University of Glasgow

MSc Bioinformatics, Polyomics and Systems Biology

Initial Investigations into Multi-file MS2LDA

Francesca Young

Student Number: 2226768Y

Supervisor: Dr Simon Rogers

A report submitted in partial fulfilment of the requirements for the MSC Bioinformatics, Polyomics and Systems Biology Degree at The University of Glasgow.

August 2016

Summary

Untargeted metabolomics has become an important tool in biomedical research. However its potential is severely restricted by the lack of suitable computational methods available for metabolite identification. Current methods focus on using reference databases which are leaving the majority of metabolites unidentified and vast amounts of potential information untapped. The MS2LDA pipeline currently under development is a novel approach that has adapted Latent Dirichlet Allocation, a text mining algorithm to mining MS data. It identifies molecular substructures that are found across the dataset in a completely unsupervised way. It is able to cluster structurally related metabolites and to assign substructure annotations to individual metabolites. A recent adaption of MS2LDA has been implemented to process multiple samples into a single model.

In this project I make the initial investigations into the new information available from the models generated. This information is contained in a set of parameters that are extracted from the data by multi-file MS2LDA. I explored the potential of using these parameters as a measure of the differential expression of substructure content. I demonstrated that this reduced set of features could be used to group samples correctly using both PCA and hierarchical clustering. I then developed a set of tools for the statistical analysis and visualisation of the differential expression measured by these parameters. Finally I used these tools to analyse models of large datasets to validate both the tools developed and the new MS2LDA extension. The new tools were able to find identify regulated motifs that made biochemical sense both using the statistical tests and an interactive PCA plot. I demonstrate that analysis of these parameters have significant potential to be used as a high level tool for phenotyping on substructure.

Table of Contents

Acknowledgements:

I would like to thank my supervisor Dr Simon Rogers for his support and guidance throughout this project.

I would also like to thank Joe Wandy and Justin Van Der Hooft for their advice and suggestions.

Abbreviations

LC- MS	Liquid Chromatography Mass Spectrometry
MS1	Mass Peak in the mass spectrum when ions of a specified m/z are detected.
MS2	Tandem MS/MS spectra from a single MS1 peak (not necessarily a single ion)
DDA	Data Dependant Acquisition. Mode of data collection in tandem mass spectrometry in which a fixed number of precursor ions are stage MS/MS analysis.
PCA	Principal Component Analysis
DE	Differential Expression the difference in Motif distribution across samples
LDA	Latent dirichlet allocation
MS2LDA	Pipeline in which LDA has been adapted for MS2 data.

1 Introduction

Metabolomics is the analysis of the complete small molecule content or metabolome of a biological system. The metabolome is downstream of genetic and environmental influences and is directly linked to phenotype giving a snap shot of the current state of a biological system. Metabolomics close link to phenotype makes it an ideal tool in a wide range of applications, including cancer research, precision medicine and plant sciences.

Metabolomics is the youngest and least developed of all the 'omics' technologies, at present metabolomics struggles to identify or characterise even a small fraction of the metabolome. Current methods can result in in less than 2% of metabolites being identified (Silva et al., 2015). The main methods available for processing this data use comparison to reference spectra in public databases.

This inability for metabolite identification severely limits the potential of this technique. There is an urgent need for new computational solutions to expand the power of metabolomics to enable it to become as routine as genetic sequencing or RNA-seq. This requires the development of novel approaches that can extract the underlying biochemical information from the data without necessarily achieving full structural annotation. (Treutler et al., 2016),

MS2LDA (Van der Hooft et al. 2016) is an innovative approach that applies topic modeling to the problem of information retrieval from MS/MS datasets. Topic modeling aims to uncover common themes or topics within a cohort of text by identifying the co-occurrence of words. By applying this approach to MS data we aim to find common substructures in the metabolome. Latent Dirichlet Allocation (LDA) does this by uncovering patterns of co-occurring m/z peaks in the MS2 spectra for all the fragmented molecules in all the sample. This pipeline enables metabolites to be clustered according to their substructure and resulted in about 70% of the 1000 fragmented MS1 features being annotated by at least one biochemically relevant sub-structure. It has reduced the complex data from 1000 MS2 spectra to 300 motifs of which 30 were annotated. This can help in identifying all the MS1 features present in the experiment that are in a common pathway.

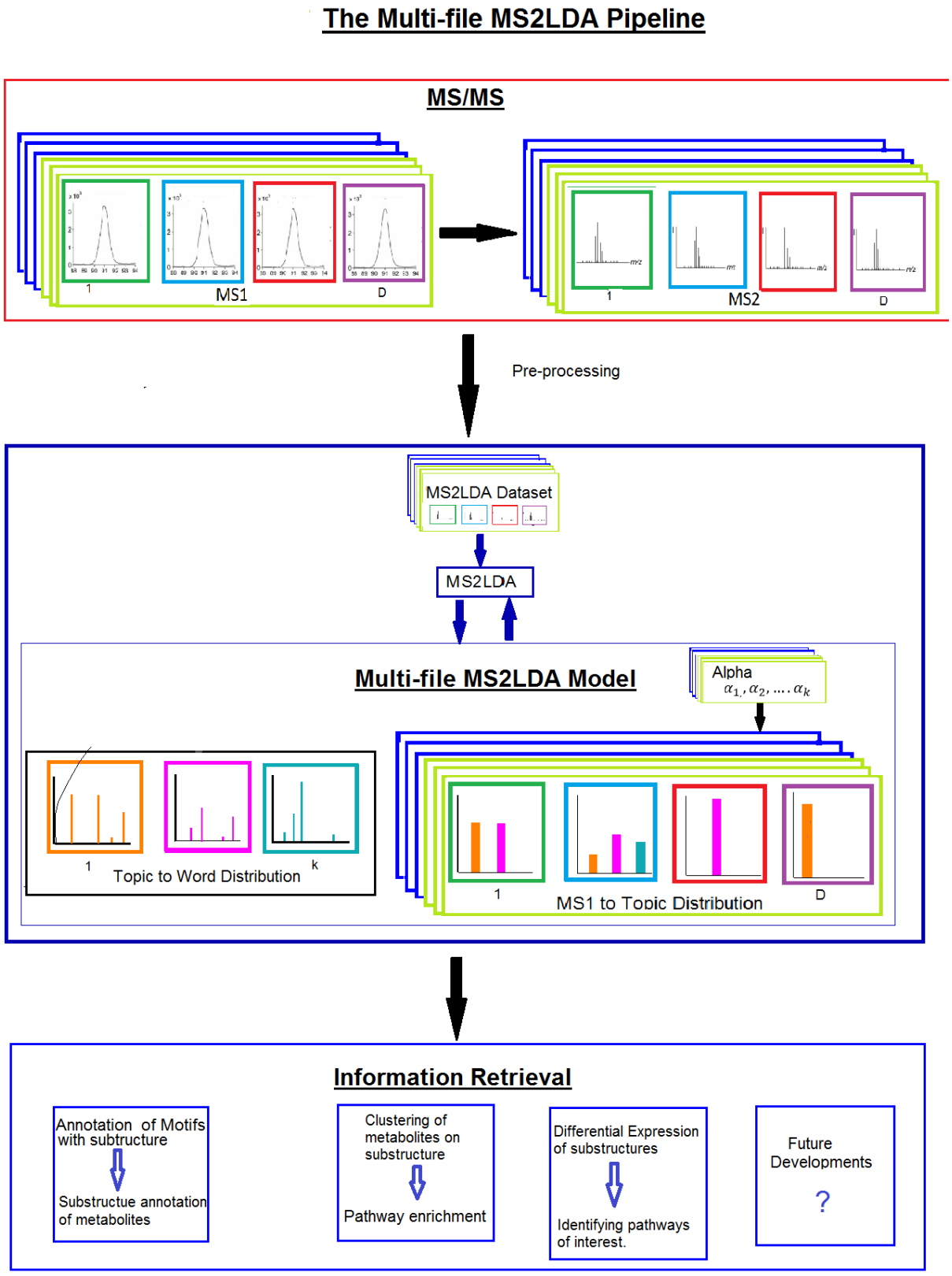


Figure 1-1 The Multi-file MS2LDA pipeline. This shows the MS2LDA pipeline for **S** samples in two groups. The MS2 data for each metabolite in each sample are processed together into 1 model with **k** topics, a topic distribution for each MS2 spectra in each sample and one set of alphas of length k.

A prototype has recently been developed to extend the MS2LDA pipeline to be able to process multiple samples into one model (Figure 1.1). In this model each sample has its own set of parameters, alpha, which are determined by the distribution of the motifs within the sample. This model only generates one set of motifs meaning that their distributions can be directly compared across all the samples.

