Utilities for Mass Spectrometry Analysis of Proteins

User's Manual

Version 2.2.0

May 2022

To download Utilities for Mass Spectrometry Analysis of Proteins visit:
www.umsap.nl

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Utilities for Mass Spectrometry Analysis of Proteins (UMSAP) is a graphical user interface (GUI) designed to speed up the post-processing of data obtained during mass spectrometry studies involving proteins. The program is not intended to analyze a mass spectrum or a mass chromatogram, neither to identify the peaks in a mass spectrum. The main objective is the fast post-processing of the vast amount of data generated in mass spectrometry experiments involving proteins after peak identification have been performed.

The program is organized in modules with each module performing a single type of data post-processing. The reason for this clear separation is the high dependency between the type of mass spectrometry experiment performed and the way in which the resulting data must be post-processed. The modules are designed in such a way that the required user input is minimized but still users can control every aspect of the analysis. Currently, the software contains three modules, but several others are already planned.

1.1 Citing Utilities for Mass Spectrometry Analysis of Proteins

If results obtained with UMSAP are published in any way, please acknowledge the use of UMSAP by including the following sentence:

"Utilities for Mass Spectrometry Analysis of Proteins was created by Kenny Bravo Rodriguez at the University of Duisburg-Essen and is currently developed at the Max Planck Institute of Molecular Physiology."

Any published work, which uses UMSAP, should include the following reference:

Kenny Bravo-Rodriguez, Birte Hagemeier, Lea Drescher, Marian Lorenz, Michael Meltzer, Farnusch Kaschani, Markus Kaiser and Michael Ehrmann. (2018). Utilities for Mass Spectrometry Analysis of Proteins (UMSAP): Fast post-processing of mass spectrometry data. Rapid Communications in Mass Spectrometry, 32(19), 1659–1667.

Electronic documents should include a direct link to the official web page of UMSAP at: www.umsap.nl

1.2 Acknowledgments

I would like to thank all the persons that have contributed to the development of UMSAP, either by contributing ideas and suggestions or by testing the code. Special thanks go to: Dr. Farnusch Kaschani, Dr. Juliana Rey, Dr. Petra Janning and Prof. Dr. Daniel Hoffmann.

In particular, I would like to thank Prof. Dr. Michael Ehrmann.

Chapter 2

Obtaining and Installing Utilities for Mass Spectrometry Analysis of Proteins

2.1 Obtaining Utilities for Mass Spectrometry Analysis of Proteins

UMSAP is distributed free of charge for anyone interested in using it. To obtain a copy of the software just register at www.umsap.nl and go to the Download page.

No extra software or packages are needed for UMSAP to properly work. So far, UMSAP have been tested in macOS 10.14.6 and 12.3 and Windows 10. Support for some Linux distributions will be available in the future.

2.2 Installing Utilities for Mass Spectrometry Analysis of Proteins

Windows

Unzip the file you just downloaded from www.umsap.nl. Then, copy the folder UMSAP to the location in your file system where you want to keep it. Finally, create a shortcut to the executable file UMSAP.exe found inside the main folder UMSAP. That is all. You are now ready to use UMSAP.

macOS

Unzip the file you just downloaded from www.umsap.nl. Then, just move the UMSAP.app folder to /Applications/. That is all. You are now ready to use UMSAP.

Depending on the security settings in macOS, it may be needed to explicitly allow UMSAP to be opened the first time the app is used.

2.3 Uninstalling Utilities for Mass Spectrometry Analysis of Proteins

UMSAP will not create any installation file in your computer. Therefore, the only thing you need to do, to completely uninstall UMSAP, is to delete the folder UMSAP.app in macOS or UMSAP in Windows. You should also delete any shortcut pointing to the executable file of UMSAP and the configuration file .umsap_config.json in your home folder. That is all.

Chapter 3

Workflow in Utilities for Mass Spectrometry Analysis of Proteins

When you start UMSAP, the program will display the main window (Figure 3.1). From this window you can access all the modules and utilities either by the menu entries: Modules and Utilities or by the corresponding buttons on the right side list. A complete description of each module and utility is given in the following chapters.

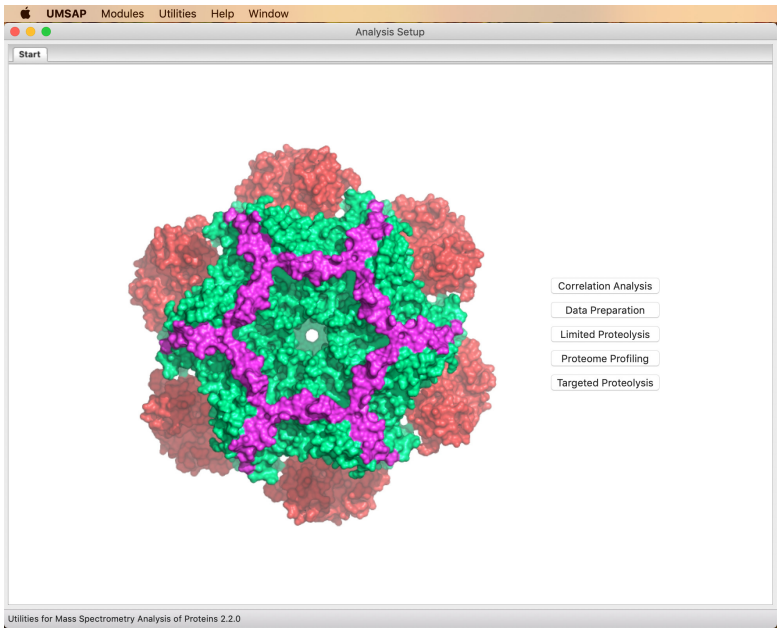

Figure 3.1: The main window of UMSAP. From this window users can access all the available Modules and Utilities.

3.1 The input files

UMSAP has two main input files. One file contains the detected peptide sequences after all peak assignments have been completed, and the other file contains the detected proteins. The program expects these files to be plain text files containing a table with the data. Columns in the files are expected to be tab separated. The first row in the files is expected to contain only the names of the columns. There is no limit in the amount and type of data present in the Data files. However, each module will expect certain columns to be present. Columns not needed by the modules will simply be ignored.

