COMPUTATIONAL METHODS FOR ANNOTATION OF METABOLITES IN UNTARGETED LC-MS ANALYSIS

Lappeenranta-Lahti University of Technology LUT

Bachelor's Program in Computational Engineering, Bachelor's Thesis

2023

Atte Lihtamo

Examiner: Associate Professor Xin Liu

Supervisor: M.Sc. (Tech.) Anton Klåvus

ABSTRACT

Lappeenranta-Lahti University of Technology LUT
School of Engineering Sciences
Computational Engineering

Atte Lihtamo

Computational methods for annotation of metabolites in untargeted LC-MS analysis

Bachelor's thesis
2023
28 pages, 6 figures, 3 tables, 2 appendices
Examiner: Associate Professor Xin Liu
Supervisor: M.Sc. (Tech.) Anton Klåvus
Keywords: metabolomics, computational annotation, lc-ms analysis

Metabolomics has become a viable approach to elucidate alterations in the small-molecule chemicals of various biological samples under changing environmental conditions, as well as to assess difference in nutritional values between processed and unprocessed foods, just to name a few. With the recent development of computational methods, there is a great potential for advancing this field. This thesis, conducted on behalf of Afekta Technologies Ltd, aims to investigate the applicability of existing computational techniques in Afekta's workflow, rather than to develop a novel approach. Specifically, two tools, namely MetFrag and Sirius, were evaluated using Afekta's in-house data set. While MetFrag exhibited limited accuracy in predicting test data, Sirius demonstrated promising performance and could potentially streamline the annotation of metabolites. It is worth mentioning that despite the progress in this area, achieving a fully automated annotation process remains a challenging task. The development of complex machine learning methods would be necessary to achieve this goal.

TIIVISTELMÄ

Lappeenrannan-Lahden teknillinen yliopisto LUT
School of Engineering Sciences
Laskennallinen tekniikka

Atte Lihtamo

Laskennalliset menetelmät metaboliittien tunnistamiseen kohdentamattomassa LC-MS analyysissä

Kandidaatintyö
2023
28 sivua, 6 kuvaa, 3 taulukkoa, 2 liitettä
Tarkastaja: Associate Professor Xin Liu
Ohjaaja: M.Sc. (Tech.) Anton Klåvus
Hakusanat: metabolomiikka, laskennallinen tunnistus, lc-ms analyysi
Keywords: metabolomics, computational annotation, lc-ms analysis

Metabolomiikan avulla voidaan selvittää esimerkiksi millaisia muutoksia minkä tahansa orgaanisen aineen aineenvaihdunnassa havaitaan ympäristön vaihdellessa, tai selvittää tarkemmin ruoka-aineiden terveellisyyteen liittyviä ominaisuuksia sekä eroja esimerkiksi prosessoidussa ja prosessoimattomassa ruoassa. Tämä kandidaatintyö on tehty Afekta Technologies Oy:lle. Työssä käydään läpi metaboliittien tunnistuksessa käytettävien laskennallisten menetelmien nykytilanne ja testataan voiko niitä käyttää helpottamaan asiantuntijan manuaalista tunnistustyötä. Työssä ei kehitetä uutta työkalua metaboliittien tunnistukseen. Testattavaksi päätyi kaksi työkalua, MetFrag ja Sirius, joita testattiin Afektan datan avulla. MetFrag ei kyennyt tekemään tarpeeksi luotettavia tunnistuksia testidatasta, kun taas Siriuksen tunnistuskyky vaikuttaa lupaavalta ja siitä olisi hyötyä tunnistusprosessin apuna. Huolimatta viime vuosien aikana tapahtuneesta edistyksestä alalla, täysin automatisoituun ratkaisuun ei ole vielä olemassa sopivaa työkalua. Tällaisen menetelmän kehittäminen vaatisi monimutkaisia koneoppimismalleja.

LIST OF SYMBOLS AND ABBREVIATIONS

.jar	Java archive file
.ms	a custom spectra file for Sirius
Afekta	Afekta Technologies Ltd.
BDE	bond dissociation energy
CASMI	Critical Assessment of Small Molecules Identification
DNA	deoxyribonucleic acid
ESI	electrospray ionization
GC	gas chromatography
HILIC	hydrophilic interaction liquid chromatography
LC	liquid chromatography
LC-MS	liquid chromatography-mass spectrometry
log P	a logarithmic ratio between two solvents
m/z	mass-to-charge ratio
MKL	multiple kernel learning
MS	mass spectrometry
MS/MS	tandem mass spectrometry
NMR	nuclear magnetic resonance
RNA	ribonucleic acid
RP	reverse-phase chromatography
RT	retention time
SMILES	simplified molecular input line entry specification
SVM	support vector machine

CONTENTS

1 Introduction

1.1 Background

In the late 20th century, the development of research methods enabled scientists to delve deeper into the functions of living organisms by studying increasingly smaller molecules. Metabolomics is the newest field of study among 'omics' disciplines, which are presented in Figure 1. Whereas genomics studies DNA, transcriptomics studies RNA, and proteomics studies the proteome, Weckwerth and Kahl [31] state metabolomics aims to identify the metabolite complement from any biological sample. Oliver et al. [21] set the foundation to the term "metabolomics" in 1998, and since then numerous papers have been published to explore the possibilities of metabolomics research [1]. Metabolomics has since become an indispensable tool for answering various biological questions, such as discovering drug actions, exploring biomarkers in medical science, and better modeling of biological systems like plants or animals.

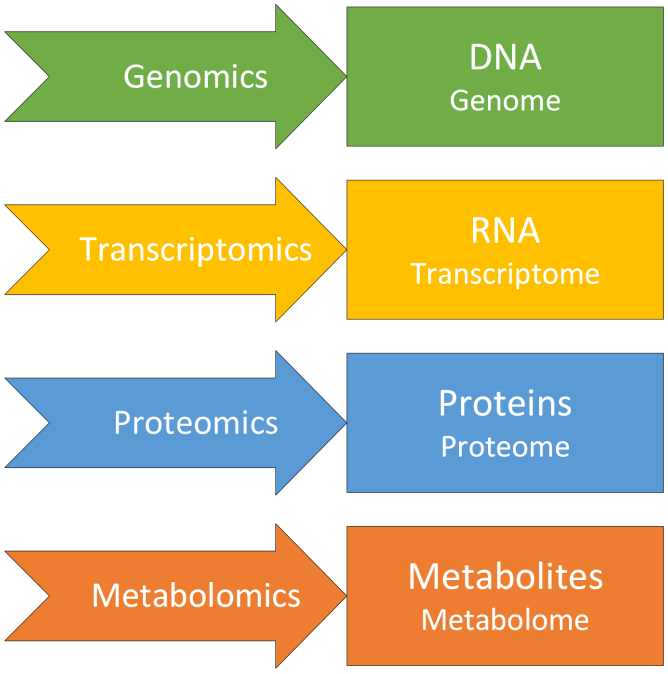

Figure 1. A diagram depicting the omics technologies and their objects of study. Adapted from: [19]

Metabolomics continues to face numerous challenges, particularly on the data analysis front [5]. To start off, the chemical diversity of metabolome is vast and a part of it remains yet unknown. Therefore, the untargeted approach also requires the identification

of unknown metabolites, which is a complex task. These in addition to multiple varying workflows lead to complex data and a non-standardized manual annotation of the results. The lack of proper database standards and current fractioned databases are also problems in metabolomics research. According to Johnson and Gonzalez [12], the annotation of metabolites is the most challenging and bottlenecked aspect of the study. It takes special knowledge and significant time from biochemists, making the automation of annotation a necessity for future advancements in metabolomics.

