Automated recommendation of metabolite substructures from mass spectra using frequent pattern mining

Aida Mrzic[1,2], Pieter Meysman[1,2], Wout Bittremieux[1,2], Kris Laukens[1,2,*]

1 Department of Mathematics and Computer Science, University of Antwerp

2 Biomedical Informatics Research Center Antwerp (biomina), University of Antwerp/Antwerp University Hospital

** kris.laukens@uantwerpen.be*

Despite the increasing importance of metabolomics approaches, the structural elucidation of metabolites from mass spectral data remains a challenge. Although several reliable tools to identify known metabolites exist, identifying compounds that have not been previously seen is a challenging task that still eludes modern bioinformatics tools. Here, we describe an automated method for substructure recommendation from mass spectra using pattern mining techniques. Based on previously seen recurring substructures our approach succeeds in identifying parts of unknown metabolites. An important advantage of this approach is that it does not require any prior information

concerning the metabolites to be identified, and therefore it can be used for the (partial) identification of unknown unknowns. Using association rule mining we are able to recommend valid substructures even for those metabolites for which no match can be found in spectral libraries or structural databases. We further demonstrate how this approach is complementary to existing metabolite identification tools, achieving improved identification results. The method is called MESSAR (MEtabolite SubStructure Auto-Recommender) and is implemented as a free online web service available at http://www.biomina.be/apps/MESSAR/.

Key words: metabolomics, metabolite identification, substructure, mass spectrometry, pattern mining

1 Introduction

Metabolomics is the discipline that deals with the high-throughput analysis of metabolites, i.e. small biomolecules, with highly relevant applications in drug and biomarker discovery [4, 26]. However, the identification of metabolites from a biological sample remains a major bottleneck in metabolomics, with a vast number of potentially interesting metabolites that are still unknown [12, 19].

The standard method for metabolite identification is mass spectrometry (MS), preceded by a separation technique, such as gas chromatography (GC) or liquid chromatography (LC). Single-stage MS measures the mass-to-charge ratios (m/z) of intact metabolites. To obtain information beyond the mass of a molecule, tandem mass spectrometry (MS/MS) is used. During MS/MS the m/z of the product ion fragments of isolated metabolites are recorded, yielding additional structural information. The traditional way to identify an observed metabolite through spectral library searching works by comparing the measured MS/MS spectra to (historic) spectra of previously identified

compounds stored in a spectral library and selecting the best match. A drawback of spectral library searching is that it is only possible to obtain a valid identification for a given MS/MS spectrum if the spectral library contains a corresponding reference measurement [19]. Unfortunately, the size of spectral libraries is necessarily limited: reference spectra need to be explicitly generated from (often synthetic) compounds, which takes substantial effort and is expensive. Consequently, only a somewhat limited number of known unknowns can be effectively identified in this manner.

Recent approaches have moved beyond the use of spectral libraries in an attempt to identify additional metabolites from a biological sample, for example by using molecular structural databases, which are larger than spectral libraries by several orders of magnitude [12]. Here, the experimental spectra are compared to fragmentation spectra that are predicted from molecular structures [3, 8–10, 14, 20, 28]. Nevertheless, these approaches are still limited as well as they can only identify molecules that are present in the used database and as they crucially rely on completely correct fragmentation predictions.

Besides identifying metabolites by searching in structural databases, there are methods that aim to predict certain structural properties from mass spectral data [5, 6, 10]. Other approaches identify structurally similar molecules [2], group similar spectra [29], mine spectra for specific user-defined properties [15] or help in the manual annotation of mass spectra [16]. Notably, FingerID [10, 22] uses supervised machine learning to predict molecular fingerprints, i.e. bit vectors where each bit represents the presence (or absence) of a certain structural property of the molecule. Its successor CSI:FingerID [8] further improves the fingerprint prediction using high-resolution MS/MS data and uses the fingerprints to rank candidates retrieved from a molecular database [10]. MS2LDA, a recent approach by van der Hooft et al. [23, 24], employs text mining techniques to discover patterns across fragmentation spectra, which can be used to aid the *de novo* annotation of unknown unknowns. A drawback of this approach, however, remains that

the extracted patterns still need to be structurally annotated based on expert knowledge and matched to the reference spectra, a time-consuming and complex manual process.

In the presented work we introduce a new, automated approach for substructure recommendation from MS/MS spectra based on frequent itemset mining. Frequent itemset mining is a class of data mining techniques that is specifically designed to discover co-occurring items in transactional data sets [17]. Our approach does not rely on the full metabolite being seen before, but looks for commonly observed substructures to identify part of an unknown metabolite. Instead of using predefined molecular fingerprints in a supervised framework, we generate general substructures in an unsupervised fashion by breaking chemical bonds of the molecules. The advantage of this approach is that it allows an automated annotation of metabolite substructures based on fragmentation data to provide a (partial) identification of unknown unknowns. Substructure recommendations are computed from associations between spectral features and structural features derived from a high-quality set of mass spectrum identifications for known metabolites. Based on these data pattern mining techniques are used to detect which substructures are associated with certain fragments and fragment mass differences.

The substructure recommender is available as a free online web service which can be accessed at http://www.biomina.be/apps/MESSAR.

2 Methods

An overview of the presented substructure recommendation workflow is depicted in figure 1. In brief, molecular substructures are first generated for a database of metabolites. These substructures are then combined into a single data set with fragment ions and mass differences between fragment ions extracted from previously identified MS/MS spectra in a spectral library. We then apply frequent pattern mining techniques to this data set to infer which substructures are associated with certain fragments and fragment mass differences. This results in a list of recommendations of the form:

peak p is associated with substructure s with frequency f_p and confidence c_p,

mass difference md is associated with substructure s with frequency f_{md} and confidence c_{md}.

Structural information

We start with a set of metabolites for which both experimental MS/MS data and molecular structures are available. This molecular data is referred to as data set M. For every metabolite $m \in M$, a set of substructures S_m is created using the "breaking of retrosynthetically interesting chemical substructures" (BRICS) fragmentation rules [7], while discarding trivial substructures consisting of only a single atom. We will refer to the set of substructures created by BRICS fragmentation for metabolite m as S_m, and to the set of all substructures for any metabolite m in data set M as $S_M = \{S_m \mid \forall\, m \in M\}$.

