Chapter 4

Incorporating Clustering Information into Peak Alignment

4.1 Introduction

In liquid chromatography measurements, peaks can experience non-linear shift in retention time (RT) values across runs [48]. RT 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). In direct matching, several potential matches may be present for a peak from one run to another, but because the elution order of correspondent peaks may swap across runs [18], the candidate peaks nearest in distance are not necessarily the correct match.

As described in Section 2.4.1, ionization product (IP) peaks are the set of chemically-related peaks produced from the mass spectrometry measurement of a single compound, such as a peptide fragment (in the case of proteomics) or a metabolite (in metabolomics). Examples of IP peaks are isotope peaks, multiple adduct and deduct peaks, and fragment peaks. IP peaks of the same compound have similar chromatographic peak shapes as they co-elute from the column. Such information could potentially be used to improve matching since a group of IP peaks in one run should generally be aligned to another group of IP peaks in the other run. In the direct-matching approach (discussed in Section 2.4.2), correspondent peaks from one run to another are directly matched to each other without first correcting for RT drift (instead the assumption on RT noise is built into the distance/similarity function used for matching). A direct matching method can take this structural information of IP peaks into account in order to improve alignment, however none of the direct matching methods discussed in Section 2.4.2 exploit this information.

In this chapter, we propose clustering IP peaks that share similar RT values together. This clustering information is used to modify the similarity score matrix used for matching to bring groups of IP peaks that should be matched closer, with the key assumption that groups of co-eluting peaks corresponding to the same metabolite are generally preserved across runs. This idea is illustrated in Figure 4.1 and further introduced in Sections 4.3 and 4.3.2.

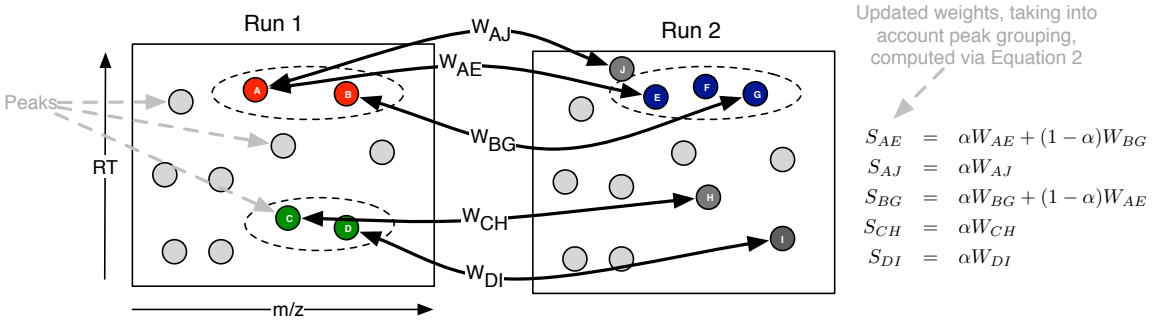

Figure 4.1: Illustrative example of the incorporation of grouping information into the similarity score. Each node in the figure is a peak feature, and dotted ovals represent groups of IP peaks, e.g. isotopes, fragments, etc. Initially weights (e.g. W_{AE}) are computed for pairs of peaks (one from each run) with m/z and RT within pre-defined thresholds. These weights are converted into an overall score by incorporating grouping information. For example, peak pairs (A, E) and (B, G) are both within the threshold. Because A and B are in the same group, and E and G are in the same group, the weights between pairs (A, E) and (B, G) are upweighted. Peak J is not related to any peaks that could be matched with A's IP peaks and the similarity between A and J is therefore downweighted (because $\alpha \leq 1$). The same applies to similarities between pairs (C, H) and (D, I).

As shown in Figure 4.1, initial weights are computed between pairs of peaks in the two runs that are within m/z and RT tolerances (e.g. W_{AE} and W_{AJ}). When related peak information is added, the similarity between peaks A and E is increased due to peak A being related to another peak (B) that is similar to a peak (G) related to E. On the other hand, the similarity between A and J is not increased as J does not have any IP peaks that could potentially be matched to peaks related to A. In other words, we are proposing using the structural dependencies present between peaks in each run to modify the similarity scores and improve alignment performance: the more peaks related to A that could be matched to peaks related to E, the more likely it becomes that A should be matched to E.

4.2 Related Work

Direct matching is introduced in Section 2.4.2, while the grouping of IP peaks is introduced in Section 2.4.3.

Statement of Original Work

The work from this chapter has been published in *Bioinformatics* [49]. The author proposed and implemented the idea of incorporating clustering information into a direct-matching alignment method. The author also performed the evaluation of performance of the proposed approach against the baseline methods.

4.3 A Direct Matching Method That Incorporates Clustering Information

Our proposed alignment method combines a novel similarity score with maximum weighted bipartite matching. This results in pairwise alignments which can be, if desired, extended to multiple alignments with hierarchical merging strategy. In such merging strategies, having an accurate initial pairwise alignments is important because of its influence on the final multiple alignment results. Here, we describe a direct matching approach to performing alignment of peaks across two LC-MS runs.

A peak feature refers to a tuple of $(m/z, RT)$ produced as output after the initial peak detection stage of LC-MS data. Here, m/z is the mass-to-charge value and RT the retention time value of a peak feature. Suppose we wish to align run A containing N_A peaks with run B containing N_B peaks. Alignment between two runs can be represented as a matching problem on a bipartite graph G, where nodes in the graph are the features, edges are the potential correspondence between features and the weights on the edges are the similarity scores (entries in S) between features. In SIMA [27], the Gale-Shapley algorithm [50] is used to find a stable matching in G. A matching is stable if there are no two features in different runs that would prefer to be matched to each other than to their currently matched partners. Since the stable matching is computed based on ranked preference, valuable information could be discarded as distances between features are converted to ranks. As such, we prefer to use a method that maximises the total sum of similarity scores of matched features (maximum weighted matching).

The benefit of maximum weighted bipartite matching in solving the peak correspondence problem has been studied in [29] in their LWBMatch tool. LWBMatch shows that such matching method, coupled to a local regression method, is able to align runs having large and systematic drifts in RT values. The well-known Hungarian algorithm [51] attributed to Kuhn and Munkres is used in LWBMatch to solve this problem. The time complexity of the Hungarian algorithm is $O(n^3)$, where n is the number of peaks in the larger set. While the Hungarian algorithm's implementation can be improved to $O(n^2 log\, n)$ by using Fibonacci heaps for the shortest path computation, the polynomial time complexity required in this

ing mixture model clustering (MWM). The time complexity of the mixture-model clustering step in MWM is $O(N)$ where N is the number of features in the run being clustered. We took 2000 posterior samples, discarding the first 1000 samples during the burn-in period. The number of samples were chosen to ensure convergence to the stationary distribution during inference.

Fraction	Total Features	MW	MWG	MWM
000	10606	9	12	2700
020	2135	1	2	524
040	2188	2	2	540
060	3342	2	3	825
080	2086	2	2	505
100	1326	1	2	321

Table 4.9: Example measured execution time in seconds on fractions of the P1 dataset

4.6 Conclusion

In this chapter, we have proposed a novel peak matching method that incorporates related peak information to improve alignment performance. The method takes related peak information in the form of peak-by-peak binary or real-valued similarity matrices and as such is independent of the particular method used to compute these. The method fits into the category of direct matching approaches — those alignment methods that do not perform an explicit time-warping phase. Our experimental results demonstrate the potential of this approach. From the training results, we see evidence of performance improvement across all evaluated datasets by incorporating grouping information into the matching process in the proposed manner. With the exception of the metabolomic dataset, both the greedy and model-based clustering approaches evaluated in our experiments rely only on the RT information for grouping IP peaks. By looking at the testing performance, our results also explore the ability of the evaluated methods to generalise on different runs using less than optimal parameters. This is important because in the actual analytical situation of LC-MS data, neither the optimal parameters nor the alignment ground truth is known.

