Research article

Journal of MASS SPECTROMETRY

Received: 20 April 2012 Revised: 1 October 2012 Accepted: 11 October 2012 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI 10.1002/jms.3123

MetFusion: integration of compound identification strategies

Michael Gerlich* and Steffen Neumann

Mass spectrometry (MS) is an important analytical technique for the detection and identification of small compounds. The main bottleneck in the interpretation of metabolite profiling or screening experiments is the identification of unknown compounds from tandem mass spectra.

Spectral libraries for tandem MS, such as MassBank or NIST, contain reference spectra for many compounds, but their limited chemical coverage reduces the chance for a correct and reliable identification of unknown spectra outside the database domain.

On the other hand, compound databases like PubChem or ChemSpider have a much larger coverage of the chemical space, but they cannot be queried with spectral information directly. Recently, computational mass spectrometry methods and *in silico* fragmentation prediction allow users to search such databases of chemical structures.

We present a new strategy called MetFusion to combine identification results from several resources, in particular, from the *in silico* fragmenter MetFrag with the spectral library MassBank to improve compound identification. We evaluate the performance on a set of 1062 spectra and achieve an improved ranking of the correct compound from rank 28 using MetFrag alone, to rank 7 with MetFusion, even if the correct compound and similar compounds are absent from the spectral library. On the basis of the evaluation, we extrapolate the performance of MetFusion to the KEGG compound database. Copyright © 2013 John Wiley & Sons, Ltd.

Supporting information may be found in the online version of this article.

Keywords: metabolomics; integrated identification; MassBank; MetFrag; *in silico* fragmentation

Introduction

Soft ionization mass spectrometry, often coupled to liquid chromatography (LC-ESI-MS), has been established as an important analytical technology in several applications, such as metabolomics or screening of unknowns in the environmental sciences.[1,2] In untargeted approaches, complex samples are analyzed by LC-ESI-MS and can lead to elucidation of metabolites in biosynthetic pathways,[3,4] discovery of biomarkers,[5,6] prediction of disease states or detection of emerging pollutants in water samples.[7,8] However, these compounds are only characterized by their mass-to-charge ratio (m/z) and retention time, and subsequent identification requires substantial effort.

Tandem mass spectra provide valuable structural hints for the identification and structure elucidation of compounds and can be obtained from ion-trap or hybrid instruments, such as triple-quadrupole (QqQ) or quadrupole-time-of-flight (QqTOF). Collision-induced dissociation (CID) is a common fragmentation method for small compounds, resulting in a detailed fragmentation spectrum.[9]

These characteristic fragmentation patterns are available from spectral libraries such as MassBank,[10] HMDB,[11] where version v2.5 contains 2654 compounds with three MS/MS reference spectra on average, and NIST'11[12] that provides an MS/MS library with a total of 95 409 spectra representing 5843 compounds, including dipeptides and tripeptides, and METLIN,[13] which contains 48 596 high-resolution spectra for 10 076 metabolites as of February 2012.

MassBank is the first open community repository for mass spectral data (including spectral information, as well as analytical conditions) and provides both a web interface for human interaction and an application programming interface for programmatic access to the data and search functions. MassBank contains spectra from different instruments, including QqTOF, QqQ and ion-trap from different vendors. Most compounds are measured under various analytical conditions, for example in both positive and negative mode, or at several collision energies. The federated architecture of MassBank provides access to distributed data contributed by various institutes. There are approximately about 13 623 spectra high-resolution ESI-spectra representing about 2000 compounds in MassBank as of February 2012 (including redundancies, where the same compound was measured with different analytical settings on various instruments). A sample query result is shown in Figure 1.

Compound databases like PubChem,[14] KEGG[15] or ChemSpider[16] provide information on a huge number of both natural products and synthetic compounds. Although these databases excel in terms of chemical information (measured and predicted chemical properties, structure information and for some also assay results), they do not support queries using mass spectral measurements. The acquisition of reference spectra of all known compounds is unfeasible because of the enormous efforts required and the limited availability of commercial standards.

To alleviate the problem of limited availability of reference spectra, computational mass spectrometry tools with *in silico* spectra prediction have been developed.[17] *MetFrag* is a free and open-source program for compound identification based

* *Correspondence to: Michael Gerlich, Department of Stress and Developmental Biology, Leibniz Institute of Plant Biochemistry, Germany. E-mail: mgerlich@ipb-halle.de*

Department of Stress and Developmental Biology, Leibniz Institute of Plant Biochemistry, Germany

Copyright © 2013 John Wiley & Sons, Ltd.

