University of Glasgow | School of Computing Science

Porting of a Mass2LDA tool and visualiser into FrAnK (Fragment Annotation Toolkit)

Claire Thompson, 2027347t

School of Computing Science
Sir Alwyn Williams Building
University of Glasgow
G12 8QQ

Level 4 Project — March 25, 2016

Abstract

Glasgow Polyomics is a research facility at the University of Glasgow specialising in the generation and analysis of large scale biological datasets. One of the 'omics' areas researched by Glasgow Polyomics is metabolomics, the study of the set of metabolites present within an organism. To aid this work a toolkit - Polyomics Metabolomics Pipeline (PiMP) - was created in order to assist researchers in the analysis of metabolomic data. A web application, FrAnK (Fragment Annotation Kit), was also developed as a project within PiMP for the annotation of mass spectral peaks using fragmentation spectra. Independently of FrAnK a standalone tool, Mass2LDA, was developed for the analysis of fragmentation data by finding substructure motifs within datasets.The aim of this project is to port the Mass2LDA functionality and selected visualisation features into the FrAnK framework.

Education Use Consent

I hereby give my permission for this project to be shown to other University of Glasgow students and to be distributed in an electronic format. **Please note that you are under no obligation to sign this declaration, but doing so would help future students.**

Name: _____ Signature: _____

Contents

Chapter 1

Introduction and Background

1.1 Introduction

Within the field of metabolomics, mass spectrometry analysis is regularly used to quantify and identify the small molecule composition of organisms. This process, however, produces large complex datasets that can be difficult to analyse. In order to aid this process, software toolkits can be developed in order to analyse and identify these small molecules, called metabolites, within these data sets. Glasgow Polyomics has recently developed an application, 'PiMP' (polyomics metabolomic pipeline) to serve this purpose. Within pimp, a web application, 'FrAnK' (Fragment Annotation Kit) was developed to aid researchers in the identification of metabolites using fragmentation patterns generated from mass spectrometry analysis. The design of FrAnK allows for easy extensibility by plugging in different 'annotation tools' where annotation refers to the process of identifying molecules from fragmentation spectra. Independently of FrAnK, a toolkit - Mass2LDA - was developed for the analysis of fragmentation data by finding substructure motifs within datasets. The aim of this project is to port the Mass2LDA functionality and selected visualisation features into the FrAnK framework as a new annotation tool. Before details of the project are discussed, the relevant background information will be covered.

1.2 Glasgow Polyomics

Glasgow Polyomics is a research facility at the University of Glasgow specialising in the generation and analysis of large scale biological datasets. Within the group are experts in many of the 'omics' disciplines including genomics (the study of DNA and genetic information), transcriptomics (the study of RNA), protemoics (large scale study of proteomes) and metabolomics (the study of the set of metabolites present within an organism). Glasgow Polyomics is also supported by software engineers and bioinformaticians to produce software for the analysis of datasets within these disciplines[2].

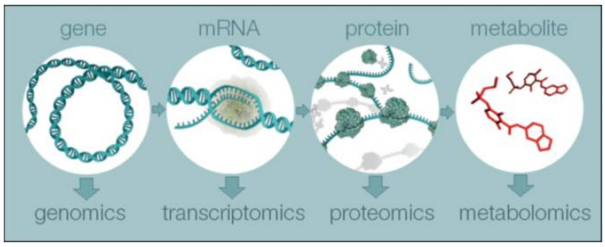


Figure 1.1: Overview of the four major 'omics' fields [4].

The software developed by Glasgow Polyomics concerning this project is the PiMP/ FrAnK pipeline which lies within the discipline of metabolomics.

1.3 Metabolomics

Metabolomics is an emerging field concerning the study of small molecules within organisms. The complete set of these small molecules, which are the intermediates and products of metabolism, collectively make up the metabolome[5]. Examples of metabolites include gluscose in the metabolism of sugars and amino acids in the biosynthesis of proteins. Analysing these small molecules can provide information regarding the entire physiology of an organism and has many applications in the areas of medicine (e.g. the study of disease) and in agricultural studies (e.g.the development of new pesticides). The best results for metabolomic studies involve the identification of the masses of these molecules, typically measured using analytical techniques such as mass spectrometry (MS).

1.3.1 Mass spectrometry (MS)

Mass spectrometry is an analytical technique used to separate components of a sample and determine the molecular mass of those components. The process of mass spectrometry involves the following components – a sample inlet to inject the sample to be analysed (stage1), an ionisation source to produce ions from the sample (stage 2), a mass analyser to separate these ions (stage 3), a detector to record the ions passing through (stage 4) and a data system to produce the mass spectrum (stage 5) [8].

Figure 1.2: Mass Spectrometry Process

At stage 2, the sample is bombarded with a stream of electrons causing the molecules to form positively charged ions (known as molecular ions). These ions will be relatively unstable resulting in the breakdown of the ions into smaller pieces through a process called fragmentation. It is these fragments that will then go on to appear as a line on the mass spectrum. Because there are many ways the molecular ion can be fragmented, the spectrum produced will display a wide range of detected masses as shown below.

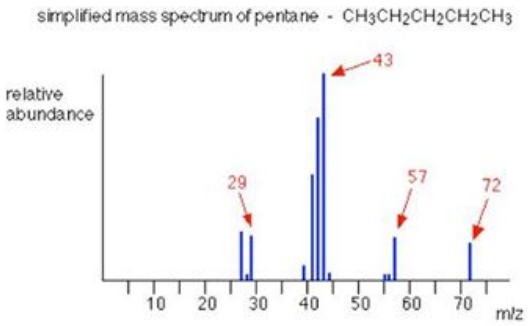


Figure 1.3: Mass Spectrum example

1.3.2 Tandem MS

Tandem Mass Spectrometry (MS/MS) involves multiple stages of mass spectrometry analysis with fragmentation occurring in between the stages. MS/MS can capture greater detail regarding the structure of metabolites that cannot be captured by a single stage of MS analysis and is therefore regularly used in metabolomic studies[7]. After the first run of mass spectrometry analysis, selected MS1 peaks are broken down further through the process of fragmentation. These resulting fragments are then separated and are detected in a second stage of mass spectrometry analysis (MS2). An MS1 spectrum therefore refers to the peaks detected from the first mass spectrometry analysis and an MS2 spectrum refer to the peaks detected from the fragmentation of a selected MS1 peak.

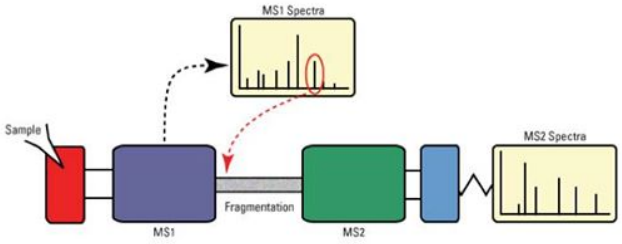


Figure 1.4: Tandem MS

1.4 Software for metabolomic analysis

1.4.1 PiMP/ FrAnK

MS-metabolomic experiments typically generate large quantities of data which can be difficult and time consuming to analyse manually. Software toolkits can therefore be developed to help the analysis of these datasets and ultimately, in the case of metabolomics, help in the process of quanitifying and identifying metabolites.

PiMP is an integrated, web-enabled application developed by Glasgow Polyomics that provides tools to standardise and automate the analysis of metabolomic data. Running within the PiMP environment, FrAnK (Fragment Annotation Kit) was developed in order to aid researchers in identifying small biological metabolites using fragmentation patterns generated from mass spectrometry analysis. Currently within FrAnK several annotation tools exist including MassBank, NIST, LCMS DDA Network Sampler and Precursor Mass Filter which all aim to annotate (identify) metabolites within mass spectrometry data. For example, the MassBank annotation tool takes MS2 spectrum data and sends this as a query to MassBank, an online data base containing mass spectral data[3].

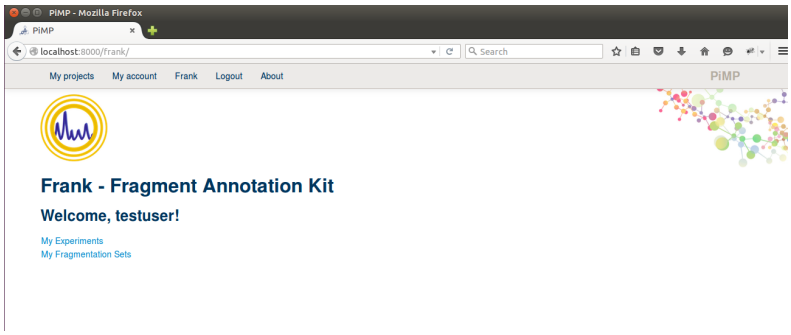


Figure 1.5: FrAnK main screen

Independently of FrAnK, a standalone application, Mass2LDA was developed that finds substructure motifs (recurring patterns) within fragmentation data to aid in the analysis of fragmentation data. This project will concentrate on porting the Mass2LDA application and visualisation into FrAnK as a new annotation tool.

1.4.2 Mass2LDA Tool

Topic Modelling

Topic models represent a class of computer programs that automatically extracts topics from texts with the purpose of discovering the hidden structure in large archives of documents. Topic modelling algorithms can be applied to many different types of data and have previously been applied to various areas including social networks, images and genetic data [6]. In Mass2LDA, a topic modelling algorithm, specifically Latent Dirchlet Allocation (LDA) is applied to mass spectrometry data.

LDA

LDA (Latent Dirichlet Allocation) is a case of topic modelling proposed by Blei, Ng and Jordan in 2003 [9]. LDA is based on the assumption that documents are made up of multiple topics and these topics generate words based on their probability distribution.

In more detail, LDA assumes that a document is generated using the following steps:

- The number of words is determined in a document (e.g. 10 words).
- The mixture of topics within that document is determined (e.g. the document consists of 1/2 topic 'animals' and 1/2 topic 'fruit').
- Based on the probability distribution of each topic, the word slots within the document are populated (to clarify, words within topics have different probabilities of appearing. For example, the 'animals' topic may contain the words 'cat', and 'horse' with 20% and 15% probabilities respectively and account for 1/2 of the words in the document. The document will be populated according to these probabilities).

LDA then backtracks based on this assumption to calculate which topics initially created these documents.

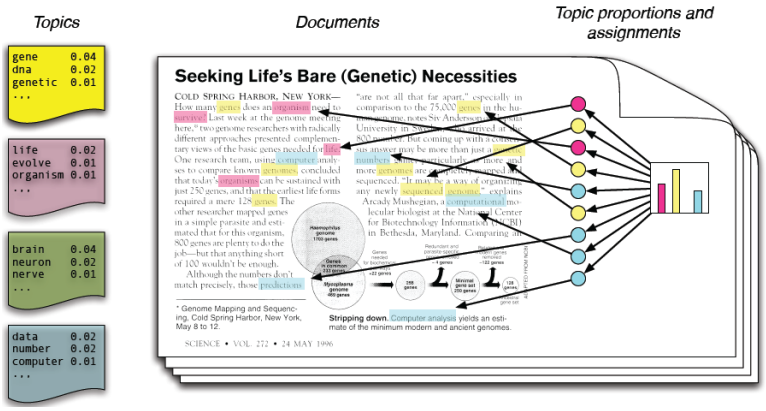


Figure 1.6: In the article "Seeking Life's Bare (Genetic) Necessities, different words have been highlighted according to the topic they belong to and the word distributions within these topics can be seen on the left. The plot on the right represents the topic distribution within the document [6].

The process of generating an LDA model involves several steps - preparing data for input to the LDA analysis (e.g. in text mining ”cleaning data” removing stop words etc), constructing a document-term matrix (in order to understand how frequently each term occurs within a document) and finally applying the LDA analysis using the document term matrix as input. This will be covered in more detail in the implementation chapter in the context of Mass2LDA.

1.4.3 Mass2LDA

In the Mass2LDA tool, LDA analysis is applied to fragmentation data in order to identify groups of fragments that commonly occur together (Mass2Motifs). The identification of Mass2Motifs allows the user to determine chemical relationships between molecules based on groups of fragments they share in common.

Below, we can see the classic LDA for text and the adapted Mass2LDA for mass spectrometry data. In classic LDA for text, documents are broken down into topics based on words that frequently co-occur. In the diagram, we can see that the topics generated from the LDA analysis are classified as ’football related topic’, ’Business related topic’ and ’environment related topic’. In Mass2LDA, rather than documents, fragmentation spectra is broken down into fragments that commonly occur together which are termed mass2motifs. In the diagram, we can see that the mass2motifs generated from Mass2LDA analysis can be classified as ’Asparagine-related’, ’Hexose-related’ and ’Adenine-related’.