Note that our method relies on grouping of IP peaks, and this introduces additional user-defined parameters. However, as our experiments have shown, in some settings, it may be much easier to produce good groupings of IP peaks than accurately determine RT window parameters (the same grouping parameters were used for all evaluation datasets in the case of mixture-model clustering). Depending on the nature of the data, parameters relating to within-run characteristics (e.g. RT window for grouping IP peaks) may be more likely to generalise across runs and experiments than parameters relating to between-run character-

istics (particularly RT). For example, changes in the liquid chromatography (LC) column would likely result in related-peaks still co-eluting but could significantly change the absolute RT.

It would be interesting to investigate in greater detail any performance improvements that can be obtained from using other peak grouping methods, such as [55] that uses a mixture model of peak shape correlations or [37] that considers the dependencies between adduct and isotopic peaks when clustering. Exploring alternative approximate matching algorithms (such as the scaling algorithm in [52], which provides a $(1 - \epsilon)$ approximation of the maximum weighted matching in optimal linear time for any ϵ) and evaluating the benefits of incorporating different clustering approaches into our proposed alignment method are avenues for future work. Finally, the different alignment methods evaluated in this chapter also suffer from variable behaviours depending on the order of the runs being aligned [7]. This is particularly true in the case of alignment of multiple runs (typical in large-scale LC-MS studies), where the final alignment results are often constructed through merging of intermediate alignments of pairwise runs. Different alignment methods may employ a different merging approach, for example, Join merges the intermediate results towards a reference run while SIMA allows the possibility of using a greedy hierarchical merging scheme. Systematic evaluation on how the chosen merging scheme may influence alignment performance is beyond the scope of this chapter and is an item for future work.

A limitation of the proposed approaches lies in the fact that the clustering of IP peaks are performed based on RT only. The valuable information present in the m/z domain is not used for clustering. The grouping of IP peaks based on their m/z information is less straightforward as peaks that are related (sharing close RT values) do not necessarily have m/z values that are close to each other. In the evaluation on the complex metabolomic dataset, we observe that the proposed approach using RT clustering manages to improve training performance (due to overfitting) but fails to produce any statistically significant improvement in the testing performance due to its limited generalisation ability. In the next chapter, we address this issue by focusing specifically on metabolomics and proposing a clustering model that explicitly takes into account the chemical relationship between IP peaks in LC-MS-based metabolomics.

Chapter 5

Precursor Clustering of Ionisation Product Peaks

5.1 Introduction

Chapter 4 explores the idea of using the clustering of ionization product (IP) peaks to modify the similarity scores used for matching with the aim of improving alignment results. However, the MWG and MWM methods in Chapter 4 lies on clustering based on the retention time alone. Valuable information present in the mass-to-charge (m/z) domain and also in the chemical relationships of IP peaks is not used for clustering. In this work, we extend upon the methods in the previous chapter and propose a novel Bayesian mixture model (PrecursorCluster) to cluster IP peaks based on m/z and RT values. The key difference from the mixture model RT clustering introduced in the previous chapter lies in how PrecursorCluster uses a set of user-defined transformation rules to relate peaks to a common precursor mass, allowing IP clusters to be formed through the grouping of peaks that share chemically meaningful relationships.

Building upon the clustering results returned by PrecursorCluster, two alternative alignment methods (illustrated in Figure 5.1) are introduced for aligning IP clusters across runs: **(i)** Cluster-Match, a fast direct-matching method of IP clusters that uses the posterior precursor mass and RT values of IP clusters to compute the approximate maximum-weighted matching of the IP clusters and **(ii)** Cluster-Cluster, a second-stage clustering model that constructs alignment by means of grouping IP clusters according to their likelihood of being assigned to the same top-level cluster (corresponding to metabolites shared across all runs). In this manner, IP clusters assigned to the same top-level cluster are considered to be matched. The actual alignment between their member peaks are established by grouping member peaks that share the same IP type. The Bayesian approach in Cluster-Cluster also allows us to incorporate additional information for alignment in a principled manner by adding likelihood

terms. As an example, we illustrate this in Cluster-Cluster by including a likelihood term on the different adduct types of IP peaks assigned to an IP cluster. This allows IP clusters to be placed in the same top-level cluster — and correspondingly having their member peaks matched — only if the characteristic adduct 'signatures' of IP clusters are similar.

The aim of this chapter is to evaluate whether through the proposed methods, the matching of IP clusters can improve upon the matching of peaks alone. For the purpose of evaluations, two benchmark datasets of standard and beer mixtures, alongside their associated alignment ground truth and a list of 14 adduct transformations in positive ionization mode, were used. Using precision, recall and F_1-score as evaluation measures, the performance of the proposed method of matching IP clusters (Cluster-Match) were compared against the direct matching of peak features (MW) and its variant (MWG) in the previous Chapter 4 that modifies the similarity matrix used during matching to bring together group of peaks related by RT closer during matching. Additionally, the probabilistic matching results produced by Cluster-Cluster is also described, demonstrating that it is possible to use its output to extract aligned peaksets with varying degrees of confidence. Cluster-Cluster were evaluated with and without the adduct signature term to determine whether through the addition of that likelihood term, we can obtain better alignment results.

5.2 Related Work

It is suggested in [18] that the objective function used for alignment can be improved by operating on groupings of IP peaks rather than using individual peaks. In addition, [56] proposes minimising an objective function that uses groups of isotopic peaks as objects to be matched, but does not provide any implementation or evaluation on the effectiveness of the proposed objective function. In MetAssign [37], a Bayesian mixture model was introduced to perform the identification of a set of observed peaks based on how well they fit the theoretical mass spectrum of a metabolite computed from a given formula. While the groupings of related peaks extracted from PrecursorCluster can potentially be used in a similar manner as MetAssign to perform a more robust annotation of metabolites present the sample, here we investigate its uses in improving the alignment step. Unlike MetAssign, PrecursorCluster does not require a prior library of possible metabolite formulas to be specified to perform ionization product clustering, relying only on prior chemical knowledge of which ionization transformations are expected to be present in the data. CAMERA [33] approaches the problem of ionization product clustering from a graph-theoretic perspective. In CAMERA, peak features are nodes in a graph, and edges are drawn between nodes if their scores are greater than a predefined threshold. The graph is clustered to find highly-connected subgraphs, and edges in the subgraph are annotated by known rules of chemical transformations. Unlike

CAMERA, PrecursorCluster is a fully probabilistic model, relying on Bayesian inference to update the probabilities of which LC-MS peak features can be explained by which transformations into IP clusters. This additional information can be used to provide an estimate to the uncertainty of IP annotations. The Bayesian model proposed in PrecursorCluster can also be easily extended to incorporate additional sources of information (e.g. chromatographic peak shapes) for clustering peaks in a different manner.

Since alignment is such an important part of the data preprocessing steps, it is useful to be able to robustly identify the uncertainty or confidence in the alignment results. In the absence of ground truth information (typically the case in untargeted metabolomics experiment), the user measures alignment quality through manual inspection or by comparing and visualising the summary statistics (e.g. median, standard deviation of retention time) across different replicates. Alignment methods that can produce matching confidence values is a big research gap that, to our knowledge, has not been addressed by any of existing direct-matching tools. Tools such as MAVEN [57] assigns quality scores to individual peaks by training a predictive model on expert-annotated training data of peak quality metrics, but this does not extend to scoring groups of peaks. Other approach like [58] computes the Pearson correlations between intensity profiles of all peaks across replicates. Moving from these approaches towards a robust method that can provide confidence values for groups of aligned peaks across many label-free experiments is challenging research problem.

