Finding Characteristic Substructures for Metabolite Classes

Marcus Ludwig[1], Franziska Hufsky[1,2], Samy Elshamy[1], and Sebastian Böcker[1]

1 **Chair for Bioinformatics, Friedrich-Schiller-University, Jena, Germany,** {m.ludwig,franziska.hufsky,sebastian.boecker}@uni-jena.de
2 **Max Planck Institute for Chemical Ecology, Beutenberg Campus, Jena, Germany**

Abstract

We introduce a method for finding a characteristic substructure for a set of molecular structures. Different from common approaches, such as computing the maximum common subgraph, the resulting substructure does not have to be contained in its exact form in all input molecules. Our approach is part of the identification pipeline for unknown metabolites using fragmentation trees. Searching databases using fragmentation tree alignment results in hit lists containing compounds with large structural similarity to the unknown metabolite. The characteristic substructure of the molecules in the hit list may be a key structural element of the unknown compound and might be used as starting point for structure elucidation. We evaluate our method on different data sets and find that it retrieves essential substructures if the input lists are not too heterogeneous. We apply our method to predict structural elements for five unknown samples from Icelandic poppy.

1998 ACM Subject Classification J.2 Physical Sciences and Engineering (Chemistry)

Keywords and phrases metabolites, substructure prediction, mass spectrometry, FT-BLAST

Digital Object Identifier 10.4230/OASIcs.GCB.2012.23

1 Introduction

The rapidly developing field of metabolomics deals with the detection, identification and quantification of low molecular-weight compounds (typically below 1000 Da). All organisms synthesize many different metabolites and a large portion of them remain uncharacterized regarding their structure and function [25]. Metabolites cover a wide array of compound classes and, due to their physical and chemical properties, show large structural diversity. The analysis and identification of small molecules plays an important role in biomarker discovery, diagnostics, pharmaceutical chemistry, and functional genomics [9, 21, 41].

Currently, no single instrumental platform can analyze all metabolites. Mass spectrometry (MS) is a key technique to analyze small molecules. It is orders of magnitude more sensitive than nuclear magnetic resonance (NMR). MS enables high throughput experiments and the amount of data produced is hard to process and analyze manually [19]. Usually, a combination of a chromatography with a fragmentation technique is used to obtain information beyond the compound mass. Most common combinations are gas chromatography MS (GC-MS) using electron impact (EI) fragmentation and liquid chromatography MS (LC-MS) using collision-induced dissociation (CID) for fragmentation.

The identification of unknown compounds is a major bottleneck in the interpretation of metabolomics MS data. When the compound is unknown, comparison with library entries or spectra of known standards will result in imprecise or incorrect hits, or no hits at all [10, 20].

German Conference on Bioinformatics 2012 (GCB'12).
Editors: S. Böcker, F. Hufsky, K. Scheubert, J. Schleicher, S. Schuster; pp. 23–38

OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

For GC-MS spectra Demuth *et al.* [7] propose a method for finding similar compounds in a spectral library if the sample molecule is not contained. Further, methods for the identification of chemical substructures of the unknown sample molecule based on the direct comparison of fragmentation spectra have been proposed [17, 34, 37, 39]. For the automated analysis of LC-MS there are few computational pioneering studies [13, 14, 26, 42].

Besides classification of unknown compounds, not much progress has been made towards the *de novo* interpretation of mass spectra of small molecules, that cannot be found in any, not even a structural, database. When manually analyzing fragmentation mass spectra experts annotate fragmentation peaks and pathways to explain the data and determine the molecular structure. Böcker and Rasche [3] developed a method for the *de novo* interpretation of such spectra, resulting in hypothetical fragmentation trees. Fragmentation tree nodes are annotated with the molecular formula of the fragments, whereas edges represent fragmentation events, that is, neutral or radical losses. Here the computational analysis does not require any knowledge about compound structures or databases of compound structures or mass spectra. Only lists of common and implausible losses are required as expert knowledge about fragmentation mechanisms. Recently, methods to calculate fragmentation trees from multiple MS and GC-MS data have been developed [16, 32].

Rasche *et al.* [28] propose local tree alignments for the automated comparison of fragmentation trees. A local tree alignment contains those parts of the two trees where similar fragmentation cascades occurred. The authors could show, that this method is superior to spectral comparison in applications such as database searching. Fragmentation tree alignments even allow for inter-dataset comparisons for data sets measured on different instruments [28]. The authors present FT-BLAST, a database search tool for the identification of unknown metabolites based on fragmentation tree alignment. The received hit lists contain compounds with large structural similarity to the unknown metabolite. The common substructure of these compounds can be a starting point for its structural elucidation. A similar approach has been proposed by [31].

Finding the largest substructure common to a collection of graphs is denoted as maximum common subgraph (MCS) problem [29]. This problem is NP-hard even for two graphs [11]. There exists a rich literature in chemoinformatics and molecular modelling on this topic, often targeted at searching in molecular databases [27]. See Brown [5] for an introduction to chemoinformatics, and Raymond and Willet [29] for a review of exact and approximation algorithms for the maximum common subgraph problem. When computing the MCS for a set of molecules no deviation to the subgraphs in the molecules is allowed. This rather strict definition may not reflect the chemical similarity between compounds [36]. Due to their different physical and chemical properties, the structure of metabolites is heterogeneous even within the same compound classes. Finding the maximum common substructure to the hit lists received from FT-BLAST will often result in a very small structure that is not meaningful, or even in a single (carbon-) atom or an empty graph. An alternative approach is searching for frequent substructures in the molecule set. Frequent subgraph mining has been studied extensively in the last decade [4, 12, 15, 18, 23, 24, 43].

In this work, we loosen the strict definition of a common substructure that has to be contained in its exact form in either some or all of the input molecules. We rather try to find a *characteristic substructure* that is build of frequent (representative) substructures and reflects the specific features of the molecule set. We present a method to compute this structure and evaluate it on different compound classes and hit lists received from FT-BLAST. We find that our method is suitable to deal with structural outliers and retrieves characteristic substructures if the input lists are not too heterogeneous. We use this method to predict

substructures for unknown samples from Icelandic poppy.

In addition we developed *MoleculePuzzle* to combine this characteristic substructure with information from fragmentation trees, in silico fragmentation prediction [13, 26, 33, 42], rule based substructure prediction [17], or structural isomer generators [8]. This tool can help with the assembly of structural pieces to elucidate the structure of an unknown compound.