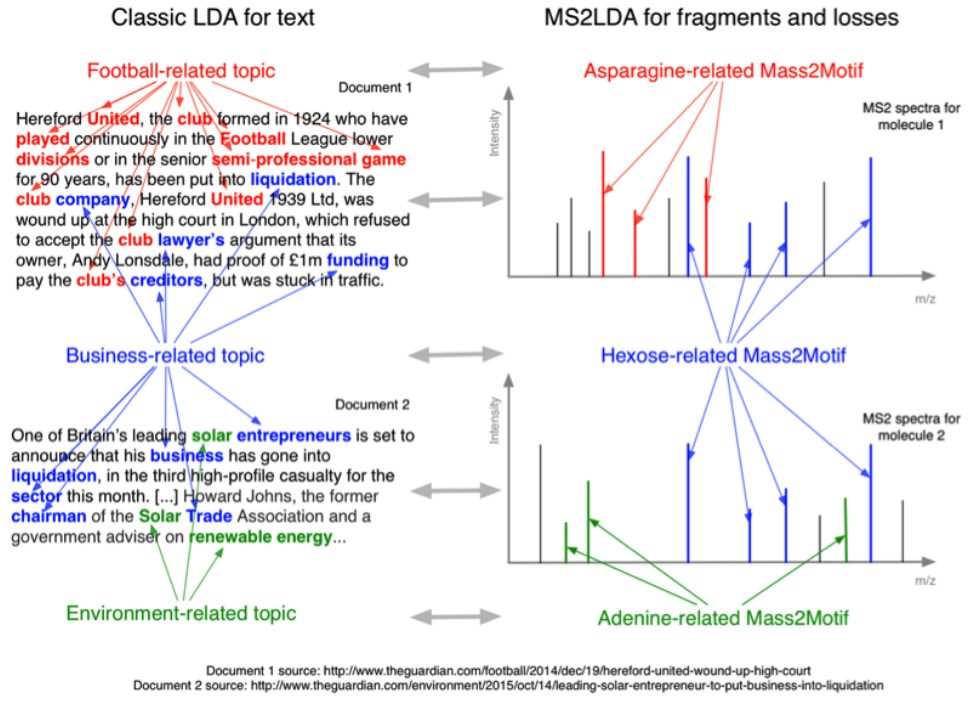


Figure 1.7: Mass2LDA compared to class LDA

To summarise, by applying LDA, Mass2LDA groups fragments that regularly occur together, identifies these groups and allows the user to identify relationships between molecules based on these shared common groupings.

Chapter 2

Requirements

The following requirements were gathered as a result of discussions at weekly project meetings with Dr Simon Rogers and Joe Wandy. Additionally, a formal requirements meeting was arranged in order to discuss requirements for the visualisation where Dr Justin van der Hooft, Mass Spectrometry Technologist at Glasgow Polyomics, was also present. Note, because of the limited time available, the expectation was that only a set number of these requirements would be implemented and these will be highlighted in the requirements documentation. However, it was still worthwhile documenting the full set of requirements because of future work that is likely to take place.

The requirements were prioritised according to the MoSCoW methodology and categorised as either a port requirement (Ms2ldaPort-), visualisation requirement (Ms2ldaVis-) or non-functional requirement (Ms2ldaNF-1). This section contains a summary list - full details of each requirement can be found in Appendix A.

Must Have

In order for this project to be considered complete these requirements must be satisfied.

Requirement ID	Description
Ms2ldaPort-01,	Add new annotation tool option
Ms2ldaPort-02	Create Mass2LDA form
Ms2ldaPort-03	Create ms2lda button to display on fragmentation set screen
Ms2ldaPort-04	LDA Analysis - feature extraction
Ms2ldaPort-05	LDA Analysis - MS2 intensity
Ms2ldaPort-06	LDA Analysis - Run Analysis
Ms2ldaPort-08	Save the Mass2LDA object to a file
Ms2ldaVis-01	MS2 spectrum plot
Ms2ldaVis-02	MS2 spectrum plot - prev and next buttons
Ms2ldaVis-10	Network Graph
Ms2ldaNF-1	The system should run on Ubuntu 15.04

Table 2.1: Requirements - Must Have

Should Have

Requirements that are deemed important but not essential to the completion of the project.

Requirement ID	Description
Ms2ldaPort-07	LDA Analysis - Do thresholding
Ms2ldaVis-03	MS2 spectrum plot - highlight mass loss between MS2 and MS1.
Ms2ldaVis-04	MS2 spectrum plot - update when node selected on Network Graph
Ms2ldaVis-05	Parent Ion plot
Ms2ldaVis-06	Parent Ion plot - mass2motif colours
Ms2ldaVis-07	Parent Ion plot - highlight mass loss between MS2 and MS1.
Ms2ldaVis-08	Parent Ion plot - update the data when a new MS1 Parent peak is selected on the MS2 Spectrum graph
Ms2ldaVis-19	Network Graph (widget 2) - degree slider
Ms2ldaVis-22	General layout - 3 different widgets
Ms2ldaVis-24	Thresholding - allow dynamic thresholding via visualisation.
Ms2ldaVis-25	Parent Ion plot - grey out peaks not associated with a mass2motif

Table 2.2: Requirements - Should Have

Could Have

Requirements that are deemed desirable but not necessary to the completion of the project.

Requirement ID	Description
Ms2ldaVis-09	Parent Ion plot - distinguish overlapping MS2 peaks
Ms2ldaVis-12	Topic plot (widget 1) - overlapping
Ms2ldaVis-13	Topic plot (widget 1) - space
Ms2ldaVis-14	Topic plot (widget 1) - external database feed
Ms2ldaVis-18	Network Graph (widget 2) - drag topic

Table 2.3: Requirements - Could Have

Would Like To Have

Requirements that are deemed desirable but not necessary to the completion of the project.

Requirement ID	Description
Ms2ldaVis-11	Multiple files
Ms2ldaVis-16	Network Graph - improve usability
Ms2ldaVis-17	Network Graph (widget 2) - grey out features
Ms2ldaVis-21	Network Graph (widget 2) - colouring of nodes.
Ms2ldaVis-23	General layout - ability to select widgets

Table 2.4: Requirements - Would Like To Have

Chapter 3

Design

3.1 Design Approach

The main objective of this project was to port the existing stand-alone Mass2LDA analysis tool into the FrAnK framework. The secondary objective was to add functionality to the visualisation (the amount of which was dependent on the remaining time available in the project). With regards to the port, the main changes related to porting the existing code or implementing equivalent functionality on to the Django web framework which FrAnK has been implemented on (see below for more details on Django). The sections below summarise the design changes that needed to be taken into account.

3.2 Existing Mass2LDA tool

The existing Mass2LDA tool is written in Python and R and runs standalone on Mac and Linux machines. The tool is split into three main sections -

1. Data Processing and Transformation. This section is written in R and takes as input MzXML format and MzML format files holding peak data. It then detects MS1-MS2 pairings and aligns the MS2 fragments across different fragmentation spectra saving the results in dataframes.
2. Mass2Motif discovery. Written in Python, the existing Mass2LDA takes as input the dataframes from the data processing stage and performs an LDA analysis on the data. Thresholding can also be applied to the data to aid visualisation. The resulting output dataframes from this stage contain the Mass2Motif document and topic data.
3. Visualisation. Several graphical widgets are provided that allow exploration of the generated Mass2Motifs and a select number of these will be ported over to the new visualisation.

3.3 FrAnK

3.3.1 Use of Django within FrAnK

Django is an open-sourced web development framework which allows users to build and maintain Web applications using a Model-Template-View (MTV) approach. MTV is way of developing software so that the code for defining and accessing data (the model) is separate from the presentation layer (the template) which in turn is separate from the user interface (the view). FrAnK currently uses Django as its web framework, therefore the exsisting Mass2LDA functionality had to be adapted to suit this framework.

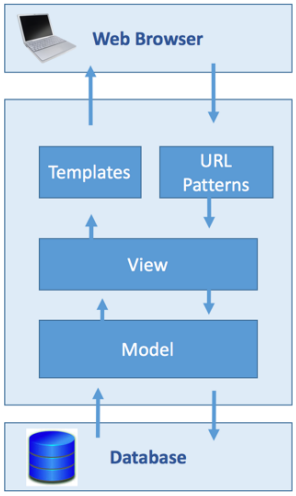


Figure 3.1: Django overview

View

The function to build the view is defined in views.py. This was modified in FrAnK to build the context dictionary needed for the new Mass2LDA visualisation using the data generated from the data analysis stage. This will be discussed in further detail in the Implementation section.

Templates

A new visualisation screen template, 'ms2lda_vis.html' was created which contains the scripts needed for the visulation (the Network Graph, MS2 Spectra plot and Parent Ion plot).

Models

Accessing the data that is needed to be passed to the view is handled in the model layer. MySQL is used for PiMP/FrAnK. A new model was not created specifically for Mass2LDA since it does not store results in the database - data is stored as files in Ms2lda_data directory instead. Mass2LDA will, however, access existing data models e.g. peak data, fragmentations set data from the database.

URL Patterns

The URL Patterns will be updated to provide a mapping between the new Mass2LDA template and the context dictionary building carried out in views.py .

3.3.2 Site Map

Appendix B. provides details of the complete site map for FrAnK but the specific areas affected with the port of Mass2LDA was as follows -

1. Fragmentation Set - if the annotation listed was created using Mass2LDA then the visualisation can be run from this screen
2. Create Annotation Query - user will now have forms to create a Mass2LDA annotation.
3. New Mass2LDA visualisation screen

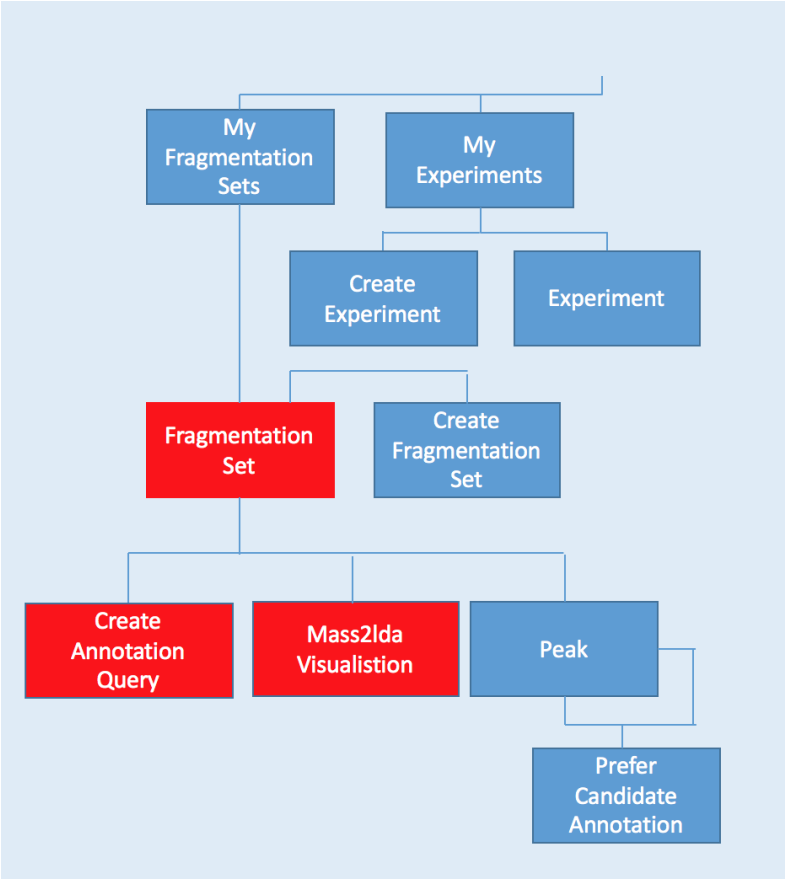


Figure 3.2: Site Map Changes

3.3.3 User parameters

The various parameters needed to run Mass2LDA are currently hard-coded. The Views section in FrAnK will be updated so that new forms will exist to allow the user to define selected parameters both for the LDA analysis and the visualisation.

3.3.4 Data Input

The MS1/MS2 data input needed for the LDA analysis and subsequent visualisation is currently read in from xml files. This will change after the port so that the MS1/MS2 peaks are read from a specific fragmentation set

held in the database. Note that this will not require any changes to the existing models in FrAnK but will make use of the existing ones.

3.3.5 LDA analysis

This will be updated to make use of the modified input and use of annotation objects within FrAnK . The actual method to perform the LDA analysis will be ported from the legacy Mass2LDA tool .

3.3.6 Annotation tool selection

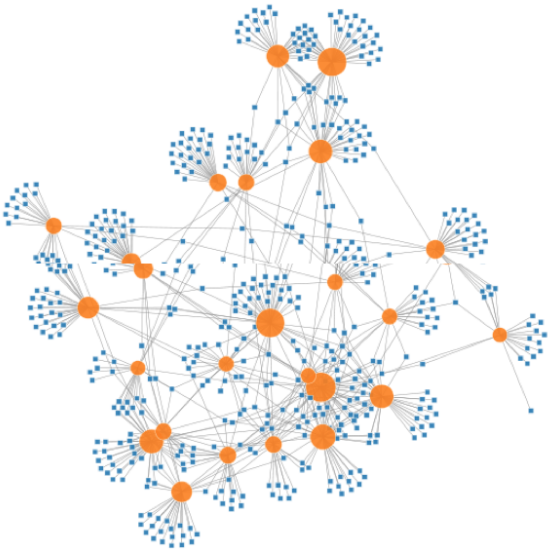
Within FrAnK, the user has the ability to select an annotation tool to analyse fragmentation data. This will be updated to allow the user to select Mass2LDA or any of the already existing annotation tools.