The subject of identifying and quantifying uncertainty has been extensively investigated in the problem of multiple sequence alignment (MSA) for genomics and transcriptomics. The probability on MSA alignment allows researchers to focus on regions of the genome that are difficult to align, potentially revealing evolutionary insights as such regions have high alignment uncertainty that can be the result of e.g. the lack of conserved sequences. [59] attempt to quantify the alignment uncertainty of the popular MSA tool ClustalW [60], based on evaluations using synthetic data, and concludes that between half to all columns in their benchmark MSA results contain alignment errors. [61] construct a score that reflects the consensus between all possible pairwise alignments in T-COFFEE, while [62] propose GUIDANCE, a confidence measure obtained from pertubations of guide trees. Statistical approaches that provide a measure of confidence in alignment results have also been explored by [63] and [64], where the MSA results and phylogeny are constructed simultaneously, thus eliminating the need for a guide tree.

Despite the clear benefits of alignment uncertainty quantification in the sequence domain, the challenge of quantifying alignment results remains relatively unaddressed for the alignment of multiple runs in LC-MS-based-omics. Uncertainties on aligned peaksets can be used by metabolomic researchers to flag the low-probabilities peaksets from further analysis. The flip side of this is high-probabilities peaksets, that we are more confident of as being aligned correctly, can be selected for subsequent analysis in the pipeline or as the focus of manual

validation in a targeted manner if they are revealed as corresponding to metabolites having an interesting differential change across samples. Several recent feature-based alignment methods incorporate probabilistic modelling as part of their workflow, making it possible to extract some form of scores or probabilities on the alignment results. These methods are often limited to the alignment of two runs, which is not a realistic assumption in actual LC-MS experiments. For example, [65] propose a model for pairwise peak matching. Matching confidence can be obtained from the model in form of posterior probability for any peak pair in two runs, however constructing multiple alignment results in [65] still requires a greedy search to find candidate features within m/z and RT-RT tolerances to a predetermined set of 'landmark' peaks. [66] describe PeakLink, a workflow for alignment that performs an initial warping using a fourth-degree polynomial. PeakLink poses the pairwise matching problem as a binary classification problem, where a Support Vector Machine (SVM) is trained based on an alignment ground truth derived from MS-MS information and used to differentiate matching and non-matching candidate pairs to produce the actual alignment results. While not explicitly included in the output of PeakLink, a matching score can be extracted from the SVM that represents how far each candidate pair is from the decision boundary. Note that these scores are not well-calibrated in the probabilistic sense, thus making comparisons of matching scores less straightforward. PeakLink is also not extended to the problem of aligning multiple runs, although [66] state that it would be possible to do so with the choice of a suitable reference run.

Statement of Original Work

The work from this chapter has been submitted for review to *Bioinformatics*. The author proposed and implemented the idea of clustering IP peaks by their transformations, and also the matching of the resulting IP clusters to construct alignment. The author also performed the evaluation of performance of the proposed approach against the baseline methods.

5.3 Methods

The workflow is illustrated in Figure 5.1. A novel Bayesian model, **PrecursorCluster**, is introduced to group related peaks into IP clusters (Section 5.3.1). Each LC-MS run is processed separately through PrecursorCluster — the model takes as input the list of m/z, RT and intensity values of peak features and the list of user-defined transformations and produces as output the set of IP clusters per run. Alignment of IP clusters across runs are performed through **Cluster-Match** (Section 5.3.2) or **Cluster-Cluster** (Section 5.3.3). From Cluster-Match, a list of aligned peaksets (the set of peak features matched across runs) is

Table 5.3: Precision, recall and F_1 values from Cluster-Cluster for randomly selected sets of 2, 3 and 4 Standard runs (averaged) and the Beer runs for various l and thresholding levels $th = \{0.30, 0.60, 0.90\}$. Best results from Cluster-Match and the result of running Cluster-Cluster without the adduct fingerprint term are shown for comparison. Note that for Cluster-Cluster, the results come from using one set of potentially sub-optimal parameters for the second-stage clustering.

Dataset	l	Best Cluster-Match			Cluster-Cluster (CC)				CC (without adduct term)
		Avg. Prec.	Avg. Rec.	Avg. F_1	Threshold	Avg. Prec.	Avg. Rec	Avg. F_1	Avg. F_1
Standard	2	0.93	0.92	**0.93**	0.30	0.96	0.95	0.95	**0.95**
					0.60	0.98	0.93	**0.96**	0.93
					0.90	1.00	0.90	0.94	0.80
Standard	3	0.89	0.90	**0.89**	0.30	0.82	0.91	0.84	**0.86**
					0.60	0.86	0.88	**0.86**	0.85
					0.90	0.89	0.81	0.84	0.62
Standard	4	0.87	0.92	**0.89**	0.30	0.81	0.92	0.85	**0.89**
					0.60	0.84	0.89	0.85	0.86
					0.90	0.90	0.83	**0.86**	0.65
Beer 3 runs	3	0.92	0.89	**0.91**	0.30	0.76	0.77	**0.77**	**0.79**
					0.60	0.88	0.67	0.76	0.68
					0.90	0.94	0.54	0.68	0.63

posterior samples, Gibbs sampling for PrecursorCluster requires 20 minutes to process one Standard run on an Intel Core i5, 3.3GHz PC. Runs are processed independently and can be parallelized. In Cluster-Match, the matching of IP clusters via MW has a time complexity of $O(m \log n)$ time, where n and m are the number of vertices and edges in the bipartite graph to be solved, translating to a wall clock of less than a minute for each run. Cluster-Cluster requires longer computational time. With 1000 posterior samples per top-level bin, the processing of 2 Standard runs requires approximately half an hour. Each top-level bin can also be processed in parallel.

5.6 Conclusions

We have proposed an integrative workflow that performs the precursor clustering of ionization product peaks and uses that to improve alignment. The PrecursorCluster model introduced is a data reduction process that can reduce the number of peaks to IP clusters based on a list of possible ionization transformation types. The clustering information extracted from PrecursorCluster can be used to improve other steps in the pipeline too. For instance, metabolite identification, currently the main bottlenecks in high-throughput metabolomics, might be improved through analyzing IP clusters as the objects of interest rather than individual peak features. In this chapter, the PrecursorCluster model is optimised on metabolomics data by focusing on the set of adduct transformations as ionization product transformations. However this does not preclude the model from being applied to other MS-based omics as

well. For instance, adduct peaks are less of a problem in proteomics data, and since peptide fragments being fragmented are larger than metabolites, the resulting proteomic spectra often contains more isotopic peaks. This rule on isotopic transformations can also be incorporated as part of PrecursorCluster.

One of the key assumption made in PrecursorCluster is that the peak with M+H transformation must be the peak having the largest intensity in the IP cluster (other peaks are not allowed to join the IP cluster if their intensity values are smaller than the M+H peak). While this is a reasonable assumption to make, some IP peaks do not obey this modelling assumption. An alternative clustering methods that are more flexible and does not have the intensity constraint can be considered, e.g. by allowing peaks to form IP clusters if they can be transformed (within tolerance) to any potential precursor mass that other peaks also jointly 'vote' for. The weight of a potential precursor mass can then be updated based on the likelihood of the set of peaks having valid transformation paths to that precursor mass.

