Abstract

In recent years, the large-scale untargeted studies of compounds that serve as workers in the cell (proteins) and the by-products of essential life-sustaining chemical processes (metabolites) have provided insights into a wide array of fields, such as medical diagnostics, drug discovery, personalised medicine and many others. Measurements in such studies are routinely performed using liquid chromatography mass spectrometry (LC-MS) instruments, resulting in a set of peaks having mass-to-charge, retention time (RT) and intensity values. Before further analysis is possible, the raw LC-MS data has to be processed in a data pre-preprocessing pipeline. In the alignment step of the pipeline, peaks from multiple LC-MS measurements have to be matched. In the identification step, the identity of compounds that generate the observed peaks have to be assigned. Using tandem mass spectrometry, fragmentation peaks characteristic to a compound can be obtained and used to help establish the identity of the compound. Alignment and identification are challenging because the true identities of the entire set of compounds in the sample are unknown, and a single compound can produce many observed peaks, each with a potential drift in its retention time value. However, observed peaks are not independent — there exists structural dependencies among the observed peaks as multiple peaks are related through being attributed to the same underlying compound.

The aim of this thesis is to introduce methods in which related peaks can be grouped and to use these groupings to improve alignment and assist in identification. Firstly, we introduce a generative model to group related peaks by their retention time and use this information to influence direct-matching alignment, bringing related peak groups closer during matching. Investigations using benchmark datasets shows that this approach produces a better alignment result (Chapter 4). Secondly, we consider mass information and introduce a model that performs the grouping of related peaks in the same LC-MS run by explainable mass relationships, RT and intensity. Through a second-stage matching process, the resulting groups from the model can be used to produce better alignment results on benchmark metabolomics datasets, In addition, uncertainties in matched peaksets can also be extracted from the model (Chapter 5). Next, we improve upon the two-stage process described before and introduce a model that performs the flexible clustering of related peaks within and across multiple LC-MS runs at once, allowing for the matching of peaks and their respective uncertainties to be naturally extracted from the model (Chapter 6). Finally, we look at fragmentation data and introduce the application of topic modelling to model groups of related fragmentation peaks that potentially correspond to substructures shared by metabolites (Chapter 7). This final section corresponds to work in progress and points to many interesting avenues for future research.

metabolites are much more than that! (:

assays / methods
Instrument = mass spectrometer

difficult phrase
why not: by shared common biochemical substructures

Chapter 1

Introduction

Liquid chromatography combined with mass spectrometry (LC-MS) has emerged as one of the most popular methods of measurements in the untargeted study of proteins (proteomics) and metabolites (metabolomics). Proteins and metabolites serve as crucial building blocks in the body and play a vital role in the cellular maintenance of any organism. Metabolomics in particular is regarded as the -omics that is the closest to the phenotype: changes to the physical traits of an organism is often expressed in the metabolome. Understanding and characterising the proteome and metabolome provide important insights into the working of any biological system.

Before the raw LC-MS data can be used for further analysis, it has to be processed in a data pre-processing pipeline. This starts from the initial step of peak detection, where the observed peaks having m/z, retention time (RT) and intensity values are extracted from the raw data. The two important steps that follow after peak detection are the alignment and identification steps. In most studies, multiple samples are obtained and measured (producing biological replicates) or alternatively a sample is run through the LC-MS instruments multiple times (producing technical replicates). Alignment refers to the matching of these peaks across multiple LC-MS runs. In identification, we seek to associate the information on which compounds generate the observed peaks. Fragmentation data, where parent peaks are processed through a second-stage mass spectrometry, provides an additional information as to the identity of a compound in the form of patterns of fragment peaks that are characteristic to the compound.

In many cases, LC-MS data pre-processing is challenging. The lack of knowledge in the complete composition of compounds in a sample means that we do not know for certain which compounds are present in the sample. Compounds ionise differently during mass spectrometry, while a single compound can produce multiple observed peaks, making data interpretation difficult as there is no one-to-one correspondence between the observed peaks and the compounds that generate them. While the m/z information of a peak is generally

preserved across runs, retention time drift means the observed RT values can vary among peaks produced on different instruments or even peaks produced on the same instrument but measured at a different time period. This makes alignment difficult. Identification using fragmentation data is also hampered by the limited coverage of spectral databases to compare the observed fragmentation pattern against.

However, peaks generated from the same underlying compound are not independent. They are structurally related in a chemical manner e.g. through being the ionisaton product peaks of the same compound. We reason that this structural dependencies can be used to improve alignment. In a similar manner, fragmentation spectra, which provides the characteristic fingerprints of compounds, also contains structural information where a subset of fragment peaks may correspond to a shared chemical substructure in a class of compounds. In this thesis, we show that through generative modelling, the structural dependencies of these peaks can be revealed and exploited to improve or enhance the alignment and identification steps. Moreover through generative modelling, alignment uncertainties can also be quantified, allowing the user to control the level of uncertainty they desire from matched peaksets.

1.1 Thesis Statement

Untargeted liquid chromatography mass spectrometry data pre-processing is a challenging task that is often subjected to errors and inaccuracies. Much of this can be attributed to the complexity of the LC-MS data itself and also to the lack of knowledge as to which compounds are present in the sample. However, the structural dependencies in the observed peak data means that through generative modelling, we can explain the relationships between peaks, allowing us to produce groups of related peaks that can be used to improve or enhance the alignment and identification steps of LC-MS data pre-processing.

1.2 List of Contributing Papers

The work described in this thesis has led to the following publication:

1. Wandy, J., Daly, R., Breitling, R., Rogers, S. (2015). "Incorporating peak grouping information for alignment of multiple liquid chromatography-mass spectrometry datasets". *Bioinformatics*, 31(12), 1999-2006.

Additionally the following manuscripts are still under review:

1. Wandy, J., van der Hooft, J. J., Rogers, S. (2016). "Ionization Product Clustering to Improve Peak Alignment in LC-MS-based Metabolomics" submitted to *Bioinformatics*.
2. van der Hooft, J. J., Wandy, J., Barrett, M., Burgess, K. V., Rogers, S. (2016). "Topic Modeling for Untargeted Substructure Exploration in Metabolomics" submitted to *Proceedings of the National Academy of Sciences* (PNAS).

Chapter 4 of this thesis is based on the first published paper, Chapters 5 and 7 are based on the two manuscripts that are under review.

1.3 Overview of Thesis and Research Contributions

The contributions of this thesis are:

- A method that combines direct-matching and related peak grouping information to improve alignment.
- A generative model that groups related peaks in the same LC-MS run by their ionisation product (IP) relationships, producing IP clusters. This is described alongside methods that use the resulting IP clusters to produce a better alignment.
- A generative model that groups related peaks in the same and across LC-MS runs in a flexible manner. From this model, we can extract alignment and furthermore, it allows for the probabilities of matching of certain peaksets to be quantified.
- A study to the application of topic modelling to model substructures in fragmentation data.

The remainder of this thesis is structured as follows:

- **Chapter 2** discusses the background literature that this thesis is built upon. In particular, the chapter explains the nature of the LC-MS data and the necessary pre-processing steps before the data can be used for further analysis, including the challenges faced in the data pre-processing steps.
- **Chapter 3** introduces probabilistic modelling, with a particular focus on the construction of mixture models and other related generative models that are used in the rest of the thesis.

- **Chapter 4** presents an approach that combines matching and clustering information to produce a better alignment result.
- **Chapter 5** presents a generative model that performs the clustering of ionisation product peaks into IP clusters, and introduces ways these IP clusters can be used to produce a better alignment result.
- **Chapter 6** introduces a hierarchical mixture model that can be used to perform the clustering of related ionisation product peaks across multiple runs, allowing for the uncertainties in the matching of peaks to be extracted from the model.
- **Chapter 7** presents the application of topic modelling to capture the structural dependencies of peaks in fragmentation data. It also introduces a visualisation module that can be used to aid in the analysis of the result from the model.
- **Chapter 8** presents a summary of the work and contributions. It also highlight the avenues for future research based on the work done so far, and finally it concludes this thesis.

Chapter 2

Computational Mass Spectrometry Analysis

2.1 Introduction

The three major types of macromolecules that are fundamentally essential to all life on Earth: deoxyribonucleic acid (DNA), ribonucleic acid (RNA) and proteins. The central dogma of molecular biology states that *DNA is transcribed into RNA, which is translated into proteins.* Since its initial proposal, the central dogma model has been challenged and expanded to acknowledge other factors that can influence the transcription and translation processes. For instance, the reverse flow of information from RNA to DNA is possible but was not in the initial model. Nevertheless, the central dogma is broadly useful to explain how genetic information can flow in a biological system, starting from DNA to RNA to proteins.

DNA is the basic storage unit of genetic information. In a rather simplified view, the flow of information in a biological system begins from the double-helix strands of the DNA as the starting point. A DNA strand consisting of a series of linked nucleotides subunits. Each nucleotide is a molecule composed of a sugar molecule (deoxyribose), a phoshoric acid and a nitrogenous base. The base in DNA can be either adenine (A), thymine (T), guanine (G) or cytosine (C), and together they form the four well-known 'alphabets' of the DNA. Bases are complementary in their pairing through hydrogen bonds, such that A pairs only with T, and G with C. It is this pairing that produces the double helix structure of the DNA.

Regions of the DNA that code for specific proteins are called genes, however DNA is not the direct template for protein synthesis. Rather, DNA is *transcribed* into RNA. The same information is encoded in RNA as its originating DNA strand, but with the crucial difference that the subunits (nucleotides) of RNA has ribose as the sugar molecule and uracil substituted in place of thymine as one of the bases. In this manner, the four alphabets of RNA are adenine

(A), uracil (U), guanine (G) and cytosine (C).

After the transcription process, a class of RNA molecules known as the messenger RNA (mRNA) serves as the template for protein synthesis. Compared to the relatively inert DNA, mRNA is biochemically active and allows for genetic information to be transferred to outside the nucleus. The ribosome, a part of the translational apparatus of the cell, then reads mRNA and *translates* it into proteins. A sequence of three RNA nucleotides, terms a codon, codes for a particular amino acid, which is the building block of proteins. Proteins serve critical roles in an organism by participating in nearly all cellular processes: performing cellular maintenance, catalysing chemical reactions and carrying other functions essential to life. Proteins also serve as the biochemical machineries involved in carrying out DNA replication and the transcription and translation processes themselves to produce more proteins.

In total, there are 20 different types of amino acids used as the building blocks of proteins (Table 2.1). By allowing multiple codons to encode for the same amino acid, redundancies are built to deal with transcription errors. For instance both 'AAT' and 'AAC' codons correspond to the asparagine amino acid. An amino acid consists of a central carbon atom surrounded by an amine group ($-NH_2$), a carboxylic group (-COOH) and a side chain specific to the amino acid. Through the loss of water molecule, amino acids can be chained to each other through peptide bonds. A short chain of amino acid residues form a peptide, and in a longer chain, they fold into a fixed structure to form a protein. The function of a protein is directly determined by its three-dimensional structure. As each amino acid can be described by a unique letter drawn from a set of 20 chemical alphabets in Table 2.1, a protein can be succinctly described by a string of its peptides.

Apart from proteins, numerous other chemical reactions essential for sustaining life also happen inside a cell, including crucially, the breaking of organic compounds into energy and the production of other cellular building blocks involved in the transcription and translation processes. Together these chemical reactions comprise the *metabolism* of an organism. In catabolic reactions, large organic molecules within a cell are broken into energy and smaller molecules. These serve as the input to anabolic reactions, producing the basic building blocks of a cell such as proteins and nucleic acids. Both anabolic and catabolic reactions are usually catalysed by enzymes, and together these two reactions comprise the metabolism of an organism. *Metabolites* are small molecules (usually defined as less than 1000 Da) involved during or produced as the by-products of metabolism. Through the help of various enzymes, metabolites are transformed from one form to another in a series of chemical reactions as part of the metabolic pathways. Some examples of common metabolites are the various amino acids, fatty acids, vitamins, carbohydrates and many others. The overall set of metabolites that can be found within an organism is collectively called the *metabolome*.

As illustrated in Figure 2.1, each sub-field of computational biology focuses on the entities

Amino Acids	RNA Codons	Amino Acids	RNA Codons
Isoleucine (I)	AUU, AUC, AUA	Serine (S)	UCU, UCC, UCA, UCG, AGU, AGC
Leucine (L)	CUU, CUC, CUA, CUG, UUA, UUG	Tyrosine (Y)	UAU, UAC
Valine (V)	GUU, GUC, GUA, GUG	TrypUophan (W)	UGG
Phenylalanine (F)	UUU, UUC	Glutamine (Q)	CAA, CAG
Methionine (M)	AUG	Asparagine (N)	AAU, AAC
Cysteine (C)	UGU, UGC	Histidine (H)	CAU, CAC
Alanine (A)	GCU, GCC, GCA, GCG	Glutamic acid (E)	GAA, GAG
Glycine (G)	GGU, GGC, GGA, GGG	Aspartic acid (D)	GAU, GAC
Proline (P)	CCU, CCC, CCA, CCG	Lysine (K)	AAA, AAG
Threonine (T)	ACU, ACC, ACA, ACG	Arginine (R)	CGU, CGC, CGA, CGG, AGA, AGG

Table 2.1: The 20 amino acids and the RNA codons that encode them.

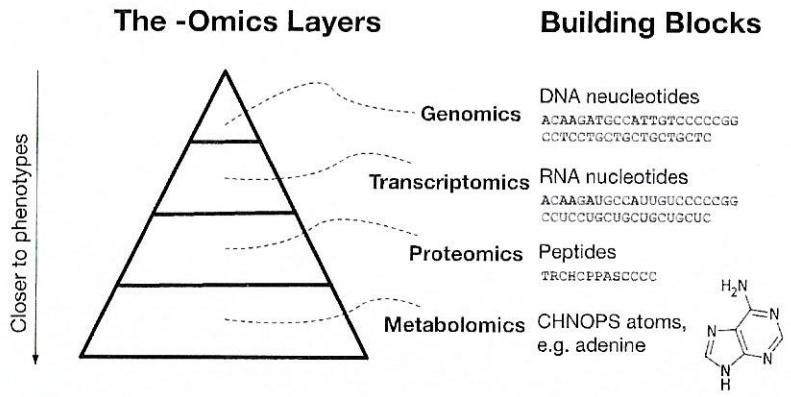

Figure 2.1: The layers of -omics and their building blocks.

and processes involved in a stage of the central dogma. Genomics is concerned with the large-scale study of the entire DNA in the organism (the genome) and how the genes encoded in the genome interact with each other. Transcriptomics focuses on understanding the complete set of mRNA (the transcriptome), particularly those that correspond to protein-encoding genes and measurements on their abundance in the sample. Proteins and their large-scale identifications and quantifications are studied in proteomics. Metabolomics studies the metabolome on a large scale, usually for the purpose of identifying and quantifying the differences of metabolite compositions in a particular organism or tissue under various experimental or physiological conditions.

