Mass Spectrometry
DOI: 10.5702/massspectrometry.S0037

Original Article

Molecular Formula Identification Using Isotope Pattern Analysis and Calculation of Fragmentation Trees

Kai Dührkop,* Franziska Hufsky, and Sebastian Böcker

Lehrstuhl für Bioinformatik, Friedrich-Schiller-Universität Jena, Ernst-Abbe-Platz 2, 07743 Jena, Germany

We present the results of a fully automated *de novo* approach for identification of molecular formulas in the CASMI 2013 contest. Only results for Category 1 (molecular formula identification) were submitted. Our approach combines isotope pattern analysis and fragmentation pattern analysis and is completely independent from any (spectral and structural) database. We correctly identified the molecular formula for ten out of twelve challenges, being the best automated method competing in this category.

Please cite this article as: K. Dührkop, *et al.*, Molecular Formula Identification Using Isotope Pattern Analysis and Calculation of Fragmentation Trees, Mass Spectrom (Tokyo) 2014; 3(3): S0037; DOI: 10.5702/massspectrometry.S0037

Keywords: CASMI 2013, mass spectrometry, small molecules, molecular formula identification, isotope patterns, fragmentation patterns, fragmentation trees

(Received April 1, 2014; Accepted May 14, 2014)

INTRODUCTION

The identification of small molecules in a high throughput manner plays an important role in many areas of biology and medicine. Mass Spectrometry (MS) is a dominant technology for high-throughput analysis of metabolites and other small molecules[1–3] and is orders of magnitude more sensitive than nuclear magnetic resonance spectroscopy. A major challenge in metabolomics is the low proportion of detected analytes with unambiguously assigned chemical structures. Such unknown metabolites are neither contained in spectral databases nor in molecular structure databases, making their identification and structural elucidation difficult.

Recent approaches[4–7] for small molecule identification seek to replace searching in spectral libraries by searching in molecular structure databases such as KEGG (Kyoto Encyclopedia of Genes and Genomes) and PubChem, which are and will be more comprehensive than spectral libraries. In order to reduce the search space, and with it, false identifications, it is useful to determine the molecular formula of the compound before starting a search. Here, we describe how isotope pattern analysis in combination with calculation of fragmentation trees can be used to identify the molecular formula of a compound. Both methods, and also the combination thereof, are fully automated *de novo* approaches, requiring neither a spectral library nor a molecular structure database. Consequently, we explore a huge search space of molecular formulas including completely unknown molecules.

The Critical Assessment of Small Molecule Identification (CASMI) contest is a benchmark dataset for the identification and structural elucidation of small molecules. Our paper is submitted as part of the CASMI[1] contest 2013. More detailed descriptions of the methods have previously been published elsewhere.[8–14] An updated description of the method and scoring will be published soon.

We restricted ourselves to the first category of the contest, that is the determination of the correct molecular formula. It is commonly believed that structure elucidation of fully unknown molecules is impossible using MS techniques alone. Identification of the molecular formula of the compound can, however, serve as a basis for subsequent structure elucidation. The first category consists of twelve challenges: five compounds measured on a TOF instrument (LC-IT-TOF, Shimadzu), six compounds measured on a Thermo Orbitrap instrument and one compound measured on a FT-ICR instrument (APEX-II FT-ICR, Bruker Daltonics). Each measurement consists of MS and MS^2 spectra.

We submitted molecular formulas for all twelve challenges and obtained the correct formula ranked first for ten challenges. This result was only exceeded by one other team which, however, participated with manually determined molecular formulas. Consequently, we are the best automated method competing in the CASMI 2013 contest.

METHODS

To determine the molecular formula of a measured compound, we use a combined automated analysis of the isotope pattern (see Isotope pattern analysis Section) in the MS spectra and the fragmentation pattern (see Fragmentation pattern analysis Section) in the MS^2 spectra. Our method is a fully automated *de novo* approach and does not perform

*Correspondence to: Kai Dührkop, *Lehrstuhl für Bioinformatik, Friedrich-Schiller-Universität Jena, Ernst-Abbe-Platz 2, 07743 Jena, Germany*, e-mail: kai.duehrkop@uni-jena.de

[1] http://casmi-contest.org/2013/

any database search—neither in spectral databases nor in compound databases.

