Natural Products Plug and Play with Chemical Substructures

Linking Natural Product Molecular Substructures to MS data by *in silico* generation of mass-based NP substructures

Author: Rutger Ozinga (930102638010)
Programme: MSc Bioinformatics
Supervisors: Dr. Justin J.J. van der Hooft, Dr. Marnix Medema
Examiner: Dr. Dick de Ridder
Institution: Wageningen University & Research, Bioinformatics department
Thesis period: June 2018 - March 2019, 36 ECTS

Abstract

In the field of metabolomics the annotation and classification of unannotated molecules is still a large bottleneck, this is due to a lack of reference spectra and tools helping with the structural annotation of these molecules. Therefore, efforts have been made to create novel tools to accelerate this process and make it more accurate. The goal of this project is to add to these efforts by combining two tools, CFM-id and MS2LDA, to speed up the identification substructures in mass spectra with *in silico* data. This was done with CFM-id, the output of this tool was used to create Mass2Motifs, patterns of fragments and losses that are concurrently observed, with MS2LDA. Using this information from the Mass2Motifs we can observe which motifs match to a spectra to gain a better understanding of what the molecule is composed of. By making the pipeline that uses CFM-id, it allowed for a quick and streamlined path from SMILES to a MGF that can be used to in MS2LDA. Of the created flavonoid motifs, around 40 of them show to have overlap with motifs belonging to flavonoid related datasets containing real mass spectra. These motifs have been annotated with the help of both the already annotated motifs of the matching data and annotation by hand with the help of MAGMa. Findings show that there is a lot of information to be gained from applying *in silico* spectra to create Mass2Motifs. Which in turn should greatly improve the rate and scale of structural annotation can be achieved.

Introduction

The study of metabolomics focuses on the full collection of metabolites present in an organism or biological sample[1]. Metabolomics is the last step in the process from gene to phenotype[2]. Therefore, metabolomics is crucial for understanding the mechanisms inside organisms[2]. Metabolomics can be divided into two different groups, targeted- and untargeted metabolomics[1]. With targeted metabolomics the goal is to quantify the already known and annotated metabolites with the help of reference databases[1]. Untargeted metabolomics on the other hand focuses on the entire metabolome including the unknowns[1,3]. The field of metabolomics has grown a lot in the last number of years, however major bottlenecks still remain[3,4]. One of the most important tools in metabolite identification is mass spectrometry (MS)[2]. A commonly used type of MS in metabolomics, called tandem mass spectrometry (MS/MS)[5], provides mass spectra from which a lot of metabolite data can be derived. Although MS/MS is highly capable for picking up the presence of metabolites, the issue currently is that the analysis and interpretation of these spectra is difficult[6]. For the identification of known metabolites reference MS/MS tools and libraries are used such as MassBank[7], ReSpect[8], NIST[9] and Global Natural Product Social Molecular Networking (GNPS)[10]. Even with the amount of databases there still is a lack of public available MS/MS reference spectra, as well as a lack of tools that help with the structural annotation of molecules[11]. Therefore it becomes hard to identify the metabolites present in MS/MS spectra. However, in recent years progress has been made with the creation of some new tools and approaches to assist with the identification and annotation of metabolites in spectra[5,11].

This project builds upon the tools cfm-if and MS2LDA that Allen et al.[5] and van der Hooft *et al*.[11] created. To tackle the problems that small molecule identification and annotation are facing, MS2LDA is a tool that uses the machine learning method latent Dirichlet allocation (LDA) to create Mass2Motifs out of fragmentation spectra[11]. Mass2Motifs can be seen as patterns that identify common structures found in a set of data, these patterns can belong to a specific substructure or region that reoccurs in the entered set of spectra[11]. Instead of identifying molecules with the fragmentation spectra[12], this method aims to use the Mass2Motifs to identify the different substructures. These motifs can then be used to help with the structural annotation of the molecule they belong to.[11]

Another development in the metabolomics field is the creation of *in silico* spectra to create large amount of simulated data which can then hopefully can be used to compensate for the lack of real-world reference spectra[5]. Competitive fragmentation modelling of EMI-MS/MS spectra for putative metabolite identification tool or CFM-ID[5] is one such tools which aims to produce *in silico* spectra based on entered molecules. This is done with machine learning making use of the support vector machine (SVM) method[5]. The program uses the probabilities of the fragmentation points of the molecule and creates spectra based on the most likely fragmentations[5]. The spectra are created with multiple collision energy levels, making it possible change the collision energy levels leading more fragmentation in case the energy gets higher[5]. By combining these two tools we will attempt to predict experimental data with Mass2Motifs that are created by using only *in silico* data. The main objectives for this project are to improve molecular substructure annotation and identification. We will be using the CFM-ID to predict the fragmentation of molecules and generate tandem mass (MS/MS) spectra, next the created spectra will be analyzed with MS2LDA. This tool creates Mass2Motifs based on the recurring patterns in the *in silico* spectra. These are patterns composed of fragments and losses observed in the data. Once this is done, we take these motifs and try to identify what they are with the use of experimental data and a database containing predicted substructures. Because the motifs and the spectra used to create the motifs are all *in silico*, it is need to compare the Mass2Motifs to experimental data in order to confirm that these motifs are actually picking up the substructures in real MS/MS data. The goal of this project is to see if possible to create a pipeline that can speed up the structural annotation of molecules with the use of *in silico* spectra.

Materials and Methods

All tools made in this project can be found at https://github.com/NP-Plug-and-Play-Scripts.
The tools and pipelines utilised in this project were made with Java version 1.8.0_191 and Python version 2.7.15.

Natural product database

The natural product database (NP DB) was created by S. Stokman who combined the data of multiple Np databases into a single one, this includes Super Natural II[13], ChEBI[14], HMDB[15], Np Atlas[16], GNPS[10] and others as seen in figure 1. This database contains the data of around 320.000 molecules and is the main source of data used in this project. The database contains a number of tables with the main one used for this being the structure table which contains the ID, SMILES, Classification and InChIKey. This database was also expanded upon at the end of this project to include the created data, check the appendix for the layout of the expanded database as well as a link to github for the code.

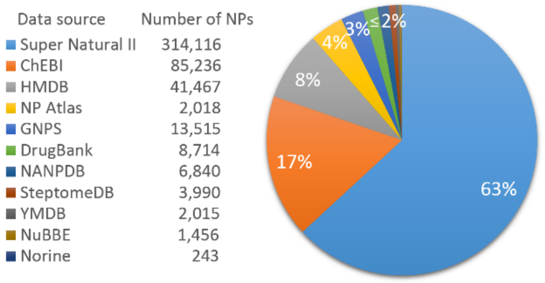


Figure 1 Natural product database data sources. The main bulk of the data is coming from Super Natural 2 with a number other smaller databases making up the rest of the database. Here they are indicated with the number of molecules and what percentage of the total data that they make up.

NpDb Extractor

To keep the search-process of the database manageable a General user interface was made to quickly obtain data. This interface was made with Java Swingx and uses the aforementioned NP DB to which it connects. Based on user input returns the entries in the databases corresponding to the selected data. The selection of data is based on the selected classifications. For example, a user can select a super class like benzenoids and get a list of all the data that has the super class benzenoids. The user can then save the data using a save button leading to a file containing the structure ID and SMILES string. This data is ready to be put through the cfm-pipeline. On github these are available as runnable jar and full project (made in Netbeans) either find them with the link above or look in the first part of the appendix.

MGF Format

Mascot generic format or MGF is a format for MS/MS data files[17]. These files contain the information of a spectrum, usually having multiple headers followed by m/z – intensity pairs making it very readable while keeping everything ordered. The start and stop of a spectrum are indicated with “BEGIN IONS” and “END IONS” everything in between these two lines belongs to one spectra[17,18]. The minimum information to be in a spectra is the precursor mass, the charge, a title and at least one m/z – intensity pair. However more information can be added to the spectrum, this leads to a very flexible file type while also making it hard to use MGF files made by others since they might include or exclude different information[18]. The appendix includes examples of some MGF formats.

CFM-id

Competitive fragmentation modelling of ESI-MS/MS spectra for putative metabolite identification or CFM-ID[5] is a probabilistic generative model that uses machine learning to learn parameters with the use of MS/MS data. CFM can be used for two different tasks, the first being the prediction of mass spectra of molecules using their chemical structure (in the form of a SMILES string or InChI) and secondly by identifying putative molecules[5]. For the present project CFM-id is mainly used to create *in silico* spectra and subsequently to work with these generated spectra. There are currently a number of different tools available for the generation of spectra mainly using one of two different methods. These methods use 1) a rule based model in which uses thousands of manually curated rules to predict spectra or 2) a combinatorial fragmentation procedure which enumerates all the possible fragments of the original structure, and then making a spectra out of those. CFM-id is different as it establishes its spectra on the likelihood of fragmentation. Based on the benchmark results when comparing CFM-id to MetFrag[5,19] and FingerID[5,20], CFM appears to outperform than both existing methods. The generation of spectra with CFM-id can be done in two ways: SE-CFM (single energy) and CE-CFM (combined energy). The difference between these two methods is that SE-CFM uses fragments the given molecule with only one energy level while CE-CFM creates three different spectra each using a different collision energy level (10,20,40V). This gives a better representation of reality since MS/MS spectra are normally viewed at multiple collision energies. The output of CFM-id can be created in multiple formats such as mzML[21], mzXML[22] and MGF[17]. The latter two are alternative ways of saving the mass spectra data. While mzML is widely used it isn’t as interpretable as MGF is, the same counts for mzXML.

To run CFM-id you need to provide it with three files.

1. the input file containing the ID’s and smiles.
2. A parameter file which the user can create themselves with a set of real mass spectra on which it trains to recognize the breaking points[5]. In case the user doesn’t want to or is unable to create their own parameter file they can use one of the pre trained models provided by CFM-id.
3. The last file that needs to be added is a configuration file, this file contains the settings on what rules CFM should follow when creating the *in silico* spectra. This file can be supplied by the user as well or downloaded off the CFM-id site.

When running CFM on the command line more options can be given. These options change if fragments are annotated and also make it so the probability threshold can be changed. This is to prune unlikely fragmentations that fall below the threshold. To get an installer of CFM-id please see the github section of the appendix.

RDKit

RDKit is an open source toolkit for cheminformatics, it has a large amount of functions that makes it easier to manipulate chemical structures by allowing the user to change aromaticity, neutralize, kekulize, add and remove molecules, cluster molecules and more[23]. RDKit can be used with Python, C# and Java. However, the latter two lack any examples on how to work with it whereas python has numerous sites with examples on how to use the different functions. For this project we chose to use it for the neutralization and kekulization options which are used in the CFM-pipeline. This is due to the inability of CFM-id to work with charged SMILES and Molconvert to work with aromatic SMILES.

Molconvert & InChIKeys

Molconvert is a tool in the JChem toolkit that is used to convert between molecular file types[24]. It was used in this project to change the SMILES in to InChIKeys. The InChIKeys are unique identifiers for molecules they are created since the InChI and SMILES notations get very lengthy whereas InChIKeys are always 27 characters long and will always be unique for one molecule. The first 14 characters are the hash encoding for the molecular skeleton of the molecule the 8 characters after that are the hash encoding of the remaining layers such as the stereochemistry. After this there is a letter for the flag which indicates the type of InChIKey, this can be either standard or non-standard[25]. Next is another letter which indicates the version of the InChIKey starting with A for version 1, B for version 2 etc. last is a letter indicating the whether or not a molecule is protonated or deprotonated, with N meaning 0 protons M being -1 and O being +1 the lower in the alphabet the more negative charges the molecule has and the higher in the alphabet the more protons it has[25]. In the appendix github section a standalone version of the InChIKey pipeline can be found.