1.2 Goals and delimitation

This thesis was conducted for Afekta Technologies Ltd. (Afekta), one of the world's leading enterprises in metabolomics, specializing in plant-based foods and phytochemicals. They offer a complete metabolomics analysis service using state-of-the-art technologies to analyze almost any biological sample. Afekta utilizes liquid chromatography-mass spectrometry (LC-MS), which is one of the primary methods for metabolomics research and therefore this thesis focuses primarily on LC-MS, although other methods are mentioned.

Recent advancements in the development and assessment of automatic annotation tools for metabolites, such as those evaluated in the Critical Assessment of Small Molecules Identification (CASMI) contest and the study by Blaženović et al. [3], provide a promising foundation. However, due to variability in analysis methods, further testing is required for finding the optimal tool for a specific use. As such, this thesis has two main objectives. Firstly, to assess the current status of computational tools for metabolite identification in LC-MS analysis. Secondly, to identify a tool to use in practice for Afekta's data processing pipeline. This is achieved by testing selected tools using an in-house data set of pure compounds.

To ensure a manageable workload, the thesis has several limitations. Testing is restricted to existing tools, and the development of a novel tool is outside the scope of the thesis. Testing tools with unknown features would require expertise in biochemistry, thus this kind of comprehensive testing is not part of this thesis. Furthermore, automatisation or the implementation of a tool into the data processing pipeline are also excluded.

1.3 Structure of the thesis

The thesis is structured in a way that first introduces the reader to the topic, and sets out the goals and limitations. The subsequent three sections, in addition to the Introduction section, cover the major areas of focus. Section 2 describes metabolomics as research field and and provides an overview of the most commonly used methods in it, with a particular emphasis on LC-MS and the metabolite identification process. Section 3 provides comprehensive technical specifications employed in the testing phase, including details of the data and software used. Section 4 reports on the results obtained from the testing phase. To conclude, a critical analysis of the results and a concluding discussion are presented at the end of the thesis.

2 The principles of metabolomics research

2.1 LC-MS metabolomics

Small molecules, called metabolites, are formed in cells, tissues and body fluids as a product of metabolism [5]. According to Johnson et al. [13], metabolites are responsible for vital cellular functions, such as producing and storing the energy, transducting signals and programmed death of cells. Metabolomics is the study of metabolites, and it can be used for characterizing them and the metabolic pathways in the sample. It is not as straightforward as analyzing the genome or the proteome since both environment and microflora affect and change the metabolome [18, 31]. Lutz et al. [18] describe three major research methods used for analyzing samples, those being nuclear magnetic resonance (NMR) and either liquid chromatography (LC) or gas chromatography (GC) combined with mass spectrometry (MS). An overview of a LC-MS instrument is presented in Figure 2.

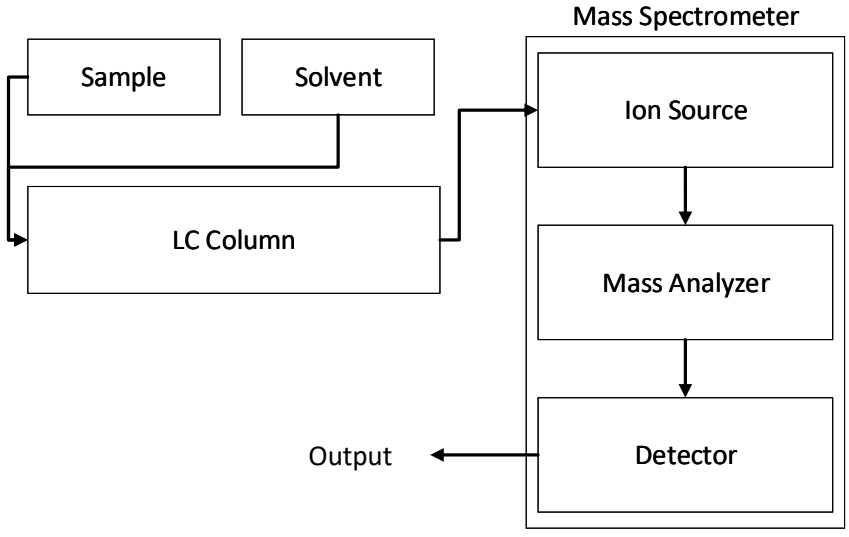

Figure 2. A diagram of the LC-MS instrument. Adapted from: [19]

Chromatography is a fundamental analytical technique that relies on the selective inter-

action of molecules with stationary and mobile phases. In LC, molecules are separated based on their polarity. The principle of LC is similar to the phenomenon observed when dissolved molecules in coffee stain clothes, as they interact with the solid molecules in the surface. A polar stationary phase is used to retain polar molecules that are dissolved in a mobile liquid solvent. By gradually increasing the polarity of the mobile solvent, more polar molecules are dissolved, resulting in different retention time (RT)s in the column. This enables separation and identification of the molecules of interest.[2]

The two most popular columns are hydrophilic interaction liquid chromatography (HILIC) and reverse-phase chromatography (RP). In RP, the mobile phase which becomes increasingly hydrophobic, passes through a stationary phase that is hydrophobic in nature. This mode is termed "reverse" as compared to the normal phase, wherein molecules that are more hydrophobic exhibit a greater RT. On the other hand, HILIC is a more complicated version of the normal phase, which uses relatively hydrophobic bulk eluent as a solvent. The order of dissolving molecules is more or less opposite of the RP mode, thus these two modes complement each other well.[2, 9]

The second part of the instrument, MS, contains three major parts: an ion source, a mass analyzer and a detector. Initially, metabolites are subjected to ionization in the ion source, a process that converts the molecules into ions that can be subsequently detected by the mass analyzer. The mass-to-charge ratio (m/z) of the metabolites is then identified. Typically, in LC-MS, the ionization process involves both positive and negative ions to capture a broader range of chemical properties of the metabolites. An electrospray ionization (ESI) is widely considered the most suitable ion source for metabolomics, due to its ability to generate a higher quantity of intact molecular ions that can facilitate identification.[32]

