Chapter 7

Substructure Discovery in Tandem Mass Spectometry Data

7.1 Introduction

As the results from Chapter 5 shows, the ionization product (IP) types of many observed peaks are often unknown and therefore the molecular mass of metabolites that generate these peaks are also unknown. This makes identification difficult as mass is often a required information when querying metabolite identities against publicly-available databases, such as KEGG [72] and PubChem [73]. In addition, while modern mass spectrometry instruments can be highly accurate up to 3 parts-per-million (ppm), even a mass accuracy of 1 ppm is not sufficient to reliably determine the elemental composition (formula) of a metabolite [74] during database queries. The presence of isomers (metabolites having the same formula and mass but are structurally different from each other) suggests that relying on mass alone, the same peak might be incorrectly matched to multiple isomeric metabolites. Retention time (RT) may help to distinguish certain isomers that have different elution profiles, but RT drift, a main challenge in alignment, means observed RT values can vary across different chromatographic platforms and cannot be easily used as a characteristic information in public databases during identification. Apart from a small number of metabolites present in a standard solution that can be identified with a high degree of confidence (as they produce measured peaks having reliably known m/z and RT values), information on the mass and RT values alone are not enough to establish the identity of many metabolites in untargeted studies.

Fragmentation spectra are the results of chaining two stages of mass spetrometry steps. In data-dependent acquisition, a precursor or parent (MS1) peak is selected according to a certain criteria, frequently the top-N most intense peaks in a scan, for further fragmentation. This produces for each fragmented parent peak a distinct pattern of fragment (MS2) peaks.

Fragmentation patterns can be used to aid identification through the matching of a query spectrum to a database of reference spectra. In recent years, a growing number of fragmentation spectra databases have been made public, including METLIN [75], ChemSpider [76] and MassBank [77]. However, mass spectral databases are not comprehensive and contain only a small number of known metabolites. The large variance in submitted spectra further limits potential matches as sensible results can only be obtained when matching spectra generated from measurement platforms having similar characteristics (for e.g., produced through the same ionization method under a similar mass accuracy). According to [78], approximately 2% of spectra in an untargeted metabolomics experiment can be matched and subsequently identified – a small number in contrast of the vast collection of metabolites that comprise the metabolic pathways of an organism.

Multiple metabolites can share the same chemical substructure. For example, carboxylic acid (Figure 7.1) is a generic substructure shared by many amino acids and organic acids. During fragmentation, a characteristic fragment peak 46 Da away from the parent peak — representing the combined loss of CO and H2O due to the breaking of the neutral carboxyl group (COOH) from the molecular ion — can be expected to occur in the spectra of metabolites sharing carboxylic acid as a substructure. Knowledge of the constituent substructures that comprise a metabolite, particularly of the larger and more specific substructures, can also be used to provide a hint as to the overall identity of the metabolite. Classification method, such as Support Vector Machine, decision tree and neural networks [79, 80, 81, 82], have been trained to learn spectral features that represent substructures and predict the presence or absence of substructures from fragmentation spectra. Combined with information from the parent peak (such as the m/z, RT values and IP types if available), this provides additional information that can aid in the identification of metabolites that cannot be resolved through the traditional method of spectral database matching alone.

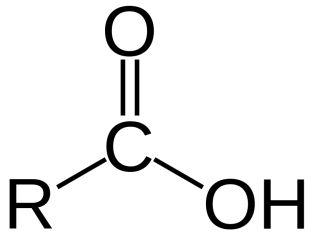

Figure 7.1: The carboxylic acid substructure. In the diagram, R refers to the residue, which is the rest of the metabolite attached to this substructure.

A common shortcoming of these classification approaches higlighted before is the need of the supervised training of the classifier (classification-based approaches may fail to generalise well to new dataset produced from different analytical platforms). Based on the assumption that fragmentation spectra contain fragment peaks that represent shared substructures

of metabolites, we propose a workflow that applies the Latent Dirichlet Allocation (LDA) model to spectral fragmentation data. The proposed workflow produces the decomposition of fragmentation spectra (equivalently a document in standard LDA) into the set of *Mass2Motifs* (equivalently a topic in standard LDA). Here, a Mass2Motif is defined to be the recurring set of fragment peaks and neutral losses that potentially correspond to a biochemically-relevant substructure shared by many metabolites. Unlike the classification-based methods highlighted earlier, the decomposition of fragmentation spectra into Mass2Motifs is achieved in an unsupervised manner. The MS2LDA workflow is introduced in Section 7.4.

