Chapter 2

Computational Biology

Simon: more refs, perhaps pull the proteomics and metabolomics stuff apart a bit more, careful with bold claims and mention NMR

This chapter provides the background knowledge necessary to understand the basic principles of mass-spectrometry-based analysis as applied to large-scale untargeted biological studies. 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 [3] and [4]. For literature surveys on the different steps that comprise an LC-MS data processing pipeline, the reader is directed to [5, 6, 7, 8] for metabolomics and [9, 10, 6] for proteomics.

2.1 Computational Biology

The central dogma of molecular biology states that *DNA is transcribed into RNA, which is translated into proteins*. Together, deoxyribonucleic acid (DNA), ribonucleic acid (RNA) and proteins comprise the three major types of macromolecules that are fundamentally essential to all life on Earth.

DNA is the basic storage unit of genetic information. In the central dogma, information flow begins with the double-helix strands of the DNA as the starting point. A DNA strand consists of a series of linked nucleotides subunits, where 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 re-

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.

actions 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 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. 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 four possible alphabets in the genome and transcriptome. In proteomics, the object of interest, proteins, is a chain of amino acid residues, with 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 **C**arbon, **H**ydrogen, **N**itrogen, **O**xygen, **P**hospor and **S**ulphur (CHNOPS) commonly used. Moving through the -omics layers in Figure 2.1 and getting closer the phenotypes introduces greater complexity due to the increased number of ways to putting the building blocks of that -omics layer.

Unlike the genome that is relatively static, the proteome and metabolome of an organism are

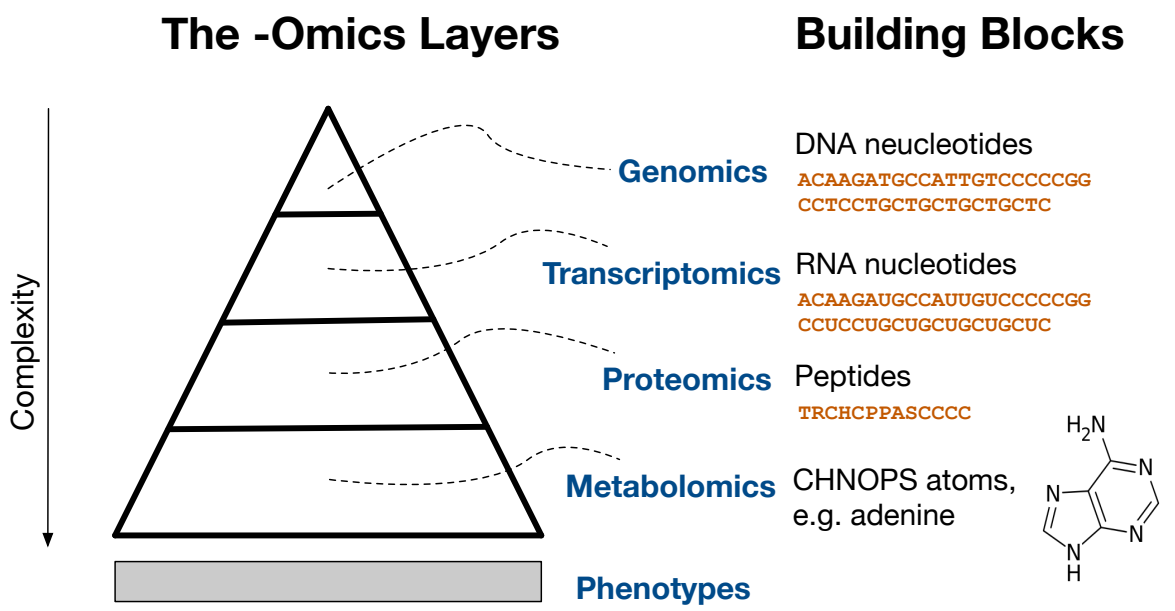

Figure 2.1: Successive -omics that get closer to the phenotype increases in complexity.

also considerably more dynamic. The expression of proteins and metabolites are governed by various complex, interacting factors. In a process called post-translational modification [11], 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 [12] or environmental factor [13]. 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 phenotype [14], as changes to the metabolome often results in changes to the physically observed properties (phenotypes) of that organism. 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 [15]. 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 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 electron, 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 are electrically charged, and they also possess 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, has a spin of 1/2 and can therefore be measured by NMR spectroscopy.

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. Depending on the ionisation mode used, the compound produce positively or negatively charged ions. These 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)** quadropole 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 [16], 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 [16]. As it stands, the two approaches are often seen as complementary rather than competitive.

In direct injection mass spectrometry, the entire compounds in the sample are introduced into the MS at once 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 low abundance compounds not being ionised and detected in the mass spectra [6]. Separating compounds as they gradually elute 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 as an additional information to deduce their identities [17]. 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).