2.2 Untargeted LC-MS analysis

LC-MS can be used in two complementary approaches for metabolic analysis. The targeted analysis identifies and quantifies a few selected metabolites with prior knowledge of the metabolome. In contrast, an untargeted approach is used to obtain the broadest range of metabolites without any prior knowledge. In the untargeted approach, the goal is to study all metabolites that can be measured by the LC-MS instrument. The output data from the instrument contains molecular features that represent metabolites. These features need to be separately identified to associate them with their metabolite origin. Identifications can be divided into four sections based on if they are expected and able to be identified. Figure 3 depicts the classification of the identifications.[13, 25]

LC-MS has multiple advantages in untargeted analysis [18]. Firstly, it is a product-sensitive instrument that can detect thousands of so-called features although not all features can be identified as compounds. Secondly, a tandem mass spectrometry (MS/MS) spectra can be produced to help with the identification as described in Section 2.3. Thirdly, samples do not require any chemical preparation which could produce unwanted side products. LC-MS has also some disadvantages compared to other methods. It is neither quantitative nor reproducible and takes the most time to perform out of all methods. The LC-MS analysis is rather semi-quantitative than quantitative, meaning that multiple ions may correspond to different fragments from the same molecule, making full quantification difficult [14]. Analysis can not be reproduced accurately, since the environment has been changed and RTs of the same sample between batches may differ. When comparing data from two different analyses, a batch correction should be applied to align RTs.

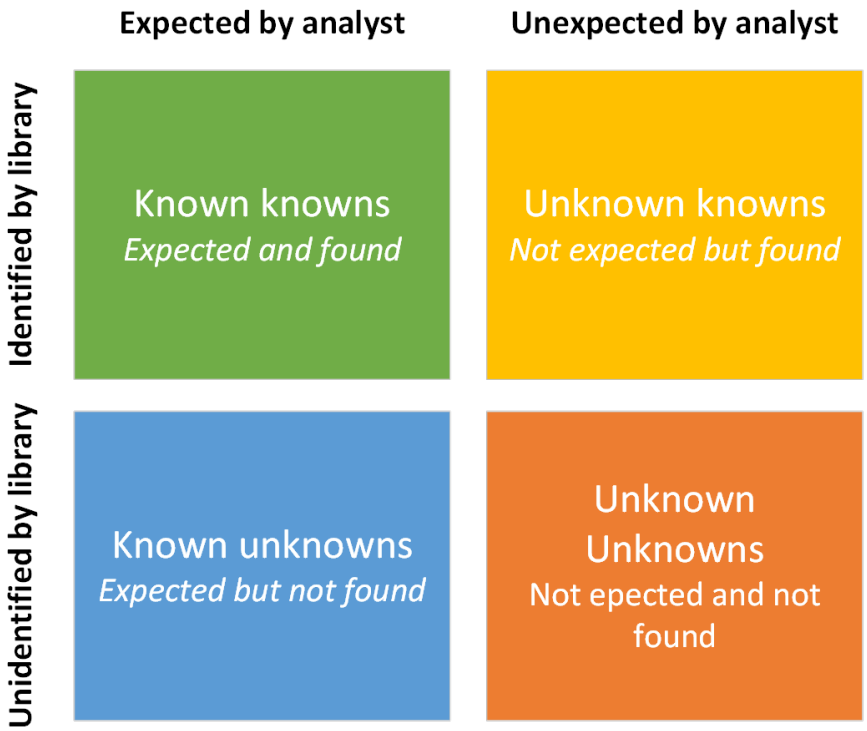

Figure 3. "Rumsfeld Quadrants" showing the intersection of yes/no answers for whether analysts expect a compound to be identified in the sample (prior probability) and whether it was identified in a library search. Adapted from: [29]

After analyzing the sample, the raw data needs pre-processing before annotation can be performed. Zhou et al. [32] stated the following procedure, which is shown in Figure 4. First, outliers are screened in case some acquired peaks deviate significantly from the majority. Improved data quality can be achieved by removing noise while preserving

peaks in various ways, such as performing a baseline correction with Savitzky–Golay filter to remove the unwanted effect of baseline shift caused by increasing RTs. This procedure is called filtering. Next, raw continuous data is transformed into centroided discrete data where one peak represents each ion. This so called peak detection has two major advantages, it further reduces the noise and also reduces data dimensions without significant information loss. Even if samples would be identical replicates of each other, some variation in RTs always exists. Therefore, the peaks are aligned after the peak detection so a comparison can be done across all samples. Finally, data is normalized to remove unwanted system bias. Relative abundances are calculated by adjusting other ions' intensities based on their ratios to internal standards added to the peaks. Multiple tools are available for this purpose, as most of the instrument manufacturers provide their own software [32]. In addition, there are free tools available for pre-processing LC-MS data such as MS-DIAL [30], Mzmine [22] and XCMS [28].

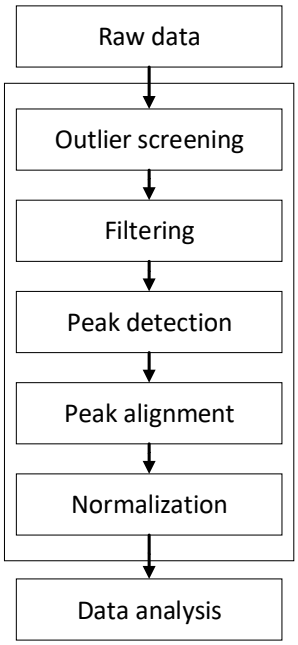

Figure 4. Pre-processing procedure of raw LC-MS data

2.3 Annotation of metabolites

To obtain biological information from LC-MS data, metabolites need to be annotated. According to Chaleckis et al. [5], basic criteria for annotation are m/z and RT, which can be retrieved from MS and LC, respectively. This information is not usually enough for identification due to the millions of theoretical structures one chemical formula can have. To elucidate chemical formula and structure, MS/MS spectra can be used. It is acquired by colliding neutral molecules with the target molecule [11]. These fragments are then subjected to another round of MS, which yields a spectrum of fragments' masses and intensities. The annotation can then be done by matching the spectra to mass spectral libraries. For example, the world's largest open chemistry database, PubChem, contains over 113 million entries of unique chemical structures [20].

Table 2 outlines the different levels of the annotation set by the Metabolomics Standards Initiative with a new level 0 annotation [3]. Level 0 is the complete identification which requires a complete 3d structure and stereochemistry information. Level 1 requires two of RT, m/z and MS/MS to define the 2d structure confidently. Level 2 differs from level 1 only by the confidence level of the structure. In Level 3, only one piece of information is available, leading to possible structural matches. Level 4 annotation is unknown but can be distinguished or quantified. It's important to note that level 0 is considered as identification, while subsequent levels are referred to as annotation. In practice, annotations up to level 2 have a MS/MS spectrum available and thus can be annotated.