Isotope pattern analysis has been described in detail by Böcker *et al.*[11]; fragmentation pattern analysis by Böcker and Rasche.[8] The computation was done using the latest version of Sirius[2] which will be released soon and made available on our website[2].

Candidate molecular formula generation

We first decompose the monoisotopic peak using a maximal allowed mass deviation of 20 ppm for negative and 15 ppm for positive TOF data, 5 ppm for Orbitrap and 2 ppm for FT-ICR data.

The list of candidate formulas is then filtered using Senior's rule.[15] We use the six elements most abundant in metabolites carbon (C), hydrogen (H), nitrogen (N), oxygen (O), phosphorus (P), and sulfur (S), as well as the four halogen elements fluorine (F), chlorine (Cl), bromine (Br), and iodine (I). To speed up computations, we set upper bounds for some of the elements: F, I, and S are restricted to six occurrences per molecule; Cl and P are restricted to three occurrences per molecule; Br is restricted to one occurrence per molecule. For molecules with mass greater than *m/z* 900 we use only CHNOPS (upper bounds as given above). Without any knowledge of the ionization of the compound, it is not possible to restrict our search to formulas with integer ring double bond equivalent (RDBE) values.[16] Computing the RDBE as described in refs. 16 and 17, we filter out all candidates with RDBE values lower than −0.5.

We do not use a molecular structure database for compound filtering, that is we explore all possible molecular formulas including potential undiscovered molecules. Sizes of the candidate sets for all compounds are given in Table 2. For comparison, we also point out the much lower number of candidates after filtering using PubChem, which is not employed by our method.

Isotope pattern analysis

For all candidate molecular formulas, we compute a theoretical isotope pattern by convoluting isotope distributions and match it to the measured spectrum. We assume that the measured compounds contain naturally distributed isotopes. Matching of peaks is a trivial task, as we have usually only one peak per nominal mass. If we have more than one peak per nominal mass, we select the peak with the highest intensity and assume the others to be noise. Peaks with relative intensities measured below 1% are omitted.

To score the similarity between measured spectra and theoretical isotope patterns, we calculate likelihoods and posterior probabilities using Bayesian Statistics. For each matched pair of measured and theoretical isotope peaks, we calculate likelihoods from deviation in mass and intensity.

Mass deviations are assumed to be normally distributed as $\mathcal{N}(0, \sigma_{\text{mass}})$. We expect the standard deviation to increase with decreasing peak intensity. The logic behind this assumption is that accurate peak picking becomes increasingly tedious for low intensity peaks. We define a parameter σ_{ppm} to be the expected standard deviation of the instrument in parts per million (ppm). The parameter σ_{mass} of the normal distribution is determined depending on the measured peak intensity f:

$$\sigma_{\text{mass}}(f) = \sigma_{\text{ppm}} \cdot 10^{-6} \cdot (1.5 - 0.5 \cdot f)$$

The probability of observing the mass deviation between the measured peak M and the theoretical peak m is calculated as

$$\mathbb{P}\left(M \middle| m\right) = \operatorname{erfc}\left(\frac{|M-m|}{\sqrt{2}\,\sigma_{\text{mass}}(f)}\right)$$

Calibration issues are very common in mass spectrometry leading to shifts in the measured peak masses. To avoid summing up this systemic errors we use a slightly different approach for calculating posterior probabilities for mass deviations: For each peak but the monoisotopic peak, we subtract the monoisotopic mass from the peak mass and calculate the probability of the peaks using a distribution of mass differences. This distribution of mass differences is again a normal distribution with half the standard deviation $\mathcal{N}(0, \sigma_{\text{mass}}/2)$. The probability is calculated as described above. As expected standard deviations of the instruments we chose $\sigma_{\text{ppm}}=2$ for Orbitrap, $\sigma_{\text{ppm}}=1$ for FT-ICR, and $\sigma_{\text{ppm}}=6$ for TOF (see Table 1).