In addition, certain modules use other input files as well. The modules Targeted Proteolysis and Limited Proteolysis use fasta files containing the sequences of the recombinant and native proteins used in the experiments. The first sequence found in the fasta file is assumed to be the sequence of the recombinant protein. The second sequence found in the fasta file is assumed to be the sequence of the native protein. All other sequences found in the fasta file are discarded. If the sequence of the native protein is given UMSAP performs a sequence alignment between the native and recombinant sequences. The alignment allows UMSAP to translate the results obtained with the residue numbers of the recombinant protein to the residue numbers of the native protein. This is done to facilitate future comparison of results between different recombinant proteins of the same native protein. However, when analyzing the results of the alignment UMSAP assumes that the recombinant and native sequences differs only in the N and C-terminal tags while the sequence between the tags is identical. If this is not the case, e.g. there are point mutations or insertion/deletion in the sequence of the recombinant protein no native sequence file should be given to UMSAP.

The Targeted Proteolysis module may also use a local PDB file.

3.2 The output files

Results generated by UMSAP will be saved in two folders and a file with extension .umsap (Figure 3.2). Direct manipulation of the umsap file and files within these folders should be avoided. UMSAP provides a way to manage them through the UMSAP Ctrl window (Chapter 4). Nevertheless, all the files created by UMSAP are plain text files with json or cvs (tab separated) format, in order for users to be able to read their content. Changing the content of the files is highly discouraged as this will lead to errors in the reliability and visualization of the results with UMSAP.

The folder Input_Data_Files contains a copy of the input files used for the analysis in the project. When adding a new analysis to the project, the new input files used will be copied to the Input_Data_Files folder. The date and time of the analysis will be added to the name of the file to avoid overwriting existing files inside the folder.

The folder Steps_Data_Files contains a folder for each analysis in the project. These folders contain the main results for the analysis as well as a step by step account of the calculations and any further analysis performed after the main results were created.

The .umsap file contains information about all the analysis in the project and allows

managing the project and the visualization of the results. An unlimited number of analysis can be added to any given .umsap file. UMSAP will never overwrite or replace an .umsap file, instead new analysis will be added to the selected .umsap file.

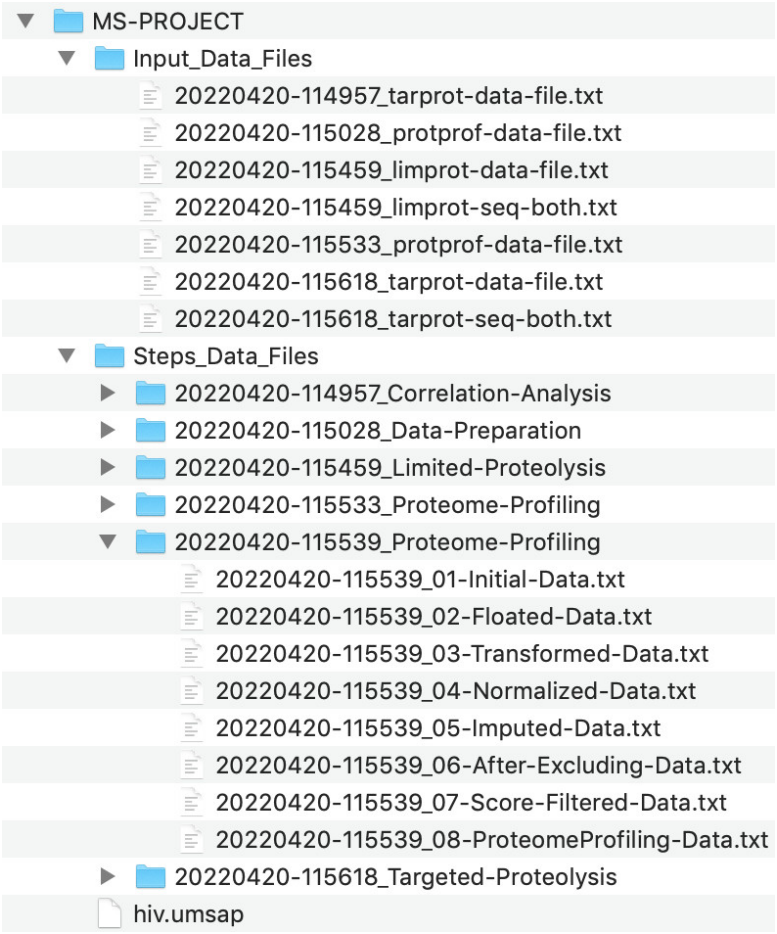

Figure 3.2: Structure of the output generated by UMSAP. Results are saved in the Steps_Data_Files folder. The .umsap file allows managing and visualizing the results.

3.3 Using Utilities for Mass Spectrometry Analysis of Proteins

Once the input files are ready to be analyzed, using UMSAP is straightforward. Just open the program and select a module or utility. In the new tab, fill in the needed information and hit the Start Analysis button at the bottom of the tab. Depending on the amount of data and the complexity of the analysis to perform it may take a few minutes for the program to complete the task at hand. While the analysis is running, a window, containing a progress bar, will appear. This window will give a rough guess of the remaining time needed to complete the current analysis and will report any error

encountered. It will be helpful if users send a crash report to umsap@umsap.nl, so we can correct them.

In order to make the program as user-friendly as possible help messages will pop up from buttons and labels. The help messages will contain a brief description of what is the button or label for and what input is expected from the user. In this way, users can find basic information about a particular element of the interface without needing to go to the manual or online tutorials. If more information is needed, users may consult the manual or click the Help button at the bottom of the module/utility tab to read an online tutorial.

Depending on the module or utility just run, new windows will be created to show a graphical representation of the results. All plots support to zoom into a rectangular selection of the plot and to reset the zoom level.

3.4 Navigating through Utilities for Mass Spectrometry Analysis of Proteins

The entries Modules and Utilities will be available in the menu of every window. The Modules entry in the menu gives direct access to all modules. The same is true for the Utilities entry. These menu entries are the fastest way to access all the functions in UMSAP. In a typical UMSAP session, users will work with different independent windows simultaneously. The windows have descriptive names, so users can quickly guess the content of any window. The scheme of the windows name is *File Name - Utilities or Module Name - ID of the Analysis*. For example, the window with name *hiv.umpap - Target Proteolysis - 20220420-115618 - Cleavage Sites* will be displaying the Targeted Proteolysis analysis with ID 20220420-115618 - Cleavage Sites from file hiv.umsap.