2 Molecule graphs

We model the chemical structures of molecules using an undirected, labeled graph $M = (V, E)$ with vertices V representing the atoms of the molecule, and edges E representing the covalent bonds. Vertices are labeled with the corresponding element. Edge labels consider single or multiple bonds. In the following, we call these graphs *molecule graphs*. Molecule graphs only reflect the topology of the molecule and do not indicate the geometric distance between a pair of atoms (vertices). As hydrogen atoms are of limited importance for elucidating the core structure of a molecule, they are ignored in the following.

Two graphs are *isomorphic* if there is a one-to-one correspondence between their vertices such that an edge exists between two vertices in one graph if and only if an edge exists between the two corresponding vertices in the other graph. A (vertex-) *induced subgraph* of M is a set of vertices $S \subseteq V$ and those edges from M that connect two vertices from S. A graph $M_{1,2}$ is a *common induced subgraph* of graphs M_1 and M_2 if $M_{1,2}$ is isomorphic to induced subgraphs of M_1 and M_2. A *maximum common induced subgraph* (MCIS) is the common induced subgraph with maximum number of vertices. The related *maximum common edge subgraph* (MCES) is a subgraph consisting of the largest number of edges common to both M_1 and M_2.

Since the structures of metabolites are heterogeneous, finding the MCS (either MCIS or MCES) will often result only in small subgraphs or even a single vertex. In this work we loosen the strict definition of an MCS and try to find a *characteristic substructure* CS, that is a graph reflecting the characteristics of a set of molecule graphs best.

3 Methods

To compute a *characteristic substructure* CS for a set $\mathcal{M}$ of m molecule graphs $M_1, \ldots, M_m$ we start with the computation of representative paths, since paths can be found fast in (molecule) graphs. The relative frequency $h_{\mathcal{M}}(p)$ of a path p is the number of molecule graphs $M_i \in \mathcal{M}$ that contain a path q that is isomorphic to p divided by the total number of molecule graphs. We call a path *representative* if it has a relative frequency above a certain threshold $h_t \in [0, 1]$. For the computation of representative paths edge labels are ignored.

A path p induces a subgraph $M(p)$ containing all edges from M that connect two vertices v_i, v_j from p. We call this subgraph *path structure* P. Note that two isomorphic paths do not necessarily have to induce two isomorphic subgraphs (see Figure 1). Path structures contain the information that is necessary to construct the characteristic substructure. The relative

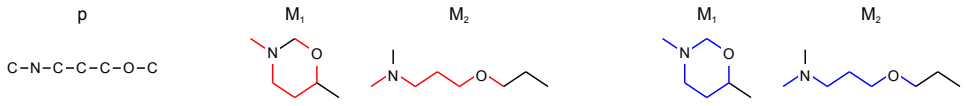

Figure 1 Molecules M_1 and M_2 both contain a path isomorphic to path p (red). The two isomorphic paths induce two subgraphs in M_1 and M_2 that are not isomorphic (blue).

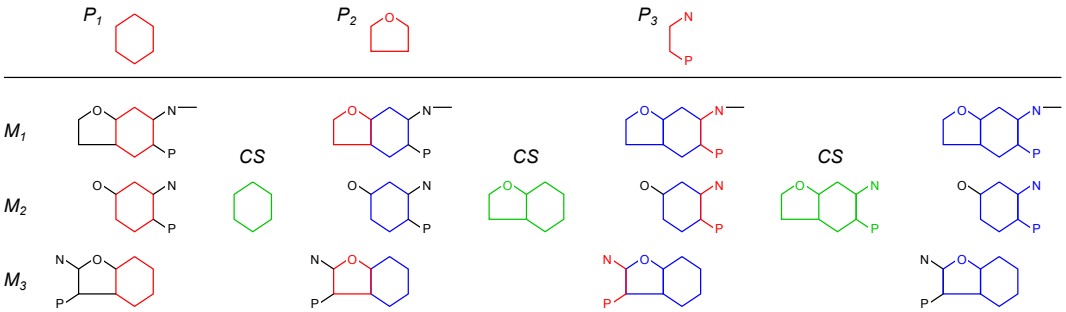

Figure 2 Adding path structures to the *characteristic substructure* CS. The figure is not a snapshot of a specific stage of the algorithm but rather an illustration of how to build up the CS from several path structures. C atoms are implicit. The drawn structures are not meant to depict real molecules. First, all subgraphs isomorphic to the first path structure P_1 are located in the molecule graphs (red) and P_1 is used as skeletal structure to build CS (green). We store the locations of the subgraphs (blue) to locate the subgraphs isomorphic to the next path structures. The second path structure P_2 is only isomorphic to subgraphs in M_1 and M_2 (red). Location of the subgraphs is the same in both molecules. Path structure P_2 is added to CS at this location. Path structure P_3 is isomorphic to subgraphs in all molecule graphs (red). The subgraphs in M_1 and M_2 are located at the same position while the subgraph in M_3 is located somewhere else. Path structure P_3 is added to CS at the most frequently occurring location and the subgraphs in M_1 and M_2 are marked (blue).

frequency $h_{\mathcal{M}}(P_i)$ of a path structure P_i is the number of molecule graphs $M_j \in \mathcal{M}$ that contain a subgraph that is isomorphic to P_i divided by the total number of molecule graphs. We call a path structure *representative* if it has a relative frequency above a certain threshold h_t. For the computation of representative path structures, edge labels are reconsidered.

For a certain length l we compute all representative paths of length l, find all representative path structures induced by these paths, and add these path structures to the characteristic substructure. We start with a certain length l_{start} decreasing it stepwise until we find a path structure that can be added to the characteristic substructure. Afterwards we increase the step size to s to skip path structures that add only few information to the characteristic substructure. We stop if $l < l_{\text{end}}$ since adding shorter paths seems to worsen the results.

Assume we have computed all representative path structures for a certain length l. We sort these path structures by their relative frequency in descending order $P_1, \ldots, P_n$. Lets assume that each path structure is isomorphic to at most one induced subgraph in each molecule graph. We start building the *characteristic substructure* CS with the most frequent path structure P_1 using it as skeletal structure. After P_1 is added to CS we remember the locations of all subgraphs within the molecule graphs that are isomorphic to P_1 (see Figure 2).