For intensities, we assume that the logarithmized ratio between measured and theoretical intensity is normally distributed. Peak intensities are normalized such that they sum to 1 and log-ratios between the measured and theoretical intensities are then computed. We define the standard deviation of this ratio depending on the intensity of the peak:

$$\sigma_{\text{int}}(f) = \log\left(\frac{f+\delta_{\text{int}}}{f}\right)$$

where δ_{int} is the expected intensity deviation for this instrument. For Orbitrap and FT-ICR data we set $\delta_{\text{int}}=0.03$, for TOF instruments we chose $\delta_{\text{int}}=0.01$.

Finally, a bias in the intensities of Orbitrap data was detected, which leads to an underestimation of the monoisotopic peak intensity.[11] We compensate for this by adding an offset of 0.01 to the peak intensities of Orbitrap spectra before renormalizing them. The probability of observing a deviation in intensity between a measured intensity f and a theoretical intensity p is calculated as

$$\mathbb{P}\left(f \middle| p\right) = \operatorname{erfc}\left(\frac{\log(f/p)}{\sqrt{2}\,\sigma_{\text{int}}(f)}\right)$$

Table 1. Chosen parameters for the instruments.

Instrument	Ion mode	Allowed mass deviation	σ_{ppm}	δ_{int}	Intensity offset
Orbitrap	Positive Negative	5 ppm	2 ppm	0.03	0.01
APEX-II FT-ICR	Positive Negative	2 ppm	1 ppm	0.03	0
IT-TOF, QTOF	Positive	15 ppm	6 ppm	0.01	0
	Negative	20 ppm	6 ppm	0.01	0

All Orbitrap instruments (Exactive Orbitrap and LTQ Orbitrap Velos) are denoted as Orbitrap and all instruments using a TOF (LCMS-IT-TOF and Xevo QTOF) are simply outlined as TOF. For TOF instruments we additionally distinguish between positive and negative ion mode as in our experience negative spectra often have lower mass accuracy.

[2] http://bio.informatik.uni-jena.de/sirius2/

All parameters σ_{ppm}, δ_{int} and the peak intensity offset compensation were not optimized for the contest data but are rather chosen *ad hoc* by educated guesses. The isotope pattern score is the sum of the logarithmized probabilities of each peak.

Fragmentation pattern analysis

Analysis of the MS^2 spectra is done using fragmentation trees.[8] We follow the evaluation protocol described in ref. 13. For each compound, we merge all MS^2 spectra to a single fragmentation spectrum and remove all peaks with relative intensity lower than 0.5%.

For each candidate formula a fragmentation graph is built, containing vertices for all possible decompositions of the fragment peaks (each molecular formula within the mass accuracy of the instrument), and edges for all possible fragmentation reactions between the peak explanations.

Vertices are scored as log-odds of the likelihood to observe this mass deviation and the likelihood that the peak is noise. As above, the mass deviations are modeled according to a normal distribution using the instrument specific parameter σ_{ppm} as standard deviation.

The noise probability is modeled as a Pareto distribution with parameter $x_{\min}=0.05$ and $k=1$ for all instruments. A heteroatom-to-carbon distribution, gleaned from the KEGG COMPOUND database,[18] adds prior probabilities to the vertices to score formulas according to their "chemical reasonability."

We weight the edges using a scoring scheme loosely based on the logarithmized likelihood that a certain fragmentation reaction occurs. We give positive scores for a few common losses that were learned from the data, and penalize implausible losses and radicals.[13] In addition, edges are penalized in case losses have a high mass using a distribution of loss masses. Details to the learning of common losses and the loss mass distribution will be published elsewhere.

We then compute a subtree of maximum weight, which explains each peak by at most one molecular formula and which assigns a unique fragmentation reaction for the generation of each fragment peak.[14] Note that the method implicitly decides, whether a peak is noise or not. The score of the tree is the sum of its edge weights.

Data processing

The first category of the CASMI challenge consists of MS and MS^2 spectra of twelve compounds measured on various instruments. The compound masses range from *m*/*z* 210 to *m*/*z* 936. The parameters of our method are chosen with regard to the instrument and ionization mode (see Table 1). For challenges 15 and 16 we only use the spectra measured on the Exactive Orbitrap instrument, as the Xevo TQ instrument did not provide high mass accuracy.