A list of current shortcuts is given in Table 3.1.

3.5 Backward compatibility

Unfortunately, UMSAP 2.2.0 is not capable to read any file generated with previous versions of UMSAP.

Shortcut	Action	Window
Alt+Cmd+L	Create the Limited Proteolysis tab	All
Alt+Cmd+P	Create the Proteome Profiling tab	All
Alt+Cmd+T	Create the Targeted Proteolysis tab	All
Cmd+R	Read umsap file	All
Cmd+C	Copy	Text and List boxes
Cmd+X	Cut	Text and List boxes
Cmd+V	Paste	Text and List boxes
Cmd+A	Select all	Text and List boxes
Cmd+P	Show Data Preparation results	Results plot
Cmd+D	Duplicate result window	Results plot
Cmd+E	Export data	Results plot
Cmd+I	Export image	Results plot
Cmd+K	Clear all selections	Results plot
Cmd+A	Add analysis	UMSAP Ctrl
Cmd+X	Delete analysis	UMSAP Ctrl
Cmd+E	Export analysis	UMSAP Ctrl
Cmd+U	Reload file	UMSAP Ctrl
Cmd+Z	Reset the zoom on a plot	Selected plot
Alt+Shift+I	Export all images	Multiple plots
Alt+Shift+Z	Reset all zooms	Multiple plots
Shift+I	Export main plot image	Multiple plots
Shift+Z	Reset main plot zoom	Multiple plots
Alt+I	Export secondary plot image	Multiple plots
Alt+Z	Reset secondary plot zoom	Multiple plots
Cmd+A	Show all peptides	Limited Proteolysis
Cmd+L	Toggle Band/Lane selection mode	Limited Proteolysis
Cmd+S	Export sequence alignments	Limited Proteolysis
Shift+A	Add label to Volcano plot	Proteome Profiling
Shift+P	Toggle Pick label / Select protein	Proteome Profiling
Shift+Cmd+A	Apply all Filters	Proteome Profiling
Shift+Cmd+F	Auto apply all Filters	Proteome Profiling
Shift+Cmd+R	Remove selected Filters	Proteome Profiling
Shift+Cmd+Z	Remove last applied Filter	Proteome Profiling
Shift+Cmd+X	Remove all Filters	Proteome Profiling
Shift+Cmd+C	Copy Filters	Proteome Profiling
Shift+Cmd+V	Paste Filters	Proteome Profiling
Shift+Cmd+S	Save Filters	Proteome Profiling
Shift+Cmd+L	Load Filters	Proteome Profiling
Shift+Cmd+E	Export filtered data	Proteome Profiling
Cmd+S	Export sequence alignments	Targeted Proteolysis

Table 3.1: List of built-in keyboard shortcuts. Windows users should replace Cmd with Ctrl.

Chapter 4

UMSAP Control

The UMSAP Control windows shows the content of an .umsap file (Figure 4.1).

4.1 The interface

The analysis contained in the selected .umsap file are displayed in alphabetical order and grouped by the analysis type. The checkboxes to the left of the names of the Utilities and Modules allow creating the corresponding window showing the results available for the selected Utility or Module.

Each analysis in the file is represented by the user-provided Analysis ID. Unfolding any ID will display all the configuration values provided by the user prior to running the analysis. In addition, a left click over any Analysis ID will create the corresponding tab in the Analysis Setup window (Figure 3.1) and populate all fields with the values in the selected analysis. This is the fastest way to configure the analysis tab to rerun an analysis with slight changes in the configuration options. After rerunning an analysis or simply adding a new analysis to the .umsap file, the window will be automatically updated to display the new results.

4.2 The Tools menu

The UMSAP Control windows allows also to manage the content of the selected .umsap file. Currently, it is possible to Add (Cmd+A) analysis from a different .umsap file, to Delete (Cmd+X) analysis from an existing .umsap file and to Export (Cmd+E) the analysis in an .umsap file to a new .umsap file.

Adding analysis from an .umsap file to the already opened .umsap file will result in the addition of the new information to the already opened .umsap file and in the copy of the necessary files and folders to folders Input_Data_Files and Steps_Data_Files. During this process there is a small chance to end up with duplicated file and/or folder names or Analysis ID. In this case, UMSAP will rename the file/folder/Analysis ID to avoid any overwriting and will update any reference in the .umsap file to the files/folders that

were renamed.

Deleting any analysis from an .umsap file will also result in the removal of the files and/or folders referenced in the deleted analysis. Files in Input_Data_Files are only deleted if they are not referenced by any remaining analysis. Deleting all analysis in an .umsap file will result in the removal of the .umsap file and folders Input_Data_Files and Steps_Data_Files. If the folder containing the project is empty after deleting all UMSAP files and folders the project folder is also deleted.

Exporting some or all analysis in an opened .umsap file to an already existing .umsap file is not possible. When exporting the selected analysis to a project folder containing an Input_Data_Files and/or Steps_Data_Files folder, UMSAP will create a new folder in the selected project folder and export all the information to this empty folder.

