Mending the pieces: Computational mass spectrometry for small molecule fragmentation

Franziska Hufsky[a,b], Kerstin Scheubert[a], Sebastian Böcker[a]

[a] *Chair of Bioinformatics, Friedrich Schiller University, Ernst-Abbe-Platz 2, Jena, Germany*
[b] *Max Planck Institute for Chemical Ecology, Beutenberg Campus, Jena, Germany*

Abstract

The identification of small molecules from mass spectrometry (MS) data remains a major challenge in the interpretation of MS data. Computational aspects of identifying small molecules range from searching a reference spectral library to the structural elucidation of an unknown. In this review, we concentrate on five important aspects of the computational analysis: First, we briefly review the above-mentioned search in spectral libraries. Second, we investigate the rule-based *in silico* prediction of a fragmentation spectrum from the molecular structure of the compound. Third, we cover combinatorial fragmentation where a molecular structure is used to "explain" an observed spectrum. Fourth, we review the prediction of structural features using Machine Learning techniques. Finally, we look into the concept of fragmentation trees, which allows a true *de novo* analysis of fragmentation data. We find that novel computational methods may overcome the boundaries of spectral libraries, either by searching in the more comprehensive molecular structure databases, or by not requiring any databases at all.

This is a preprint of the article: Franziska Hufsky, Kerstin Scheubert and Sebastian Böcker. Computational mass spectrometry for small molecule fragmentation. Trends Anal Chem, 53:41-48, 2014.

1. Introduction

Metabolomics covers the detection, identification, and quantification of compounds of low molecular weight. Identification of metabolites poses a problem as, unlike proteins, these small molecules are usually not made up of building blocks, and the genomic sequence does not reveal information about their structure. Thus, a huge number of metabolites remain uncharacterized with respect to their structure and function [1].

Mass spectrometry (MS), typically coupled with chromatographic separation techniques, is a key analytical technology for high-throughput analysis of small molecules [1]. It is orders of magnitude more sensitive than nuclear magnetic resonance (NMR). Beyond information on the mass of the molecule, the compound can be fragmented and masses of the fragments recorded, revealing certain information about the structure of a compound. Several analytical techniques have been developed, where tandem mass spectrometry is usually combined with liquid chromatography MS (LC-MS) [2], whereas gas chromatography MS (GC-MS) is coupled with electron impact (EI) fragmentation [3]. Given the huge amount of data produced in a high-throughput experiment, the manual interpretation of fragmentation spectra is time-intensive and often impractical [1]. So, an important aspect of small molecule MS is the automated processing of the resulting fragmentation mass spectra.

Searching in libraries of reference spectra provides the most reliable source of identification. But this is only the case if the library contains a fragmentation spectrum from a reference compound measured on a similar instrument [4]. Unfortunately, spectral libraries are vastly incomplete. Recent approaches tend to replace searching in spectral libraries by searching in the more comprehensive molecular structure databases. Kind and Fiehn [5] give a survey of structure elucidation techniques for small molecules using mass spectrometry, whereas Scheubert et al. [6] review computational methods for this task. In this review, we focus on the five basic approaches of dealing with metabolite fragmentation data, which are: (a) searching spectral libraries; (b) rule-based *in silico* fragmentation spectrum prediction; (c) mapping the fragmentation spectrum to the compound structure (combinatorial fragmentation); (d) predicting structural features and compound classes; and (e) fragmentation trees.

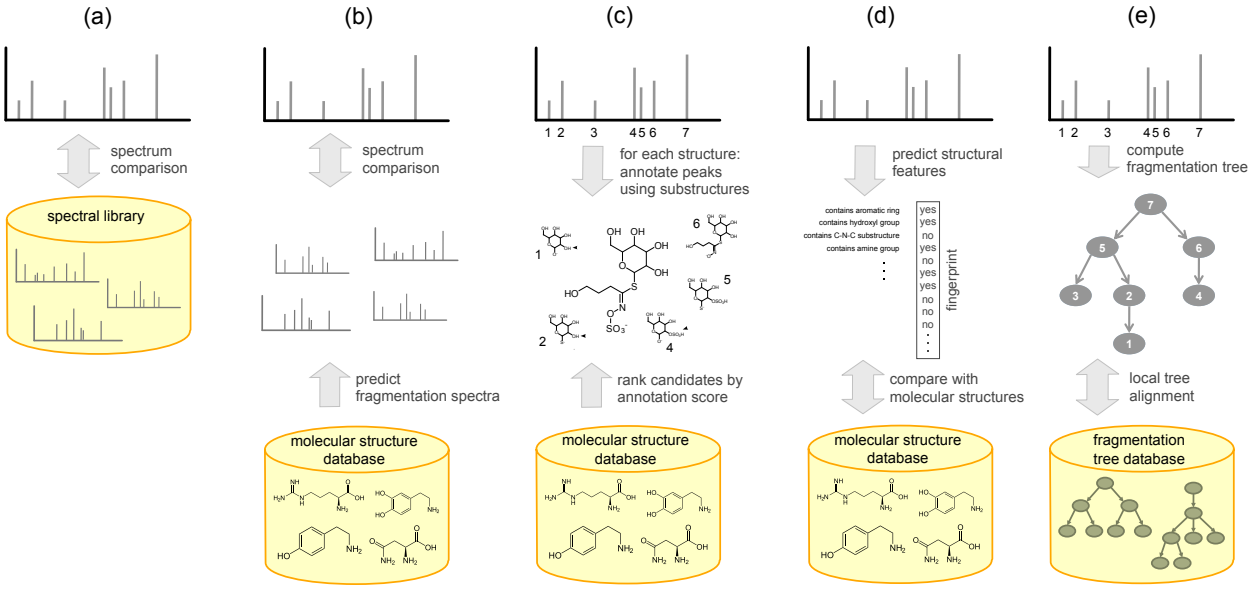

Figure 1: The five basic approaches of dealing with metabolite fragmentation data: (a) searching spectral libraries; (b) fragmentation spectrum prediction; (c) combinatorial fragmentation; (d) predicting structural features; and (e) fragmentation trees.