By definition path structure P_i is a common induced subgraph of at least h_t molecule graphs $M_j \in \mathcal{M}$. To add P_i to CS we have to compare the locations of these subgraphs with the location of the subgraphs of recently added $P_1, \ldots, P_{i-1}$. Comparison can be done easy, since we have stored the locations of all recently added path structures in the molecule graphs. Path structure P_i gets the location in CS that most of the subgraphs have within the molecules. Again, we memorize this subgraph location for the molecules in which it occurs (see Figure 2).

If a path structure P_i is isomorphic to more than one subgraph in a molecule graph we want to add all subgraphs instead of choosing one. If at least $k \geq 1$ isomorphic subgraphs within the same molecule graph in at least $|\mathcal{M}| \cdot h_{\text{iso}}$ molecule graphs occur, we proceed as

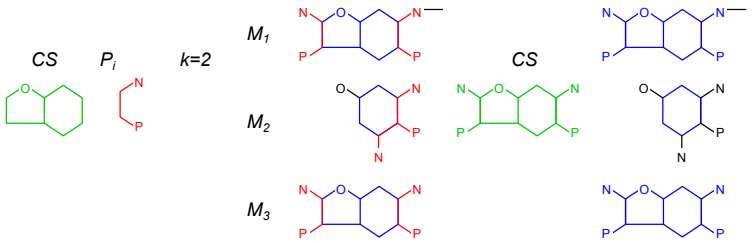

Figure 3 Adding a path structures that is isomorphic to $k = 2$ subgraphs in all molecule graphs. C atoms are implicit. The drawn structures are not meant to depict real molecules. In all molecules we consider the two subgraphs isomorphic to P_i as a single subgraph (red). This subgraph is disconnected in M_1 and M_3, and connected in M_2. We add the most occurring subgraph to CS (green) and mark its location in M_1 and M_3 (blue).

follows: In all molecules that contain exactly k subgraphs isomorphic to P_i we consider these k subgraphs as a single subgraph. This subgraph is either connected or disconnected. We compare the location of these subgraphs and add the most occurring to CS as described above (see Figure 3).

4 MoleculePuzzle

Computing a *characteristic substructure* for a set of molecules that have presumably large structural similarity to an unknown compound will greatly support its structure elucidation. In combination with the information from the fragmentation tree this substructure can be used to "puzzle" the molecular structure of the unknown compound. If need be, molecular isomer generators [8] can be used to enumerate all structural isomers of the annotated fragment formulas.

We present a puzzle tool to assemble molecular structures (see Figure 7 in the Appendix). *MoleculePuzzle* is a JAVA based plugin for the SIRIUS[2] framework[1] [2]. It is based on JChemPaint [22] which provides an interface that allows to draw chemical compounds. JChemPaint already contains templates for ring structures and different bond types and allows atom coloring, different render settings, as well as loading and storing structures from and into various file formats. In addition, it offers automated highlighting of chemically incorrect structures, which simplifies structure assembly.

We extend JChemPaint to work with a list of molecular structures which are transformed into puzzle pieces. These pieces can be added to the painting panel to modify the structures and fit them together. JChemPaint's drag and drop feature indicates when it is possible to connect one structure to another. *MoleculePuzzle* helps to assemble several structural pieces to a full structure.

5 Experimental results and discussion

We evaluate our method on three different data sets: First, we compute characteristic substructures for molecules from the same compound class and compare these structures to all molecules within this class. Second, we compute the characteristic substructures for molecules in the hit lists received from the FT-BLAST search tool for a reference data set

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

and compare them to the structure of the query compound. Third, we use our method on FT-BLAST hit lists for structure prediction of unknowns from Icelandic poppy.

To evaluate the quality of the computed substructures we use binary fingerprint representations to compute Tanimoto similarity scores (Jaccard indices) [30] and Tversky similarity scores [38]. We use fingerprints of the PubChem database [40] in the Chemistry Development Toolkit version 1.3.7 [35] for our computations. To compute the asymmetric Tversky similarity we weight the features only occurring in the CS with 1 and these only occurring in input molecules with 0. Tversky similarity gives an indication whether the CS is a substructure of the input molecules. Tanimoto similarity indicates whether the CS reconstructs important features of the input list.

In the following, we choose $l_{\text{start}} = 20$, $l_{\text{end}} = 5$, $s = 4$, $h_{\text{iso}} = 0.8$, $h_t = 0.8$. We implemented the algorithm in Java and carried out all computations on a quad-core Intel® Core™ i7-820QM with 8 GB RAM under Windows 7 operating system.

5.1 Characteristic substructures for compound classes

This data set consists of 395 molecules from 15 different compound classes (based on MeSH[2] categories) downloaded from PubChem Compound[3] (see Table 3 in the Appendix for CIDs). For each compound class, we compute the characteristic substructure (see Figure 8 in the Appendix). Computation required 6.4 s on average.

We calculate a score Tanimoto$_{CS}$ that is the average of Tanimoto scores of all input molecules to the CS (see Table 1). To report the average Tanimoto structural similarity score of a class, we calculate pairwise Tanimoto scores for all molecules in the class. Tanimoto$_{\text{class}}$ is the average of all these pairwise Tanimoto scores. This score may be seen as an upper bound for Tanimoto$_{CS}$. Tversky$_{CS}$ is the average of Tversky scores of all input molecules to the CS.