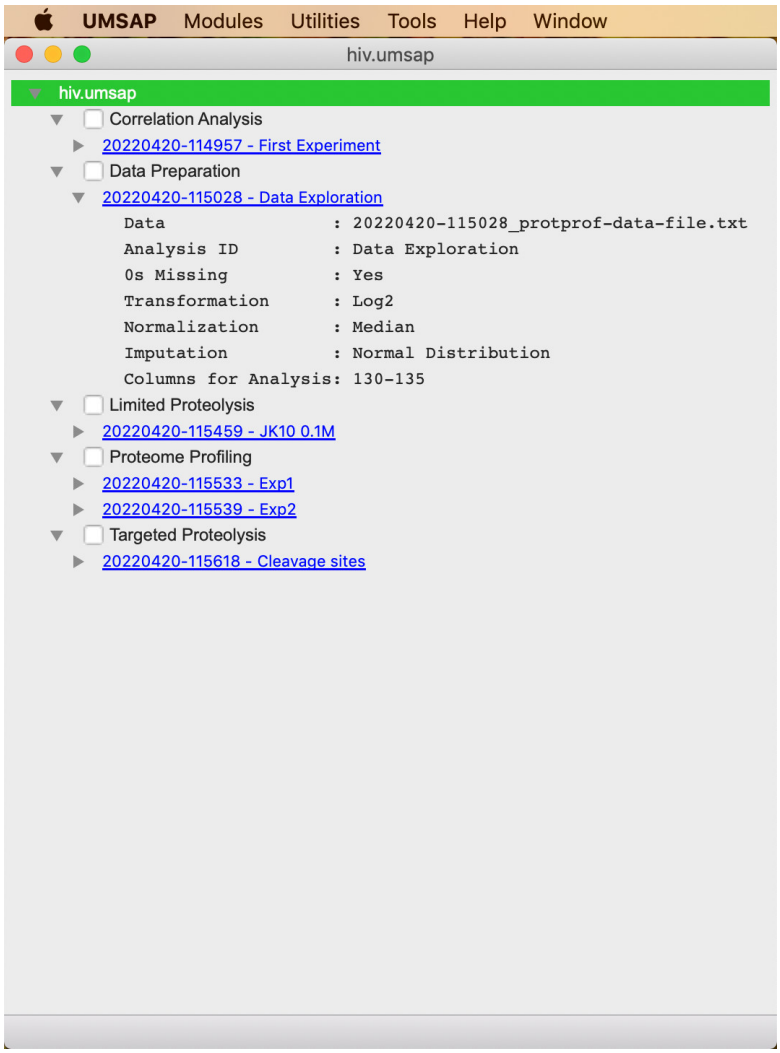

Figure 4.1: The UMSAP Control window. The content of the selected .umsap file is shown in alphabetically order. The window allows managing the content of the .umsap file and to visualize the results of the analysis in the file.

Chapter 5

Correlation Analysis

The Correlation Analysis utility calculates the correlation in the MS data used as input for UMSAP.

5.1 The interface

The Correlation Analysis tab is divided in four regions (Figure 5.1).

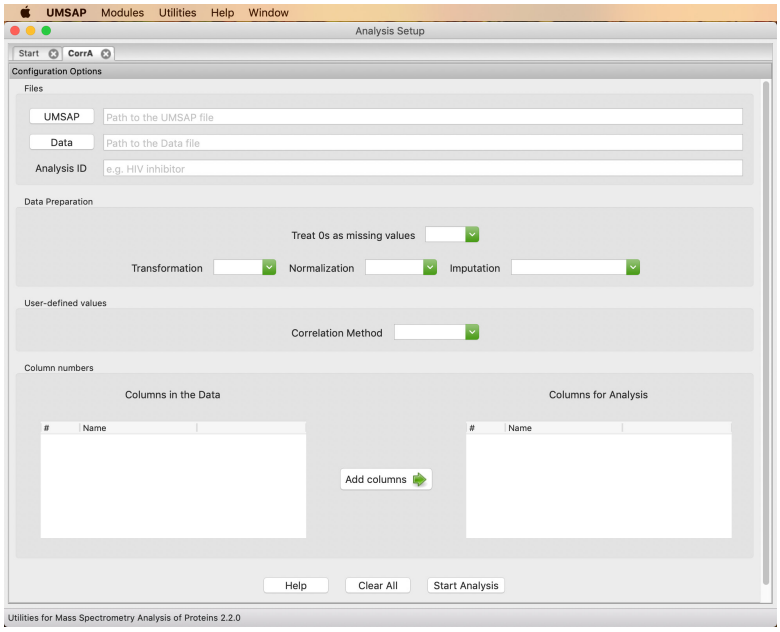

Figure 5.1: The Correlation Analysis tab. This tab allows to perform a correlation analysis of the data contained in a given Data file.

Region Files contains two buttons and a text field. Here users select the input and output files for the analysis.

1. The button UMSAP allows users to browse the file system to select the location and

name of the .umsap file. When selecting an already existing .umsap file the operating system will ask if it is ok to replace the file, the answer can be yes since UMSAP will never overwrite or replace an .umsap file, instead the new analysis will be added to the already existing file. Only .umsap files can be selected here.

2. The button Data allows users to browse the file system to select the input data file that will be used for the analysis. The Data file is expected to be a plain text file with tab separated columns and the name of the columns in the first row of the file. In addition, columns to be analyzed must contain only numbers and must be of the same length. Only .txt files can be selected here.

3. The text field Analysis ID allows users to provide an ID for the analysis to be run. The date and time of the analysis will be automatically added to the beginning of the name.

Region Data Preparation contains four dropdown boxes. Here users select how the data in the Data file should be prepared before starting the analysis.

1. The dropdown Treat 0s as missing values allows user to define how to handle 0 values present in the Data file.

2. The dropdown Transformation allows user to select the Transformation method to be applied to the data.

3. The dropdown Normalization allows users to select the Normalization method to be applied to the data.

4. The dropdown Imputation allows user to select the Imputation method used to replace missing values in the data.

Region User-defined values contains one dropdown box.

1. The dropdown Correlation Method allows users to select the correlation method to use.

Region Column numbers contains two lists and a button. Here users select the columns in the Data file to be used in the Correlation Analysis.

1. The list to the left will display the names of the columns present in the selected Data file. The list is automatically filled once the Data file is selected. Rows in the list can not be deleted, except in the case of loading a different Data file or using the Clear All button at the bottom of the tab. Selected rows can be copied with the Cmd+C shortcut.

2. The list to the right will contain the columns in the Data file that will be used for the Correlation Analysis. This list must contain at least two rows for the analysis to proceed. Selected rows in this list can be deleted with the Cmd+X shortcut and rows already copied from the list in the left can be pasted with the Cmd+V shortcut. While pasting the rows, duplicate rows will be silently discarded. Importantly, the order of the rows and columns in the matrix containing the correlation coefficients will be the same as the order of the columns in this list. Therefore, users are advised to fill the list in such a way that replicates of the same experiment are consecutive to each other in the list.

3. The button Add columns will add the selected rows in the left list to the right list. The rows will be added to the right list in the same order as they are selected in the left list. Duplicate rows will be silently discarded.

The bottom of the tab contains three buttons.