The original MS2LDA has been shown to be able to aid metabolite identification based on substructure content and to cluster metabolites based on these substructures. In this project I investigate the potential of this new MS2LDA extension to to reveal biochemically meaningful differences in the chemical substructures present in samples.

This project had three different phases of exploration and is part of an ongoing development of the MS2LDA pipeline. The initial exploration started off by looking at the differences in alphas between samples. I then generated a model of mixed beer and urine samples and looked at the models ability to separate these two groups. The second phase involved developing statistical tools to analyse and visualise the differential expression of the motifs across groups of samples and to annotate the motifs. The final phase involved using the developed code on two models generated from large datasets to validate both the code and the potential of the multi-file MS2LDA pipeline to find biochemically relevant DE.

2 MS2LDA - Latent Dirichlet Allocation for Mass Spectrometry Data

Topic modeling is an efficient algorithm for making sense of large volumes of unlabelled data, it is a powerful tool for exploring and making inferences about cohorts of documents. It has been successfully applied to text document summarisation and classification and to information retrieval. It can be adapted to search for patterns in any sorts of discrete data

and has been successfully applied to social networks, images and genetic data (Sakthivel et al., 2014)

The MS2LDA is a software pipeline currently being developed by Simon Rogers, Joe Wandy, Justin van der Hooft, and others to apply the LDA algorithm to metabolomic MS/MS fragmentation data. In tandem MS (MS/MS) data dependant acquisition a small proportion of the most 'interesting' MS1 peaks are selected for further fragmentation. This gives rise to a signature m/z spectra arising from the fragments (and neutral loss data) of the parent molecule which can yield information about its chemical substructure.

The advantages of LDA over other topic models is the flexibility introduced by the dirichlet parameters. Metabolites can contain more than one chemical substructure, this is reflected by the results from MS2LDA that identified out of the 600 MS2 spectra the majority contained 2 or more motifs. This method allows 3 things

- Clusters molecules on sub-structure which allows us to infer information about the molecules containing this topic. The added flexibility of LDA means a molecule made up of more than one substructures will be clustered with each relevant topic.
- Groups of metabolites which have similar structures will have the effect of clustering metabolites from a common pathway together e.g. drug metabolism
- Can be used to look for clusters of substructures that change together across conditions for example disease/non-disease, drug response/no drug response.

The analogy between LDA applied to text documents and MS2 data is shown in Figure -2-1 . In this adaption each of the MS2 spectra for a molecule is treated as an individual documents. All the MS2 documents for a sample become the cohort of documents. The individual m/z peaks and the neutral losses become the words and the relative abundance of the peaks the frequency of the word within that document. The neutral loss intensity is equal to the intensity of the fragment from which it was calculated. The topics are motifs of co-occurring m/z words. This has been successfully applied to single beer samples to extract information about chemical sub-structures that are common within the MS2 data for an

MS1 peaks. Figure -2-1 shows examples of the three substructures identified as motifs (and subsequently manually annotated).

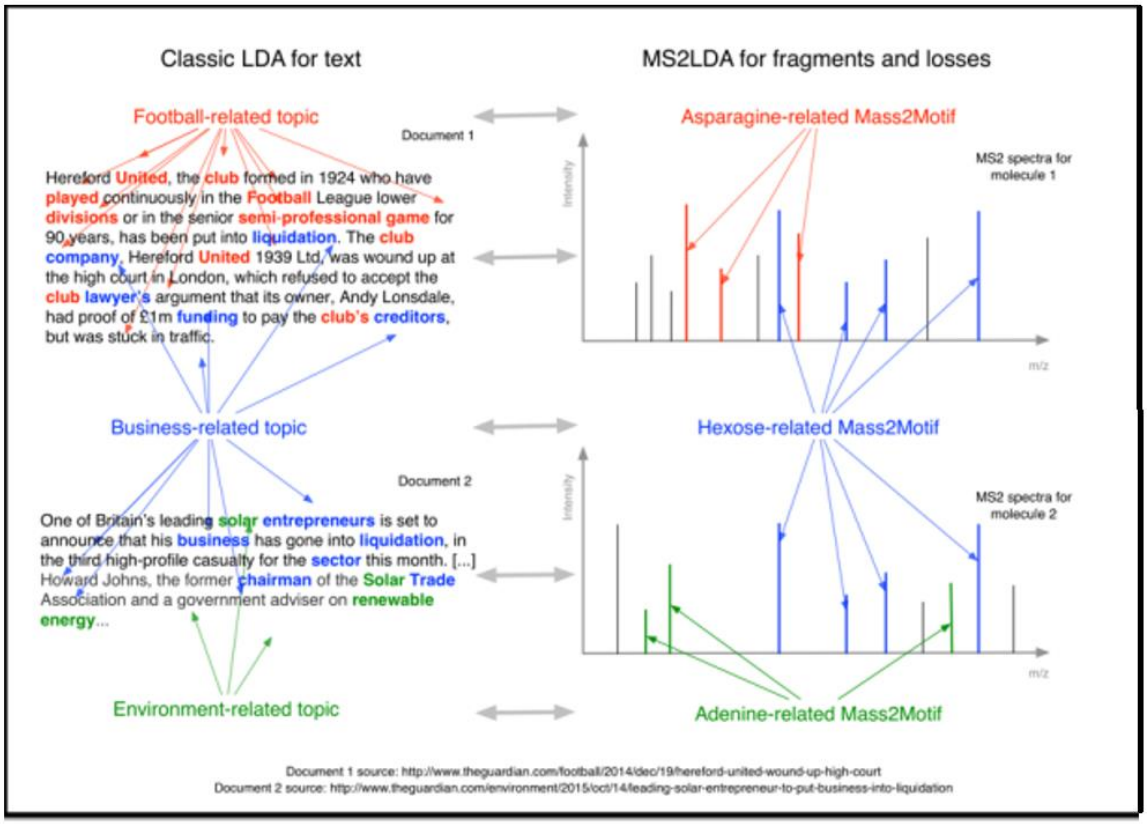


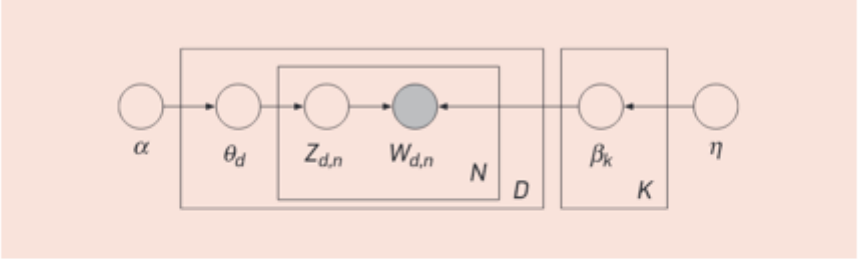
Figure -2-1. The analogy between LDA applied to text and MS2LDA applied to MS data. (van der Hooft, 2016)

Text LDA	MS2LDA
Cohort- a collection of documents	All MS2 spectra for an experiment
Document	A single MS2 spectra
Topic co-occurring words across cohort	Motif co-occurring peaks across MS2 spectra.
Word	Word - a fragment or loss mass
Word frequency	Word frequency- calculated from the MS2 peak intensity.

Table 2-1 Translation table for terms between standard textual LDA and the MS2LDA adaption

2.1 The Latent Dirichlet Algorithm

LDA is a probabilistic topic modeling algorithm that takes the observed the words within a collection of documents and learns the underlying thematic structure contained within this cohort. The learnt model is a set of parameters that represents each document as a mixture of these topics and each topic as a distribution over the word along with two Dirichlet parameters, **α** and **η** that determine these distributions. This can be done in a completely unsupervised way with no prior knowledge of the topics. In effect it is a dimensionality reduction tool, reducing massive cohorts of documents to a distribution over a relatively small number of motifs. These parameters can then be used for clustering, classification and information retrieval.



Graphical model of LDA. Each node is a random variable. The hidden variables are unshaded and the observed variables – the words of a document – are shaded. The rectangles are "plate" notation which donates replication i.e. there are K topics β. The N plate denotes the collection of words within the documents, the D plate denotes the collection of documents within the collection. (Blei 2012)

Figure 2-2 Graphical Representation of the LDA model.

LDA can be represented by the graph, as shown in **Error! Reference source not found.**. It is a generative model and assumes the documents have been generated in the following way. In MS2LDA a document is a MS2 spectra for a single molecule and a topic is a motif of MS2 fragments and losses, see Table 2-1 for the translation of the terms used between the two applications of LDA.

