrument and the inter-laboratory transferability of a tandem mass spectral reference library: 2. Optimization and characterization of the search algorithm. J Mass Spectrom 44(4):494–502
30. Mylonas R, Mauron Y, Masselot A, Binz P-A, Budin N, Fathi M, Viette V, Hochstrasser DF, Lisacek F (2009) X-Rank: a robust algorithm for small molecule identification using tandem mass spectrometry. Anal Chem 81(18):7604–7610
31. Smith CA, Maille GO, Want EJ, Qin C, Trauger SA, Brandon TR, Custodio DE, Abagyan R, Siuzdak G (2005) METLIN: a metabolite mass spectral database. In: Proceedings of the 9th international congress of therapeutic drug monitoring and clinical toxicology, Louisville, Kentucky, vol 27, pp 747–751
32. Reemtsma T (2009) Determination of molecular formulas of natural organic matter molecules by (ultra-) high-resolution mass spectrometry: status and needs. J Chromatogr A 1216(18):3687–3701
33. Mohamed R, Varesio E, Ivosev G, Burton L, Bon-ner R, Hopfgartner G (2009) Comprehensive analytical strategy for biomarker identification based on liquid chromatography coupled to mass spectrometry and new candidate confirmation tools. Anal Chem 81(18):7677–7694
34. Böcker S, Rasche F (2008) Towards *de novo* identification of metabolites by analyzing tandem mass spectra. Bioinformatics 24: T49–T55, Proc. of European Conference on Computational Biology (ECCB 2008)
35. Advanced Chemistry Development, Inc (2010) ACD/MS Fragmenter. http://www.acdlabs.com/products/adh/ms/ms_frag/
36. Pelander A, Tyrkkö E, Ojanperä I (2009) In silico methods for predicting metabolism and mass fragmentation applied to quetiapine in liquid chromatography/time-of-flight mass spectrometry urine drug screening. Rapid Commun Mass Spectrom 23(4):506–514
37. Tyrkkö E, Pelander A, Ojanperä I (2010) Differentiation of structural isomers in a target drug database by LC/Q-TOFMS using fragmentation prediction. Drug Test Anal 2(6):259–270
38. Highchem, Ltd (2010) Mass Frontier. http://www.highchem.com/massfrontier/mass-frontier.html
39. Horai H, Arita M, Ojima Y, Nihei Y, Kanaya S, Nishioka T (2009) Traceable analysis of multiple-stage mass spectra through

precursor-product annotations. In: Grosse I, Neumann S, Posch S, Schreiber F, Stadler PF (eds) GCB. Lecture notes in informatics (GI), vol 157, pp 173–178
40. Heinonen M, Rantanen A, Mielikäinen T, Kokkonen J, Kiuru J, Ketola RA, Rousu J (2008) FiD: a software for *ab initio* structural identification of product ions from tandem mass spectrometric data. Rapid Commun Mass Spectrom 22(19):3043–3052
41. Hill AW, Mortishire-Smith RJ (2005) Automated assignment of high-resolution collisionally activated dissociation mass spectra using a systematic bond disconnection approach. Rapid Commun Mass Spectrom 19(21):3111–3118
42. Heinonen M, Rantanen A, Mielikäinen T, Pitkänen E, Kokkonen J, Rousu J (2006) *Ab initio* prediction of molecular fragments from tandem mass spectrometry data. In: Proceedings of the German conference on bioinformatics (GCB 2006). Lecture notes in informatics, pp 40–53
43. Böcker S, Rasche F, Steijger T (2009) Annotating fragmentation patterns. In: Proceedings of the workshop on algorithms in bioinformatics (WABI 2009). Lecture notes in computer science, vol 5724. Springer, Heidelberg, pp 13–24
44. Hill DW, Kertesz TM, Fontaine D, Friedman R, Grant DF (2008) Mass spectral metabonomics beyond elemental formula: chemical database querying by matching experimental with computational fragmentation spectra. Anal Chem 80(14):5574–5582
45. Wolf S, Schmidt S, Müller-Hannemann M, Neumann S (2010) In silico fragmentation for computer assisted identification of metabolite mass spectra. BMC Bioinformatics 11(1):148
46. Levsen K, Schiebel H-M, Terlouw JK, Jobst KJ, Elend M, Preiss A, Thiele H, Ingendoh A (2007) Even-electron ions: a systematic study of the neutral species lost in the dissociation of quasi-molecular ions. J Mass Spectrom 42(8):1024–1044
47. Alex A, Harvey S, Parsons T, Pullen FS, Wright P, Riley J-A (2009) Can density functional theory (DFT) be used as an aid to a deeper understanding of tandem mass spectrometric fragmentation pathways? Rapid Commun Mass Spectrom 23(17):2619–2627
48. Wright P, Alex A, Nyaruwata T, Parsons T, Pullen F (2010) Using density functional theory to rationalise the mass spectral fragmentation of maraviroc and its metabolites. Rapid Commun Mass Spectrom 24(7):1025–1031