1. The button Help leads to an online tutorial about Correlation Analysis in UMSAP.
2. The button Clear All will delete all user input from the tab.
3. The button Start Analysis starts the Correlation Analysis.

5.2 The analysis

First, UMSAP will check the validity of the user provided input. Then, columns in the right list are read from the Data file. The columns must contain only numbers and the same amount of rows must be found in all columns. Failing to comply with this will result in the program aborting the analysis. After this, all steps selected in the Data Preparation region are carried out (Chapter 6). Finally, the correlation coefficients are calculated using the selected method. If any of the coefficients cannot be calculated, then the corresponding coefficient is set to NA.

5.3 The result window

The correlation coefficients resulting from a Correlation Analysis will be shown as a color coded matrix (Figure 5.2). Values between -1 to 0 will be shown in shades of red, 0 will be shown as white and values between 0 to 1 will be shown in shades of blue. NA values will be shown in green. The columns and rows of the matrix are the column names used to calculate the correlation coefficients. Information about a specific matrix element can be obtained by simply placing the mouse pointer over the matrix element.

5.4 The Tools menu

The Tools menu in the window showing the correlation coefficients allows user to view any of the Correlation Analysis contained in the selected .umsap file or to modify the appearance of the displayed plot. For example, the column numbers can be displayed instead of the column names or the color bar can be hidden. In addition, only a subset of the columns can be shown using the Select Columns entry.

The Tools menu also allows duplicating the window (Cmd+D) for easier comparison of two or more analysis, checking the Data Preparation steps of the analysis (Cmd+P), creating an image of the plot (Cmd+I), exporting the correlation coefficient matrix to a tab separated CSV file (Cmd+E) and resetting the zoom level of the plot (Cmd+Z).

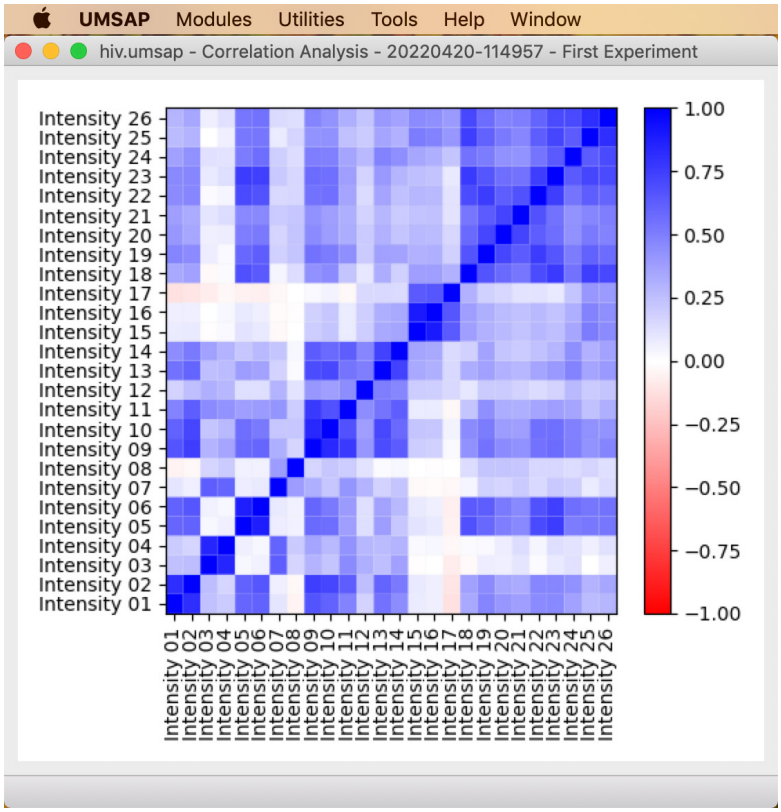

Figure 5.2: The Correlation Analysis result window. The correlation coefficients are shown as a color coded matrix. Values between −1 to 0 are shown in shades of red, 0 is shown in white and values between 0 to 1 in shades of blue. NA values are shown in green.

Chapter 6

Data Preparation

The Data Preparation utility allows exploring the distribution of the data in the selected Data File and the impact that different Data Preparation options have over the data.

6.1 The interface

The Data Preparation tab is divided in two main regions (Figure 6.1).

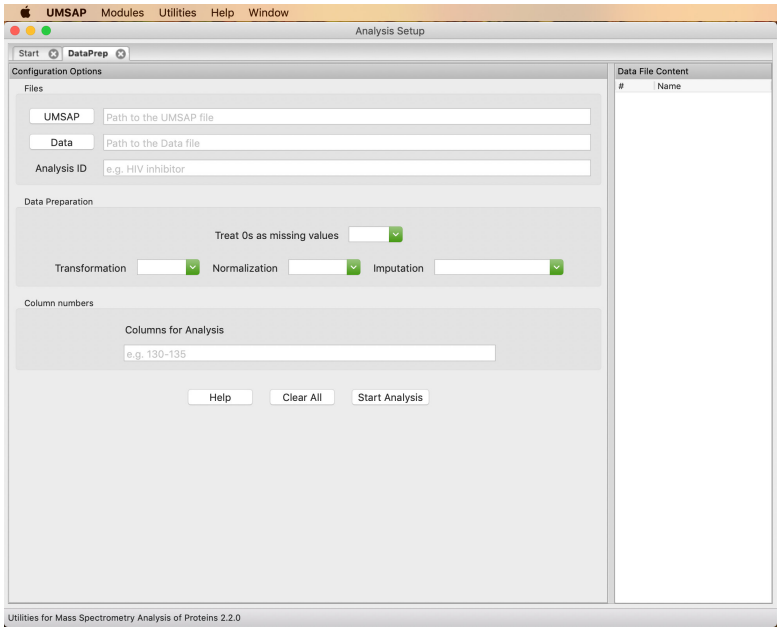

Figure 6.1: The Data Preparation tab. This tab allows to perform a statistical exploration of the data contained in a given Data file.

The Data File Content region holds only a list to show the name of the columns in the selected Data File. The list will be automatically filled after selecting the file.

The Configuration Options region contains all the fields needed to configure and run

the analysis.

Section Files contains two buttons and a text field. Here users select the input and output files for the analysis.