3.3.7 Visualisation

The legacy Mass2LDA visualisation consists of a number of screens and widgets that have been implemented using a combination of html, css, matplotlib and D3 (a JavaScript library for manipulating documents based on data and creating data visualisations in web browsers).The visualisation is extended from the topic modelling visualisation interface LDAVis [10] and allows the user to explore Mass2Motifs in MS2 data that have been identified by Mass2LDA. The decision was made to include the following three widgets in the new visualisation -

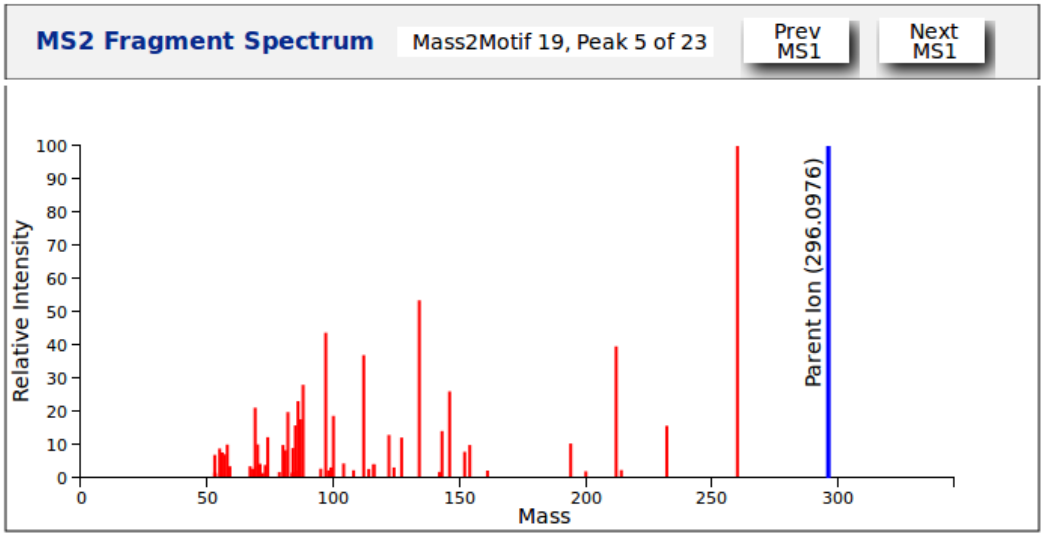
Network Graph

In the Network Graph, the user is able to select Mass2Motifs (represented by orange circles) which will highlight the MS1 peaks associated with this mass2motif (represented by the small blue squares). This allows the user to explore the identified Mass2Motifs within the dataset and how they connect to other Mass2Motifs.



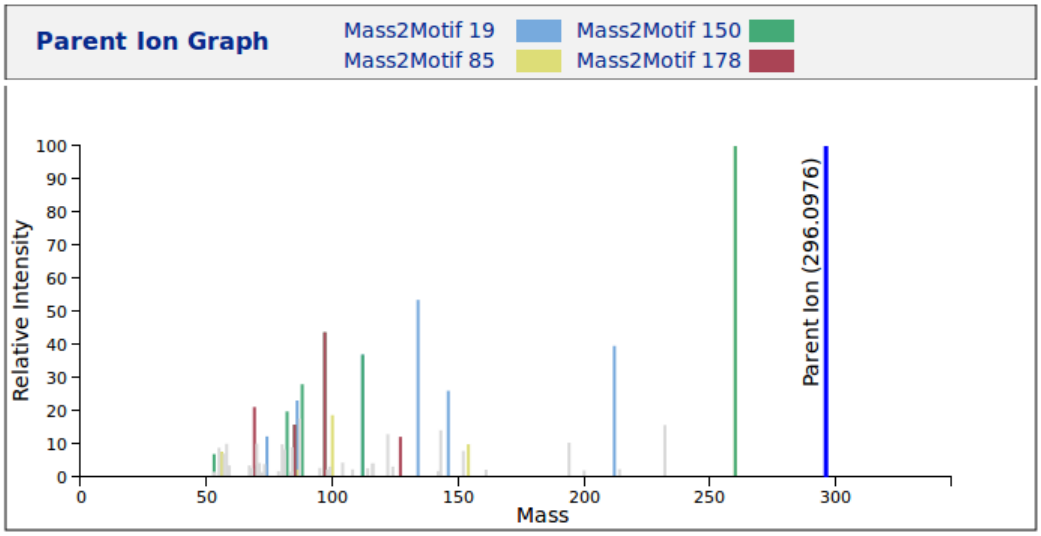
MS2 Spectra plot

The MS2 spectra plot will display the MS2 Spectra for a given MS1 parent peak. Within the plot, the user will be able to scroll through all MS1 parent peaks whose spectrum contains the mass2motif selected on the network graph. Hovering over the fragment peaks will also display the mass losses against the parent peak.



Parent Ion plot

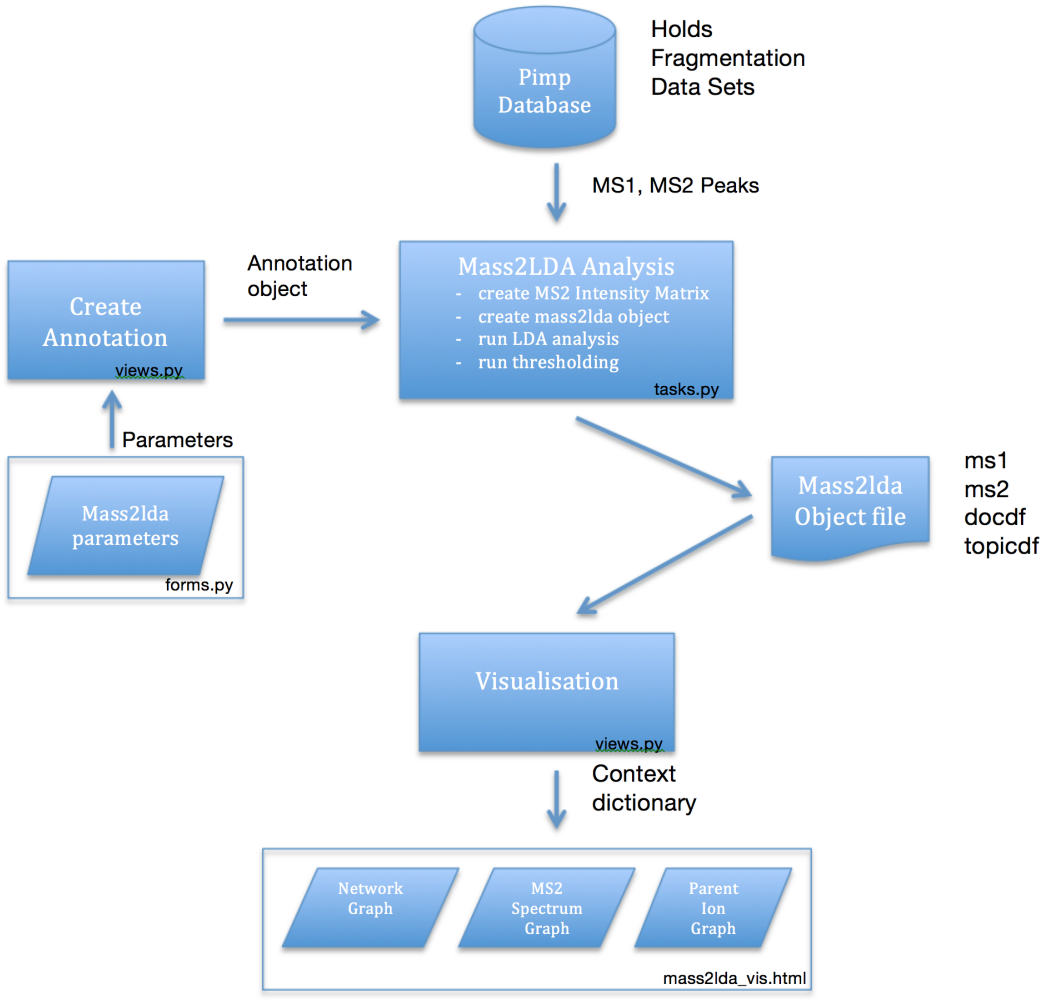
The Parent Ion plot will graph the MS2 peaks for the parent ion currently displayed on the MS2 Spectra plot. It will then colour each fragment within this spectra according to the Mass2Motif it is associated with (if any). If a MS2 peak does not have an associated Mass2Motif then it will still be displayed but greyed out. Unlike in the MS2 plot, multiple colours are needed in the parent ion plot to colour the fragments. Because it is these colours that allow users to differentiate between peaks, it is important to use colours that can be seen by the majority of users including colour-blind users. The colours that were chosen were selected due to the fact that they are distinct as possible both in normal vision and also in colour blind vision [1].



3.4 Summary Architecture Diagram for Mass2LDA FrAnK running within FrAnK

The following diagram shows the overall architecture of the Mass2LDA tool within FrAnK.
In summary -

- The fragmentation data is held in the PiMP database (the fragmentation data set used during the development has been beer3pos which was generated from an analysis of IPA style of beer).
- The MS1 and MS2 peaks are extracted from these fragmentation sets and stored in dataframes to be used as input to LDA analysis.
- The parameters used within the LDA analysis are defined as part of the creation of an annotation (a new form has been created to define the parameters - see Appendix D).
- An LDA analysis and threshold, based on the parameters provided, is carried out on the data. A Mass2LDA object is created to hold the results in a .project file.
- When the user presses the visualisation button a context dictionary is created holding the graph and plot data and a new Mass2LDA visualisation screen displayed.



Chapter 4

Implementation

4.1 Development Overview

It was clear at the start of the project that there was considerable risk because of the learning curve that had to come up both in terms of the subject matter - Mass Spectrometry - and the technologies involved - Python, Django, JavaScript , D3, NetworkX etc. most of which I had minimal knowledge.

As a result of this, the decision was made to take an Agile approach to the development aiming for small incremental releases of functionality every two weeks. The functionality to be implemented was reviewed and prioritised at the start of each Sprint (based on the discussions with Dr Simon Rogers, Joe Wandy and Dr Justin van der Hooft) and the highest priority items selected for the following sprint. The overall approach, however, was to focus on completing the port into FrAnK first and then work on the visualisation changes after with any time remaining.

The sprints that were worked on turned out as follows -

Sprint Release Plan				
Sprint	Start Date	End Date	ID	Functionality
#1	12-Oct-15	25-Oct-15		
			1	Investigaton
#2	26-Oct-15	08-Nov-15		
			2	Development/runtime environment setup
			3	Mass2LDA tool add into Frank
#3	09-Nov-15	22-Nov-15		
			4	Lda analysis, LDA fragmentation, MS2 Intensity
#4	23-Nov-15	06-Dec-05		
			5	Complete Lda analysis, LDA fragmentation, MS2 Intensity
			6	Requirements clarification
#5	07-Dec-15	20-Dec-15		
			7	D3, javascript investigation,
			8	Merge mass2lda into forms
Christmas Holiday				

#6	18-Jan-16	31-Jan-16		
			9	Network Graph Screen
#7	01-Feb-16	14-Feb-16		
			10	MS2 Spectrum Graph
#8	15-Feb-16	28-Feb-16		
			11	Parent Ion Graph
			12	Dissertation write-up (investigation)
#9	29-Feb-16	13-Mar-16		
			13	Defect fixing
			14	Testing
			15	Evaluation
			12	Dissertation write-up (first pass)
#10	14-Mar-16	25-Mar-16		
			14	Testing (complete)
			12	Dissertation write-up (final)
			16	Demo
Submission deadline 25th March 2016				

Figure 4.1: Sprint Release Plan

4.2 Development/runtime environment setup

A setup guide exists to install the FrAnK pipeline on both a Linux (Ubuntu 15.04 & Fedora 21) and Mac OS X Yosemite operating systems. The install includes a variety of packages - Java, R, XCode, Anaconda Python, virtualenvwrapper for python, PiMP, PiMPDB and Celery.
Initially the decision was made to install the pipeline on a Mac OSX system but this resulted in a considerable number of issues during the install and as a result the decision was made to install on Ubuntu 15.04 instead and all subsequent development and testing was done on this environment alone.

4.3 Port of Mass2LDA in to FrAnK

The port into FrAnK consisted of three main parts -

1. Addition of a new annotation tool and specification of parameters for the tool
2. After a new annotation has been created, run the LDA analysis and feature extraction on the fragmentation data and save the results.
3. Format the data, update the urls and provide context dictionary for the visualisation screens.