CFM-pipeline – a pipeline to turn SMILES them in to *in silico* spectra

In order to get from database info to a MGF used as input for MS2LDA a number of steps need to be taken. A pipeline was created to do this task, it will take an input csv file containing ID, SMILES pairs as well as a settings file that contains the preferred settings for the experiment. Here the pre trained model for CFM, the cut-off value along with other settings will be decided. This tool modifies and edits the data so it is prepared to be submitted to MS2LDA[11].

Extraction: First of all the data needs to be extracted from the database. This is done with the NPDB extractor.

Neutralisation and splitting: The csv file is then put in the CFM-pipeline. After one of the earlier runs the pipeline it turned out that around a 1000+ molecules were being lost per dataset if the entered csv file was just entered in to CFM-id. The reason for the loss of data turned out to be the inability of CMF-id to fragment molecules that were charged. This led to the implementation of a neutralization step, which was done with the help of RDKit[23]. With this intervention the loss of molecules was greatly reduced in most data sets, going down to around 500 molecules lost per dataset. The output is then saved in a new file which contains the ID,SMILES and neutralized SMILES. Another problem arose in the flavonoid data, this dataset consisted of a large amount of SMILES that contained multiple molecules. SMILES can contain multiple molecules in a single string, this being indicated with a "." On the location were a seperate molecule starts. Cfm-id is not able to process these leading to only 3000 of the around 9000 flavonoid molecules being able to pass through it. Therefore only the longest sub SMILES was taken from the SMILES, this is then considered as the main molecule. Next the large file is split in to equal parts so they can undergo the next steps simultaneously with the help of multiprocessing which greatly reduces the process time of Molconvert and CFM-id.

InChIKey creation: Molconvert [24] is then used to turn both the original and neutralized smiles in to InChIKeys. However SMILES can be denoted in their aromatic form, this lead to molconvert[24], which wants to have a static representation of the molecule figure 2, not being able to handle certain molecules with aromatic rings due to the delocalization of bonds[26]. Thus RDKit was used again to turn the molecules to their kekulized form which has localized bonds. All the info is stored in a new file, which contains the ID, SMILES, neutral SMILES, InChIKey, neutral InChIKey.

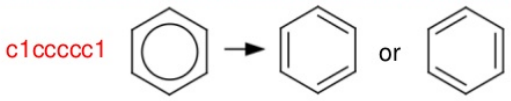


Figure 2: example of a delocalized benzene ring and the two possible kekule versions of it.

Spectra creation: Once all files are done receiving their InChIKeys the files are ready to be entered in to CFM-id. For these runs the settings of the CFM-pipeline were put the CE-CFM mode to obtain all three energy level, the output type is MGF. The parameter model used was the pre trained model by CFM-id, param_output0.log. This file is available on the CFM-id site. The parameter file was obtained here as well. For the command line settings the probability threshold was put on 0.001, the default value, and the Annotate_fragments option was set on 0 meaning the fragments were not annotated. This was done due to reduce the run time.

Peak normalization: Next the spectra intensities that came out of CFM-id are normalized with the formula $\sum((x_a - \mu) / \sigma)$[27]. Any normalised intensity of 2 or an intensity below -2 is set to 2 or -2 respectively. This was done to have the normalized values are on a scale from -2 to 2. Next all the values are multiplied by 225 to put the values on a scale of -450 to 450 and finally 450 is added to all values to make a scale of 0 to 900. This is done to make more pronounced differences in the intensities as well as making them more in line with actual spectra[27]. **T**he old intensities are replaced with the created normalized intensities.

Peak merging: The normalized spectra are then put in to a merger which takes the three spectra each that belong to one molecule (each were fragmented on a different energy level with the CE-CFM method) and merges these in to a single spectra. This is done by adding all the peaks in to a single spectra and then combining the ones that have the same mass, the combined intensity of those spectra are then divided by the number of peaks that were combined. So if we had two peaks with the m/z of 78.114 then the combined intensity would be divided by 2. The reason for the merging of the peaks is to increase the amount of information in a single spectrum. Because MS2LA will find common patterns in the spectra, the more information is added the more informative Motifs should come out of it. It also helps as an extra step to conserve relevant peaks. A low energy spectra with a peak that has a low intensity for benzene for example might get filtered out, however with the combining of the three spectra we might see that the mid and high energy spectra have the benzene peak as well but with a much higher intensity thus conserving the relevant peaks. Once all the spectra are merged a final filtering step is done by removing all peaks with an intensity below a set threshold, 120 in this case, to remove some of the background noise from the spectra.

Adding info: The final step is the addition of extra information to the MGF. The added information comprises: 1) the ID which links to the NP database, 2) one or two InChIKeys based on if the SMILES got neutralized or not, 3) a title which describes the spectra, 4) the SMILES belonging to the spectra (also a neutral SMILES in case it was neutralized) and 5) an IUPAC name (naming of organic compounds). Once this info is added to the spectra all the files that were split into separate parts are combined in to a single file again. This file is then made after which it can be used as input for MS2LDA.

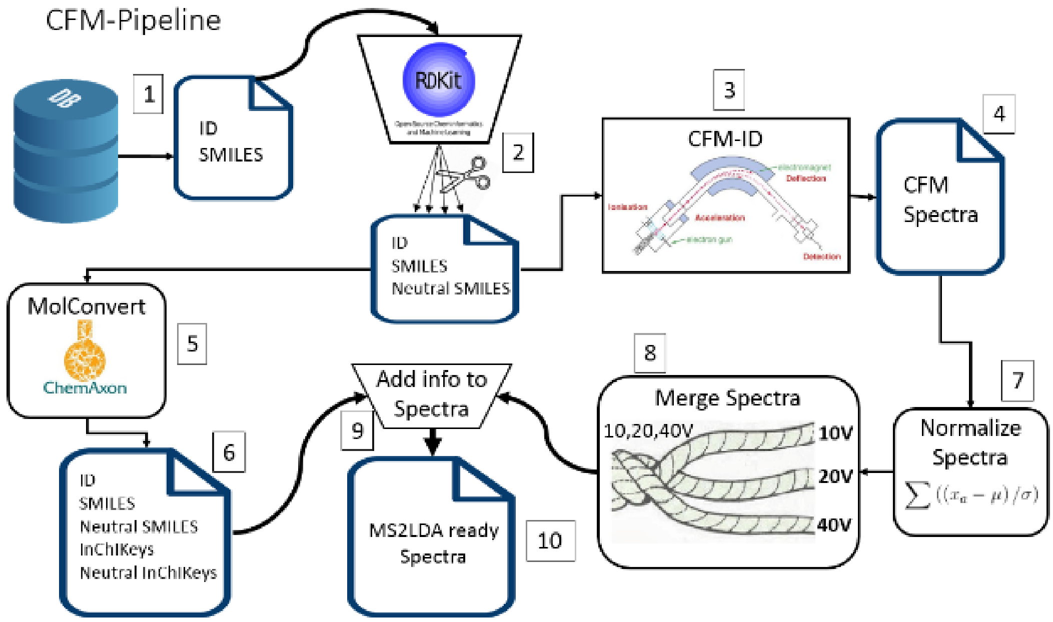


Figure 3: An overview of the pipeline. 1) The smiles are obtained from the pipeline with their ID. 2) The SMILES are neutralized so they can be submitted to CFM-ID and Molconvert, and the file is split in 10 parts. 3) CFM-ID creates spectra out of the neutralized SMILES. 4) A MGF file containing the by CFM-ID created spectra. 5) Molconvert takes the original and ,in case present, neutralized SMILES and creates InChIKeys for both. 6) A file containing the SMILES and InChIKeys for the neutral and original molecules. 7) The created spectra are normalized. 8) The spectra that came of the same molecule but fragmented on different energy levels are combined. 9) The InChIKey(s) and SMILES along with some other info such as the ID and description are added spectra. 10) The resulting file contains all information needed for MS2LDA to get the most out of these spectra.

MS2LDA & Mass2Motif

Mass2Motifs are patterns that can have a possible match to biochemically relevant molecular substructures[11]. The term was first used by van der Hooft et al[11]. The idea of Mass to motifs is that they match to a certain set of peaks that occur together often this means that if a selection of NPs contain a Mass2Motif they all contain the same molecular substructure that correspond to the pattern of the Mass2Motif. Whether this Motif is actually accurate or even matches to a relevant substructure is not something that is given, thus manual validation of these motifs is necessary. These Mass2Motifs are generated with the tool MS2LDA which uses LDA[28] on the MS/MS Spectra to find patterns in the different spectra. It looks at the recurring losses and fragments in a set of spectra and based on that it creates Mass2Motifs belonging to that set of spectra.

To start an experiment in MS2LDA the tool requires an input file to be in one of three file types mzML, MGF or MSP as well as a number of setting. These include options such as;

- The name of the field that contains the unique identifiers for the spectra in the submitted file.
- The min and max retention times to store for the MS1.
- The minimum intensities for both the MS1 and MS2 to store.
- The Number of Mass2Motifs to create.
- The Number of iterations for the LDA.

These options are all set on default values (with the exception of the unique identifier), the last two options heavily influence the run time of MS2LDA: the more Motif that are created the longer the runs will take to be completed, the same also applies for the number of iterations. Therefore the amount of motifs was set to 300 for all the datasets except the steroid dataset, for which the number was set to 400. This was done due to the large number of molecules in the dataset, increasing the number of motifs made it possible to catch more motifs that might be present in a dataset ten times the size of the others (with the exception of amino acids). Although even more info could have been collected with an even higher number of motifs the computing power of the server running MS2LDA could not handle more than 400 motifs. The number of iterations was kept on 1000.

In general, the number of motifs heavily depends on the input dataset. Too many motifs means the run time becomes longer and it also opens up the possibility of capturing motifs that are only specific for one or two spectra. On the other hand, using a very small amount of motifs can lead to a large number very general motifs that match to patterns that appear in nearly every molecule (e.g., C or O fragments). Furthermore, when submitting a set of data it is good to keep in mind that MS2LDA will make a set of motifs based on that info. So if a set of data is submitted that is very similar, for example all from the same subclass of molecules, the resulting set of motifs will have a good change of belonging to substructures frequently found in this set of subclass. On the other hand if the set of molecules is very diverse its more likely that more motifs will more frequently belong to common substructures. All aforementioned considerations need to be kept in mind when running an experiment. Once the experiments are done running more options become available such as; inspecting the motifs or spectra submitted, comparing the motifs of one experiment to others in order to find motifs with overlap. It is also possible to add more filtering options.

Molecular data sets

Five sets of molecule classes were selected from the NP DB. These sets contain the data of 18082 structures belonging to ‘Amino acids, peptides, and analogues’, 3654 Azoles, 9752 Flavonoids, 1441 Lactones and 15170 structures belonging to Steroids and steroid derivatives. The reason these sets where chosen vary, the Azole dataset was picked as it contains relatively short and simple molecules making it easier to analyse and find back substructures. Lactones were picked because of them being well researched for their presence in antibiotics[29], Flavonoids for their bioactive structures[30,31], Steroids as well for their use in medicine and their effects on organisms and health[32] and Amino acids because they are relevant for proteomics, biochemistry and many others[33,34].