For each candidate, we compute a score analyzing its isotope pattern and a score analyzing its fragmentation pattern. We combine the scores by summing them up. Molecular formulas are then ranked by combined score. We automatically transform this output to a new result list which is more suitable for this contest: The best molecular formula candidate receives score 1.0; subsequent formulas receive logarithmic decreasing scores. Formulas with scores differing more than 10% from the highest scoring formula are excluded.

For all compounds we assume $[\text{M+H}]^+$ ionization for spectra measured in positive mode and $[\text{M–H}]^-$ ionization for spectra measured in negative mode. Predicting the ionization type is out of scope for our method, but can be done, for example, using CAMERA.[19]

RESULTS

The first category of the contest consists of twelve challenges: five compounds measured on a TOF instrument (LC-IT-TOF, Shimadzu), six compounds measured on a Thermo Orbitrap instrument and one compound measured on a FT-ICR instrument (APEX-II FT-ICR, Bruker Daltonics). We submitted molecular formulas for all twelve challenges.

We identified the molecular formula of ten out of twelve compounds correctly. Only one other submission to the CASMI contest identified more compounds using manual interpretation of the mass spectra. Our method was the best performing automated approach for molecular formula identification.

For nine compounds we only reported a single molecular formula candidate since the score difference between the first and the second candidate is above 10%.

Table 2. Overview of the results for all twelve challenges.

Challenge	Molecular formula	*m*/*z*	Number of candidates		Combined rank	Rank frag. pattern	Rank isotope pattern
			Decompositions	PubChem			
1	$C_{18}H_{19}NO_4$	314.1364	234	53	1	1	5
2	$C_{14}H_{20}N_2O_3$	265.1524	42	21	1	1	2
3	$C_{13}H_{24}N_4O_4$	301.1843	76	30	1	1	10
4	$C_{15}H_{14}O$	211.1121	13	8	1	1	1
5	$C_{12}H_{18}O_4S_2$	291.0719	116	21	1	2	8
6	$C_{49}H_{93}O_{14}P$	935.6224	224	1	1	2	1
9	$C_9H_{11}Cl_3NO_3PS$	349.9337	888	13	1	1	2
10	$C_{42}H_{71}N_{11}O_{12}$	922.5351	860	5	1	1	3
11	$C_{20}H_{18}O_5$	337.1080	662	86	1	2	7
12	$C_{15}H_{10}O_5$	269.0449	368	55	—	1	80
15	$C_{10}H_5F_{17}O$	462.9973	3326	159	—	—	—
16	$C_{18}H_{20}FN_3O_4$	362.1527	175	26	1	1	27

The combined rank is based on the combined scoring of isotope pattern (rank isotope pattern) and fragmentation pattern analysis (rank frag. pattern) and is the final rank of the compound which was submitted to CASMI. Number of candidates is the size of the candidate set of molecular formulas. Our method uses all decompositions of the monoisotopic mass within the mass deviation of the instrument. Filtering molecular formulas using a molecular structure database such as PubChem would strongly reduce the number of candidates but is not employed by our method.

Our molecular formula identification is a fully automated *de novo* approach, *e.g.* we do not use compound or structural databases to restrict the number of candidate formulas. This strongly increases the complexity of the problem: For most compounds that are part of the CASMI contest, the number of candidate formulas our method has to discriminate is 2–8 fold higher than the number of candidates retrieved from PubChem (see Table 2). For challenge 10 the number of decomposable candidates within a mass range of 5 ppm is even 172 times higher than the number of molecular formulas from PubChem within the same mass range. For challenge 6, there was only a single molecular formula candidate in PubChem which makes this challenge trivial. Nevertheless, our approach identified the correct molecular formula from a much larger set of 224 candidate formulas not filtered using a molecular structure database.