Spectral information

Let P_m be the set of all MS/MS peaks p in the spectrum corresponding to a metabolite $m \in M$, then the mass of every $p \in P_m$ is discretized by rounding to one decimal place. We will refer to the sets of rounded peak masses of a spectrum of metabolite $m \in M$ and of the entire data set M as p_m and p_M, respectively.

For every $p, q \in P_m$, $p \neq q$, a mass difference $md_{pq} = |p-q|$ is calculated. Subsequently, the mass differences are discretized by rounding to the closest integer. We will refer to the sets of discretized absolute mass differences of a spectrum of metabolite $m \in M$ and of the entire datas et M as $mdiff_m$ and $mdiff_M$, respectively.

Pattern mining

Specialized pattern mining techniques are used to detect frequent substructures that can be consistently related to the occurrence of certain spectral features. Frequent substructures are identical parts of a molecular structure that frequently occur in a

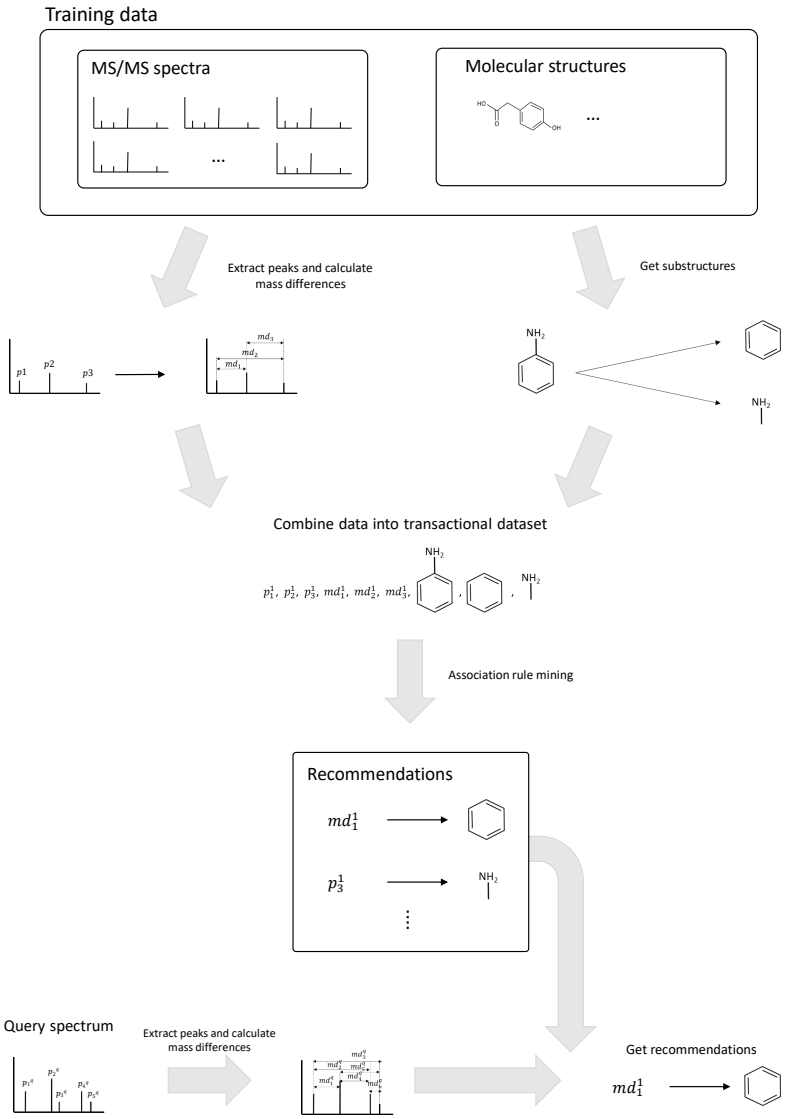

Figure 1: Pattern mining workflow to generate metabolite substructure recommendations. The training data set consists of metabolites for which both an MS/MS spectrum and its molecular substructure are known. These data are transformed into a transactional format, i.e. for every metabolite a single transaction is created which combines the spectral information (peaks and mass differences between the peaks in the MS/MS spectrum) with the corresponding structural information (substructures of the metabolite). These transactions are collected into a single transactional data set and mined for association rules. Rules that are extracted have the form *peak p (mass difference md) is associated with substructure s.*

given data set. To perform pattern mining we require a transactional data set, with each transaction a set of items (or a so-called "itemset"). In our approach, each item consists of a molecular substructure or a spectral feature extracted from an MS/MS spectrum. A transaction is then the set of all substructures, peaks and mass differences for a single molecule. The support of an itemset is defined as the number of transactions in a data set that contain that itemset. An itemset is considered frequent if its support exceeds a specified minimal threshold.

After the frequent itemsets have been determined we mine for association rules to reveal hidden relationships in the transactional data. An association rule can be expressed as $X \Rightarrow Y$, where X and Y are sets of items, and $X \cap Y = \emptyset$. X is called the body or antecedent of the rule, while Y is called the head or consequent of the rule. The rule $X \Rightarrow Y$ states that if itemset X is present in a transaction so is itemset Y, which indicates the existence of an association between X and Y. The support of an association rule $X \Rightarrow Y$ is equal to the support of $X \cup Y$. The confidence of an association rule $X \Rightarrow Y$ is the conditional probability that Y is present in a transaction given that X is also present in that transaction, and is defined as:

$$confidence(X \Rightarrow Y) = \frac{support(X \cup Y)}{support(X)} \tag{1}$$

Association rule mining is the task of finding all the association rules that are frequent and confident, i.e. identifying the association rules for which the support and confidence exceeds specified minimal support and confidence thresholds. A high support indicates that the rule applies to a large number of cases, while a high confidence indicates that the rule should often be correct.