Results and Discussion

The first goal of this project was to set up a pipeline with cfm-id in order to create *in silico* spectra that can be used to derive motifs. This step alone cost a lot of time due to difficulties with the installation of the tools required for the pipeline. This includes tools lp-solve[35] a program that changes the way operating systems handle certain calculations, without it cfm-id won't run. In order to get it working an edited version of the program was made so it would run on the servers at Wageningen University and Research. The resulting pipeline can be seen in figure 3. In order to find out if it's possible to create and link motifs to substructure test need to be ran and for this the datasets mentioned where extracted from the NPDB with the help of the created NpDb extractor.

CFM-pipeline output

The first step was to put the extracted data of the selected molecule classes through the pipeline. The MGF files that came out of the pipeline had slightly less spectra in them compared to the number obtained from the database. While the previously explained steps of neutralization, kekulization and the removal of multiple molecules in one single SMILES helped to reduce the number of lost molecules. A few molecules still remain that can't pass through CFM-id. Figure 4 shows the number of spectra remaining per dataset.

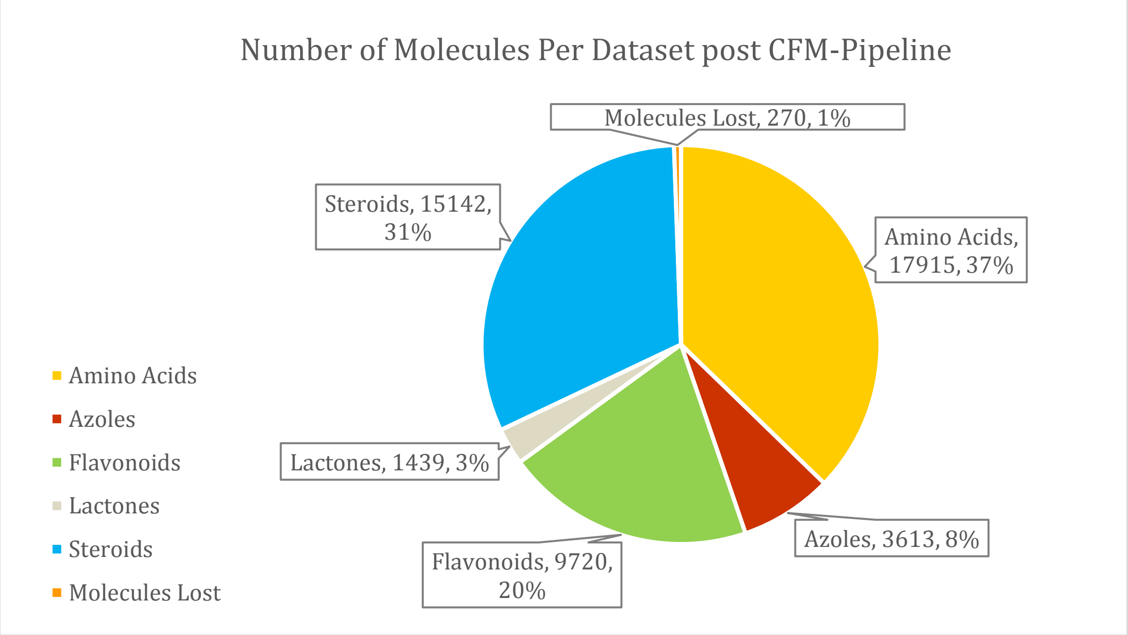


Figure 4: Shows the distribution of the data that passed through the CFM-pipeline as well as the total amount of molecules lost due to varying reasons. The values indicate the number of spectra that came out of the pipeline along with what percentage of the total they are. Of the lost molecules 167 belonged to the amino acid data, all other data sets lost around 35 molecules with the exception of lactones which lost only 2.

Spectra Validation

In the paper on CFM-id by Allen et al[5] the output spectra were already validated showing that the predicted spectra matched fairly well with the measured spectra. This resulted in weighted recall values around 70%, weighted recall being the percentage of the total peak intensity matching between the measured and predicted spectra. However due to the merging of the spectra in this project additional validation was done to ensure that the created spectra still are comparable to experimental spectra. Therefore a number of molecules were picked that had available spectra and these were compared with the CFM created spectra. In the examples below the spectra belonging to an anisole and a cholesterol of CFM-id and MassBank were taken and compared. When looking at figure 5 it can be seen that the two spectra are mostly similar with some exceptions. And while this is true for a large part of the spectra created there are also spectra that while having some peaks in common also have many different peaks. These include peaks like 39 m/z, 65m/z and 78 m/z. This can be seen in figure 6 where the cfm version of cholesterol is very different from the one obtained from MassBank North America[36].

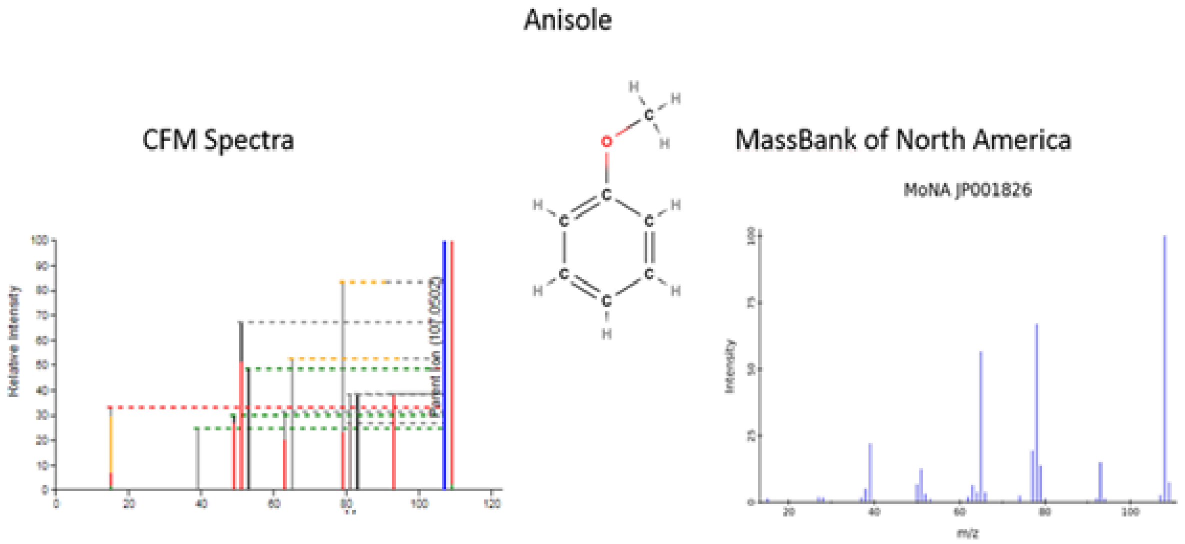


Figure 5: This figure shows the comparison between the cfm generated spectra of an Anisole and an actual spectra of an anisole obtained from MassBank North America. On average the peaks that appear in the MassBank spectra also appear in the spectra created with cfm-id. Thus a likely accurate version of the spectra was made. [*http://mona.fiehnlab.ucdavis.edu/spectra/display/JP001826*](http://mona.fiehnlab.ucdavis.edu/spectra/display/JP001826)

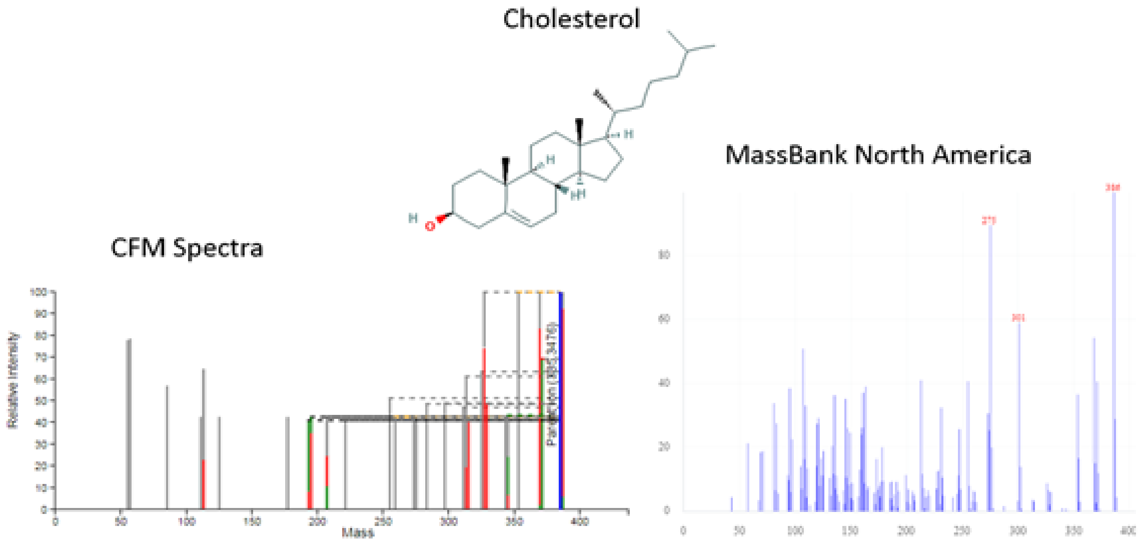


Figure 5: This figure shows the comparison of the cfm generated cholesterol spectra with MassBank North America spectra of cholesterol as can be seen while there is some overlap in peaks there is also a large amount of peaks not pressent or on different locations. [*http://mona.fiehnlab.ucdavis.edu/spectra/display/JP003478*](http://mona.fiehnlab.ucdavis.edu/spectra/display/JP003478)

Spectra variation

The reasons for why some of these generated spectra are different from the spectra obtained from MassBank could lie in the use of the pre trained model of cfm-id. The data used to train the model is a large set of molecules with little to no natural products[5]. This could mean that some of the ways and likelihood steroids and other molecules fragment are not known to the model thus giving an incorrect fragmentation. The other explanation can be that while the peaks were present in the *in silico* spectra the intensity of these peaks was to low leading to them being filtered out. However even with these somewhat incorrect spectra it can still be possible to find useful motifs. With enough spectra the recurring patterns may be retrieved and thus useful motifs can still be generated. Another example can be found in the appendix where a CFM spectra has.

Motif matching against annotated datasets

With the CFM Runs done the datasets were entered in MS2LDA. The output of this was compared to other existing datasets. For the Flavonoids the data was split in two separate parts to make it possible for them to pass through MS2LDA. These two sets were compared with each other and a number of different experimental datasets. These being the MotifDB (massbank_binned_005), Rhamnaceae_plant_extracts_KyoBin_200Motifs, GNPS-Rhamnaceae, GlobalEuphorbiaStudy, Urine38, and the Foodomics datasets. The Rhamnaceae_plant, GlobalEuphorbia and Foodomics sets contain a lot of plant data thus likely containing flavonoids as well[37]. This Resulted in a total of 116 matches of which 73 have a score above 0.7. Of these 73 motif a number of them had matches with one or more datasets. This is plotted in figure 7, here the total amount of unique matches and annotated motifs is shown. In figure 8 the number of motifs with multiple matches is shown along with the number of these that are annotated.

