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 rules extracted by pattern mining we are able to recommend valid substructures even for those metabolites for which no match can be found in spectral libraries. 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 very important applications in drug and biomarker discovery [4, 19]. 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 [10, 14].

The standard method for metabolite identification is mass spectrometry (MS) proceeded 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. In MS/MS mode, the the m/z of their 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 identi-

fied 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 [14]. Unfortunately, the size of spectral libraries is necessarily limited: reference spectra need to be explicitly generated from (often synthetic) compounds, which takes a lot of 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 attempts 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 [10]. Here, the experimental spectra are compared to fragmentation spectra that are predicted from molecular structures [3, 6–8, 12, 15, 21]. However, these approaches can also only identify molecules that are present in the used database and crucially rely on complete correct fragmentation predictions.

Besides identifying metabolites by searching in structural databases, other methods aim to identify structurally similar molecules [2] or substructures of unknown metabolites [17]. For example, a recent approach by van der Hooft et al. [17] 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, is that the extracted patterns still need to be structurally annotated based on expert knowledge and matched to reference spectra, a time-consuming and complex manual process.

Here, we introduce an 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 datasets [13]. Our approach does not rely on the full metabolite being seen before, but instead it looks for commonly observed substructures to iden-

tify part of an unknown metabolite. The advantage of this approach is that it is able to automatically annotate metabolite substructures based on fragmentation data which enables the (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 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 number of metabolites. These substructures are then combined into a single data set with mass differences between fragment ions extracted from previously identified MS/MS spectra in a spectral library. We apply frequent pattern mining techniques to this data set to infer which substructures are associated with certain fragment mass differences. This results in a list of recommendations of the form:

mass difference md is associated with substructure s with frequency f and confidence c.

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 algorithm [5]. We will refer to the set of substructures created by BRICS fragmentation for metabolite