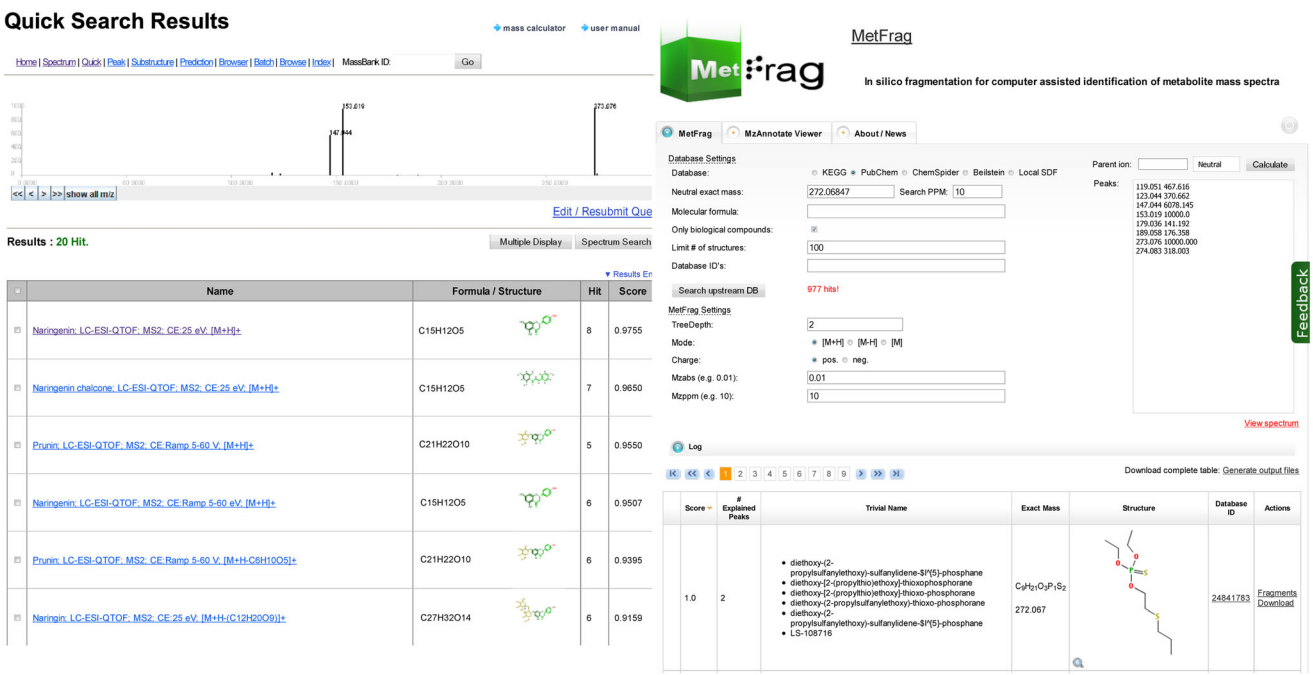

Figure 1. Individual results from MassBank and MetFrag. Left: The upper part of the MassBank screenshot shows the query spectrum, below are the resulting matches, sorted by spectrum similarity. Right: The top part of MetFrag contains the query input, the result list is presented below.

on compound databases.[18] The MetFrag web interface is shown in Figure 1. MetFrag obtains candidate structures from a compound database and matches *in silico* predicted fragments to the query spectrum. Each candidate is ranked according to a fragmentation score.

But despite the advent of *in silico* tools, reference spectra obtained under comparable analytical conditions are still the preferred way to achieve a reliable compound identification. To the best of our knowledge, there is currently no approach to integrate both strategies, where the most reliable answers of the two are returned.

In this paper, we present *MetFusion*, a strategy and system to *combine* the compound hypotheses obtained by complementary identification approaches. Here, we integrate the results from MassBank and MetFrag. This strategy combines the best of both worlds: the identification using spectral libraries if similar spectra are available and the huge chemical coverage of the compound databases queried by MetFrag.

Methods

In the following, we describe how the individual compound identification sources are queried, show the mathematical background for the integrated score and depict the web application. Subsequently, we explain the evaluation dataset and our evaluation approach to make the results more realistic and finally extrapolate the generalization to any compound in KEGG.

System architecture

The underlying assumption in MetFusion is that the correct compound is present in the compound database and consequently among the structure candidates in the MetFrag result. The idea of MetFusion is to confirm the *in silico* predicted results with spectral reference data and calculate a new integrated score for each candidate processed by MetFrag. This is depicted in the workflow shown in Figure 2.

The MassBank scores are calculated on the basis of a modified cosine distance to compute the similarity between the query spectrum and the reference spectra.[10] Results are ranked according to this spectral similarity. MassBank is accessed using a Java library application programming interface, which queries the individual servers, and passes the relevant parameters (intensity cutoff, ionization mode and instrument filter) directly to the servers.

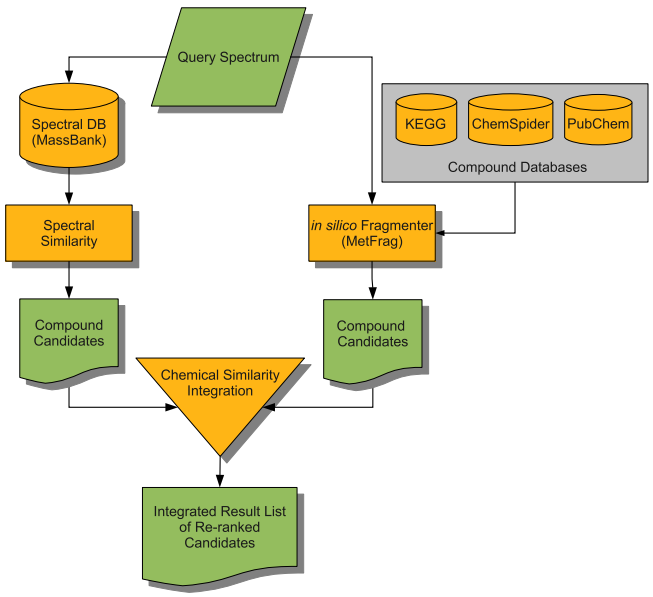

Figure 2. The MetFusion workflow: the query spectrum is passed to both the MassBank and the MetFrag query tools. Both return ranked lists, providing spectra matches and compound candidates, respectively. These lists are combined by calculating the chemical similarity between all structures. The integrated score is used to re-rank the list of MetFrag candidates from the compound database.

Copyright © 2013 John Wiley & Sons, Ltd.

Journal of MASS SPECTROMETRY

The *in silico* fragmentation is performed with an embedded MetFrag module, which queries KEGG, PubChem or ChemSpider as a compound database. In addition, local compound libraries can be used, which allows the use of in-house compound databases or mirrors of, e.g. PubChem. This is performed either by SD file upload or direct database access. Likewise, this upload allows users to submit their own generated structures as candidates for *in silico* fragmentation.