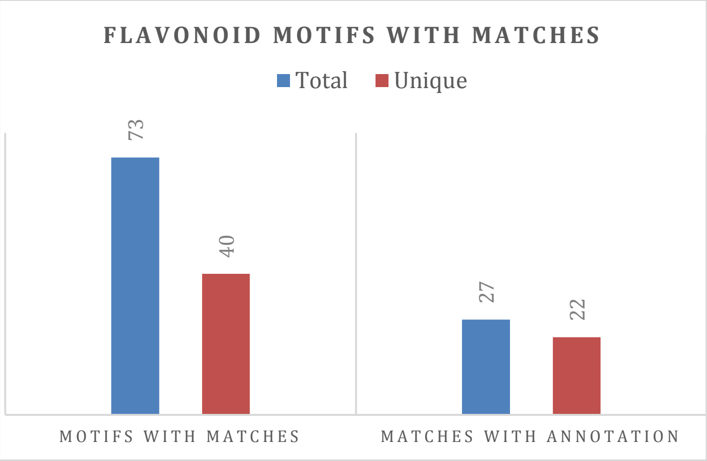


Figure 7: Shows the amount of motifs matched to the flavonoids2 dataset showing that around there are 73 matches in total with 40 of these being unique matches. Meaning 40 of the 300 motifs have matches.

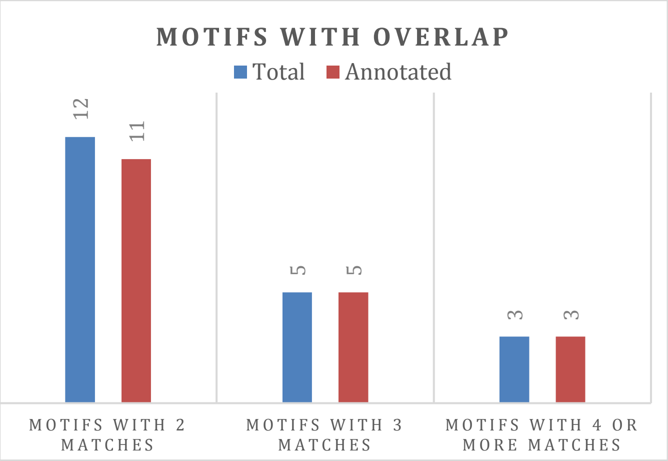


Figure 8: Shows the motifs that have multiple matches with different datasets. This is shown as the amount of in silico motifs with either 2, 3 or with 4 or more matches.

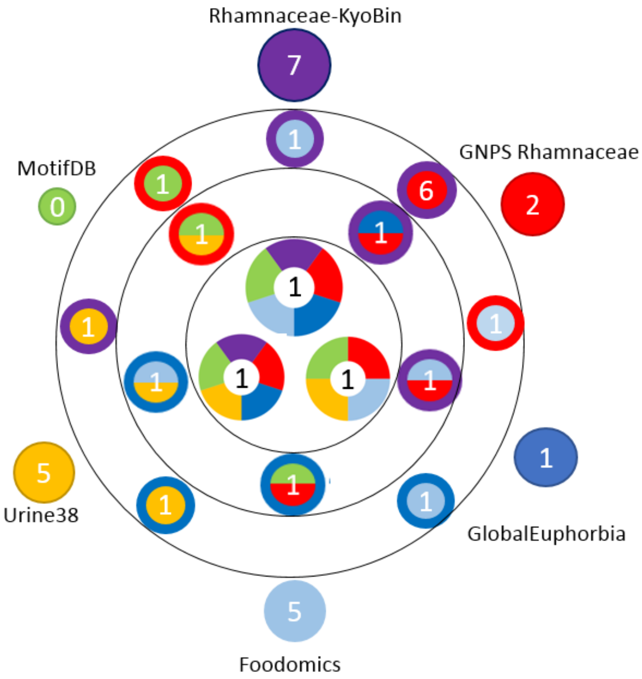


Figure 9 shows the overlap of matches between the datasets, shown in figure 8. Each point shows amount of overlap between the data. So for example a point having purple blue and red means that one motif in the cfm-flavonoids has matches with GNPS-Rhamnaceae, Global uphorbia and Rhamnaceae-KyoBin. The reason MotifDB has 0 is because MotifDB contains the data of some of these other datasets thus only having matches with other data and no unique matches.

Overlap in motif matches

One of the biggest concerns at the start of this project was if the *in silico* spectra would actually be able to produce motifs that would match to real data. With the flavonoid experiments having 300 motifs (the default amount of motifs created) having a total of 40 of these spectra covered, as seen in figure 7 with matches from experimental data is a good indication that it is working for at least other flavonoid data. Especially if it is taken in to account that of the 300 motifs a large amount of them will be overfitting to the data, thus being specific for only a few spectra. This happens a lot with the larger molecules due to them having several large fragments, as well as for molecules that are quite similar such as isomers or those belonging to the same sub class. Furthermore, as seen in figure 8 and 9, 50% of these motifs have two or more matches in different datasets. This further reinforces that the motifs created with the pipeline will most likely be able to be used on other data as well. Figure 10 also shows some of the annotated motifs with the number of overlaps. As can be seen the possible fragments of most of these motifs belong to structures often seen in plants. Which is expected due to the motifs being based on flavonoid data.

Motif Degree

Although the before mentioned motifs have one or more matches it won't be useful if the motif doesn't match to anything. Therefore the motifs were collected along with their degree, which is the amount of times they match to a spectra in the data. All of the 40 motifs with matches have a degree of 250 or higher with more than half having a degree of 500+. This shows that most of the motifs actually match with the data. See the MS2LDA section of the appendix for a graph and table of the distribution of the degree between the motifs with matches.

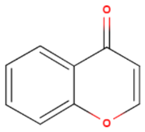
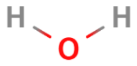
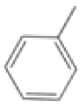
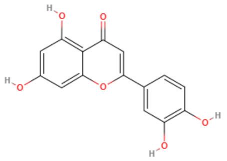
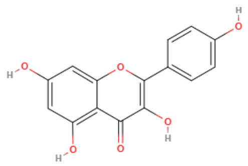
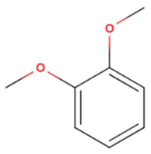
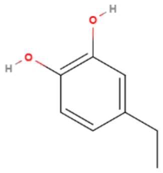
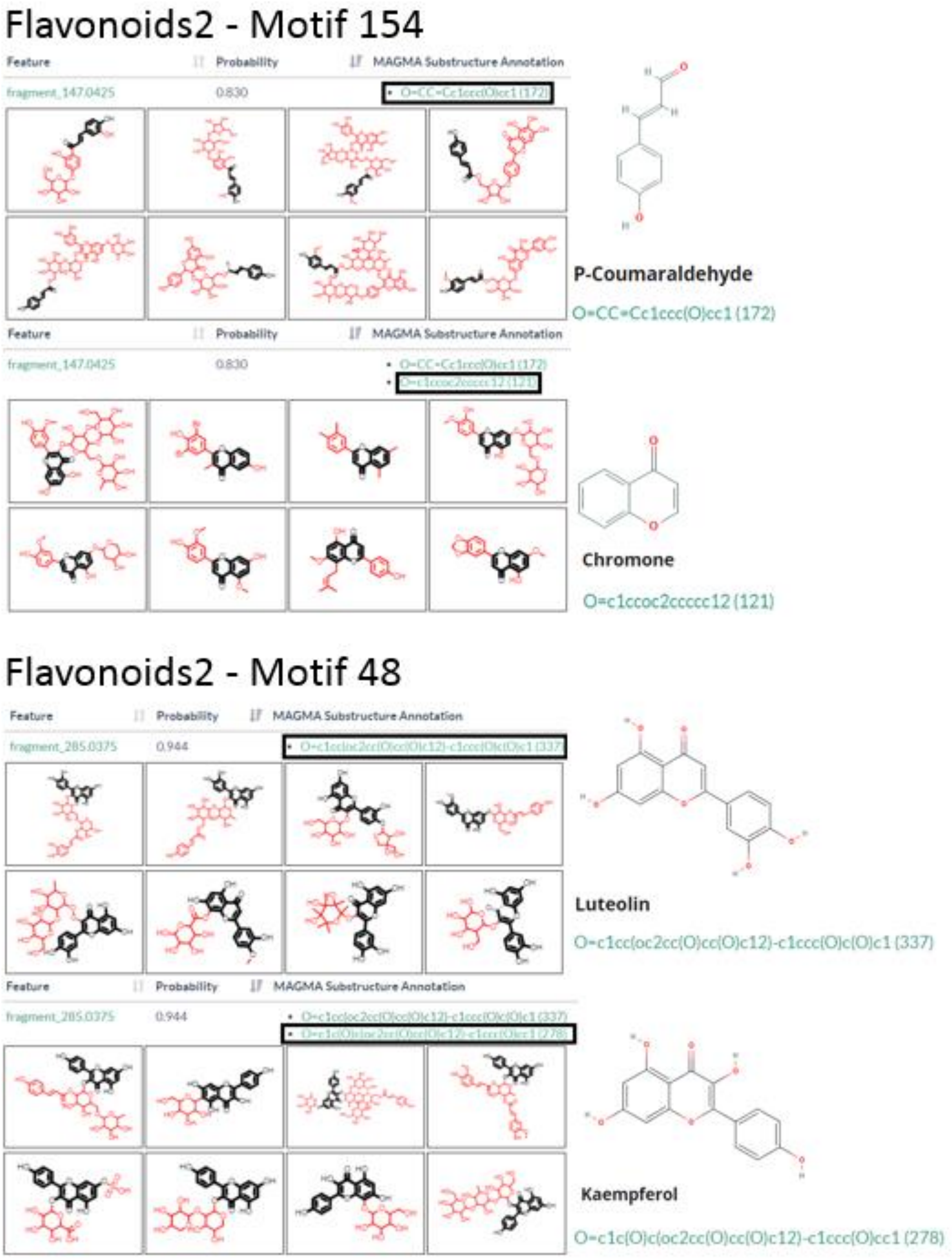
Motif with Annotation	Number of matches with other datasets	Possible fragment/loss structure
motif_154 (Possible Chromone fragment an isomer of Coumarin - $C_9H_6O_2$)	5	
motif_240 (Water loss - indicative of a free hydroxyl group (in beer often seen in sugary structures))	5	
motif_225 (Fragment indicative for aromatic compounds related to methylbenzene substructure (C_7H_7 fragment))	4	
motif_48 possible Luteolin, Kaempferol like structures ($C_{15}H_{10}O_6$) often found in plants	2	 
motif_7 possible Veratrole or 4-Ethylcatechol 137.0575 fragment ($C_8H_{10}O_2$) - 4-Ethylcatechol a constituent of roasted coffee and Veratrole is an insect attractant created by plants, being the methylated form of guaiacol	1	 

Figure 10: This table shows 5 motifs that had matches with one or more of the motifs from other databases.. Here a few of these motifs are displayed with their given annotation as well as the possible structure of the fragment. As seen, a number of these motifs have flavonoid related fragments. Thus making them good candidate motifs to try and find new flavonoids on other data. Especially since they have matches with actual datasets.