For eight challenges, the fragmentation pattern analysis using fragmentation trees was sufficient to identify the correct molecular formula (see Table 2). Using only isotope pattern analysis resulted in the correct molecular formula identification in only two cases. For challenges 5 and 11, only the combined analysis was able to identify the correct molecular formula. Overall, in this year's contest the fragmentation tree approach yielded better results than the isotope pattern analysis. This differs from the CASMI 2012 contest, where the isotope pattern analysis was able to identify the correct molecular formula for ten out of thirteen challenges.[20)]

The fragmentation pattern analysis performed well for all

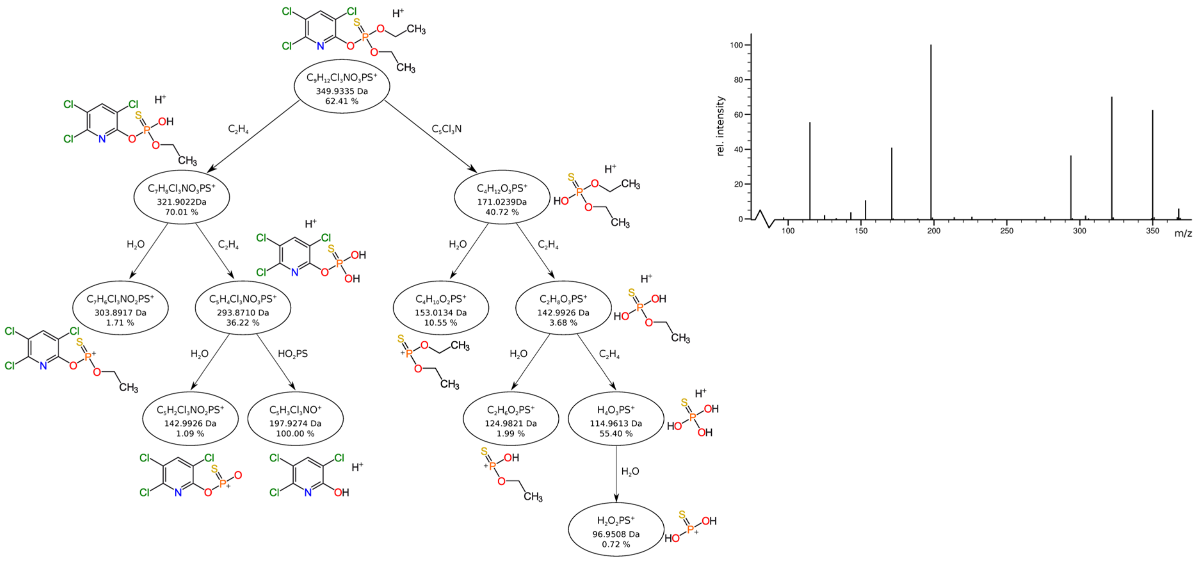

Fig. 1. Challenge 9: Fragmentation tree of Chlorpyrifos ($C_9H_{11}C_{13}NO_3PS$) an insecticide measured for challenge 9. We manually assigned substructures of the compound to the fragment formulas annotated by the vertices of the fragmentation tree. This manual assignment is an example how to use fragmentation trees as basis for subsequent manual structure elucidation of the compound.