2. Searching in spectral libraries

Given the fragmentation spectrum of an unknown metabolite, the straightforward approach for identifying the metabolite is looking up its fragmentation spectrum in a spectral library. For GC-MS, huge spectral reference libraries are routinely used; for LC-MS/MS, libraries contain fewer compounds and are limited in their availability. Database search requires a similarity or distance function for spectrum matching. Often, this is done using the "dot product" of the spectra: The spectra are treated as vectors $f = (f_1, \ldots, f_M)$ and $g = (g_1, \ldots, g_M)$, and the scalar product $\langle f, g \rangle = \sum_m f_m g_m$ is computed. This is particularly applied for unit mass accuracy data, where spectra can be directly mapped to vectors. For data with high mass accuracy, we can treat the spectra as continuous functions f, g with scalar product $\int f(m)g(m)\,dm$. Often, not the raw peak shapes are used but instead, peaks are idealized as Gaussian functions. We can also introduce a weight function, to differently weight the terms of the product depending on the mass. Often, not the dot product but the enclosed angle θ or its cosine,

$$\cos\theta = \frac{\langle f, g \rangle}{\sqrt{\langle f, f \rangle}\sqrt{\langle g, g \rangle}}$$

are reported. The spectral dot product is an advanced form of the most fundamental scoring, namely the "peak counting" family of measures that basically count the number of matching peaks. Using the dot product for library searching is among the oldest computational techniques presented in this paper, and has been developed independently of the actual task of searching for small compounds.

In 1994, Stein and Scott [7] evaluated the dot product against several other scoring systems, and found that it performed best among the them. Several authors suggested modifications of the dot product, such as giving different confidence (weight) to different peaks, see [8, 9] for two recent examples. Unfortunately, it appears to be a tough problem to consistently and significantly outperform the basic dot product and its simplest modifications.

The above scoring systems tell us which spectrum in the library best matches our query spectrum, and how to rank the remaining ones. But it cannot tell us whether this is a true or a bogus hit [10]. The reliable identification of a compound depends on the uniqueness of its spectrum. But the presence and intensity of peaks across spectra is highly correlated, as these depend on the non-random distribution of molecular (sub-)structures. For example, benzene and fulvene have similar spectra, and a fulvene query spectrum would match a benzene database spectrum [10]. Hence, structurally related compounds generally have similar mass spectra. This becomes a crucial problem when our

database contains thousands of spectra. Unfortunately, little progress has been made in establishing the confidence of a compound identification using library search [11, 12]. Citing Stein [10], the field of proteomics "has the luxury of being able to estimate 'false discovery rates' because of the ability to construct appropriate libraries of false identifications; such measures of reliability are not available for other classes of compounds." But we can also use the problem of similar spectra to our advantage: Since structurally related compounds generally have similar mass spectra, false positive hits may hint at correct "class identifications" if the true spectrum is not contained in the database [13]. Using fragmentation trees (see Sec. 6) as a detour in library searching allows us to compute such FDRs for small molecule MS.

The computational analysis of EI fragmentation spectra of small molecules via database search is generally simpler than for tandem MS data, as the fragmentation mechanisms are highly reproducible even across instruments, and reference spectra have been collected over many years [10]. On the other hand, LC-MS coupled with tandem MS fragmentation requires less sample preparation, and has other benefits such as the known precursor mass of a compound. Fragmentation by tandem MS (such as collision-induced dissociation, CID) is less reproducible, in particular across different instrument types or even instruments [14]. Only first steps have been taken towards searching tandem MS spectral libraries [15], and these libraries are much smaller than for GC-MS. Attempts have been made to create more reproducible and informative LC-MS fragmentation spectra [14, 16, 17].

For a comprehensive review on the fundamentals and difficulties of mass spectral libraries for compound identification see Stein [10].

3. Rule-based Fragmentation Spectrum Prediction

Spectral libraries are (and will always be) several orders of magnitude smaller than molecular structure databases. For example, PubChem currently contains about 30 million compounds, while even the biggest (commercial) spectral libraries, the National Institute of Standards and Technology (NIST) mass spectral library (version 11) and the Wiley Registry (9th edition) contain mass spectra for only 200 000 and 600 000 compounds, respectively. This gap may be filled by an accurate prediction of fragments (and their abundances) from the molecular structure of a compound. In this way, searching in spectral libraries can be replaced by searching in a database of theoretical mass spectra obtained from molecular structure databases. This trick has been very successfully used in proteomics for many years, as prediction of peptide fragmentation is comparatively easy.

To generate a set of candidate molecules, we can filter a molecular structure database using the molecular mass of the unknown, or even its molecular formula if already known. On the other hand, we can use molecular structure generators to create a "private database", integrating further knowledge such as substructure information.

Given a set of candidate molecular structures, spectra can be predicted by applying fragmentation rules to these structures, see Figure 2. In principle, such rules can be learned from experimental data using data mining; but until recently, experimental data was used solely to predict probabilities and, hence, intensities in the fragmentation spectrum [18, 19]. In practice, these rules are manually curated from mass spectrometry literature. First attempts for generating structural candidates and predicting their fragmentation mass spectra using general models of fragmentation, as well as class-specific fragmentation rules, have been made as part of the DENDRAL project starting in 1965 [20, 21]. However, the DENDRAL project failed in its major objective of automatic structure elucidation by mass spectral data, and research was discontinued [18]. Nowadays, there are three major commercial tools that predict MS fragmentation based on rules: *Mass Frontier* (HighChem, Ltd. Bratislava, Slovakia; versions after 5.0 available from Thermo Scientific, Waltham, USA), *ACD/MS Fragmenter* (Advanced Chemistry Labs, Toronto, Canada), and *MOLGEN-MS* [22, 23].

Rule-based prediction systems were initially developed for the prediction and interpretation of EI fragmentation data. EI spectra are highly reproducible and much is known about the fragmentation. However, this ionization technique can produce complex rearrangements during fragmentation that are relatively hard to predict. For tandem MS, the fragmentation behavior of small molecules under varying fragmentation energies is not completely understood [24]. Nevertheless, there has been a recent tendency to investigate general fragmentation rules of tandem MS and interpret the data with rule-based prediction programs, too.