First we choose:

A. A topic mixture for a document, Θ_d, according to the Dirichlet distribution **α** over a set of K topics.

B. A word mixture for a topic, $\beta_{k,}$ according to the Dirichlet distribution **η** over the vocab.

Each word in a document is then generated in a two-step process.

1. Randomly choose a topic z_{dn} according to the multinomial distribution Θ_d over topics in A.
2. Randomly choose a word from the corresponding topics multinomial distribution $\beta_{k,}$ over vocab in B.

This will generate a joint probability distribution over the observed variables, the word and the latent or hidden variables Θ_d β_k and α. The goal of the algorithm is to backtrack from the observed words to uncover these latent variables. The LDA algorithm is uses Bayesian inference to compute the conditional distribution of the hidden variables given the observed variables.

2.2 Multi-file MS2LDA

To find biochemical explanation behind phenotypic differences, such as drug response or disease states we need to be able to analyse the differences in metabolite content of samples groups. In the context of LDA we need to compare the LDA parameters such as document to motif distribution and the Dirichlet parameters, alpha, between groups.

MS2LDA has been adapted to process multiple samples into one model. In this model a set of input files each containing the MS/MS data for one sample are processed together. This produces a similar model to before with a common set of motifs, distributions over words and document motif distributions for each MS2 document. Additionally all the documents for a sample are linked by a unique set of Dirichlet parameters, alpha, a vector length k with an α_k for each of the K topics These are the parameters that define the shape of the Dirichlet distribution from where the document to motif distributions are chosen and represent the distribution of motifs within the sample. As this model only generates one set of motifs with its associated word distribution the motifs are directly comparable across samples.

In this project I have focused on analysing the differences in the alphas parameters between samples .The alphas values are the parameters of the Dirichlet whose properties enable documents to be mixtures of motifs and are the parameters for the multinomial distribution

that determine the individual document to motif distributions . These are in effect the average motif proportions across all the MS2 documents for a sample and as such are a measure of the substructure content of the samples. Because MS2LDA clusters metabolites on substructures in common and thereby metabolites on a pathway it is hoped that the DE in motif content between samples will reveal new biochemical information about perturbed pathways.

3 The Exploration

The brief of this project was to investigate the nature of the Dirichlet parameters. At the start of this project the Multi-file MS2LDA program had only just been implemented and this was the first exploration of these values and whether by analysing the differences across samples would reveal biochemically useful insights. The main goal of this project became to investigate the use of the differences in alpha values as a way of measuring motif DE across samples and to build tools that could be used in the ongoing development of this pipeline.

This project had three different stages of exploration. The rational for the direction chosen at each each stage was driven by the results of the previous one. Hence I have presented the rational, methods and results stage by stage.

3.1 Software Engineering Challenges

This project has posed a number of technical challenges for me to overcome many due to the nature of the pipeline being at the leading-edge research stage and being in very active development.

- Understanding how LDA algorithm works.
- Understanding the LDA code written in python by Joe Wandy.
- Learning python as this the language used for the development of this pipeline.
- Learning Jupyter notebooks as a tool to develop and share code.

- Learning to use cluster and writing python scripts (Appendix C)to run large LDA analysis that took 8 days to run.
- Learning to use Git and GitHub, necessary to keep up to date with MS2LDA developments and to share my code.
- Learning to use Plotly to make interactive and sharable plots.

All code was developed in Python run from Jupyter notebooks. Reusable routines were developed in separate libraries imported into the notebooks as needed.

Copies of the notebooks are in Appendix A. The library code is Appendix B. These can also be found on the disk provided and at https://github.com/youngfran/Alpha_MS2LDA

3.2 Data

I have been using LDA models generated from various combinations beer and urine MS/MS data. The beer samples from three commercial beers and one home-brewed beer were used as representative complex mixtures of diverse biochemical including Beer was chosen for initial testing for MS2LDA because it contains a complex biochemical mix including amino acids, sugars and nucleic acids. The urine samples were from stroke patients all of who had complex drugs routines and confounding health issues. The patient data including age, sex, diagnose and drug treatment was available.

3.3 Initial Exploration of the Alphas

3.3.1 Difference in individual motif proportion

Initially I looked at the differences between the alphas for a model of four beer samples. This included both difference and fold change between the individual samples. Also looked at the effect of normalising the data a by using the expected values of alpha .Notbeook1.

The alphas can be thought of as an array of size S x K, S is the number of samples and K is the number of motifs. The following definitions

Difference $$\boldsymbol{DE_k = \alpha_{k,s1} - \alpha_{ks,2}}$$ *Equation 1*

Fold Change $$FC_k = log_2\alpha_{s1} - log_2\alpha_{s2}$$ *Equation 2*

Expected Alphas $$\alpha_{sk} = \alpha_{sk} / \sum_i^k \alpha_{si}$$ *Equation 3*

3.3.2 Separation of groups with Principal Component Analysis on Alphas

Having found a good variation in the alpha values across various motifs between similar samples I generated a model for a mixture of beer and urine samples to see look how these differences varied between two distinct groups of biochemically diverse samples. My rational was that we would know if the DE observed makes biochemical sense whereas looking at either beer or urine alone would result in lower DE and fewer motifs that we could definitively explain biochemically. Multi-file MS2LDA was run on a cluster from a python script. Appendix 3.

Hierarchical Clustering and Principal Component Analysis to see if the features separated the groups.

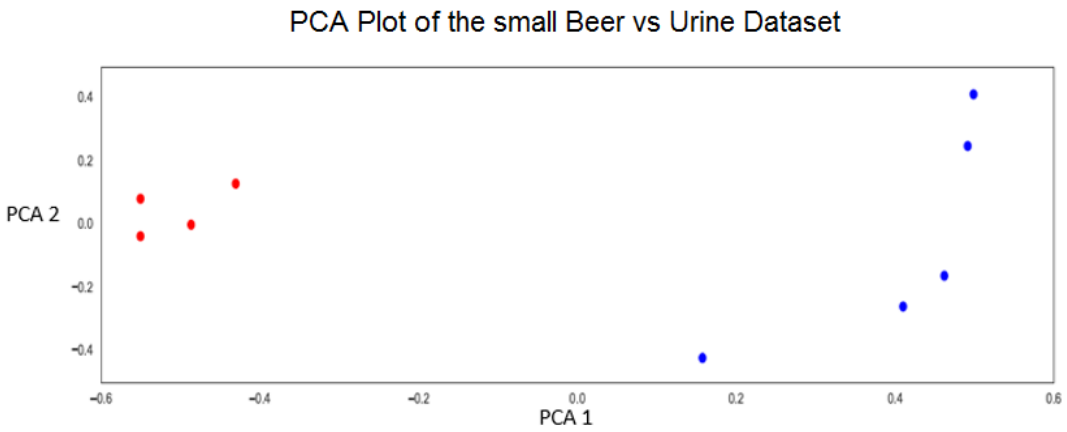


Figure 3-1 The PCA plot of the samples from the alpha values. • Beer • Urine .

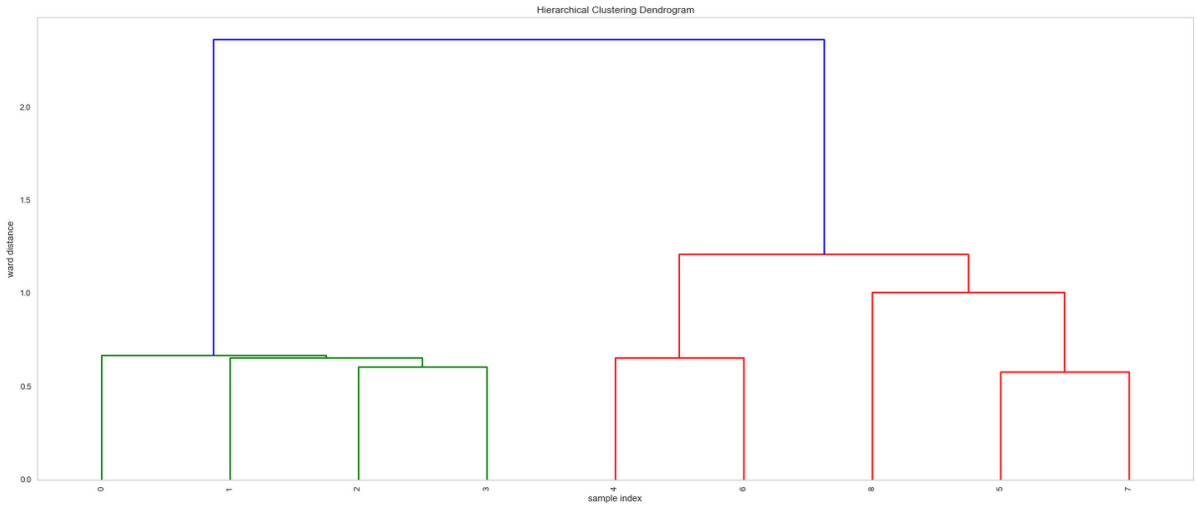


Figure 3-2 Dendrogram of the results of hierarchical clustering on the alphas by sample.