Moving through the successive -omics layers in Figure 2.1 and getting closer the phenotype introduce greater complexity due to the increased number of ways to putting the building blocks of each -omics layer together. The building blocks of the genome are the nucleotides of the DNA, while in the transcriptome, the building blocks are the nitrogenous bases that comprise the RNA. There are only four possible alphabets in the genome and transcriptome. In proteomics, the object of interest, proteins, is a chain of amino acid residues. There are 20 possible alphabets of amino acids residues listen in Table 2.1. The small molecules in metabolomics have atoms as their building blocks, with the elements Carbon, Hydrogen, Nitrogen, Oxygen, Phospor and Sulphur (CHNOPS) that can be arranged in many chemically-plausible configurations. Furthermore, unlike the genome that is relatively static, the proteome and metabolome of an organism are also considerably more dynamic. The expression of proteins and metabolites are governed by various complex, interacting factors. In a process called post-translational modification [3], proteins can be chemically modified after synthesis in a way that completely alters its structure and folding stability, e.g. through phosphorylation (the addition of a phoshate group) or methylation (the addition of a methyl group). Metabolites expression can also change in response to the cellular sytems cellular [4] or environmental factor [5]. As a result, the knowledge of the DNA sequence alone is not sufficient to predict the proteins and metabolites that may be expressed in an organism. However, the metabolome is considered closest to the physically observed properties (phenotypes) of that organism [6], so changes to phenotype are often most readily observed in the metabolome. Studying the metabolome therefore provides us with an instantaneous 'snapshot' of the chemical activities that occur in the cell, leading to an understanding of how cellular processes behaves and possibly an explanation of how certain phenotypes are expressed.

2.2 Measurement Technologies

Sequencing technologies, in particular next-generation sequencing (NGS) machines such as Illumina and Ion Torrent, have been instrumental in revolutionising genomics by making possible the high-throughput and rapid sequencing of the entire DNA sequence from a sample [7]. Transcriptome relies on DNA micro-array technologies and more recently, have been increasingly performed by NGS sequencing as well. Nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS) are the two widely used measurement technologies for proteomics and metabolomics. Before we can understand the principle behind NMR spectroscopy and mass spectrometry, we need to take a detour and talk about atoms.

Atoms are the small building blocks of matter. An atom has a nucleus at the centre, which consists of positively charged protons and neutrons with no charge. Electrons, having negative charge, are bound to the neucleus through electromagnetic force. The overall charge of the atom is therefore determined by the number of electrons and protons that it has. The atom is called a positive ion when there are more protons than electrons, otherwise it is a negative ion. Two or more atoms held via chemical bonds comprise a compound. The molecular mass of a compound is the sum of the molecular mass of its elements, measured in Dalton (Da), where one Da is $\frac{1}{12}$ of the molecular mass of the carbon element (^{12}C). Elements in nature occur as isotopes. Isotopes are naturally occurring elements that have the same number of protons (same atomic number) but different number of neutrons (different molecular masses). Each elements has many isotope species, for instance carbon has two isotopes: ^{12}C with molecular mass 12.000000 at 98.890% abundance in nature, and ^{13}C with molecular mass 13.003355 and 1.110% abundance.

NMR spectroscopy operates on the principle of measuring the energy absorption of certain nuclei as radio frequency is applied. The nucleus of an atom possesses an angular moment, called spin. A nucleus with a spin of 1/2 develops a magnetic field, and when placed in an external magnetic field, a nuclei can either align itself with the external field (a lower energy state) or against the external field (a higher energy state). In NMR spectroscopy, initially most nuclei will be in their ground state of being in alignment with the external magnetic field, but when radio waves are applied, the nuclei in the lower energy state can absorb the energy and move to the higher energy state (their spin flip). When the radio waves are removed, the energised nuclei relaxes back to the lower energy state. The fluctuation of the magnetic field during relaxation is called 'resonance' and can be measured in the form of a current in the magnetic coil around the sample, resulting in peaks in an NMR spectrum. Many isotopes naturally occurring in an organic compound, e.g. ^{1}H and ^{13}C, have a spin of 1/2 and can therefore be measured by NMR spectroscopy. From NMR measurements, signals in the time-domain is obtained. The signal is usually converted using Fourier transform from the time to the frequency domain. The resulting NMR spectrum is then processed in

a data pre-processing pipeline, which typically includes steps like baseline correction, noise filtering, peak alignment, etc. [8].

As an alternative to NMR spectroscopy, mass spectrometry operates by ionising compounds in the sample, producing charged ions that are separated by their mass-to-charge (m/z) ratio. During mass spectrometry, the compounds to be analysed (metabolites or peptide fragment) are introduced into the ionisation source of the MS, and depending on the ionisation mode used, these compounds produce positively or negatively charged ions. They travel through the mass analyser and arrive at the detector at a different rate due to each ion having different mass-to-charge (m/z) ratios. The detector measures the ions that arrive and produce signals in form of a mass spectrum, showing the relative abundance of detected ions at different m/z ratios. MS instruments can be ranked by the ascending order of their resolving powers of their mass analyser: **(1)** time-of-flight MS, **(2)** quadrupole MS, and lastly **(3)** Fourier transform ion-cyclotron MS. A higher resolving power corresponds to a better ability of the instrument to detect small differences in mass-to-charge (m/z) ratios. Having a higher resolving power is generally very useful when trying to identify which metabolites are present in the sample. Modern high-precision MS instruments have very accurate resolving power, with accuracy up to several parts-per-million. The difference between the observed mass-to-charge value to the exact-mass-to-charge value of a compound is the mass accuracy of a mass spectrometry instrument, measured in parts-per-million, i.e. mass accuracy $= 1e6 * \frac{(\text{observed m/z} - \text{exact m/z})}{\text{exact m/z}}$.

The main advantage of NMR spectroscopy over MS is that its spectra is very high reproducibility since the same compound structure always produces peaks at the same locations in the spectra. Absolute quantification of the abundance of the compounds is possible in NMR as the signal intensity in NMR spectra is directly proportional to the concentration of protons in the nucleus of the compounds. In MS, often only the relative abundance (with respect to some reference compounds of known concentration) can be obtained. However, while the resulting spectra from NMR provides information on the structure of the metabolite, certain regions in the spectra can also be crowded with many overlapping metabolite signals [9], potentially hindering identification. NMR also has a lower sensitivity than mass spectrometry, which limits the number of metabolites that can be detected from NMR spectra. For more detailed comparisons of NMR vs. MS, the reader is directed to [9]. As it stands, the two approaches are often seen as complementary rather than competitive.

In direct injection mass spectrometry, the sample is introduced into the MS at a constant flow. However the ionisation capacity of MS is limited, and in what is called the ion supression effect, compounds can compete for charges during ionisation — resulting in certain compounds not being ionised and detected in the mass spectra [10]. Separating compounds as they gradually elute from the chromatographic column at a different *retention time* (RT) into the MS is often preferred. Additionally, from chromatographic separation, the retention time of observed peak reflects the underlying biochemistry of the compounds and can serve

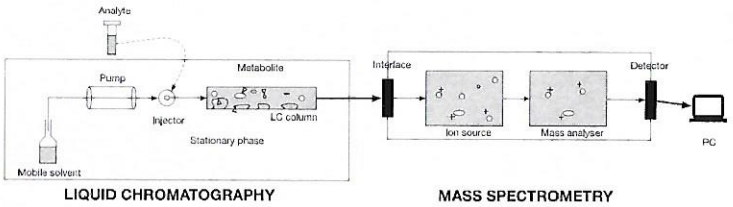

Figure 2.2: A typical LC-MS set-up. High performance liquid chromatography instruments are used to separate metabolites (by their chemical properties) in the sample before they are gradually introduced into the mass spectrometer.

as an additional information to deduce their identities [11]. Particularly in large-scale untargeted studies, MS is often coupled to a chromatographic separation technology such as liquid chromatography (LC), forming the combined set-up of LC-MS (Figure 2.2).

As illustrated in Figure 2.2, during liquid chromatography, the solvent containing the analytes (metabolites) is introduced and pumped into the stationary phase that is part of the chromatographic column. Metabolites elutes at different time through their interactions with the capillary in the column, based on their biochemical properties (e.g. their hydrophobicity, polarity, molecular shapes etc). In the LC-MS set-up, metabolites that elute from liquid chromatography are then vaporised and ionised inside the mass spectrometer. Ionisation in an LC-MS setup is usually performed via electrospray ionisation (ESI). In ESI, the sample analyte is dissolved into a solvent and sprayed through an electrospray (a highly charged needle) creating charged droplets. As the charged droplets travel through the vacuum of the MS, they evaporate, creating charged electric fields on the surfaces. In the strong electric field of the MS, ions on the surface of the droplets have enough energy to separate, generating charged molecular ions and their corresponding fragment ions. The generated ions are separated by the mass analyser inside the MS instrument according to their m/z (mass-to-charge) ratios and the detected signal abundance for a particular m/z value. As ESI requires a continuous supply of dissolved analytes, it can be directly coupled to LC, so often it is the preferred method of ionisation in LC-MS.

2.3 LC-MS Analysis in Metabolomics

The raw data produced from an LC-MS set-up is a collection of mass spectra from each scan over a range of elution time. Each MS measurement of compounds that elute at the same or similar retention time is called a scan. A mass spectrum in each scan is the two dimensional representation of m/z values of charged ions to signal intensities (Figure 2.3C).

The sum the signal intensities across all mass spectra, called the total ion chromatogram or TIC (Figure 2.3D) shows how compounds elute over time over all m/z values. The TIC plot can be too crowded, so given a specific m/z range to inspect, the extracted ion chromatogram (EIC) plot shows the total signal in that m/z range vs. RT (Figure 2.3E). The m/z range for inspection in the EIC is usually selected based on the prior knowledge of what signal a compound is supposed to produce in the spectra.

As shown in Figure 2.3B, the raw LC-MS data can also be seen as a 3D image containing peaks that can be characterised by a set of vector of m/z, retention time and intensity. This raw LC-MS data is noisy, so pre-processing has to take place before analysis can be performed and biological conclusion drawn. Generally, the main steps of LC-MS data pre-processing takes the form of a sequential pipeline shown in Figure 2.4. Note that Figure 2.4 illustrates an exemplar pipeline. In practice, many variations of this exemplar pipeline exists. For instance, the gap filling and the peak grouping steps can be omitted, the noise filtering step can be performed before peak alignment, no visualisation is produced from the output of identification, etc. The following sections explain in details the key steps of the LC-MS data processing pipeline in Figure 2.4.

2.3.1 Raw Data Importing & Peak Detection

The LC-MS data pre-processing pipeline starts with the raw data importing of vendor-proprietary format into an open XML-based format, such as mzXML [12] or mzML format [13]. Peak detection is applied to the imported LC-MS data to produce peaks. Each peak feature is characterised by its m/z, RT and intensity values. The CentWave algorithm [14] from XCMS is one of the more widely used peak detection method in metabolomics. It is particularly suitable for modern metabolomics data that are generated from instruments having a high mass accuracy. CentWave extracts regions of interest from the data. Chromatographic analysis of the EIC from each region of interest is performed using continuous wavelet transform is used to detect candidate chromatographic peaks. For each candidate peak, once its chromatographic peak boundaries have been identified, the centroid m/z value of a peak feature is defined as the weighted mean of the m/z values within the boundaries. Similarly, the intensity of a peak feature is defined as the maximal intensity value in the chromatographic peak boundaries. The signal-to-noise ratio of each candidate peak is calculated and if it is lower than the thresholded defined by the user, the candidate peak is rejected. As an alternative peak detection method, the MZmine 2 [15] software suite is also widely used.