Hill et al. [25] pioneered the identification of an unknown compound by matching the experimental tandem mass spectrum with predicted spectra of candidate compounds from a molecular structure database. They used *Mass*

Frontier 4 for the simulation of CID spectra, and identified the correct structure in 64 % of 102 cases. For each "unknown compound", they retrieved an average of 273 candidate molecular structures from the PubChem database. For the simulation of EI fragmentation spectra, Schymanski et al. [26] compared the three commercial programs, and indicated that at the time of evaluation, mass spectral fragment prediction for structure elucidation was still far from daily practical use. The authors noted that *ACD Fragmenter* "should be used with caution to assess proposed structures [...] as the ranking results are very close to that of a random number generator." Recently, Kumari et al. [27] implemented a pipeline similar to [25] for EI spectra, using *Mass Frontier* 6 for spectrum prediction and searching PubChem. Integrating other sources of information such as the retention index, they reported the correct structure for 73 % of 29 metabolites within the top 5 hits.

One major disadvantage of rule-based fragmentation prediction is that to achieve high-quality predictions, this approach requires expert-curated "learning" of fragmentation rules. Even the best commercial systems cover only a tiny part of the rules that could be known. Although novel rules are constantly added, all of these rules do not necessarily apply to a newly discovered compound. In proteomics, rule-based systems did not have much impact: There, it was apparent from the beginning that, in view of the huge search space, only methods based on combinatorial optimization can be successful. This situation is somewhat comparable to chess where certain rules are useful (such as opening databases) but ultimately, combinatorial optimization is needed to find the best move.

Instead of curating or learning real fragmentation rules, Kangas et al. [28] used machine learning to find bond cleavage rates for spectral simulation. Different from the rules learned for example during the DENDRAL project, they do not claim these predictions to be true fragmentation rules. Their *In Silico* Identification Software (*ISIS*) currently works only for lipids and does not model rearrangements of atoms and bonds.

4. Combinatorial Fragmentation

In contrast to rule-based fragmentation that simulates the fragmentation spectrum of a given compound, combinatorial fragmentation aims at explaining the peaks in a measured spectrum (see Figure 2). This is based on the assumption that most peaks result from substructures of the compound without major rearrangement. Combinatorial fragmenters use bond disconnection to find these fragments.

Early combinatorial fragmentation methods such as *EPIC* [29] and *FiD* [30] did not aim to find a molecular structure but instead, to explain each peak in a fragmentation spectrum with the most likely substructure of a *known* molecular structure. These early approaches enumerate all fragments by applying all combinations of bond cleavages. The resulting list of potential peaks is then compared to the measured peak list. This exhaustive enumeration is very slow and, hence, cannot be applied for a larger set of candidate molecular structures. Later, Wolf et al. [31] introduced the heuristic method *MetFrag* for this problem. *MetFrag* is much faster than the above-mentioned approaches, an can be applied to a full structure database to find the compound that best explains the spectrum.

At a first step in combinatorial fragmentation, costs for cleaving bonds are assigned. These costs show that some bonds break easier than others and enables the different candidate structures to be ditinguished. A major issue is to choose a suitable cost function. For example, the type of a chemical bond (single, multiple or aromatic bond) [32], standard bond energies (*FiD*) or bond dissociation energies (*MetFrag*) can be used to approximate the cost of cleaving a bond. The next step is to explain each peak in the spectrum with a substructure of minimal cost. Heinonen et al. [30] proposed a Mixed Integer Linear Program (MILP) to solve this problem, but due to the computational complexity of the problem running times explode even for medium-sized molecules. Hill and Mortishire-Smith [29] and Wolf et al. [31] pruned the search space by limiting the number of allowed cleavages. Recently, Gerlich and Neumann [33] introduced *MetFusion*, which combines *MetFrag* with spectral library search in MassBank to improve compound identification. *MetFusion* returns the correct molecular structure with median rank 10 when searching PubChem and using 1062 compounds, strongly improving upon results by solely using *MetFrag*.

One major point of concern regarding combinatorial fragmentation is that fragments resulting from structural rearrangements are not covered by this approach, or only in a limited way such as hydrogen rearrangements. In fact, this is a problem for both combinatorial and rule-based methods [29, 30, 34]. Another problem is to find a good cost function. For example, the scoring of Wolf et al. [31] that considers bond dissociation energies results in a decreasing fragment prediction accuracy when we increase the allowed number of bond cleavages. A possible explanation is that more bond cleavages do not only explain more peaks, but also generate more unlikely fragments. Ridder et al. [32]

report that even a simplistic scoring which basically assigns score 1 to single bonds, 2 to double bonds etc. outperforms the more complicated cost function of Wolf et al. [31]. This underlines that finding a suitable cost function remains an important open problem.

5. Predicting substructures and compound classes

Automated prediction of substructures or compound classes from mass spectral data can be achieved by learning spectral classifiers. The term *compound class* is not exactly defined: Molecules may fall into the same group because they share a common reactive group, a substructure, have a certain chemical property, or a similar biological function. Usually, a mixture of these class types is used in applications.

Given the spectrum of an unknown compound, a classifier gives a response telling us whether a particular substructure (or a more general chemical property) is present or not in the investigated compound (see Figure 3). In its simplest form this is a *yes/no* answer, but alternatively some score or likelihood may be reported. These classifiers have to be trained on a set of mass spectra of known reference compounds, to yield output *yes* for compounds containing the substructure and *no* for all other compounds. Each spectrum is first transformed to a fixed set of numerical *features* characterizing the spectrum. Here, finding "good features" is essential for good performance of the classifier [35]. Feature vectors of the known references, together with the yes/no answers, are fed to the classifier for training. The field of Machine Learning offers a huge number of classification methods for this purpose, such as regression methods, neural networks, support vector machines, and random forests. This process is repeated for every property that we want to predict.