Table 2. New confidence levels of compound annotations, as discussed by the Compound Identification work group of the Metabolomics Society at the 2017 annual meeting of the Metabolomics Society (Brisbane, Australia). The new edition refers to the 'Level 0' annotation; other levels remain as discussed by the Metabolomics Standards Initiative. Adapted from: [3]

Confidence Level	Description	Minimum Data Requirements
Level 0	An unambigous 3D structure: Isolated, pure compound, including full stereochemistry	Following natural product guidelines, determination of 3D structure
Level 1	A confident 2D structure: Uses reference standard match or full 2D structure elucidation	At least two orthogonal techniques defining 2D structure confidently, such as MS/MS and RT or m/z
Level 2	A probable 2D structure: Matched to literature data or databases by diagnostic evidence	At least two orthogonal pieces of information, including evidence that excludes all other candidates
Level 3	A possible 2D structure or class: Most likely structure, isomers possible, substance class or substructure match	One or several candidates possible, requires at least one piece of information supporting the proposed candidate
Level 4	An unknown feature of interest	Presence in a sample

Computational methods for MS/MS spectra generation can be divided into four categories. Quantum chemistry-based approaches generate mass spectra solely from physical and chemical information and first principles. The second category involves heuristic methods that are only suitable for predicting compound classes with reoccurring and predictive fragmentation patterns, rather than distinct structures. Reaction-based approaches utilizes reactions found in literature and is based on observed reaction pathways. The final category are machine learning-based methods, which require diverse training sets to achieve decent results. However, it is important to note that computational methods are not yet able to generate accurate MS/MS spectra for all metabolites, and experimental validation is still necessary for confident identification.[3]

3 Computational annotation tools in practice

3.1 MetFrag

MetFrag, an *in silico* molecular fragmenter, was first founded in 2010 and it has since gone through multiple algorithmic and scoring refinements [24]. The core scoring function is based on weighted characteristics of the spectra in addition to recently added annotation algorithm which is based on Bayesian modeling. Initially, candidate molecules sourced from the molecular structure database are constrained to matching m/z within a user-defined confidence interval. Optionally, if the molecular formula is known, it can be included to further restrict candidate molecules in MetFrag. In the absence of molecular formula, MetFrag does not predict it and sources the candidates based solely on m/z. According to Ruttkies et al. [24], the limitation of the search by molecular formula might lead to wrong results.

After sourcing the candidates, the final score for each candidate is calculated as

$$S_{C_{final}} = \omega_i \cdot S_{C_i}, \tag{1}$$

where ω refers to weights and S_C refers to scoring terms of different characteristics of input spectra. The original term is fragmentation score, which is defined by

$$S_{C_{Frag}} = \sum_{p=P} \frac{RelMass_p^{\alpha} \cdot RelInt_p^{\beta}}{(\sum_{b=B_f} BDE_b)^{\gamma}}, \tag{2}$$

where the relative mass $RelMass$ and the intensity $RelInt$ are from each peak p matching the generated fragment, in addition to the sum of bond dissociation energy (BDE) of all cleaved bonds b of the fragment assigned to p. The weights $\alpha = 1.84$, $\beta = 0.59$ and $\gamma = 0.47$ were optimized in the process. [24]

In the development process, several scoring terms have been added to the scoring function. The RT score is calculated via correlation of a logarithmic ratio between two solvents (log P) values of candidate structures and input RTs. For candidates, log P can be calculated separately or sourced from PubChem. For RT, log P is estimated by linear model $logP_U = a \cdot RT_U + b$ trained on log P values in the training set. Thus, the score is denoted by

$$S_{C_{RT}} = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-(logP_U - logP_C)^2/2\sigma^2}. \tag{3}$$

Let n be the number of matching peaks in inclusion or exclusion lists. Then, the match score is calculated by

$$N_{C_{Match}} = \sum_{i=1}^{n} M_i, \tag{4}$$

where $M_i \in \{0, 1\}$. Then, inclusion and exclusion scores, which allow molecular substructure restrictions, are calculated by

$$S_{C_{Incl}} = \frac{N_{C_{Match}}}{max_{C' \in L}(N_{C'_{Match}})}, \tag{5}$$

$$S_{C_{Excl}} = \frac{n - N_{C_{Match}}}{max_{C' \in L}(n - N_{C'_{Match}})}, \tag{6}$$

where $max_{C' \in L}(N_{C'_{Match}})$ is the maximal value of $N_{C_{Match}}$ of the candidates L. Additional scoring functions including suspect lists and reference information are available, but not accounted in this study. Users can also provide custom scoring functions, but they were not provided in this study.[24]

The latest addition is a Bayesian model used to estimate the likelihood of a certain fragmentation structure in a given peak. The candidate score $S_{C_{RawPeak}}$ is calculated from the resulting probability distribution. This score indicates how well a candidate can explain the m/z peaks from the training data. It denotes to

$$S_{C_{RawPeak}} = \frac{1}{-logP_L}, \tag{7}$$

where $logP_L$ is log P based on the probability distribution from the Bayesian model. Finally, the raw score is normalized giving

$$S_{C_{Peak}} = \frac{S_{C_{RawPeak}}}{max_{C' \in C}(S_{C'_{RawPeak}})}. \tag{8}$$

The same analogy can be applied to remaining losses when removing similarities between the candidate and the unknown structure, denoted by

$$S_{C_{Loss}} = \frac{S_{C_{RawLoss}}}{max_{C' \in C}(S_{C'_{RawLoss}})}. \tag{9}$$

With these new scoring functions, the performance of MetFrag is drastically improved, especially on negative modes.[23]

3.2 Sirius

Dührkop et al. [7] developed a full software, called Sirius, for pre-processing and analyzing full LC-MS data sets. It combines multiple separately developed tools, resulting in a full LC-MS analysis software which is also capable of predicting the chemical formulas and compound classes of metabolites, as well as annotate them. Although Sirius is freely available for academic use, a separate license is required for commercial purposes. Permission was obtained from the authors for testing the software.

CSI:FingerID is the module responsible for annotation predictions, which is a multiple kernel learning (MKL) method combined with support vector machine (SVM)s which was developed by Dührkop et al. [8]. Kernel functions are essential tools that enable efficient representation of linear patterns in high-dimensional spaces, allowing the kernel methods to be applied to a broad range of data types and learning tasks. The fundamental idea behind kernel methods is to embed data into a feature space that is appropriate for a particular learning task. The embedded data is then analyzed using linear algebra, geometry, and statistics to uncover patterns.[10] Sirius uses CSI:FingerID as a web service, which means calculations do not require resources from a local computer.