7.2 Related Work

Clustering is commonly used for group fragmentation spectra that are similar to each other. Clusters of spectra can be used for identification by forming a consensus spectrum and matching it against spectral databases. Molecular networking clusters MS1 peaks by their MS2 spectral similarity such that one identifiable metabolite in a cluster facilitates structural annotation of its neighbors [83, 84, 85]. However, only MS2 spectra with high overall (e.g. cosine) spectral similarity are grouped in Molecular Networking. Consequently Molecular Networking may fail to group molecules that share small substructures. In particular, spectra may be placed in different clusters if they share a small number of fragment peaks that related to a common substruture, but their overall global similarities are too different. Even for spectra placed into the same cluster, often manual analysis (by eyes) is required to select the characteristic fragment peaks that represent a potential substructure and are shared by members of the clusters. Another package, MS2Analyzer [86] mines MS2 spectra given the prior knowledge on the fragment patterns of interest to be specified in advance. While generic features, such as CO or H2O losses, will be common to many experiments, sample-specific features can be easily overlooked if they have not been specified *a priori*.

The assumption that spectral consist of building blocks that correspond to substructures is alluded in certain works but not directly mined from the data. Prior knowledge on substructures have been used for the annotations of a small number of moelcules in fragmentation data [87] and for metabolite classification in GC-MS [88, 80]. In CSI:FingerID [82], a fragmentation tree is used to predict (using Support Vector Machine) the molecular 'fingerprint', computed through the implicit assumption that fragments share substructures, of an unknown compound. The resulting fingerprint is used to improve the matching of spectrum of the unknown compound against a vast chemical database (PubChem). Implicit in these methods are the assumption that recurring patterns of fragment peaks and neutral losses values explain the presence of common biological substructures (e.g. a hexose unit, or a CO loss) shared by metabolites.

Latent Dirichlet Allocation has not been applied to metabolomics or mass spectrometry data, but it has been applied to other fields of computational biology in e.g. genomics [89], metagenomics [90], and transcriptomics [91]. In [89], DNA sequence from genomics studies is decomposed into recurring patterns of N-mers nucleotides. A topic in this context corresponds to the set of N-mers (e.g. 'ATGC' as an instance of a 4-mers) that co-occur together across the different genomic sequences of a species, and the objective of the study is characterise the sets of N-mers that corresponds to conserved genes of the species. Similarly in [90], a metagenomic read (essentially a DNA sequence) is decomposed into its topic distribution. The unsupervised decomposition of metagenomic reads into topic distributions is used to improve the binning (clustering) of reads from the same species. In [91], a sample or gene from transcriptomics studies is decomposed into multiple processes in a manner similar to how a document is decomposed into different topics in traditional LDA for text.

7.3 Statement of Original Work

The work discussed in this chapter has been submitted for publication and is still under review. Justin van der Hooft (JvdH) performed the measurements of the Beer samples through mass spectometry, generating the set of fragmentation data that can be used for topic modelling. The author contributed to the design and development of the MS2LDA workflow. This includes the development and optimisation of the feature extraction process, the implementation and testing of inference via LDA and also model validation against multinomial mixture model.

JvdH then analysed the results from MS2LDA for biochemical significance. To assist JvdH in his analysis, the author proposed and developed the visualisation module, MS2LDAVis. To improve the visualisation module, the author integrated elemental formula annotation functionalities. This includes writing a wrapper in MS2LDA to call SIRIUS [92], a Java-based elemental formula annotator. Cristina Mihailescu (CM) implemented another Python-based elemental formula annotator, which was also customised and integrated into MS2LDA by the author.

JvdH then performed molecular networking analysis on the same dataset, which was used for comparison to MS2LDA results. The author performed the identification of metabolites through matching to reference standard compounds and also the differential analysis of Mass2Motifs, and JvdH validated the results.