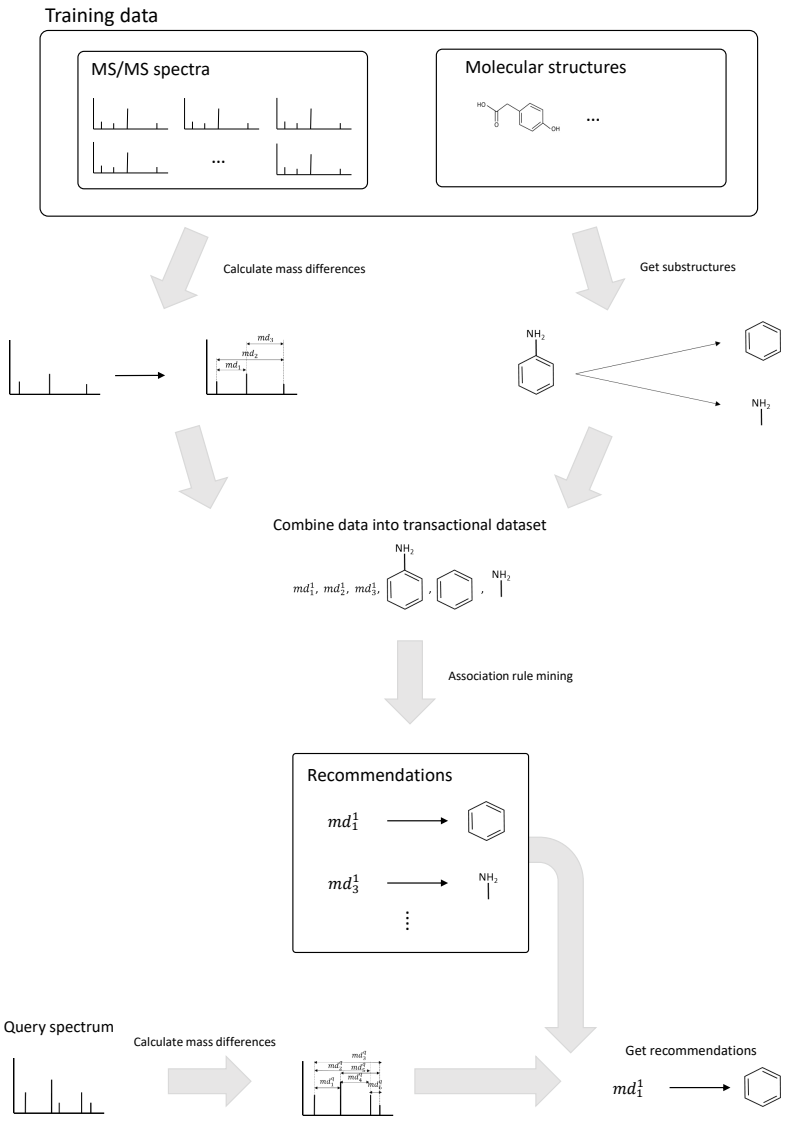

Figure 1: Pattern mining workflow to generate metabolite substructure recommendations. The training dataset consists of metabolites for which both an MS/MS spectra and its molecular substructure are known. For every metabolite a single transaction is created which consists of both spectral information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

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 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 dataset 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 mass differences. Frequent substructures are identical parts of a molecular structure that frequently occur in a given dataset. To perform pattern mining we require a transactional dataset, with each transaction a set of items (an itemset). In our approach, each item consists of a molecular substructure or a mass difference extracted from an MS/MS spectrum. A transaction is then the set of all substructures and mass differences for a single molecule. The support of an itemset is defined as the number of transactions in a dataset 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 concerned with finding all the association rules that are frequent and confident, i.e. identifying the association rules whose 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 mass differences $mdiff(m)$ and all molecular substructures $S(m)$: $T(m) = mdiff(m) \cup S(m)$. We will refer to the set of transactions $T(m)$ of all the metabolites $m \in M$ as $T(M)$. We only mine for patterns in the transactional dataset $T(M)$ that consist of both molecular substructures and 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 (md, s) are retained, with $md \in mdiff(m)$ 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 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 dataset $T(M)$ will result in a list of recommendations of the form:

mass difference md_i *can be associated with substructure* s_j *with support* f *and confidence* c.

Note that the 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 [18], 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) [20]. A total of 814 compounds for which both experimental MS/MS data and the molecular 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 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 to two transactional datasets (one dataset for each ionization mode) association rules were mined with support and confidence thresholds of 3 and 1% respectively, resulting in a total of 60,073 unique recommendations for positive ionization mode and 12,623 for negative ionization mode. Figure 2 shows some of the recommended substructures for the MS/MS spectrum of N(6)-carboxymethyllysine.

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 mass differences between the mass spectral peaks with intensity exceeding 5% of the most intense peak are calculated. These mass dif-