The initial PCA in Figure 3-1, shows that the alpha values can successfully be used to separates the samples into the two correct groups on PCA1. This was the first result the multi-file MS2LDA contained some features, the motifs, that are capable of picking up the differences between groups. His was backed up by a similar result from hierarchical clustering using 'ward' linkage The next step was to write code to analyse these differences.

3.4 Development of Tools to analysis the Differential Expression

Having demonstrated the ability of the alpha values to separate samples I developed statistical tools to analyse the differences in these values across sample groups. An example notebook is provided as a manual for the different routines.

3.4.1 Normalising Alphas

To more comparison between the alphas … The the raw alphas were normalised by using the expected alphas as defined in equation 3 above. The distribution of these were compared.

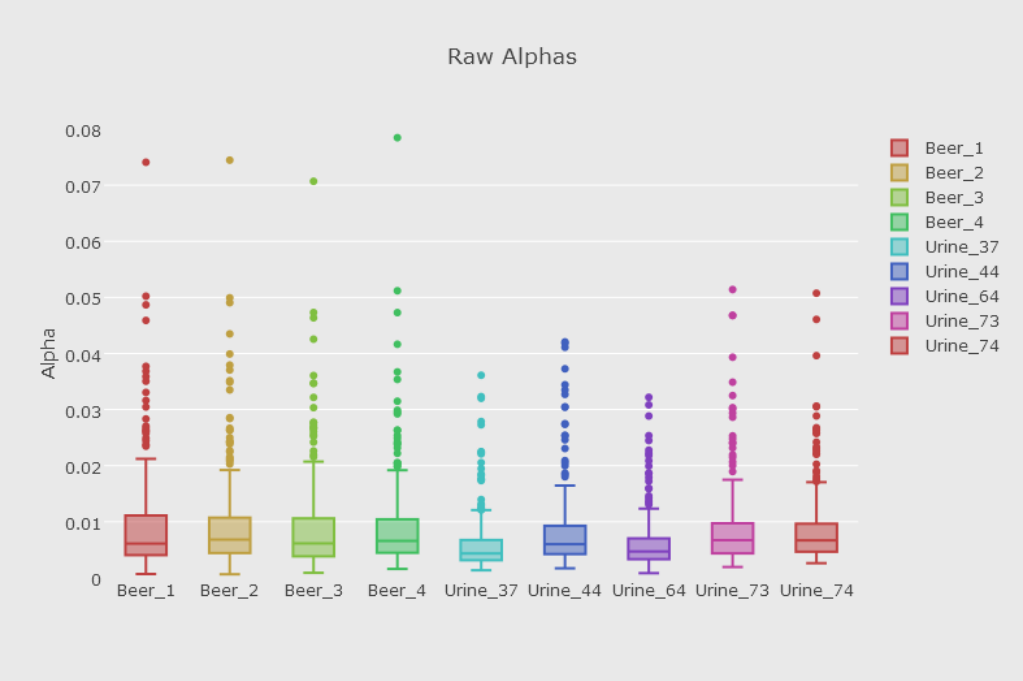


Figure 3-3 Boxplots of the distribution of the raw alphas within each sample.

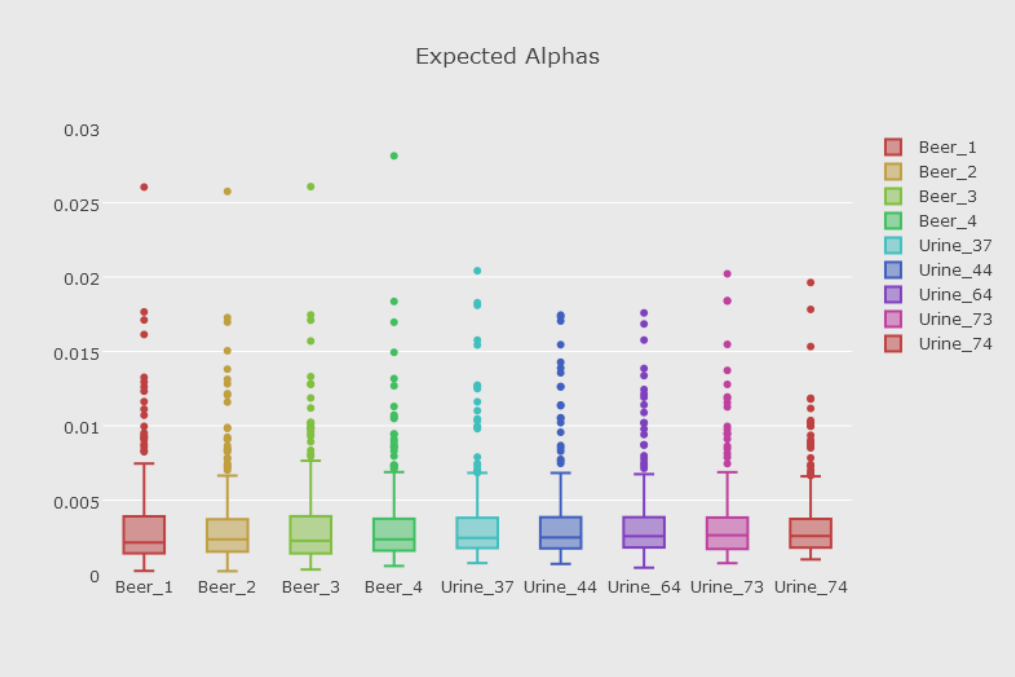


Figure 3-4 Boxplots of the distribution of the Expected alphas within each sample.

Comparing the distributions for raw alphas and expected alphas shows that using the expected values has the effect of normalising them across the files. This will help negate any difference caused by different numbers of MS2 documents in each sample. It also has the effect of smoothing the data some information will be lost. When loading the data you can choose to use expected or raw values.

3.4.2 Statistical analysis

The function alpha_stats in lda_alpha_stats.py takes the alphas and groupings as parameters and calculates the statistics below. The T-test and Mann Whitney test routines from SciPy open source libraries were used.

- Fold-change calculated as in Equation 2 above from the mean of each group.
- Parametric T-test. This assumes normal distributions with equal variance in the groups
- Mann Whitney non parametric test p-value. This should have the advantage that it is more robust to outliers.
- Benjamini-Hochberg p-value correction for multiple hypothesis testing.
- Display a table of the results

This was applied to a model from the large dataset of Beer-v-Urine and Table 3-1 shows the results sorted on T-test adjusted p-value. Interesting motifs can then be displayed using topic boxplots on the topic object. For example *Figure 3-9 Boxplots of the motifs most significantly up-regulated in Beer in the BvU-large dataset. F*or the code and associated notebook see Example Notebook ?.

	topic	log2FC	TT p-value	TT p_adjust	TT significant	MW p-value	MW p_adjust	MW significant
30	30	-3.079550	1.681395e-20	7.546097e-19	True	5.006010e-08	3.623123e-07	True
294	294	-2.580850	4.139424e-19	1.379808e-17	True	5.006010e-08	3.623123e-07	True
220	220	-2.484164	6.896266e-17	1.477771e-15	True	5.006010e-08	3.623123e-07	True
186	186	-2.066444	8.841274e-16	1.473546e-14	True	5.006010e-08	3.623123e-07	True
183	183	-1.909662	2.716803e-15	3.704731e-14	True	5.006010e-08	3.623123e-07	True
259	259	-1.524044	1.547488e-14	1.709347e-13	True	5.006010e-08	3.623123e-07	True
73	73	-1.242342	2.531677e-17	6.329194e-16	True	5.006010e-08	3.623123e-07	True
106	106	-1.095307	3.091287e-14	2.991568e-13	True	5.006010e-08	3.623123e-07	True
225	225	-3.497854	1.728480e-12	1.058253e-11	True	6.708507e-08	3.946181e-07	True

Table 3-1 Table 3-1

At present this is a very basic set of statistics but can be developed and added to as more functionality or refinement is required.