In our approach, for each metabolite $m \in M$ the transaction T_m consists of all peaks p_m, all mass differences $mdiff_m$ and all molecular substructures S_m: $T_m = p_m \cup mdiff_m \cup S_m$. We will refer to the set of transactions T_m of all the metabo-

lites $m \in M$ as T_M. We only mine for patterns in the transactional data set T_M that consist of both molecular substructures and peak masses or mass differences, as these combinations indicate spectral–structural associations. Our frequent itemset mining algorithm is optimized so that only those itemsets that contain at least one pair (p, s) or pair (md, s) are retained, with $p \in p_m$, $md \in mdiff_m$ and $s \in S_m$. This step reduces the (large) search space by pruning uninteresting itemsets that do not contain a combination of both spectral and structural information. Furthermore, the association rule mining step is optimized so that only those rules that contain one peak $p \in p_m$ (mass difference $md \in mdiff_m$) in its antecedent and one substructure $s \in S_m$ in its consequent are considered.

Mining for association rules in the transactional data set T_M results in a list of associations of the form:

peak p_i can be associated with substructure s_j with support f_p and confidence c_p.

mass difference md_i can be associated with substructure s_j with support f_{md} and confidence c_{md}.

Such associations can be interpreted as recommendations for unexplained spectra in which the given peaks and mass differences are observed. Note that the peak (mass difference) in the antecedent of the rule and the molecular mass of the substructure in the rule's consequent will rarely be exactly equal. Because it is very hard to accurately simulate the fragmentation of a molecule to generate its substructures due to molecular rearrangements [25], the spectral information can typically be linked to a substructure that is a part of the metabolite under consideration, even though this might not necessarily be the exact substructure responsible for that spectral information.

Substructure recommendations

Small molecule data was retrieved from the Human Metabolome Database (HMDB) [27]. A total of 814 compounds for which both experimental MS/MS data and the molecu-

lar structure are available were taken into account with no further restrictions being present. As such, this data contains a heterogeneous set of metabolites. Only those spectra labeled as 'Excellent' were used. Mass spectra generated at different collision energy levels belonging to the same metabolite were combined, while peaks from different collision energies with an m/z difference less than or equal to 0.01 Dalton were merged. We grouped the spectra based on ionization modes and retained only those peaks with relative intensity compared to the most intense peak exceeding 5%.

After converting this data into two transactional data sets, for positive ionization mode and negative ionization mode, association rules were mined with support and confidence thresholds of 3 and 1% respectively, resulting in a total of 92,597 unique recommendations for positive ionization mode and 15,278 for negative ionization mode. As an example, figure 2 shows some of the recommended substructures for the MS/MS spectrum of 3,4-Dihydroxyphenylacetic Acid.

2.1 Improving metabolite identifications

The substructure recommendations can be used to improve the accuracy of the existing tools for metabolite identification. First, to generate recommendations, for an unidentified query MS/MS spectrum all peaks with intensity exceeding 5% of the most intense peak are extracted and all mass differences between these peaks are calculated. These spectral features are then checked against the association rule database and substructures are recommended based on rules containing the matching spectral features. Taking into account possible neutral losses, we allowed a maximum margin of 1 Da for mass differences and 0.1 Da for peaks. Subsequently, the recommended substructures are ranked by the average confidence of their respective rules, as it is often the case that certain substructures will be present in more than one rule.

The recommended substructures can be used to rank candidate metabolites retrieved from a database. The ranking is done based on the similarity between the recommended

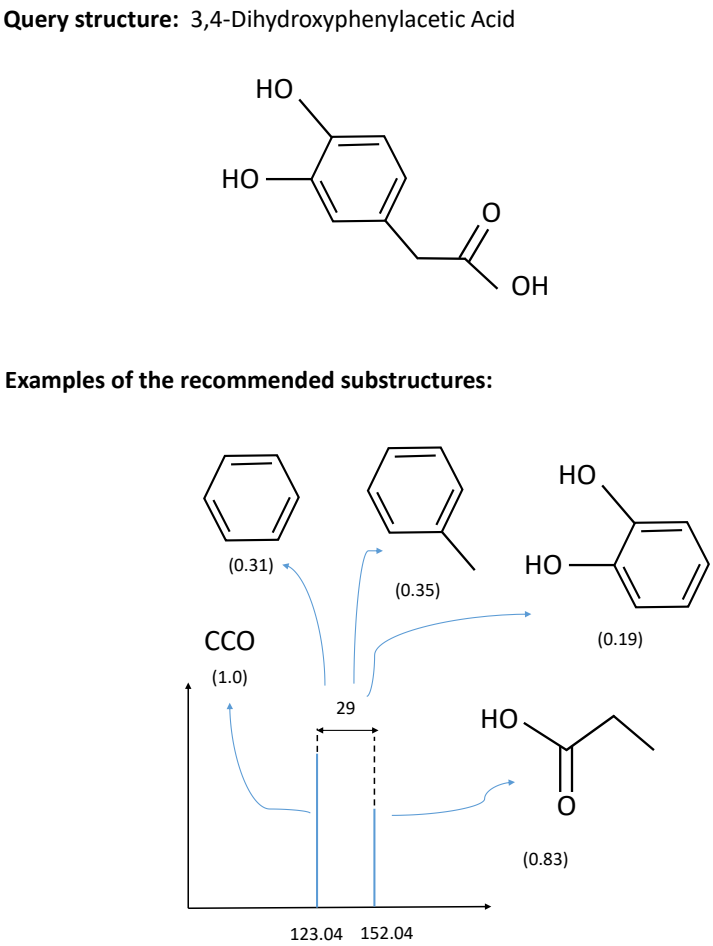

Figure 2: Query substructure of 3,4-Dihydroxyphenylacetic Acid together with the examples of recommended substructures. The confidence of the rule recommending the substructure is given in the brackets.

substructures and the candidate metabolite structure. Full metabolite structures that contain a higher number of increasingly confident substructures recommended by our approach receive a higher rank. In this fashion it is possible, for example, that a structure containing two recommended substructures is ranked higher than a structure containing three recommended substructures if the average confidence of the rules recommending the substructures is higher in the first case.