A survey of the many different approaches for peak detections can be found in [16, 17, 8], however it is important to note that most peak detection methods are sensitive to the choice of parameters [14], with a method potentially producing different results when its parameters are varied. For instance, CentWave requires as user-defined parameters the mass deviation

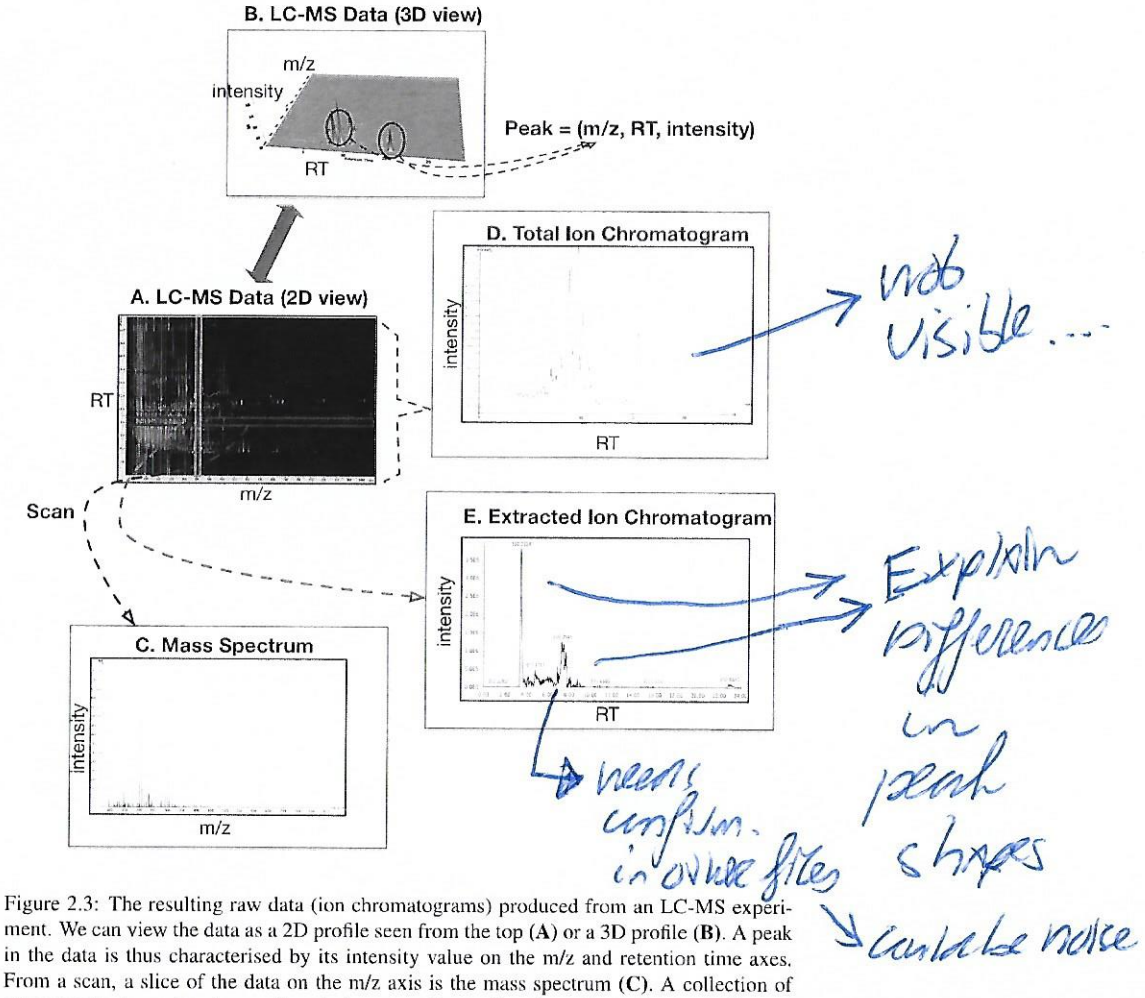

Figure 2.3: The resulting raw data (ion chromatograms) produced from an LC-MS experiment. We can view the data as a 2D profile seen from the top (**A**) or a 3D profile (**B**). A peak in the data is thus characterised by its intensity value on the m/z and retention time axes. From a scan, a slice of the data on the m/z axis is the mass spectrum (**C**). A collection of mass spectra is produced over the whole range of retention time. Summing over all scans produce the total ion chromatogram (TIC) (**D**), while plotting the intensity values vs. RT for a particular m/z range produces the extracted ion chromatogram (EIC) (**E**).

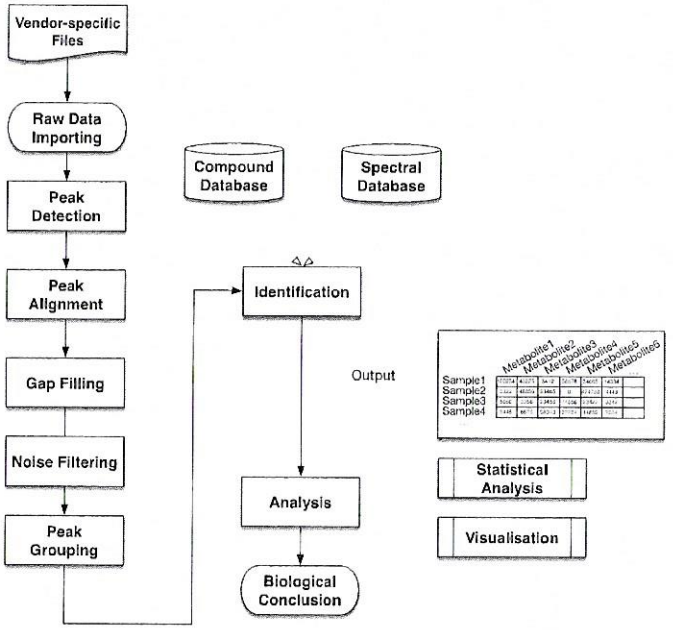

Figure 2.4: An exemplar pre-processing pipeline of LC-MS metabolomics data.

in parts-per-million (which is usually set based on the mass accuracy of instrument), the minimum width of the chromatographic peak and a signal-to-noise threshold. Setting a width that is too narrow or a signal-to-noise threshold that is too high can potentially lead to peaks that should be detected instead marked as missing.

2.3.2 Peak Alignment

Following peak detection, peak alignment is performed to match peaks that are the same across samples. An alignment method takes as input multiple lists of peaks — one from each LC-MS run — and produces as output a list of *aligned peaksets*. Each aligned peakset is a set of peaks coming from different runs that are considered to be *correspondent* and have to be matched. Alignment is necessary because experiments in biology usually involve the comparison of multiple samples. Samples can be produced as either biological or technical replicates. Biological replicates are obtained from the same organism studied under varying conditions and exposed to different factors (e.g. treatment or no treatment). Biological replicates are necessary to determine entities that are differentially expressed across samples. In contrast, technical replicates are obtained from the same sample analysed multiple times. Technical replicates are necessary to account for the variability and measurement errors throughout the experiment. In this manner, each replicate, whether biological or technical, is measured through the LC-MS instrument. This produces an LC-MS run for each replicate.

An initial approach towards alignment of multiple LC-MS runs would be to spike a known amount of internal standards into each sample before running them through the LC-MS instruments. Standards are compounds of known concentration that produce peaks at well-defined m/z and RT values. The peaks generated from these standards can be used as 'landmark' peaks to linearly shift the retention time in each sample, usually against a reference sample. Alternatively, stable-isotope labelling experiments exploit the fact that atoms have isotopes, which when measured in mass spectrometry, produce a distinct pattern of peaks that follow the binomial distribution. This information can be used to aid peak alignment and identification. In a labelling experiment, two samples are prepared: one from cells that grow in a normal medium and another from cells that grow in isotopic reagents. The two samples are combined and measured as a single LC-MS run. A metabolite from the normal medium and its corresponding isotopic counterpart have the same chemical formula and structure and hence will appear at close retention time, however the distinctive pattern of peaks produced from the isotopic metabolite makes it possible to trace the peaks back to the metabolite that produce them [18]. This makes alignment easier. However, labelled experiments consume expensive reagents, are more difficult to prepare and harder to compare across laboratories and to various mass spectral databases online for identification. Consequently, it is common

for large-scale untargeted LC-MS experiments, where the identities of the metabolites of interest are not known in advance, to be performed label-free without relying on such labelling information. This is called label-free experiments. To be comparable, the results from these label-free experiments need to be aligned, using peak alignment methods.

Broadly speaking, the main challenge in the peak alignment stage of label-free experiments is the poor reproducibility of retention time, with potentially large non-linear shifts and distortions across LC-MS runs produced from different analytical platforms or even the same platform over time [19]. Consequently, most alignment methods correct for those shifts and distortions by finding a mapping function f that maps peaks from one run to another. Depending on how they find f, alignment methods can be divided into two broad categories: **(1)** warping-based methods and **(2)** direct-matching methods.

Warping-based Alignment Methods

Warping-based methods seek to model the RT drifts between runs. In the past, many warping-based methods operate by aligning the whole ion chromatograms (profile data) directly before peak detection. Since this alignment step is performed before peak detection, warping-based methods that operate on profile data do not depend on the correctness of detected peaks. In this manner, the profile data being aligned is reduced to a simpler form by using the total ion chromatograms (TIC) as a representation of the entire data — frequently ignoring the rich information present in the m/z dimension of LC-MS data. As a consequence, warping-based methods that rely on profile information alone might not perform well for the alignments of the typical LC-MS data produced from complex mixtures – frequently having a lot of peaks of different m/z values co-eluting at similar retention times.

Many warping-based methods that operate on profile data are based on dynamic programming. In dynamic programming, all possible local solutions are evaluated but computed only for each sub-problem. In theory, this allows for an optimal global solution to be obtained efficiently. In practice, exact dynamic programming solutions are often intractable when a large number of runs need to be aligned at once due to their high time complexity when aligning multiple profile data simultaneously. As such, many of these methods aligns runs in a hierarchical pairwise manner. Some examples of well-known warping-based methods that operate on profile data are highlighted below:

1. **Dynamic Time Warping (DTW)** [20] performs a pairwise alignment of runs using the RT information only. The TICs being aligned are first discretised along the RT axes. Finding the alignment path is accomplished by setting up an alignment matrix and obtaining the best warping path that minimises the global distance in the alignment matrix. Three weight factors that computes the penalty for matches, expansion and

compression are defined. The optimal warping path is obtained by applying dynamic programming principle and tabulating intermediate results in the alignment matrix (in a manner similar to global sequence alignment for DNA sequences). The best warping path can then be read by backtracking from the final entry of the alignment matrix to the start.

2. **Correlation Optimised Wrapping (COW)** [21] operates in a manner similar to DTW by using the discretised TICs. COW divides the RT axes of replicates into segments. Each segment boundary can change within some user-specified slack parameter. COW then produces an alignment by finding the path across segments that has the highest sum of correlations. An alignment matrix is set up, and different segment boundaries can be shifted to maximise the global correlations between the two replicates being aligned using dynamic programming. In [22], COW is combined with a component detection algorithm (CODA [23]) that removes noisy signal and background noise from the mass chromatograms, aligning only regions containing high-quality information.

3. **Parametric Time Warping (PTW)** [24] produces pairwise alignment by using a second degree polynomial for mapping time between chromatograms. Coefficients of the polynomial are optimised by minimising the sum of squared residuals between the reference and aligned chromatograms. PTW performs much faster than COW. However, the quadratic polynomial model proposed in PTW, while simpler to describe, might not be sufficient to capture the complexity in non-linear retention time drifts across LC-MS data [19]. Semi-parametric Time Warping (STW) extends upon PTW and uses a series of B-splines as the mapping function. Optimising the warping coefficients in STW is done iteratively.

4. **Continuous Profile Mode (CPM)** [25] aligns multiple LC-MS data in a time series using a hidden Markov model-based approach. Each observed chromatogram profile is considered to be a time series of noisy signals sampled from a canonical latent profile. Parameters of the model are trained using the Expectation-Maximisation algorithm. The actual alignment of observed profiles to the latent profile is done using Viterbi algorithm. Compared to previous pair-wise methods such as DTW, CPM alignment is more robust since it aligns multiple LC-MS data simultaneously.

Since untargeted metabolomic experiments often produce a large number of runs, all of which need to be aligned as correctly as possible, most of the recent advances in warping-based methods are based on aligning peaks — a reduced representation of the raw LC-MS data obtained as the outcome of the peak detection step. Operating on peaks makes it easier to incorporate mass, intensities and other structural information that can potentially help

improve the alignment result. By extracting a smaller set of features from complex LC-MS raw data, often it is easier and faster to align many runs at once. To deal with the non-linear nature of retention time shifts in LC-MS data, a approach is to attempt to fit a regression curve on the peaks — usually using all the features observed across run or by selecting a certain subsets of all peaks. Some examples of well-known warping-based methods that operate on peaks are highlighted below:

1. **XCMS** [26] XCMS is one of the oldest tool used in metabolomics for processing mass spectrometry data and metabolite profiling. Alignment is XCMS is performed in two stages: peak matching and retention time correction. During the peak matching stage, the m/z axis is divided into discrete fixed-width overlapping bins. The alignment algorithm constructs a Gaussian kernel density estimation of the peaks inside each bin. This results in groups of peaks ('meta-peaks') that are close in their masses. Groups that do not contain enough peaks across samples are discarded. Next, during the retention time correction stage, well-behaved groups are selected as landmark peaks. The median retention time of each group is calculated, and the deviation from the median for each peak is used to train a local regression model. The resulting regression is used to correct for peak deviations.

2. **OpenMS** [27] OpenMS alignment works by first selecting a replicate that has the highest number of features. This replicate is used as the reference replicate, against which all other replicates are aligned against (in a star-like manner). The actual alignment process is divided into following two phases: superposition and consensus. During the superposition phase, the alignment algorithm tries to find the parameter for an affine transformation that maximises the number of features mapped from the reference replicate to the other replicates. An object recognition algorithm, called pose clustering, is used for this purpose. Additional information – such as m/z, RT and intensity dimension – is considered during the clustering process. The subsequent consensus phase then produces the actual alignment between matching features across replicates, using nearest-neighbour criteria.

3. **MZmine's RANSAC Aligner** [15] The RANSAC aligner is an alignment method developed part of the MZmine 2 software suite, used for the processing of metabolomics data. Random Sample Consensus (RANSAC) works by constructing a local regression model that maps retention time from one replicate to another. Once retention time correction has been performed, the actual matching of peaks across runs are performed greedily (using the older Join Aligner in MZmine 2). RANSAC Aligner is an iterative, non-deterministic algorithm, so there can be variations in the final alignment results. This non-determinism comes from the random sampling in the construction of the candidate model using the RANSAC algorithm[28].

Direct-matching Alignment Methods

Direct matching methods, which skip the warping step and seek to establish the correspondence of peaks across runs directly, can be preferred due to their simplicity, while still offering good performance [29]. Most direct matching methods consist of two stages: computing feature similarity and using this similarity to match peaks across runs. A wide range of feature similarity measures have been proposed to compare the m/z and RT values of two peaks, including normalised weighted absolute difference [15], cosine similarity [30], Euclidean distance [31], and Mahalanobis distance [32]. Once similarity has been computed, feature matching can be established through either a greedy or combinatorial matching method. Direct matching approaches therefore require that the peak detection step has already been completed, and the correctness of aligned peaksets depend on the output of the peak detection step. In fact, *all* steps that operate on peaks are similarly dependent on the correctness of the preceding peak detection step. In the presence of chemical and technical noises in the raw LC-MS data, relying on detected peak might serve to provide informative features rather than operating on the entire profile data [10].