Table 1 Quality of the characteristic substructures for 15 metabolite classes. Tanimoto$_{\text{class}}$ is the average of all pairwise Tanimoto scores within the compound class. Tanimoto$_{CS}$ (Tversky$_{CS}$) is the average Tanimoto (Tversky) score of all input molecules to the CS.

compound class	# of molecules	Tanimoto$_{\text{class}}$	Tanimoto$_{CS}$	Tversky$_{CS}$	running time (s)
2-acetylaminofluorenes	12	0.88	0.82	0.99	1.6
adenines	44	0.81	0.71	0.97	8.1
benzothiadiazines	25	0.64	0.54	0.99	4.3
chlorothiazides	24	0.58	0.51	0.98	7.1
cytosines	22	0.62	0.45	0.98	1.5
erythromycins	24	0.77	0.53	0.99	8.0
glucosinolates	66	0.67	0.52	1	4.9
guanines	18	0.62	0.26	0.97	2.7
lipids	22	0.92	0.37	1	28.0
neuraminic acids	14	0.77	0.57	0.97	1.8
peroxides	24	0.59	0.18	0.98	4.7
pregnadienes	26	0.79	0.68	0.99	12.5
thymines	24	0.66	0.49	1	0.9
trichothecenes	26	0.86	0.74	1	8.9
uraciles	24	0.56	0.46	1	0.7

[2] http://www.nlm.nih.gov/mesh/
[3] http://www.ncbi.nlm.nih.gov/pccompound

For all compound classes, Tversky_{CS} is at least 0.97, indicating that the CS is contained in all molecules to a large extend. For 9 compound classes, Tanimoto_{CS} is above 0.5. We argue that $\text{Tanimoto}_{CS} > 0.5$ indicates good quality of the result, as large parts of the input structures have been reconstructed. From the remaining, we found that for three classes the difference between Tanimoto_{CS} to $\text{Tanimoto}_{\text{class}}$ is less than 0.2, so the classes already seem to be very heterogeneous and therefore the resulting substructures are only small. For the other three compound classes the difference is larger than 0.2: For guanines our method only found the 6-membered-ring but unfortunately the 5-membered-ring is missing. Thus Tanimoto_{CS} is low (0.26). Lipids are typically build of several 6-membered-rings and several carbon sidechains of different lengths. Due to the different number of rings and sidechains, our method only reconstructs a single 6-membered-ring with the typical oxygen and phosphor structure and a carbon sidechain of eleven atoms connected via a nitrogen. We argue, that this is the main substructure which can be duplicated and expanded (for example using the *MoleculePuzzle* tool). Our method works worst for the class of peroxides, which we found to be structurally more heterogeneous (at least for our method) than the Tanimoto score suggests.

For erythromycins the difference between Tanimoto_{CS} and $\text{Tanimoto}_{\text{class}}$ is also above 0.2. Erythromycins are huge compounds (56 heavy atoms on average) all containing a macrocycle (14 atoms). Our method reconstructs this ring correctly but has problems with the different sidechains, resulting in a somewhat cluttered side structure. Nevertheless, Tanimoto_{CS} is still above 0.5 and Tversky_{CS} is 0.99.

5.2 Characteristic substructures for FT-BLAST hit lists

We compute the characteristic substructures for hit lists reported by FT-BLAST [28]. Rasche *et al.* [28] evaluated their method on 97 compounds measured on an Orbitrap XL instrument using Collision Induced Dissociation (CID) or High-energy Collision Dissociation (HCD) for fragmentation. They computed fragmentation trees for all compounds and carried out a *leave-one-out* FT-BLAST search. For each compound they removed the correct answer from the database and searched for the compound in the remainder. They reported hits up to a False Discovery Rate (FDR) of 30%. For the 60 reported lists with at least two hits we compute characteristic substructures and compare them to the structure of the query compound. We also use the hit lists with FDR 20 % (57 lists with at least two hits) and 10 % (56 lists with at least two hits).

To report the average structural similarity of the hits returned by FT-BLAST, we calculate the Tanimoto score of the query compound to each hit list entry. $\text{Tanimoto}_{\text{hitlist}}$ is the

Table 2 Quality of the characteristic substructures computed for FT-BLAST hit lists. $\text{Tanimoto}_{\text{hitlist}}$ is the average of all Tanimoto scores of the query compound to the hit list entries. Tanimoto_{CS} (Tversky_{CS}) is the Tanimoto (Tversky) score of the query compound to the CS. We average $\text{Tanimoto}_{\text{hitlist}}$ over all hit lists and Tanimoto_{CS} (Tversky_{CS}) over all non empty substructures.

FDR	nr of hit lists	average hit list length	average $\text{Tanimoto}_{\text{hitlist}}$	empty CS	average Tanimoto_{CS}	average Tversky_{CS}	running time (s)
30 %	60	11	0.76	9	0.49	1	1.4
20 %	57	9	0.77	5	0.50	0.99	1.1
10 %	56	8	0.79	4	0.50	0.99	0.9

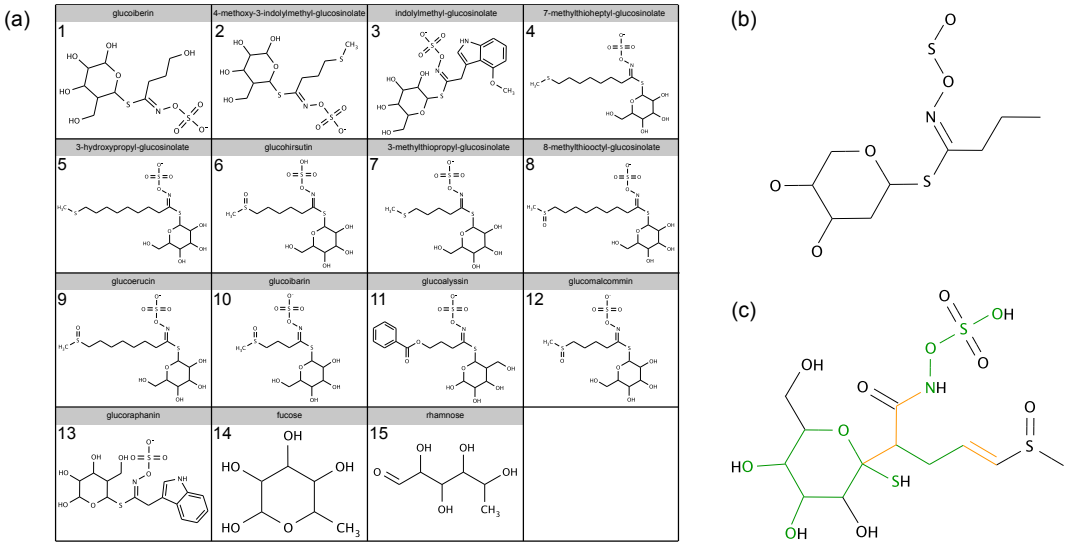

Figure 4 Characteristic substructure for the FT-BLAST hit list of glucoraphenin (CID 6443008). (a) The FT-BLAST hit list with FDR 30 % contains several glucosinolates (1-13) and two sugars (14,15). Computing the maximum common substructure would result in a very small structure. The characteristic substructure computed by our method (b) is contained in the query compound (c) (green) with slight variation (orange). Note that H atoms are ignored by our method.

average over all these scores. Again, this is somewhat an upper bound for the Tanimoto score between the query compound and the characteristic substructure (Tanimoto$_{CS}$). The average Tanimoto$_{\text{hitlist}}$ similarity for the complete dataset, using the leave-one-out strategy described above, is 0.76 for an FDR of 30 % and increases slightly with decreasing FDR to 0.79 (see Table 2).