Taking the MAP results from PrecursorCluster, we have also demonstrated how IP clustering can be used to improve alignment. Our results show that in comparison to the conventional direct-matching of peak features, the proposed approach Cluster-Match, which performs the matching of IP clusters and subsequently groups member peaks having the same IP types, allows us produce a better (more precise) alignment result. It is also noteworthy that while Cluster-Match still makes the assumption that across runs, correspondent IP clusters always exist, this assumption is more relaxed when it comes to the construction of the actual alignment of peak features. Since member peaks in matched IP clusters are grouped according to their IP types, peaks that do not have correspondent type across runs will never be matched together — even if they are close in distance and would otherwise has been matched in the conventional direct-matching scheme. In this manner, Cluster-Match can potentially produce fewer false positives in matching compared to the direct matching of peak features alone.

The proposed pairwise matching scheme used by Cluster-Match is frequently extended to the processing of multiple runs in a fairly ad-hoc manner and suffers from having to set a reference run (which can be considered another parameter to set). Producing a distance measure that works well for measuring similarities of peaks across multiple runs is non-trivial, in particular in the merged-match scheme from the sequential processing of pairs of runs. We propose the Cluster-Cluster method that addresses this issue by not requiring a reference run when constructing alignment through a second-stage clustering of IP clusters. To our knowledge, no literature has systematically evaluated the effect of choosing a different reference run for alignment or how changing the order of runs being processed might affect the alignment result, but we hypothesise that methods like Cluster-Cluster that does not require a reference run will have an advantage when aligning a large dataset — typical in modern large-scale metabolomics experiments having hundreds of runs to process.

In addition, most methods also do not take into account the uncertainties inherent in the matching of peak features across runs. Cluster-Cluster is able to return aligned peaksets at varying probabilities. Our experiments show that by setting a suitable threshold, we can extract from Cluster-Cluster results alignment results having a higher precision than what can be obtained from other methods. As future work, an interactive visualisation module can be developed to let user visualize ionization product clustering and aligned peaksets (with their probabilities) from a single graphical interface. Such module can be incorporated as part of a larger metabolomics pipeline.

A weakness of the alignment methods described in this chapter is the fact that as a second-stage clustering step, both Cluster-Match and Cluster-Cluster requires the MAP assignment of peak features into their IP clusters from PrecursorCluster. The complete uncertainties from PrecursorCluster are not propagated to the matching stage. The next chapter addresses this problem by introducing a fully-hierarchical model that performs the clustering of peak features within run and across runs at once.

Chapter 6

Hierarchical Clustering of LC-MS Peaks

6.1 Introduction

The Cluster-Cluster method introduced in Chapter 5 performs the direct-matching of peak features in ionisation product (IP) clusters that themselves have been clustered together. However, related peak features are assigned into IP clusters based on their maximum *a-posteriori* probabilities. In this chapter, we expand upon the idea of alignment as a hierarchical clustering problem by proposing **HDP-Align**, a Bayesian non-parametric model that groups related peaks within runs by their retention time (RT) and assigns them to global clusters shared across runs. Within each global cluster, peaks are further grouped by their m/z values into mass clusters, representing the various ionisation products derived from the global compound. In this manner, the local clusters in HDP-Align correspond to the within-file IP clusters from running PrecursorCluster on each run, while the global clusters in HDP-Align correspond to the top-level clusters produced from Cluster-Cluster (described in Chapter 5).

The proposed HDP-Align model introduced in this chapter allows us to infer the matching of peaks across all runs at once without the need for any intermediate merging of pairwise runs. Similar to the Cluster-Cluster model introduced in Section 5.3.2, the proposed model of HDP-Align also introduces the possibility of allowing the user to trade recall for precision from the alignment results by returning a smaller subset of the results having a higher confidence score of being correctly aligned. Figure 6.1 shows an illustration of the clustering process in HDP-Align. Additionally, the latent variables inferred in the model may correspond to chemically meaningful compounds and can be used for further analysis. Using a metabolomic dataset, we demonstrate the usefulness of such latent objects by using

the mass clusters derived from the model and a set of defined ionisation product transformations to perform the putative annotations of peaks based on their potential adduct types and metabolite identities.

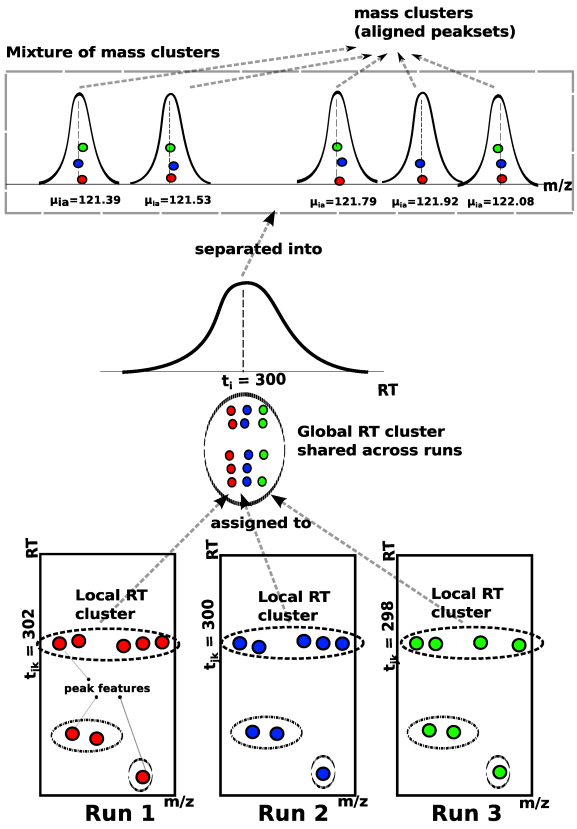

Figure 6.1: An illustrative example of how the proposed model in HDP-Align simultaneously **(1)** performs the clustering of related peak features into within-run local clusters by their RT values, **(2)** assigns the peak features to global RT clusters shared across runs, and **(3)** separates peak features into mass clusters, which correspond to aligned peaksets.

6.2 Related Work

The goal of establishing the matching of peaks across multiple runs at once can be viewed as a clustering problem, where a set of peaks can be grouped (by their m/z, RT and other suitable features) into local clusters within each run (representing all of the peaks from an individual compound), which are further grouped into global clusters shared across runs. Hierarchical clustering has been used for the matching of peak features across runs [68, 69]. In [68], peaks are hierarchically clustered based on their m/z values to construct matching across runs, while in [69], peaks from the entire dataset are pooled and a hierarchical clustering scheme based on RT only is used to group peaks into within-run local clusters, which are further grouped into across-run super clusters. Both approaches require choosing various

user-defined parameters, such as determining a suitable cut-off for the dendogram produced, deciding on a suitable linkage method and defining an appropriate distance measure between groups of peaks. In [68], no chromatographic separation is performed, so only the m/z values of peaks are used. The nature of the gas chromatography data used in [69], where retention time across runs is more reproducible, means that even without using the m/z information, good alignment performance can still be obtained. This will not be the case of LC-MS data, where retention time drift is common and the highly accurate m/z information is crucial for alignment. The proposed HDP-Align model fills this gap where both m/z and RT values, important for LC-MS peak alignment, are used for the hierarchical clustering process. The probabilistic approach employed by HDP-Align also allows us to extract confidence values from aligned peaksets.

6.3 Hierarchical Dirichlet Process Mixture Model for Alignment

The proposed model for HDP-Align is framed as a Hierarchical Dirichlet Process (HDP) mixture model [70]. Essential modifications to the basic HDP model, described in Section 3.5, were performed to suit the nature of the multiple peak alignment problem. Figure 6.2 shows the conditional dependencies between random variables in the HDP-Align model.