3.4.3 Interactive annotated PCA Plot

Developed an interactive PCA that shows both the samples coloured by group and the motifs as the loadings in the same two PCA dimensions, PCA1, PCA2. See. Notebook 2 and Method Plot_PCA in the library LDA_alphas_model.py

This method allows you to

- Add the motif's words as hover text for the motif.
- Set groups for colouring samples
    - Either from group names found in sample files. e.g. [' Beer','Urine']
    - Or from grouping info from a metadata file. E.g. patients taking/not taking a drug.
    - Set the names of these groups.
- Set clusters for colouring motifs.
    - From up/down regulated motifs
    - Results of hierarchical clustering
    - Motifs of interest e.g. from annotation information.

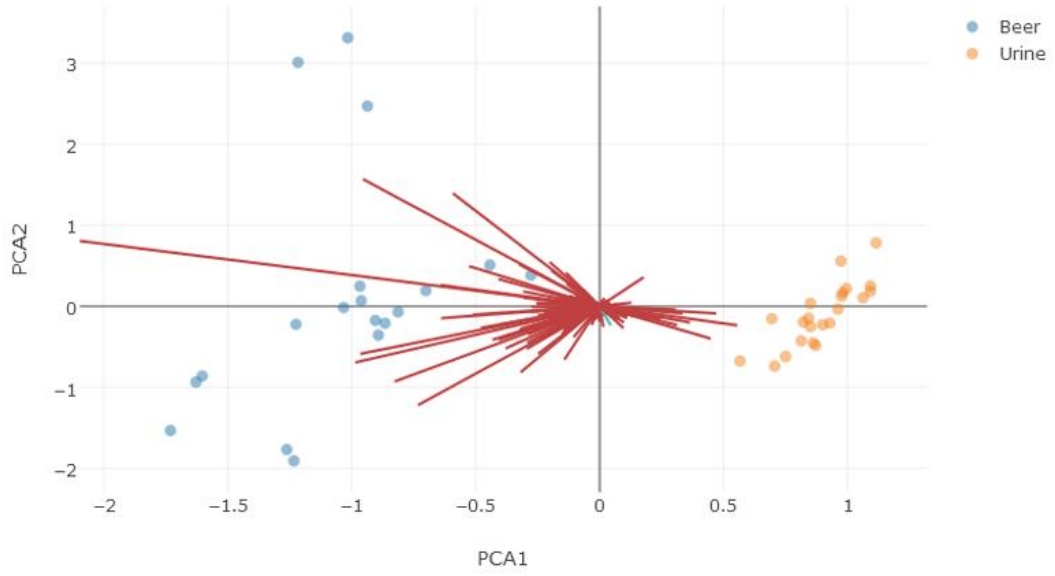


Figure 3-5 PCA of the Beer versus Urine – Large dataset.

For a fully interactive version see https://plot.ly/~fran123/222/pca-of-beervurine-large-dataset/

3.4.4 Annotation

To allow useful interpretation of the model and to check its validity the motifs need to be annotated with putative sub- structures. This has being done manually by a chemist (Justin Van Der Hooft) . A reference table of commonly seen motifs was being built up from previous models. Appendix

I developed code to automatically find hits from a model in this reference table. See Notebook Annotation and associated library lda_annotate_topic.py . This creates an object that is a list annotated topics and the words (annotated with chemical formula) in those topics which the individual words from the model can be matched against. This is done within a matching tolerance that the user can set in parts per million, ppm. Each hit is scored which is calculated as the sum of the word probabilities that match a word from the annotation.

Hits for motifs of interest can then be displayed in the notebook or a file of all annotations can be written to a CSV file see Appendix 4 . This can then be used to choose the most likely annotations.

- Plots below show a matching threshold of 20 ppm for fragments (double for loses) will pick up all the matches,
- A Score threshold of over 0.2 will remove the bulk of the low scoring hits making it easier to manually search for meaningful ones
- Scoring very basic for now, need to investigate more sophisticated methods that take account of low intensity peaks. Also probability of word in a motifs is directly related to the number of words in in the topic (as it is a distribution of words) this means single word motifs get very high scores when they match an annotation. Are they more valid than a match of lots of words but with low probability?

From the following plots in Figure 3-6 and Figure 3-7, a word matching tolerance of 20ppm and a scoring threshold of 0.3 were set as the default parameters.

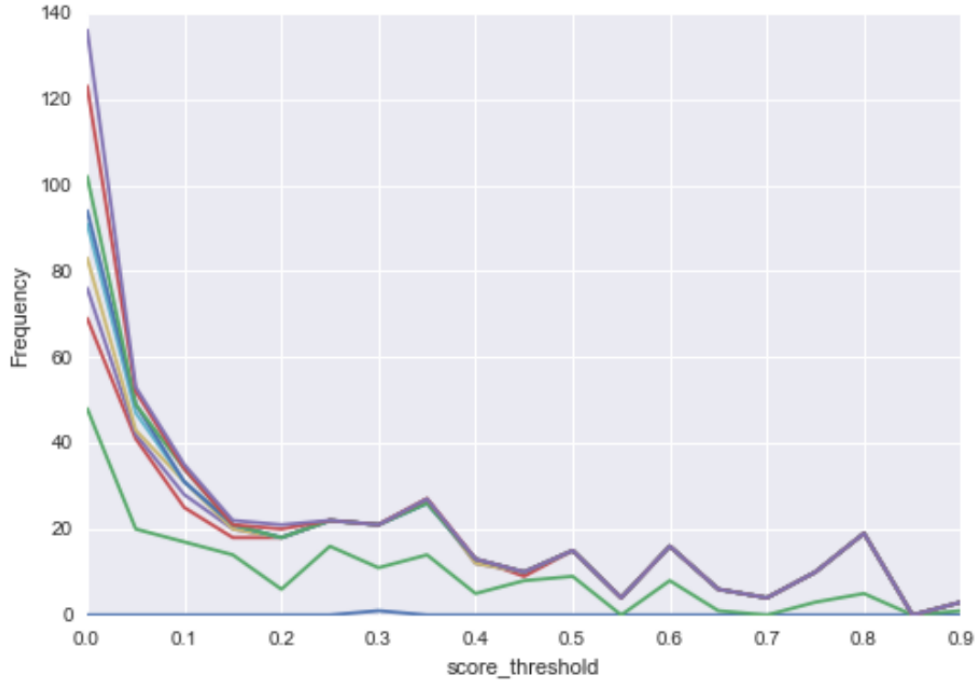


Figure 3-6 Plot of the frequency of annotation hits against score threshold, each plot is for a different ppm ranging from 0 to 50 .

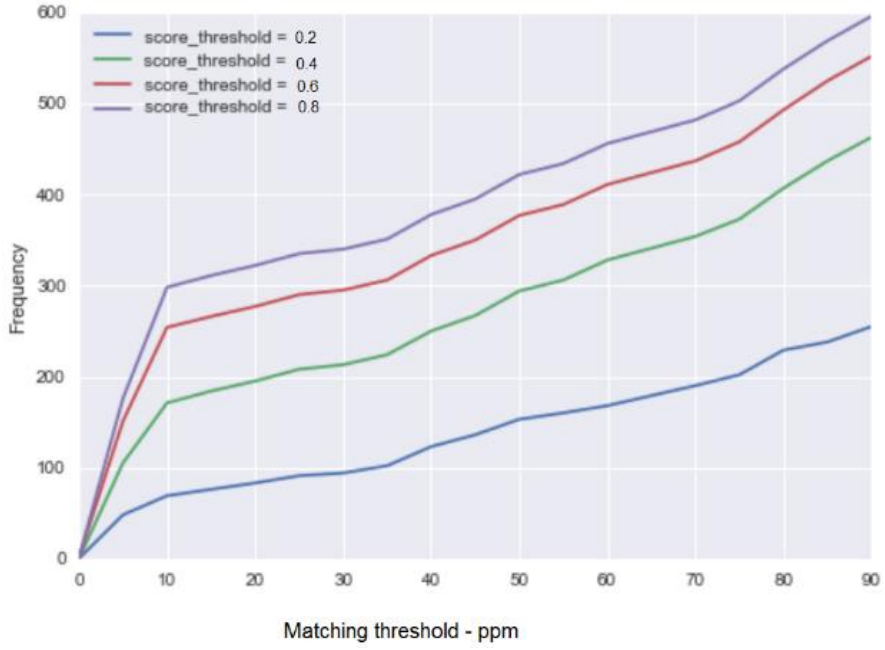


Figure 3-7 Plot of cumulative frequency of annotation matches against the ppm word matching threshold.