3 Results

3.1 MESSAR rules cover the MS2LDA patterns

MS2LDA [24] discovers patterns across fragmentation spectra in an unsupervised fashion using text mining techniques. As it operates within a similar scope as our tool, a detailed comparison is warranted. This method was used to discover patterns consisting of spectral peaks and neutral losses, which are then structurally annotated based on expert knowledge and matched to reference spectra. This resulted in a set of so-called Mass2Motifs that couple patterns in the spectra to molecular structures. These patterns can then be used to partially identify unknown spectra. While this approach does not require prior structural information, the generated patterns do have to be manually annotated mid-process. Our pattern mining approach uses similar input information to address a similar problem, but it provides an automated recommendation in turn. We compared our set of recommendation rules with the Mass2Motifs derived from four beer extract data sets [24]. We only took into account those Mass2Motifs labeled with the highest level of confidence. Neutral losses were rounded to the nearest integer. In positive ionization mode, there are 31 high-confident Mass2Motifs. Each Mass2Motif consists of at least one spectral peak and/or neutral loss and a matching substructure. Out of these 31 patterns, 6 have equivalent MESSAR recommendation rules, i.e. the spectral peak (or neutral loss) is identical to the peak (or mass difference) in the antecedent of the

Ionization mode	Complete match	Partial match	No match
Positive	6/31	25/31	0
Negative	6/16	8/16	2/16

Table 1: Fraction of MS2LDA patterns appearing in MESSAR rule database.

rule and the substructure is identical to the substructure in the consequent of the rule (table 1). The remaining 25 Mass2Motifs have partial matches in the rule database, i.e. peaks (or neutral losses) are matched to peaks (or mass differences) in the antecedent of the rules, but substructures feature a variant of the Mass2Motif substructures. For example, the Mass2Motif consisting of the spectral peak at 18.0080 and substructure O (water loss) is related to the MESSAR recommendation rule $md_{18} \Rightarrow CO$. There is no MESSAR rule containing a single oxygen atom because MESSAR explicitly only retains the substructures that consist of at least two atoms to avoid trivial recommendations. Another example of a partial match is the Mass2Motif consisting of three fragments and one neutral loss indicative of a serine substructure, where two fragments and the neutral loss have complete matches in the database while the remaining fragment is partially matched (the substructure in the MESSAR rule has one oxygen atom less), as shown in figure 3.

Out of 16 high-confident Mass2Motifs in negative ionization mode, 6 are completely matched, 8 patterns are partially matched to a MESSAR rule and 2 patterns have no match (table 1). Both Mass2Motifs with no match contain the sulphate group substructure for which MESSAR rules were found, but none with the peaks documented in the Mass2Motifs.

In summary, almost all high-confident Mass2Motifs are at least partially covered by the recommendation rules generated using the MESSAR approach, despite that both sets of patterns originate from different data sets. Despite their similar performance, an important advantage of MESSAR over MS2LDA is that MESSAR does not require

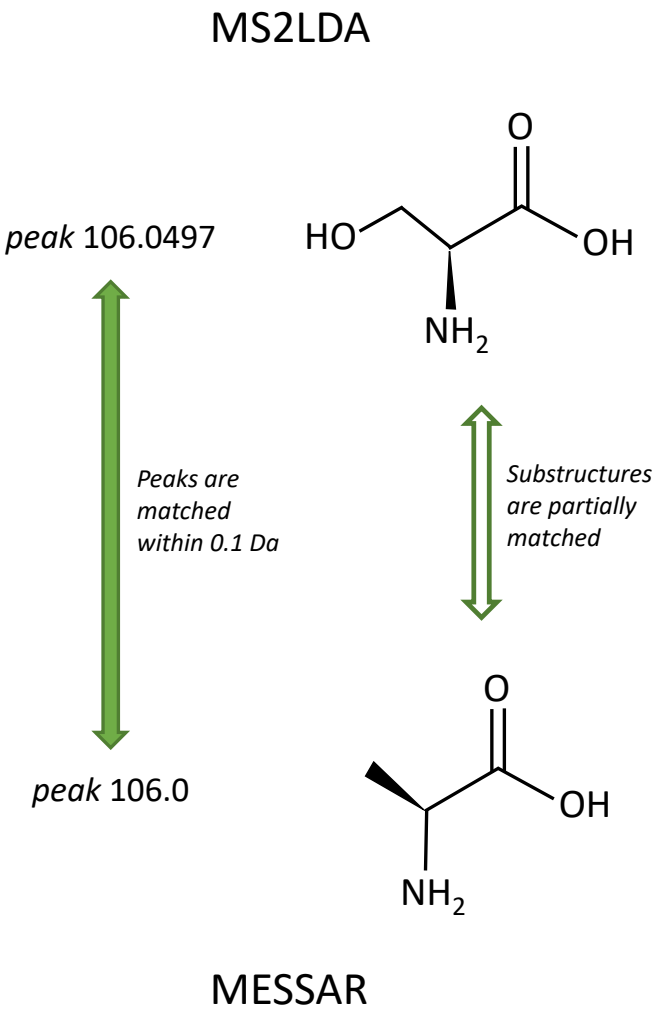

Figure 3: Example of a partial match between an MS2LDA pattern indicative for a serine substructure and the corresponding MESSAR rule. The partial match exhibits identical spectral properties (within our mass margin) and partially matched substructures.

manual curation or integration of expert knowledge to generate its rules.

3.2 MESSAR rules correspond to true substructures

To evaluate the accuracy of the generated recommendations we used data provided for the previous two Critical Assessment of Small Molecule Identification (CASMI) challenges (CASMI2014, CASMI2016) [21]. CASMI is an open contest in which participants have to identify the molecular formula and chemical structure for molecules of natural as well as synthetic origins based on mass spectrometry data. The winner is determined by the number of correctly predicted structure identifications. This remains a significant challenge, as for some metabolites the true structure was even missing from all predictions, by any of the competing teams. We used MS/MS spectra from in total 60 metabolites from CASMI2014 and CASMI2016.

Using MESSAR we generated recommended substructures from the spectra of these 60 compounds. These recommendations were then compared to the true structures. As a measure of the quality of the prediction a Fisher's exact test was performed to see whether the recommendations are statistically significant compared to simply randomly assigning substructures to spectra. Figure 4 shows the p-values for this test, which indicates that our method is able to provide relevant recommendations. Only for a single compound the p-value exceeded 0.05, and that compound proved to be hard to identify for some of the CASMI contestants as well as two out of seven contestants did not list the true structure at the top of their predictions.