Our input consists of J input files, indexed by $j = 1, ..., J$, corresponding to the J LC-MS runs to be aligned. Each j-th input file contains N_j peak features in total, which can be separated into K_j local clusters of related-peak features. In a j-th file, peak features are indexed by $n = 1, ..., N_j$ and local clusters are indexed by $k = 1, ..., K_j$. Across all files, we assign each local cluster k in file j to a global cluster $i = 1, ..., I$, where I is the total number of global clusters, using the indicator variable v, as described in the following paragraph. A global cluster corresponds to the compound of interest during LC-MS analysis, e.g. metabolite or peptide fragment, that is present across runs, while local clusters are realisations of the global clusters in a specific run. Finally, within each global cluster i, we can further group peak features by their m/z values into A mass clusters (indexed by $a = 1, ..., A$). Each mass cluster therefore corresponds to the ionization product peaks coming from the different runs that are produced by a global compound during mass spectrometry.

We use the indicator variable $\boldsymbol{z}_{jnk} = 1$ to denote the assignment of peak n in file j to local cluster k in that file. Similarly, $\boldsymbol{v}_{jni} = 1$ if peak n in file j is assigned to global cluster i, and $\boldsymbol{v}_{jnia} = 1$ if peak n in file j is assigned to mass cluster a linked to metabolite i. Let d_j be the list of observed data of peak features in file j, $d_j = (\mathbf{d}_{j1}, \mathbf{d}_{j2}, ..., \mathbf{d}_{jn})$ where $\mathbf{d}_{jn} = (x_{jn}, y_{jn})$ with x_{jn} the RT value and y_{jn} the log m/z value of the peak feature. The log of m/z value is here used as the m/z error is assumed to increase linearly with the observed m/z value [71].

Dataset	Benchmark (SIMA, Join)
P1 Frac 000	$T_{(m/z)} = \{1.0, 1.1, ..., 2.0\}$, $T_{rt} = \{10, 20, ..., 180\}$ s
P1 Frac 020	
P1 Frac 040	
P1 Frac 060	
P1 Frac 080	
P1 Frac 100	
Glycomic	$T_{(m/z)} = \{0.05, 0.1, 0.25\}$,$T_{rt} = \{5, 10, ..., 120\}$ s
Metabolomic	$T_{(m/z)} = \{0.001, 0.01, 0.1\}$, $T_{rt} = \{5, 10, ..., 120\}$ s

Table 6.3: Parameters used for the benchmark methods (SIMA, Join).

6.6 Results and Discussions

The performance of the evaluated methods methods on the different datasets are presented in Sections 6.6.1 and 6.6.2. Additionally, an example of the further annotations for the putative adduct type and metabolite identity that can be produced by HDP-Align is also shown in Section 6.6.2.

6.6.1 Proteomic (P1) Results

Figure 6.4 shows the results from performance evaluation on the Proteomic (P1) dataset. We see that both benchmark methods (SIMA and Join) produce a wide range of performance depending on the parameter values for $(T_{(m/z)}, T_{rt})$ chosen. Sensitivity to parameter values is expected on this dataset due to the low mass accuracy in the MS instrument that produces the data and the high RT drifts present across runs (further details in [23]). HDP-Align performs well on several fractions (particularly fractions 040, 060, 080, 100) with precision-recall performance close to the optimal performance attainable by the benchmark methods. On all fractions, HDP-Align is also able to produce higher-precision results compared to the benchmark methods by reducing recall through setting the appropriate values for the threshold t. The primary benefits of quantifying alignment uncertainties is realised here as the well-calibrated probability scores on the matching confidence of aligned peak features produced by HDP-Align allows the user to choose which point along the PR curve to operate on. It is less obvious how this can be accomplished in the benchmark methods by varying the RT (T_{rt}) and m/z ($T_{m/z}$) thresholding parameters, if at all possible.

The P1 datasets also represent the most challenging alignment scenario as they have the largest RT drift and low mass accuracy in comparison the glycomic and metabolomic data. We use the largest (P1 Fraction 000) and the smallest (P1 Fraction 100) from P1 to examine how well our chain converges during Gibbs sampling.

Figure 6.5 (left) shows the traceplots of the number of global clusters from three randomly

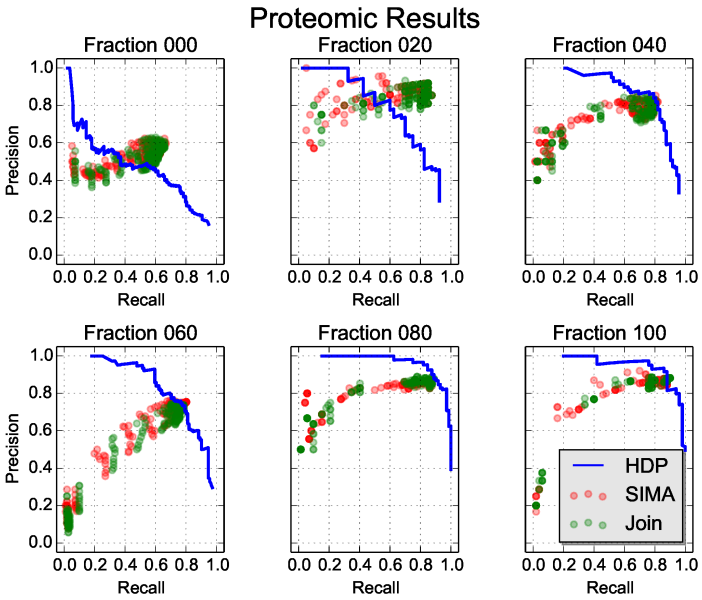

Figure 6.4: Precision-recall values on the different fractions of the Proteomic (P1) dataset.

initialised MCMC chains when running HDP-Align on the largest P1 Fraction 000 dataset containing 10606 features. From the jumps in the number of global clusters in the traceplots, we see some evidence of bad mixing in the chains. This is explained by the fact that in our sampler, we do not allow for the block reassignment of a group of peaks (that are together placed in a local cluster) into a new global cluster. Consequently, a global cluster can only be deleted when all its individual peak features have completely moved elsewhere. This leads to the slow convergence and poor mixing of the model. An inspection of the distributions of global clusters after burn-in from the three chains, shown in Figure 6.5 (right), also suggests that the chains have not fully converged yet.

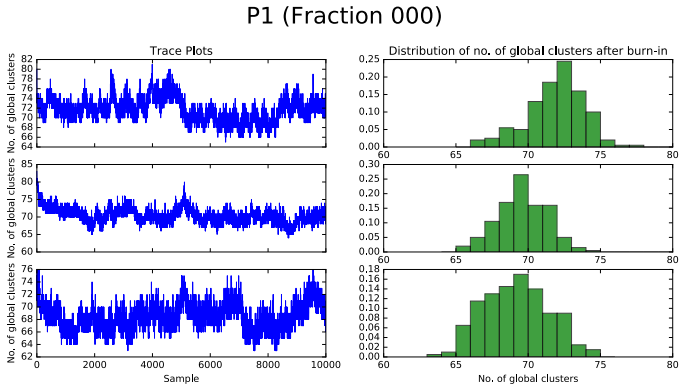

Figure 6.5: Traceplots from three randomly initialised MCMC chains for the number of global clusters across all posterior samples for the largest fraction (000) from the Proteomic (P1) dataset.