Below is an example of the annotation information displayed in the notebook . The U_ and B_ indicate which annotation table the annotation comes from.

```
 Topic  217  hits 2
 Model Words: ['Frag_110.06009', 'Frag_152.07019', 'Loss_79.95625', 'Frag_232.02748', 'Frag_65.
03878', 'Loss_96.98348', 'Loss_121.96743', 'Loss_138.99373', 'Frag_249.05499', 'Frag_92.0494',
'Frag_111.04408']

 Score: 0.390748502715 number of words matching 3 [0, 4, 9]
U_Mass2Motif related to Paracetamol ? many ion products associated to this topic (abundant spec
ies ? two real paracetamol metabolites): Frag 110.060588884 (C6H8NO),Frag 65.0391251607 (C5H5),
Frag 92.0500241981 (C6H6N),

 Score: 0.330408592525 number of words matching 2 [1, 18]
B_Unclear what these fragments relate to.: Frag 152.071153571 (C8H10NO2),Frag 134.060588884 (C8
H8NO),
```

Figure 3-8 Example of a motif annotation

3.5 Validation

The final stage in this project was to use the tools developed to analyse two models and to check the validity of these tools.

The code generated above is imported into notebooks to process two models.

3.5.1 Beer versus Urine Large Dataset

This BvU model was generated from 19 beer samples and 22 Urine samples. Again the rationale behind using mixed samples was that it should be easier to establish that the alpha DE could be explained biochemically or not as some substructures will be more prevalent or absent in one group or the other.

Below are the results of the statistical tests showing the most significant DE motifs (Mann Whitney)

	topic	log2FC	TT p-value	TT p_adjust	TT significant	MW p-value	MW p_adjust	MW significant
30	30	-3.079550	1.681395e-20	7.546097e-19	True	5.006010e-08	3.623123e-07	True
294	294	-2.580850	4.139424e-19	1.379808e-17	True	5.006010e-08	3.623123e-07	True
220	220	-2.484164	6.896266e-17	1.477771e-15	True	5.006010e-08	3.623123e-07	True
186	186	-2.066444	8.841274e-16	1.473546e-14	True	5.006010e-08	3.623123e-07	True
183	183	-1.909662	2.716803e-15	3.704731e-14	True	5.006010e-08	3.623123e-07	True
259	259	-1.524044	1.547488e-14	1.709347e-13	True	5.006010e-08	3.623123e-07	True
73	73	-1.242342	2.531677e-17	6.329194e-16	True	5.006010e-08	3.623123e-07	True
106	106	-1.095307	3.091287e-14	2.991568e-13	True	5.006010e-08	3.623123e-07	True
225	225	-3.497854	1.728480e-12	1.058253e-11	True	6.708507e-08	3.946181e-07	True

Table 3-2 Table of the top most significant motifs of the BvU large dataset.

The following boxplots are for the most up and down regulated motifs.Figure 3-9Figure 3-10

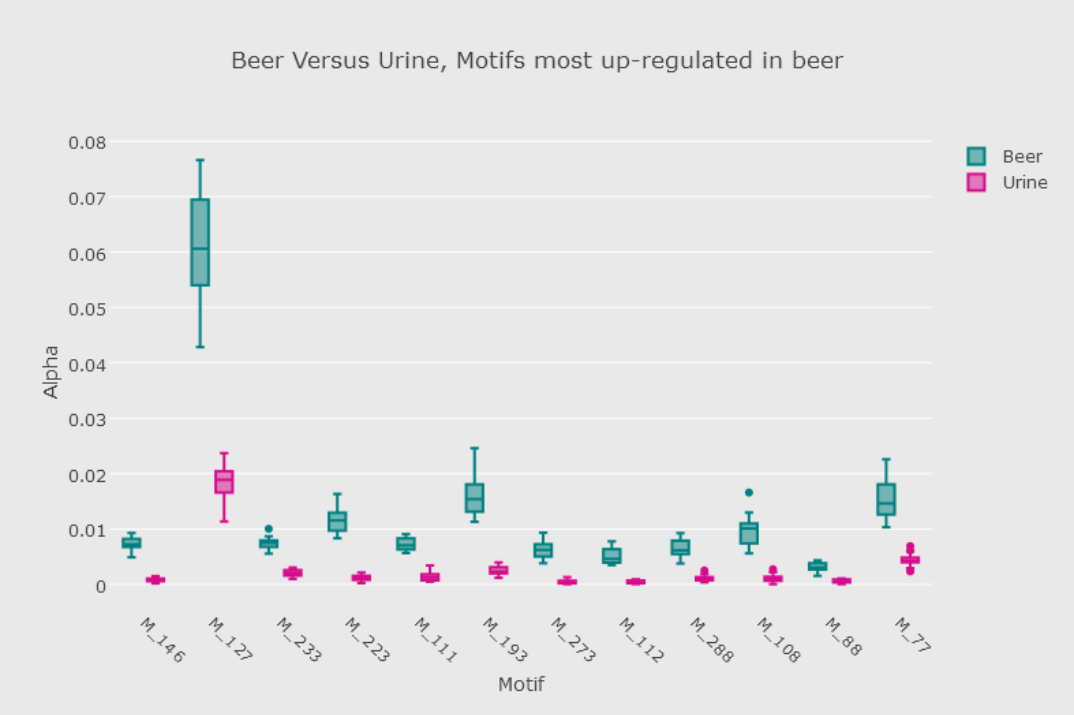


Figure 3-9 Boxplots of the motifs most significantly up-regulated in Beer in the BvU-large dataset.

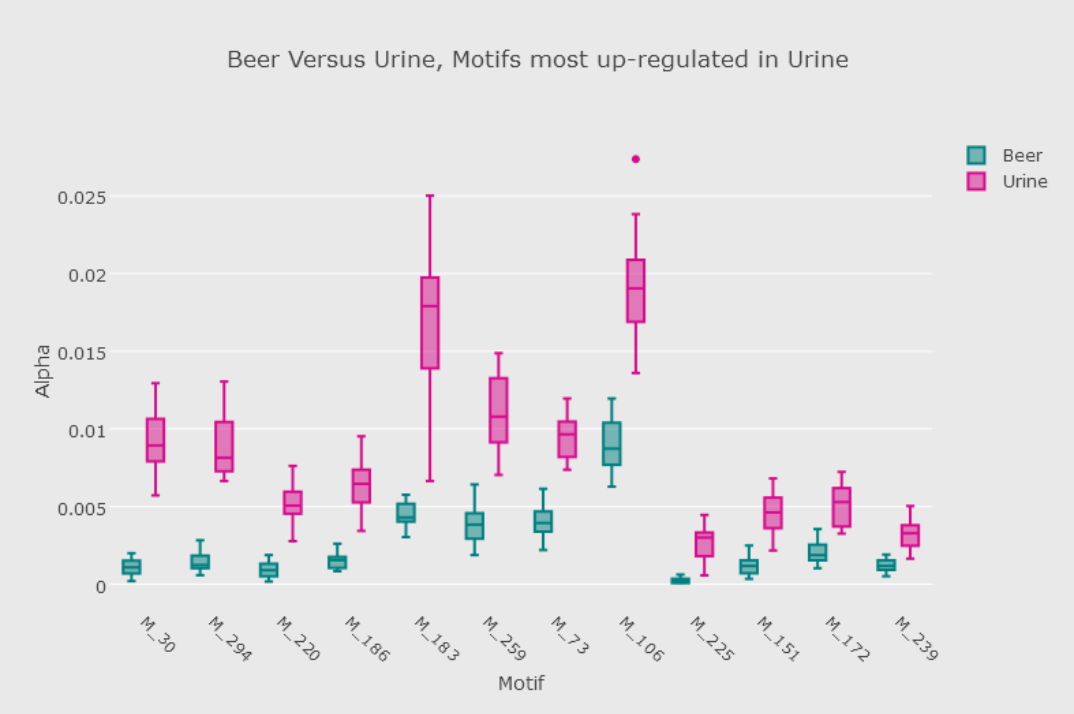


Figure 3-10 Boxplots of the motifs most significantly up-regulated in Urine in the BvU-large dataset.

3.5.1.1 Motif annotation with putative substructures.

The PCA plot was used (without the results of the statistical analysis) by a chemist (Justin van der Hooft) to get confirmation that the distribution of the motifs made biochemical sense. By checking the word content (MS2 fragment spectra) he identified several motifs that were loaded (i.e. pointed) towards either beer or urine that should be more highly expressed in one or other of the groups.