Many approaches have been proposed for direct matching of peaks. Greedy direct-matching methods work by making a locally optimal choice at each step, in the hope that this will lead to an acceptable matching solution in the end. RTAlign in MSFACTs [33] merges all runs and greedily groups features into aligned peaksets within a user-defined RT tolerance. Join Aligner [15] in MZmine 2 merges successive runs to a master peaklist by matching features greedily according to their similarity scores within user-defined m/z and RT windows. Similarly, MassUntangler [31] performs nearest-distance matching of features, followed by various intermediate filtering and conflict-resolutions steps. Recent advances in direct matching methods have also posed the matching task as a combinatorial optimisation problem. Simultaneous Multiple Alignment (SIMA) [32] uses the Gale-Shapley algorithm to find a stable matching in the bipartite graph produced by joining peaks (nodes) from one run with peaks from another run that are within certain m/z and RT tolerances. [34] explores the application of the classical Hungarian algorithm to find the maximum weighted bipartite matching. BIPACE [30] establishes correspondence by finding the maximal cliques in the graph. SMFM [35] uses dynamic programming to compute a maximum bipartite matching under a relaxed bijective mapping assumption for time mapping.

As the output of direct-matching methods is the list of aligned peaksets itself, this class of methods can be used as an independent alignment method or as a second-stage process that follows a warping-based method. Once RT drift have been corrected in warping-based methods, it is often easier to establish the actual correspondence of peaks. Seen differently, if a good correspondence between peaks can be established, finding a warping function that maps the retention time from one run to another also becomes easier. In this manner, both

approaches to alignment — whether warping-based or direct-matching — complements each other. It is worth noting, however, that the final goal of alignment is not correcting retention time but establishing the matching of correspondent peaks across runs. In this manner, direct-matching methods directly addresses the core of the alignment problem.

Direct-matching methods can also be categorised depending on whether they require a user-defined reference run to be specified. When such reference is necessary, the full alignment of multiple runs is constructed through successive merging of pairwise runs towards the reference run (e.g. MZmine2's Join aligner in [15]. Alternatively, methods that do not require a reference run can either operate in a hierarchical fashion – where the final multiple alignment results are constructed in a greedy manner by merging of successive pairwise results following a guide tree (e.g. SIMA [32]) – or by pooling features across runs and grouping similar peaks in the combined input simultaneously (e.g. the *group()* function of XCMS in [26]).

2.3.3 Gap Filling & Noise Filtering

From the alignment results, certain peaks might be missing from an aligned peakset. The gap filling step recovers this missing signal from the raw data. A peak may be missing as it was not detected in the peak detection step (due to having a low intensity or a poor chromatographic peak shapes). Once gap filling has been performed, noise filtering is performed. Filtering can be performed based many criteria, e.g. using a threshold on the intensity to remove low-intensity peaks that are likely to be noise.

2.3.4 Peak Grouping

In the peak grouping stage, the sets of peaks that are chemically-related to each-other are grouped. During ionisation in mass spectrometry, a single metabolite alone can produce multiple peaks (e.g. isotopic peaks, adduct peaks and fragment peaks) that are all chemically-related to each other. Following [2], we call this set of peaks the *ionisation product* (IP) peaks of the compound. In particular, the presence of naturally occurring isotopes (e.g. ^{13}C) means a single compound can produce a pattern of peaks with m/z and intensity that follow the isotopic distributions of the atomic elements of the compound [36]. Similarly, the formation of adducts (the addition of a molecule ion to another) means that within a mass spectrum, certain adduct peaks, generated from the same compound, can be explained by the set of adduct transformations [37]. As they co-elute from the column, these IP peaks are expected to have similar chromatographic peak shapes, and therefore they share similar RT values. In [38], an analogous concept of 'derivative peaks' is defined to be the set of peaks that elute at the same retention time, show a strong correlation between their chromatographic peak

shapes, have mass differences that can be explained by known chemical relationships and have intensity values that can be correlated across different runs.

A common use of the peak grouping stage is as a data filtering procedure prior to identification. This is because the naive assumption that each observed peak corresponds to a single compound will produce too many false positives in the identification step that follows in the pipeline — particularly when identification is made based on querying by mass alone to large public compound databases, such as KEGG or PubChem. Following the idea of derivative peaks in [38], the mzMatch software suite [39] detects IP peaks based on a greedy clustering scheme. Peaks having the largest intensity are clustered to others sharing chromatographic peak shape correlations above a certain user-defined threshold. This is repeated until all peaks are processed. In [40], the same idea is exploited in the form of a mixture model to cluster peaks based on their chromatographic peak shape correlations. CAMERA [41] performs the annotations of ionisation product species on groups of peaks, based on constructing a similarity graph and detecting highly-connected subgraphs in the graph. IP peaks are annotated on the subgraphs based on how their masses can be explained by a set of user-defined chemical rules. In [2], IP peaks are grouped along the RT dimension using a sliding window and along the m/z dimension using k-means clustering. The grouping induced by these methods are used as a form of data filtering by discarding peaks that are deemed irrelevant. Groups can also be used to aid identification, as explained in the following section, but often this information is not used.

2.3.5 Peak Identification

In a general sense, peak identification refers to the process of annotating a label that tells us which peaks are associated to which metabolite. As shown in Figure 2.4, the output from the identification step is a matrix where each row in the matrix corresponds to a biological or technical sample, each column a metabolite, and entries in the matrix are the intensity of the detected metabolite in each sample. Untargeted identification is challenging in untargeted metabolomic studies due to the vast number of metabolites present in sample and the diversity in elements that comprise a metabolite. Unlike the genome that has four nucleotide bases as its sole alphabets, or proteins with twenty one amino acids as their building blocks, metabolites are harder to characterise structurally, the basic building blocks of a metabolite are atoms (commonly CHNOPS) that can be arranged in a variety of configurations in a single molecule alone (Figure 2.1).

The term 'identification' can be overloaded with many different meanings, e.g. is it the definite annotation of a compound label to a peak or a putatively assigned label? The Chemical Working Group of the Metabolomics Standards Initiative proposed four levels of identifications for the reporting of metabolite identifications [42] that have been accepted to a varying

degree by the community. In this scheme, the most confident Level 1 identification is obtained through the comparison of the observed peaks against those generated from a set of chemical standards (a solution containing compounds of known concentration). A putative Level 2 identification is obtained from comparison against publicly available spectral libraries. Level 3 identification seeks to confirm the chemical class of the compounds, while a Level 4 of no identification is assigned unknown compounds. In its most basic form, both Level 1 and 2 identifications are performed by taking the neutral masses of observed peaks and matching them against the list of masses from a database of compounds, which may range in size from just a few hundreds of metabolites to as large as tens of thousands of compounds or more. The database for matching may be constructed for the standard compounds or the public database. Having a high mass accuracy is therefore crucial for identification as it reduces the size of possible alternatives that can be matched.

In untargeted metabolomics, the lack of knowledge in the composition of metabolites in the sample means that, apart from the small number of metabolites confidently identified as the authentic standard compounds, the putative annotations of metabolite identities that are assigned to a peak might be the result of incorrect matching against the compound database. This leads to false positive identifications and consequently incorrect biological conclusions. Creating a larger authentic standard to facilitate more confident identifications is constrained by time and cost and can never be comprehensive enough to include all metabolites of interest in an untargeted study. Another challenge of identification is even at the very high mass accuracy of 1 ppm, the number of possible formulae matched by accurate mass is still too large to allow for definite metabolite identifications [43]. Identification is particularly difficult for metabolites present in low abundance in the samples. Relying on mass alone for untargeted identification is also problematic as different metabolites may produce peaks having the same measured m/z values, and as in the case of isomers, the same precursor mass can therefore be matched to multiple possible formulae. Retention time drift, a main challenge in alignment, means RT values vary across different chromatography platforms and laboratories and cannot be easily used as a characteristic identifying information in a public compound database. Incomplete knowledge on the metabolites expected to be present in the sample, coupled with the complexity of the sample being analysed itself, means identification is challenging [44], with more metabolites being putatively identified (Level 2) than very confidently identified (Level 1), but the majority of metabolites can only be identified based on their class (Level 3) or not at all (Level 4). Even for the putatively identified metabolites, their manual verification is a laborous and time-consuming process, often serving as the primary bottleneck in large-scale untargeted metabolomic studies [44, 45]. In particular, false positives from identification is a major concern in the data pre-processing step.

To reduce false positives, additional information can be incorporated into the identification process. In particular, identification can also be performed on the basis of a group of ionisa-

tion product peaks, rather than on individual peaks alone, although this is often not exploited in many tools. As discussed before, tools such as CAMERA [41] can produce a group of IP peaks. From this group, the precursor mass that corresponds to the molecular ion mass of the compound can be deduced. This can be used for matching against a compound database, allowing for a set of peaks to be identified rather than individual peaks alone. Other sources of information that can help identification include using the predicted RT of a compound [46, 11, 47], but matching the predicted RT values against the observed RT data that contain drifts might be challenging too. Probabilistic methods that use prior information of a known set of formulae to annotate peaks by explainable transformations have also been proposed [48, 49], but often have difficulties scaling up to large-scale experiments to be of practical use.

Fragmentation through tandem MS or MS^n instruments is another way to provide further information to aid identification. As suggested by its name, tandem MS requires two MS analysers operating in tandem. Ions resulting from the initial fragmentation of metabolites in the first MS analyser are selected for further fragmentation in the second MS analyser. The ions selected for the first MS analyser stage are called the precursor ions. In data-dependent acquisition (DDA), precursor ions within some small m/z windows are selected based on some predetermined rules (such as fragmenting the top few most intense precursor peaks in each scan). As a result, typically a small percentage, e.g. less than a fifth of all precursor peaks in the full-scan mode data are selected for MS-MS fragmentation. Peaks that are generated from the fragmentation of the precursor ions in the second MS stage are called product ions. Fragmentation spectra of product ions are often used as the unique 'fingerprint' identifiers of the structural composition of the precursor ions. An alternative to DDA is the data-independent acquisition (DIA), where no selection of precursor ions needs to be specified as all peaks within a defined m/z range are fragmented. DIA results in a more complex fragmentation spectra due to multiple metabolites being fragmented together in the same m/z window, and require sophisticated analysis strategy to deconvolve the signals from the noise.

A fragmentation spectrum of interest can be identified through matching against **(1)** a database of public reference spectra or **(2)** a database of theoretical spectra generated in an in-silico manner [50]. Examples of public databases are KEGG [51], Massbank [52] and ChemSpider [53]. Frequently, a combination of matching against a public database and in-silico theoretical spectra is used to ensure the largest coverage of compounds during matching. The actual matching process is often established in a greedy manner, heuristically through agreement against a set of well-validated fragmentation rules or combinatorially by minimising a cost/distance function. In the combinatorial case, heuristic rules are still applied to reduce the exponentially-growing search space to allow matching to run in acceptable time. However, fragmentation cannot be used in all cases as not all metabolomics experi-

ments include fragmentation as part of their data acquisition process — due to cost or other resource constraints. Publicly available databases have a limited coverage in the number of submitted spectra. Often spectra in public databases are contributed from a wide variety of instruments, further limiting potential matches as matching is often possible only for spectra generated on similar platforms. Large variance in the mass accuracy and characteristics of submitted spectral library entries further limit potential matches as a query match can only be made against spectra generated from similar platforms and mass accuracies. Unwanted spectral peaks (due to e.g. the presence of contaminants and noise in the sample) present in the database may also lead to incorrect spectral matching. Fragmentation and its challenges are further discussed in Chapter 7.

2.3.6 Analysis

The last step in preprocessing of LC-MS data is the normalisation and visualisation of data. Normalisation is essential for removing any possible variation and systematic bias to allow for comparisons of differential levels of expressions of metabolites across samples. Statistical analysis is performed with visualizations in order to draw useful inferences from data – a step that is crucial in confirming or rejecting biological hypotheses. At this stage, the data is normalised to correct for systematic variations before statistical analysis. Spiked-in compounds that do not occur naturally are used for this purpose. Since the spiked-in compounds are expected to have equal concentration in all samples, they can be used to normalise peak areas in samples. Statistical analysis, such as t-test, ANOVA and principal component analysis, can then be performed on the normalised peaks across samples. The goal of statistical analysis is to answer biological hypothesis posed by life-science researchers. During the analysis, it is common to place the result obtained from metabolomic studies on the larger biological context by mapping them onto some biological pathways ([54, 55]) or in relation to other -omics studies ([56, 57]).

While targeted metabolomics focuses on a handful of specific metabolites, untargeted studies (such as in [58] and [46]) attempt to perform a global analysis of metabolites in the samples under study. Understanding the metabolome in an untargeted study is a challenging task due to the complex interactions of metabolites in the metabolome. Identification of specific metabolites are frequently not the final goal in untargeted metabolomics, rather it is the discovery of metabolites or groups of metabolites that are differentially expressed or correlated to the expression of specific physical traits being studied. Of particular interest is the detection of metabolites that act as disease biomarkers. The presence or absence of such metabolites can provide an indication to the corresponding presence or absence of disease in the organism [59]. Differences caused by genetic variations are also highly visible as changes in the metabolite composition of an organism. These could be quantified through

differential analysis that compares the expression levels (abudance) of metabolites across samples. The resulting differential analysis provides biologists with a better understanding of the metabolic pathways in the cell and how they respond to perturbations. Differential analysis also underpins many practical applications of systems biology, such as nutritional research [60], drug discovery [61] and even in an integrative approach that combines genomics and metabolomics to obtain a more comprehensive picture of living organisms [57]. Visualisation of the identified metabolites can also be performed by mapping metabolites to well-known pathways from databases such as KEGG [62] or MetaCyc [63]. Identified metabolites at this stage can also be integrated with the reconstructed metabolic information from other -omics [64] to allow for a rapid generation of biological hypotheses.