Each recommendation was assigned a confidence score based on the associations between substructures and peaks or mass differences found within the data set. Recommendations with a higher confidence can be assumed to have a higher chance of giving a correct recommendation. The recommendations can therefore be ranked by this confidence value for each CASMI metabolite. Figure 5 shows the mean receiver operator characteristic (ROC) curve for the CASMI data recommendations calculated from 60

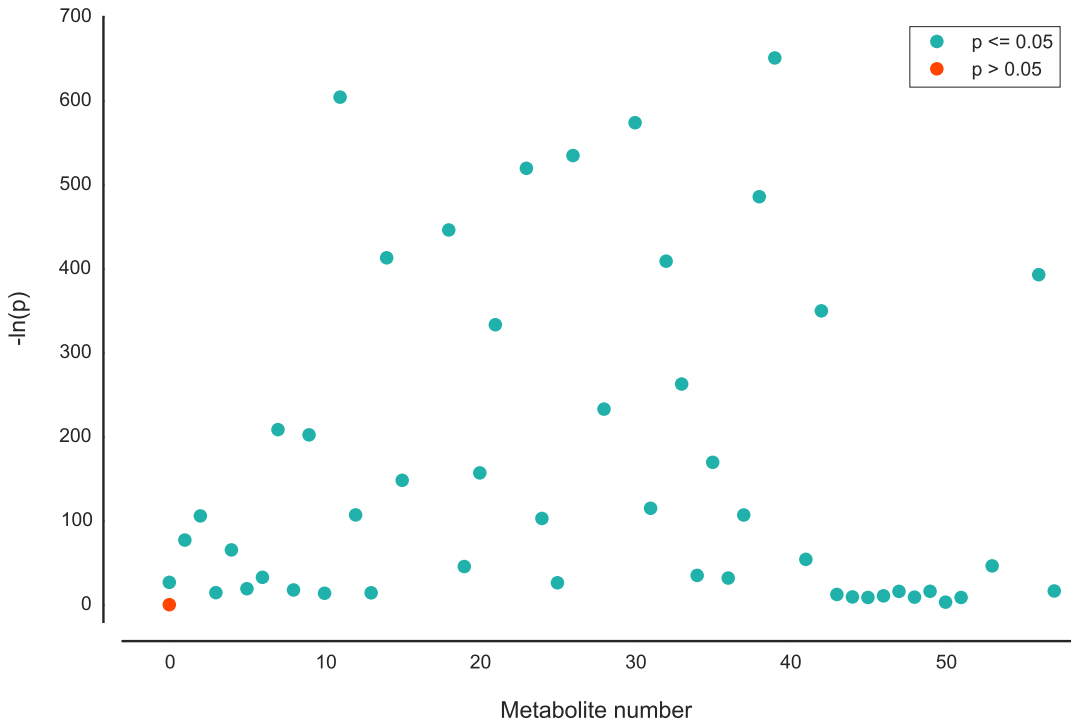

Figure 4: Enrichment of recommended substructures in the CASMI molecular structures. The p-values are given for each of the 60 molecules tested based on a Fisher's exact test. Green dots denote significant enrichment compared to random assignment, red dots denote non-significant enrichment.

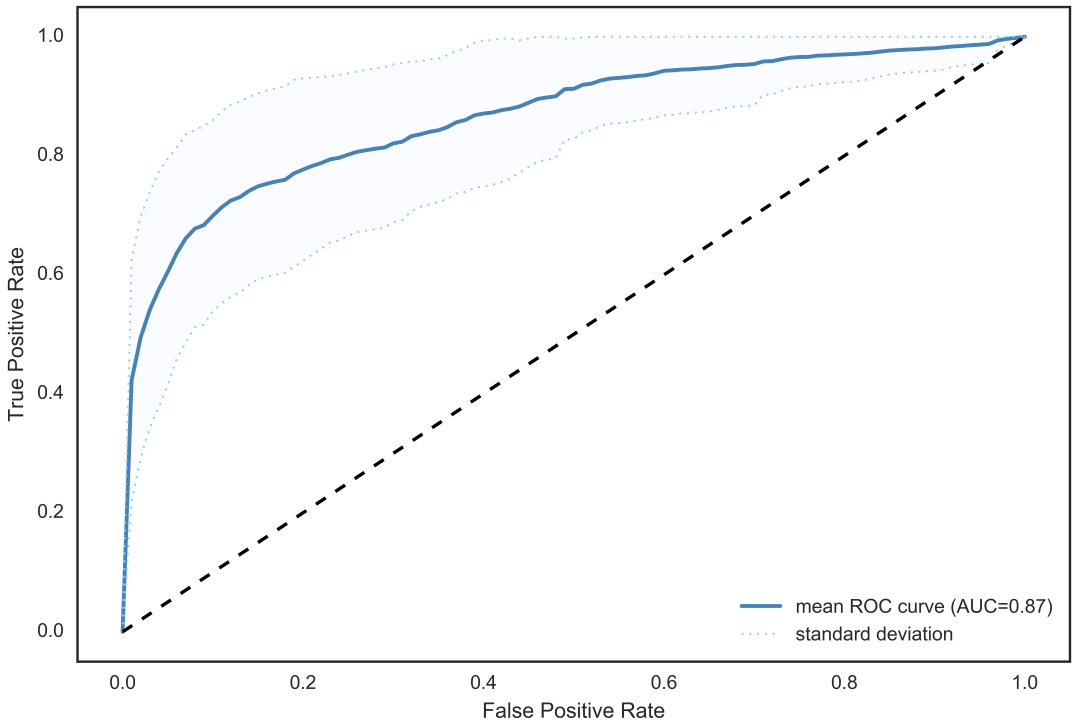

Figure 5: Mean ROC curve for CASMI metabolites. The ROC curve shows that our method can accurately distinguish substructures present in a metabolite from those that are not present based on the average confidence of the recommendations. It plots the true positive rate (TPR) as a function of the false positive rate (FPR) for different average confidence thresholds.

curves, with each curve evaluating the recommendations for a single compound. The ROC results show that recommendations with a higher confidence are more likely to be true for each of these metabolites than those with a lower confidence. These findings suggest that the average confidence of the recommendations provides a good measure of the quality for the recommended substructures.