Flavonoids2 - Motif 154

Feature	Probability	MAGMA Substructure Annotation
fragment_147.0425	0.830	• O=CC=Cc1ccc(O)cc1 (172)

P-Coumaraldehyde

O=CC=Cc1ccc(O)cc1 (172)

Feature	Probability	MAGMA Substructure Annotation
fragment_147.0425	0.830	• O=CC=Cc1ccc(O)cc1 (172) • O=c1ccoc2ccccc12 (121)

Chromone

O=c1ccoc2ccccc12 (121)

Flavonoids2 - Motif 48

Feature	Probability	MAGMA Substructure Annotation
fragment_285.0375	0.944	• O=c1cc(oc2cc(O)cc(O)c12)-c1ccc(O)c(O)c1 (337)

Luteolin

O=c1cc(oc2cc(O)cc(O)c12)-c1ccc(O)c(O)c1 (337)

Feature	Probability	MAGMA Substructure Annotation
fragment_285.0375	0.944	• O=c1cc(oc2cc(O)cc(O)c12)-c1ccc(O)c(O)c1 (337) • O=c1c(O)c(oc2cc(O)cc(O)c12)-c1ccc(O)cc1 (278)

Kaempferol

O=c1c(O)c(oc2cc(O)cc(O)c12)-c1ccc(O)cc1 (278)

Figure 11: The figure above shows motifs 154 and 48 of the flavonoids2 dataset. When looking at the most likely fragments of these motifs we find that for motif 154 the fragment 147.0425 has the highest probability of being present when the motif is found in a spectra. The MAGMa annotation given to this fragment is as seen above likely related to P-Coumarldehyde or Chromone, the first being observed 172 times in the flavonoid data and the second is observed 121 times. As for motif 48 the most likely fragment is 285.0375, this is annotated with MAGMa as likely being the isomers Luteolin (observed 337 times) and Kaempferol (observed 278 times), both being related to flavonoids. While these MAGMa annotations help with the annotation of the motif it is not guarantee that in other data these annotations are correct.

Motif annotation

While a number of the motifs of the flavonoid date were given an annotated of a matching motifs with an annotation, there were motifs that had a match with no annotation. For these time was spend to try and annotate these motifs by hand. At first this process was slow and tedious due to having to look at the masses of the fragments belonging to the motif and the structures of the molecules that belonged to the spectra. And even when a possible annotation could be given it was still uncertain that this was the case. That changed when the flavonoid datasets and the steroid dataset were processed by MAGMA Substructure annotation[38]. This online application annotates multistage Mass spectrometry data[38]. By annotating these datasets it allowed for easier identification of fragments part of motifs thus leading to more accurate ways to annotate the data. For extra insight in the data look at the MS2LDA section of the report, here a table of all 40 matches can be found along with their annotation (if they have one). Also a link to the flavonoid experiments can be found in the appendix.

Discussion of Motif Annotation

As seen in figure 11, with the help of MAGMa[38] the two fragments of motif 154 with the highest probability were inspected. Previously the motif was annotated by hand and given the initial annotation of Chromone as seen in figure 11. While this annotation is partially right it is not an accurate representation of what can be found with the motif. So a new annotation should be given that also takes the other possibilities shown here in to account, or make it more general as “Coumarin related”. While this method of annotation works really well there is something to keep in mind when looking at these motifs. While it works when comparing these motifs generated with flavonoid *in silico* spectra to motifs and data of other flavonoid spectra. This does however not guarantee that these *in silico* motifs will only match to flavonoid data and that the annotations hold when they annotate to other natural product spectra. For example a match between one of these *in silico* flavonoid motifs and real steroid motifs can be found but they both have a different annotation due to the motif having similar fragments. However when the *in silico* steroid motifs and the *in silico* flavonoid motifs were matched, only the motifs that should be present in both (such as benzenaldahyde) matched. Also none of the flavonoid specific motifs were found in this comparison. For more info look at the MS2LDA section of the appendix here a table with the annotated hits can be found along with a link to the experiment. What also happened in some cases is that a motif would only have a single fragment, while these motifs can be useful for prediction this can lead to the motif matching to fragments of the same weight. While this can also happen to motifs with more fragments, the chances of, for example all five features being the same in a flavonoid motif and a steroid motif is way lower than the chance of the same thing happening when both motifs have a single feature.

Conclusion

The tools and knowledge that were produced by this project will make it easier to quickly produce a set of spectra and motifs that can be utilized to identify new molecules. The best way to create new Mass2Motifs for structural annotation for molecules belonging to a certain class would be to first use the NP DB created by S. Stokman to obtain the desired data. These can be searched with NpDb extractor by searching on the classification. The classifications are currently in the progress of being improved by O. Hoekstra who is working on a project related to this. With this data is then entered in this pipeline and turned in to spectra. Next MS2LDA can create the motifs and MAGMa will annotate the substructures to allow for easy annotation. This will result in a set of motifs that can be used to give a first impressions on what these unannotated molecules are made of. The results obtained in this project show great promise. They show that a lot of information to be obtained from Mass2Motifs created with *in silico* spectra, this is done in relative short time frame using publicly available data. However, effort should be made to create a pre-trained model for CFM-id trained on real natural product spectra. This way the probability of obtaining more accurate spectra is increased, which in turn will likely yield more accurate Mass2Motifs with more predictive capabilities. Also with the help of projects like MAGMa[38], the rate at which the created motifs are annotated can be improved immensely. Thus adding this to more datasets will further improve our insight of the quality of created motifs. Other projects like the work done by Lai *et al.* [39] also created a workflow to aid in the structural annotation of molecules. A good first step would be to take the annotated structures they mentioned such as: N-methyl-uridine monophosphate, lysomonogalactosyl-monopalmitin and N-methylalanine, and create our own motifs based on the classes these molecules belong to. The resulting motifs can then be annotated with MAGMa. The results can then be compared to the results of Lai *et al.*[39] as an extra validation step for both projects. Due to this project highly relying on *in silico* spectra the extra validation will be especially valuable. To conclude, along with other advances in the field of metabolomics this project shows it has a lot of potential to improve the efficient structure annotation and classification of unknown molecules. Especially when combined efforts of other projects such as the one of S. Stokman, O. Hoekstra and L. Ridder.

Acknowledgements

I would like to thank Justin J. J. van der Hooft for the opportunity to work on this project and both him and Marnix Medema for the time and knowledge they provided during the meetings and when asked for help. Lars Ridder for the MAGMa annotation of the flavonoid and steroid data. And Oscar Hoekstra and Sam Stokman for the support they provided in the form of help and explaining their respective projects.

Sources

1. Roberts, L. D., Souza, A. L., Gerszten, R. E. & Clish, C. B. Targeted metabolomics. *Curr. Protoc. Mol. Biol.* (2012). doi:10.1002/0471142727.mb3002s98
2. Dettmer, K., Aronov, P. A. & Hammock, B. D. Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews* (2007). doi:10.1002/mas.20108
3. Peisl, B. Y. L., Schymanski, E. L. & Wilmes, P. Dark matter in host-microbiome metabolomics: Tackling the unknowns-A review. *Analytica Chimica Acta* (2018). doi:10.1016/j.aca.2017.12.034
4. Dunn, W. B. *et al.* Mass appeal: Metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics* (2013). doi:10.1007/s11306-012-0434-4
5. Allen, F., Greiner, R. & Wishart, D. Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics* **11,** 98–110 (2015).
6. Ma, Y., Kind, T., Yang, D., Leon, C. & Fiehn, O. MS2Analyzer: A software for small molecule substructure annotations from accurate tandem mass spectra. *Anal. Chem.* (2014). doi:10.1021/ac502818e
7. Horai, H. *et al.* MassBank: A public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* (2010). doi:10.1002/jms.1777
8. Sawada, Y. *et al.* RIKEN tandem mass spectral database (ReSpect) for phytochemicals: A plant-specific MS/MS-based data resource and database. *Phytochemistry* (2012). doi:10.1016/j.phytochem.2012.07.007
9. The National Institute of Standards and Technology. Available at: https://www.nist.gov/srd/nist-standard-reference-database-1a-v17. (Accessed: 16th July 2018)
10. Wang, M. *et al.* Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology* (2016). doi:10.1038/nbt.3597
11. van der Hooft, J. J. J., Wandy, J., Barrett, M. P., Burgess, K. E. V. & Rogers, S. Topic modeling for untargeted substructure exploration in metabolomics. *Proc. Natl. Acad. Sci.* (2016). doi:10.1073/pnas.1608041113
12. Allard, P. M. *et al.* Integration of Molecular Networking and In-Silico MS/MS Fragmentation for Natural Products Dereplication. *Anal. Chem.* (2016). doi:10.1021/acs.analchem.5b04804
13. Banerjee, P. *et al.* Super Natural II-a database of natural products. *Nucleic Acids Res.* (2015). doi:10.1093/nar/gku886
14. Degtyarenko, K. *et al.* ChEBI: A database and ontology for chemical entities of biological interest. *Nucleic Acids Res.* **36,** 344–350 (2008).
15. Wishart, D. S. *et al.* HMDB: The human metabolome database. *Nucleic Acids Res.* **35,** 521–526 (2007).
16. Victor Aniebok, M. B. NP Atlas. (2018). Available at: https://www.npatlas.org/joomla/.
17. Kirchner, M., Steen, J. A. J., Hamprecht, F. A. & Steen, H. MGFp: An Open Mascot Generic Format Parser Library Implementation. *J. Proteome Res.* **9,** 2762–2763 (2010).
18. Fiehn Lab. MGF Files. (2016). Available at: https://fiehnlab.ucdavis.edu/projects/lipidblast/mgf-files.
19. Ruttkies, C., Schymanski, E. L., Wolf, S., Hollender, J. & Neumann, S. MetFrag relaunched: Incorporating strategies beyond in silico fragmentation. *J. Cheminform.* (2016). doi:10.1186/s13321-016-0115-9
20. Alotaibi, S. J. & Argles, D. FingerID: A new security model based on fingerprint recognition for personal learning environments (PLEs). in *2011 IEEE Global Engineering Education Conference, EDUCON 2011* (2011). doi:10.1109/EDUCON.2011.5773128
21. Martens, L. *et al.* mzML—a Community Standard for Mass Spectrometry Data. *Mol. Cell. Proteomics* (2011). doi:10.1074/mcp.R110.000133
22. Lin, S. M., Zhu, L., Winter, A. Q., Sasinowski, M. & Kibbe, W. A. What is mzXML good for? *Expert Review of Proteomics* (2005). doi:10.1586/14789450.2.6.839
23. Landrum, G. RDKit: Open-source cheminformatics. Available at: http://www.rdkit.org.
24. ChemAxon. Molconvert version 18.23.0.
25. Heller, S. R., McNaught, A., Pletnev, I., Stein, S. & Tchekhovskoi, D. *InChI, the IUPAC International Chemical Identifier*. *Journal of Cheminformatics* **7,** (Journal of Cheminformatics, 2015).
26. O¸ÄôDonnell, T. *Design and Use of Relational Databases in Chemistry*. *Design and Use of Relational Databases in Chemistry* (2010). doi:10.1201/9781420064438
27. Joy, S. (Albertus M. C. Normalization. (2015). Available at: https://www.researchgate.net/post/How_do_i_normalize_data_from_0_to_1_range.
28. Campbell, J. C., Hindle, A. & Stroulia, E. in *The Art and Science of Analyzing Software Data* (2015). doi:10.1016/b978-0-12-411519-4.00006-9
29. Barnes, E. M. (a) Antibiotics. *J. (Royal Soc. Heal.* (2007). doi:10.1177/146642405707700807
30. Hidalgo, M., Sánchez-Moreno, C. & de Pascual-Teresa, S. Flavonoid-flavonoid interaction and its effect on their

antioxidant activity. *Food Chem.* (2010). doi:10.1016/j.foodchem.2009.12.097

31. Heim, K. E., Tagliaferro, A. R. & Bobilya, D. J. Flavonoid antioxidants: Chemistry, metabolism and structure-activity relationships. *Journal of Nutritional Biochemistry* (2002). doi:10.1016/S0955-2863(02)00208-5
32. Panepistēmio tēs Krētēs., S. *et al. International journal of molecular medicine. International Journal of Molecular Medicine* (2015).
33. Maleknia, S. D. & Johnson, R. in *Amino Acids, Peptides and Proteins in Organic Chemistry* (2011). doi:10.1002/9783527631841.ch1
34. Vertes, A. in *Medical Applications of Mass Spectrometry* (2008). doi:10.1016/B978-044451980-1.50010-0
35. Eikland, Kjell Notebaert, P. LP Solve.
36. MassBank North America. Available at: http://mona.fiehnlab.ucdavis.edu/.
37. Wandy, J. *et al.* Ms2lda.org: Web-based topic modelling for substructure discovery in mass spectrometry. *Bioinformatics* (2018). doi:10.1093/bioinformatics/btx582
38. Ridder, L., van der Hooft, J. J. J. & Verhoeven, S. Automatic Compound Annotation from Mass Spectrometry Data Using MAGMa. *Mass Spectrom.* (2014). doi:10.5702/massspectrometry.s0033
39. Lai, Z. *et al.* Identifying metabolites by integrating metabolome databases with mass spectrometry cheminformatics. *Nat. Methods* (2018). doi:10.1038/nmeth.4512

Supplementary

Github links.