2.3.7 Mass Spectrometry Analysis in Proteomics

LC-MS analysis in proteomics proceeds largely in the same manner as to the data pre-processing pipeline in Figure 2.4. However, the key difference between proteomics and metabolomics lies in sample preparation. In the mass spectrometry analysis of proteins, the samples to be analysed come either in the form of tissues or as body fluids, such as urine, plasma and serum, with each different type of sample demand an appropriate sample handling protocol. Next, cells extracted from the sample are broken down, allowing proteins to be isolated from other constituent parts of the cell, for instance the DNA, lipids and other metabolites that are present. The purified proteins are then separated. Traditional 2-D gel electrophoresis method allows proteins to be separated according to their size (molecular mass) in one axis and according to their isoelectric points (the pH where the molecule carries no electrical charges) on another. Because 2D-GE approach is tedious and time-consuming, liquid chromatograph mass spectrometry has gotten more popular as the preferred separation technology as it enables the large-scale high-throughput separation of thousands of proteins in a single chromatographic run. Enzymes that can cut the peptide bonds, such as trypsin, are then used to digest proteins into shorter peptide fragments. Using certain enzymes, the cleavage of the peptide bonds happen at specific and predictable spots, allowing well-defined and easily identifiable peptide fragments to emerge. For instance by using trypsin as the digestion enzyme, the cleavage of the protein happens after each arginine or lysine amino acid is encountered, unless a proline amino acid comes next.

Identification of peptide sequences in proteomics largely proceeds in the same manner as metabolomics. Different set of tools and public databases are queried for matching. In particular, the problem of peptide identification from fragmentation data is referred to as peptide mass fingerprinting [65]. As proteins are cleaved into peptides that are unique, the resulting fragmentation spectra are also expected to be unique to a protein. The theoretical peptide spectra can then matched against a reference spectra library. In practice, the resulting fragmentation spectra are not entirely unique and multiple hits can be returned from the spectra library, particularly in the case of libraries that have a large number of records. The fact that the peptide sequence of a protein is known and digestion enzyme produces cuts at predictable spots means identification through a comparison to a *de novo* peptide sequences is possible in proteomics. Additionally, it is also more common in proteomics than metabolomics for an initial separation process, called pre-factionation, to be performed on the digested peptides using liquid chromatography. This divides the entire sample into multiple *fractions* of compounds that elute at different retention time, which can then be ran separately through the LC-MS instrument for mass fragmentation analysis in a manner similar to metabolomics analysis. Certain fractions can be selected for further analysis, leading to a simpler set of data to deal with.

2.4 Conclusion

Data processing has major impact on the outcome of quantitative label-free LC-MS analysis [66]. Even the choice of the software tools itself, with differing implementation details, affect the outcome. In particular, label-free experiments pose many challenges when analysing many LC-MS runs. Peaks from different runs can experience a potentially non-linear shift in retention time across chromatograms [67]. There is often a large amount of variations in the retention times across the replicates. Retention time variation could be due to instrument-specific factors (the condition of the chromatographic column itself. including flow rate variations, gradient slope and temperature [22]) or experiment-specific factors (e.g. instrument malfunctions or columns that need be replaced mid-experiment). Both factors are difficult to control, even in a careful experimental setting. Consequently, a single peak from one run can have several potential matching peaks in another run, while having no matches in another run. This is exacerbated by the uncertainties introduced due to parameter selections in the preceding steps before alignment in the pipeline. As a result, replicates produced by different LC-MS platforms or from different laboratories cannot be easily aligned to each other.

Since large-scale untargeted metabolomics study can generate a huge number of samples (see [58, 46]), having a reliable and accurate peak alignment step during data pre-processing is important. Peaks that are improperly aligned can lead to false positives, and especially for untargeted label-free metabolomic experiments, the presence of even relatively small errors in any steps preceding the identification stage (including alignment) can result in significant differences to the final analysis and biological conclusions. Errors or uncertainties inadvertently produced in any sub-step before identification would be carried forward forward in the pipeline. Improper pre-processing steps can also introduce variabilities that obscure important biological variations of metabolites themselves.

Software tools that deals with LC-MS data in proteomics and metabolomics usually operate in a modular and serial manner, where successive transformations occur to the raw LC-MS data as it goes through the data pre-processing pipeline. However, it is important to note that despite the apparently serial pre-processing manner shown in Figure 2.4, the actual pipeline workflow employed by the user is often iterative. For example, it is often the case that certain low intensity metabolites are expected to be present in the identification result, but are found to be missing. This requires the user to revisit each step of the pipeline, experiment with the numerous user-defined parameters and threshold values used for the peak detection, alignment, gap filling, noise filtering and identification step to troubleshoot this issue. Each step of the exemplar pipeline in Figure 2.4 is therefore dependent on the steps that come before it. However, at the moment, each step in the pipeline exists independently and information from one step is not used to improve the performance of the subsequent steps in the pipeline.

This chapter has provided the necessary background knowledge to understand the basic principles of mass-spectrometry-based analysis as applied to large-scale untargeted biological studies, but it is far from complete. A particular emphasis is given to the application of mass spectrometry techniques in the field of metabolomics. For further readings on mass spectrometry as an analytical platform, the reader is directed to more comprehensive textbooks such as [68] and [69]. For literature surveys on the different steps that comprise an LC-MS data processing pipeline, the reader is directed to [17, 10, 70, 8] for metabolomics and [65, 71, 10] for proteomics.

Chapter 3

Probabilistic Modelling

3.1 Introduction

As described in Chapter 2, the raw data produced from liquid chromatography mass spectrometry (LC-MS) measurements has to be processed through a data pre-processing pipeline before further analysis. From the peak detection step, we obtain points on the ion chromatograms having mass-to-charge (m/z), retention time (RT) and intensity values. We call each point a *peak*. The nature of LC-MS measurements means that a compound being analysed generates multiple peaks. At the heart of this thesis is the grouping of these peaks that are structurally or chemically related, and using the grouping to improve other steps (such as the alignment and identification steps) in the pipeline. The problem of finding these groups of related peaks can be approached as an unsupervised learning problem. In the unsupervised learning approach, broadly speaking our task is to separate peaks into *clusters*, where members of the cluster are related through sharing some commonalities, e.g. from being the ionisation products of the same compound or from sharing chemical substructures.

Numerous methods exist to perform data clustering in an unsupervised manner [72, 73]. In probabilistic modelling, one way to do this is to try and explain the generative process that produces the observed data. This results in a generative model. Peaks generated from the same underlying cause in the model can then be assigned to the same cluster. Modelling the data in this manner has some advantages in comparison to other distance-based clustering methods, such as e.g. hierarchical clustering that has also been applied to peak data [74, 75]. A generative model provides more than just clustering. It is often easier to extract from a generative model a hint as to *why* the observed data points are clustered, and this insight can be very useful in certain applications. Additionally, through specifying the appropriate likelihood functions, generative modelling also provides a flexible way of specifying how data points should be clustered, while prior assumptions can be incorporated into the model in a principled manner.

Generative modelling has been applied to LC-MS data. In [40], mixture model clustering is used to cluster LC-MS peaks in the same run by their chromatographic profiles. Mixture models are the building blocks of more complex generative models. In mixture models, it is assumed that the observed data can be explained by the presence of some latent variables. These variables are 'latent' as they are not directly observed, rather their presence is inferred from the observed data. The assumption made in [40] is that ionisation product peaks that are related share similar chromatographic profiles. Given N peaks in the data, the method computes the pairwise Pearson correlation values for the chromatographic profiles of all peaks, resulting in an N-by-N matrix of Pearson correlation values. The likelihood of an entry in this matrix is described by a mixture of two components: an exponential-type distribution to describe the correlation values of peaks in the same cluster and a Gaussian distribution to describe the correlation values of peaks in different clusters.

Along a similar line, mixture model clustering is also used in MetAssign [49] to perform the probabilistic annotations of ionisation product types and formulae to peak data. It is assumed in MetAssign that a prior knowledge of the form of known formulae is provided. Theoretical peaks are then generated using the provided formulae. The likelihood of an observed peak is computed based on how well the observed m/z, RT and intensity values fit the theoretical peaks. Other applications of probabilistic modelling on mass spectrometry data include modelling the assignment of formulae to peaks [48, 76], modelling the fragmentation events of tandem mass spectrometry data, where the separation is performed using liquid chromatography (CMF-ESI, [77]) or gas chromatography (CFM-EI, [78]). Machine learning techniques in general have also been applied to mass spectrometry data, e.g. for the predictions of retention time [46, 11, 47] and the characteristic fingerprints of compounds from fragmentation data [79, 80].

3.2 Mixture Model Clustering

As an example of the thinking process behind generative modelling, we see that during liquid chromatography, metabolites are separated by their chemical properties. From mass spectrometry, ionisation product peaks are produced from the same metabolites. These peaks will co-elute and have similar chromatographic profiles, including broadly similar RT values. A group of observed peaks having similar retention time (RT) values can be therefore be modelled as being generated by the same metabolite, and in this case, although the metabolite is not directly observed, its presence can be inferred based on the observed data. Peaks that are related to the same compound can therefore be clustered according to their RT values. Let our LC-MS run be represented as $\mathbf{y} = \{y_1, y_2, ... y_n\}$ where each y_n is the RT value of a peak.

A principled way to model a generative process is through Bayesian inference. Suppose θ is the parameter of interest to the generative process that produces the data $\mathbf{y}$. In Bayesian inference, we begin by specifying a prior distribution over the model parameter θ. Through the application of Bayes rule, this prior distribution is updated by the likelihood of seeing the observed data given our prior hypothesis on θ, resulting in a posterior distribution:

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}, \theta)}{p(\mathbf{y})} = \frac{p(\mathbf{y}|\theta)p(\theta)}{p(\mathbf{y})} = \frac{p(\mathbf{y}|\theta)p(\theta)}{\int_\theta p(\mathbf{y}|\theta)p(\theta)\ d\theta} \tag{3.1}$$

In eq. (3.1), $p(\mathbf{y}, \theta)$ is the joint distribution between the data $\mathbf{y}$ and the model parameter θ. This can be factorised into a product of $p(\mathbf{y}|\theta)$, which is the likelihood of observing the data $\mathbf{y}$ given the model parameter θ, and $p(\theta)$, which is the prior distribution on the model parameter θ. Normalising the joint distribution by the marginal likelihood or evidence $p(\mathbf{y})$ produces the posterior distribution $p(\theta|\mathbf{y})$, which is the probability of model parameter θ given the data. Inferring model parameters given the observed data is usually what we are interested in.

Using the posterior distribution, we can make a prediction on a new peak, y_{new} by averaging over all values of θ. This results in the posterior predictive distribution:

$$p(y_{new}|\mathbf{y}) = \int_\theta p(y_{new}|\theta)p(\theta|\mathbf{y})\ d\theta \tag{3.2}$$

In many cases, the integral in eqs. (3.1) and (3.2) cannot be solved analytically and have to be approximated through maximum likelihood or sampling-based approaches.

We now introduce mixture modelling for this example peak data. Probabilistic mixture model represents each cluster by a probability distribution, with a distribution being a component in the mixture model. Our resulting Gaussian mixture model for the peak data follows from [81]: it starts from having a finite number of components (denoted by K) and is later extended in Section 3.3 to an infinite mixture model, where the number of components is unbounded. The generative process for this finite mixture model can be written as the following. The conditional dependencies of random variables in the finite mixture model is also shown in Figure 3.1A.

$$\begin{aligned} \boldsymbol{\pi}|\boldsymbol{\alpha} &\sim Dir(\boldsymbol{\alpha}) \\ z_{nk} = 1|\pi_k &\sim \boldsymbol{\pi} \\ \mu_k|\mu_0 &\sim \mathcal{N}(\mu_k|\mu_0, \sigma_0^2) \\ y_n|z_{nk} = 1, \mu_k &\sim \mathcal{N}(y_n|\mu, \sigma^2) \end{aligned} \tag{3.3}$$

We now explain the model specification in eq. (3.3). First we assume that peaks that are related are generated by the same component in the mixture model. Let the variable $k = 1, ..., K$ index the mixture components. The choice of which probability distribution to

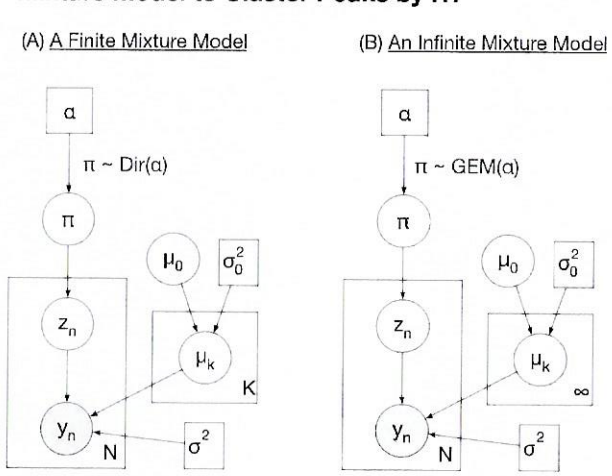

Figure 3.1: Graphical models of (1) a finite mixture model, which is extended into (2) an infinite mixture model, to cluster peaks by their retention time (RT) values. Circles denotes random variables, squares denote fixed parameters, while the shaded node denotes an observed peak's RT.

use as a component is usually determined by the type of observed data. Each observed data point y_n can be considered to a random variable drawn from the generating probability distribution. Assuming that each data point is generated by a univariate Gaussian distribution, we denote by $y_n|\mu, \sigma^2 \sim \mathcal{N}(\mu, \sigma^2)$ that y_n is distributed as a Gaussian distribution having the mean μ and the variance σ^2 (as an alternative parameterisation, precision, i.e. the inverse variance ($\frac{1}{\sigma^2}$) can also be used, with a higher precision meaning a narrower distribution). The probability density function for this univariate Gaussian distribution is given by:

$$\mathcal{N}(y_n|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_n-\mu)^2} \tag{3.4}$$

For a single peak, its mixture model likelihood is therefore given by:

$$p(y_n|\boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(y_n|\mu_k, \sigma^2) \tag{3.5}$$

where π_k is the mixture proportion (the positive weight for each component) and μ_k is mean for that component. $\boldsymbol{\pi} = \{\pi_1, ..., \pi_k\}$ denotes the vector over all mixture proportions that sums to one ($\sum_{k=1}^{K} \pi_k = 1$).