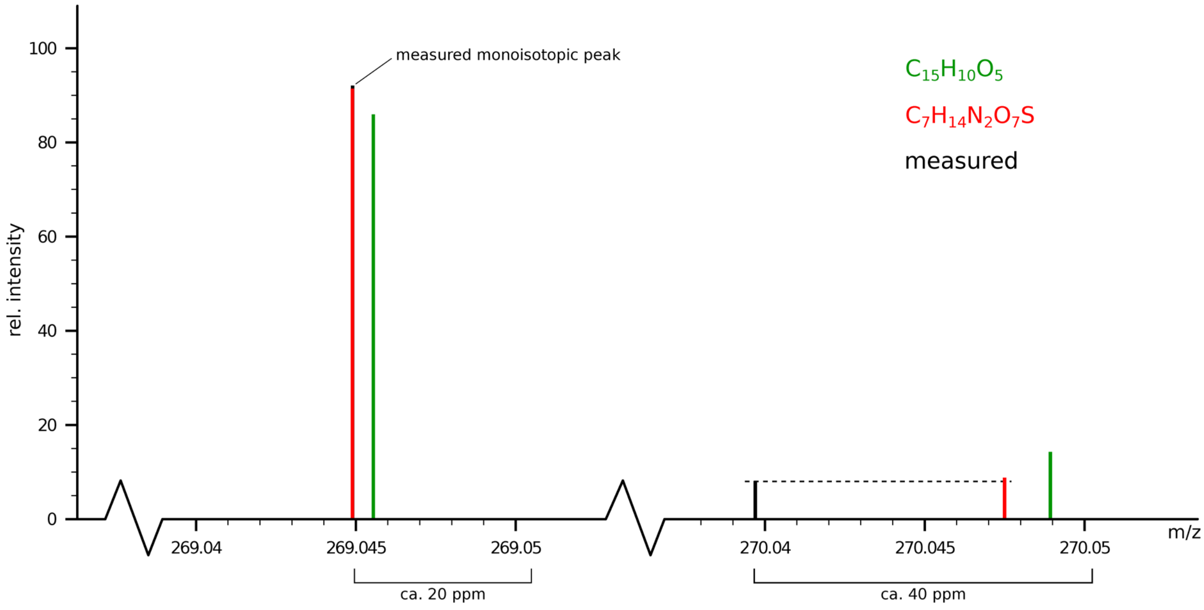

Fig. 2. Challenge 12: Comparison of the measured isotope pattern from challenge 12 (black), the calculated isotope pattern of the correct compound Baicalein $C_{15}H_{10}O_5$ (green), and the calculated isotope pattern of the best matching molecular formula $C_7H_{14}N_2O_7S$ (red). The mass of the monoisotopic peak of the measurement almost completely matches the monoisotopic peak of the best matching molecular formula with a mass deviation of only 0.17 ppm. That is why the red peak almost covers the black peak.

instrument types. The only instrument-specific parameters of the method are the allowed mass deviation and the standard mass deviation σ_{ppm}, which demonstrates the robustness of the fragmentation tree approach against different instruments or experimental setups.

To give an example of a fragmentation pattern analysis using fragmentation trees, we manually assigned substructures of the compound from challenge 9 to the annotated fragments in the fragmentation tree (see Fig. 1). We can explain every fragment in the tree with a substructure of the compound.

For challenges 12 and 15, our method did not identify the correct molecular formula. In the following, we discuss the complications and problems we have run into with these two compounds.

Challenge 12—Baicalein: For this compound, the correct molecular formula was ranked first using fragmentation tree computation. Unfortunately, the isotope pattern measured for this compound looks somewhat erroneous: The calculated isotope pattern of molecular formula $C_7H_{14}N_2O_7S$ matches the measured spectrum much better than the calculated isotope pattern of the correct molecular formula $C_{15}H_{10}O_5$ (see Fig. 2). Although the M+1 peak of the $C_7H_{14}N_2O_7S$ isotope pattern has a high mass deviation from the M+1 peak in the measured pattern, there is a very high similarity of the intensities of both peaks between these two isotope patterns. In contrast, both peaks have much higher intensity deviations between the measured isotope pattern and the isotope pattern of the correct molecular formula $C_{15}H_{10}O_5$. Furthermore, mass deviations for both, the monoisotopic peak and the M+1 peak, are even worse. Thus, combining the results from fragmentation pattern analysis with the results from isotope pattern analysis, our method ranks $C_7H_{14}N_2O_7S$ at first position, although the correct molecular formula $C_{15}H_{10}O_5$ has the best fragmentation tree.

Challenge 15—2-(Perfluorooctyl)ethanol: This compound with molecular formula $C_{10}H_5F_{17}O$ contains 17 fluorine atoms. As mentioned above, we restricted the number of fluorine atoms to a maximum of six atoms per compound. Therefore, the correct molecular formula was not contained in the candidate formula set. However, such perfluorinated compounds are rather uncommon in metabolomics.

All computations were executed on a quad-core Intel i5-3570, 3.4 GHz with 8 GB memory using Java 7. Computations took 37 min in total for all 12 challenges, where most of the running time was spent on challenge 10 (34 min). Note that there are different possibilities to speed up computation, *e.g.*, computing only fragmentation trees for the 20 best isotope patterns. However, fast computation was not relevant for this contest.

CONCLUSION

We presented the results of our combined isotope and fragmentation pattern analysis for molecular formula identification as part of the CASMI 2013 contest. Our method is a fully automated *de novo* approach which requires no user interaction, and accesses no compound databases or spectral libraries. It is the best automated method competing in the first category of the CASMI 2013 contest, identifying the correct molecular formula for ten out of twelve compounds.