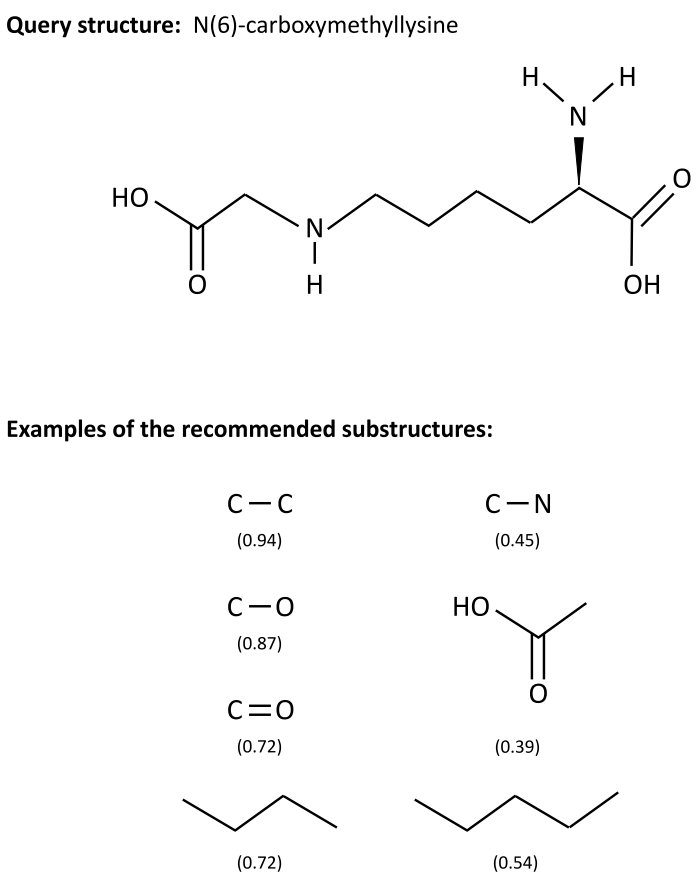

Figure 2: Query substructure of N(6)-carboxymethyllysine together with the examples of recommended substructures. The average confidence of the rules recommending the substructure (given that several different rules recommend the same substructure) is given in the brackets.

ferences are then checked against the association rule database and substructures are recommended based on rules containing the same mass differences. Taking into account possible neutral losses, we allowed a maximum mass difference of 2 Da. 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.

Using these substructure recommendations MS/MS spectrum identification results generated by third-party tools can be reranked. The reranking is done based on the similarity between the top 5 recommended substructures and the full metabolite structure predicted by the identification tool whose rankings we want to improve. 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 Recommendation 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) [16]. 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 in these two editions. We used MS/MS spectra

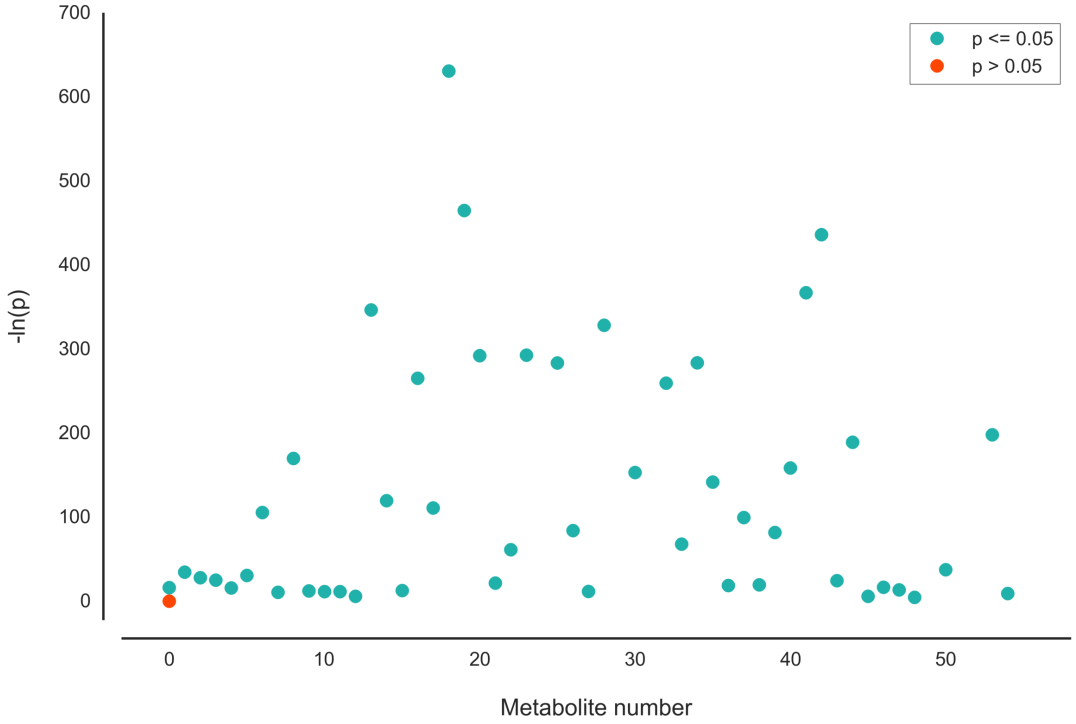

Figure 3: Enrichment of recommended substructures in the CASMI molecular structures. The P-values are given for each of the 58 molecules tested based on a Fisher's exact test. Green dots denote significant enrichment compared to random assignment, red dots denote non-significant enrichment.

from 58 metabolites in total from CASMI2014 and CASMI2016. Two compounds were excluded as they contained only one peak after preprocessing.