The query spectrum of an unknown compound is transformed to a feature vector using the same transformation as for the training data. The features are given to the substructure classifiers to predict the fingerprint of the molecule, that is, a vector of *yes/no* answers indicating which substructures are present or not. The predicted fingerprints can be directly used to characterize the class and properties of the measured metabolite. Heinonen et al. [36] went one step further, and used the predicted fingerprints to retrieve and score candidate molecules from large molecular databases such as PubChem.

The above idea was pioneered by Venkataraghavan, McLafferty and van Lear in 1969 [37]: The authors presented an automated approach "to identify the general nature of the compound and its functional groups." Several other approaches are built on the same grounds; we will mention just two here: Kwok et al. [38] and Scott and coworkers [39–41] use machine learning to predict the nominal molecular weight of an unknown compound or, more precisely, the mass difference to the detectable fragment peak of highest mass. The Varmuza feature-based classification approach for EI spectra [35] uses a set of mass spectral classifiers to recognize the presence/absence of 70 substructures or general structural properties in the compound. But to reach a precision of at least 90%, each spectrum had to be rejected by 40–70% of the classifiers [35]. Hence, for about half of the compounds we have no information regarding a particular substructure or structural property. To set things straight, we note that the individual evaluations performed by Varmuza and Werther [35] look better than these numbers suggest. In fact, this approach has found wide acceptance in the community as it is part of the NIST software. Much later, Hummel et al. [42] learned decision trees using mass spectral features and retention index information from the Golm Metabolome Database (GMD). Using these trees they predict frequent substructures and subdivide compounds into different compound classes.

Whereas the above methods are targeted towards GC-MS and EI fragmentation, the approach of Heinonen et al. [36] targets LC-MS and CID fragmentation. The characterizing fingerprint of the unknown metabolite is predicted from the mass spectrum using a kernel-based approach and matched against a molecular structure database. Using QqQ MS data and searching the smaller Kyoto Encyclopedia of Genes and Genomes (KEGG) database, they identified the correct molecular structure in about 65% of the cases, from an average of 27 candidates (293 compounds measured on an Orbitrap LTQ instrument).

6. Fragmentation Trees

The most simplistic description of the fragmentation process is that all fragments have to be part of the precursor; to this end, fragment molecular formulas have to be sub-formulas of the precursor formula. This crude view provides no information about the dependency between fragments. Given the molecular structure of the compound and the

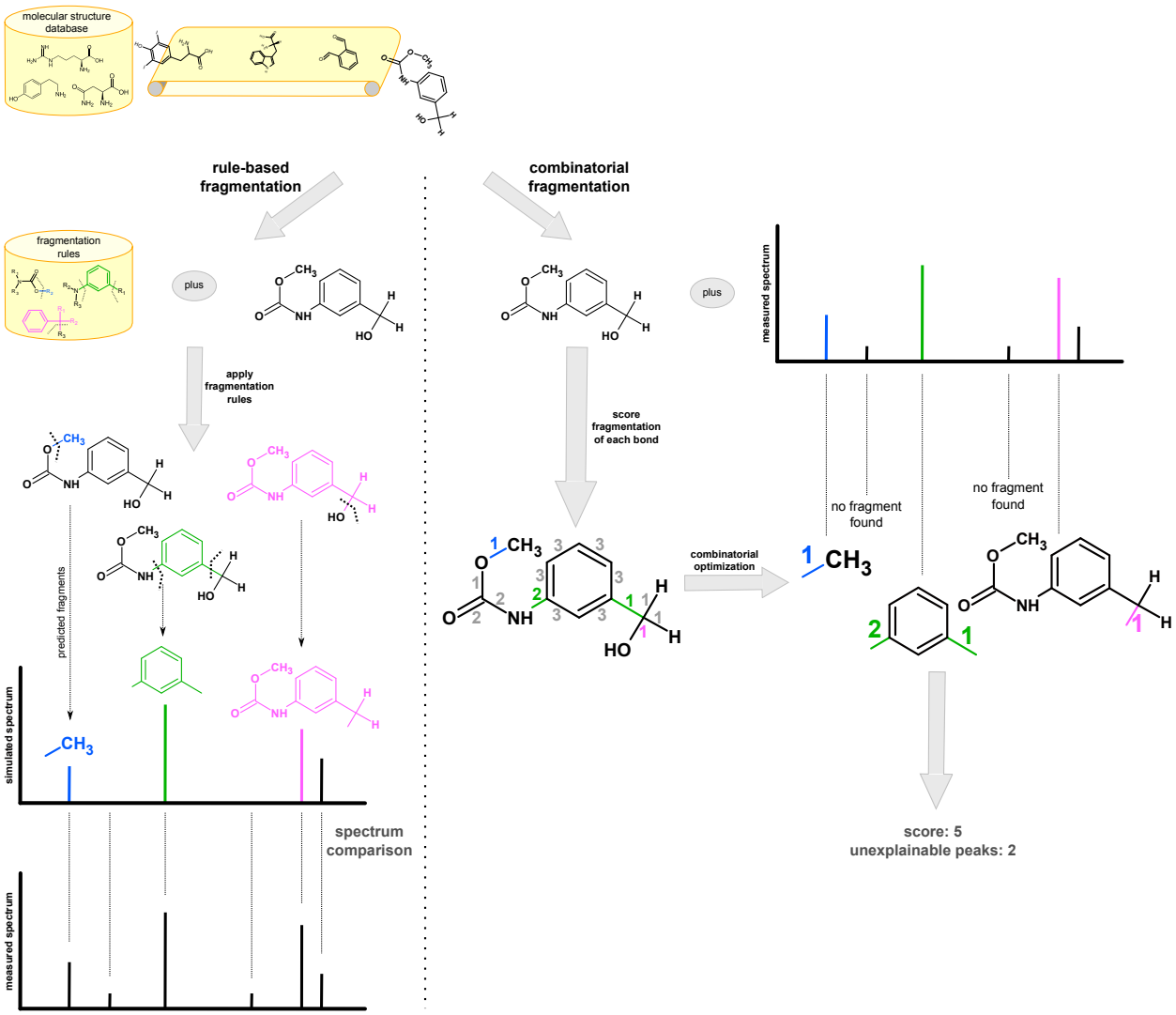

Figure 2: In silico fragmentation: Given a set of known molecular structures, spectra can be predicted by applying fragmentation rules to these structures (left). The simulated spectrum is then compared to the measured spectrum for ranking candidates. In contrast, combinatorial fragmentation (right) attempts to explain the peaks in the measured spectrum. Costs for cleaving are assigned to all bonds in the structure. Each peak in the spectrum is explained with a substructure of minimal cost.