During the training process, the computation of kernel weights was carried out using a MKL algorithm. This approach offers an advantage over utilizing a single kernel by combining multiple kernels to improve the accuracy of the model. The algorithm takes a set of kernels $K = \{K_k | K_k \in \mathbb{R}^{n \times n}, k = 1, ..., q\}$ computed from n data points of training data as input and produces a set of m chemical fingerprint properties denoted by $Y \in \{-1, +1\}^{n \times m}$. These chemical fingerprints are unique identifiers indicating the presence of a particular molecule. Following this, SVMs were trained for each property in the generated fingerprints. These SVMs can then be utilized to predict the presence of a particular property in the fingerprint of an unknown compound.[27]

Metabolites' molecular fragmentations are not well known since they are able to fragment at almost any chemical bond. To estimate these fragmentations, CSI:FingerID uses so-called fragmentation trees. These trees can be described as directed graphs, where the root node is the candidate molecular formula and the other nodes are the formulas of sub-molecules of the parent molecule. If the correct molecular formula is not given in advance, Sirius generates the candidate formulas using the Senior's rule [26] with the most common elements in metabolites [6]. Then, weights for each edge are calculated by the logarithmic likelihood that a certain fragmentation reaction occurs given the observed MS/MS spectrum. This likelihood is calculated with the chemical properties of the

molecular formula in addition to the intensity and mass deviation of the fragmented peak as well as the loss mass as proposed in Kind and Fiehn [15]. Finally, the maximum weight subtree is searched from the graph giving the score of the candidate molecular formula from the sum of edge weights.[4, 27]

These fragmentation trees, which can be perceived as a representation of the original spectra, are then used in kernel functions. CSI:FingerID uses a total of 12 kernels. Shen et al. [27] defined 11 of these kernels based on different loss-, node- and path-based characteristics of fragmentation trees, such as the presence and intensities of the trees. In addition to these, a probability product kernel defined by Kondor and Jebara [16] is used. Let $T_x = (V_x, E_x)$ be a fragmentation tree for each given spectrum x, and let r and v be the root node and child node, respectively. For all nodes $v \in V_x$, the intensity of the corresponding peak is denoted by $i_x(v)$, and for each edge $e \in E_x$ the intensity of the terminal node is denoted by $i_x(e)$. For path-based kernels, let $D[u, v]$ be a dot product table between two trees. All used kernels are listed below.

- LB: Loss binary, presence of a loss l in a fragmentation tree T_x, denoted by $K^{LB}(x) = 1_{l \in \lambda(E_x)}$
- LC: Loss count, number of losses in a fragmentation tree, denoted by $K^{LC}(x) = N_x(l)$
- LI: Loss intensity, takes the average intensity of the terminal nodes i and a loss in a fragmentation tree into account, denoted by $K^{LI}(x) = \frac{1}{N_x(l)} \sum_{e \in E_x \lambda(e) = l} i_x(e)$
- RLB: Root loss binary, presence of root loss $\xi = r - v$ in a fragmentation tree, denoted by $K^{RLB}(x) = 1_{l \in \xi_x}$
- RLI: Root loss intensity, takes the intensity of a root loss into account, denoted by $K^{RLI}(x) = i_x(r - l)$ if $r - l \in V_x$, zero otherwise
- NB: Nodes binary, the presence of node in a fragmentation tree denoted by $K^{NB}(x) = 1_{v \in V_x}$
- NI: Nodes intensity, takes the intensity of the node into account if it is present in a fragmentation tree, denoted by $K^{NI}(x) = i_x(v)$ for $v \in V_x$, zero otherwise
- CPC: Common path counting, uses the amount of identical sequence of losses for the subtrees between two fragmentation trees, denoted by $D[u, v] = \sum(1 + D[a, b])$
- CP2: Common paths of length 2, uses the amount of identical sequence of losses for the subtrees between two fragmentation trees whose depths are 2, denoted by $D[u, v] = \sum(1 + D[a, b])$

- CPK: Common paths with PPK score instead of sum of the paths, details omitted
- CSC: Common subtree counting, uses the number of common subtrees between two fragmentation trees $D[u, v] = \sum(2 + D[a, b]) - 1$
- PPK: Probability product kernel, defined as a kernel between propability distributions of the two spectra $K^{PPK}(p, p') = \int_{\mathbb{R}} p(x)p'(x)dx$

When predicting the unknown fingerprints, first the similarities of unknown spectra against all kernels are computed. Then using the trained SVMs, prediction is done resulting in candidate fingerprints, which can then be scored by matching them to molecular structure databases. First, the candidate structures are sourced from the molecular structure database using m/z and given or predicted molecular formula as well as the computed chemical fingerprint. These structures are then converted into binary fingerprints. Then, Bayesian tree network developed by Ludwig et al. [17] is used to determine the similarities between fingerprints.[8]

3.3 Data

The dataset used in the tests contains in-house standards which are separately analyzed solutions of pure compounds, i.e. the metabolites in the data are known with absolute certainty. Using unique labels for annotated standards is crucial for matching the predictions with the correct annotations. However, metabolite naming is not always straightforward, as many metabolites have multiple commonly used names. Therefore, it is important to use specific and unique labels for each metabolite to avoid confusion. In this study, PubChem IDs were chosen as labels for the data because PubChem is the world's largest open chemistry database and it was used for retrieving candidate metabolites for annotation. Using PubChem IDs as labels ensured that each metabolite had a unique identifier, which made it easier to match the predictions with the correct annotations.

Data includes metadata, such as RT, m/z and the molecular formula from each metabolite as well as normalized and centroided MS/MS spectra. Figure 5 presents a centroided MS/MS spectrum which is the core of the annotation process. The spectrum is interpreted as relative intensities in all of the tools used, therefore normalized intensities can be used. These spectra can be represented as two-dimensional data containing the relative intensity of each m/z.