Running HDP-Align on the P1 Fraction 000 data and collecting 10000 posterior samples

requires several days of walltime. The main factor affecting the running time of HDP-Align is the total number of peaks across all runs to be processed and the number of samples produced during Gibbs sampling. In each iteration of Gibbs sampling, HDP-Align removes a peak from the model, updates parameters of the model conditioned on every other parameters, and reassigns a peak into RT and mass clusters. In practice, additional time will also be spent on various necessary book-keeping operations, such as deleting empty local and global clusters that are no longer required, updating internal data structures, etc. Running a longer chain may not be not entirely practical in actual analytical situation — particularly in comparison to the speed of the baseline methods that completes in minutes. Despite this poor mixing, as the results show in Figure 6.4, we still see some evidence that by reducing recall, it is still possible to extract from HDP-Align alignment results having a higher precision than what the baseline methods can achieve.

Inspecting the diagnostic plots Figure 6.6 for the smaller P1 Fraction 100 dataset, we observe a better mixing behaviour. The traceplots that are less jumpy and the distributions of the number of global clusters that are more consistent across the three chains. This may be due to the smaller (1326) number of features in this dataset. On this dataset, HDP-Align took two hours to process. This is a reasonable time a user can tolerate for a data analysis pipeline to complete (although still significantly longer than the baseline methods that require seconds to complete). Again from Figure 6.4, the results for this P1 Fraction 100 dataset show that by trading off recall, we can obtain a higher precision than the baseline methods.

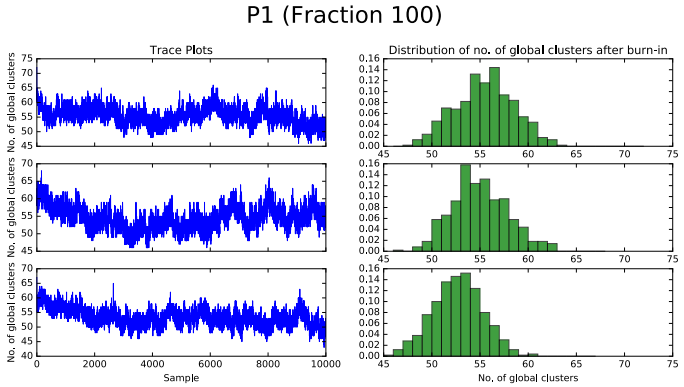

Figure 6.6: Traceplots from three randomly initialised MCMC chains for the number of global clusters across all posterior samples for the smallest fraction (100) from the Proteomic (P1) dataset.

6.6.2 Glycomic and Metabolomic Results

Figures 6.7 and 6.8 show the results from experiments on the Glycomic and Metabolomic datasets. Similar to the Proteomic dataset, a range of precision-recall values can be observed

in the results for the benchmark methods on the two datasets. Consistent with our expectation, reducing the tolerance window on the retention time produces a smaller recall value, however this does not necessarily result in a better alignment precision. The performance of HDP-Align, using the same set of parameters on both datasets, come close to the optimal results from the benchmark methods, while still allowing the user to control the desired point along the precision-recall curve to operate on.

The results for the Glycomic dataset (Figure 6.7) also show some additional results on how the measured precision-recall values might change depending on the strictness of what constitutes an alignment item during performance evaluation. This is accomplished by gradually increasing the value for l that determines the size of the feature combinations enumerated from a method's output. For example, l=2 considers all pairwise combinations of features from the method's output during performance evaluation, while $l = 4$ considers all combinations of size 4, and so on. Figure 6.7 shows that as l is increased, parameter sensitivity seems to become more of an issue for the benchmark methods, with more parameter sets having lower precisions in the results. Across all ls evaluated, parameter pairs that produce the best alignment performance (points with high precision and recall values) are generally small $T_{(m/z)}$ and large T_{rt} values. Examples of parameter pairs that produce the best and worse performance for SIMA are shown in Figure 6.8. The results here appear to suggest the importance of having high mass precision during matching. Importantly, we see from Figure 6.7 that the performance of HDP-Align remains fairly consistent as l is increased.

The Metabolomic dataset also provides us with additional results in form of annotations of putative adduct type and metabolite identities. A thorough evaluation on the quality of such annotations, in comparison to e.g. the workflow proposed in [2], is beyond the scope of this chapter and would likely necessitate using a different and more appropriate evaluation dataset. Instead, we present an example of the further analysis performed by HDP-Align (as proposed in Section 6.4.4) on the resulting clustering objects after inference. Figure 6.9 shows a global RT cluster where peak features across runs have been grouped by their RT and m/z values. Within this global cluster, peak features are further separated into 6 mass clusters – corresponding to ionisation products produced by the global cluster during mass spectometry. In Figure 6.9, mass cluster A and B contain features aligned from several runs but they do not have any other mass cluster sharing a possible precursor mass. Mass cluster C and D share a common precursor mass (292.12696) and can thus be annotated by the adduct type that produce the transformation. Similarly, mass cluster E and F share a common precursor mass at 383.14278. Queries to a local KEGG database are issued based on the precursor mass values, producing several compound identities that can be putatively assigned to the global RT cluster. It is a strength of the HDP-Align approach that this putative identification step appears very naturally from the alignment results.

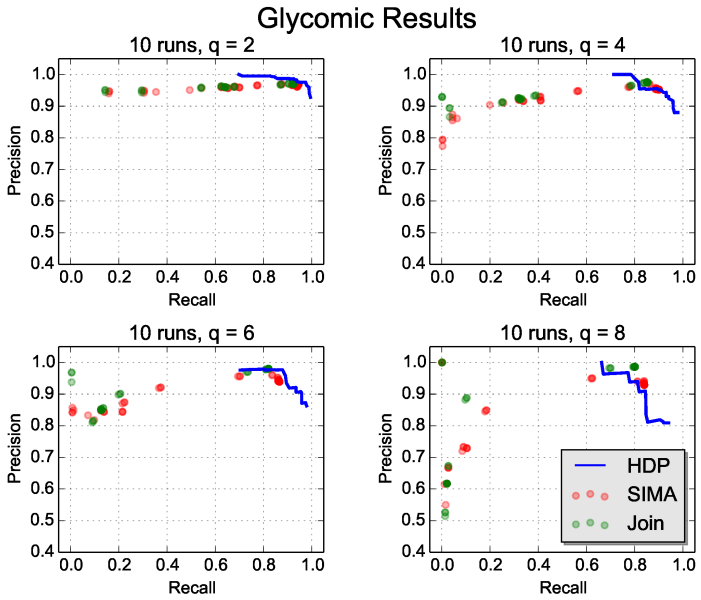

Figure 6.7: Precision-recall values on the alignment of 10 runs from the Glycomic dataset when q (the strictness of performance evaluation as described in Section 5.4.2) is gradually increased.

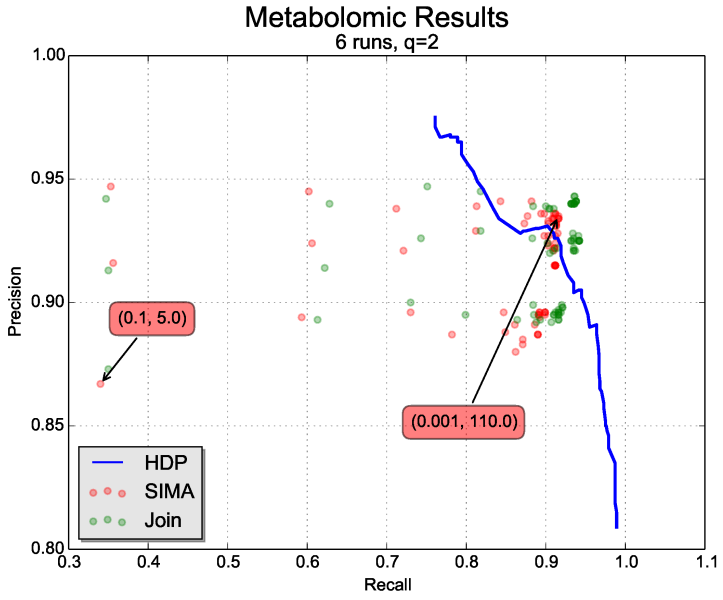

Figure 6.8: Precision-recall values on the alignment of 6 runs from the Metabolomic dataset. The parameter values $(T_{m/z}, T_{rt})$ that produce the best and worst performance in SIMA are also annotated in the Figure (red boxes).