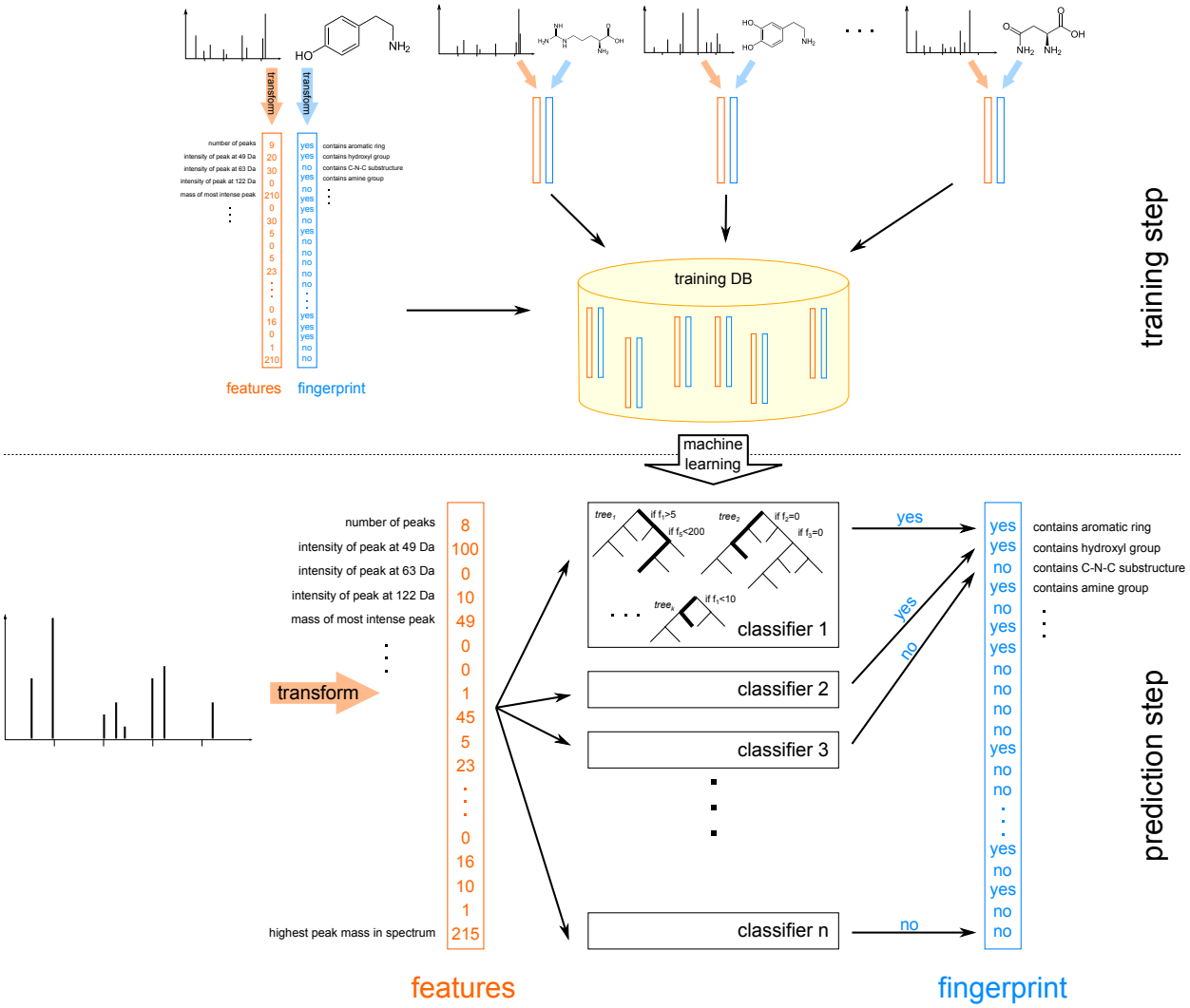

Figure 3: Feature-based substructure prediction. In the training step (top), pairs of mass spectra and corresponding molecular structures are transformed to feature vectors (describing the spectra) and fingerprints (describing the structures). These are used to train the classifiers using, for example, random forests. In the prediction step (bottom), a query spectrum is transformed to a feature vector using the same transformation. These features are passed to the classifiers, which decide whether a substructure is present or not in the investigated compound.

measured fragmentation spectrum, an MS expert can assign peaks to compound fragments and derive a "fragmentation diagram". This is infeasible if we do not know the molecular structure, or when thousands of spectra have to be analyzed.

Fragmentation trees are similar to experts' "fragmentation diagrams" but are extracted directly from the data, without knowledge about a compound's structure. A fragmentation tree consists of nodes, corresponding to the precursor and fragments, and directed edges connecting the nodes. Each node is annotated with the fragment's molecular formula; edges are implicitly annotated with molecular formulas of losses. Given a fragmentation spectrum (possibly merged from several spectra at different energies) of an unknown compound, a fragmentation tree can be computed using combinatorial optimization; besides the fragmentation spectrum, only the molecular formula of the unknown compound is required. For the resulting tree, each node "explains" a peak in the measured fragmentation spectrum: That is, the mass difference between the node's molecular formula and the observed peak mass is below the assumed mass accuracy. To this end, fragmentation trees introduce an "annotation layer" on top of the raw fragmentation data.