For each hit list, we compute the characteristic substructure (see Table 2 for an overview of the results). Computation required on average 1.4 s for the hit lists with FDR 30 % (average hit list length 11), 1.1 s for hit lists with FDR 20 % (average hit list length 9), and 0.9 s for hit lists with FDR 10 % (average hit list length 8). We calculate Tanimoto$_{CS}$ (Tversky$_{CS}$), that is the Tanimoto (Tversky) similarity between the query compound and the characteristic substructure. For the hit lists with FDR 10 % and 20 % Tanimoto$_{CS}$ was 0.5 on average and for FDR 30 % it was about the same, namely 0.49. Tversky$_{CS}$ was 0.99 for 10 % and 20 % FDR and even increased to 1.0 for FDR 30 %. Since using the hit lists with smaller FDR seems not to improve the results significantly, we use the FDR 30 % hit lists reported by Rasche *et al.* [28] for further evaluation (see Table 2 in [28]).

The method works best if many compounds of the same class are included in the hit lists (as for sugars, zeatins and glucosinolates). Nevertheless, the method also computes good results even if some outliers are contained in the list. For example the hit list for an anthocyanin (CID 44256805) contains benzopyrans, amino acids, anthocyanins and one sugar. The method finds the main component, a 6-membered-ring with an oxygen and a carbon side atom. This ring is contained several times in the structure of the query compound. Another example is glucoraphenin (see Figure 4). Computing the maximum common substructure would have result in a very small, as fucose and rhamnose do not share the large structure with the remaining hits. Our method computes a characteristic substructure which is close to the query compound.

Very heterogeneous hit lists are hard to process. For 9 query compounds the characteristic

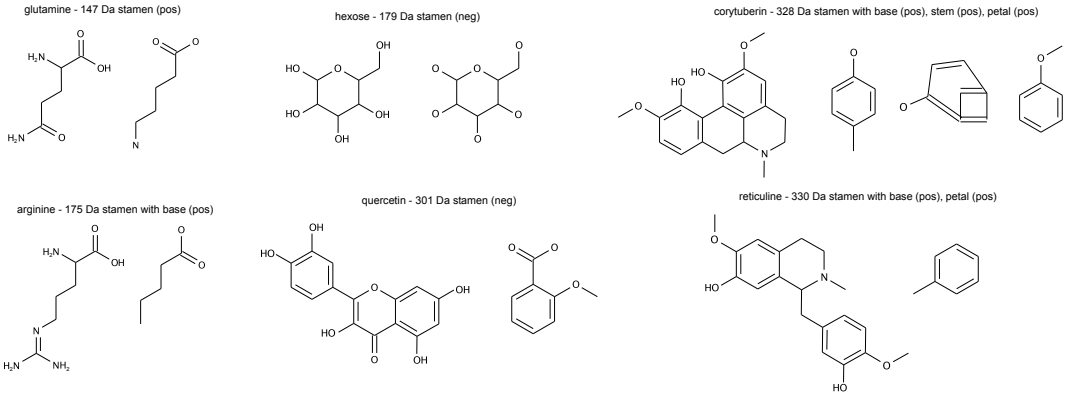

Figure 5 Resulting characteristic substructures for the FT-BLAST hit lists for the six compounds from Icelandic poppy that were manually identified. For each sample, the query compound (left) is compared to the resulting substructure(s) (right). For corytuberin the hit lists for the different samples from stamen, stem and petal supply different characteristic substructures that could be further combined with the *MoleculePuzzle* tool. For reticuline the samples from stamen and petal result in the same substructure, while no substructure was found for the stem sample. Processing the hit lists of the samples for the unspecified palmatine derivates (370 Da and 386 Da) results in the same substructure as for reticuline. Note that H atoms are ignored by our method.

substructure was empty. For other compounds (for example bergapten), the characteristic substructure looks somewhat cluttered.

5.3 Characteristic substructures for unknowns from Icelandic poppy

As a real-world example Rasche *et al.* [28] tried to identify unknown metabolites from Icelandic poppy (*P. nudicaule*). They computed fragmentation trees for 32 compounds and compared them with the Orbitrap data set as reference using FT-BLAST. Again, they reported hits up to an FDR of 30 %. We use this lists, since smaller FDRs seem not to improve the results for the previous data set.

Eight compounds from the dataset were manually identified by Rasche *et al.* [28]. For arginine, glutamine, quercetin and a hexose the correct compound was in the hit list. FT-BLAST results for reticuline (330.17 Da) and corytuberine (328.15 Da) included chemical precursors of these alkaloids. Two other unknowns (370 and 386 Da) were manually classified as palmatine-derivatives. For all these compounds our method predicts correct substructures (see Figure 5).

We use our method to predict substructures for the remaining unknown compounds in this data set. For one compound only one hit is received and substructure prediction is not necessary. For seven compounds the hit lists are very heterogeneous ($\text{Tanimoto}_{\text{hitlist}}=0$) and our method finds no characteristic substructure. This can be attributed to the small database (97 compounds) we are searching in. For the remaining five samples we compute characteristic substructures (see Figure 6) that can be used for further structure elucidation of the compounds.