In this model, each mixture component is set to have an unknown mean μ_k but a known variance σ^2. The choice of setting an unknown μ_k but a fixed variance for σ^2 is motivated by the following reasonable modelling assumptions: **(1)** the retention time drift of observed peaks is broadly similar across the compounds being measured, and **(2)** this parameter can be set by the user based on his knowledge on the characteristic RT drifts of the LC instrument. Each cluster mean μ_k is assumed to be generated independently by a prior Gaussian distribution, parameterised by the mean μ_0 and the variance σ_0^2. Let $\boldsymbol{\mu} = \{\mu_1, ..., \mu_k\}$ be the vector over all component means. This results in:

$$p(\boldsymbol{\mu}|\mu_0, \sigma_0^2) = \prod_{k=1}^{K} \mathcal{N}(\mu_k|\mu_0, \sigma_0^2) \tag{3.6}$$

We also require another random variable z_{nk} to store the assignment of peak n to cluster k, i.e. $z_{nk} = 1$ if peak n is assigned to cluster k and 0 otherwise. Each peak is assumed to be generated independently by exactly one mixture component ($\sum_k z_{nk} = 1$). For a peak, its entire cluster assignments can be stored in a vector $\boldsymbol{z}_n$ of length K, where only k-th entry has a value of 1 (at $z_{nk} = 1$). $\boldsymbol{z}_n$ is assumed to be generated from a multinomial distribution having the parameter vector $\boldsymbol{\pi}$. This multinomial distribution has the probability mass function given by:

$$p(\boldsymbol{z}_n|\boldsymbol{\pi}) = C \prod_{k=1}^{K} \pi_k^{z_{nk}} \tag{3.7}$$

where C is the multinomial coefficient, given by $\frac{(\sum_k z_{nk})!}{\prod_{k=1}^{K} z_{nk}!}$. Since $\boldsymbol{z}_n$ has only one draw from the multinomial, C evaluates to 1 and can be dropped. Now, let $\boldsymbol{Z}$ be the set of all indicator vectors for all peaks. This results in the following likelihood for all the peak assignment vectors:

$$\begin{aligned} p(\boldsymbol{Z}|\boldsymbol{\pi}) &= \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \\ &= \prod_{k=1}^{K} \pi_k^{c_k} \end{aligned} \tag{3.8}$$

where $c_k = \sum_n z_{nk}$ is the count of peaks assigned to the k-th cluster. Collectively for all peaks, the joint likelihood of the observed data and the cluster assignments is:

$$p(\boldsymbol{y}, \boldsymbol{Z}|\boldsymbol{\mu}, \boldsymbol{\pi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} [\pi_k \ \mathcal{N}(y_n|\mu_k, \sigma^2)]^{z_{nk}} \tag{3.9}$$

In our generative model, a prior distribution is also placed on $\boldsymbol{\pi}$. Due to its conjugacy to the multinomial distribution, a Dirichlet distribution parameterised by the vector $\boldsymbol{\alpha}$ =

$[\alpha_1, \alpha_2, ..., \alpha_k]^T$ is a suitable prior. This results in:

$$p(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} \propto \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} \tag{3.10}$$

We can now state the complete joint likelihood of the model. Putting together the individual terms in eqs. (3.6)-3.10) and their respective independence assumptions, we obtain the joint probability distribution of the model parameters and data $p(\boldsymbol{y}, \boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\pi}|\boldsymbol{\alpha}, \mu_0, \sigma_0^2)$, which can be factorised into:

$$p(\boldsymbol{y}, \boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\pi}|\boldsymbol{\alpha}, \mu_0, \sigma_0^2) = p(\boldsymbol{y}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\pi}) p(\boldsymbol{Z}|\boldsymbol{\pi}) p(\boldsymbol{\pi}|\boldsymbol{\alpha}) p(\boldsymbol{\mu}|\mu_0, \sigma_0^2) \tag{3.11}$$

3.2.1 Gibbs Sampling for a Finite Mixture Model

Given the joint distribution in eq. (3.11), we are interested to infer the posterior distribution on the assignments $\boldsymbol{Z}$, the mixture proportions $\boldsymbol{\pi}$ and the cluster means $\boldsymbol{\mu}$. This is given by:

$$p(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}|\boldsymbol{y}, \boldsymbol{\alpha}, \mu_0, \sigma_0^2) = \frac{p(\boldsymbol{y}, \boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\pi}|\boldsymbol{\alpha}, \mu_0, \sigma_0^2)}{p(\boldsymbol{y}|\alpha, \mu_0, \sigma_0^2)} \tag{3.12}$$

Substituting eq. (3.11) into the numerator of eq. (3.12) results in the following posterior distribution over the parameters that we want to infer:

$$p(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}|\boldsymbol{y}, \alpha, \mu_0, s_0) \propto p(\boldsymbol{y}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\pi}) p(\boldsymbol{Z}|\boldsymbol{\pi}) p(\boldsymbol{\pi}|\boldsymbol{\alpha}) p(\boldsymbol{\mu}|\mu_0, s_0) \tag{3.13}$$

In many cases for the more interesting and complex models, the posterior distribution (such as the one in eq. 3.13) and also the posterior predictive distribution cannot be derived analytically. Various methods, such as the EM algorithm [82], can be used to perform posterior inference in a mixture model, but throughout this thesis, we will use Gibbs sampling, an instance of Markov chain Monte Carlo (MCMC) methods. Gibbs sampling approximates the target posterior distribution by sequentially updating each random variable conditioned on all other random variables in the model. This requires deriving the *conditional distribution* of each random variable that we want to infer. In some cases, obtaining these conditional distributions can be challenging, although the process can be simplified by the independence assumptions of our model (e.g. in assuming that the cluster means are independent) and through the use of the appropriate conjugate prior distributions. Here we describe the steps required to construct a Gibbs sampler for the mixture model defined in eq. (3.3).

As the initial step in our Gibbs sampler, we initialise the cluster means $\mu_1, \mu_2, .., \mu_k$ and the mixture proportion $\boldsymbol{\pi}$ by sampling from their respective prior distributions. Then we sequentially sample for new values of $\boldsymbol{Z}$, $\boldsymbol{\mu}$ and $\boldsymbol{\pi}$ from the conditional distributions listed below.

1. We can update $\boldsymbol{z}_n$, the membership vector for peak n, by updating each of its k-th individual entry, i.e. z_{nk}. Simplifying eq. (3.9) to consider just one n-th peak, we obtain the following after normalisation:

$$P(z_{nk} = 1|\boldsymbol{\pi}, y_n, \mu_k) = \frac{\pi_k \ \mathcal{N}(y_n|\mu_k, \sigma^2)}{\sum_{k=1}^{K} \pi_k \ \mathcal{N}(y_n|\mu_k, \sigma_2)} \tag{3.14}$$

2. As the next step, we also need to update each μ_k conditioned on the membership vectors $\boldsymbol{Z}$ and the hyperparameters μ_0 and σ_0. Consider one k-th cluster, and let $\boldsymbol{x}_k = \{x_1, x_2, ..., x_m\}$ be the set of peaks currently assigned to cluster k. The variable m indexes over the member peaks of cluster k, and there are M_k such peaks. Their joint likelihood is given by $p(\boldsymbol{x}_k|\mu_k)$. As defined in eq. (3.6), we assume that each μ_k is independent given its conjugate prior $\mathcal{N}(\mu_0, \sigma_0^2)$. The posterior distribution on $p(\mu_k|\boldsymbol{x}_k, \mu_0)$ is therefore:

$$\begin{aligned} p(\mu_k|\boldsymbol{x}_k, \mu_0) &\propto p(\boldsymbol{x}_k|\mu_k) \cdot p(\mu_k|\mu_0) \\ &= \prod_{m=1}^{M_k} \mathcal{N}(x_m|\mu_k, \sigma^2) \cdot \mathcal{N}(\mu_k|\mu_0, \sigma_0) \\ &= \prod_{m=1}^{M_k} \frac{1}{\sqrt{2\sigma^2\pi}} exp\left(\frac{-(x_n - \mu_k)^2}{2\sigma^2}\right) \cdot \frac{1}{\sqrt{2\sigma_0^2\pi}} exp\left(\frac{-(\mu_k - \mu_0)^2}{2\sigma_0^2}\right) \end{aligned} \tag{3.15}$$

Since $p(\mu_k|\boldsymbol{x}_k, \mu_0)$ is a product of Gaussians, the posterior is proportional to another Gaussian, parameterised by say $\mathcal{N}(\tilde{\mu}, \tilde{\sigma}^2)$. Equating this with eq. (3.15) results in:

$$exp\left(\frac{-(\mu_k - \tilde{\mu})^2}{2\tilde{\sigma}^2}\right) \propto exp\left(\frac{-\sum_{m=1}^{M}(x_m - \mu_k)^2}{2\sigma^2} + \frac{-(\mu_k - \mu_0)^2}{2\sigma_0^2}\right) \tag{3.16}$$

Simplifying eq. (3.16) and completing the squares, we obtain the following parameters for $\mathcal{N}(\mu_k|\tilde{\mu}, \tilde{\sigma}^2)$:

$$\tilde{\mu} = \tilde{\sigma}^2\left(\frac{\sum_{m=1}^{M} x_m}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right), \quad \tilde{\sigma}^2 = \frac{1}{\frac{M}{\sigma^2} + \frac{1}{\sigma_0^2}} \tag{3.17}$$

3. Finally we also need to update the mixture proportion $\boldsymbol{\pi}$. Putting together the multinomial likelihood and Dirichlet prior in eqs. (3.8) and (3.10), we obtain a conditional

distribution for $\boldsymbol{\pi}$ that is another Dirichlet distribution, parameterised by $[\alpha_1 + c_1, \alpha_2 + c_2, ..., \alpha_k + c_k]^T$. Each entry in this parameter vector is influenced by two values: the pseudo-count contribution from α_k and the actual counts of peaks currently assigned to cluster k from c_k.

$$\begin{aligned} p(\boldsymbol{\pi}|\boldsymbol{\alpha}, \boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{\pi}) \cdot p(\boldsymbol{\pi}|\boldsymbol{\alpha}) &\propto \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} \cdot \prod_{k=1}^{K} \pi_k^{c_k} \\ &= \prod_{k=1}^{K} \pi_k^{\alpha_k + c_k - 1} \\ &= Dir(\alpha_1 + c_1, \alpha_2 + c_2, ..., \alpha_k + c_k) \end{aligned} \tag{3.18}$$

In Gibbs sampling, each newly updated value is immediately used before sampling for the next value. This sampling of each random variable is repeated until convergence. Often a certain number of initial samples are discarded during the *burn-in* period. Since successive samples are correlated, a certain *thinning* interval is also used to reduce the number of samples used. The resulting samples can now be used to approximate the true posterior distribution of the model. Frequently, the marginal distribution of the random variable of interest is studied. Particularly for our case, often we are interested in the probability of any pair of peaks (or even a set of peaks) to be placed in the same component since, as the subsequent chapters will show, this has a direct application to the problem of peak alignment.

3.2.2 Collapsed Gibbs Sampling for a Finite Mixture Model

As we have chosen conjugate prior distributions on the mixture proportion $\boldsymbol{\pi}$ and also the cluster mean μ_k, it is possible for us to integrate (*collapse*) $\boldsymbol{\pi}$ and μ_k from the model during Gibbs sampling. This results in a collapsed Gibbs sampler (CGS) where we need not sample $\boldsymbol{\pi}$ and μ_k explicitly. Collapsing has also been shown to lead to a better model convergence [82]. It will also help in the next section when we want to extend our finite mixture model (where K the number of components is specified) to an infinite mixture model (where the number of components is unbounded) as we do not need to explicitly sample an infinite-dimensional vector $\boldsymbol{\pi}$.

Specifically in this CGS implementation, we aim to marginalise $\boldsymbol{\pi}$ and μ_k by integrating them out from the conditional probability for z_{nk}, the assignment of peak n to cluster k. Collapsing $\boldsymbol{\pi}$ introduces dependencies among all the $\boldsymbol{z}_n$ random variables, so we introduce another notation $\boldsymbol{Z}^-$ to mean all other $\boldsymbol{z}_n$s except the one for the current n-th peak being sampled upon. Similarly, $\boldsymbol{y}^-$ denotes the RT values for other peaks apart from y_n. The

conditional distribution for z_{nk} in the CGS is given by:

$$p(z_{nk} = 1, \boldsymbol{\pi}|\boldsymbol{Z}^-, \boldsymbol{y}, \mu_0, \boldsymbol{\alpha}) \propto p(y_n|\boldsymbol{Z}^-, \boldsymbol{y}^-, \mu_0) \cdot P(z_{nk} = 1|\boldsymbol{Z}^-, \boldsymbol{\alpha}) \tag{3.19}$$

We consider both terms of eq. (3.19) separately.

1. The first term on the right hand side of eq. (3.19) is the likelihood of y_n to be assigned to cluster k. Here, we no longer need to sample for μ_k as we are integrating over all values of μ_k using the posterior distribution $\mathcal{N}(\mu_k|\tilde{\mu}, \tilde{\sigma}^2)$ defined in eq. (3.16). Instead we can directly compute this likelihood by:

$$\begin{aligned} p(y_n|\boldsymbol{Z}^-, \boldsymbol{y}^-, \mu_0) &\propto \int p(y_n|\mu_k) \cdot p(\mu_k|\boldsymbol{Z}^-, \boldsymbol{y}^-, \mu_0) \; d\mu_k \\ &\propto \int \mathcal{N}(y_n|\mu_k, \sigma^2) \cdot \mathcal{N}(\mu_k|\tilde{\mu}, \tilde{\sigma}^2) \; d\mu_k \\ &\propto \mathcal{N}(y_n|\tilde{\mu}, \sigma^2 + \tilde{\sigma}^2) \end{aligned} \tag{3.20}$$

   where $\tilde{\mu}$ and $\tilde{\sigma}$ are defined in eq. (3.17).

   As an alternative parameterisation, we can also rewrite $\tilde{\mu}$ and $\tilde{\sigma}^2$ using precision (inverse variance) $\tau = \frac{1}{\sigma^2}$ and $\tau_0 = \frac{1}{\sigma_0^2}$ to replace the variances. The expression in eq. (3.17) then becomes:

$$\begin{aligned} \tilde{\mu} &= \tilde{\sigma}^2 \left(\frac{\sum_{m=1}^{M} x_m}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \right) = \frac{\tau \sum_{m=1}^{M} x_m + \mu_0 \tau_0}{M\tau + \tau_0} \\ \tilde{\sigma} &= \frac{1}{\frac{M}{\sigma^2} + \frac{1}{\sigma_0^2}} = \frac{1}{M\tau + \tau_0} \end{aligned} \tag{3.21}$$

   In the Gibbs samplers for the mixture models in the later chapters, this parameterisation using precision is what we will use.