The molecular formula of the unknown compound can be determined analyzing its isotope pattern [43] or, again, by computing fragmentation trees for *all possible* molecular formulas of the precursor. In fact, fragmentation trees were initially introduced for this task [44]. In 2011, Rasche et al. [45] found that fragmentation trees are reasonable descriptions of the fragmentation process and, hence, can also be used to derive further information about the unknown compound.

For a given fragmentation spectrum, optimization is used to find the tree that, according to some scoring function, *best* explains the observed spectrum. Unfortunately, this is impeded by the size of the search space: The fragmentation spectrum of a compound may be explained by numerous fragmentation trees. The number of fragmentation trees we have to consider can easily reach 10^{100} even for small compounds [46], a number much larger than the number of atoms in the observable universe. This problem is especially pronounced for compounds above 500 Da, or fragmentation spectra with many peaks; it is less pronounced but, nevertheless, present for high mass accuracy measurements; finally, it is aggravated if we consider elements besides CHNOP. To find the best (and hopefully correct) fragmentation tree, combinatorial optimization is used, as it has been for numerous problems in bioinformatics and cheminformatics. Unfortunately, finding an optimal fragmentation tree is proven to be computationally hard [47]. This severely complicates the design of swift algorithms for the problem. Nevertheless, algorithms have been developed that guarantee to find the optimal solution and are also swift in practice [47]: Usually, the optimal fragmentation tree can be found in a matter of seconds.

To compare two unknown compounds based on their fragmentation spectra, the corresponding fragmentation trees can be aligned [48]. By this, similar fragmentation cascades in the two trees are identified and scored. Even for compounds that cannot be identified, we can derive useful information from this, such as a clustering. We have depicted the workflow from fragmentation spectra to fragmentation tree alignments in Fig. 4. The hierarchical clustering in Fig. 4 was derived solely from the fragmentation data; even if half or more of the compounds had been unknowns, the method would have produced the same clustering. A method similar in spirit [49] first maps fragmentation trees to feature vectors, then compares the feature vectors using standard measures such as Tanimoto similarity. Here, the problem is to find universally applicable feature vectors, as we do not know beforehand what is present in our sample.

Fragmentation trees must not be confused with "spectral trees" for multiple stage mass spectrometry [50], that describe the relationship between the MS^n spectra of a single compound. Similarly, "fragmentation trees" in [51] contain no information regarding the descent of fragment peaks inside a single spectrum.

7. Challenges and future perspectives

For the maturation of metabolomics, the lack of freely available mass spectrometric reference data needs to be filled, as it is required for training and evaluation of novel computational methods. The metabolomics and small-molecule research community should follow examples from other research areas such as proteomics, where computational analysis could mature much faster. The availability of free experimental data was crucial for the development of genomics, and the proteomics community has adopted similar standards with the Amsterdam principles [52]. Regarding free data sharing, metabolomics is still in the dark ages.

After the first approaches for automated analysis of fragmentation spectra 40 years ago, research mostly came to a halt for many years. But over the last five years, several new ideas and approaches were developed that deal

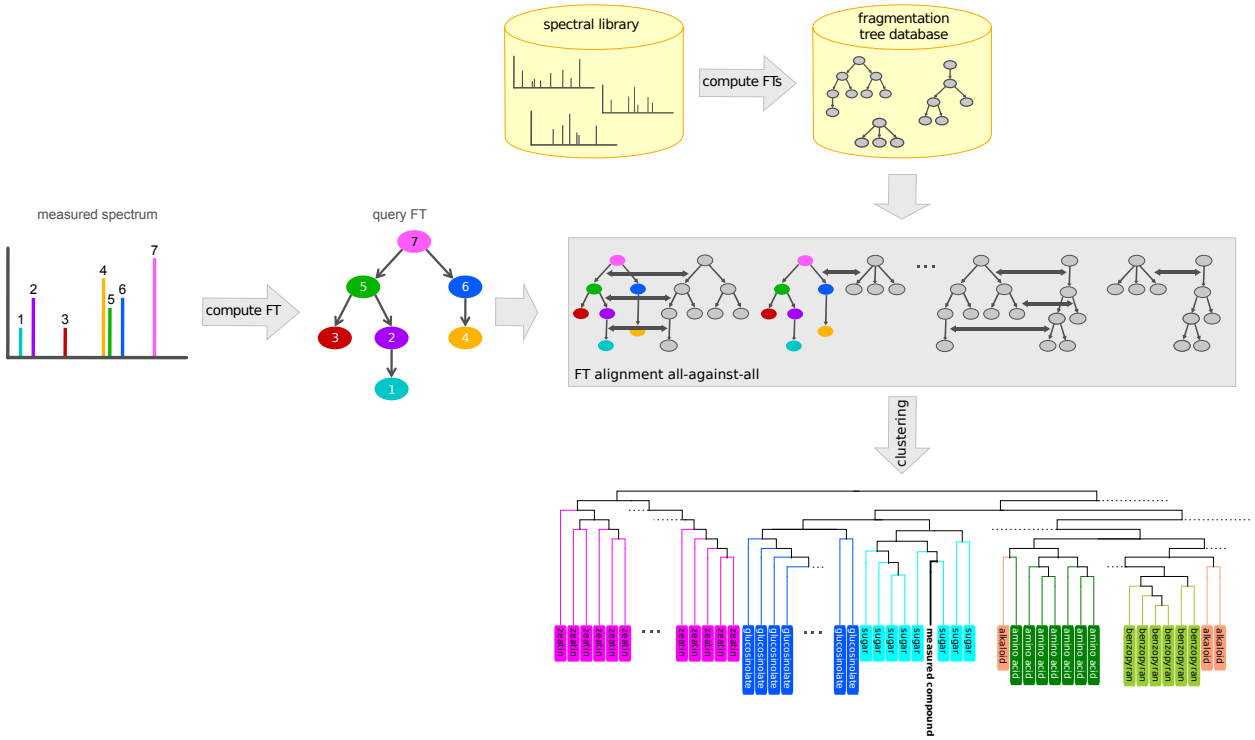

Figure 4: Fragmentation tree alignment for compound classification. A fragmentation tree is computed from the measured spectrum. The tree is aligned to a database of fragmentation trees in an all-against-all manner. The compounds are clustered based on the resulting similarity scores. Similar compounds (belonging to the same compound class) cluster together. The class of the unknown compound can be concluded from the cluster it falls into.

with fragmentation mass spectra not contained in a spectral library. As these approaches are still in their infancy, it is hard to predict their success and further development potential in the near future. However, several of the above-mentioned approaches share the idea to replace searching in spectral libraries by searching in the more comprehensive molecular structure databases, by predicting fragmentation spectra from molecular structures [25, 28], by explaining the experimental spectrum with fragments obtained from molecular structures [31], or by predicting structural features and comparing them with molecular structures [36]. For the computation of fragmentation trees [45, 46, 53], neither spectral libraries nor molecular structure databases are needed. This *de novo* approach targets "true unknowns" and aims at overcoming the limits of the "known universe of organic chemistry".