1. The button UMSAP allows users to browse the file system to select the location and name of the .umsap file. When selecting an already existing .umsap file the operating system will ask if it is ok to replace the file, the answer can be yes since UMSAP will never overwrite or replace an .umsap file, instead the new analysis will be added to the already existing file. Only .umsap files can be selected here.

2. The button Data allows users to browse the file system to select the input data file that will be used for the analysis. The Data file is expected to be a plain text file with tab separated columns and the name of the columns in the first row of the file. In addition, columns to be analyzed must contain only numbers and must be of the same length. Only .txt files can be selected here.

3. The text field Analysis ID allows users to provide an ID for the analysis to be run. The date and time of the analysis will be automatically added to the beginning of the name.

Section Data Preparation contains four dropdown boxes. Here users select how the data in the Data file should be prepared before starting an analysis.

1. The dropdown Treat 0s as missing values allows user to define how to handle 0 values present in the Data file.

2. The dropdown Transformation allows user to select the Transformation method to be applied to the data.

3. The dropdown Normalization allows users to select the Normalization method to be applied to the data.

4. The dropdown Imputation allows user to select the Imputation method used to replace missing values in the data.

Section Column numbers contains a text field. Here users specify the Columns in the Data File to be used during the Data Preparation steps. Only integers can be accepted here. Column numbers can be copied (Cmd+C) and paste (Cmd+V) from the selected rows in the list on region Data File Content or just type the numbers.

The bottom of the region contains three buttons.

1. The button Help leads to an online tutorial about Correlation Analysis in UMSAP.

2. The button Clear All will delete all user input from the tab.

3. The button Start Analysis starts the Correlation Analysis.

6.2 The analysis

First, UMSAP will check the validity of the user provided input. After this, the selected Data File is read and the following steps are taken:

1. The content of all specified columns in Data File is checked to make sure only numbers are found in them. 0 values present in the columns are left or remove depending on the selected value for field Treat 0s as missing values.
2. The indicated Transformation method is applied to the selected columns.
3. The indicated Normalization method is applied to the transformed data.
4. The indicated Imputation method is applied to the normalized data.

The results from the four steps is saved, so users can check the effect of the selected workflow over the data. Currently, only one method is implemented for the Transformation, Normalization and Imputation of the Data, respectively. The only alternative is to skip the corresponding step. The methods available will be expanded in the near future. All steps are column wise applied.

6.3 The result window

The window showing the results from a Data Preparation workflow is divided in three regions (Figure 6.2).

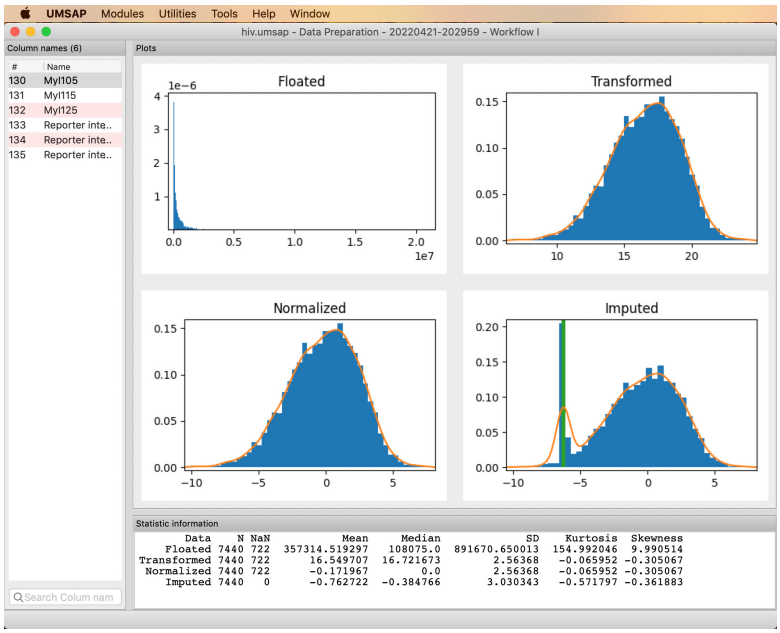

Figure 6.2: The Data Preparation result window. Histograms for the initial, transformed, normalized and imputed data are shown.

Region Column names shows a table with the number (0 based) and name of the analyzed columns.

Region Plots shows the results from the Data Preparation workflow in four histograms for the selected column in region Column names. The histograms are created for the initial, transformed, normalized and imputed data. They show the probability density as blue bars and the calculated probability density function in orange. The green bars

in the Imputed histogram represent the imputed values.

Region Statistic information shows a description of the data for the selected column in region Column names.

6.4 The Tools menu

The Tools menu in the window showing the results from a Data Preparation workflow allows user to view any of the Data Preparation analyzes contained in the selected .umsap file. The Tools menu also allows duplicating the window (Cmd+D) for easier comparison of two or more set of results, creating an image of the plots (Alt+Shift+I), exporting the data shown to a tab separated CSV file (Cmd+E) and resetting the zoom level of the plots (Alt+Shift+Z).

Chapter 7

Limited Proteolysis

The Limited Proteolysis module is designed to post-process the results from an enzymatic digestion performed in two steps. The first step is assumed to be a limited proteolysis in which a large protein is split in smaller fragments. The fragments are then separated using a SDS-PAGE electrophoresis. Finally, selected bands from the gel are submitted to a full enzymatic digestion and the generated peptides are analyzed using mass spectrometry.

The main objective of the module is to identify the protein fragments generated in the initial limited proteolysis from the peptides found in the MS analyzed gel spots. This is achieved by performing an equivalence test(*1*) between the peptides in the selected gel spots and a control spot containing the full length target protein. In this way, peptides leaked from one gel spot to another can be eliminated. Several replicates of the experiment are expected.

7.1 Definitions

Before explaining in detail the interface of the module and how does the module work, let's make clear the meaning of some terms that will be used in the following paragraphs.

- *Recombinant protein*: actual amino acid sequence used in the mass spectrometry experiments. It may be identical to the native sequence of the Target protein under study or not.
- *Native protein*: full amino acid sequence expressed in wild type cells.
- *Detected peptide*: any peptide detected in any of the mass spectrometry experiments including the control experiments.
- *Relevant peptide*: a detected peptide with a Score value above a user defined threshold, see page 38.
- *Filtered peptide*: a relevant peptide with equivalent intensities in the control and a given gel spot at the chosen significance level.