4.3.1 Addition of a new annotation tool and parameter selection

1. Updates were made to FrAnK via the Admin login to create an additional annotation tool - Mass2LDA - to select from. After the user has selected a fragmentation set from the 'My Fragmentation Sets' screen they will then be taken to the 'Annotation Sets' screen where the user will have a pull down menu option to select the annotation tool to use when creating the new annotation

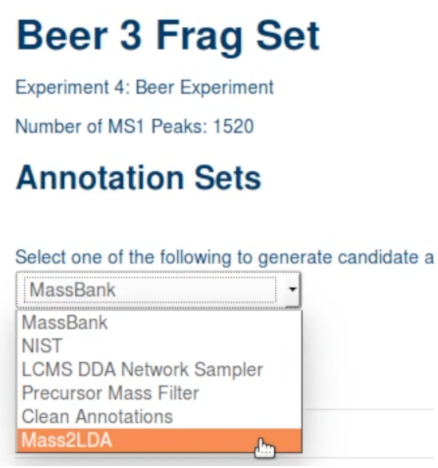


Figure 4.2: Annotation Tool Options

2. A new query form - Mass2LDAQueryForm - was created in forms.py tol display the parameters that can be entered by the user for the Mass2LDA annotation tool (see Appendix D). This form will be displayed when the user presses the 'Create New Annotation Query' button on the 'Annotation Sets' screen.

3. Details of the annotation are saved into an annotation query object (including the paramters, fragmentation set and the tool used). The method set_annotation_query_parameters in views.py was updated so that it extracted the parameters from the Mass2LDA form and added these to an annotation_query_object.

4. The parameters that can be specified by the user are split between those used for pre-filtering, e.g. Grouping Tolerance, and those used in the lda analysis, e.g. Gibbs sampling number. The full set of parameters are -

Pre-filtering Parameters

Parameter	Default	Minimum	Maximum
Minimum MS1 Intensity	0	0	300000
Minimum MS2 Intensity	0	0	300000
Minimum MS1 Retention Time	0	0	50
Maximum MS1 Retention Time	2000	1	2000
Grouping Tolerance	7	1	10
Scaling Factor	100	0	100
Polarity	Positive	Negative	Positive

Table 4.1: Pre-filtering Parameters

LDA Analysis Parameters

Parameter	Default	Minimum	Maximum
Alpha Model Parameter	0.1	0.0	50.0
Beta Model Parameter	0.01	0.0	50.0
Gibbs Sampling number	3	1	1000
Mass2motif count	300	1	10000

Table 4.2: LDA Analysis Parameters

5. The screen that displays the annotations that have been created for a Fragmentation Set - fragmentation_set.html - was updated so that if the annotation tool used was Mass2LDA then there is now a button displayed labelled 'Mass2LDA visualisation' which will allow the user to run the visualisation for that annotation i.e. the Network Graph, MS2 Spectrum and Parent Ion plots. See Appendix C. for an example of the Fragmentation Set screen.

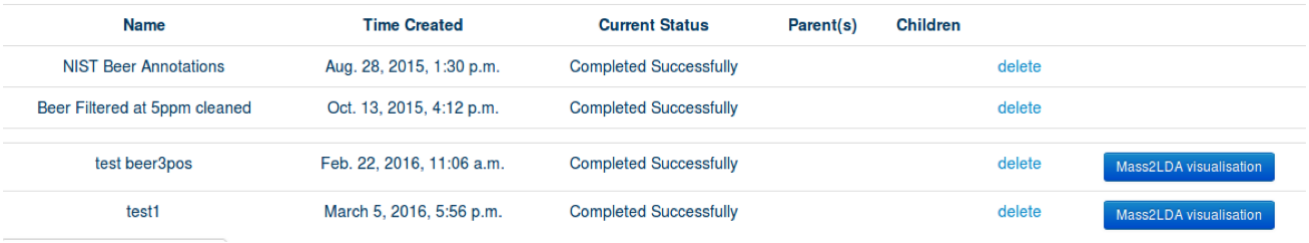


Figure 4.3: Mass2LDA Visualisation button

4.3.2 LDA Feature Extraction and Analysis

After the user has specified the annotation parameters and pressed the Retrieve Annotations button on the Mass2LDA form an LDA Feature Extraction and Analysis will be performed on the Fragmentation Set based on the parameters specified.
A function was created - run_mass2lda_analysis - in tasks.py that performs the following steps -

1. Extracts the fragmentation set details (as specified in the annotation query) from the database
2. Extracts the MS1 peaks of the fragmentation set from the database (using the polarity specified as a parameter) and saves them into a pandas dataframe.
3. Extracts the MS2 peaks of the fragmentation set from the database and saves them into a pandas dataframe.
4. Calculates the MS2 fragment (Word) values - groups the MS2 peaks into fragment buckets based on the grouping tolerance which had been passed in as parameter.
5. Create a pandas dataframe holding the MS2 Intensity Matrix which is the intensity of ms2 peaks for a given ms1 peak where the column is the ms1 peak and the rows are the ms2 fragments
6. Save the MS1, MS2 and MS2 intensity dataframes into a Mass2LDA object. The Mass2LDA class is defined in lda_for_fragments.py and has been copied across unmodified from the existing Mass2LDA code (this will need to be included in the port as a dependency).
7. Run an LDA analysis on the data. This uses the 'run_lda' method from the Mass2LDA object and takes as parameters the number of topics, the alpha and beta model.
8. Run thresholding on the data. This uses the 'do_thresholding' method from the Mass2LDA object which will limit the number of topics identified according to the frequency they appear across spectra. The output from this will two attributes in the Mass2LDA object - docdf and topicdf - which will hold the parent MS1 peaks in each topic and the fragment buckets in each topic respectively.
9. Saves the Mass2LDA object off to file so that it can be used later in the visualisation. Each annotation created for Mass2LDA will be saved in a file named 'mass2lda_annotation id.project' in the $HOME/mass2lda_data/ directory. The decision to save the data off to disc as opposed to the PiMP database was made because of the size of .project file, the fact that it is not an overly complex dataset and there was no requirement to perform complex queries on the data.

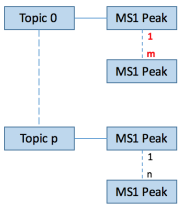
4.3.3 Url updates and visualisation context dictionary

The Django framework uses urlpatterns to handle urls/web addresses. The urlpatterns data structure in urls.py was updated to add a new entry - mass2lda_vis - which will call the method views.mass2lda_vis to build the context dictionary and display the new Mass2LDA visualisation screen (this url will be triggered when the user presses the Mass2LDA visualisation button on Fragmentation Set screen).
The context dictionary used by the new Mass2LDA screen is built by the function get_mass2lda_vis_context_dict in views.py. This will perform the following functions -

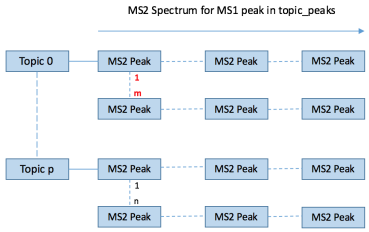
- Retrieve the Mass2LDA object from the saved file $HOME/mass2lda_data/mass2lda_annotation id.project
- Create the data needed for the MS2 Spectra and Parent ion plots. This will use the ms2, docdf and topicdf attributes from the Mass2LDA object and will create three datasets (made up of mass/intensity pairs) that will be used to generate the data for the plots.

  The datasets created were -

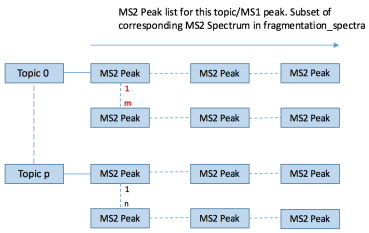
  – topic_peaks: Holds a nested list of topics, and each topic will point to a list of MS1 peaks in that topic. This will be used by both the MS2 Spectra and Parent Ion plots.



  – fragmentation_spectra: A nested list that holds the complete MS2 spectrum for each MS1 peak. There will be an MS2 Spectrum list associated with the corresponding MS1 entry in topic_peaks. This will be used by the MS2 Spectra plot.



  – topic_fragments: a nested list of the MS2 fragments that exist for each topic. There will be a MS2 fragment list associated with the corresponding MS1 entry in topic_peaks. This will be used by the Parent Ion plots.



- Build the data needed to display the Network Graph. The graph uses the NetworkX python software package and the data will consist a collection of nodes along with edges for those nodes. The NetworkX plots are created using the existing function get_json_from_docdf in visualisation/networkx/lda_visualisation.py (this will need to be included as a dependency from the legacy Mass2LDA in the port). The input to this will be the docdf attribute from the Mass2LDA object.

- In summary, the context dictionary passed over to the mass2lda_vis.html will consist of the following -

```
'fragment_slug'   : fragmentation_set_name_slug,
'annotation_name': annotation_query.name,
'graph'           : json.dumps(json_data),
'topic_data'      : json.dumps(topic_data)
```

  where,

  1. fragment_slug, annotation_name are used to display details of the annotation on the screen.
  2. graph data contains the nodes & edges for the Network Graph.
  3. topic_data holds the mass/intensity pairs in topic_peaks, fragmentation_spectra and topic_fragments needed for the MS2 Spectrum and Parent Ion plots.

4.4 Visualisation

The visualisation for the Mass2LDA annotation tool now consists of three graphs - Network Graph, MS2 Spectrum plot and the Parent Ion plot. The graphs/plots use a combination of html, css, JavaScript, D3 and NetworkX to display the data and the scripts for all graphs/plots are held in the mass2lda_vis.html file.

4.4.1 Network Graph

This plot presents a visual representation of mass2motif nodes and their links. It is written using NetworkX language and is primarily a re-use of the Network Graph scripts from the original Mass2LDA tool (graph.html). The input to the graph will be a dataset of nodes and edges between the nodes which, for the purposes of the port, is passed in json format via the context dictionary.

Note that the the 'degree' parameter which filters the nodes that are displayed on the graph defaults to 20. This should be enhanced with a slider on the screen to allow dynamic changing of the threshold. This has been included as a requirement for a future release - see Appendix A for further details.

The user has the option to select a node on the plot and this will result in the data displayed on the MS2 Spectrum plot and Parent Ion plot being updated to display data related to the selected mass2motif.
See Appendix C.1 for example screen.

4.4.2 MS2 Spectrum plot

This plot uses D3 to display the full MS2 spectrum for each of the MS1 Parent peaks in a selected mass2motif. The user will have the option to hover the mouse over an MS2 peak and further information on the peak (mass and relative intensity) and the loss mass between the MS2 Peak and its MS1 Parent peak will be displayed.

The scripts to display the plot are held in a function display_ms2_spectrum_graph in the file mass2lda_vis.html. See Appendix C.2 for example screen.

4.4.3 Parent Ion plot

This will use D3 to display, for a selected MS1, the MS2 peaks that are present in all mass2motif that contains that parent. Each of the mass2motif displayed on the plot will be represented by a separate colour.
As with the MS2 Spectrum plot, the user will have the option to hover the mouse over an MS2 peak and get details of its mass and relative intensity displayed and mass against its parent peak.
The scripts to display the plot are held in a function display_parent_ion_topics_graph in the file mass2lda_vis.html. One of the first tasks it will do is to create the peaks to be plotted on the plot. It is written in JavaScript and will use the following algorithm in to run down the datasets and retrieve the relevant data needed -

```
For each topic
    Get the list of ms1 peaks for this topic
    If current peak on display is in the ms1 list
           Extract the topic frag peaks and add to plot data
```

When deciding how to generate the data for this plot several options were considered. The first option was to generate data for all potential MS1/mass2motif combinations that could be displayed and pass this through the context dictionary. This would, however, require a large quantity of data to be generated and would add to the noticeable amount of time already being taken to build the context dictionary before the visualisation screen is displayed. Only one set of data will be on display at any one point and it is therefore not necessary to generate all data before the visualisation is loaded - the necessary data can be generated on an on-demand basis when the MS2 graph is updated by selection of the prev/next buttons. The decision was therefore made to generate the data dynamically with the above algorithm using javascript in the Mass2LDA visualisation. This algorithm will extract the relevant data on-demand from the topic_peaks, fragmentation_spectra and topic_fragments datasets described above.

See Appendix C.3 for example screen.

Chapter 5

Evaluation / Testing

5.1 Evaluation

To carry out the evaluation, a meeting was arranged with Joe Wandy and Dr Justin van der Hooft in order to discuss the progress of the port and identify changes to be made in future developments.

During the evaluation meeting, copies of the set of requirements gathered during the scope of the project were distributed and each of these requirements discussed in turn. When considering each requirement, it was discussed as to whether the requirement had been achieved, any difficulties that had occurred in completing the requirement and any changes that needed to be made. In order to evaulate this, both Joe and Justin were given access to the final product. Additional requirements had been discussed during the requirements gathering meeting that were not intended to be worked upon during the scope of this project and these have been recorded in the requirements as 'not included' (see Appendix A). Therefore in the evaluation, only the requirements specified as being included in the scope of the project were discussed.