The two failed challenges indicate potential directions to further improve our method. On the one hand, we need to predict the alphabet of potential elements for decomposing the compounds mass using machine learning. Training high quality predictors, however, requires an independent training set. On the other hand, we need an improved scoring for combining isotope pattern analysis and fragmentation pattern analysis. Predicting the quality of an isotope pattern or fragmentation pattern may help to decide which of the two analyses is more reliable.

Acknowledgements

We thank Aleš Svatoš from the Max Planck Institute for Chemical Ecology, Jena, for helping us with the manual assignment of fragment structures for visualization of challenge 9.

REFERENCES

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

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

3) R. L. Last, A. D. Jones, Y. Shachar-Hill. Towards the plant metabolome and beyond. *Nat. Rev. Mol. Cell Biol.* 8: 167–174, 2007.

4) F. Allen, R. Greiner, D. Wishart. Competitive fragmentation modeling of ESI-MS/MS spectra for metabolite identification. Preprint, Cornell University Library, 2013. arXiv: 1312.0264.

5) M. Gerlich, S. Neumann. MetFusion: Integration of compound identification strategies. *J. Mass Spectrom.* 48: 291–298, 2013.

6) M. Heinonen, H. Shen, N. Zamboni, J. Rousu. Metabolite identification and molecular fingerprint prediction *via* machine learning. *Bioinformatics* 28(18): 2333–2341, 2012. Proc. of European Conference on Computational Biology (ECCB 2012).

7) L. J. Kangas, T. O. Metz, G. Isaac, B. T. Schrom, B. Ginovska-Pangovska, L. Wang, L. Tan, R. R. Lewis, J. H. Miller. *In silico* identification software (ISIS): A machine learning approach to tandem mass spectral identification of lipids. *Bioinformatics* 28: 1705–1713, 2012.

8) S. Böcker, F. Rasche. Towards *de novo* identification of metabolites by analyzing tandem mass spectra. *Bioinformatics* 24: I49–I55, 2008. Proc. of European Conference on Computational Biology (ECCB 2008).

9) S. Böcker, M. Letzel, Zs. Lipták, A. Pervukhin. Decomposing metabolomic isotope patterns. In Proc. of Workshop on Algorithms in Bioinformatics (WABI 2006), Vol. 4175 of *Lect. Notes Comput. Sci.* pp. 12–23. Springer, Berlin, 2006.

10) S. Böcker, Zs. Lipták, M. Martin, A. Pervukhin, H. Sudek. DECOMP—from interpreting mass spectrometry peaks to solving the Money Changing Problem. *Bioinformatics* 24: 591–593, 2008.

11) S. Böcker, M. Letzel, Zs. Lipták, A. Pervukhin. SIRIUS: Decomposing isotope patterns for metabolite identification. *Bioinformatics* 25: 218–224, 2009.

12) F. Rasche, A. Svatoš, R. K. Maddula, C. Böttcher, S. Böcker. Computing fragmentation trees from tandem mass spectrometry data. *Anal. Chem.* 83: 1243–1251, 2011.

13) F. Rasche, K. Scheubert, F. Hufsky, T. Zichner, M. Kai, A. Svatoš, S. Böcker. Identifying the unknowns by aligning fragmentation trees. *Anal. Chem.* 84: 3417–3426, 2012.

14) I. Rauf, F. Rasche, F. Nicolas, S. Böcker. Finding maximum color-

ful subtrees in practice. *J. Comput. Biol.* 20: 1–11, 2013.

15) J. K. Senior. Partitions and their representative graphs. *Am. J. Math.* 73: 663–689, 1951.

16) T. Kind, O. Fiehn. Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics* 8: 105, 2007.

17) H. E. Dayringer, F. W. McLafferty. Computer-aided interpretation of mass spectra. STIRS prediction of rings-plus-double-bonds values. *Org. Mass Spectrom.* 12: 53–54, 1977.

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

19) C. Kuhl, R. Tautenhahn, C. Böttcher, T. R. Larson, S. Neumann. CAMERA: An integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal. Chem.* 84: 283–289, 2012.

20) K. Dührkop, K. Scheubert, S. Böcker. Molecular formula identification with SIRIUS. *Metabolites* 3: 506–516, 2013.