Acknowledgements

F. Hufsky supported by the International Max Planck Research School Jena.

References

References

[1] G. J. Patti, O. Yanes, G. Siuzdak, Innovation: Metabolomics: The apogee of the omics trilogy, Nat Rev Mol Cell Biol 13 (2012) 263–269.

[2] J. F. Xiao, B. Zhou, H. W. Ressom, Metabolite identification and quantitation in LC-MS/MS-based metabolomics, Trends Analyt Chem 32 (2012) 1–14.

[3] O. Fiehn, Extending the breadth of metabolite profiling by gas chromatography coupled to mass spectrometry, Trends Analyt Chem 27 (2008) 261–269.

[4] S. Neumann, S. Böcker, Computational mass spectrometry for metabolomics – a review, Anal Bioanal Chem 398 (2010) 2779–2788.

[5] T. Kind, O. Fiehn, Advances in structure elucidation of small molecules using mass spectrometry, Bioanal Rev 2 (2010) 23–60.

[6] K. Scheubert, F. Hufsky, S. Böcker, Computational mass spectrometry for small molecules, J Cheminform 5 (2013) 12.

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

[8] I. Koo, X. Zhang, S. Kim, Wavelet- and fourier-transform-based spectrum similarity approaches to compound identification in gas chromatography/mass spectrometry, Anal Chem 83 (2011) 5631–5638.

[9] S. Kim, I. Koo, X. Wei, X. Zhang, A method of finding optimal weight factors for compound identification in gas chromatography-mass spectrometry, Bioinformatics 28 (2012) 1158–1163.

[10] S. E. Stein, Mass spectral reference libraries: An ever-expanding resource for chemical identification, Anal Chem 84 (2012) 7274–7282.

[11] S. E. Stein, Estimating probabilities of correct identification from results of mass spectral library searches, J Am Soc Mass Spectrom 5 (1994) 316–323.

[12] J. Jeong, X. Shi, X. Zhang, S. Kim, C. Shen, An empirical bayes model using a competition score for metabolite identification in gas chromatography mass spectrometry, BMC Bioinformatics 12 (2011) 392.

[13] W. Demuth, M. Karlovits, K. Varmuza, Spectral similarity versus structural similarity: Mass spectrometry, Anal Chim Acta 516 (2004) 75–85.

[14] E. Champarnaud, C. Hopley, Evaluation of the comparability of spectra generated using a tuning point protocol on twelve electrospray ionisation tandem-in-space mass spectrometers, Rapid Commun Mass Spectrom 25 (2011) 1001–1007.

[15] H. Oberacher, M. Pavlic, K. Libiseller, B. Schubert, M. Sulyok, R. Schuhmacher, E. Csaszar, H. C. Köfeler, 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 (2009) 494–502.

[16] P. Goodley, Maximizing MS/MS Fragmentation in the Ion Trap Using CID Voltage Ramping, Technical Report 5988-0704EN, Agilent Technologies, 2007.

[17] C. Hopley, T. Bristow, A. Lubben, A. Simpson, E. Bull, K. Klagkou, J. Herniman, J. Langley, Towards a universal product ion mass spectral library – reproducibility of product ion spectra across eleven different mass spectrometers, Rapid Commun Mass Spectrom 22 (2008) 1779–1786.

[18] J. Gasteiger, W. Hanebeck, K.-P. Schulz, Prediction of mass spectra from structural information, J Chem Inf Comput Sci 32 (1992) 264–271.

[19] B. Fan, H. Chen, M. Petitjean, A. Panaye, J.-P. Doucet, H. Xia, S. Yuan, New strategy of mass spectrum simulation based on reduced and concentrated knowledge databases, Spectrosc Lett 38 (2005) 145–170.

[20] J. Lederberg, Topological mapping of organic molecules, Proc Natl Acad Sci U S A 53 (1965) 134–139.

[21] J. Lederberg, How DENDRAL was conceived and born, in: ACM Conf. on the History of Medical Informatics, History of Medical Informatics archive, 1987, pp. 5–19.

[22] A. Kerber, R. Laue, M. Meringer, K. Varmuza, MOLGEN-MS: Evaluation of low resolution electron impact mass spectra with MS classification and exhaustive structure generation, Adv Mass Spectrom 15 (2001) 939–940.

[23] A. Kerber, M. Meringer, C. Rücker, CASE via MS: Ranking structure candidates by mass spectra, Croat Chem Acta 79 (2006) 449–464.

[24] E. Werner, J.-F. Heilier, C. Ducruix, E. Ezan, C. Junot, J.-C. Tabet, Mass spectrometry for the identification of the discriminating signals from metabolomics: Current status and future trends, J Chromatogr B 871 (2008) 143–163.

[25] 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 80 (2008) 5574–5582.