3.3 Substructure recommendations provide additional insights

One of the most common analysis methods for metabolite spectra is to search the full spectrum in a mass spectral database. To examine whether or not the substructure recommendations from MESSAR provide us with relevant additional information compared

to full spectral matching, we used the MassBank mass spectral database [11] to identify 60 metabolites of the CASMI data. MassBank returned matches for only 7 metabolites out of 60. In addition, out of these 7 metabolites only one was identified correctly (ibuprofen) in the MassBank search results. In contrast, as shown above, we are able to generate relevant recommendations even for those spectra for which no similar spectrum can be found in a spectral library.

3.4 Comparison to FingerID

Another common analysis approach for metabolite spectra is to use supervised models to predict the presence or absence of fingerprints. We therefore compared our unsupervised method with a supervised fingerprint-based approach, namely FingerID [10, 22]. For candidate molecule retrieval we used the PubChem database [13], which contains all of the CASMI compounds. First, for each of the 60 metabolites in the CASMI data set we retrieved all possible candidate matches from PubChem whose precursor mass falls within the given mass tolerance provided by the CASMI challenge. Given the PubChem version we used (2011), this resulted in 21 out of 60 molecules (18/21 in positive mode) for which the candidate list contained the correct identification. Next, we generated fingerprints using FingerID and substructures using MESSAR and evaluated how this information can be used to rank the candidates retrieved from PubChem. For FingerID we computed a fingerprint for each MS/MS spectrum based on the set of 528 unique fingerprints from OpenBabel (FP3, FP4 and MACCS) [18] and eliminated those that appear in all molecules or do not appear at all [10, 22]. This resulted in a final number of 299 fingerprints. We followed the original authors' recommendations and only trained fingerprint prediction model applicable to a positive ionization mode [8]. Next, we ranked the candidate molecules retrieved from PubChem based on the similarity with the predicted fingerprints. Similarly, we obtained MESSAR recommendations for all MS/MS spectra and used these to rank each of the candidate molecules. We then compared the

FingerID		MESSAR	
Avg.rank	rank≤20	Avg.rank	rank≤20
1799	0/18	1456	3/18

Table 2: Comparison of FingerID and MESSAR metabolite identification using PubChem.

FingerID ranks with the MESSAR ranks for each of the 18 spectra. Table 2 shows the ranking results for both FingerID and MESSAR. Although in general the ranking is rather poor for both tools this is to be expected considering the size of the PubChem database. Nevertheless, the MESSAR ranks are somewhat better than the FingerID ranks. For MESSAR both the average rank is slightly lower and the number of true structures that were ranked within the top 20 is higher. A direct comparison of the rankings does reveal that they perform roughly the same, as can be seen in figure 6.

This shows that the performance of our method is at least comparable to that of a recent fingerprint-based supervised method given the same low-resolution training data. Furthermore the results suggest that the two approaches are likely complementary, each providing users with a different set of information that can be used to infer the true molecular structure.

3.5 Substructure recommendations improve existing prediction tools

As MESSAR does not provide full metabolite identifications but only recommends substructures, its main applicability lies in assisting a *de novo* annotation of mass spectra and in improving existing metabolite prediction tools. To evaluate the suitability of substructure recommendations in combination with existing tools we used the MAGMa search engine, a state-of-the-art metabolite structure identification tool based on a structural database [20], which was the winner of CASMI2014. For the MAGMa structural search again the PubChem database [13] was used. For each of the spectra in the CASMI data set we retained the 50 highest ranked MAGMa predictions, which were subsequently

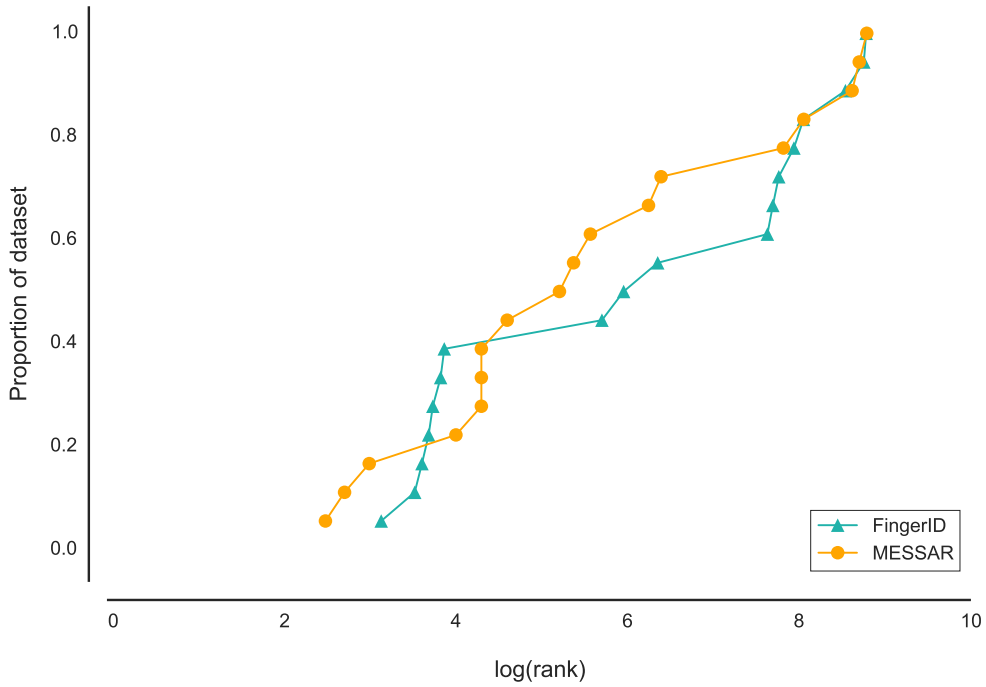

Figure 6: Rank distribution of CASMI metabolites for FingerID and MESSAR. The ranks are shown in log scale.