After obtaining recommended substructures for the 58 compounds, these recommendations were 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 3 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 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 on this confidence value for each CASMI metabolite. Figure 4 shows the mean receiver operator characteristic (ROC) curve for the CASMI data recommendations calculated from 58 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.

3.2 Partial matching of substructures provides additional insights

As previously mentioned, 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 partial matching used to generate substructure recommendations provides us with relevant additional information compared to full spectral matching we used the MassBank mass spectral database [9] to identify 58 metabolites of the CASMI data. MassBank returned matches for only 7 metabolites out of 58. 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.

3.3 Substructure recommendations improve existing prediction tools

As our method 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

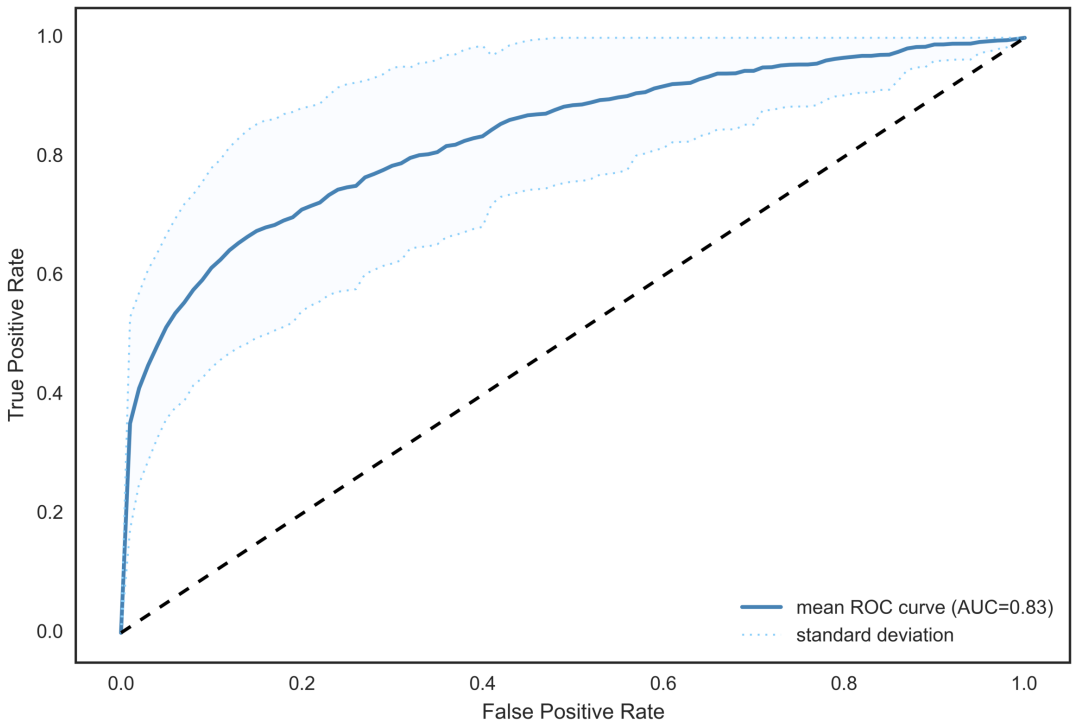

Figure 4: Mean ROC curve for CASMI metabolites. The ROC curve shows that our method can accurately distinguish substructures present in a metabolite from those that are not present based on the average confidence of the recommendations. It plots the true positive rate (TPR) as a function of the false positive rate (FPR) for different average confidence thresholds.