2. The second term on the right hand side of eq. (3.19) is the prior probability of assigning the n peak to cluster k. Again we do not have to sample for $\boldsymbol{\pi}$ as we integrate over all values of $\boldsymbol{\pi}$. Our desired conditional probability is given in eq. (3.22). By definition, $P(z_{nk} = 1|\boldsymbol{\pi})$ is π while $p(\boldsymbol{\pi}|\boldsymbol{Z}^-, \boldsymbol{\alpha})$ is the posterior Dirichlet defined in eq. (3.18). This results in:

$$\begin{aligned} P(z_{nk} = 1|\boldsymbol{Z}^-, \boldsymbol{\alpha}) &= \int P(z_{nk} = 1|\boldsymbol{\pi}) \cdot p(\boldsymbol{\pi}|\boldsymbol{Z}^-, \boldsymbol{\alpha}) \; d\boldsymbol{\pi} \\ &= \frac{c_k + \alpha_k}{\sum_{k=1}^{K} c_k + \alpha_k} \end{aligned} \tag{3.22}$$

   A derivation for the integral in eq. (3.22) can be found in Ch. 24 of [83]. In the result of eq. (3.22), c_k denotes the number of data points (peaks) currently assigned to cluster

7.7 Substructure Discoveries Across Many Fragmentation Files

Manual inspection of the results revealed that many Mass2Motifs, related to the same substructures, are consistently present in two or more beers. This is despite each sample being processed independently through MS2LDA. For example, the hexose-related Mass2Motifs are present in all positive ionization mode beer files with degrees from 58 to more than 100 in each beer. The results suggest that we can jointly model the presence or absence of Mass2Motifs across many input files at once, eliminating the necessary but tedious matching of Mass2Motifs across files if they were to be inferred independently for each input file.

7.7.1 Multi-file LDA Model

Metabolomics dataset consist of fragmentation spectra in multiple input files, where each file is generated from measurements of a technical or biological replicate. Here we introduce an extension of the standard LDA model that allows for Mass2Motifs, the distributions over fragment and loss features, to be shared across files. Within each file, fragmentation spectra have their own file-specific probabilities of observing certain Mass2Motifs. When only a single input file is provided, the proposed extended model reduces to the standard LDA model. The conditional dependences of this model, which we call the multi-file LDA model, is shown in Figure 7.11 and described below.

We follow the notations used for the standard LDA model in Section 7.4, but expand them slightly to index the different files. w^f_{dn} refers to the n-th fragment or loss feature in the d-th fragmentation spectra in file f. Each w^f_{dn} is assigned to the k-th Mass2Motif, a multinomial distribution over the entire vocabulary of fragment and loss features, through the indicator variable z^f_{dn}, so $z^f_{dn} = k$ if feature n from fragmentation spectra d in file f is assigned to the k-th Mass2Motif. The probability of seeing certain Mass2Motifs for each d-th fragmentation spectra in file f is then drawn from a multinomial distribution with a parameter vector $\boldsymbol{\theta}^f_d$. This parameter vector $\boldsymbol{\theta}^f_d$ is in turn drawn from a prior Dirichlet distribution having the parameter vector $\boldsymbol{\alpha}^f$. Note that unlike the standard LDA model, each file now has its own prior Dirichlet distribution parameterised by $\boldsymbol{\alpha}^f$ and all documents in the same file has their document-to-topic distributions drawn from the same prior Dirichlet specific to the file.

$$z^f_{dn}|\boldsymbol{\theta}^f_d \sim Multinomial(\boldsymbol{\theta}^f_d) \tag{7.5}$$

$$\boldsymbol{\theta}^f_d|\boldsymbol{\alpha}^f \sim Dir(\boldsymbol{\alpha}^f) \tag{7.6}$$

As in the case of standard LDA, the k-th multinomial distribution for a Mass2Motif is still characterised by the parameter vector $\boldsymbol{\phi}_{z^f_{dn}}$, with $\boldsymbol{\phi}_{z^f_{dn}}$ drawn from a prior Dirichlet distribu-

Multi-file Latent Dirichlet Allocation

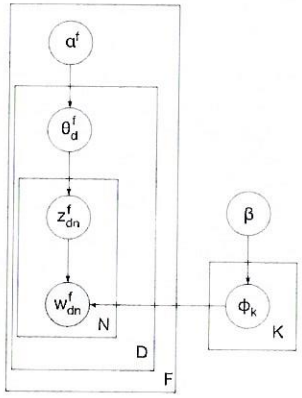

Figure 7.11: Graphical model of the multi-file LDA model. The addition to the standard LDA model is the plate on F that denotes an index over the files, $f = 1, ..., F$. Circles denotes random variables, while the shaded node denotes the observed word value.

tion that is global to all files, parameterised by the vector $\boldsymbol{\beta}$.

$$w^f_{dn}|\boldsymbol{\phi}_{z^f_{dn}} \sim Multinomial(\boldsymbol{\phi}_{z^f_{dn}}) \tag{7.7}$$

$$\boldsymbol{\phi}_k|\boldsymbol{\beta} \sim Dir(\boldsymbol{\beta}) \tag{7.8}$$

Inference in the multi-file LDA model is again performed via a collapsed Gibbs sampling scheme. The conditional probability of $P(z^f_{dn} = k|w^f_{dn}, ...)$ of the assignment of feature n in spectra d file f to Mass2Motif k is given by eq. (7.9).

$$P(z^f_{dn} = k|w^f_{dn}, ...) \propto P(w^f_{dn}|z^f_{dn} = k, ...)P(z^f_{dn} = k|...) \tag{7.9}$$

where ... denotes any other parameters being conditioned upon but not explicitly listed. Similar to the derivation of standard LDA, we can marginalise over all $\boldsymbol{\phi}_k$ parameters in the likelihood term, $P(\boldsymbol{w}^f_{dn}|z^f_{dn} = k, ...)$ of eq. (7.9), to obtain:

$$P(w^f_{dn}|z^f_{dn} = k, ...) \propto \frac{\sum_f c^f_{kn} + \beta_n}{\sum_n \sum_f c^f_{kn} + \beta_n} \tag{7.10}$$

where $\sum_f c^f_{kn}$ is the total number of the n-th feature from all files currently assigned to Mass2Motif k (this count excludes the current feature being sampled in the current iteration

of Gibbs sampler). For the prior term $P(z^f_{dn} = k|...)$, marginalising over all θ^f_d parameters produces as in the standard LDA:

$$P(z^f_{dn} = k|...) \propto c^f_{dk} + \alpha^f_k \tag{7.11}$$

with c^f_{dk} the number of features from document n in file f currently assigned to Mass2Motif k, excluding the current feature being sampled. Putting the prior and likelihood terms together, the following predictive distribution is obtained for the assignment of feature n from document d file f to Mass2Motif k:

$$P(z^f_{dn} = k|w^f_{dn}, ...) \propto (c^f_{dk} + \alpha^f_k) \cdot \frac{\sum_f c^f_{kn} + \beta_n}{\sum_n \sum_f c^f_{kn} + \beta_n} \tag{7.12}$$

In each iteration of the Gibbs sampling, the information on the current feature n in spectra d file f being sampled is removed. Reassignment of the feature to a Mass2Motif is then performed by sampling z^f_{dn} from the distribution specified by eq. (7.12). Given z, the predictive distribution for the d-th spectrum over the Mass2Motifs, θ^f_d, is obtained from the expectation of the Dirichlet-Multinomial distribution defined in eqs. (7.5)-(7.6):

$$\theta^f_{dk} = \frac{c^f_{dk} + \alpha^f_k}{\sum_k c^f_{dk} + \alpha^f_k} \tag{7.13}$$

where c^f_{dk} is the count of features from spectra d in file f assigned to Mass2Motif k.

For each spectra, the multinomial count vector $\mathbf{c}^f_d$, of features from the spectra that are assigned to the different Mass2Motifs, is a sample from the Dirichlet-Multinomial distribution defined in eqs. (7.5)-(7.6). Given all the $\mathbf{c}^f_1, \mathbf{c}^f_2, ...\mathbf{c}^f_D$ vectors in the file, the parameter α^f of the Dirichlet-Multinomial distribution of spectra-to-Mass2Motifs in file f can be estimated by maximizing the log likelihood, $log \prod_{d=1}^{D} p(\mathbf{c}^f_d|\alpha^f)$. An iterative procedure to approximate this is described in [140].

In a similar manner to standard LDA, each k-th Mass2Motif, the predictive distribution over features, ϕ_k, can be obtained as the expectation of the Dirichlet-Multinomial distribution defined in eqs. (7.7)-(7.8):

$$\phi_{kn} = \frac{c_{kn} + \beta_n}{\sum_n c_{kn} + \beta_n} \tag{7.14}$$

where c_{kn} is the count of the n-feature from all files that are assigned to Mass2Motif k.

7.7.2 Results & Discussion

On the dataset of four Beer extracts in positive ionisation mode processed through multi-file LDA using the same hyperparameters as the individual LDA. For data interpretation, initially, the same threshold values on t_θ and t_ϕ were selected as the previous single-file analysis (0.05 and 0.01 respectively). Table 7.4 shows the results of five global Mass2Motifs that could be matched to the individual LDA results in Section 7.6.2. The results in Table 7.4 shows that multi-file LDA produces comparable results on the Mass2Motifs composition. This is entirely expected given that the four Beer extracts used for evaluation share similar metabolic profiles and correspondingly, have many substructures in common.

Mass2Motif	Annotation	Top Features Above Threshold
M2M_17	Ferulic acid substructure	**fragment_177.05478**, **fragment_89.03865**, **fragment_145.02844**, **fragment_117.03319**, loss_58.98941, fragment_163.03887, **fragment_149.05998**, loss_88.09967
M2M_155	Histidine substructure	**fragment_110.07161**, **fragment_156.07687**, fragment_83.06041, **fragment_93.04511**, fragment_82.05246, fragment_209.10558, **fragment_95.06057**, loss_167.08663, fragment_81.04494, loss_191.0615
M2M_115	Leucine substructure	**fragment_86.09653**, **fragment_132.10165**, fragment_69.07013, fragment_332.112, fragment_143.11763
M2M_95	Water loss substructure	**loss_18.01031**, fragment_314.0859, fragment_296.07259
M2M_232	Asparagine substructure	**fragment_136.06231**, loss_162.03459, **fragment_119.0354**, loss_162.00534, fragment_137.04623

Table 7.4: Five global Mass2Motifs inferred from multi-file LDA. For each Mass2Motif, the top features above threshold are listed. Features characterised as key to the substructure from the previous individual LDA analyses are shown in bold.

Information from all files now contribute to the inference of global Mass2Motifs. The fact that global Mass2Motifs that are consistent with our previous characterisation in Section 7.6.2 still emerge suggests the same underlying patterns of fragment and loss features to be present in each Beer extract. Figure 7.12 shows four example fragmentation spectra originating from different Beer extracts — jointly inferred by multi-file LDA as containing the Mass2Motif characterised as the ferulic acid substructure. While this can be achieved from independently running LDA on each file, the tedious matching process of common Mass2Motifs across files can now be eliminated. Inspections on the degree (the number of spectra associated to a Mass2Motif above the user-defined threshold t_θ) of the five Mass2Motifs in Table 7.4 revealed that with a minor adjustment to t_θ, the same sets of fragmentation spectra previously associated to the listed Mass2Motifs can all be recovered.

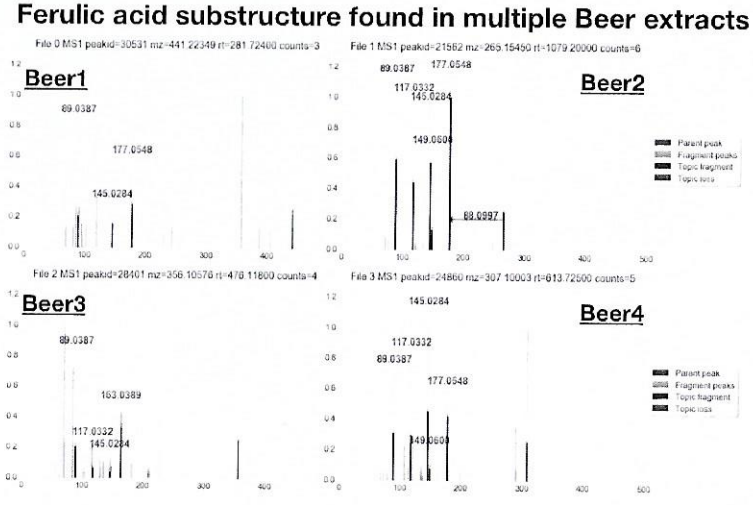

Figure 7.12: Fragmentation spectra from different Beer extracts found by multi-file LDA to contain the same Mass2Motif (M2M˙17) characterised as the ferulic acid substructure.

From each posterior sample, we can also obtain the updated $\boldsymbol{\alpha}^f$ for the different Mass2Motif across all files. As $\boldsymbol{\alpha}^f$ is the asymmetric parameter that serves as the pseudo-count in the Dirichlet-Multinomial distribution of spectra-to-Mass2Motifs, a high value of $\boldsymbol{\alpha}^f_k$ for a particular k means that a specific Mass2Motif is more likely for each spectra in file f. Figure 7.13 shows the plot of posterior alpha values for the Mass2Motifs characterised as the ferulic acid, histidine and leucine substructures. Inspections of the comparisons in Figure 7.13 may lead to interesting biological hypothesis that explains e.g. why the ferulic acid

substructure is more likely for the spectra in the third beer file compared to the others.

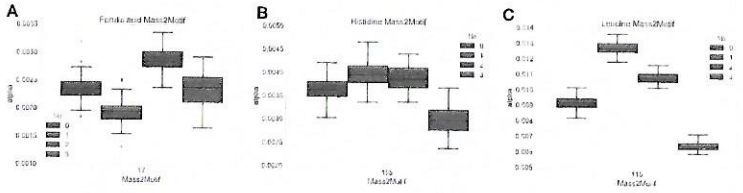

Figure 7.13: Posterior alpha values for the **A)** ferulic acid, **B)** histidine and **C)** leucine Mass2Motifs across the different beer files.