MetFusion requires a spectrum with *m/z* and intensity values as query input and passes the provided settings to the corresponding databases. MetFrag settings can also be adjusted, most importantly the allowed *mass deviation* for the generated fragments.

The core of MetFusion is implemented as a Java library, which is used both by the command line and web interface. Both the MassBank and MetFrag queries are performed in separate threads and run in parallel.

Integration of spectral matches, in silico scores and chemical similarity

The identification strategies return two individual lists of spectra matches and candidate compounds, both with associated scores. The spectra scores are combined into a *spectral summary*. This is an aggregation of similar spectra and their respective chemical similarity to a candidate compound.

The spectral library can contain multiple measurements of a single compound or its isomeric variants, so we use an InChIKey-based filtering of the original MassBank result list, which only retains the highest-scoring record for each compound constitution. This is also justified because distinguishing between stereoisomers is hardly possible with mass spectrometry alone. For this filter step, we rely on the connectivity information stored in the first block of the InChIKey.[19]

Equation (1) describes the integrated MetFusion score s_c: for each MetFrag candidate c, we calculate s_c as a sum of the MetFrag score f_c and the 'spectral summary' on the basis of the scores m_j for all MassBank results j, and the chemical similarity t_{cj} between MetFrag candidate c and each MassBank result j. The number of results from MetFrag is denoted by N, and the number of MassBank results is denoted by M. This leads to an $N \times M$ matrix of chemical similarities. An excerpt of such a matrix can be seen in Figure 3.

The chemical similarity t_{cj} between a MetFrag candidate c and each MassBank result j allows us to determine how similar each pair of compounds is. This provides a validation of *in silico* generated spectra with measured spectra, based not on spectral similarity but rather on the chemical similarity between the corresponding compounds represented by their spectra. This approach results in the integrated score, allowing us to rank the MetFrag candidates with an additional level of information.

We use the sigmoid function *sig(x)* shown in Equation (2) to introduce a non-linear behavior, which reduces the influence of mediocre spectral matches and chemical similarities. Further information about the sigmoid function is available in Supplemental Material S-1, describing the impact of the parameters β and γ.

$$s_c = \underbrace{\alpha * f_c}_{MetFrag} + \overbrace{(1-\alpha) * \sum_{j=1}^{M} sig(m_j * t_{cj})}^{\text{"spectral summary"}} \quad (1)$$

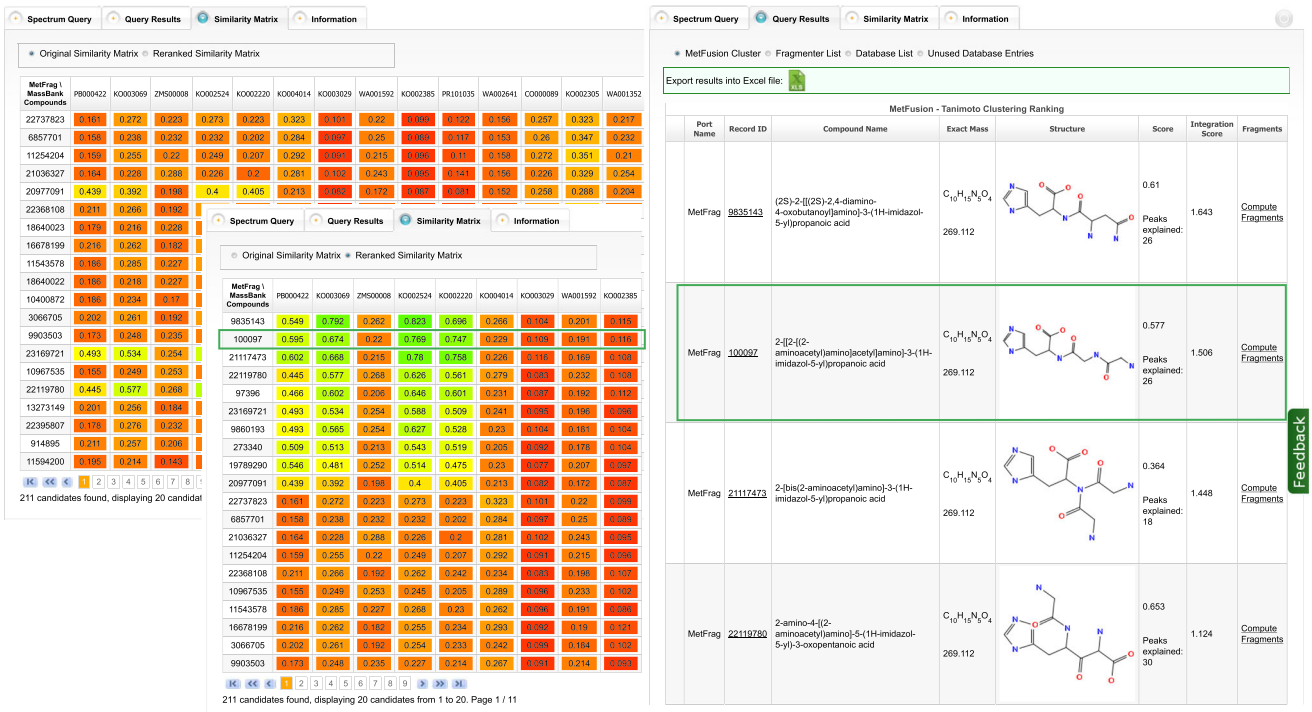

Figure 3. The MetFusion web application. Left: The background shows an excerpt of the similarity matrix for a query with the Gly-Gly-His spectrum (NIST# 1012075, CID: 100097) prior to re-ranking. Columns contain MassBank results, and rows correspond to MetFrag results. Each cell shows the chemical similarity between 0 (red) to 1 (green). The correct structure appears at tied rank 23 (not visible). None of the top MetFrag candidates show high chemical similarity. Overlaid is the similarity matrix after re-ranking. Here, several MetFrag candidates have a reasonable similarity to MassBank results, and the correct candidate is circled in green. The combination via chemical similarity improves the rank of the correct structure to two. Right: The head of the re-ranked MetFusion output, showing the results with structure formula, database link and scores.