For liquid chromatography, the mobile solvent containing the analytes (metabolites) is introduced and pumped into the stationary phase of the chromatographic column. As mentioned before, 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, cre-

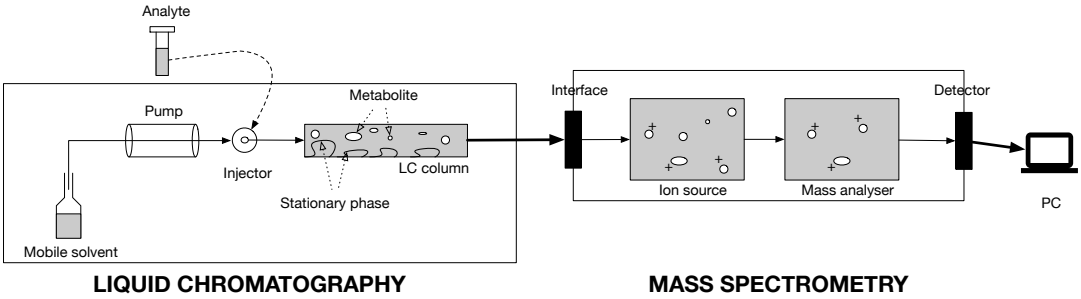

Figure 2.2: A typical LC-MS set-up. High performance liquid chromatography instruments are used to separate metabolites (by their chemical properties) in the sample before they are gradually introduced into the mass spectrometer.

ating 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. A scan is obtained for each MS measurement of compounds that elute at the same or similar retention time. 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.