6 Conclusion

We have developed a method for computing a characteristic substructure for a set of molecules. Different from the maximum common substructure this substructure does not have to be

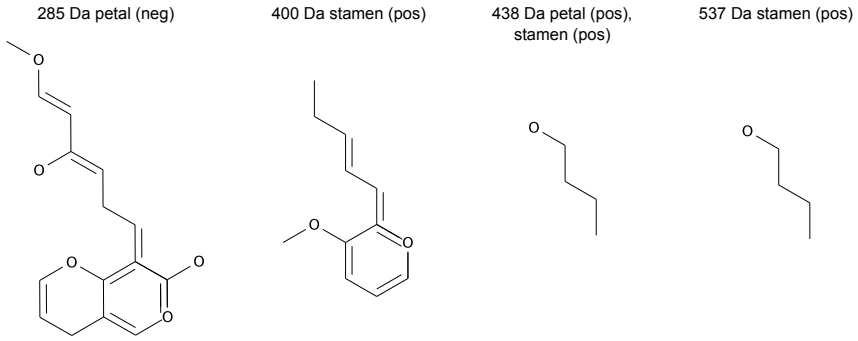

Figure 6 Substructure prediction for five samples from Icelandic poppy that are unknown. FT-BLAST hit list for 285 Da petal (neg) contains six compounds with $\text{Tanimoto}_{\text{hitlist}} = 0.6$. The hit list for 400 Da stamen (pos) contains six compound with $\text{Tanimoto}_{\text{hitlist}} = 0.4$. For both compounds we compute large characteristic substructures. The hit lists for 438 Da petal (pos) and 438 Da stamen (pos) both contain ten compounds with $\text{Tanimoto}_{\text{hitlist}} = 0.1$. FT-BLAST hit list for 537 Da stamen (pos) contains twelve compounds with $\text{Tanimoto}_{\text{hitlist}} = 0.1$. The resulting characteristic substructures for these three samples are rather small.

contained in its exact form in all input molecules. Finding characteristic substructures is an important step in the identification of unknown metabolites. It is part of a pipeline which is based on the computation of fragmentation trees from mass spectral data. Fragmentation tree alignment allows to find similar, not necessarily identical, compounds in a library search. Characteristic substructures of these hits may be key structural elements of the unknown metabolite.

We have evaluated our method on classes of molecules and FT-BLAST hit lists. We found that our method reconstructs many structural features contained in the input lists. Different from finding the maximum common substructure our method can deal with structural irregularities. Nevertheless, if the input lists are to heterogeneous no characteristic substructure can be predicted. We used the Tanimoto score for estimating structural homogeneity, but by visual inspection we found that some input lists are structurally more heterogeneous (at least for our method) than the Tanimoto score suggests.

We have predicted substructures for five unknown samples from Icelandic poppy. Substructure prediction is strongly dependent on the FT-BLAST hit lists. The reference data set for FT-BLAST used in the study of Rasche *et al.* [28] was pretty small. The resulting hit lists will become more homogeneous as more reference compounds become available, and we expect that the predicted substructures will be better and probably larger.

The presented method is not taking atom valences into account, but rather adopts the bond order from the input molecules. Bond orders can be corrected using automated bond order assignment [1,6]. The characteristic substructure and information from fragment formulas can be assembled in our *MoleculePuzzle* tool with results from in silico fragmentation prediction [13, 26,33,42], rule based substructure prediction [17], or structural isomer generators [8]. Together, the two presented tools form an important step towards the structure elucidation of unknown compounds.

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

References

1 Sebastian Böcker, Quang Bao Anh Bui, and Anke Truss. Computing bond orders in molecule graphs. *Theoretical Computer Science*, 412(12–14):1184–1195, 2011.

2 Sebastian Böcker, Matthias Letzel, Zsuzsanna Lipták, and Anton Pervukhin. SIRIUS: Decomposing isotope patterns for metabolite identification. *Bioinformatics*, 25(2):218–224, 2009.

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

4 Christian Borgelt and Michael R. Berthold. Mining molecular fragments: Finding relevant substructures of molecules. In *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM 2002)*, page 51. IEEE Computer Society, 2002.

5 Nathan Brown. Chemoinformatics — an introduction for computer scientists. *ACM Comput Surv*, 41(2):8, 2009.

6 Anna Katharina Dehof, Alexander Rurainski, Quang Bao Anh Bui, Sebastian Böcker, Hans-Peter Lenhof, and Andreas Hildebrandt. Automated bond order assignment as an optimization problem. *Bioinformatics*, 27(5):619–625, 2011.

7 W. Demuth, M. Karlovits, and K. Varmuza. Spectral similarity versus structural similarity: mass spectrometry. *Anal Chim Acta*, 516(1-2):75–85, 2004.

8 Jean-Loup Faulon, Donald P. Visco, and Diana Roe. Enumerating molecules. In Kenny B. Lipkowitz, Raima Larter, and Thomas R. Cundari, editors, *Reviews in Computational Chemistry*, volume 21, chapter 3, pages 209–286. John Wiley & Sons, Inc., 2005.

9 O. Fiehn, J. Kopka, P. Dörmann, T. Altmann, R. N. Trethewey, and L. Willmitzer. Metabolite profiling for plant functional genomics. *Nat Biotechnol*, 18(11):1157–1161, 2000.

10 Oliver Fiehn. Extending the breadth of metabolite profiling by gas chromatography coupled to mass spectrometry. *Trends Analyt Chem*, 27(3):261–269, 2008.

11 Michael R. Garey and David S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. 1979.

12 Ehud Gudes, Solomon Eyal Shimony, and Natalia Vanetik. Discovering frequent graph patterns using disjoint paths. *IEEE Trans Knowl Data En*, 18(11):1441–1456, 2006.

13 Markus Heinonen, Ari Rantanen, Taneli Mielikäinen, Juha Kokkonen, Jari Kiuru, Raimo A Ketola, and Juho Rousu. FiD: a software for ab initio structural identification of product ions from tandem mass spectrometric data. *Rapid Commun Mass Spectrom*, 22(19):3043–3052, 2008.

14 Dennis W. Hill, Tzipporah M. Kertesz, Dan Fontaine, Robert Friedman, and David F. Grant. Mass spectral metabonomics beyond elemental formula: Chemical database querying by matching experimental with computational fragmentation spectra. *Anal Chem*, 80(14):5574–5582, 2008.

15 Jun Huan, Wei Wang, and Jan Prins. Efficient mining of frequent subgraphs in the presence of isomorphism. In *Proceedings of the Third IEEE International Conference on Data Mining (ICDM 2003)*, page 549, 2003.

16 Franziska Hufsky, Martin Rempt, Florian Rasche, Georg Pohnert, and Sebastian Böcker. De novo analysis of electron impact mass spectra using fragmentation trees. *Anal Chim Acta*, 739:67–76, 2012.