Originally, the data contains 240 metabolites. After removing the duplicate data entries,

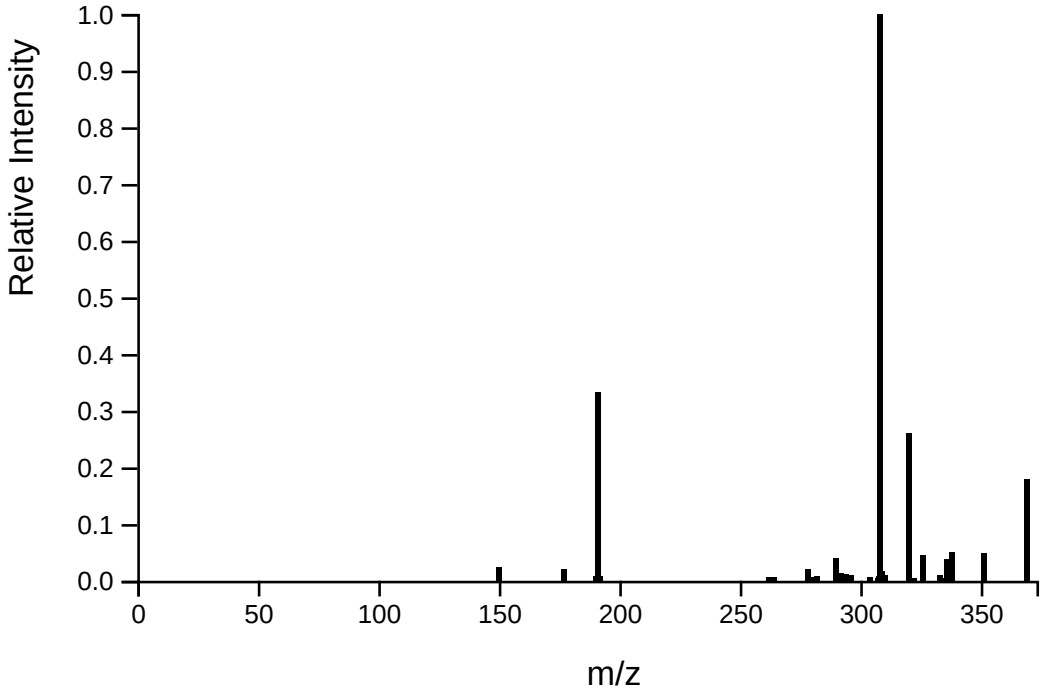

Figure 5. Example MS/MS spectrum generated in Sirius

sourcing PubChem IDs, and removing rows missing IDs and simplified molecular input line entry specification (SMILES) strings, which are needed in the tools used, 155 unique compounds were left to be used in the testing set. MetFrag has a defined set of precursor ions available, which means it can not predict all compounds. From the testing set, Metfrag was able to process 137 compounds. Whereas Sirius was able to process all 155 compounds.

3.4 Testing

The conducted study aimed to evaluate the performance of the chosen software tools. Both of the chosen software can be used in a targeted approach, where the molecular formula is given as an input. Both of them are also capable of a more non-targeted approach, where the molecular formula is not given in advance. The first round of testing was a targeted search where the molecular formulas were provided. The goal was to see how well the software can predict the correct metabolite based on given molecular formula. The second round of testing was otherwise similar to the first test case, but molecular formulas were not given as input. It was conducted to check how well the software performs in a

more complex case. This allowed for the assessment of the tools' performance in a more complex case, simulating a real project where molecular formulas are not always known in advance.

Both tools generate multiple potential predictions, ranked from most to least likely, for each unknown feature based on the scores of the tools. Each prediction also contains PubChem IDs for predicted annotation. Results were evaluated based on the rank of the correct annotation among the predicted annotations. The annotation was interpreted as correct if the labeled ID was found from the predicted IDs of potential prediction. In addition, Sirius returns a confidence score that represents the tool's certainty about its annotations. This score could potentially reduce manual workload by correlating negatively with the ranks.

MetFrag requires a separate spectra file and a parameter file as input, which contains the necessary metadata for the tool to function. The spectra file is a basic tab-separated file containing intensities of fragmentation peaks for each m/z. Examples for both files are in Appendix 1. MetFrag was run as a Java archive file (.jar) in the command line, and commands were executed with a custom parallel wrapper function in R to achieve faster execution. Sirius, on the other hand, uses a custom spectra file for Sirius (.ms) which combines the metadata and the spectrum into a single file. These files were generated with custom R function. While Sirius has a command line tool, these tests were performed using the graphical user interface, since it was a faster approach to a small study like this one. An example of the .ms file can be found in Appendix 2.

4 Results

All results have been summarized in the summary table, which was done based on how many correct annotations are found within top $n \in \{1, 5, 10, 20\}$. The summary table from the first experiment is presented in Table 3. The results show that Sirius outperformed MetFrag in all levels. Considering the top 20 candidates, Sirius found the correct match from 44 % of metabolites, whereas MetFrag was able to only correctly match 10 % of metabolites. Sirius achieves the same results in the top 10 as in the top 20. The number of correct annotations for MetFrag is reduced by 47 % when considering the top 10 predictions. The correct annotation can still be found from 41 % in the top 5 of Sirius predictions. MetFrag achieves only 1.3 % on this level. Sirius can correctly annotate 32 out of 155 metabolites whereas MetFrag is not capable of correct annotations.

Table 3. Summary of the results from the targeted experiment.

	Sirius		**MetFrag**	
Rank	Match %	Total matches	Match %	Total matches
1	20.6 %	32	0.0 %	0
5	41.3 %	64	1.3 %	2
10	43.9 %	68	5.2 %	8
20	43.9 %	68	9.7 %	15

The second experiment results are summarized in table 4. When chemical formulas were not provided in advance, MetFrag was not able to make any correct annotations, even outside the considered top candidates. On the other hand, Sirius achieved almost identical results when chemical formulas were not provided. Similar to the first experiment, Sirius did not find anything new in the top 20 candidates. However, for the top 10 and top 5, the number of matches decreased by 10% and 11%, respectively. Despite this decrease, the number of correct annotations remained the same as in the first experiment.

Table 4. Summary of the results from the non-targeted experiment.

	Sirius		**MetFrag**	
Rank	Match %	Total matches	Match %	Total matches
1	20.6 %	32	0.0 %	0
5	36.8 %	57	0.0 %	0
10	39.4 %	61	0.0 %	0
20	39.4 %	61	0.0 %	0

Regarding the scoring performance, MetFrag and Sirius showed different results. Met-

frag overall score is a value $\in$ [0,1], and in both experiments, it had a large number of confident predictions (score > 0.9) for each compound. On the other hand, Sirius confidently predicted the right chemical formula and often one annotation was separate from the others in all scoring functions.

To assess the effectiveness of the confidence score of Sirius, Figure 6 shows the distribution of confidence scores against the corresponding ranks. Pearson correlation $\rho = 0.23$ was calculated to determine if correct annotation can be found from top ranks based on the confidence score of Sirius. One outlier, where the rank was 24, was removed from the correlation test.