 Copyright © 2013 John Wiley & Sons, Ltd.

$$sig(x) = \frac{1}{1 + e^{(\beta*(x-\gamma))}} \quad (2)$$

The 'spectral summary' for the candidate c is then the sum of all MassBank scores m_j, weighted by their chemical similarity t_{cj} to the candidate c. The MassBank spectral scores m_j use a modified cosine distance in the range from 0 to 1, where values ≥0.65 indicate reasonable spectral similarity.[10]

For the chemical similarity calculation, we use the Chemistry Development Kit (CDK, version 1.4.7).[20] The chemical similarity t_{cj} between the molecular fingerprints (CDK standard fingerprint with 1 024 bit length) of the compounds c and j is calculated using the Tanimoto (also known as Jaccard) coefficient.[21,22]

The balance between the individual identification approaches is determined by the weight α, where $\alpha = 1$ uses exclusively the MetFrag scores and $\alpha = 0$ results in a compound library search for those compounds that have the most similar high-scoring MassBank hits. Although both individual MetFrag and MassBank scores fall in the range of 0–1, the MetFusion result score has no upper bound and depends on the original MetFrag score f_c, the number of spectral database hits and the corresponding chemical similarity. The lower limit of the MetFusion score is 0.

Evaluation method and dataset

MetFusion was evaluated on a dataset of 1099 spectra, containing compounds ranging in molecular weight from 89 to 837 Da. A wide range of compound classes is covered, including flavonoids, steroids, amino acids, carboxylic acids, glucosides, drugs and toxins. Nine hundred and eighteen spectra were measured with a single collision energy (such as 10, 20, 30, 40 and 50 eV). The remaining 181 spectra were created by merging spectra at several collision energies for a single compound. The corresponding spectra were measured on the same instrument type, and only the collision energies differed. In this way, more informative peaks are present in a merged spectrum. The use of merged spectra for similarity search is also recommended by the MassBank consortium.[10]

The reference spectra used to evaluate MetFrag[18,23] are a subset of this evaluation dataset. All spectra are available from MassBank; for details, see Supplemental Material S-7.

The dataset contains 37 spectra, which contain only the precursor ion information, resulting from soft ionization with 10 eV. The results presented in the main article exclude these spectra, but the complete results can be found in Supplemental Materials S-5 to S-8.

For each test spectrum, we determine the rank of the correct candidate obtained with MetFusion. For the evaluation, we consider all different configurations of a candidate compound as a single constitution because the compound databases often include several stereoisomers and unspecified stereo configurations. We again use an InChIKey-based filtering of the candidate list.

The *relative ranking position* (RRP) describes the position of the correct compound in relation to the whole result set.[24,25]

$$RRP = \frac{1}{2}\left(1 - \frac{BC - WC}{TC - 1}\right) \quad (3)$$

In Equation (3), BC denotes the number of candidates that have a higher MetFusion score than the correct compound. WC denotes the number of candidates that have a lower score. TC denotes the total number of candidates, i.e. the number of MetFrag results N.

We have defined Equation (3) such that an RRP of 1.0 is equivalent to the correct compound at the first position, this value also implies that no other compound is ranked first. If the compound is ranked last, this results in an RRP of 0.0. If all compounds share the same score, this results in an RRP of 0.5.

We also report the median rank of the correct solution, which indicates how many candidates have to be considered before the correct solution appears in the web application. If several compounds (including the correct solution) have an identical score, we use the most conservative approach and report the maximum (worst case) rank of equally scored candidates.

Simulation of real world queries for training and evaluation

We cannot use the 'normal' operation mode of MetFusion to evaluate the identification performance, as all test spectra are also present in MassBank. If we did so, MetFusion would be simply 'too good'. This is because of the fact that querying MassBank with spectra from our dataset is guaranteed to find matches at the top positions, as these spectra are present in MassBank. The correct candidate would also have a Tanimoto similarity of 1.0, which would favor its scoring even more because the parameter optimization would result in a scoring function strongly biased towards MassBank.

To avoid this, we simulated the identification of *unknown* spectra: we removed not only the query spectrum from the MassBank results in our evaluation but also any spectra whose compounds were above a certain chemical similarity to the query compound. This filtering approach provides controlled conditions for the evaluation because for an independent set of evaluation spectra, we also would need to specify to what degree the compounds are present in the reference library. We used 0.7 as the most stringent Tanimoto similarity threshold, which removed on average 2.4 MassBank records for each test spectrum. Less stringent filters are 0.8 and 0.9, which removed on average 1.8 and 1.3 results, respectively. A threshold of 1.0 removed only the correct compound. So with just one set of test spectra, we can evaluate our approach against several levels of completeness of the spectral library used and find to what degree the identification depends on the reference spectra.

MetFusion web application and availability

The MetFusion application is available at http://msbi.ipb-halle.de/MetFusion/ and features a user-friendly interface.