[26] 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 81 (2009) 3608–3617.
[27] S. Kumari, D. Stevens, T. Kind, C. Denkert, O. Fiehn, Applying in-silico retention index and mass spectra matching for identification of unknown metabolites in accurate mass GC-TOF mass spectrometry, Anal Chem 83 (2011) 5895–5902.
[28] 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 (2012) 1705–1713.
[29] A. W. Hill, R. J. Mortishire-Smith, Automated assignment of high-resolution collisionally activated dissociation mass spectra using a systematic bond disconnection approach, Rapid Commun Mass Spectrom 19 (2005) 3111–3118.
[30] M. Heinonen, A. Rantanen, T. Mielikäinen, J. Kokkonen, J. Kiuru, R. A. Ketola, J. Rousu, FiD: A software for ab initio structural identification of product ions from tandem mass spectrometric data, Rapid Commun Mass Spectrom 22 (2008) 3043–3052.
[31] S. Wolf, S. Schmidt, M. Müller-Hannemann, S. Neumann, In silico fragmentation for computer assisted identification of metabolite mass spectra, BMC Bioinformatics 11 (2010) 148.
[32] L. Ridder, J. J. J. van der Hooft, S. Verhoeven, R. C. H. de Vos, R. van Schaik, J. Vervoort, Substructure-based annotation of high-resolution multistage MS^n spectral trees, Rapid Commun Mass Spectrom 26 (2012) 2461–2471.
[33] M. Gerlich, S. Neumann, MetFusion: integration of compound identification strategies, J Mass Spectrom 48 (2013) 291–298.
[34] D. L. Sweeney, Small molecules as mathematical partitions, Anal Chem 75 (2003) 5362–5373.
[35] K. Varmuza, W. Werther, Mass spectral classifiers for supporting systematic structure elucidation, J Chem Inf Comp Sci 36 (1996) 323–333.
[36] M. Heinonen, H. Shen, N. Zamboni, J. Rousu, Metabolite identification and molecular fingerprint prediction via machine learning, Bioinformatics 28 (2012) 2333–2341. Proc. of *European Conference on Computational Biology* (ECCB 2012).
[37] R. Venkataraghavan, F. W. McLafferty, G. E. van Lear, Computer-aided interpretation of mass spectra, Org Mass Spectrom 2 (1969) 1–15.
[38] K.-S. Kwok, R. Venkataraghavan, F. W. McLafferty, Computer-aided interpretation of mass spectra. III. Self-training interpretive and retrieval system, J Am Chem Soc 95 (1973) 4185–4194.
[39] D. R. Scott, Rapid and accurate method for estimating molecular weights of organic compounds from low resolution mass spectra, Chemometr Intell Lab 16 (1992) 193–202.
[40] D. R. Scott, Pattern recognition/expert system for mass spectra of volatile toxic and other organic compounds, Anal Chim Acta 265 (1992) 43–54.
[41] D. R. Scott, A. Levitsky, S. E. Stein, Large scale evaluation of a pattern recognition/expert system for mass spectral molecular weight estimation, Anal Chim Acta 278 (1993) 137–147.
[42] J. Hummel, N. Strehmel, J. Selbig, D. Walther, J. Kopka, Decision tree supported substructure prediction of metabolites from GC-MS profiles, Metabolomics 6 (2010) 322–333.
[43] S. Böcker, M. Letzel, Zs. Lipták, A. Pervukhin, SIRIUS: Decomposing isotope patterns for metabolite identification, Bioinformatics 25 (2009) 218–224.
[44] S. Böcker, F. Rasche, Towards de novo identification of metabolites by analyzing tandem mass spectra, Bioinformatics 24 (2008) I49–I55. Proc. of *European Conference on Computational Biology* (ECCB 2008).
[45] F. Rasche, A. Svatoš, R. K. Maddula, C. Böttcher, S. Böcker, Computing fragmentation trees from tandem mass spectrometry data, Anal Chem 83 (2011) 1243–1251.
[46] F. Hufsky, M. Rempt, F. Rasche, G. Pohnert, S. Böcker, De novo analysis of electron impact mass spectra using fragmentation trees, Anal Chim Acta 739 (2012) 67–76.
[47] I. Rauf, F. Rasche, F. Nicolas, S. Böcker, Finding maximum colorful subtrees in practice, in: Proc. of Research in Computational Molecular Biology (RECOMB 2012), volume 7262 of *Lect Notes Comput Sci*, Springer, Berlin, 2012, pp. 213–223.
[48] F. Rasche, K. Scheubert, F. Hufsky, T. Zichner, M. Kai, A. Svatoš, S. Böcker, Identifying the unknowns by aligning fragmentation trees, Anal Chem 84 (2012) 3417–3426.
[49] M. Rojas-Chertó, J. E. Peironcely, P. T. Kasper, J. J. J. van der Hooft, R. C. H. de Vos, R. J. Vreeken, T. Hankemeier, T. H. Reijmers, Metabolite identification using automated comparison of high-resolution multistage mass spectral trees, Anal Chem 84 (2012) 5524–5534.
[50] M. T. Sheldon, R. Mistrik, T. R. Croley, Determination of ion structures in structurally related compounds using precursor ion fingerprinting, J Am Soc Mass Spectrom 20 (2009) 370–376.
[51] M. Rojas-Chertó, P. T. Kasper, E. L. Willighagen, R. J. Vreeken, T. Hankemeier, T. H. Reijmers, Elemental composition determination based on MS^n, Bioinformatics 27 (2011) 2376–2383.
[52] H. Rodriguez, M. Snyder, M. Uhlén, P. Andrews, R. Beavis, C. Borchers, R. J. Chalkley, S. Y. Cho, K. Cottingham, M. Dunn, T. Dylag, R. Edgar, P. Hare, A. J. R. Heck, R. F. Hirsch, K. Kennedy, P. Kolar, H.-J. Kraus, P. Mallick, A. Nesvizhskii, P. Ping, F. Pontén, L. Yang, J. R. Yates, S. E. Stein, H. Hermjakob, C. R. Kinsinger, R. Apweiler, Recommendations from the 2008 International Summit on Proteomics Data Release and Sharing Policy: the Amsterdam principles, J Proteome Res 8 (2009) 3689–3692.
[53] K. Scheubert, F. Hufsky, F. Rasche, S. Böcker, Computing fragmentation trees from metabolite multiple mass spectrometry data, J Comput Biol 18 (2011) 1383–1397.