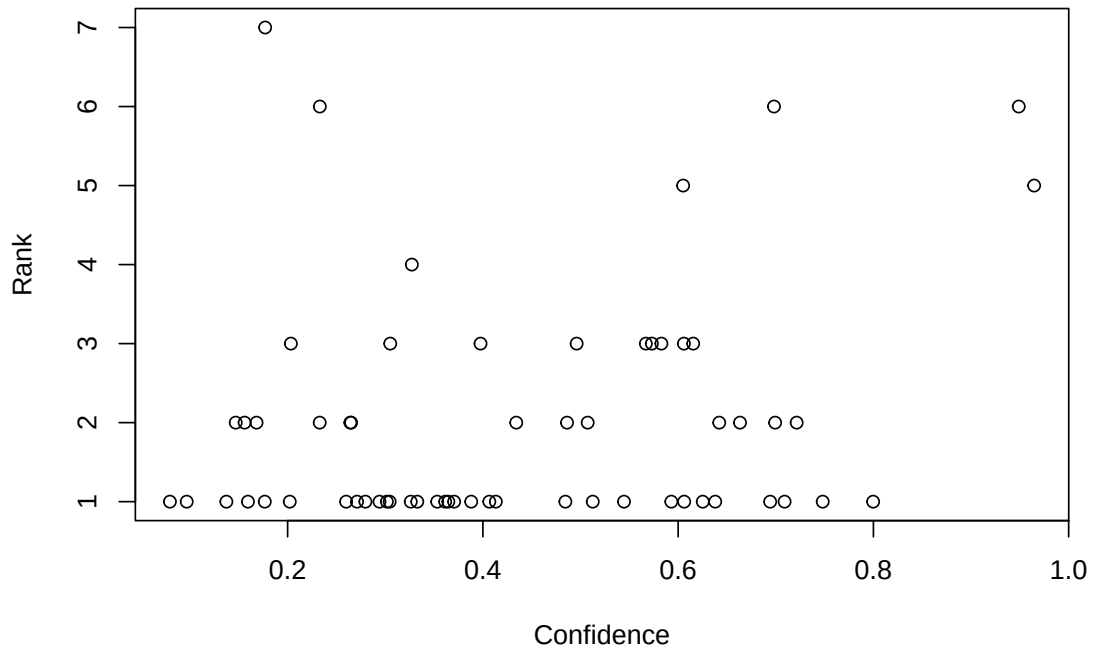

Figure 6. Sirius confidence scores plotted against corresponding ranks. Pearson $\rho = 0.23$.

In addition, despite not accurately calculating the time taken to process 155 spectra due to not being able to make equal measurements, time comparison were performed. It was found that the processing time for MetFrag was a couple of hours, whereas Sirius took only a few minutes.

5 Discussion

Based on this study, it is clear that MetFrag falls short when compared to Sirius. To start with, the tool is limited to being able to process a narrow set of the most common precursor ions. The precursor ions used in Afekta projects contain a wider selection of ions which limits the usability of MetFrag in this case. While MetFrag can make a few targeted annotations correctly, it is not consistent enough to be used in practice since the ratio of correct annotations is low and the ranks for those are too high. Checking all predictions would not make the workload easier for the expert. Additionally, the argument made by Ruttkies et al. [24] that limiting the chemical structure database search by molecular formula may lead to incorrect annotations is proven wrong in the second test case, where MetFrag is not able to annotate anything without molecular formulas. The score type weights for MetFrag were not optimized, since results the results from these tests suggest that the tool is not worthy of further testing.

On the other hand, Sirius showed promising results in the annotation process. Although it may not reach the level of fully automated annotation, it was able to correctly annotate every fifth compound and approximately 40 % of all correct annotations were found within the top 10 annotations. This makes Sirius a useful tool for the expert in the annotation process. One approach for semi-automated annotation could have been to rely on the confidence score of Sirius whether an annotation is correct. However, the correlation test results prove that the confidence score is not accurate. Although the correlation test did not show a negative correlation between confidence score and ranks, other factors suggest confidence in the prediction ability of Sirius. For example, the fact that correct annotations were not found outside the top 10 predictions relates to confidence in the prediction ability of Sirius. It is possible that the metabolites that Sirius cannot annotate or that are found in higher ranks may be too complex for the current approach. Additionally, correct annotations between targeted and non-targeted approaches stay the same, which also implies that simple metabolites can be annotated confidently. Moreover, the minimal difference between targeted and non-targeted accuracy proves that Sirius is able to predict the molecular formula of the compound accurately.

The next step in this study would be testing Sirius in practice with unknown features. If it turns out successful, an automated pipeline could be applied to ensure a faster work process. One additional solution could be combining multiple tools in the same pipeline to obtain different annotations from different tools. In this case, as MetFrag configuration is too complex to run efficiently and the tool is too slow to run in practice, combining tools was not tested further. Nevertheless, it can be an efficient solution in the future.

6 Conclusions

Based on the results of the experiments with MetFrag and Sirius, it can be concluded that metabolomics data is too complex to be accurately modeled with basic computational methods. Currently available annotation tools can assist in the demanding task of annotating metabolites but are not yet fully capable of automating the process. The development of reliable tools in this field is still in the early stages, with only a few tools meeting all of the necessary requirements and successfully producing results. Furthermore, the documentation for many of these tools is often inaccurate or outdated. As metabolomics research continues to advance, there is a pressing need for further development of computational tools that can accurately and reliably annotate metabolites. Only through continued improvement and innovation in this field will researchers be able to fully realize the potential of metabolomics for understanding biological systems and diseases.

REFERENCES

[1] Alseekh, S. and Fernie, A. (2018). Metabolomics 20 years on: what have we learned and what hurdles remain? *The Plant Journal*, 94(6):933–942.

[2] Bird, I. (1989). High performance liquid chromatography: principles and clinical applications. *BMJ: British Medical Journal*, 299(6702):783–787.

[3] Blaženović, I., Kind, T., Ji, J., and Fiehn, O. (2018). Software tools and approaches for compound identification of lc-ms/ms data in metabolomics. *Metabolites*, 8(2).

[4] Böcker, S. and Rasche, F. (2008). Towards de novo identification of metabolites by analyzing tandem mass spectra. *Bioinformatics*, 24(16):i49–i55.

[5] Chaleckis, R., Meister, I., Zhang, P., and Wheelock, C. E. (2019). Challenges, progress and promises of metabolite annotation for lc-ms-based metabolomics. *Current Opinion in Biotechnology*, 55:44–50.

[6] Dührkop, K., Hufsky, F., and Böcker, S. (2014). Molecular formula identification using isotope pattern analysis and calculation of fragmentation trees. *Mass Spectrometry*, 3(Special_Issue_2):S0037–S0037.

[7] Dührkop, K., Fleischauer, M., Ludwig, M., Aksenov, A., Melnik, A., Meusel, M., Dorrestein, P., Rousu, J., and Böcker, S. (2019). Sirius 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nature Methods*, 16:299–302.

[8] Dührkop, K., Shen, H., Meusel, M., Rousu, J., and Böcker, S. (2015). Searching molecular structure databases with tandem mass spectra using csi:fingerid. *Proceedings of the National Academy of Sciences of the USA*, 112(41):12580–12585.