The query spectrum should contain at least *m/z* and *intensity* values. The three column MassBank peaklist format is supported as well. Additional search parameters allow users to modify the search behavior for compound database and spectral database, respectively. The results are presented in a table with 20 entries per page, ordered by decreasing MetFusion score. It is also possible to download all results as a spreadsheet, which contains the result lists of MetFusion, MetFrag and MassBank, with corresponding scores and images of the molecular structures and the computed similarity matrix. Users can then add this report to the supplemental information of publications to support their findings.

It is possible that several candidate compounds obtain identical scores and thus have tied ranks. To improve the overview of the MetFusion result list in the web application, we perform a

Copyright © 2013 John Wiley & Sons, Ltd.

Journal of MASS SPECTROMETRY

structural clustering of all compounds that have the same MetFusion score and a Tanimoto fingerprint similarity ≥0.95 and join them into a cluster. This applies in particular to stereo isomers, which can in general not be distinguished with mass spectrometry. In contrast to our evaluation, the web application does not perform any filtering of the candidates because for a downstream analysis (such as citation counts), the full candidate list could be relevant. The clustered results can be expanded and viewed in detail by the user.

The web interface is based on Java Server Faces 2, ICEfaces 2 (component library with AJAX capabilities) and an Apache Tomcat 7 server.

The application is suitable for browsers with JavaScript enabled.

The MetFusion implementation is available as a Java library from the project repository at https://github.com/mgerlich/MetFusion, which can be used to perform batch searches on a local computer or cluster. The code is available under the open-source GPL license.

Results and discussion

MassBank contains spectra from various MS instruments and chromatography types. For this paper, we focused on ESI tandem MS spectra. This includes 13 instrument types with a total of 13 623 spectra as of February 2012. The result size for a MassBank query was limited to the best 100 records.

MetFrag was used with two different values for the *mzabs* parameter. This parameter defines the allowed absolute mass deviation between the *in silico* generated fragments and the measured peaks. Spectra that were measured on high-resolution devices with good accuracy used *mzabs* = 0.0, so only the relative *mzppm* error threshold is used. For less accurate spectra, we increased this value to *mzabs* = 0.01, allowing a broader range for the exact mass of generated fragments to match. The additional parameter *mzppm* was set to 10 ppm in all cases. After filtering for unique InChIKeys, we found that the result list contained on average 1247 candidates per query spectrum from the PubChem database.

Optimization of scoring function parameters

The integrated scoring function has three internal parameters: α balances the *in silico* prediction and the spectral summary, and β and γ determine the shape of the sigmoid function in the spectral summary. For an optimal choice of these parameters, we performed a parameter scan using the complete set of 1 099 spectra for each of the filter thresholds. The resulting parameter sets are shown in Supplemental Material S-2. To assess the stability and generalization of such a parameter optimization, we performed a tenfold cross-validation. Across all ten partitions, we obtained very similar optimal parameter combinations when optimizing the mean rank of the correct compound, which suggests that the scoring function is robust to parameter and data variations. The detailed results of the cross-validation are shown in Supplemental Material S-2.

For the remainder of the paper and for the web application, we have chosen the parameter set obtained with the similarity filtering threshold of 0.9. The corresponding optimal parameters are thus $\alpha = 0.3$, $\beta = -9$ and $\gamma = 0.6$. Additional information on the performance of MetFusion is available in Supplemental Material S-3 to S-8.

Examples: Gly-Gly-His and naringenin

First, we have selected two example query spectra to demonstrate MetFusion and discuss the results.

We selected a spectrum for the tripeptide Glycine-Glycine-Histidine (Gly-Gly-His, Pubchem CID: 100097) with 42 peaks, measured on a Micromass Quattro Micro QqQ device with nominal mass resolution from the NIST 2008 database. MassBank has very little spectral information on dipeptides and almost none for tripeptides or polypeptides. However, the basic amino acids are present in MassBank. So the challenges for MetFusion are to deal with the low mass resolution and the lack of a reference spectrum for this compound.

Gly-Gly-His has an exact mass of 269.112 Da. We modified the MetFrag parameters and increased *mzabs* to 0.1 Da and *mzppm* to 30 ppm to account for the low resolution spectrum.

With MetFrag alone, the top ranked candidates explain up to 35 fragments, and many have purine or furan substructures. The correct structure explains 26 fragments and is returned at tied rank 23. The first MassBank hits contain spectra for Histidine (155.069 Da, best score 0.856), Carnosine (a dipeptide of β-Alanine and Histidine, 226.106 Da) and L-Homocarnosine (240.122 Da, score of 0.798). Figure 3 shows the similarity matrix dominated by chemical similarities < 0.3.

Although none of the MassBank hits fully resemble the tripeptide of interest, the basic building blocks and their corresponding characteristic fragment peaks provide enough information to obtain the higher rank of the correct compound. The rank of Gly-Gly-His is improved from rank 23 in MetFrag to rank 2 in MetFusion. After MetFusion was run, the visual inspection of the similarity matrix helps to interpret the result and avoid some pitfalls. In another example, for the naringenin chalcone (CID: 155802) spectrum, both MetFrag and MassBank results also contain the related naringenin (CID: 932). The spectrum PB000129 of naringenin chalcone has a MassBank score of 0.98 compared with the spectrum PB000125 of naringenin. The ring break of naringenin chalcone leads to very low chemical similarity scores, thus promoting the rank of naringenin with intact rings and its higher spectral and chemical similarity towards naringin (spectrum PB000804, CID: 442428), favoring naringenin over naringenin chalcone. Additional information is available in Supplemental Material S-3.