Motif	TT p_adjust	Annotation	Comment
223	5.69E-20	loss of [hexose-H2O]	Beer contains many more glycosylated compounds than urine
191	4.15E-13	hexose fragments	As above.
77	7.46E-15	proline	Amino acid that is abundant in many beers
267	1.75E-11	phenylalanine	Amino acid that is abundant in many beers
162	7.81E-05	histidine	Beer contains many more histidine conjugates
198	1.59E-06	?	Points towards the only 3 Belgium beers, They could be related to the hop metabolites

217	0.257063	paracetamol related	Not to be expected in beers!
126	0.000675	glucuronidation	Common biotransformation in humans, not so much in plants
26	5.28E-07	acyl carnitines	Not to be expected in beers, and present in all urines
73	6.33E-16	acetylation	Common biotransformation in humans
106	2.99E-13	trimethylamine	Betaine related fragment quite a few human metabolites have this methylated amine substructure
279	3.18E-07	trimethylamine oxide	See motif_106, related substructure
259	1.71E-13	valerolactam substructure	Likely formed from acetylamine substructure present in acetylated spermine/spermidine and other related amines ? expected in urines
294	1.38E-17	creatinine	A common metabolite in urine ? not to be expected in beers

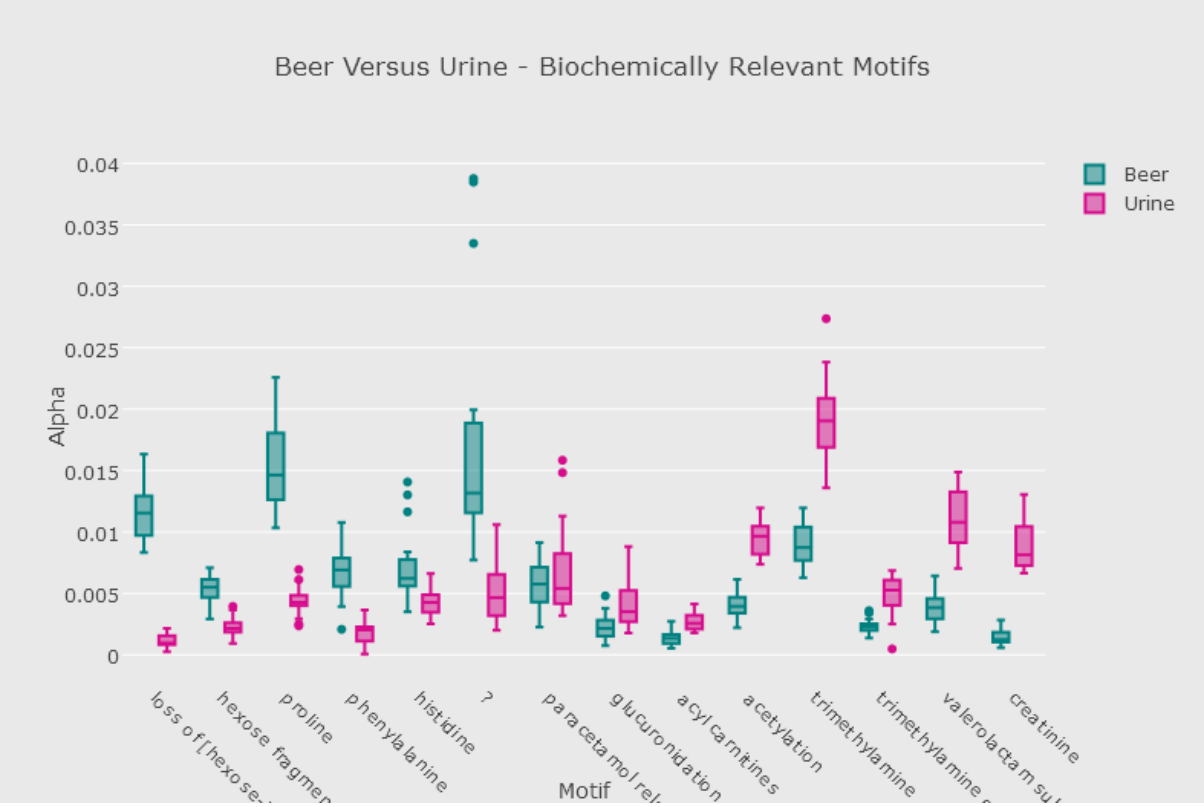


Figure 3-11 Boxplots of Biochemically Relevant Motifs with Putative Annotations

Motif 217 which has been annotated as paracetamol by Justin has a p-value of 0.26 and so is not significantly different as can be seen from the plot. This motif has been matched to paracetamol in the urine annotation table and to a beer annotation of unknown origin with the word fragment 152.07. Two more motifs were matched to paracetamol related substructures with motif 14 and motif 229.

MOTIF	P-VALUE	LOG_2FC (B/U)	SCORE	WORD HITS
217	0.26	- 0.25	0.33	3 (/3 anno words)
14	0.001	0.65	0.79	7 (/7 anno words)
229	0.009	0.319	0.82	3 (/3 anno words)

Table 3-3 The motifs annotated with paracetamol related substructures.

3.5.2 Urine large Data set

This model was generated from 22 Urine samples. Patient data was included and information from this was used to group the data in various ways to see if any of the motifs were significantly DE.

Table 3-4 Table of the Motifs that are significant (MW p-value) on Statin/non Statin grouping.Table 3-4 shows the motifs that were significant when grouping the samples on whether the patient was on Statins. Figure 3-9 Boxplots of the motifs most significantly up-regulated in Beer in the BvU-large dataset.Figure 3-9 shows the boxplots for these motifs and finally shows a section of the PCA where the motifs significant on gender are coloured in green and the motifs significant on use of Statins are coloured blue. It shows that the loadings for statin significant motifs are much smaller than those for gender.

	topic	log2FC	TT p-value	TT p_adjust	TT significant	MW p-value	MW p_adjust	MW significant
279	279	1.207274	0.006223	0.933450	False	0.004796	1.0	False
118	118	-0.613863	0.001998	0.599464	False	0.009726	1.0	False
138	138	1.468946	0.016060	0.996560	False	0.012174	1.0	False
295	295	0.900391	0.028053	0.996560	False	0.018754	1.0	False
71	71	0.461335	0.022787	0.996560	False	0.028258	1.0	False
108	108	0.781870	0.111802	0.996560	False	0.028258	1.0	False
280	280	0.469529	0.035758	0.996560	False	0.034401	1.0	False
81	81	0.999406	0.047081	0.996560	False	0.034401	1.0	False
123	123	0.336054	0.031875	0.996560	False	0.041650	1.0	False

Table 3-4 Table of the Motifs that are significant (MW p-value) on Statin/non Statin grouping.

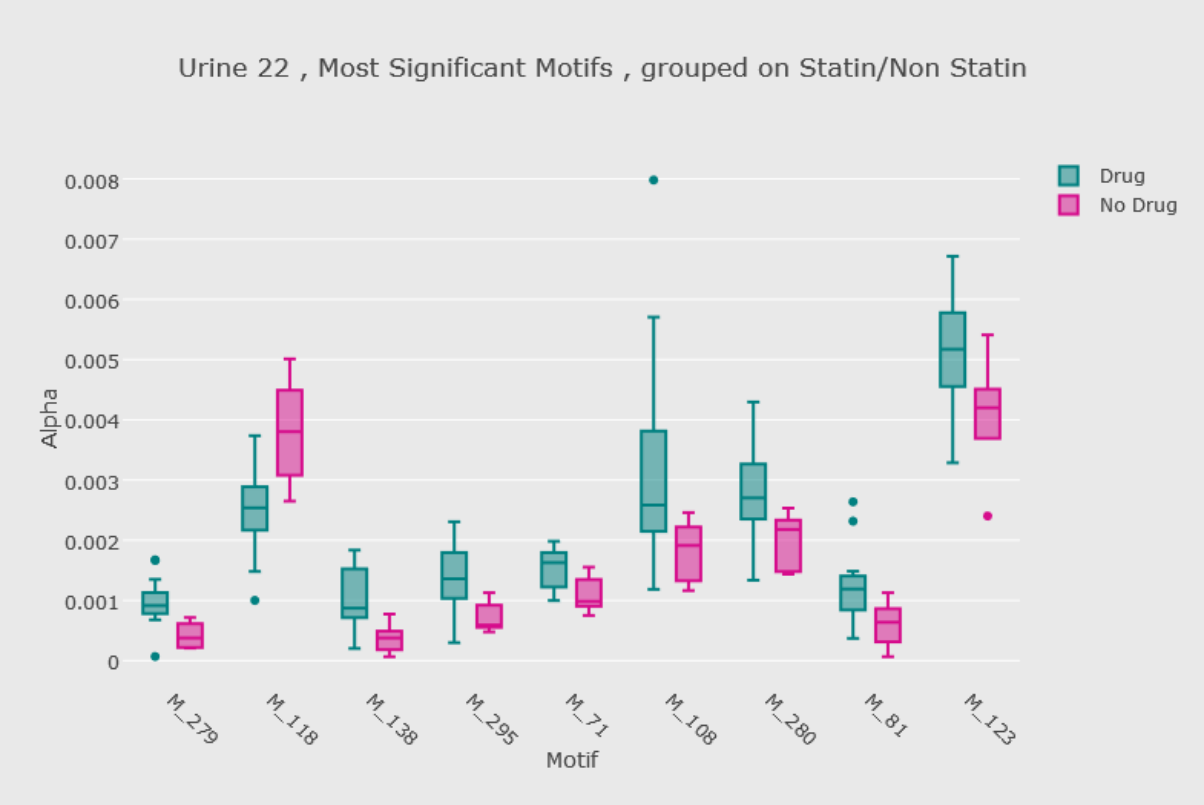


Figure 3-12 Box plots of the significantly regulated motifs on statin/non-statin use