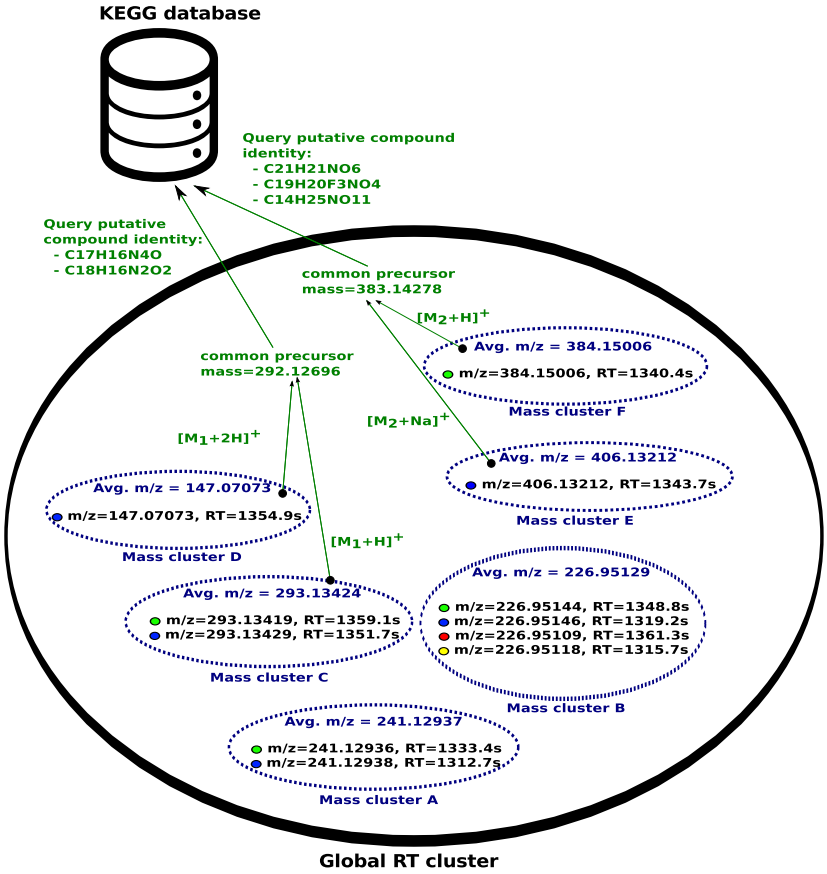

Figure 6.9: Example analysis that can be performed on the clustering objects inferred in a Gibbs sample from HDP-Align. The outer black oval denotes a global RT cluster (generally corresponding to a metabolite compound), while the smaller dotted ovals within denote mass clusters (labelled as mass cluster A, B, C, D, E, F). Peak features are denoted by the filled circle, with the fill colour indicating the originating run of a peak. Green colours denote additional analysis steps that can be performed on the mass cluster objects.

6.7 Conclusion

In this chapter, we present HDP-Align, a hierarchical non-parametric Bayesian model that simultaneously performs the within-run clustering and across-run clustering of peak features. As a natural consequence of the clustering process, the direct matching of peak features can be extracted from the model. In addition, the clustering objects from the model can also be used for further analysis in the pipeline as they potentially correspond to the actual chemical compounds that generate the generate. Similar to the two-stage clustering methods (Cluster-Cluster) introduced in the previous chapter, the HDP-Align model is able to produce well-calibrated probability scores on the matching confidence of aligned peak features (evidenced by the increasing precision and decreasing recall as the threshold t is increased). This is accomplished by casting the multiple alignment problem of LC-MS peak features as a hierarchical clustering problem. Matching confidence can be obtained based on the probabilities of co-eluting peak features to be placed under the same mass component in the same global cluster. Experiments based on datasets from real proteomic, glycomic and metabolomic experiments show that HDP-Align is able to produce alignment results competitive to the benchmark direct-matching alignment methods, with the added benefit of being able to provide a measure of confidence in the alignment quality. This can be useful in real analytical situations, where neither the optimal parameters nor the alignment ground truth is known to the user.

A primary weakness of HDP-Align lies in the long computational time required to produce results. This is due to the slow mixing of the chains, as the consequence of our incremental Gibbs sampling that samples one variable at a time. The split-and-merge MCMC algorithm for the HDP proposed in [72] may help to improve sampling performance and is an avenue for future work. The actual running time for the sampling can also be improved by taking the lessons from the Cluster-Cluster approach introduced in the previous chapter, for instance it may be possible to partition the data into subsets of peaks based on their retention time as only peaks within a certain RT tolerance should ever be clustered and matched to each other. The key insight of the HDP-Align model lies in the way related IP peaks are modelled as within-file clusters in a single run but the model also allows these within-file clusters to be generated by globally-shared clusters spanning multiple runs. The results presented in the current chapter suggest the method shows enough promise to warrant the effort to speed it up.

Additional sources of information present in the LC-MS data, such as chromatographic peak shapes, can also be used to improve alignment performance and subsequent analyses that follow. The mixture of mass components used in HDP-Align with a more appropriate mass model, such as that in MetAssign [37] that specifically takes into account the interdependency structure of peaks. Alternatively, it may be possible to build the transformation

rules employed by the PrecursorCluster model (from the previous chapter) into HDP-Align. However, such modifications will introduce even more complexity to an already complex model, requiring a more sophisticated inference scheme and perhaps an even longer running time.

Through comparisons against benchmark methods, our studies have also investigated the effect of sub-optimal parameter choices on alignment performance. While beyond the scope of our paper, we agree with [18, 53] that thorough investigations into the influence of numerous configurable parameters (prevalent in nearly all LC-MS data processing pipeline) on the resulting biological conclusions are of utmost importance. This should be followed by the development of methods to minimise or automatically-tune such configurable parameters. Despite the abundance of new methods proposed for LC-MS data pre-processing, relatively few studies have been done on the subject of quantifying uncertainties and alleviating the burden of parameter optimisations during actual data analysis. One way to minimise the number of parameters is through the integration of multiple steps in the typical LC-MS pipeline into fewer steps. Our proposed model in HDP-Align can potentially be extended in this manner, as evidenced by the metabolomic dataset results where we directly use the clustering objects inferred from the model to perform further analysis on putative adduct and metabolite type annotations. While the proposed annotation approach in Section 6.4.4 is fairly simple, it can be easily extended to more sophisticated annotation strategies, such as in CAMERA [33]. This will be particularly useful when we aim to extend the proposed model in HDP-Align into a single inferential model that encompasses many intermediate steps in a typical LC-MS data processing pipeline.

Our experiments of taking the clustering objects from HDP-Align and using them to assign metabolite identities to the clusters through matching to a compound database also shows the potential of HDP-Align in assisting compound identification by allowing identifying labels to be assigned to a group of matched peaks from several runs at once. However, identification of metabolites, particularly in large-scale untargeted experiments, is challenging. The next chapter explores this in greater details and proposes the use of a different type of structural information, present in mass spectrometry fragmentation data, to improve identification.