Evaluation with benchmark dataset

We performed the evaluation on both the reduced dataset of 1062 spectra, which excluded spectra that contain only precursor ion information, as well as the complete dataset with 1099 spectra. The latter are available in Supplemental Material S-5. We first queried MetFrag separately and use these results as baseline. As stated before, we applied the InChIKey-based filter step to the candidate lists prior to the MetFusion combination to remove duplicated constitutions. Here, the correct compound had a median tied rank position of 28 and a mean rank position of 164. The corresponding median RRP is 0.959, and the mean RRP is 0.886. The discrepancy between mean and median shows that the distribution is skewed and the low performance for several compounds increases the mean considerably.

For the evaluation of our MetFusion strategy, we used the simulated real world queries for the evaluation dataset. We chose the similarity filter of 0.9 as the basis for the optimization and evaluation. With this setting, we obtain the correct solution among the

 Copyright © 2013 John Wiley & Sons, Ltd.

top 2% (median RRP 0.991) in the result list or at an absolute rank 7 (median). Without filtering the correct compound from the MassBank results, we obtain a median RRP of 1.

With the most restrictive filtering we used (similarity threshold 0.7), the median RRP drops to 0.986, with a median rank of 10. This filter setting removed on average 2.4 spectra from the MassBank results, and in one case up to 23. With these results, one can expect to find the correct solution on the first page of the result list of the web application. The other (less pessimistic) filter settings are shown in Table 1.

The main advantage of this approach is the combination of two separate identification approaches: (1) Instead of dealing with multiple interfaces, all results are available in a single application. More importantly, (2) MetFusion does neither depend solely on *in silico* prediction nor on the possibly poor coverage of reference spectra. A distinct advantage is that the spectrum search from MassBank will not only retrieve spectra from compounds with the actual precursor mass but also includes related compounds with different masses that share similar fragment peaks. These peaks can be attributed to similar structural features of a compound. MetFrag usually retrieves the candidates on the basis of the precursor mass or elemental composition, so all candidates of a query will have the same mass. Hence, if the correct compound is contained in the compound database, it is also included in the MetFusion result list.

Please also note that in this evaluation, we used PubChem to demonstrate the ability to process large compound databases, although for metabolomics applications, many of the candidates will be irrelevant. Generally, an experimentalist will have additional prior information, which can be used to ignore candidates that could not occur in the sample under investigation.

These results show that combining an *in silico* approach with curated reference measurements can directly improve compound identification and give the best of both worlds.

Extrapolation to KEGG

The results presented earlier show the performance of MetFusion on the benchmark dataset from MassBank. But what performance can we expect for arbitrary compounds in, e.g. KEGG? This depends on the number of 'similar' reference spectra available in MassBank for each KEGG compound so that the results in Table 1 can be extrapolated to the KEGG compound database.

We calculated the pairwise chemical similarity between compounds in MassBank and KEGG. Using the last publically available KEGG COMPOUND snapshot (15 499 entries as of 24 June 2011) and a local MassBank database of 5 063 compound structures for which ESI reference spectra are available, we found that for 2690 KEGG entries, there is a MassBank record with a Tanimoto similarity of 0.9 or better. Additional information is available in Supplemental Material S-4.

Under the assumption that our compound selection in the test data is unbiased and that Table 1 can be generalized to all KEGG compounds, we would expect that half of these 2690 compounds can be ranked among the first seven MetFusion results, even if they are searched against the whole of PubChem. If we relax the restrictions, we find that for 5513 entries in KEGG, there is a MassBank spectrum with a Tanimoto similarity of at least 0.7, so the extrapolation from Table 1 suggests a median rank of 10 for the correct compound.

We were able to validate this extrapolation on the subset of 180 unique compounds from our dataset that also provide a KEGG identifier. We used the identical settings as for the full benchmark dataset and also retrieved the candidates from PubChem. The results are shown in Table 1. Although the RRPs are slightly lower, the absolute ranks are even slightly better. One reason is that PubChem returned fewer candidates for the compounds also present in KEGG.

These calculations are just an extrapolation under several assumptions, which cannot be taken for granted. If a compound is not amendable to mass spectrometry, e.g. because of low ionization efficiency, the identification is impossible with MetFusion, and other analytical methods have to be used. Secondly, the extrapolation assumes the same performance on compounds not contained in the benchmark dataset. Although there are several diverse compound classes in the evaluation dataset we used, the benchmark data could be biased and MetFusion could have a different performance (lower, but also higher) for classes not taken into account here. On the other hand, because of the distributed nature of MassBank, we only considered those structures that were available in our local database mirror, so the number of KEGG structures for which similar reference spectra are available is definitely higher than 2690.

Table 1. Results of MetFusion for the 1062 spectra dataset with the (artificially) filtered MassBank

		Similarity filter				
		0.7	0.8	**0.9**	1	none
MetFusion	Rank	10	8	7	4	1
	RRP	0.986	0.990	0.991	0.993	1
KEGG only	Rank	8	6	6	4	1
	RRP	0.976	0.984	0.987	0.989	1

Both the median rank and the median relative ranking position (RRP) are shown for a given filter stringency. In addition, the results for 180 unique KEGG compounds are presented.

Conclusions

The MetFusion approach was developed to combine the knowledge from reference spectra with the *in silico* prediction tool MetFrag for structure identification in metabolomics studies and small molecules in general. We showed that merging this information via chemical similarity improves the position of the correct compound from rank 28 to rank 7 compared with *in silico* prediction alone. This improvement is even more remarkable, given that in the evaluation, we made sure that reference spectra of the correct (and similar) compounds were excluded. Solely relying on the spectral library as identification strategy would result in no – or a wrong – identification for these cases.

As this paper used existing benchmark data from spectral libraries, we have described a metabolomics workflow elsewhere.[26] The MetShot approach first obtains a list of peaks of interest from metabolite profiling and statistical analysis and acquires high-quality tandem mass spectra, which can be converted to MetFusion batch query files.