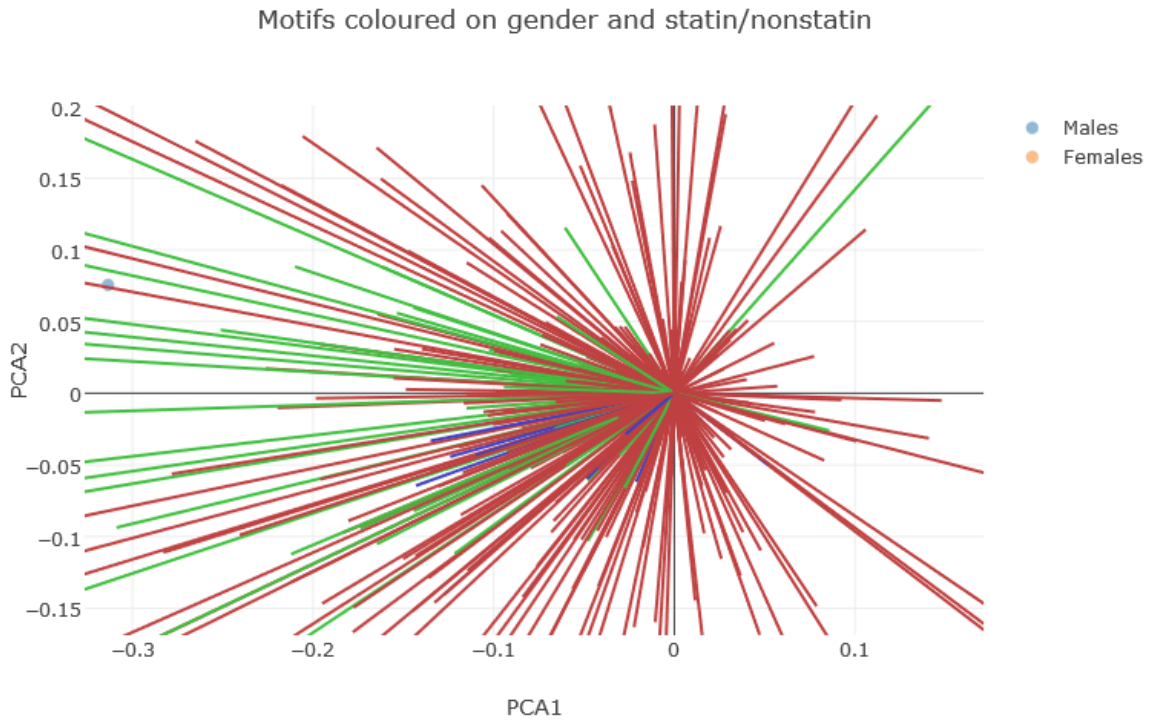


Figure 3-13 PCA plot of urine22 data, zoomed in to show the motifs in blue that are significant when patients grouped on Statin/non-Statin. The green motifs are significant on gender. A fully interactive version of this plot can be found at https://plot.ly/~fran123/224/motifs-coloured-on-gender-and-statinnonstatin. /

4 Discussion

The aim of this project was to make the initial investigations into the models generated by the multi-file MS2LDA pipeline that is currently in development. The main focus was to explore how new parameters generated in these models could be used to make substructure comparisons across sample groups, and whether this DE could be used to make biochemically informative inferences.

During the initial exploratory stage I used PCA to show that the alphas can separate the sample groups this implies that the total substructure content of the samples can reveal information about the groups. By generating a model of widely diverse biochemical samples we were able to establish that

The development of tools to analyse the DE of motifs made identifying motifs of interest a straight forward process. The time taken from loading a model to obtaining a list of candidate motifs for further investigation being a matter of minutes. The information is displayed to facilitate their analysis. This includes the results of statistical tests in table form and boxplots to show the distribution of the alpha values within the groups. The annotation facility allows you to easily check the meaning of a motif in terms of substructure. If an annotation is not available the fragments and losses displayed can be used to search databases such as MassBank (Horai et al., 2010)

The interactive PCA tool gives an alternative way of investigating the DE of motifs. The loadings of the motifs that contribute to the distribution of the samples in the PCA dimensions are an alternative measure of the differences in motif content of the samples. Again this motif content is directly related to the substructure makeup of the samples. This has been used by Justin find biochemically relevant motifs.

The results from using these tools to analyse larger datasets validated that they were able to identify biochemically relevant motifs. Both the statistical tools and the interactive PCA were independently able to pick out motifs that were DE across the groups that made biochemical sense when annotated for the beer versus urine model. When applied to the urine model motifs with significant DE were identified the statistical significance, as indicated by the p-values, of the DE was much lower than in the Beer and urine model as would be expected . Further work needs carried out to annotate these motifs to see if any of these DE substructure makes sense.

5 Conclusion

In this project I set out to investigate the models produced by the newly developed multi-file MS2LDA pipeline and to develop tools to investigate the differences between samples. Through various stages of this exploration I was able to demonstrate that the use of the alpha values as a measure of the substructure content of the samples allowed the differential expression (DE) of the motifs to be analysed. By looking for significantly regulated motifs we are able extract biochemically useful information that is linked to phenotype.

The tools developed will up the the analysis of the models and the annotation matching tool successfully automates the initial steps of motif annotation with subst. As the database of substructure annotations grows this initial annotation will improve and allow us to focus on less common or more interesting substructures. The interactive PCA has been used to generate results and along with a boxplot have already been used in a presentation at meeting.

This has resulted in a clearer picture of the strengths of the MS2LDA method to mine biochemically insightful information and ideas for future research. I have given clear evidence that the multi-file MS2lda pipeline is able to produce biochemically meaningful results. This validates the tools developed in this project as well as providing further validation of the MS2LDA approach.

Further Work

This code has been developed as part of an active research project and as such continual improvements and additions will be required. Integrating the tools developed here into the established MS2LDA toolset would be the first step.

There are also several areas where these tools would benefit from improvement.

- The interactive PCA plot could be improved with further hover text such as patient information such as sex, age and diagnosis. If the motifs have been coloured by the different groupings used to test for significance a legend for this should be included.
- The annotation table should be continually developed with the addition of commonly observed motifs and refinement of those already in it. The scoring system will need development so the choice in annotation could become unsupervised. This may including a measure of number of words matched so that even very low probability words contribute to the score.

Further investigation into the use of the alpha values to extract useful information from the model is also required. The use of clustering to group motifs would be an obvious next step in this investigation and would reveal information about which substructures and therefor potentially pathways that are similarly affected by perturbation.

6 References

Justin J.J. van der Hooft, Joe Wandy, Michael P. Barrett, Karl E.V. Burgess, and Simon Rogers (2016)
Topic Modeling for Untargeted Substructure Exploration in Metabolomics,
Submitted for publication

Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., Ojima, Y., Tanaka, K., Tanaka, S., Aoshima, K., et al. (2010). MassBank: a public repository for sharing mass spectral data for life sciences. J. Mass Spectrom. JMS *45*, 703–714.

Sakthivel, N.R., Nair, B.B., Elangovan, M., Sugumaran, V., and Saravanmurugan, S. (2014). Comparison of dimensionality reduction techniques for the fault diagnosis of mono block centrifugal pump using vibration signals. Eng. Sci. Technol. Int. J. *17*, 30–38.

Silva, R.R. da, Dorrestein, P.C., and Quinn, R.A. (2015). Illuminating the dark matter in metabolomics. Proc. Natl. Acad. Sci. *112*, 12549–12550.

Treutler, H., Tsugawa, H., Porzel, A., Gorzolka, K., Tissier, A., Neumann, S., and Balcke, G.U. (2016). Discovering Regulated Metabolite Families in Untargeted Metabolomics Studies. Anal. Chem. *88*, 8082–8090.