- *Fragment*: group of filtered peptides with no gaps when their sequences are aligned to the sequence of the recombinant/native protein.

For example, there are three fragments in the alignment shown below. The first fragment is formed by sequences 1 to 3 since there is no gap in the sequence MKKTAIAIAVAL. SEQ4 forms the second fragment because there is a gap between the last residue in SEQ3 and the first residue in SEQ4 and another gap between the last residue in SEQ4 and the first residue in SEQ5. For the same reason SEQ5 forms the third fragment.

```
REC.PROT  MKKTAIAIAVALAGFATVAQAASWSHPQFEKIEGRRDRGQKTQSAPGTLS    50
SEQ1      MKKTAIAIAV..........................................    10
SEQ2      ..KTAIAIAV..........................................     8
SEQ3      .....IAIAVAL........................................     7
SEQ4      ..............ATVAQAASWS............................    10
SEQ5      ....................................DRGQKTQSAPG...    11
```

7.2 The input files

The Limited Proteolysis module requires a Data file containing the detected peptides and a sequence file containing the amino acid sequence of the recombinant protein used in the study. Both files must follow the guidelines specified in Section 3.1. In short, the Data file must have a tabular format with tab separated columns and the name of the columns are expected as first row. The Sequence file is expected to contain at least one sequence and to be FASTA formatted. If more than one sequence is found in the Sequence file the first sequence will be taken as the sequence of the Recombinant protein and the second sequence will be taken as the sequence of the Native protein. All other sequences are discarded.

7.3 The interface

The tab of the Limited Proteolysis module is divided in two regions (Figure 7.1).

The Data File Content region holds only a table to show the name of the columns in the selected Data File. The table will be automatically filled after selecting the file. Selected rows in the table can be copied (Cmd+C) and pasted (Cmd+V) to the text fields in region Configuration Options.

The Configuration Options region contains all the fields needed to configure and run the analysis.

Section Files contains three buttons and a text field. Here users select the input and output files for the analysis.

1. The button UMSAP allows users to browse the file system to select the location and name of the .umsap file. When selecting an already existing .umsap file the operating system will ask if it is ok to replace the file, the answer can be yes since UMSAP will

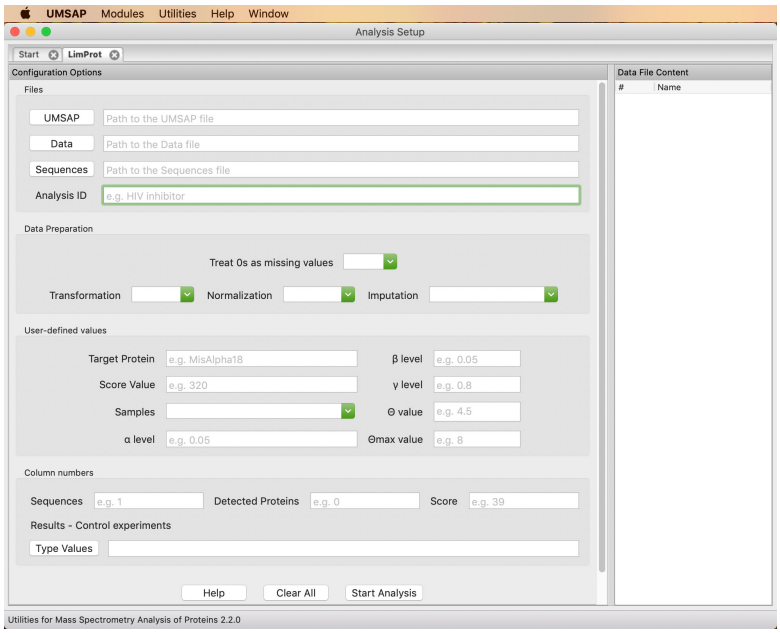

Figure 7.1: The Limited Proteolysis module tab. This tab allows users to perform the analysis of the results obtained during a two steps enzymatic proteolysis experiment where the products from the first limited digestions are separated using SDS-PAGE electrophoresis.

never overwrite or replace an .umsap file, instead the new analysis will be added to the already existing file. Only .umsap files can be selected here.

2. The button Data allows users to browse the file system to select the input data file that will be used for the analysis. The Data file is expected to be a plain text file with tab separated columns and the name of the columns in the first row of the file. In addition, columns to be analyzed must contain only numbers and must be of the same length. Only .txt files can be selected here.

3. The button Sequences allows users to browse the file system to select the FASTA file containing the sequence of the Recombinant protein and the Native protein. The FASTA file must contain at least one sequence.

4. The text field Analysis ID allows users to provide an ID for the analysis to be run. The date and time of the analysis will be automatically added to the beginning of the name.

Section Data Preparation contains four dropdown boxes. Here users select how the data in the Data file should be prepared before starting the analysis.

1. The dropdown Treat 0s as missing values allows user to define how to handle 0 values present in the Data file.

2. The dropdown Transformation allows user to select the Transformation method to be applied to the data.

3. The dropdown Normalization allows users to select the Normalization method to be

applied to the data.

4. The dropdown Imputation allows user to select the Imputation method used to replace missing values in the data.

Section User-defined values contains seven text fields and one dropdown box. Here users configure the Limited Proteolysis analysis to be run.

1. The text field Target Protein allows users to specify the protein of interest. Users may type here any unique protein identifier present in the Data file. The search for the Target Protein is case-sensitive, meaning that eFeB is not the same as efeb.

2. The text field Score Value allows users to define a threshold value above which the detected peptides will be considered as relevant. The Score Value is an indicator of how reliable was the detection of the peptide during the MS experiments. The value given to UMSAP depends on the program generating the Data file. Only one real number equal or greater than zero will be accepted here. A value of zero means all detected peptides belonging to the Target Protein will be treated as relevant peptides.

3. The dropdown Samples allows users to specify whether samples are independent or paired. For example, samples are paired when the same Petri dish is used for the control and experiments.

4–8. The parameters α, β, γ, Θ and Θmax are used to configure the equivalence test(*1*) performed to identify peptides in the selected gel spots with equivalent intensity values to the control spots (Section 7.4). α, β and γ must be between 0 and 1. The value of Θ is optional. If left blank UMSAP will calculate a value for each peptide based on the intensity values found in the Data file. If given then the given value will be used for each peptide. Θmax is the maximum value to consider the intensity values in the gel spot and control as equivalent.