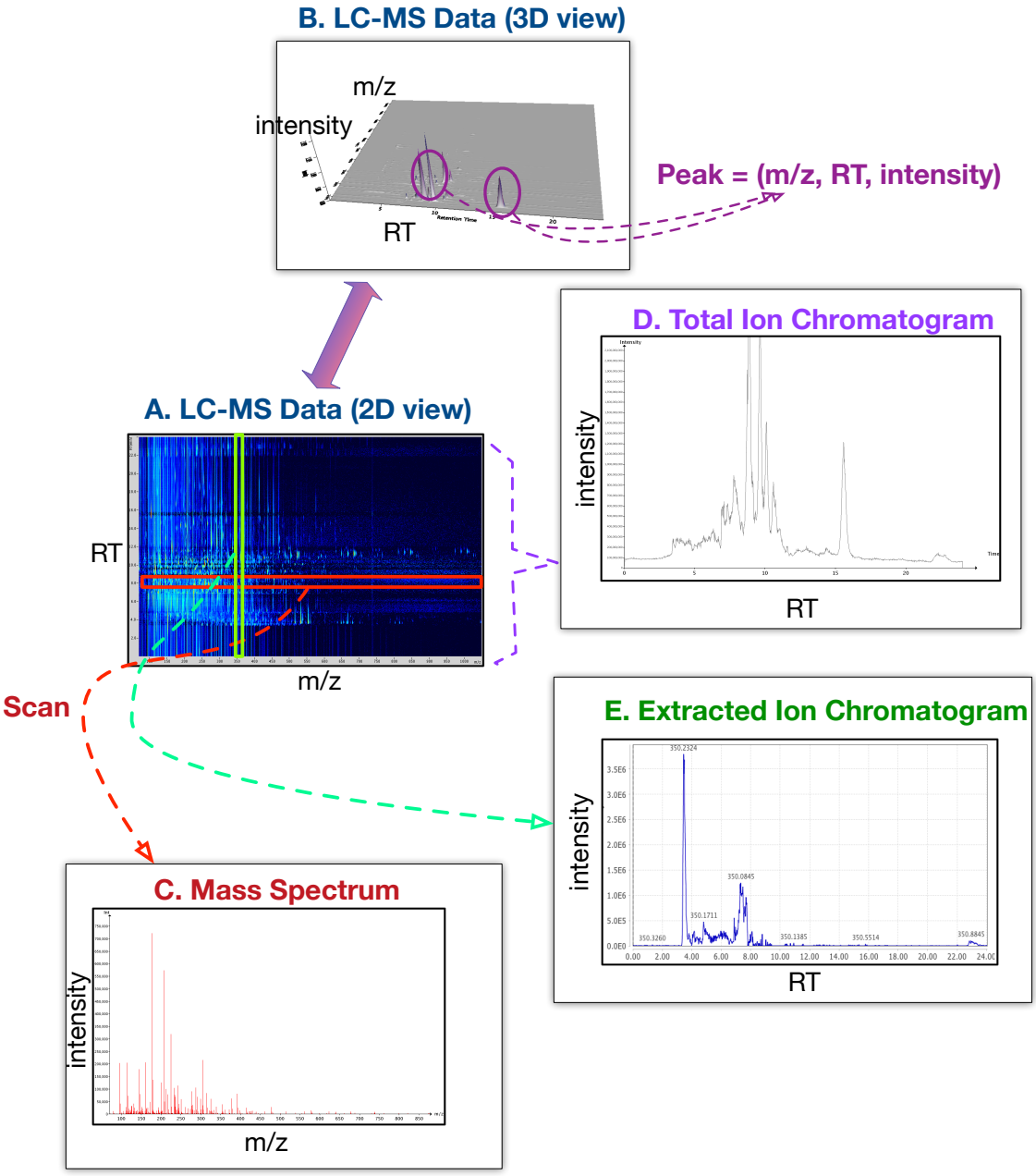

Figure 2.3: The resulting raw data (ion chromatograms) produced from an LC-MS experiment. We can view the data as a 2D profile seen from the top **(A)** or a 3D profile **(B)**. A peak in the data is thus characterised by its intensity value on the m/z and retention time axes. From a scan, a slice of 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)**.

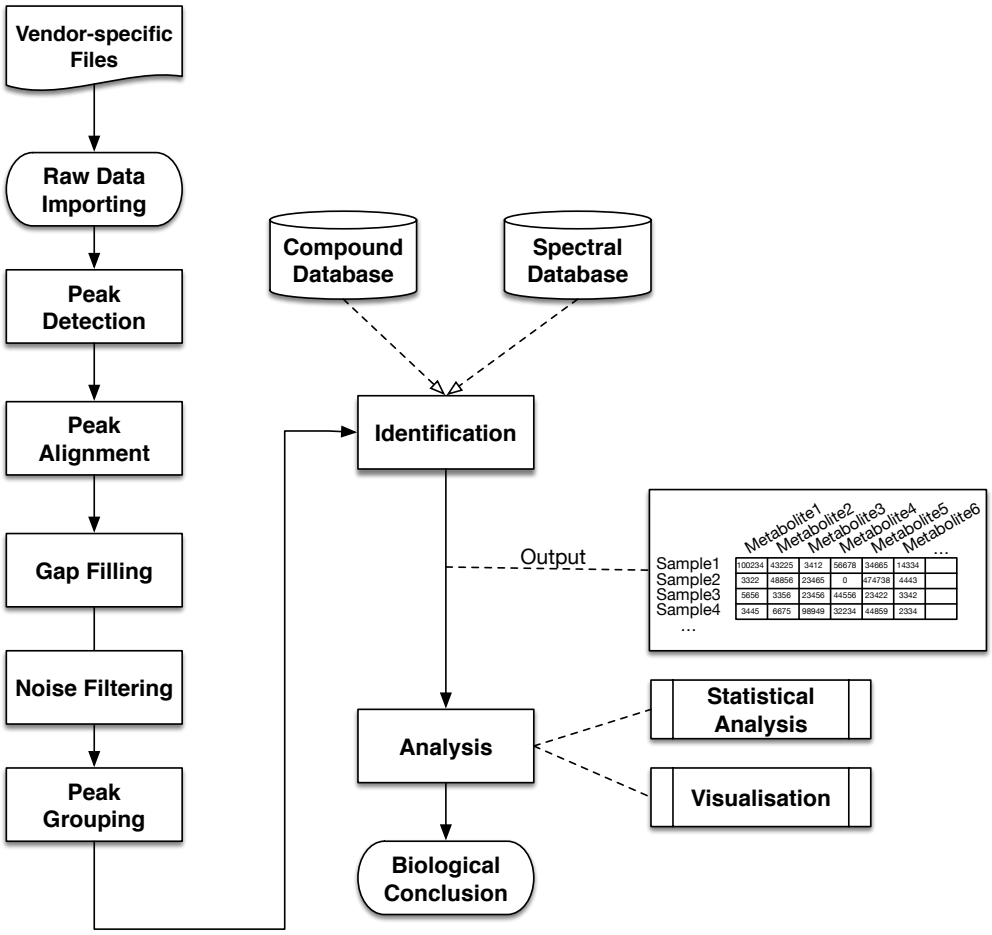

Figure 2.4: An exemplar pre-processing pipeline of LC-MS metabolomics data.

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 [18] or mzML format [19]. Peak detection is applied to the imported LC-MS data to produce peak features. Each peak feature is characterised by its m/z, RT and intensity values. The CentWave algorithm [20] 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. Chro-

matographic 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 [21] software suite is also widely used.