For all files go to this link: https://github.com/NP-Plug-and-Play-Scripts

CFM-Pipeline:

This contains the main pipeline created in this project

https://github.com/NP-Plug-and-Play-Scripts/CFM-Pipeline

NpDb Extractor:

Keep in mind that this project requires Java. Contains the NpDb Extractor.

- https://github.com/NP-Plug-and-Play-Scripts/NpDbExtractor-runnable
- https://github.com/NP-Plug-and-Play-Scripts/NpDbExtractor

InChIKey Pipeline:

While not described in the project itself, a standalone version of the InChIKey creator was made to help O. Hoekstra and S. Stokman with the creation of InChIKeys. Feel free to use this pipeline. On the github there should also be an added manual on how to set it up and run it.

- https://github.com/NP-Plug-and-Play-Scripts/inchiKeyCreatorPipeline

CFM-Workplace installation:

In order to use the pipeline a workplace installer script was made this makes a folder containing all the dependencies for CFM-id, along with extra folder for storage of the input and results.

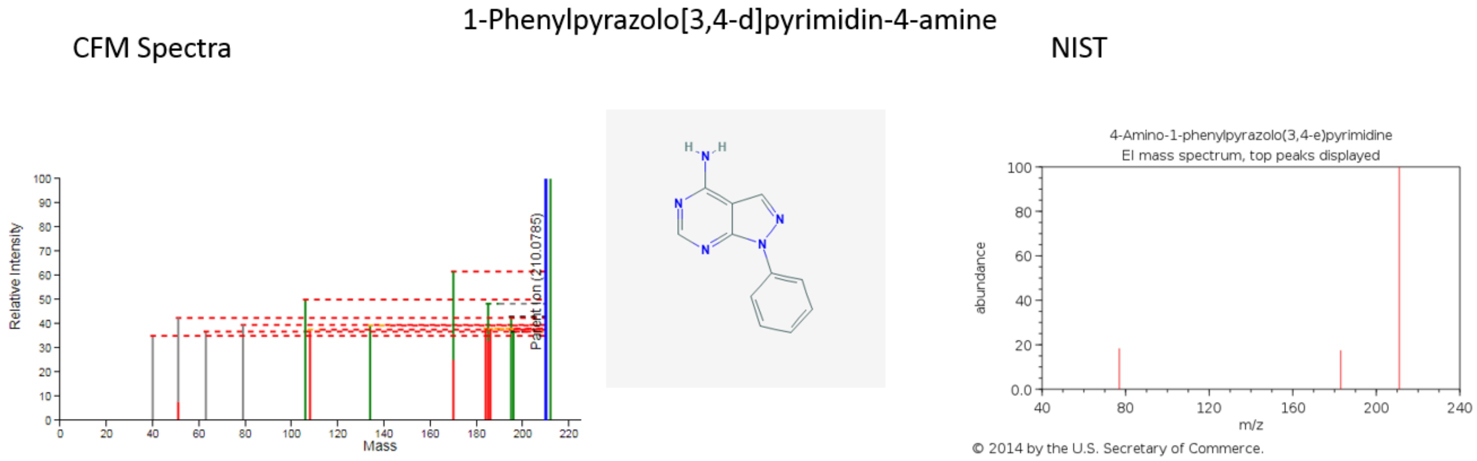
https://github.com/NP-Plug-and-Play-Scripts/Bash-scripts/blob/master/cfm-install.sh

NpDb Expansion

An expansion to the existing NPDB. Will add 6 tables that can store the cfm spectra as well as the motifs created with MS2LDA. Project contains a manual explaining how to run the data as well as the required scripts.

https://github.com/NP-Plug-and-Play-Scripts/NpDb_expansion

CFM-ID Additional tables and graphs


CFM Spectra

1-Phenylpyrazolo[3,4-d]pyrimidin-4-amine

NIST

Appendix figure 1: Another example of a cfm spectra compared to a spectra obtained from NIST. While the peaks around 80 and 180 appear in both the spectra, the CFM predicted spectra has more peaks in this case. Showing that it's not always the case that real spectra have more peaks. In the comparisons made it seemed to vary per comparison, this again could be due to the trained model.

MS2LDA Additional info tables and graphs

Experiment numbers of the created and used data:

All experiments should be available at http://ms2lda.org/basicviz/. At the time of writing this report, opening the larger datasets makes the server overload from time to time so open with caution.

Data	Experiment id on MS2LDA
cfm_Flavonoids2 (most used dataset)	839
cfm_Flavonoids	838
cfm_neutralised-Steroids	810
cfm_Lactones_neutralised	818
cfm_azoles_neutralised	821
Amino_acids_and_peptide_analogs_part1of4	875
Amino_acids_and_peptide_analogs_part2of4	883

Additional tables

All Motifs with Matches above 0.7 probability including their annotation.

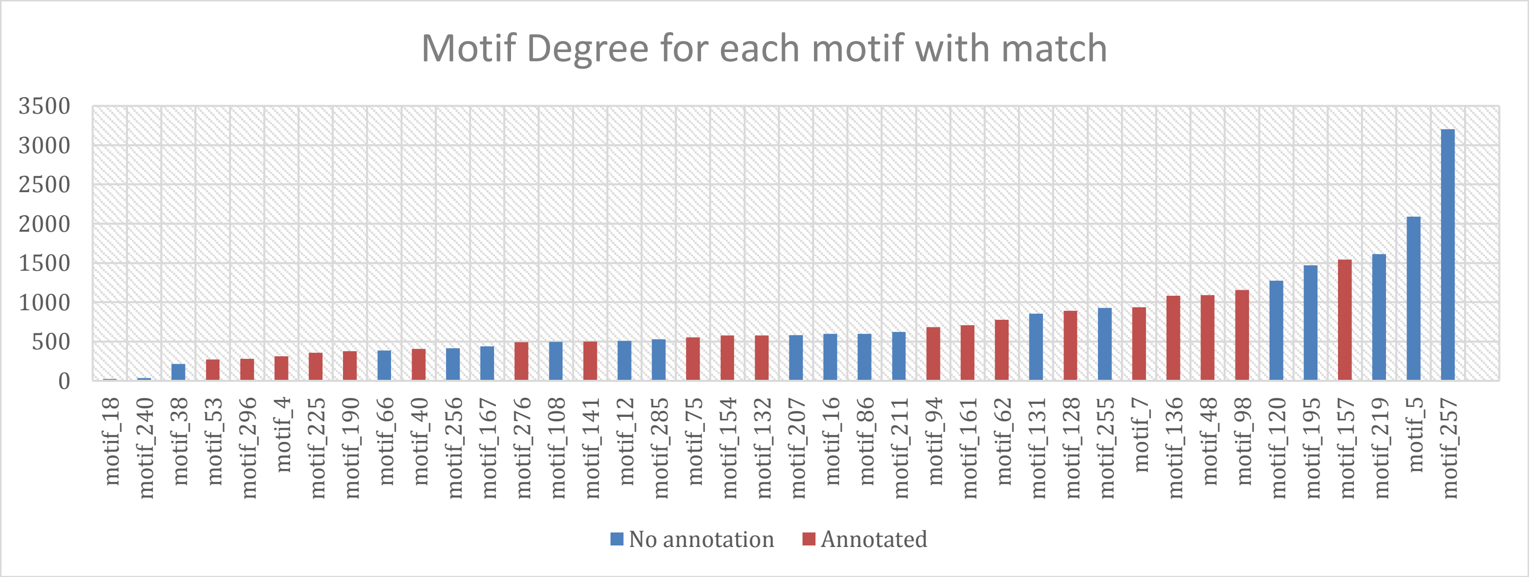
Mass2Motif (Annotation)	Best Match (Annotation)	Best Match (Experiment)	Match Score
motif_108 (None)	motif_355 (None)	Foodomics_beverage_5_10_0_100_400_1000	0.707
motif_12 (None)	motif_108 (None)	GlobalEuphorbiaStudy	0.702
motif_120 (None)	motif_102 (None)	Foodomics_beverage_5_10_0_100_400_1000	0.96
motif_128 (protocatechuoyl-related)	gnps_motif_19.m2m ((5-Hydroxy-2 2-dimethyl-4-oxo-3 4-dihydro-2H-chromen-7-yl)oxy substructure)	GNPS-MS2LDA_integration_Rhamnaceae_noMS1PeakListProvided_withMotifDB	0.967
motif_128 (protocatechuoyl-related)	motif_117 (protocatechuoyl-related)	Rhamnaceae_plant_extracts_KyoBin_200Motifs_MS1_peaktable	0.972
motif_128 (protocatechuoyl-related)	motif_177 (None)	Foodomics_beverage_5_10_0_100_400_1000	0.977
motif_131 (None)	motif_113 (None)	Rhamnaceae_plant_extracts_KyoBin_200Motifs_MS1_peaktable	0.842
motif_132 (vanilloyl-related)	motif_191 (vanilloyl-related)	Rhamnaceae_plant_extracts_KyoBin_200Motifs_MS1_peaktable	0.898
motif_132 (vanilloyl-related)	motif_252 (None)	GNPS-MS2LDA_integration_Rhamnaceae_noMS1PeakListProvided_withMotifDB	0.838