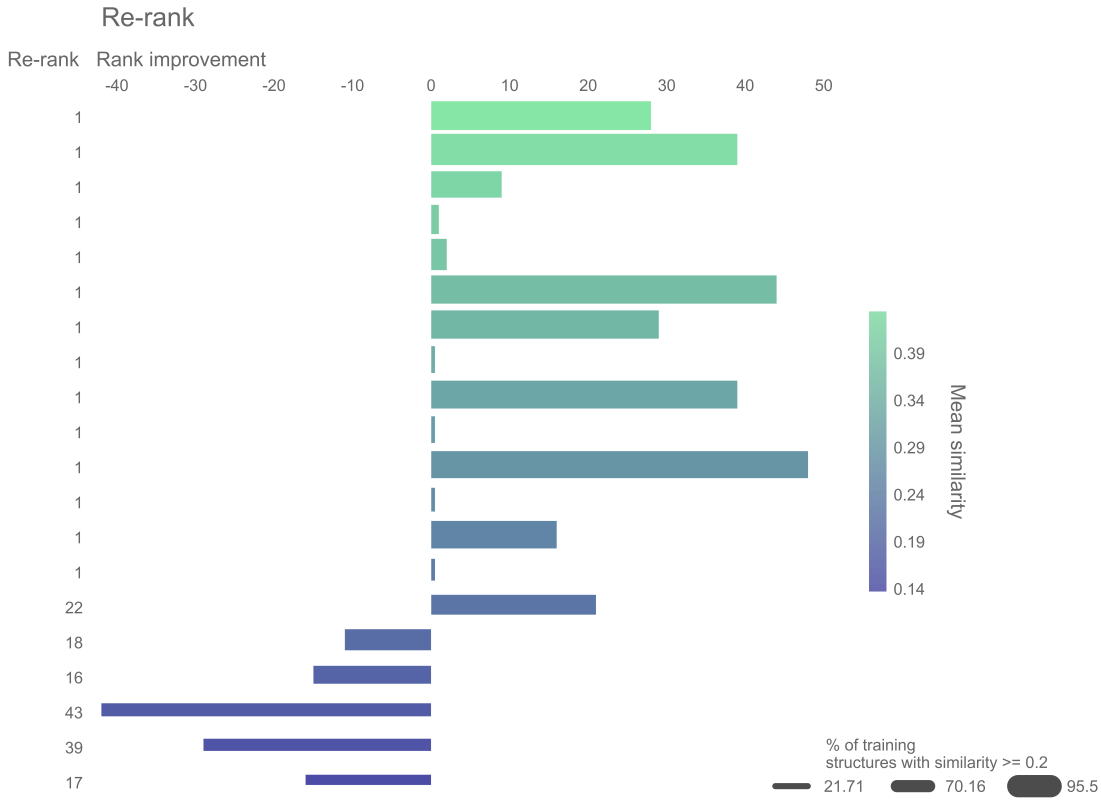

Figure 7: Identification improvements by reranking the MAGMa structure predictions. The horizontal axis shows the rank improvement, while the final rank is shown in the first column. Furthermore, for every CASMI metabolite the average similarity between the training data set and that metabolite is shown, as well as the percentage of training structures for which the similarity with that metabolite is larger than 0.2 (20%). The similarity is calculated by comparing molecular fingerprints [1], which are formed by bit vectors with each bit representing the presence (or absence) of a certain structural property of the molecule.

reranked based on our recommended substructures.

Figure 7 shows the results of the reranking. Out of 20 metabolites for which the true structure was ranked among the top 50 MAGMa identifications, 11 correct identifications were ranked higher by combining the substructure recommendations with the MAGMa identifications, with 10 correct identifications receiving rank 1. Furthermore, four correct identifications which were top ranked retained this rank, while five correct identifications were ranked lower. The rank decrease for these last five correct identifications can be explained by the significant dissimilarity between their structures and the data set used

to generate the recommendation rules. Although our approach does not require that a fully matching structure is present in the training data set, it still requires repeated observations of matching substructures. Unfortunately, because only a limited number of high-quality metabolomics MS/MS spectra are publicly available, the data set used within this study consisted of less than one thousand metabolites. As a result, for substructures that are missing from the data set, or that only occur a few times, there is insufficient data to learn the relationship between the substructure and its spectra. By increasing the size of the MS/MS data set to be mined we expect that the performance of our method will increase further. Nevertheless, despite the decrease in identification performance for a few molecules due to a lack of suitable training data, the provided rerankings on average greatly improve the MAGMa predictions.

4 Conclusions

We have introduced a novel pattern mining-based approach to recommend metabolite substructures from MS/MS spectra. The aim of this method is to provide ranked recommendations to the end user regarding the origins of unexplained mass spectra. We have shown that our method succeeds in recommending substructures even for those spectra for which no match can be found in mass spectral libraries. Therefore, our tool can be used to assist in the *de novo* annotation of metabolites not present in mass spectral or structural databases. In addition, the recommendations can be combined with existing tools for metabolite structure prediction to improve the accuracy of the compound identifications. An important advantage is that, as opposed to expert-driven substructure recommendations, our method is fully automated. It is freely available at http://www.biomina.be/apps/MESSAR.

References

[1] MACCS structural keys; MDL Information Systems Inc., 14600 Catalina Street, San Leandro, CA 94577.

[2] A. Aguilar-Mogas, M. Sales-Pardo, M. Navarro, R. Guimerá, and O. Yanes. imet: A network-based computational tool to assist in the annotation of metabolites from tandem mass spectra. *Analytical Chemistry*, 86(6):3474–3482, 2017.

[3] F. Allen, R. Greiner, and D. Wishart. Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics*, 11(1):98–110, 2015. ISSN 1573-3890.

[4] E.G. Armitage and C. Barbas. Metabolomics in cancer biomarker discovery: Current trends and future perspectives. *Journal of Pharmaceutical and Biomedical Analysis*, 87:1 – 11, 2014. ISSN 0731-7085. Review Papers on Pharmaceutical and Biomedical Analysis 2013.

[5] B. Curry and D.E. Rumelhart. Computer-aided interpretation of mass spectra. information on substructural properties from stirs. *Tetrahedron Computer Methodology*, 3(3/4):213–237, 1990.

[6] H.E. Dayringer, G.M. Pesyna, R. Venkataraghavan, and F.W. McLafferty. Computer-aided interpretation of mass spectra. information on substructural probabilities form stirs. *Organic Mass Spectrometry*, 11(5):529–542, 1976. ISSN 1096-9888.

[7] J. Degen, C. Wegscheid-Gerlach, A. Zaliani, and M. Rarey. On the art of compiling and using 'drug-like' chemical fragment spaces. *ChemMedChem*, 3:1503–1507, 2008.

[8] K. Dührkop, H. Shen, M. Meusel, J. Rousu, and S. Böcker. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proceedings of*

the National Academy of Sciences of the United States of America, 112(41):12580–12585, 2015. ISSN 0027-8424.