17 Jan Hummel, Nadine Strehmel, Joachim Selbig, Dirk Walther, and Joachim Kopka. Decision tree supported substructure prediction of metabolites from GC-MS profiles. *Metabolomics*, 6(2):322–333, 2010.

18 Akihiro Inokuchi, Takashi Washio, and Hiroshi Motoda. An apriori-based algorithm for mining frequent substructures from graph data. In Jan Komorowski Djamel A. Zighed and Jan Zytkow, editors, *Proceedings of the 4th European Conference on Principles of Data*

Mining and Knowledge Discovery (PKDD 2000), volume 1910 of *Lect Notes Comput Sci*, pages 13–23. Springer, Berlin, 2000.

19 Haleem J Issaq, Que N Van, Timothy J Waybright, Gary M Muschik, and Timothy D Veenstra. Analytical and statistical approaches to metabolomics research. *J Sep Sci*, 32(13):2183–2199, 2009.

20 Tobias Kind and Oliver Fiehn. Advances in structure elucidation of small molecules using mass spectrometry. *Bioanal Rev*, 2(1-4):23–60, 2010.

21 Sing Teang Kong, Hai-Shu Lin, Jianhong Ching, and Paul C Ho. Evaluation of dried blood spots as sample matrix for gas chromatography/mass spectrometry based metabolomic profiling. *Anal Chem*, 83(11):4314–4318, 2011.

22 Stefan Krause, Egon Willighagen, and Christoph Steinbeck. JChemPaint - using the collaborative forces of the internet to develop a free editor for 2D chemical structures. *Molecules*, 5(1):93–98, 2000.

23 Michihiro Kuramochi and George Karypis. An efficient algorithm for discovering frequent subgraphs. *IEEE Trans Knowl Data En*, 16(9):1038–1051, 2004.

24 Siegfried Nijssen and Joost N. Kok. The Gaston tool for frequent subgraph mining. *Electron Notes Theor Comput Sci*, 127(1):77–87, 2005.

25 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.

26 Anna Pelander, Elli Tyrkkö, and Ilkka Ojanperä. In silico methods for predicting metabolism and mass fragmentation applied to quetiapine in liquid chromatography/time-of-flight mass spectrometry urine drug screening. *Rapid Commun Mass Spectrom*, 23(4):506–514, 2009.

27 Syed Asad Rahman, Matthew Bashton, Gemma L Holliday, Rainer Schrader, and Janet M Thornton. Small molecule subgraph detector (SMSD) toolkit. *J Cheminform*, 1(1):12, 2009.

28 Florian Rasche, Kerstin Scheubert, Franziska Hufsky, Thomas Zichner, Marco Kai, Aleš Svatoš, and Sebastian Böcker. Identifying the unknowns by aligning fragmentation trees. *Anal Chem*, 84(7):3417–3426, 2012.

29 John W Raymond and Peter Willett. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J Comput Aided Mol Des*, 16(7):521–533, 2002.

30 D. J. Rogers and T. T. Tanimoto. A computer program for classifying plants. *Science*, 132(3434):1115–1118, 1960.

31 Miguel Rojas-Cherto, Julio E. Peironcely, Piotr T. Kasper, Justin Johan Jozias van der Hooft, Ric C. H. De Vos, Rob J. Vreeken, Thomas Hankemeier, and Theo Reijmers. Metabolite identification using automated comparison of high resolution ms(n) spectral trees. *Anal Chem*, 84(13):5524–5534, 2012.

32 Kerstin Scheubert, Franziska Hufsky, Florian Rasche, and Sebastian Böcker. Computing fragmentation trees from metabolite multiple mass spectrometry data. In *Proc. of Research in Computational Molecular Biology (RECOMB 2011)*, volume 6577 of *Lect Notes Comput Sci*, pages 377–391. Springer, Berlin, 2011.

33 Emma Louise Schymanski, Christine M J Gallampois, Martin Krauss, Markus Meringer, Steffen Neumann, Tobias Schulze, Sebastian Wolf, and Werner Brack. Consensus structure elucidation combining GC/EI-MS, structure generation and calculated properties. *Anal Chem*, 84(7):3287–3295, 2012.

34 Stephen Stein. Chemical substructure identification by mass spectral library searching. *J Am Soc Mass Spectrom*, 6:644–655, 1995.

35 Christoph Steinbeck, Christian Hoppe, Stefan Kuhn, Matteo Floris, Rajarshi Guha, and Egon L Willighagen. Recent developments of the Chemistry Development Kit (CDK) - an open-source java library for chemo- and bioinformatics. *Curr Pharm Des*, 12(17):2111–2120, 2006.

36 Yoshimasa Takahashi, Yuzuru Satoh, Hajime Suzuki, and Shin-ichi Sasaki. Recognition of largest common structural fragment among a variety of chemical structures. *Anal Sci*, 3:23–28, 1987.

37 Hiroshi Tsugawa, Yuki Tsujimoto, Masanori Arita, Takeshi Bamba, and Eiichiro Fukusaki. GC/MS based metabolomics: development of a data mining system for metabolite identification by using soft independent modeling of class analogy (SIMCA). *BMC Bioinformatics*, 12:131, 2011.

38 Amos Tversky. Features of similarity. *Psychol Rev*, 84(4):327–352, 1977.

39 K. Varmuza and W. Werther. Mass spectral classifiers for supporting systematic structure elucidation. *J Chem Inf Comp Sci*, 36(2):323–333, 1996.

40 Yanli Wang, Jewen Xiao, Tugba O Suzek, Jian Zhang, Jiyao Wang, and Stephen H Bryant. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res*, 37(Web Server issue):W623–W633, 2009.

41 David S Wishart. Current progress in computational metabolomics. *Brief Bioinform*, 8(5):279–293, 2007.

42 Sebastian Wolf, Stephan Schmidt, Matthias Müller-Hannemann, and Steffen Neumann. In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics*, 11:148, 2010.

43 Xifeng Yan and Jiawei Han. gSpan: Graph-based substructure pattern mining. In *Proceedings of the Second IEEE International Conference on Data Mining (ICDM 2002)*, pages 721–724, 2002.