A survey of the many different approaches for peak detections can be found in [22, 5, 8], however it is important to note that most peak detection methods are sensitive to the choice of parameters [20], with a method potentially producing different results when its parameters are varied. For instance, CentWave requires as user-defined parameters the mass deviation 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 peak features that are the same across samples. An alignment method takes as input multiple lists of peak features — one from each LC-MS run — and produces as output a list of *aligned peaksets*. Each aligned peakset is a set of peak features 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, labelling experiment can also be done by chemically labelling metabolites in two samples with isotopic reagents. The samples are all mixed before the LC-MS experiment and combined, they are measured as a single LC-MS run. The same metabolites from two samples would generally appear at close retention time, making alignment easy. 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 [23]. Consequently, most alignment methods correct for those shifts and distortions by finding a mapping function f that maps peak features 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)** [24] 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)** [25] 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 [26], COW is combined with a component detection algorithm (CODA [27]) that removes noisy signal and background noise from the mass chromatograms, aligning only regions containing high-quality information.

3. **Parametric Time Warping (PTW)** [28] 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 [23]. 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)** [29] 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 peak features — a reduced representation of the raw LC-MS data obtained as the outcome of the peak detection step. Operating on peak features 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 peak features — 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 peak features are highlighted below:

1. **XCMS** [30] 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** [31] 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** [21] 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 peak features 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[32].

Direct-matching Alignment Methods

Direct matching methods, which skip the warping step and seek to establish the correspondence of peak features across runs directly, can be preferred due to their simplicity, while still offering good performance [33]. Most direct matching methods consist of two stages: computing feature similarity and using this similarity to match peak features 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 [21], cosine similarity [34], Euclidean distance [35], and Mahalanobis distance [36]. 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 peak features are similarly dependent on the correctness of the 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 [6].

Many approaches have been proposed for direct matching of peak features. 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 [37] merges all runs and greedily groups features into aligned peaksets within a user-defined RT tolerance. Join Aligner [21] 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 [35] 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) [36] 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. [38] explores

the application of the classical Hungarian algorithm to find the maximum weighted bipartite matching. BIPACE [34] establishes correspondence by finding the maximal cliques in the graph. SMFM [39] 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 peak features. Seen differently, if a good correspondence between peak features 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 peak features 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 [21]. 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 [36]) – or by pooling features across runs and grouping similar peaks in the combined input simultaneously (e.g. the *group()* function of XCMS in [30]).

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 [40]. 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 [41]. 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 [42], 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 [42], an implementation is provided in the mzMatch software suite [43] to detect 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 [44], the same idea is exploited in the form of a mixture model to cluster peaks based on their chromatographic peak shape correlations. CAMERA [45] 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.

2.3.5 Peak Identification

From a group of IP peaks that have been grouped from the peak grouping step, 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 in order to annotate peaks by the identities of the metabolite that generate them. Fragmentation spectra data, if available, is also used at this stage to provide an additional supporting evidence to the identity of the 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 primary technique for identification involves the querying of the precursor mass of a compound to the set of chemical possibilities in a compound database.

Having a high mass accuracy is crucial for identification as it reduces the size of possible alternatives that can be matched. However, 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 [46]. Identification is particularly difficult for metabolites present in low abundance in the samples. To reduce false positives, identification can also be performed using the group of ionisation product peaks rather than on individual peak features alone. Fragmentation through tandem MS or MS^n instruments can be used to provide further fragmentation information for metabolite 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.

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 ([47, 48]) or in relation to other -omics studies ([49, 50]).

While targeted metabolomics focuses on a handful of specific metabolites, untargeted studies (such as in [51] and [52]) 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 [53]. 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 [54], drug discovery [55] and even in an integrative approach that combines genomics and metabolomics to obtain a more comprehensive picture of living organisms [50].

2.3.7 Mass Spectrometry Analysis in Proteomics

LC-MS analysis in proteomics proceeds largely in the same manner as to the data preprocessing 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 [9]. 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 [56]. 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 [57]. 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 [26]) 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 [51, 52]), 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.

In the coming chapters, we explore the idea of using the information from peak grouping to improve the direct-matching alignment step that follows (Chapter 4). In Chapter 5, a probabilistic model is proposed to group ionisation product peaks by the set of user-defined chemical transformations. This grouping is again used to improve the alignment step. In Chapter 6, a hierarchical model is proposed to perform the grouping of ionisation product peaks across multiple runs, constructing alignment as a natural output and allowing for matching uncertainties of aligned peaksets to be returned to the user. Finally, in Chapter 7, an application of topic modelling is proposed to decompose fragmentation spectra into a set

of co-occuring fragment peaks, allowing for better hypothesis generations in the identification of metabolites present in the sample. Common to all the chapters are the idea that many peaks that exist in LC-MS data are not independent but can be explained by some underying latent variables that potentially correspond to peptides or metabolites. From the peak grouping stage, these peaks can be grouped, and this information can be used to improve the LC-MS data pre-processing pipeline.