[9] M. Gerlich and S. Neumann. MetFusion: Integration of compound identification strategies. *Journal of Mass Spectrometry*, 48(3):291–298, 2013.

[10] M. Heinonen, H. Shen, N. Zamboni, and J. Rousu. Metabolite identification and molecular fingerprint prediction through machine learning. *Bioinformatics*, 28(18): 2333–2341, 2012. ISSN 1367-4803.

[11] H. Horai, M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, Y. Ojima, K. Tanaka, S. Tanaka, K. Aoshima, Y. Oda, Y. Kakazu, M. Kusano, T. Tohge, F. Matsuda, Y. Sawada, M. Y. Hirai, H. Nakanishi, K. Ikeda, N. Akimoto, T. Maoka, H. Takahashi, T. Ara, N. Sakurai, H. Suzuki, D. Shibata, S. Neumann, T. Iida, K. Tanaka, K. Funatsu, F. Matsuura, T. Soga, R. Taguchi, K. Saito, and T. Nishioka. Massbank: A public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry*, 45(7):703–714, 2010. ISSN 1076-5174.

[12] F. Hufsky, K. Scheubert, and S. Böcker. Computational mass spectrometry for small-molecule fragmentation. *Trends in Analytical Chemistry*, 53:41–48, 2014.

[13] S. Kim, P.A. Thiessen, E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. Shoemaker, J. Wang, B. Yu, J. Zhang, and S. Bryant. Pubchem substance and compound databases. *Nucleic Acids Res*, 44(D1):D1202–D1213, January 2016.

[14] L. Li, R. Li, J. Zhou, A. Zuniga, A.E. Stanislaus, Y. Wu, T. Huan, J. Zheng, Y. Shi, D.S. Wishart, and G. Lin. MyCompoundID: using an evidence-based metabolome library for metabolite identification. *Analytical chemistry*, 85(6):3401–3408, March 2013. ISSN 1520-6882.

[15] Y. Ma, T. Kind, D. Yang, C. Leon, and O. Fiehn. MS2Analyzer: A software

for small molecule substructure annotations from accurate tandem mass spectra. *Analytical Chemistry*, 86(21):10724–10731, 2014.

[16] N.G. Mahieu, J.L. Spalding, S.J. Gelman, and G.J. Patti. Defining and detecting complex peak relationships in mass spectral data: The mz.unity algorithm. *Analytical Chemistry*, 88(18):9037–9046, 2016.

[17] S. Naulaerts, Meysman P., W. Bittremieux, T.N. Vu, W. Vanden Berghe, B. Goethals, and K. Laukens. A primer to frequent itemset mining for bioinformatics. *Briefings in Bioinformatics*, 16(2):216–231, March 2015.

[18] N.M. O'Boyle, M. Banck, C.A. James, C. Morley, T. Vandermeersch, and G.R. Hutchison. Open babel: An open chemical toolbox. *Journal of Cheminformatics*, 3(1):33, 2011. ISSN 1758-2946.

[19] G.J. Patti, O. Yanes, and G. Siuzdak. Innovation: Metabolomics: the apogee of the omics trilogy. *Nature Reviews Molecular Cell Biology*, 13(4):263–269, 2012.

[20] L. Ridder, J.J.J. van der Hooft, S. Verhoeven, R.C. de Vos, R. van Schaik, and J. Vervoort. Substructure-based annotation of high-resolution multistage MS(n) spectral trees. *Rapid Communication in Mass Spectrometry*, 26(20):2461–2471, 2012.

[21] E.L. Schymanski, C. Ruttkies, M. Krauss, C. Brouard, T. Kind, K. Dührkop, F. Allen, A. Vaniya, D. Verdegem, S. Böcker, J. Rousu, H. Shen, H. Tsugawa, T. Sajed, O. Fiehn, B. Ghesquière, and S. Neumann. Critical assessment of small molecule identification 2016: automated methods. *Journal of Cheminformatics*, 9 (1):22, 2017. ISSN 1758-2946.

[22] H. Shen, N. Zamboni, M. Heinonen, and J. Rousu. Metabolite identification through machine learning–tackling casmi challenge using FingerID. *Metabolites*, 3(2):484–505, 2013.

[23] J.J.J. van der Hooft, J. Wandy, F. Young, S Padmanabhan, K Gerasimidis, K.E.V. Burgess, M.P. Barrett, and S. Rogers. Unsupervised discovery and comparison of structural families across multiple samples in untargeted metabolomics. *Analytical Chemistry*, in press.

[24] J.J.J. van der Hooft, J. Wandy, M.P. Barrett, K.E.V. Burgess, and S. Rogers. Topic modeling for untargeted substructure exploration in metabolomics. *Proceedings of the National Academy of Sciences of the United States of America*, 113(48):13738–13743, 2016.

[25] E. Werner, JF. Heilier, C. Ducruix, E. Ezan, C. Junot, and JC. Tabet. Mass spectrometry for the identification of the discriminating signals from metabolomics: Current status and future trends. *Journal of Chromatography B*, 871(2):143 – 163, 2008. ISSN 1570-0232. Hyphenated Techniques for Global Metabolite Profiling.

[26] D.S. Wishart. Emerging applications of metabolomics in drug discovery and precision medicine. *Nature Reviews Drug Discovery*, 15(7):473–484, 2016. ISSN 1474-1776.

[27] D.S. Wishart, T. Jewison, A.C. Guo, M. Wilson, C. Knox, and et al. The human metabolome database in 2013. *Nucleic Acids Res*, 41(D1):D801–D807, January 2013.

[28] S. Wolf, S. Schmidt, M. Müller-Hannemann, and S. Neumann. In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics 2010*, 148:11, 2010. doi: 10.1186/1471-2105-11-148.

[29] J.Y. Yang, L.M. Sanchez, C.M. Rath, X. Liu, P.D. Boudreau, N. Bruns, E. Glukhov, A. Wodtke, R. de Felicio, A. Fenner, W.R. Wong, R.G. Linington, L. Zhang, H.M. Debonsi, W.H. Gerwick, and P.C. Dorrestein. Molecular networking as a dereplication strategy. *Journal of Natural Products*, 76(9):1686–1699, 2013.