A Appendix

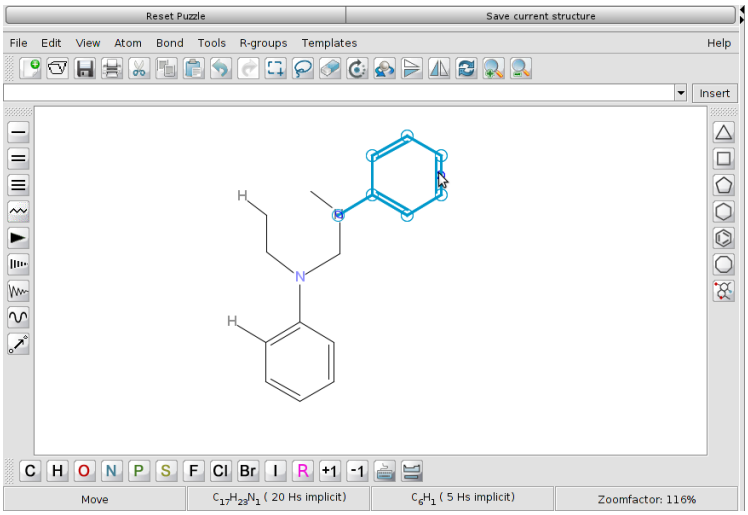

Figure 7 Screenshot of the *MoleculePuzzle* plugin for the SIRIUS[2] framework. Several structural pieces can be puzzled together to receive the full structure of a molecule.

Table 3 PubChem CIDs for the 395 molecules from 15 compound classes.

compound class	# molecules	PubChem CIDs
glucosinolates	66	23682211, 25244590, 46173878, 9548633, 9548621, 9576240, 9576416, 656498, 9548605, 656539, 656538, 9576241, 46173877, 44237373, 656547, 46173876, 44237368, 5281133, 656555, 656557, 46173880, 44237257, 656543, 656523, 44237260, 656527, 6325242, 9548892, 25244892, 441524, 46173879, 6442557, 5281135, 656568, 46173882, 9576738, 46173875, 656545, 656525, 7098673, 17756749, 656541, 656531, 46173881, 44237206, 5281136, 44237203, 9548619, 5281138, 46173884, 25245521, 6443008, 5281134, 656537, 25244538, 25244201, 17756744, 5281139, 46173883, 25246161, 25245774, 25244220, 25243874, 9548618, 656562, 656548
adenines	44	3083432, 3083316, 3082029, 3081390, 3080770, 3080762, 3036950, 703739, 466837, 465383, 440867, 25246029, 15938965, 12358355, 7059571, 7058055, 41211, 34768, 32014, 10238, 9687, 6083, 6076, 1913, 224, 50909893, 44134557, 25244014, 25201135, 23615303, 23615194, 23421209, 22848660, 16078938, 9578273, 9589376, 6992262, 6452236, 6449870, 6426627, 5748329, 5491933, 5399013, 4617095
cytosines	22	455597, 5276954, 597, 452713, 441224, 374908, 492030, 492031, 500131, 6473860, 471292, 477169, 477168, 467421, 455604, 455603, 455602, 455598, 16727509, 477170, 455605, 455601
guanines	18	764, 374910, 160219, 129161, 133387, 161069, 145817, 129136, 406591, 130450, 478537, 25082899, 471293, 485625, 485629, 485626, 485624, 195385

Continued on next page

Table 3 – continued from previous page

compound class	# molecules	PubChem CIDs
thymines	24	1135, 452067, 451954, 6439704, 135557, 6450954, 452068, 452063, 452715, 607039, 196362, 6477693, 370631, 406592, 6477695, 6477692, 492029, 465896, 445213, 495405, 374907, 457298, 485384, 452946
uraciles	24	1174, 6194, 18323, 9412, 5386, 68216, 13712, 5899, 5360852, 5282192, 688297, 6971263, 6029, 69672, 13268, 208432, 453162, 1177, 452948, 125110, 3067786, 55281, 456513, 54929
lipids	22	440885, 46173186, 46173313, 50909837, 46173444, 25246183, 25246208, 25200834, 46173394, 25246224, 3083382, 25202130, 25244473, 51351771, 51351791, 46173153, 50909852, 46173409, 127960, 13831140, 440886, 160295
benzothiadiazines	25	62940, 43148, 22425, 2910, 2348, 2343, 2122, 12933, 5560, 4870, 4121, 44154271, 44152755, 44151672, 6441852, 871720, 216293, 198367, 194167, 188359, 174783, 173791, 107748, 72070, 71652
trichothecenes	26	11968047, 11968045, 6540635, 6450461, 6444304, 6438947, 6438478, 6437354, 6437353, 45266518, 11969549, 6431315, 6321400, 5459303, 5284461, 3000635, 529495, 442403, 442400, 30552, 50987470, 50986319, 44144558, 44144549, 44144548, 11969469
chlorothiazides	24	127085, 116034, 107748, 71656, 62940, 2348, 2720, 5560, 3639, 50987261, 44151672, 44147212, 24847808, 23717274, 11354874, 3083286, 3083063, 242921, 216293, 193444, 188359, 174783, 172393, 159328
erythromycins	24	5284534, 3033819, 447043, 429694, 84029, 55185, 24847865, 17753754, 5748242, 5282045, 3002190, 44629874, 16212992, 6713919, 6426643, 83935, 83933, 12560, 44629879, 25102720, 17753750, 11969952, 9604450, 6915744
peroxides	24	5497123, 5464098, 641668, 637882, 45266618, 16760624, 16219283, 5311493, 445049, 1035, 45027791, 45027789, 44202131, 44145773, 25245484, 25244877, 25244708, 23690934, 22169438, 16394563, 6476300, 6543478, 6450800, 6454765
pregnadienes	26	63043, 63042, 63041, 62961, 45006164, 44266812, 24867475, 20054915, 23671691, 11957468, 11954369, 16490, 9793, 9782, 9642, 5876, 9571040, 6714002, 6713977, 6452749, 5388957, 5388959, 656804, 633091, 443958, 443936
2-acetylaminofluorenes	12	5897, 22469, 5896, 168033, 135827, 130776, 130694, 119334, 108117, 22722, 19347, 17270
neuraminic acids	14	46878426, 23679065, 20112027, 16760374, 6857396, 448209, 445063, 444885, 439197, 349960, 60855, 18292, 8565, 906

Figure 8 Characteristic substructures for the 15 different compound classes.