motif_136 (Fragments indicative for kaempferol/glycosylated kaempferol substructure)	mb_motif_42.m2m (Fragments indicative for kaempferol/glycosylated kaempferol substructure)	GNPS-MS2LDA_integration_Rhamnaceae_noMS1PeakListProvided_withMotifDB	0.997
motif_136 (Fragments indicative for kaempferol/glycosylated kaempferol substructure)	mb_motif_42.m2m (Fragments indicative for kaempferol/glycosylated kaempferol substructure)	MotifDB	0.997
motif_141 ([1 3-dihydro-2-benzofuran-1-yl]pyrrolidine substructure)	gnps_motif_34.m2m ([1 3-dihydro-2-benzofuran-1-yl]pyrrolidine substructure)	GNPS-MS2LDA_integration_Rhamnaceae_noMS1PeakListProvided_withMotifDB	0.891
motif_154 (Possible Chromone fragment an isomer of Coumarin - C9H6O2)	gnps_motif_37.m2m (Fragments indicative for cinnamic/hydroxycinnamic acid substructure)	GNPS-MS2LDA_integration_Rhamnaceae_noMS1PeakListProvided_withMotifDB	0.714
motif_154 (Possible Chromone fragment an isomer of Coumarin - C9H6O2)	mb_motif_20.m2m (Fragments indicative for cinnamic/hydroxycinnamic acid substructure)	MotifDB	0.714
motif_154 (Possible Chromone fragment an isomer of Coumarin - C9H6O2)	motif_124 (Fragments indicative for cinnamic/hydroxycinnamic acid substructure)	Foodomics_beverage_5_10_0_100_400_1000	0.992
motif_154 (Possible Chromone fragment an isomer of Coumarin - C9H6O2)	motif_444 (None)	GlobalEuphorbiaStudy	0.746
motif_154 (Possible Chromone fragment an isomer of Coumarin - C9H6O2)	motif_62 (None)	Rhamnaceae_plant_extracts_KyoBin_200Motifs_MS1_peaktable	0.863
motif_157 (Benzaldehyde (105.0325) fragment C7H6O)	motif_128 (Benzoyl substructure (likely from N-benzoyl) related Mass2Motif)	Urine38_POS_mzML_standardLDA_005binned	0.92
motif_157 (Benzaldehyde (105.0325) fragment C7H6O)	motif_144 (None)	GlobalEuphorbiaStudy	0.972
motif_157 (Benzaldehyde (105.0325) fragment C7H6O)	motif_172 (None)	Foodomics_beverage_5_10_0_100_400_1000	0.981
motif_16 (None)	motif_112 (None)	Urine38_POS_mzML_standardLDA_005binned	0.786
motif_161 (Coumaric acid - H2O)	motif_120 (Coumaric acid - H2O)	Rhamnaceae_plant_extracts_KyoBin_200Motifs_MS1_peaktable	0.927
motif_161 (Coumaric acid - H2O)	motif_204 (None)	GNPS-MS2LDA_integration_Rhamnaceae_noMS1PeakListProvided_withMotifDB	0.813
motif_167 (possible 1,3-Butadiene,Butane or Butene related fragment(C4H8) 53.0375)	motif_295 (None)	Urine38_POS_mzML_standardLDA_005binned	0.969
motif_18 (None)	motif_214 (None)	Urine38_POS_mzML_standardLDA_005binned	0.785
motif_190 (Sugar related (small) fragments)	motif_134 (Sugar related (small) fragments)	Urine38_POS_mzML_standardLDA_005binned	0.945
motif_195 (None)	motif_112 (None)	Rhamnaceae_plant_extracts_KyoBin_200Motifs_MS1_peaktable	0.896
motif_207 (None)	motif_85 (None)	Foodomics_beverage_5_10_0_100_400_1000	0.992
motif_211 (None)	motif_55 (None)	Rhamnaceae_plant_extracts_KyoBin_200Motifs_MS1_peaktable	0.872
motif_219 (None)	motif_134 (None)	Rhamnaceae_plant_extracts_KyoBin_200Motifs_MS1_peaktable	0.871
motif_219 (None)	motif_191 (None)	GNPS-MS2LDA_integration_Rhamnaceae_noMS1PeakListProvided_withMotifDB	0.827
motif_225 (Fragment indicative for aromatic compounds related to methylbenzene substructure (C7H7 fragment))	gnps_motif_52.m2m (Fragment indicative for aromatic compounds related to methylbenzene substructure (C7H7 fragment))	GNPS-MS2LDA_integration_Rhamnaceae_noMS1PeakListProvided_withMotifDB	0.891
motif_225 (Fragment indicative for aromatic compounds related to methylbenzene substructure (C7H7 fragment))	mb_motif_30.m2m (Fragment indicative for aromatic compounds related to methylbenzene substructure (C7H7 fragment))	MotifDB	0.891
motif_225 (Fragment indicative for aromatic compounds related to methylbenzene substructure (C7H7 fragment))	motif_114 (Fragment indicative for aromatic compounds related to methylbenzene substructure (C7H7 fragment))	Urine38_POS_mzML_standardLDA_005binned	0.89
motif_225 (Fragment indicative for aromatic compounds related to methylbenzene substructure (C7H7 fragment))	motif_207 (None)	Foodomics_beverage_5_10_0_100_400_1000	0.707
motif_240 (None)	gnps_motif_43.m2m (Water loss - indicative of a free hydroxyl group	GNPS-MS2LDA_integration_Rhamnaceae_noMS1PeakListProvided_withMotifDB	0.743

	â€“ (in beer often seen in sugary structures))		
motif_240 (None)	mb_motif_22.m2m (Water loss - indicative of a free hydroxyl group â€“ (in beer often seen in sugary structures))	MotifDB	0.743
motif_240 (None)	motif_173 (None)	Rhamnaceae_plant_extracts_KyoBin_200Motifs_MS1_peaktable	0.705
motif_240 (None)	motif_206 (None)	Urine38_POS_mzML_standardLDA_005binned	0.743
motif_240 (None)	motif_415 (None)	GlobalEuphorbiaStudy	0.719
motif_255 (None)	motif_99 (None)	Foodomics_beverage_5_10_0_100_400_1000	0.961
motif_256 (None)	motif_185 (None)	Rhamnaceae_plant_extracts_KyoBin_200Motifs_MS1_peaktable	0.737
motif_257 (None)	motif_3 (None)	Urine38_POS_mzML_standardLDA_005binned	0.932
motif_257 (None)	motif_425 (None)	GlobalEuphorbiaStudy	0.951
motif_276 (Fragments indicative for ethylphenol (i.e. resulting from Tyramine in beer) or the structurally related paramethylmethoxybenzene (MassBank) substructure)	gnps_motif_21.m2m (Fragments indicative for ethylphenol substructure (i.e. resulting from Tyramine â€“ MzCloud))	GNPS-MS2LDA_integration_Rhamnaceae_noMS1PeakListProvided_withMotifDB	0.928
motif_276 (Fragments indicative for ethylphenol (i.e. resulting from Tyramine in beer) or the structurally related paramethylmethoxybenzene (MassBank) substructure)	mb_motif_19.m2m (Fragments indicative for ethylphenol (i.e. resulting from Tyramine in beer) or the structurally related paramethylmethoxybenzene (MassBank) substructure)	MotifDB	0.928
motif_276 (Fragments indicative for ethylphenol (i.e. resulting from Tyramine in beer) or the structurally related paramethylmethoxybenzene (MassBank) substructure)	motif_343 (None)	GlobalEuphorbiaStudy	0.913
motif_285 (None)	motif_16 (None)	Rhamnaceae_plant_extracts_KyoBin_200Motifs_MS1_peaktable	0.732
motif_296 (None)	motif_282 (Mass2Motif related to caffeoylquinic acids (177 and 145 mass fragments))	Foodomics_beverage_5_10_0_100_400_1000	0.787
motif_296 (None)	motif_31 (None)	GlobalEuphorbiaStudy	0.807
motif_38 (None)	motif_2 (None)	Urine38_POS_mzML_standardLDA_005binned	0.801
motif_4 (Flavonoid core fragments (m/z 151))	motif_140 (Flavonoid core fragments (m/z 151))	Rhamnaceae_plant_extracts_KyoBin_200Motifs_MS1_peaktable	0.749
motif_4 (Flavonoid core fragments (m/z 151))	motif_202 (None)	GNPS-MS2LDA_integration_Rhamnaceae_noMS1PeakListProvided_withMotifDB	0.735
motif_40 (None)	gnps_motif_24.m2m (Fragments indicative for tyrosine related substructure (MzCloud))	GNPS-MS2LDA_integration_Rhamnaceae_noMS1PeakListProvided_withMotifDB	0.723
motif_40 (None)	motif_301 (None)	Foodomics_beverage_5_10_0_100_400_1000	0.852
motif_48 (possible Luteolin, Kaempferol like structures (C15H10O6) often found in plants)	motif_104 (None)	Rhamnaceae_plant_extracts_KyoBin_200Motifs_MS1_peaktable	0.816
motif_48 (possible Luteolin, Kaempferol like structures (C15H10O6) often found in plants)	motif_175 (None)	GNPS-MS2LDA_integration_Rhamnaceae_noMS1PeakListProvided_withMotifDB	0.873
motif_5 (None)	motif_224 (None)	Urine38_POS_mzML_standardLDA_005binned	0.784
motif_5 (None)	motif_80 (None)	Rhamnaceae_plant_extracts_KyoBin_200Motifs_MS1_peaktable	0.782
motif_53 (4-Methyl-6-oxo-6H-benzo[c]chromen-3-yl substructure)	gnps_motif_71.m2m (4-Methyl-6-oxo-6H-benzo[c]chromen-3-yl substructure)	GNPS-MS2LDA_integration_Rhamnaceae_noMS1PeakListProvided_withMotifDB	0.843
motif_62 (phthalate substructure)	gnps_motif_64.m2m (phthalate substructure)	GNPS-MS2LDA_integration_Rhamnaceae_noMS1PeakListProvided_withMotifDB	0.888
motif_62 (phthalate substructure)	motif_245 (None)	GlobalEuphorbiaStudy	0.887
motif_62 (phthalate substructure)	motif_96 (None)	Rhamnaceae_plant_extracts_KyoBin_200Motifs_MS1_peaktable	0.781
motif_66 (None)	motif_52 (None)	Rhamnaceae_plant_extracts_KyoBin_200Motifs_MS1_peaktable	0.876
motif_7 (possible Veratrole or 4-Ethylcatechol 137.0575 fragment (C8H10O2) - 4-Ethylcatechol a constituent of roasted coffee and Veratrole is an insect attractant	motif_142 (None)	Foodomics_beverage_5_10_0_100_400_1000	0.96

created by plants, being the methylated form of guaiacol)			
motif_75 (Emodin related Motif)	motif_160 (None)	GNPS-MS2LDA_integration_Rhamnaceae_noMS1PeakListProvided_withMotifDB	0.919
motif_75 (Emodin related Motif)	motif_40 (Emodin related Motif)	Rhamnaceae_plant_extracts_KyoBin_200Motifs_MS1_peaktable	0.911
motif_86 (None)	motif_11 (None)	Rhamnaceae_plant_extracts_KyoBin_200Motifs_MS1_peaktable	0.879
motif_94 (Fragments indicative for dihydroxylated benzene ring substructure (MzCloud) â€“ C6H5O2 fragment corresponds to positively charged fragment with two hydroxyl groups.)	gnps_motif_55.m2m (Fragments indicative for dihydroxylated benzene ring substructure (MzCloud) â€“ C6H5O2 fragment corresponds to positively charged fragment with two hydroxyl groups.)	GNPS-MS2LDA_integration_Rhamnaceae_noMS1PeakListProvided_withMotifDB	0.868
motif_94 (Fragments indicative for dihydroxylated benzene ring substructure (MzCloud) â€“ C6H5O2 fragment corresponds to positively charged fragment with two hydroxyl groups.)	mb_motif_38.m2m (Fragments indicative for dihydroxylated benzene ring substructure (MzCloud) â€“ C6H5O2 fragment corresponds to positively charged fragment with two hydroxyl groups.)	MotifDB	0.868
motif_94 (Fragments indicative for dihydroxylated benzene ring substructure (MzCloud) â€“ C6H5O2 fragment corresponds to positively charged fragment with two hydroxyl groups.)	motif_131 (None)	Urine38_POS_mzML_standardLDA_005binned	0.857
motif_98 (possible Pyrocatechol fragment (C6H6O2))	motif_102 (None)	Rhamnaceae_plant_extracts_KyoBin_200Motifs_MS1_peaktable	0.778
motif_98 (possible Pyrocatechol fragment (C6H6O2))	motif_271 (None)	Foodomics_beverage_5_10_0_100_400_1000	0.908

Motif degree table and graph



Appendix : This figure shows the degree for each motif of the flavonoid data that has one or more matches. Degree means the amount of times this motif was found in the flavonoids2 dataset, this dataset contains around 4500 spectra so a degree around 450 would mean the motif is observed in 10% of all the spectra. The top motif 257 has a degree of around 3200 which is explained by it matching to a OH fragment which is obviously pretty common.