A summary of the results of the evaluation is as follows -

Requirement ID	Comments
MS2ldaMerge -01 (Add new annotation tool option)	Complete
MS2lda Merge -02 (Create Mass2LDA forms)	Some parameter input validation was incorrect (e.g. MS1 intensities). These values were discussed with Justin and appropriate parameter values have been now included. One of the parameters was not being passed in the code (gibbs sampling number). Following the evaluation, this has now been fixed.
MS2lda Merge -03 (Create ms2lda button to display on fragmentation set screen)	Complete
MS2ldaMerge-04 (LDA analysis feature extraction)	Complete
MS2ldaMErge-05 (LDA analysis - MS2 Intensity)	Complete
MS2ldaMerge -06 (LDA analysis ?-Run analysis)	Complete. Note - Check gibbs sampling number is passed properly as noted above.
MS2ldaMerge-07 (LDA analysis do thresholding)	Complete however currently carried out when annotation is created. Would be better to have this dynamically (this would have been good to have included however time restrictions meant this was not included).

Requirement ID	Comments
MS2ldaMerge-08 (Save Mass2LDA object to a file)	Complete
MS2ldaVis-01 (MS2 spectrum graph)	Complete
MS2lda-vis-02 (MS2 spectrum graph scrolling functionality)	Complete. It was noted it would be better to change the cursor when hovering over a button to make it clearer. Following the evaluation, this was implemented.
MS2ldaVis-03 (losses functionality)	Complete
MS2ldaVis-04 (Update MS2 spectrum when mass2motif is selected on network graph)	Complete
MS2ldaVis-05 (Parent ion graph)	Complete
MS2ldaVis-06 (Parent ion graph - mass2motif colours)	Complete
MS2ldaVis-07 (Parent ion graph losses)	Losses shown on plot however currently not coloured according to mass2motif.
MS2ldaVis-08 (Parent ion graph - update when new MS1 is displayed on MS2 graph)	Complete
MS2ldVis-09 (Parent ion graph - distinguish overlapping MS2 peaks)	Still to be implemented - routine to distinguish provided by Joe and needs to be incorporated into code.
MS2ldaVis-10 (Network graph)	Complete
MS2ldaVis-12 (Overlapping)	Complete
Ms2ldaVis-22 (3 widgets)	Complete

Table 5.1: Evaluation discussion

Overall, from the evaluation it was concluded that the majority of the requirements that were intended to be completed had been but some still requiring minor adjustments to be classified as completed. The visualisation that had been completed so far satisfied the requirement for 3 basic widgets that interacted with each other however will be updated in future developments to improve functionality and add additional widgets. The changes to be made in future developments (gathered from the evaluation and the meetings that took place throughout the project) are documented in the next chapter.

5.2 Testing

The approach taken for the testing covered the following areas. Although extensive testing would have been desirable, given the time restrictions, the decision was made to concentrate on carrying out a wide breadth of tests rather than a comprehensive in-depth set of tests.

5.2.1 Comparison against legacy Mass2LDA tool

The first area of testing was to compare the results of the visualisation in the legacy Mass2LDA tool against the visualisation in the new Mass2LDA tool. In order to do this, one of the project files (beer3pos.project) from the legacy Mass2LDA tool was loaded into the new Mass2LDA tool and the results of both visualisations were compared.

Test ID	Description	Result	Comments
Ms2ldaTest1-1	Set degree parameter on legacy to 20 and compare mass2motif displayed against the mass2motif on new Network Graph visualisation	PASSED	Note that 'degree' parameter on new screen defaults to 20.
Ms2ldaTest1-2	Chose a number of Mass2Motif and compare the MS1 peaks against those displayed in the new MS2 Spectra visualisation	PASSED	
Ms2ldaTest1-3	Chose a number of Mass2Motif and compare the MS2 Spectra for selected MS1 Parents against those displayed in the new MS2 Spectra visualisation	PASSED	
Ms2ldaTest1-4	Chose a number of Mass2Motif and compare the MS Fragments against those displayed in the new Parent Ion visualisation	PASSED	

5.2.2 Parameter/forms testing

Each of the parameters in the Mass2LDA forms screen was tested to ensure only valid values for each parameter could be entered and that valid values entered in the form were reflected in the visualisation results.

Test ID	Description	Result	Comments
Ms2ldaTest2-1	Ensure defaults values listed in Section 4.3.1 are displayed	PASSED	
Ms2ldaTest2-2	Ensure only values within permitted range listed in Section 4.3.1 can be entered	PASSED	
Ms2ldaTest2-3	Test Minimal MS1 Intensity parameter	PASSED	Set a filter of peaks greater than 250,000 and checked ms1 dataframe in .project file to confirm.
Ms2ldaTest2-4	Test Minimal MS2 Intensity parameter	FAILED	Set a filter of peaks greater than 2000 and checked ms2 dataframe in .project file to confirm. Results not as expected and further invesitgation required.
Ms2ldaTest2-5	Test Minimal MS1 Retention Time parameter	PASSED	Set filter of 10 seconds and checked ms1 dataframe in .project file to confirm.
Ms2ldaTest2-6	Test Maximum MS1 Retention Time parameter	PASSED	Set filter of 600 seconds and checked ms1 dataframe in .project file to confirm.

Test ID	Description	Result	Comments
Ms2ldaTest2-7	Test Grouping Tolerance parameter	PASSED	
Ms2ldaTest2-8	Test Scaling Factor parameter	PASSED	Compared results in df dataframe of default against modified parameter objects
Ms2ldaTest2-9	Test Polarity parameter	PASSED	Checked both Positive and Negative Polarity . Compared ms1 results in .project file to confirm.
Ms2ldaTest2-10	Test Alpha Model parameter	PASSED	
Ms2ldaTest2-11	Test Beta Model parameter	PASSED	
Ms2ldaTest2-12	Test Gibbs Sampling parameter		Set Gibbs Sampling to 10. Results appear to be correct but may need further analysis to confirm. (The trace on the console showed that 10 samples were taken from the LDA analysis.
Ms2ldaTest2-13	Test Mass2Motif count parameter	PASSED	Tested with Mass2Motif values of 50, 300, 2000. Note that,for the purpose of these tests, the 'degree' value was set to 0 (normally set at 20) so the results of the test could be fully verified.
Ms2ldaTest2-14	Ensure only annotations created for Mass2LDA tool have the 'Mass2LDA Visualisation' button displayed beside them	PASSED	

Table 5.2: Test - parameters / forms

Note that the following steps were taken when checking parameter values -

1. A test harness was written that takes as parameter the name of a .project file and outputs the following csv files for the following attributes that were in the Mass2LDA file - ms1.csv, ms2.csv, df.csv, docdf.csv, topicdf.csv.
2. Parameter under test was changed e.g. minimum ms1 intensity, in new annotation form and a new annotation created.
3. After the annotation had been created the .project file from the mass2lda_dir was saved in to a test results directory.
4. The test harness was then run against the .project file to create the csv files.
5. The contents of the csv files were then checked - e.g. for minimum ms1 intensity the ms1.csv - to ensure that all entries had a ms1 intensity greater than the specified value.

5.2.3 Data analysis output compared to visualisation

Output from the data analysis stage was compared against the expected visualisation results. A utility was written that takes as input a .project file (from the Mass2LDA_data directory) and outputs the ms1, ms2, docdf and topicdf dataframes into csv files. This data was then compared to the results displayed on the new visualisation.

Test ID	Description	Result	Comments
Ms2ldaTest4-1	Create new annotation, run utility against the .project file. Select a number of Mass2Motif from the visualisation and check against ms1/ms2/docdf/topicdf that the correct MS1 parent peaks have been displayed on the MS2 Spectra plot for those mass2motif.	PASSED	Checked a number of mass2motif from docdf to confirm the correct MS1 peaks associated with that mass2motif are displayed on the MS2 Fragment Spectrum plot.
Ms2ldaTest4-2	Create new annotation, run utility against the .project file. Select a number of MS1 peaks from the visualisation and check against ms1/ms2/docdf/topicdf that the correct MS2 peaks have been displayed on the MS2 Spectra plot.	PASSED	Selected MS1 and associated MS Spectra from Spectrum plot. Checked relevant details from ms1 and ms2 dataframes to ensure that the details on the plot are correct.
Ms2ldaTest4-3	Create new annotation, run utility against the .project file. Select a number of MS1 peaks from the MS2 Spectra plot and check against ms1/ms2/docdf/topicdf that the correct MS2 peaks are displayed in the Parent Ion are coloured correctly against the correct mass2motif	PASSED	Selected MS1 peak on MS2 Spectrum plot. In docdf determined which Mass2Motif had this peak. Then checked the MS2 peaks listed in topicdf are displayed correctly on graph and that any MS2 peak in the spectrum which is not associated with a mass2motif is displayed in grey on the graph.

Table 5.3: Test - Data Analysis compared to visualisation

5.2.4 Fragmentation sets

During development beer3pos was used to verify functionality. This was expanded to use other fragmentation sets during the test phase.

Test ID	Description	Result	Comments
Ms2ldaTest3-1	Select Beer3 Frag set. Create annotations and ensure results are as expected.	PASSED	Note that this was the default fragment set used during development.

Table 5.4: Test - Fragmentation Set

Test ID	Description	Result	Comments
Ms2ldaTest3-2	Select Walkthrough Frag set. Create annotations and ensure results are as expected.	PASSED	Used default parameter values and changed 'degree' parameter to 0. Results appear to be correct but limited experience of this fragment set so possibly needs more investigation.

Table 5.5: Test - Fragmentation Set

5.2.5 Visualisation

The Network Graph, MS2 Spectra plot and Parent Ion plot were tested against the requirements detailed in Appendix A

Test ID	Description	Result	Comments
Ms2ldaTest5-1	Run the visualisation. Ensure on the Network Graph any mass2motif can be selected and the relevant data will be displayed on the MS2 Spectra and Parent Ions plots.	PASSED	
Ms2ldaTest5-2	Run the visualisation. Ensure on MS2 Spectra plot it is possible to scroll through all MS1 Parent peaks in the selected Mass2Motif	PASSED	
Ms2ldaTest5-3	Run the visualisation. Ensure on MS2 Spectra plot that moving the cursor over MS2 peak will display the mass loss again the parent peak.	PASSED	
Ms2ldaTest5-4	Run the visualisation. Ensure on MS2 Spectra plot that selecting a new MS1 Parent Peak to be displayed will also update the Parent Ion plot with the details of that MS1 parent peak	PASSED	
Ms2ldaTest5-5	Run the visualisation. Ensure on Parent Ion plot that the correct MS2 peaks are displayed for each mass2motif that contains that ms1 peak.	PASSED	
Ms2ldaTest5-6	Run the visualisation. Ensure on Parent Ion plot that any MS2 peaks not associated with a mass2motif are still displayed but are greyed.	PASSED	

Table 5.6: Test - Visualisation

Chapter 6

Discussion and Conclusion

6.1 Conclusion

In summary, this project concentrated on porting a standalone tool Mass2LDA into the FrAnK framework in the form of a new annotation tool. Additionally, a basic visualisation was developed consisting of three widgets (Network graph, MS2 spectrum plot and Patent ion plot) that interact with and update each other.

6.2 Future work

During the course of the project, potential improvements and future work for the Mass2LDA tool were identified and some of which have been noted below.

6.2.1 Multiple Files/ Projects

As discussed during the formal requirements meeting, it would be useful to develop Mass2LDA to support multiple files. Functionality would be implemented to allow the user to compare these files through the visualisation to, for example, identify common features. Note that during this implementation input is taken from fragmentation sets held within the database instead of files therefore this requirement could be met by loading multiple saved annotations (which may have been created from different fragmentation sets) and analysing these.

6.2.2 External database feed

Several online mass spectral databases exist where the user can take MS2 information and use this as queries to the database. It would be useful to implement a way of extracting relevant information from Mass2LDA to use as input to an external database such as mzcloud. Currently, some of the annotation tools exsisting in FrAnK perform this functionality and therefore being able to share projects between annotation tools to perform different functions may be useful.

6.2.3 Dyanmic Thresholding

Thresholding is currently hardcoded prior to loading the visualisation. It would be useful to able to do this dynamically as part of the visualisation.

6.2.4 Annotation

It should be possible to add annotation labels, for instance, identifying for a given MS2 fragmentation spectra the topics corresponding to actual chemical substructures.

6.2.5 Mass2motif Features frequencies histogram

Within the legacy Mass2LDA, two Mass2Motif feature frequencies histograms exsist providing information on:

- Displaying the count of each Mass2Motif associated fragment or loss within the fragmentation spectra explained by the Mass2motif
- Overall frequency of the fragments within the dataset that can be explained by the currently selected Mass2Motif.

This should be ported over to the new visualisation.

6.3 Reflection

Starting this project without a background in mass spectrometry analysis and in many of the range of technologies involved in the FrAnK pipeline required a considerable amount of background investigation. However, as a result of this research my knowledge of these areas has expanded considerably and I feel that I have a much broader awareness of the software engineering process. Being able to follow a project life cycle through the stages of investigation, assessing requirements, planning, design, implementation, testing and documenting has allowed me to connect my theoretical knowledge to a practical application which will give me a solid basis for project development in the future.

6.4 Acknowledgements

I would like to thank my supervisor Dr Simon Rogers, Joe Wandy and Dr Justin Van der Hooft for their continued support and input throughout this project.