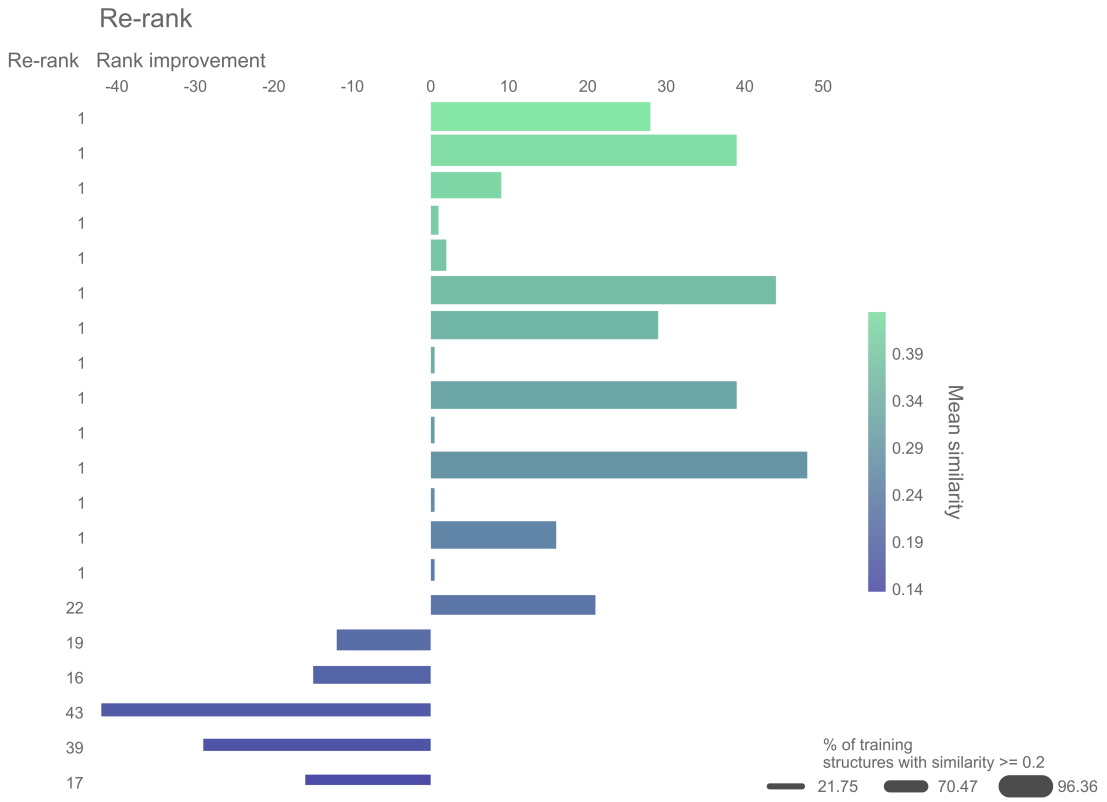

Figure 5: Identification improvements by reranking the MAGMa structure predictions. The horizontal axis shows the rank improvement, while the final rank is shown in the first column. Furthermore, for every CASMI metabolite the average similarity between the training dataset and that metabolite is shown, 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.

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 [15], which was the winner of CASMI2014. For the MAGMa structural search the PubChem database [11] was used, which contains all of the CASMI compounds. For each of the spectra in the CASMI dataset we retained the 50 highest ranked MAGMa predictions, which were subsequently reranked based on our recommended substructures.

Figure 5 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 dataset used to generate the recommendation rules. Although our approach does not require that a fully matching structure is present in the training dataset, 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 dataset used within this study consisted of less than one thousand metabolites. As a result, for substructures that are missing from the dataset, 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 dataset 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 and 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] Antoni Aguilar-Mogas, Marta Sales-Pardo, Miriam Navarro, Roger GuimerÃă, and Oscar 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. doi: 10.1021/acs.analchem.6b04512.

[3] Felicity Allen, Russ Greiner, and David Wishart. Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics*, 11(1):98–110, 2015. ISSN 1573-3890. doi: 10.1007/s11306-014-0676-4.

[4] Emily G. Armitage and Coral Barbas. Metabolomics in cancer biomarker discovery: Current trends and future perspectives. *Journal of Pharmaceutical and Biomedical Analysis*, 87:1 – 11, 2014. ISSN 0731-7085. doi: http://dx.doi.org/10.1016/j.jpba.2013.08.041. URL http://www.sciencedirect.com/science/article/pii/S0731708513004044. Review Papers on Pharmaceutical and Biomedical Analysis 2013.