This table shows all the motifs with matches and their exact degree.

Motif	Degree	Motif	Degree	Motif	Degree
motif_257	3204	motif_211	621	motif_190	377
motif_5	2089	motif_16	597	motif_225	357
motif_219	1613	motif_86	597	motif_4	312
motif_157	1544	motif_207	582	motif_296	280
motif_195	1471	motif_132	578	motif_53	270
motif_120	1275	motif_154	576	motif_38	214
motif_98	1158	motif_75	554	motif_240	37
motif_48	1091	motif_285	528	motif_18	21
motif_136	1081	motif_12	507		
motif_7	936	motif_141	501		
motif_255	929	motif_108	495		
motif_128	891	motif_276	491		
motif_131	855	motif_167	437		
motif_62	777	motif_256	413		
motif_161	707	motif_40	406		
motif_94	684	motif_66	387		

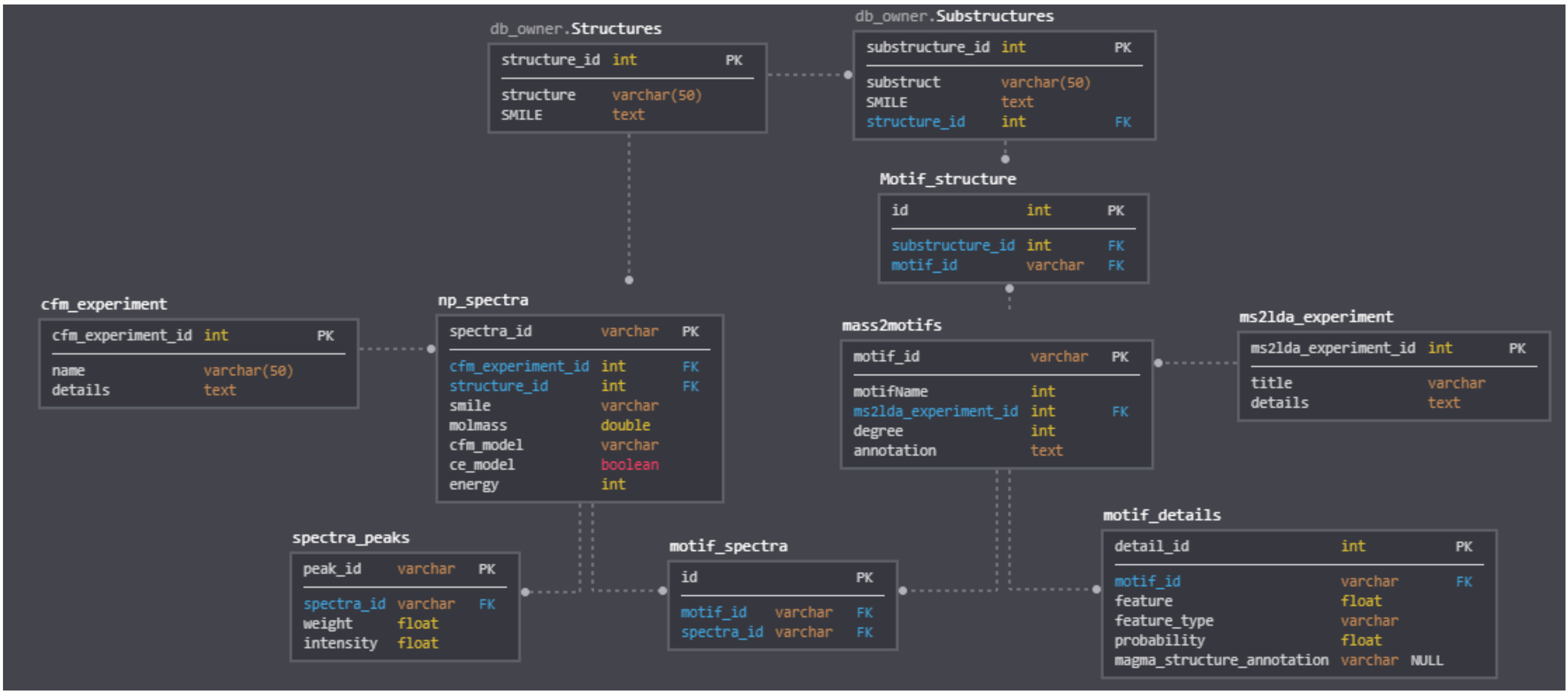
Matches Between Steroids (810) and Flavonoids2(839) – Showing only matches for unannotated structures or common structures.

	Best Match (Annotation)	Best Match (Experiment)	Match Score
motif_150 (None)	motif_257 (possible OH- fragment)	cfm_Flavonoids2	1
motif_202 (None)	motif_195 (None)	cfm_Flavonoids2	1
motif_281 (None)	motif_167 (possible 1,3-Butadiene,Butane or Butene related fragment(C4H8) 53.0375)	cfm_Flavonoids2	1
motif_178 (None)	motif_221 (None)	cfm_Flavonoids2	0.999
motif_388 (None)	motif_268 (None)	cfm_Flavonoids2	0.978
motif_2 (None)	motif_177 (None)	cfm_Flavonoids2	0.97
motif_87 (None)	motif_295 (None)	cfm_Flavonoids2	0.965
motif_230 (None)	motif_124 (None)	cfm_Flavonoids2	0.943
motif_22 (None)	motif_19 (None)	cfm_Flavonoids2	0.933
motif_263 (None)	motif_117 (None)	cfm_Flavonoids2	0.933
motif_187 (None)	motif_140 (None)	cfm_Flavonoids2	0.928
motif_111 (None)	motif_133 (None)	cfm_Flavonoids2	0.926
motif_357 (None)	motif_103 (None)	cfm_Flavonoids2	0.922
motif_63 (None)	motif_221 (None)	cfm_Flavonoids2	0.911
motif_247 (None)	motif_156 (None)	cfm_Flavonoids2	0.904
motif_349 (None)	motif_98 (possible Pyrocatechol fragment (C6H6O2))	cfm_Flavonoids2	0.895
motif_278 (None)	motif_131 (None)	cfm_Flavonoids2	0.889
motif_361 (None)	motif_234 (None)	cfm_Flavonoids2	0.864
motif_145 (None)	motif_168 (None)	cfm_Flavonoids2	0.83

motif_23 (None)	motif_191 (None)	cfm_Flavonoids2	0.821
motif_294 (None)	motif_175 (None)	cfm_Flavonoids2	0.816
motif_260 (None)	motif_58 (None)	cfm_Flavonoids2	0.811
motif_81 (None)	motif_191 (None)	cfm_Flavonoids2	0.794
motif_359 (None)	motif_3 (None)	cfm_Flavonoids2	0.793
motif_268 (None)	motif_16 (None)	cfm_Flavonoids2	0.786
motif_229 (None)	motif_18 (None)	cfm_Flavonoids2	0.779
motif_266 (None)	motif_27 (None)	cfm_Flavonoids2	0.768
motif_65 (None)	motif_157 (Benzaldehyde (105.0325) fragment C7H6O)	cfm_Flavonoids2	0.753
motif_72 (None)	motif_281 (None)	cfm_Flavonoids2	0.74
motif_354 (Mass2Motif related to methoxylated benzene ring)	motif_109 (C7H8O Fragment could be a Anisole, Para-cresol or Metacresol)	cfm_Flavonoids2	0.729
motif_333 (None)	motif_38 (None)	cfm_Flavonoids2	0.718
motif_98 (None)	motif_64 (None)	cfm_Flavonoids2	0.714
motif_189 (None)	motif_131 (None)	cfm_Flavonoids2	0.706
motif_68 (None)	motif_267 (None)	cfm_Flavonoids2	0.701

Natural Product database expansion

The figure below shows the new tables included in the database.



db_owner.Structures

structure_id	int	PK
structure	varchar(50)	
SMILE	text	

db_owner.Substructures

substructure_id	int	PK
substruct	varchar(50)	
SMILE	text	
structure_id	int	FK

Motif_structure

id	int	PK
substructure_id	int	FK
motif_id	varchar	FK

cfm_experiment

cfm_experiment_id	int	PK
name	varchar(50)	
details	text	

np_spectra

spectra_id	varchar	PK
cfm_experiment_id	int	FK
structure_id	int	FK
smile	varchar	
molmass	double	
cfm_model	varchar	
ce_model	boolean	
energy	int	

mass2motifs

motif_id	varchar	PK
motifName	int	
ms2lda_experiment_id	int	FK
degree	int	
annotation	text	

ms2lda_experiment

ms2lda_experiment_id	int	PK
title	varchar	
details	text	

spectra_peaks

peak_id	varchar	PK
spectra_id	varchar	FK
weight	float	
intensity	float	

motif_spectra

id		PK
motif_id	varchar	FK
spectra_id	varchar	FK

motif_details

detail_id	int	PK
motif_id	varchar	FK
feature	float	
feature_type	varchar	
probability	float	
magma_structure_annotation	varchar	NULL

Appendix : shows the new database additions. Main focus points are the np_spectra, spectra peaks mass2motifs and motif detail tables. These will contain the created spectra along with peaks and relevant data and the mass to motifs along with the fragments and losses, these also include extra details.

Currently the tables Motif_structure and Motif_spectra, which are linking tables, are not yet present in the database due to time constraints however it should not be hard to implement these.

MGF Format

Here are two examples obtained from MGF files, the first is one that comes straight out of CFM-id and the second is the result of the pipeline. (some peaks were left out to shorten the length of the examples a bit)

```
BEGIN IONS
PEPMASS=82.05309819
CHARGE=1+
TITLE=NP_ID_265938;Energy0;[M+H]+;In-silico MS/MS by CFM-ID;
42.03382555 1.166094178
52.01817548 1.431702494
54.03382555 7.291185703
56.04947561 8.850370785
66.03382555 1.53076755
83.06037464 77.30416099
END IONS
```

```
BEGIN IONS
IUPAC=Not_Added
ID=NP_ID_206371
TITLE=EnergyCombined 10eV 20eV 40eV;[M+H]+;In-silico MS/MS by CFM-ID;
PEPMASS=1343.350826
CHARGE=1+
SMILES=COc1cc(C=CC(=O)OC2C(OC3C(Oc4cc5c(OC6OC(COC(=O)CC(=O)O)C(O)C(O
)C6O)cc([O-])cc5[o+]c4-
c4ccc(O)c(O)c4)OC(COC(=O)C=Cc4ccc(OC5OC(CO)C(O)C(O)C5O)cc4)C(O)C3O)OCC(
O)C2O)cc(OC)c1[O-]
NEUTRAL_SMILES=COc1cc(C=CC(=O)OC2C(OC3C(Oc4cc5c(OC6OC(COC(=O)CC(=O
)O)C(O)C(O)C6O)cc(O)cc5[o+]c4-
c4ccc(O)c(O)c4)OC(COC(=O)C=Cc4ccc(OC5OC(CO)C(O)C(O)C5O)cc4)C(O)C3O)OCC(
O)C2O)cc(OC)c1O
InChIKey=QRNIDVBVORPNBX-UHFFFAOYSA-M
NEUTRAL_InChIKey=QRNIDVBVORPNBX-UHFFFAOYSA-O
41.00219107 418
68.99710569 345
87.00767038 453
147.0440559 320
165.0546206 333
179.0702706 421
1299.360996 428
1325.340261 900
1343.350826 492
END IONS
```