Appendices

Appendix A

Requirements

A.1 Port Requirements

Id:	**Ms2ldaPort-01**
Requirement:	Add new annotation tool option
Description:	Add mass2lda option into fragmentation set screen from annotation sets drop down menu
Priority:	Must Have
Comments	
Scope:	Included

Id:	**Ms2ldaPort-02**
Requirement:	Create mass2lda form
Description:	Create mass2lda form using the Django forms mechanism to display the parameters needed for mass2lda annotation tool. Parameters must be saved in an annotation query object.
Priority:	Must Have
Comments	Parameters to be included: • Pre filtering: Minimal MS1 intensity, Minimal MS2 intensity, Min MS1 retention time, Max MS1 retention, Grouping tol, Scaling factor. • LDA analysis: Alpha model parameter, beta model parameter, Gibbs sampling number, max no. Mass2Motif
Scope:	Included

Id:	**Ms2ldaPort-03**
Requirement:	Create ms2lda button to display on fragmentation set screen
Description:	If annotation created is a mass2lda analysis, then add ms2lda button beside annotation on the fragmentation set screen which will open visualisation in a new page.
Priority:	Must have
Comments	
Scope:	Included

Id:	**Ms2ldaPort-04**
Requirement:	LDA Analysis - feature extraction
Description:	Peak data (MS1 and MS2) to be extracted from the specified fragmentation set in the database and save in pandas dataframes.
Priority:	Must have
Comments	Data will be extracted from the database – not mzXML files. Assumption is that all relevant data has been loaded into a fragmentation set in the database.
Scope:	Included

Id:	**Ms2ldaPort-05**
Requirement:	LDA Analysis – MS2 intensity
Description:	Create MS2 intensity matrix using fragments based on grouping tol.
Priority:	Must have
Comments	
Scope:	Included

Id:	**Ms2ldaPort-06**
Requirement:	LDA Analysis – Run Analysis
Description:	Run LDA analysis using parameters specified in Mass2lda form (No. of Mass2Motif, alpha, beta and Gibb's sampling).
Priority:	Must Have
Comments	
Scope:	Included

Id:	**Ms2ldaPort-07**
Requirement:	LDA Analysis – Do thresholding
Description:	Output from thresholding will be docdf and topicdf data structures. Note that thresholding is done once only at the time the annotation is created.
Priority:	Should Have
Comments	
Scope:	Included

Id:	**Ms2ldaPort-08**
Requirement:	Save the mass2lda object to a file
Description:	Mass2lda object should be saved to a file for use later in visualisation.
Priority:	Must Have
Comments	
Scope:	Included

A.2 Visualisation Requirements

Id:	**Ms2ldaVis-01**
Requirement:	MS2 spectrum plot
Description:	On the visualisation screen provide a plot which displays MS2 spectrum for a given MS1 peak.
Priority:	Must Have
Comments	
Scope:	Included

Id:	**Ms2ldaVis-02**
Requirement:	MS2 spectrum plot – prev and next buttons
Description:	MS2 spectrum plot – provide previous and next buttons to allow the user to scroll through all the MS1 peaks in a selected Mass2Motif. When a new MS1 peak is selected the full MS2 Spectrum for that peak will be displayed.
Priority:	Must Have
Comments	
Scope:	Included

Id:	**Ms2ldaVis-03**
Requirement:	MS2 spectrum plot – highlight mass loss between MS2 peak and MS1 Parent Peak.
Description:	The user should be able to place the cursor over any MS2 peak in the plot and it will display details of the mass loss between that peak and the MS1 parent peak.
Priority:	Should Have
Comments	
Scope:	Included

Id:	**Ms2ldaVis-04**
Requirement:	MS2 spectrum plot – update mass2motif data when node selected on Network Graph
Description:	If the user clicks on a new node (mass2motif) on the Network Graph the information displayed on the MS2 Spectrum Plot should be updated to reflect the data from the selected Mass2Motif.
Priority:	Should Have
Comments	
Scope:	Included

Id:	**Ms2ldaVis-05**
Requirement:	Parent Ion Graph
Description:	Provide a Parent Ion graph that, for a given MS1 Parent Peak displayed on the MS2 Spectrum graph, will display the mass2motif fragments for all mass2motifs that has that MS1 in its Topic Peak list.
Priority:	Must Have
Comments	
Scope:	Included

Id:	**Ms2ldaVis-06**
Requirement:	Parent Ion Plot – mass2motif colours
Description:	Each set of mass2motif fragments will be represented by a different colour on the plot.
Priority:	Should Have
Comments	
Scope:	Included

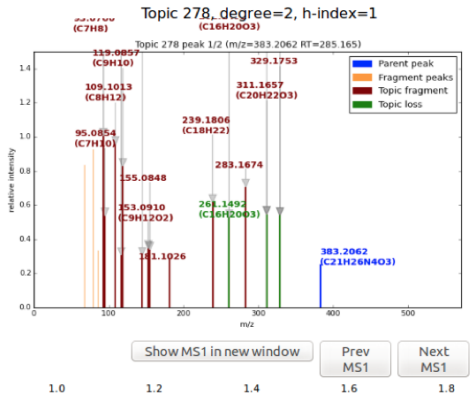
Id:	**Ms2ldaVis-07**
Requirement:	Parent Ion Plot – highlight mass loss between MS2 peak and MS1 Parent Peak.
Description:	The user should be able to place the cursor over any MS2 peak in the plot and it will display details of the mass loss between that peak and the MS1 parent peak.
Priority:	Should Have
Comments	
Scope:	Included

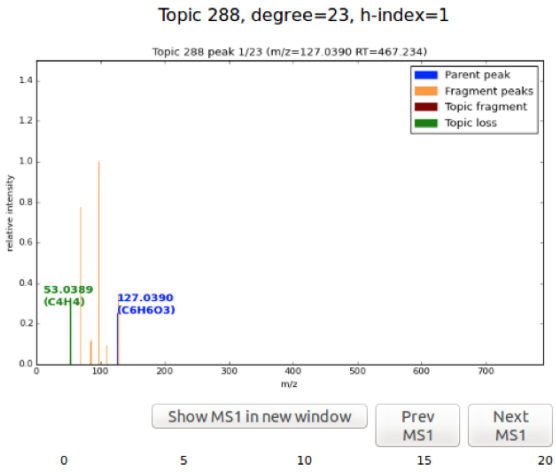
Id:	**Ms2ldaVis-08**
Requirement:	Parent Ion plot – update the data when a new MS1 Parent peak is selected on the MS2 Spectrum plot
Description:	If the user selects a new MS1 Parent peak on the Network Graph the information displayed on the Parent Ion Plot should be updated to reflect the data for that new parent peak.
Priority:	Should Have
Comments	
Scope:	Included

Id:	**Ms2ldaVis-09**
Requirement:	Parent Ion plot – distinguish overlapping MS2 peaks
Description:	If different mass2motif have the MS2 peak on display this peak should be coloured appropriately to reflect the overlap.
Priority:	Could Have
Comments	
Scope:	

Id:	**Ms2ldaVis-10**
Requirement:	Network Graph
Description:	A Network Graph should be displayed on the visualization screen showing the mass2motif nodes and associated MS1 parent peaks.
Priority:	Must Have
Comments	
Scope:	Included

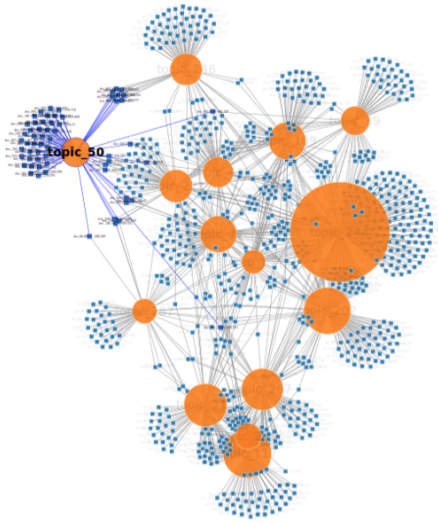
Id:	**Ms2ldaVis-11**
Requirement:	Multiple files
Description:	Ideally like to be able to load a file containing 20 files that have been analysed into the visualisation and carry out functionality that will compare these files identifying common features etc.
Priority:	Would Like To Have
Comments	
Scope:	Not Included

Id:	**Ms2ldaVis-12**
Requirement:	Topic graph (widget 1) - overlapping
Description:	Fix issues with overlapping and illegibility e.g. topic 278 below. 
Priority:	Could Have
Comments	
Scope:	Not applicable.

Id:	**Ms2ldaVis-13**
Requirement:	Topic graph (widget 1) - space
Description:	Optimise usage of space (too much white space in some topics) e.g. topic 288 
Priority:	Could Have
Comments	
Scope:	Not included.

Id:	**Ms2ldaVis-14**
Requirement:	Topic graph (widget 1) – external database feed
Description:	Include a button that will copy relevant information (all mass intensities for the topic) for input into an external databases such as mzcloud
Priority:	Could have
Comments:	
Scope:	Not Included

Id:	**Ms2ldaVis-16**
Requirement:	Network graph – make more user friendly
Description:	In general, make the graph more user friendly (size, colouring, density etc.). When highlighting a node sometimes it can be hard to see which other ones are actually connected.
Priority:	Would Like To Have
Comments:	
Scope:	Not included.

Id:	**Ms2ldaVis-17**
Requirement:	Network graph (widget 2) – grey out features
Description:	Grey out features that aren't highlighted (currently quite hard to read when a topic is selected as the irrelevant topics/documents are not transparent enough), e,g. 
Priority:	Would Like To Have
Comments:	To perhaps add a would-like feature: export good-quality picture (png or the like) of the network as the user coloured and selected it...
Scope:	Not Included.

Id:	**Ms2ldaVis-18**
Requirement:	Network graph (widget 2) – drag topic
Description:	Select a topic and be able to drag this to the side without it bouncing back to the centre.
Priority:	Could Have
Comments:	
Scope:	Not included

Id:	**Ms2ldaVis-19**
Requirement:	Network graph (widget 2) – degree slider
Description:	Move degree slider beside network animation.
Priority:	Should Have
Comments:	
Scope:	Not included

Id:	**Ms2ldaVis-21**
Requirement:	Network graph (widget 2) – colouring of nodes.
Description:	Colouring of nodes: ability to right click and node and this will colour everything has an edge from this selected node. (Does this functionality already exist?)
Priority:	Would Like To Have
Comments:	
Scope:	Not included

Id:	**Ms2ldaVis-22**
Requirement:	General layout – 3 different widgets
Description:	Will currently implement as 3 different widgets (topic widget, parent ion widget and graph widget) and will decide on layout later.
Priority:	Should Have
Comments:	
Scope:	Included

Id:	**Ms2ldaVis-23**
Requirement:	General layout – ability to select widgets
Description:	Implement some form of ability to select what widgets appear on screen (e.g. tick box options).
Priority:	Would Like To Have
Comments:	
Scope:	Not Included

Id:	**Ms2ldaVis-24**
Requirement:	Thresholding – allow it to be changed dynamically as part of the visualisation.
Description:	Currently thresholding is done once before visualisation. This should be changed so that it is done once and sliders are added to the screen so the thresholding can be altered and the visualisation refreshed.
Priority:	Should Have
Comments:	
Scope:	Not Included

Id:	**Ms2ldaVis-25**
Requirement:	Parent Ion plot – grey out peaks not associated with a mass2motif
Description:	The parent ion plot should display the complete MS2 Spectra and and peaks not associated with a mass2motif should be greyed out instead having a colour associated with it.
Priority:	Should Have
Comments	
Scope:	Included