The metabolomics standards initiative has defined four levels of confidence in metabolite identification.[27] The most confident level 1 identification requires the comparison of an unknown compound with authentic standards under the same analytical conditions, and level 2 can be achieved by a comparison of spectral data with literature or database information of the same compound. MetFrag alone can be considered to achieve the

Copyright © 2013 John Wiley & Sons, Ltd.

Journal of MASS SPECTROMETRY

annotation of compound classes, resulting in a level 3 identification. With MetFusion, we can often achieve a level between two and three, even if a reference spectrum of the actual unknown is absent from MassBank. Beyond MetFusion, additional steps such as the comparison of retention time and predicted logP ranges, UV spectra or filtering for known substructures can further improve the confidence of the identification. Some of these aspects have been evaluated elsewhere.[28]

The approach is generally applicable to any identification strategies that return compound structures and can be modified for other spectral libraries (such as Metlin, HMDB or GMD[29]) as well as other identification strategies, such as the recently published analysis of fragmentation trees.[30–32] Furthermore, it is not only restricted to tandem MS spectra but can readily be applied to MS1 spectra with informative in-source fragments, MS*n* data and GC-MS spectra.

Because the number of known metabolites will grow faster than the coverage of spectral libraries, our approach to integrate multiple identification strategies will remain of high importance in the future. Even if all KEGG compounds (as of today) were available in MassBank, the huge number of 200 000 metabolites estimated in the plant kingdom[33] will not be part of reference libraries any time soon.

Of course, a major improvement in general would be an increase in high-resolution spectra contributions to reference databases. The analysis of the chemical similarities between MassBank and KEGG allows to prioritize future efforts and select substances for which spectra should be added to MassBank to improve the coverage of biologically relevant reference spectra.

Acknowledgements

The authors thank all contributors for providing mass spectra to MassBank. We also like to thank the CDK developers, Sebastian Wolf who developed MetFrag and Emma Schymanski for helpful discussions.

Supporting information

Supporting information may be found in the online version of this article.

References

[1] W. Dunn, A. Erban, R. Weber, D. Creek, M. Brown, R. Breitling, T. Hankemeier, R. Goodacre, S. Neumann, J. Kopka, M. Viant. Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics* **2012**, 1–23. 10.1007/s11306-012-0434-4.

[2] M. Krauss, H. Singer, J. Hollender. LC-high resolution MS in environmental analysis: from target screening to the identification of unknowns. *Anal. Bioanal. Chem.* **2010**, *397*, 943–951. 10.1007/s00216-010-3608-9.

[3] C. Böttcher, E. von Roepenack-Lahaye, J. Schmidt, C. Schmotz, S. Neumann, D. Scheel, S. Clemens. 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.* **2008**, *147*(4), 2107–2120.

[4] Y. Okazaki, M. Shimojima, Y. Sawada, K. Toyooka, T. Narisawa, K. Mochida, H. Tanaka, F. Matsuda, A. Hirai, M. Y. Hirai, H. Ohta, K. Saito. A chloroplastic UDP-glucose pyrophosphorylase from Arabidopsis is the committed enzyme for the first step of sulfolipid biosynthesis. *Plant Cell* **2009**, *21*(3), 892–909.

[5] G. Glauser, D. Guillarme, E. Grata, J. Boccard, A. Thiocone, P.-A. Carrupt, J.-L. Veuthey, S. Rudaz, J.-L. Wolfender. 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* **2008**, *1180*(1-2), 90–98.

[6] R. Mohamed, E. Varesio, G. Ivosev, L. Burton, R. Bonner, G. Hopfgartner. Comprehensive analytical strategy for biomarker identification based on liquid chromatography coupled to mass spectrometry and new candidate confirmation tools. *Anal. Chem.* **2009**, *81*(18), 7677–7694.

[7] S. Kern, K. Fenner, H. P. Singer, R. P. Schwarzenbach, J. Hollender. Identification of transformation products of organic contaminants in natural waters by computer-aided prediction and high-resolution mass spectrometry. *Environ. Sci. Technol.* **2009**, *43*(18), 7039–7046.

[8] S. D. Richardson. Environmental mass spectrometry: emerging contaminants and current issues. *Anal. Chem.* **2012**, *84*(2), 747–778.

[9] J. M. Wells, S. A. McLuckey. Collision-induced dissociation (CID) of peptides and proteins. In Biological Mass Spectrometry, volume 402 of Methods in Enzymology **2005**, 148–185. Academic Press.

[10] H. Horai, M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, Y. Ojima, K. Tanaka, S. Tanaka, K. Aoshima, Y. Oda, Y. Kakazu, M. Kusano, T. Tohge, F. Matsuda, Y. Sawada, M. Y. Hirai, H. Nakanishi, K. Ikeda, N. Akimoto, T. Maoka, H. Takahashi, T. Ara, N. Sakurai, H. Suzuki, D. Shibata, S. Neumann, T. Iida, K. Tanaka, K. Funatsu, F. Matsuura, T. Soga, R. Taguchi, K. Saito, T. Nishioka. MassBank: a public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* **2010**, *45*(7), 703–714.

[11] D. S. Wishart, D. Tzur, C. Knox, R. Eisner, A. C. Guo, N. Young, D. Cheng, K. Jewell, D. Arndt, S. Sawhney, C. Fung, L. Nikolai, M. Lewis, M.-A. Coutouly, I. Forsythe, P. Tang, S. Shrivastava, K. Jeroncic, P. Stothard, G. Amegbey, D. Block, D. D. Hau, J. Wagner, J. Miniaci, M. Clements, M. Gebremedhin, N. Guo, Y. Zhang, G. E. Duggan, G. D. MacInnis, A. M. Weljie, R. Dowlatabadi, F. Bamforth, D. Clive, R. Greiner, L. Li, T. Marrie, B. D. Sykes, H. J. Vogel, L. Querengesser. HMDB: the human metabolome database. *Nucleic Acids Res.* **2007**, *35*(suppl1), D521–526.