Chapter 7

Substructure Discovery in Tandem Mass Spectometry Data

7.1 Introduction

As the results from Chapter 5 shows, the ionization product (IP) types of many observed peaks are often unknown and therefore the molecular mass of metabolites that generate these peaks are also unknown. This makes identification difficult as mass is often a required information when querying metabolite identities against publicly-available databases, such as KEGG [73] and PubChem [74]. In addition, while modern mass spectrometry instruments can be highly accurate up to 3 parts-per-million (ppm), even a mass accuracy of 1 ppm is not sufficient to reliably determine the elemental composition (formula) of a metabolite [75] during database queries. The presence of isomers (metabolites having the same formula and mass but are structurally different from each other) suggests that when relying on mass alone, the same peak might be incorrectly matched to multiple isomeric metabolites. Retention time (RT) might help to distinguish certain isomers that have different elution profiles, but RT drift, a main challenge in alignment, means observed RT values can vary across different chromatographic platforms and cannot be easily used as a characteristic information in public databases during identification. Apart from the small number of metabolites present in a standard solution that can be identified with a high degree of confidence (as they produce measured peaks having reliably known m/z and RT values), information on the mass and RT values alone are not enough to establish the identity of many metabolites in untargeted studies.

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

Fragmentation patterns can be used to aid identification through the matching of a query spectrum to a database of reference spectra. In recent years, a growing number of fragmentation spectra databases have been made public, including METLIN [76], ChemSpider [77] and MassBank [78]. However, mass spectral databases are not comprehensive and contain only a small number of known metabolites. The large variance in submitted spectra further limits potential matches as sensible results can only be obtained when matching spectra generated from measurement platforms having similar characteristics (for e.g., produced through the same ionization method under a similar mass accuracy). According to [79], approximately 2% of spectra in an untargeted metabolomics experiment can be matched and subsequently identified – a small number in contrast of the vast collection of metabolites that comprise the metabolic pathways of an organism.

Multiple metabolites can share the same chemical substructure. For example, carboxylic acid (Figure 7.1) is a generic substructure shared by many amino acids and organic acids, such as acetic acid (CH_3COOH) that is commonly found in vinegar or butyric acid ($CH_3(CH_2)_2COOH$) that is present in butter. During fragmentation in positive ionization mode, the neutral carboxyl group (COOH) breaks from the parent ion, forming CO and H2O (the extra hydrogen in H2O comes from the addition of a positively charged proton, H+, during ionization). From this, we can expect to observe a characteristic neutral loss in the spectra of metabolites that share carboxylic acid as a substructure. In a fragmentation spectrum, this will be represented by a fragment peak that is 46 Da smaller than the mass of the parent peak. Thinking generatively, observing a neutral loss of 46 Da therefore provides a hint that a fragmentation spectrum is generated from the measurement of a metabolite that contains carboxylic acid as a substructure.

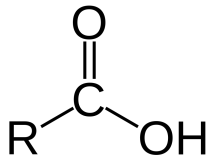

Figure 7.1: The carboxylic acid substructure. In the diagram, R refers to the residue, which is the rest of the metabolite attached to this substructure.

As illustrated by the very simplified example above, the knowledge of the constituent substructures that comprise a metabolite, particularly of the larger and more specific substructures, can be used to provide a hint as to the overall identity of the metabolite. Classification method, such as Support Vector Machine, decision tree and neural networks [80, 81, 82, 83], have been trained to learn spectral features that represent substructures and predict the presence or absence of substructures from fragmentation spectra. Combined with information from the parent peak (such as the m/z, RT values and IP types if available), this provides additional information that can aid in the identification of metabolites that cannot be resolved

through the traditional method of spectral database matching alone.

A common shortcoming of these classification approaches higlighted before is the need of the supervised training of the classifier (classification-based approaches may fail to generalise well to new dataset produced from different analytical platforms). Based on the assumption that fragmentation spectra contain fragment peaks that represent shared substructures of metabolites, we propose a workflow that applies the Latent Dirichlet Allocation (LDA) model to spectral fragmentation data. The proposed workflow produces the decomposition of fragmentation spectra (equivalently a document in standard LDA) into the set of *Mass2Motifs* (equivalently a topic in standard LDA). Here, a Mass2Motif is defined to be the recurring set of fragment peaks and neutral losses that potentially correspond to a biochemically-relevant substructure shared by many metabolites. Unlike the classification-based methods highlighted earlier, the decomposition of fragmentation spectra into Mass2Motifs is achieved in an unsupervised manner. The MS2LDA workflow is introduced in Section 7.4.

7.2 Related Work

Clustering is commonly used for group fragmentation spectra that are similar to each other. Clusters of spectra can be used for identification by forming a consensus spectrum and matching it against spectral databases. Molecular networking clusters MS1 peaks by their MS2 spectral similarity such that one identifiable metabolite in a cluster facilitates structural annotation of its neighbors [84, 85, 86]. However, only MS2 spectra with high overall (e.g. cosine) spectral similarity are grouped in Molecular Networking. Consequently Molecular Networking may fail to group molecules that share small substructures. In particular, spectra may be placed in different clusters if they share a small number of fragment peaks that related to a common substruture, but their overall global similarities are too different. Even for spectra placed into the same cluster, often manual analysis (by eyes) is required to select the characteristic fragment peaks that represent a potential substructure and are shared by members of the clusters. Another package, MS2Analyzer [87] mines MS2 spectra given the prior knowledge on the fragment patterns of interest to be specified in advance. While generic features, such as CO or H2O losses, will be common to many experiments, sample-specific features can be easily overlooked if they have not been specified *a priori*.

The assumption that spectral consist of building blocks that correspond to substructures is alluded in certain works but not directly mined from the data. Prior knowledge on substructures have been used for the annotations of a small number of moelcules in fragmentation data [88] and for metabolite classification in GC-MS [89, 81]. In CSI:FingerID [83], a fragmentation tree is used to predict (using Support Vector Machine) the molecular ‘fingerprint’, computed through the implicit assumption that fragments share substructures, of an unknown

compound. The resulting fingerprint is used to improve the matching of spectrum of the unknown compound against a vast chemical database (PubChem). Implicit in these methods are the assumption that recurring patterns of fragment peaks and neutral losses values explain the presence of common biological substructures (e.g. a hexose unit, or a CO loss) shared by metabolites.

Latent Dirichlet Allocation has not been applied to metabolomics or mass spectrometry data, but it has been applied to other fields of computational biology in e.g. genomics [90], metagenomics [91], and transcriptomics [92]. In [90], DNA sequence from genomics studies is decomposed into recurring patterns of N-mers nucleotides. A topic in this context corresponds to the set of N-mers (e.g. 'ATGC' as an instance of a 4-mers) that co-occur together across the different genomic sequences of a species, and the objective of the study is characterise the sets of N-mers that corresponds to conserved genes of the species. Similarly in [91], a metagenomic read (essentially a DNA sequence) is decomposed into its topic distribution. The unsupervised decomposition of metagenomic reads into topic distributions is used to improve the binning (clustering) of reads from the same species. In [92], a sample or gene from transcriptomics studies is decomposed into multiple processes in a manner similar to how a document is decomposed into different topics in traditional LDA for text.

7.3 Statement of Original Work

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

JvdH then analysed the results from MS2LDA for biochemical significance. To assist JvdH in his analysis, the author proposed and developed the visualisation module, MS2LDAVis. To improve the visualisation module, the author integrated elemental formula annotation functionalities. This includes writing a wrapper in MS2LDA to call SIRIUS [93], a Java-based elemental formula annotator. Cristina Mihailescu (CM) implemented another Python-based elemental formula annotator, which was also customised and integrated into MS2LDA by the author.

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

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 [70] 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 [102]) 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.