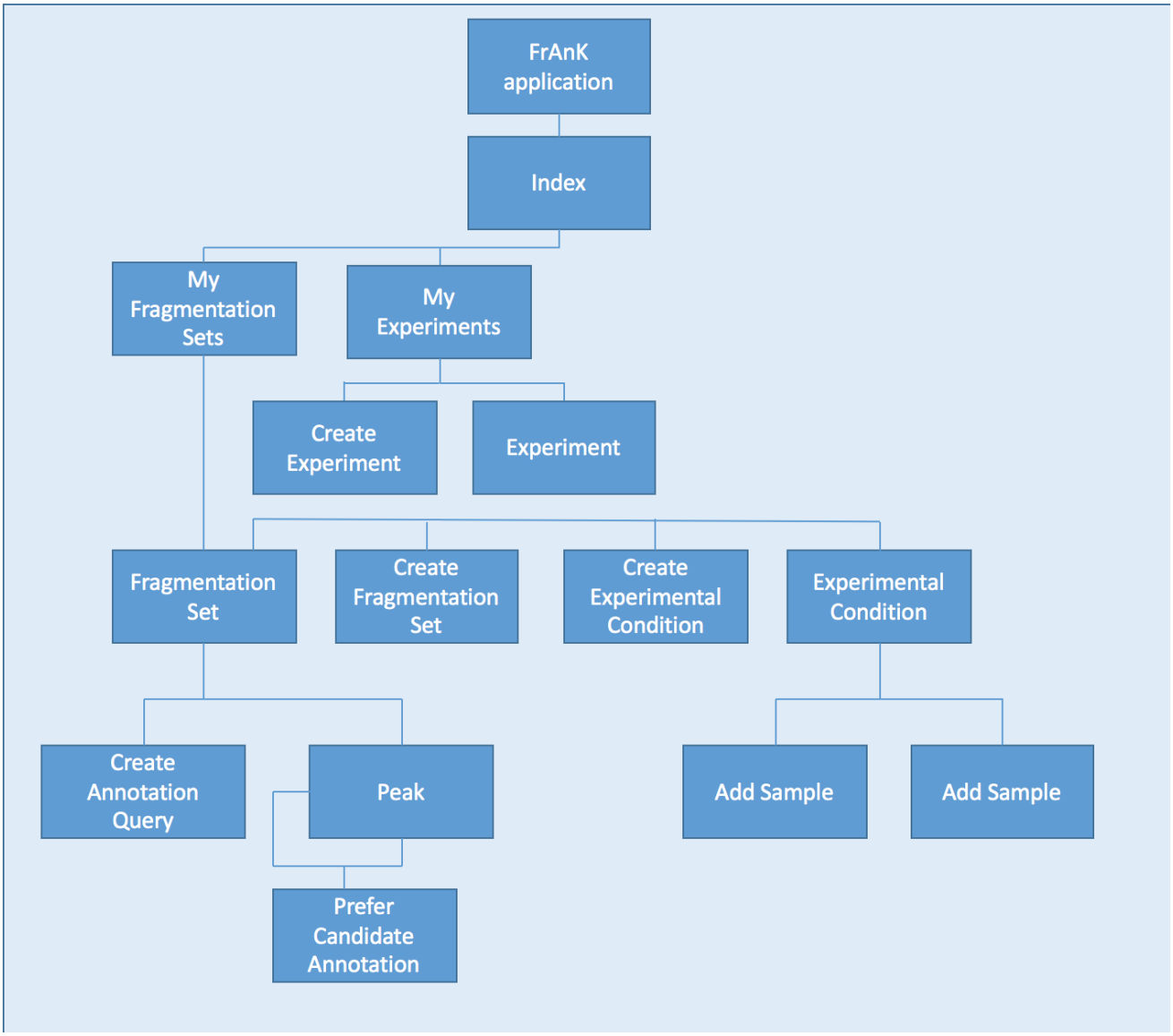
A.3 Non-Functional Requirements

Id:	**Ms2ldaNF-1**
Requirement:	The system should run on Ubuntu 15.04
Description:	
Priority:	Must Have
Comments:	The system was developed and tested this version of software.
Scope:	Included

Id:	**Ms2ldaNF-2**
Requirement:	The system should run on Mac OS X Yosemite
Description:	
Priority:	Would Like To Have
Comments:	
Scope:	Not Included