[12] S. Stein. Chemical substructure identification by mass spectral library searching. *J. Am. Soc. Mass Spectrom.* **1995**, *6*, 644–655.

[13] C. A. Smith, G. O. Maille, E. J. Want, C. Qin, S. A. Trauger, T. R. Brandon, D. E. Custodio, R. Abagyan, G. Siuzdak. METLIN: a metabolite mass spectral database. In *Proceedings of the 9th International Congress of Therapeutic Drug Monitoring and Clinical Toxicology* **2005**, *27*, 747–751. Louisville, Kentucky.

[14] E. E. Bolton, Y. Wang, P. A. Thiessen, S. H. Bryant, R. A. Wheeler, D. C. Spellmeyer. Chapter 12 PubChem: integrated platform of small molecules and biological activities. In *Annual Reports in Computational Chemistry* **2008**, *4*, 217–241. Elsevier.

[15] M. Kanehisa, S. Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **2000**, *28*(1), 27–30.

[16] H. E. Pence, A. Williams. ChemSpider: an online chemical information resource. *J. Chem. Educ.* **2010**, *87*(11), 1123–1124.

[17] S. Neumann, S. Böcker. Computational mass spectrometry for metabolomics: identification of metabolites and small molecules. *Anal. Bioanal. Chem.* **2010**, *398*(7-8), 2779–2788.

[18] S. Wolf, S. Schmidt, M. Müller-Hannemann, S. Neumann. In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics* **2010**, *11*(1), 148.

[19] H. Collier. Proceedings of the 2003 International Chemical Information Conference: Nîmes, France, 19–22 October 2003. Infonortics, **2003**.

[20] C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, E. Willighagen. The Chemistry Development Kit (CDK): an open-source Java library for chemo- and bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*(2), 493–500. PMID: 12653513.

[21] P. Willett, J. M. Barnard, G. M. Downs. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*(6), 983–996.

[22] D. Butina. Unsupervised data base clustering based on Daylight's fingerprint and Tanimoto similarity: a fast and automated way to cluster small and large data sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*(4), 747–750.

[23] D. W. Hill, T. M. Kertesz, D. Fontaine, R. Friedman, D. F. Grant. Mass spectral metabonomics beyond elemental formula: chemical database querying by matching experimental with computational fragmentation spectra. *Anal. Chem.* **2008**, *80*(14), 5574–5582.

[24] A. Kerber, M. Meringer, C. Rücker. CASE via MS: ranking structure candidates by mass spectra. *Croatica chemica acta* **2006**, *79*(3), 449–464.

 Copyright © 2013 John Wiley & Sons, Ltd.

[25] E. L. Schymanski, M. Meringer, W. Brack. Matching structures to mass spectra using fragmentation patterns: are the results as good as they look? *Anal. Chem.* **2009**, *81*(9), 3608–3617.
[26] S. Neumann, A. Thum, C. Böttcher. Nearline acquisition and processing of liquid chromatography-tandem mass spectrometry data. *Metabolomics* **2012**, 1–8. 10.1007/s11306-012-0401-0.
[27] L. W. Sumner, A. Amberg, D. Barrett, M. Beale, R. Beger, C. Daykin, T. Fan, O. Fiehn, R. Goodacre, J. L. Griffin, T. Hankemeier, N. Hardy, J. Harnly, R. Higashi, J. Kopka, A. Lane, J. C. Lindon, P. Marriott, A. Nicholls, M. Reily, J. Thaden, M. R. Viant. Proposed minimum reporting standards for chemical analysis. *Metabolomics* **2007**, *3*(3), 211–221.
[28] E. L. Schymanski, C. M. J. Gallampois, M. Krauss, M. Meringer, S. Neumann, T. Schulze, S. Wolf, W. Brack. Consensus structure elucidation combining GC/EI-MS, structure generation, and calculated properties. *Anal. Chem.* **2012**, *84*(7), 3287–3295.
[29] J. Kopka, N. Schauer, S. Krueger, C. Birkemeyer, B. Usadel, E. Bergmuller, P. Dormann, W. Weckwerth, Y. Gibon, M. Stitt, L. Willmitzer, A. R. Fernie, D. Steinhauser. GMD@CSB.DB: the Golm Metabolome Database. *Bioinformatics* **2005**, *21*(8), 1635–1638.
[30] F. Rasche, K. Scheubert, F. Hufsky, T. Zichner, M. Kai, A. Svatoš, S. Böcker. Identifying the unknowns by aligning fragmentation trees. *Anal. Chem.* **2012**, *84*(7), 3417–3426.
[31] F. Hufsky, K. Dührkop, F. Rasche, M. Chimani, S. Böcker. Fast alignment of fragmentation trees. *Bioinformatics* **2012**, *28*(12), i265–i273.
[32] M. Rojas-Cherto, J. E. Peironcely, P. T. Kasper, J. J. J. van der Hooft, R. C. H. de Vos, R. Vreeken, T. Hankemeier, T. Reijmers. Metabolite identification using automated comparison of high-resolution multistage mass spectral trees. *Anal. Chem.* **2012**, *84*(13), 5524–5534.
[33] R. A. Dixon, D. Strack. Phytochemistry meets genome analysis, and beyond. *Phytochemistry* **2003**, *62*(6), 815–816.

Copyright © 2013 John Wiley & Sons, Ltd.