[9] Hemström, P. and Irgum, K. (2006). Hydrophilic interaction chromatography. *Journal of Separation Science*, 29(12):1784–1821.

[10] Hofmann, T., Schölkopf, B., and Smola, A. J. (2008). Kernel methods in machine learning. *The Annals of Statistics*, 36(3):1171–1220.

[11] Johnson, A. and Carlson, E. (2015). Collision-induced dissociation mass spectrometry: A powerful tool for natural product structure elucidation. *Analytical Chemistry*, 87(21):10668–10678.

[12] Johnson, C. and Gonzalez, F. (2012). Challenges and opportunities of metabolomics. *Journal of Cellular Physiology*, 227(8):2975–2981.

[13] Johnson, C., Ivanisevic, J., and Siuzdak, G. (2016). Metabolomics: beyond biomarkers and towards mechanisms. *Nature reviews. Molecular cell biology*, 17(7):451–459.

[14] Katajamaa, M. and Orešič, M. (2007). Data processing for mass spectrometry-based metabolomics. *Journal of Chromatography A*, 1158(1):318–328.

[15] Kind, T. and Fiehn, O. (2007). Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC bioinformatics*, 8(1):1–20.

[16] Kondor, R. and Jebara, T. (2003). A kernel between sets of vectors. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 361–368.

[17] Ludwig, M., Dührkop, K., and Böcker, S. (2018). Bayesian networks for mass spectrometric metabolite identification via molecular fingerprints. *Bioinformatics*, 34(13):i333–i340.

[18] Lutz, N., Sweedler, J., and Wevers, R. (2013). *Methodologies for Metabolomics: Experimental Strategies and Techniques.* Cambridge University Press.

[19] Mattsson, A. (2019). Analysis of LC-MS data in untargeted nutritional metabolomics. Master's thesis, Aalto University. School of Science.

[20] National Institutes of Health (2023). Pubchem website. `https://pubchem.ncbi.nlm.nih.gov/docs/statistics`. Accessed: 15.3.2023.

[21] Oliver, S. G., Winson, M. K., Kell, D. B., and Baganz, F. (1998). Systematic functional analysis of the yeast genome. *Trends in Biotechnology*, 16(9):373–378.

[22] Pluskal, T., Castillo, S., Villar-Briones, A., and Oresic, M. (2010). Mzmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics*, 11:395.

[23] Ruttkies, C., Neumann, S., and Posch, S. (2019). Improving metfrag with statistical learning of fragment annotations. *BMC bioinformatics*, 20(1):1–14.

[24] Ruttkies, C., Schymanski, E., Wolf, S., Hollender, J., and Neumann, S. (2016). Metfrag relaunched: incorporating strategies beyond in silico fragmentation. *Journal of Cheminformatics*, 8(3).

[25] Sasse, M. and Rainer, M. (2022). Mass spectrometric methods for non-targeted screening of metabolites: A future perspective for the identification of unknown compounds in plant extracts. *Separations*, 9(12):415.

[26] Senior, J. K. (1951). Partitions and their representative graphs. *American Journal of Mathematics*, 73(3):663–689.

[27] Shen, H., Dührkop, K., Böcker, S., and Rousu, J. (2014). Metabolite identification through multiple kernel learning on fragmentation trees. *Bioinformatics*, 30(12):i157–i164.

[28] Smith, C., Want, E., O'Maille, G., Abagyan, R., and Siuzdak, G. (2006). Xcms: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical Chemistry*, 78(3):779–787.

[29] Stein, S. (2012). Mass spectral reference libraries: An ever-expanding resource for chemical identification. *Analytical Chemistry*, 84(17):7274–7282.

[30] Tsugawa, H., Cajka, T., Kind, T., Ma, Y., Higgins, B., Ikeda, K., Kanazawa, M., Van der Gheynst, J., Fiehn, O., and Arita, M. (2015). Ms-dial: data-independent ms/ms deconvolution for comprehensive metabolome analysis. *Nature Methods*, 12(6):523–526.

[31] Weckwerth, W. and Kahl, G. (2013). *The Handbook of Plant Metabolomics*. Molecular Plant Biology. John Wiley Sons, Incorporated, Hoboken.

[32] Zhou, B., Xiao, J. F., Tuli, L., and Ressom, H. W. (2012). Lc-ms-based metabolomics. *Molecular bioSystems*, 8(2):470–481.

Appendix 1. Examples of MetFrag files

Parameter file:

```
# data file containing mz intensity peak pairs (one per line)
PeakListPath = example-data.txt
# database parameters -> how to retrieve candidates
MetFragDatabaseType = PubChem
NeutralPrecursorMolecularFormula = C9H11Cl3NO3PS
NeutralPrecursorMass = 348.926284
#IonizedPrecursorMass = 349.93356
DatabaseSearchRelativeMassDeviation = 10
# peak matching parameters
FragmentPeakMatchAbsoluteMassDeviation = 0.02
FragmentPeakMatchRelativeMassDeviation = 50
PrecursorIonMode = 1
IsPositiveIonMode = True
# scoring parameters
MetFragScoreTypes = FragmenterScore,
    AutomatedPeakFingerprintAnnotationScore,
    AutomatedLossFingerprintAnnotationScore
MetFragScoreWeights = 0.378,0.488,0.134
# output
# SDF, XLS, CSV, ExtendedXLS, ExtendedFragmentsXLS
MetFragCandidateWriter = XLS
SampleName = example-1
ResultsPath = metfrag/results/
# following parameteres can be kept as they are
MaximumTreeDepth = 2
MetFragPreProcessingCandidateFilter = UnconnectedCompoundFilter
MetFragPostProcessingCandidateFilter = InChIKeyFilter
# NumberThreads = 1
```

(Continues)

Appendix 1. (continued)

Spectrum file:

96.95085	11001
114.96142	841714
124.98212	30239
142.99266	55890
153.01348	160312
171.02398	618731
197.92748	1519359
213.90458	11943
225.95878	15931
241.93590	2183
275.86044	16512
293.87101	550314
321.90224	1063710
322.90593	10373
349.93349	948212
350.93711	12565
366.80088	11456
367.80119	87667
368.79739	6865

Appendix 2. An example of Sirius .ms file

```
>compound Bicuculline
>formula C20H17NO6
>parentmass 368.113616943359
>ionization [M+H]+
>numpeaks 19
96.95085        11001
114.96142       841714
124.98212       30239
142.99266       55890
153.01348       160312
171.02398       618731
197.92748       1519359
213.90458       11943
225.95878       15931
241.93590       2183
275.86044       16512
293.87101       550314
321.90224       1063710
322.90593       10373
349.93349       948212
350.93711       12565
366.80088       11456
367.80119       87667
368.79739       6865
```