[5] 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.

[6] Kai Dührkop, Huibin Shen, Marvin Meusel, Juho Rousu, and Sebastian Böcker. Searching molecular structure databases with tandem mass spectra us-

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

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

[8] 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.

[9] Hisayuki Horai, Masanori Arita, Shigehiko Kanaya, Yoshito Nihei, Tasuku Ikeda, Kazuhiro Suwa, Yuya Ojima, Kenichi Tanaka, Satoshi Tanaka, Ken Aoshima, Yoshiya Oda, Yuji Kakazu, Miyako Kusano, Takayuki Tohge, Fumio Matsuda, Yuji Sawada, Masami Yokota Hirai, Hiroki Nakanishi, Kazutaka Ikeda, Naoshige Akimoto, Takashi Maoka, Hiroki Takahashi, Takeshi Ara, Nozomu Sakurai, Hideyuki Suzuki, Daisuke Shibata, Steffen Neumann, Takashi Iida, Ken Tanaka, Kimito Funatsu, Fumito Matsuura, Tomoyoshi Soga, Ryo Taguchi, Kazuki Saito, and Takaaki 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. doi: 10.1002/jms.1777.

[10] Franziska Hufsky, Kerstin Scheubert, and Sebastian Böcker. Computational mass spectrometry for small-molecule fragmentation. *Trends in Analytical Chemistry*, 53: 41–48, 2014. doi: 10.1016/j.trac.2013.09.008.

[11] S. Kim, PA. 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.

[12] Liang Li, Ronghong Li, Jianjun Zhou, Azeret Zuniga, Avalyn E. Stanislaus, Yiman Wu, Tao Huan, Jiamin Zheng, Yi Shi, David S. Wishart, and Guohui Lin.

MyCompoundID: using an evidence-based metabolome library for metabolite identification. *Analytical chemistry*, 85(6):3401–3408, March 2013. ISSN 1520-6882. doi: 10.1021/ac400099b. URL http://dx.doi.org/10.1021/ac400099b.

[13] 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. *Brief Bioinform*, 16(2):216–231, March 2015.

[14] Gary J. Patti, Oscar Yanes, and Gary Siuzdak. Innovation: Metabolomics: the apogee of the omics trilogy. *Nat Rev Mol Cell Biol*, 13(4):263–269, 2012. doi: 10.1038/nrm3314.

[15] L. Ridder, 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 Commun Mass Spectrom*, 26(20):2461–2471, oct 2012.

[16] Emma L. Schymanski, Christoph Ruttkies, Martin Krauss, Céline Brouard, Tobias Kind, Kai Dührkop, Felicity Allen, Arpana Vaniya, Dries Verdegem, Sebastian Böcker, Juho Rousu, Huibin Shen, Hiroshi Tsugawa, Tanvir Sajed, Oliver Fiehn, Bart Ghesquière, and Steffen Neumann. Critical assessment of small molecule identification 2016: automated methods. *Journal of Cheminformatics*, 9(1):22, 2017. ISSN 1758-2946. doi: 10.1186/s13321-017-0207-1. URL http://dx.doi.org/10.1186/s13321-017-0207-1.

[17] Justin J.J. van der Hooft, Joe Wandy, Michael P. Barrett, Karl E.V. Burgess, and Simon 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, October 2016. URL http://eprints.gla.ac.uk/130100/.

[18] Erwan Werner, Jean-FranÃğois Heilier, CÃľline Ducruix, Eric Ezan, Christophe

Junot, and Jean-Claude 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. doi: http://dx.doi.org/10.1016/j.jchromb.2008.07.004. URL http://www.sciencedirect.com/science/article/pii/S1570023208005138. Hyphenated Techniques for Global Metabolite Profiling.

[19] David S. Wishart. Emerging applications of metabolomics in drug discovery and precision medicine. *Nat Rev Drug Discov*, 15(7):473–484, July 2016. ISSN 1474-1776. URL http://dx.doi.org/10.1038/nrd.2016.32.

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

[21] 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. URL http://www.biomedcentral.com/1471-2105/11/148.