Appendix B

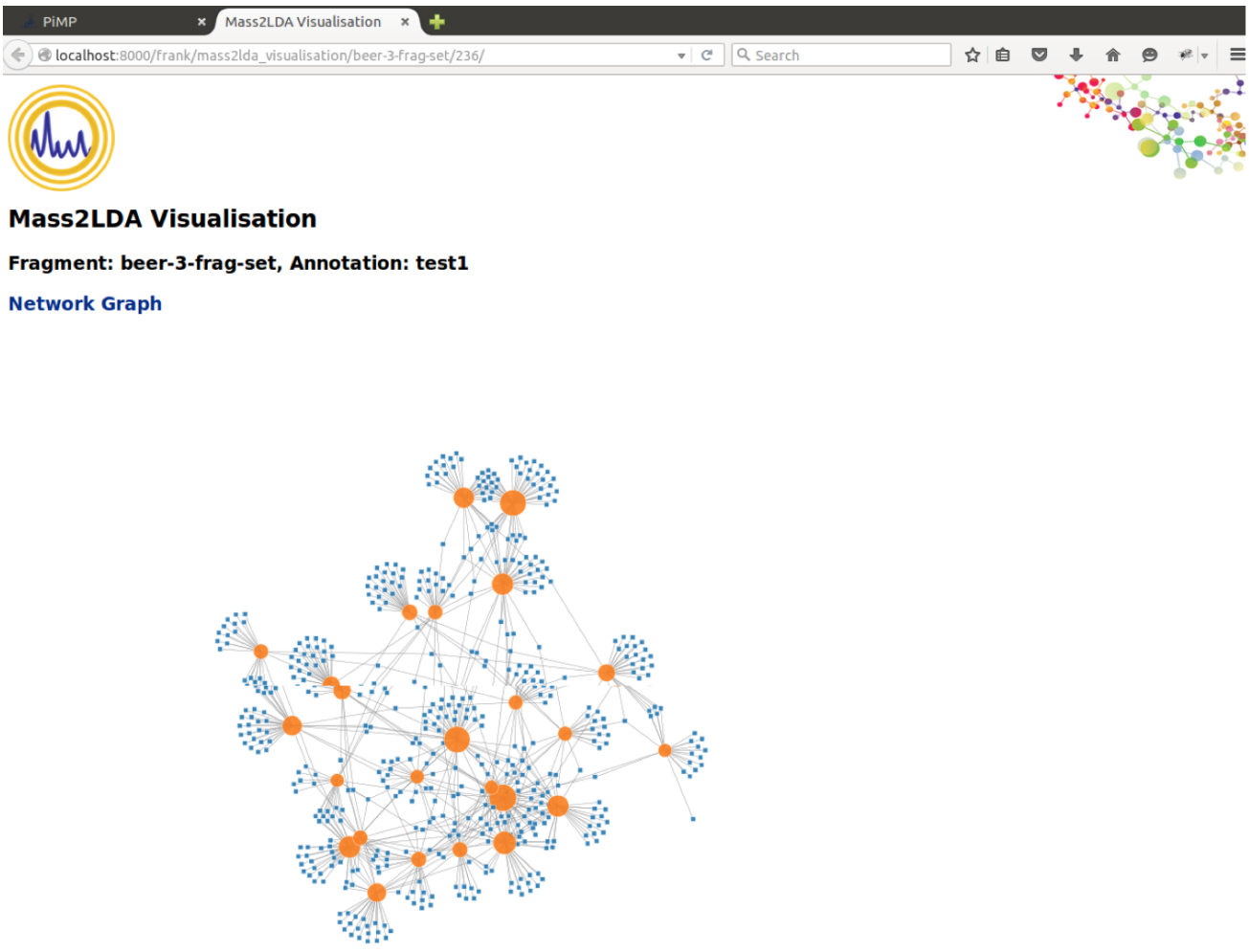
FrAnK Site Map



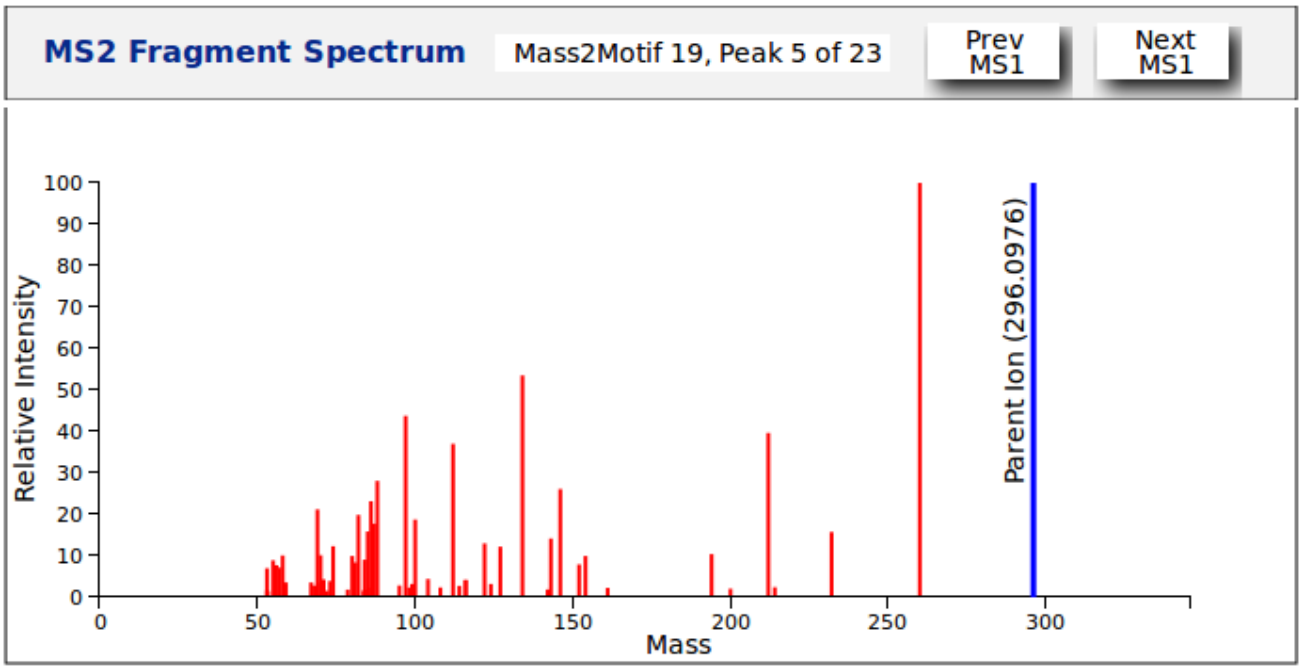
Appendix C

Visualisation Screens

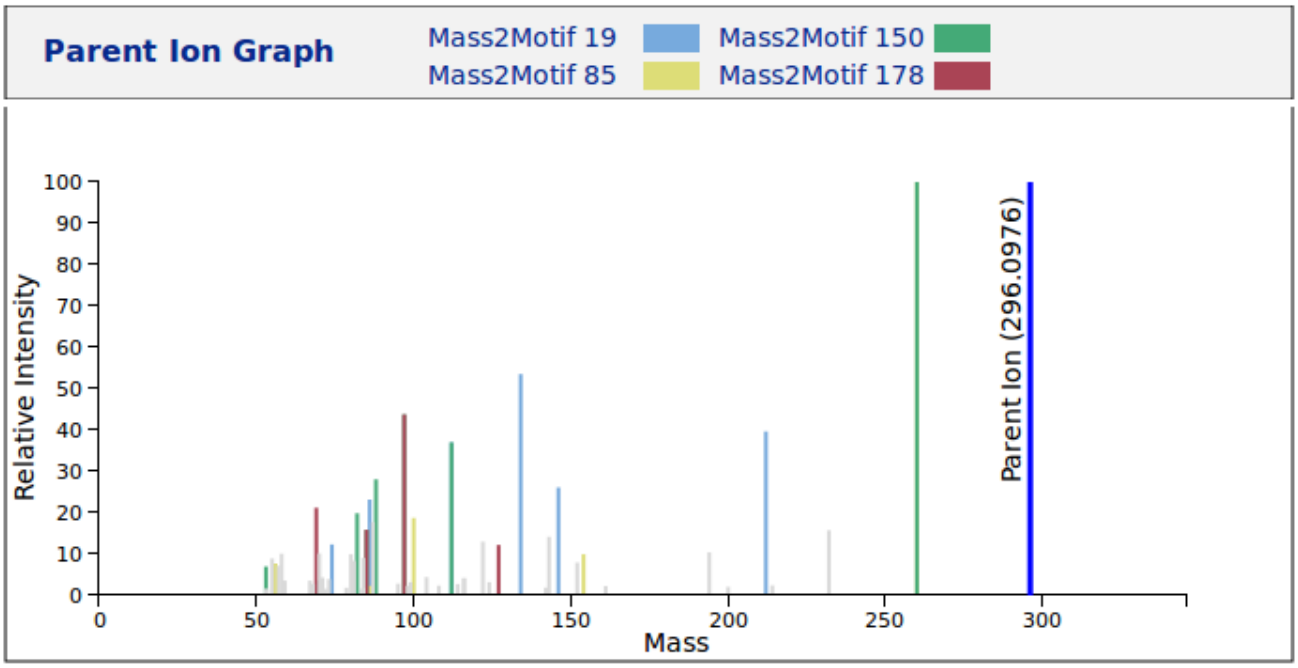
C.1 Network Graph



C.2 MS2 Spectrum Plot



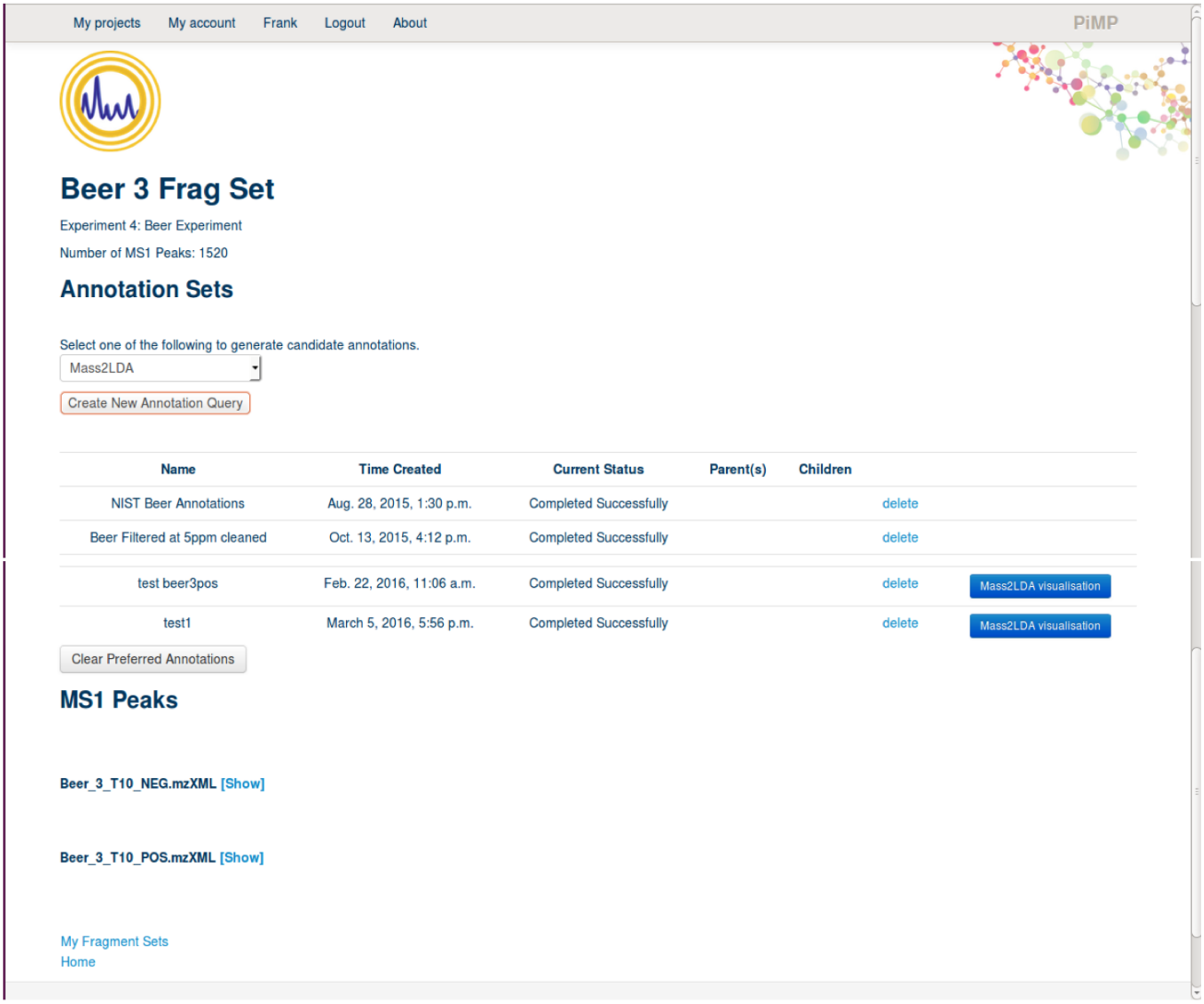
C.3 Parent Ion Plot



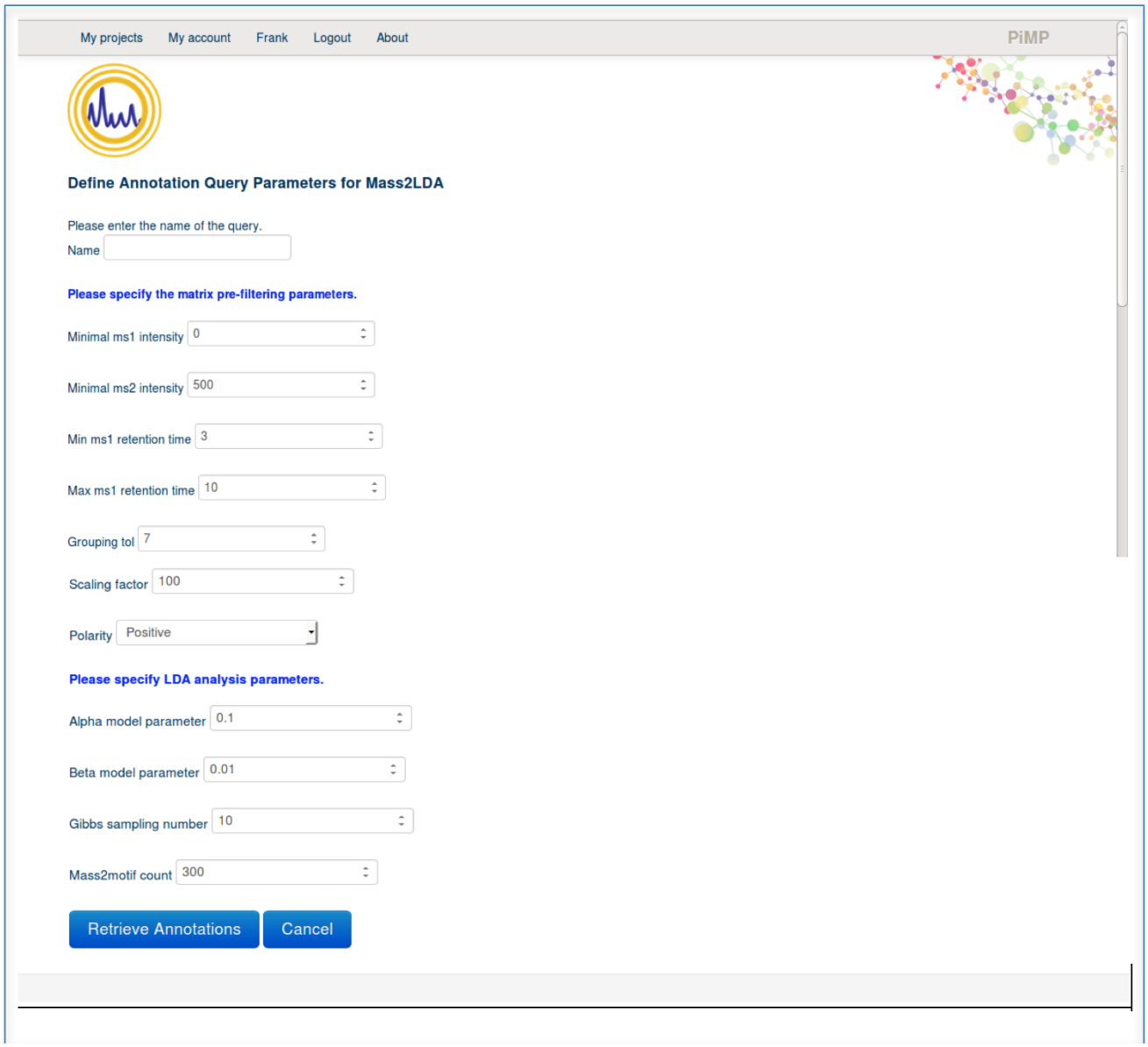
Appendix D

Forms

D.1 Fragmention Set



D.2 Mass2LDA Query Form



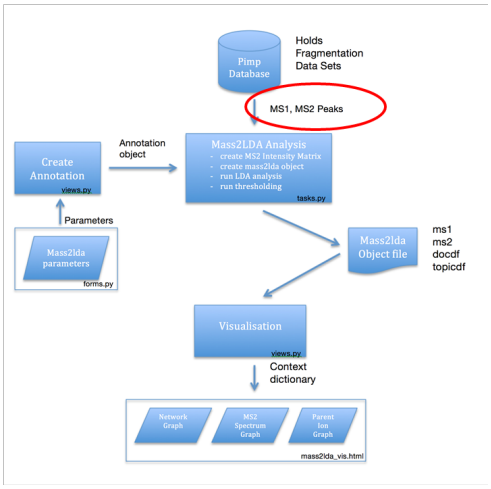
Appendix E

Data structures

E.1 MS1/MS2 data formats

DATA FORMATS – MS1, MS2 DATA

	peakID	MSnParentPeakID	msLevel	rt	mz	intensity
9091	9091	0	1	1378.24	421.1734714	405784.3125
9093	9093	0	1	1421.04	325.1775748	893404.875
9097	9097	0	1	1344.63	325.1775586	1414994.5
9101	9101	0	1	1397.97	325.1775343	2339749
9109	9109	0	1	1333.68	325.177568	1302905.875
9117	9117	0	1	1355.29	304.2481543	4683921
9120	9120	0	1	1355.29	287.221511	2314913.25
9123	9123	0	1	1145.81	265.1545303	408219.6875
9141	9141	0	1	1093.7	265.1545309	487218
9155	9155	0	1	1362.32	251.1599717	334116.7188
9171	9171	0	1	1332.44	250.0508503	1630042.625
9174	9174	0	1	1371.7	250.0508471	1987664.125

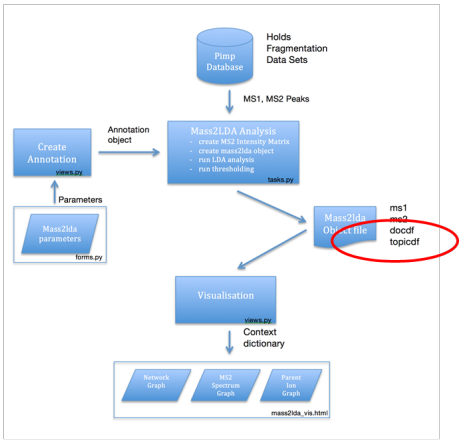


	peakID	MSnParentPeakID	msLevel	rt	mz	intensity	fragment_bin_id	loss_bin_id
9430	9430	9429	2	575.038	70.0650519	496833856	70.06505185	0
11103	11103	11094	2	603.782	104.107328	207204464	104.1073276	0
13743	13743	13742	2	528.379	58.0655513	107885408	58.06555131	0
9451	9451	9448	2	926.666	60.080932	74300856	60.08093195	0
11102	11102	11094	2	603.782	86.0965111	62708892	86.09651107	0
11101	11101	11094	2	603.782	125.000582	61858488	125.0005821	0
11100	11100	11094	2	603.782	98.9838235	44176116	98.98382347	0
15034	15034	15032	2	479.759	84.0442054	37366788	84.04420538	0
11099	11099	11094	2	603.782	60.0810425	33693744	60.08093195	0

E.2 docdf, topicdf data formats

DATA FORMATS – DOCDF, TOPICDF DATA

	120.08076_499.47_15660	121.06475_150.452_15061	121.06477_183.681_15099	121.06478_196.185_15102	121.0648_704.213_10801	121.06481_354.896_15048
0	0	0	0	0	0	0
1	0	0	0	0	0	0
2	0.219934239	0	0.562005158	0	0.14456866	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0
6	0	0	0	0	0	0
7	0	0	0	0	0	0
8	0	0	0	0	0	0
9	0	0.612033124	0	0	0	0
10	0	0	0	0	0	0



	M2M_0	M2M_1	M2M_2	M2M_3	M2M_4	M2M_5	M2M_6	M2M_7	M2M_8	M2M_9
fragment_53.01208	0	0	0	0	0	0	0	0	0	0
fragment_53.03889	0	0.01674455	0	0.01221202	0	0.02054148	0	0	0	0
fragment_53.08758	0	0	0	0	0	0	0	0	0	0
fragment_54.96366	0	0	0	0	0	0	0	0	0	0
fragment_54.97951	0	0	0	0	0	0	0	0	0	0
fragment_55.01819	0	0	0	0	0	0	0	0	0.014199	0
fragment_55.01858	0	0	0	0	0	0	0	0	0	0
fragment_55.02955	0	0	0	0	0	0	0	0	0	0
fragment_55.0422	0	0	0	0	0	0	0	0	0	0
fragment_55.05424	0	0	0	0	0	0	0	0	0	0
fragment_55.05469	0	0	0	0	0	0	0	0	0	0.04502309
fragment_55.18156	0	0	0	0	0	0	0	0	0	0

Bibliography

[1] Colour schemes. https://personal.sron.nl/ pault/colourschemes.pdf.

[2] Glasgow Polyomics Home Page. http://www.polyomics.gla.ac.uk.

[3] MassBank spectral database. http://www.massbank.jp/?lang=en.

[4] Overview of four 'omics' fields. https://www.ebi.ac.uk/training/online/course/introduction-metabolomics/what-metabolomics.

[5] Courant F, Antignac J-P, Dervilly-Pinel G, Le Bizec B. Basics of Mass Spectrometry Based Metabolomics. Proteomics. *Proteomics*, 14:2369–2388, 2014.

[6] David M. Blei. Probabilistic Topic Models. *Review articles*, 2012.

[7] Kerstin Scheubert, Franziska Hufsky, Sebastian Bocker. Computational mass spectrometry for small molecules. *Journal of Cheminformatics*, 5, 2013.

[8] Hoffmann Edmond de and Vincent Stroobant. Mass Spectrometry. . 2007.

[9] David M. Blei, Andrew Y. Ng, Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research 3*, 2003.

[10] Carson Sieverti, Kenneth E. Shirley. LDAvis: A method for visualizing and interpreting topics. *Proceedings of the Workshop on Interactive Language Learning*, 2014.