7.8 Conclusion

We have introduced MS2LDA, a pipeline that simplifies fragmentation data by exploiting the parallels between MS fragmentation data and text documents. The pipeline performs all steps required in the analysis: the preparation of a co-occurrence matrix of fragment and loss features in fragmentation spectra, the LDA analysis, and the graphical visualization of the resulting output. Evaluation of the workflow on beer extracts result in numerous informative patterns of concurrent mass fragmental and neutral loss, termed Mass2Motifs, which we could annotate as biochemically-relevant substructures. The MS2LDA approach is markedly different from other advanced spectral analysis tools as multiple Mass2Motifs can be associated with one metabolite, and determination of the key mass fragments or neutral losses that are part of a conserved structural motif is unsupervised. The application of LDA to modelling the fragmentation spectra produced by mass spectrometry instrument is exhaustively explored in this chapter. We have shown how spectra comprise of multiple substructures which can be explained by characterised Mass2Motifs. Through comparison to Molecular Networking, we demonstrated through examples how MS2LDA allows us to explain parts of a spectrum, producing a better functional annotation in contrast to spectral clustering where a spectrum can only be placed in one cluster. The differential analysis of parent ions having fragments sharing Mass2Motifs introduces the possibility of assessing changes in the expression levels of metabolites — sharing substructures explained by a characterised Mass2Motif — despite the identities of the metabolites unknown. This is particularly useful in the case of untargeted metabolomics experiments.

As future work, we envision developing a larger library of characterised Mass2Motifs from data sets produced on a diverse range of analytical platforms and different sample types. A challenge to this approach lies in the fact that mass spectrometry instruments have varying accuracy and therefore require different binning thresholds. One possible solution is define a

common space of chemical vocabulary; rather than using binned fragment and loss features; a Mass2Motif can now be defined as the distribution over chemical formulae words. Such an approach is hampered by the fact that *de novo* elemental formula assignment itself is a difficult problem, with large uncertainties as to the correctness of annotated formulae of a fragment or loss feature. A probabilistic model of formula annotation that can offers confidence values on the formulae annotation of a fragment or loss feature might be useful in this scenario as formula annotation uncertainties can then be incorporated into Mass2Motif formation in MS2LDA. Non-parametric model such as the Hierarchical Dirichlet Process [?] can also be applied for topic discovery by letting the number of Mass2Motifs to be learned from the data itself. This allows for a truly flexible system of substructure annotation where Mass2Motifs can be obtained from training the model on large public fragmentation databases, such as HDMB or MassBank. In a similar manner as our analysis in this chapter, the resulting Mass2Motifs can be characterised. New and unseen fragmentation spectra can be run using the pre-trained models with these characterised Mass2Motifs, allowing for the rapid identification of the substructure that comprise a fragmentation spectra.

An extension of the standard LDA model, in form of the multi-file LDA model, is also proposed in this chapter to handle Mass2Motif inference from multiple data sets. Such a model can be used in large-scale clinical and metabolomic studies. In this model, the prior information on which prior Mass2Motifs the user expects to see can be included into the MS2LDA workflow, allowing the LDA inference on certain known Mass2Motifs that are expected to be present in the sample while allowing others to be inferred from the data.

Other LDA-based techniques developed for text (e.g. hierarchical LDA [141]) are also likely to offer benefits as we hypothesise that Mass2Motifs can be defined in a hierarchy. For instance, generic patterns such as the loss of CO2 may lie at the top of the hierarchy of Mass2Motifs, while the more specific Mass2Motifs are formed at the bottom. It is anticipated that visualisation and the meaningful presentation of inference results will be a challenging task in such a model.

In general, we anticipate that the approach of applying topic modelling techniques to fragmentation spectra data to be particularly useful in research areas such as clinical metabolomics, pharmacometabolomics, environmental analysis, natural products research and nutritional metabolomics, as it can quickly and in an unsupervised manner recognize substructure patterns related to drugs, pollutants, and food-derived molecules, respectively.

Chapter 8

Conclusion

LC-MS based omics, such as proteomics and in particular metabolomics is important studies that play a major role in modern systems biology. However, there are a lot of challenges in the data pre-processing steps necessary before LC-MS data can be analysed. In particular, the information from the peak grouping step is often not used in the alignment and identification stage. More broadly, the presence of a grouping structure means that a set of peaks can be structurally related. Generative models can be used to induce the clustering on the peak data, revealing the latent structures that exist.

In this thesis, we have shown that using this structural information can help in alignment and identification. Our tools are generative modelling. Using this, we showed that grouping can be used to improve alignment (matching). In particular, we improve a direct matching method by by incorporating grouping information. IP clusters, corresponding to groups of peaks that are related through being the ionisation product peaks of the same metabolite, can also be matched directly – either via a direct matching scheme or through a second-stage clustering method. A generative model also can be constructed that models all peaks across all files at once, producing alignment as a result and also useful latent structures. From fragmentation data, identification can be enhanced by taking into consideration the grouping of fragmentation peaks that potentially correspond to substructures.

8.1 Summary of Contributions

This thesis makes a number of contributions, motivated by our thesis statement in Section 1.1, which is restated in the following:

Untargeted liquid chromatography mass spectrometry data pre-processing is a challenging task that is often subjected to errors and inaccuracies. Much of this can be attributed to the complexity of the LC-MS data itself and also to the lack of knowledge as to which

compounds are present in the sample. However, the structural dependencies in the observed peak data means that through generative modelling, we can explain the relationships between peaks, allowing us to produce groups of related peaks that can be used to improve or enhance the alignment and identification steps of LC-MS data pre-processing.

The thesis statement is then supported by the following contributions:

1. Chapter 4 presented a method to perform the grouping of related peaks by RT and combine this grouping information with a direct-matching method. We demonstrated on benchmark datasets with alignment ground truth how this information can be used to improve alignment.

2. Chapter 5 expands upon the grouping process in Chapter 4, where only the RT information is used, and proposes a model that takes into consideration the mass information as well when grouping related peaks. Through a set of transformation rules (specific to metabolomics data at the moment), our model produces IP clusters, where member peaks can be explained by their ionisation product transformations. We showed in Chapter 5 that IP clusters can be matched directly in place of peak features, and this produces an improved alignment performance. Additionally, uncertainties in the matching can also be quantified through a second-stage clustering of the IP clusters.

3. Chapter 6 expands upon the work in Chapter 5. Instead of having to fix the MAP cluster assignment of peaks to local clusters in the same file, we introduce the notion of a hierarchical model that allows for peaks across multiple files to be grouped. We show that modelling the data generatively in this manner and performing grouping allows us to produce alignment (matching). From the model, highly confident matched peaksets can be extracted, which may be useful in some analytical cases.

4. Chapter 7 looks at fragmentation data, produced from tandem mass spectrometry process. We show that by thinking generatively, we can explain fragment peaks by how they relate to substructures shared by metabolites. This aids in exploratory data interpretation during the identification of compounds in metabolomics data.

8.2 Future Work

There are a number of interesting future work that could follow from the results in this thesis. They are:

8.2.1 Improved Generative Models to Cluster Related Peaks

Generative modelling of peaks are demonstrated in Chapters 4 and 5 where we build a model to cluster peaks in the same file by their RT values and explainable mass transformations. However, there is more information present in the LC-MS data that is not used in our model. In particular, peaks elute from liquid chromatography and produce chromatographic profiles (the retention time value of a peak is a point along the chromatographic profile). The chromatographic profiles of related peaks should be similar, and in [40], a mixture model is proposed to cluster using chromatographic profiles. This is shown to produce improvements over the greedy approach of clustering peaks. We might also want to incorporate this information into our models, for example by changing the PrecursorCluster model from Chapter 5 and adding another likelihood term for the correlation of the chromatographic profiles. Following [40], we might we use a two-component mixtures to describe this likelihood: the first component corresponds to the likelihood of peaks to be in the same cluster, while another component describes the likelihood of peaks to be in different clusters based on their chromatographic correlations. The proposed implementation in [40] uses Gibbs sampling, and we foresee that modifying our inference procedure to accommodate this new likelihood term to be straightforward.

The proposed PrecursorCluster model in Chapter 5 also makes a fairly strong assumption that the most intense peak in the cluster must be the $M + H$ peak. This assumption may not always hold as we have seen cases where valid clusters do not have its most intense peak as the $M + H$ peak. Relaxing this assumption means more clusters may be obtained, but depending on the data, we might also see more false assignments of peaks to clusters. Performing validations on the results with and without this constraint will be challenging and require a close collaboration with a life scientist who possesses the necessary expert knowledge to validate the data. This however might point to a more flexible method where peaks can be clustered without having to make such a strong assumption.

8.2.2 Using the Generative Models for Identification

The proposed models in this thesis are generally validated against the alignment ground truth, i.e. we consider that the models produce a sensible clustering of related peaks if we can take the resulting groups and use them to obtain a good alignment performance. However, that is not the only use of the output from the models. In particular, the set of related peaks that have been grouped together and can be explained as being generated from the same latent variable (corresponding to a compound) might be used for identification. We have explored a preliminary form of this idea in Section 5 where we hand-pick clusters that correspond to Cysteic acid and melatonin, and also in Section 6 where we take some global RT clusters

and annotate them by their putative compound identities. Note that once we have assigned a putative compound identity to a clustering object, being able to annotate the entire peaks that are members of that cluster is a natural consequence of the clustering output of the model. It is worth investigating whether such an approach might bring an improved discriminative power to identification compared to identifying peaks one-by-one, as what is conventionally done at the moment. However, the lack of gold standard for identification means that this will be an extensive endeavour that again requires a close collaboration with a life scientist.

8.2.3 Data Visualisation and Interpretation

As the MS2LDAVis module in Chapter 7 shows, the interpretation of complex inference results can be daunting to the average user. Having an easy-to-use visualisation interface that displays the most pertinent information in a user-friendly manner shifts this burden of interpretation from the user to the system, and it is important when producing tools that we hope will be used and adopted by the community at large. One of the problems with the probabilistic matching results returned by the Cluster-Cluster method in Chapter 5 and also the HDP-Align method in Chapter 6 is that the result does not lend itself to easy interpretation. The conventional way of presenting a list of aligned peaksets is in the form of a table, where each row corresponds to a consensus peak (derived from the aligned peakset) and the columns are the observed intensities in the different LC-MS runs. From our output, we now obtain aligned peaksets at varying probabilities, but how about other information that we obtain from inference? From the inferred clustering structures, we obtain more than just alignment as we can also extract for e.g. the inferred ionisation product types from PrecursorCluster, the entire top-level global RT cluster from the HDP model, etc.. Displaying this information in a manner that is useful to the user requires careful considerations. For instance, we might decide to supplement the usual tabular view of peaklist with a graphical visualisation showing how peaks are explained through which ionisation product transformations and their probabilities. Integration with external database services, such as PubChem [119], is also useful in this kind of visualisation systems to obtain additional meta-data that may enhance interpretation.

8.2.4 Topic Modelling of Fragmentation Data

In our study, the multi-file LDA model proposed in Chapter 7 is applied to a metabolomics dataset containing four beer LC-MS runs. Alternative larger datasets (up to 30 LC-MS runs) from drug studies are available from our collaborators and can be run through the multi-file MS2LDA pipeline as well. This can be used to validate that indeed we can find useful

Mass2Motifs that correspond to substructures shared by drug metabolites. Also, the proposed inference procedure in Chapter 7 relies on Gibbs sampling. At the moment, the implementation will have difficulties scaling to a large number documents. Variational inference has been used in [93] to perform large-scale LDA inference, and is something we can do next for inference.

Our LDA models (both the single- and multi-file version) assumes that the number of Mass2Motifs K is known and has to be defined by the user or estimated through a cross-validation procedure (as what we have done in Chapter 7). Setting K that is too large may lead to overfitting with many small, overly specific Mass2Motifs, while setting a value for K that is too small leads to underfitting with large and generic Mass2Motifs. Hierarchical Dirichlet Process has been used as the prior in a non-parametric topic model [91] and provides a principled mechanism to let the number of Mass2Motifs to be learned from the data. This can be implemented next. Along this line, we have also seen that Mass2Motifs form *hierarchies*, with generic substructures, such as the loss of CO shared by multiple Mass2Motifs. This suggest that a hierarchical extension of the LDA model can be considered to model the data [141].

The problem of transferring Mass2Motifs that we have learned from one dataset to another is also something we need to consider as this will allow inferred and characterised Mass2Motifs to be stored in a database and applied to new, unseen data, allowing for rapid explorations of the unknowns . One way we can do this is by fixing the topic-to-word probabilities for the selected Mass2Motifs and using them when running LDA on the new data. This approach, however, is rather *ad-hoc*, and more principled approach such as [142] can be considered. For the transferring of Mass2Motifs to work, a common vocabular space over the words have to be defined on the existing data used for training and the new data. Rather than using the discretised fragment and loss features (as what we do now) that heavily depend on the mass accuracy of a particular instrument, we may explore alternative binning procedures that use the elemental formulae as the 'words' in the LDA system . The MS2LDAVis module can also be extended to allow for Mass2Motifs expressions in the different files to be compared easily. All these are the necessary building blocks that contribute towards the development of an online interactive system that the community can use to submit validated topics and apply them new dataset for a rapid exploration of the 'fragmentome' on new and unseen fragmentation data.

8.3 Summary and Conclusions

Data pre-processing is a challenging task in LC-MS preprocessing pipeline. In this thesis, we have shown how generative models can be used to explain the relationships between related peaks, allowing for groups of related peaks to be extracted. We have shown how starting

from this premise, we could propose new methods that improve on alignment and enhance identification via allowing for the untargeted exploration of fragmentation data.

Although there is a lot of work to be done still, we believe that the thesis presents a compelling case to the benefit of generative modelling of peak data. The structural information that is present in mass spectrometry data is often neglected in alignment and identification via fragmentation data. Our results show that this results in useful information that can be used to improve the quality of the preprocessing pipeline.