Section Column numbers contains four text fields. Here, users provide the column numbers in the Data file from where UMSAP will get the information needed to perform the Limited Proteolysis analysis. All columns specified in this section must be present in the Data file. Column numbers start at 0. The column numbers are shown in the table of Region Data File Content after the Data file is selected.

1. The text field Sequences allows users to specify the column in the Data file containing the sequences of the peptides identified in the MS experiments. Only one integer number equal or greater than zero will be accepted here.

2. The text field Detected Proteins allows users to specify the column in the Data file containing the unique protein identifier for the proteins detected in the MS experiments. It is in this column where the program will look for the Target Protein value given in Section User-defined values. It is important that in this column the Target Protein value is used to refer to only one protein. Only one integer number equal or greater than zero will be accepted here.

3. The text field Score allows users to specify the column in the Data file containing the Score values. It is in this column where the program will look for the values to be compared against the Score threshold given in Section User-defined values.

4. The text field Results - Control experiments allows users to specify the columns in

the Data file containing the results of the control and experiments. The button Type Values call a helper window (Figure 7.2) where users can type the information needed.

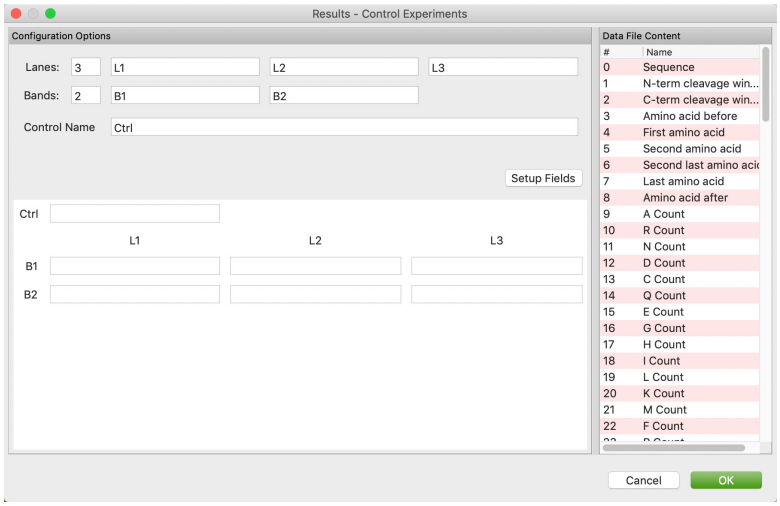

Figure 7.2: The Result - Control experiments helper window for the Limited Proteolysis module. This window allows users to specify the column numbers in the Data file containing the MS results for the selected gel spots.

The helper window is divided in two Regions. Region Data File Content will show the column numbers and names of the columns present in the selected Data file. Region Configuration Options has two sections. The upper section allows defining the number of bands and lanes of interest in the gel as well as the label for lanes, bands and control spot. The button Setup Fields creates the corresponding text fields in the bottom section to type the column numbers. Each text field should contain the column numbers with the MS results for the given gel spot. The values for the text fields should be positive integer numbers or a range of integers, e.g. 60–62 or left blank for empty gel spots. Selected rows in the table can be copied (Cmd+C) and then pasted (Cmd+V) in the text fields. Duplicate column numbers are not allowed.

7.4 The analysis

First, UMSAP will check the validity of the user-provided input and then the selected Data file is read. The columns specified in section Column numbers are extracted from the Data file. All other columns present in the Data file are discarded. After this, all steps selected in the Data Preparation section are applied to the columns specified in the text field Result - Control experiments (Chapter 6). Then, the following actions are performed.

All rows in the prepared data containing peptides that do not belong to the Target protein are removed. Then, all rows containing peptides from the Target protein but with Score values lower than the user-defined Score threshold are removed. These steps leave only relevant peptides, this means, peptides with a Score value higher than the

user-defined threshold that belong to the Target protein. For each one of these relevant peptides the equivalence test is performed (*1*).

The implementation of the equivalence test is based on the following equations:

$$s^* = s\sqrt{\frac{n-1}{\chi^2_{(\gamma,n-1)}}} \tag{7.1}$$

$$\Theta = \delta + s^* \left[t_{(1-\alpha,2n-2)} + t_{(1-\beta/2,2n-2)}\right]\sqrt{\frac{2}{n}} \tag{7.2}$$

$$(\bar{y}_1 - \bar{y}_2) \pm t_{(1-\alpha,n_1+n_2-2)} \cdot \sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \tag{7.3}$$

where s^* is an estimate of the upper confidence limit of the standard deviation, $\chi^2_{(\gamma,n-1)}$ is the (100γ)th percentile of the chi-squared distribution with $n-1$ degrees of freedom, Θ is the acceptance criterion, δ is the absolute value of the true difference between the group's mean values, t is the Student's t value, $\bar{y}$ is the measurement mean and s_p is the pooled standard deviation of the measurements calculated with:

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} \tag{7.4}$$

α, β, γ and Θ are the parameters defined in section User-defined values of region Configuration Options in the tab of the module.

In essence, for each relevant peptide, the control experiments are used to estimate the upper confidence limit for the standard deviation using Equation 7.1 and then the acceptance criterion is calculated with Equation 7.2. Finally, the confidence interval for the mean difference for the gel spot and the control is calculated with Equation 7.3 and compare to Θ. Peptides with equivalent mean intensity in at least one gel spot and the control are retained while not equivalent peptides are discarded.

If the value of Θ is given in section User-defined values of the module's tab then only the confidence interval for the mean difference is calculated, and the value is directly compared to the given Θ value. The maximum possible Θ value must always be provided. The reason for this is that when only a few replicates of the experiments are performed the calculated Θ value may be too large and then the equivalence test is easily past by all relevant peptides.

After the filtered peptide (FP) are identified the modules creates the output files.

7.5 The result window

The window showing the results from a Limited Proteolysis workflow is divided in four regions (Figure 7.3).

Region Peptide List contains a table with all FP detected during the analysis. Selecting a peptide in the table will highlight with a thick black border all gel spot in Region

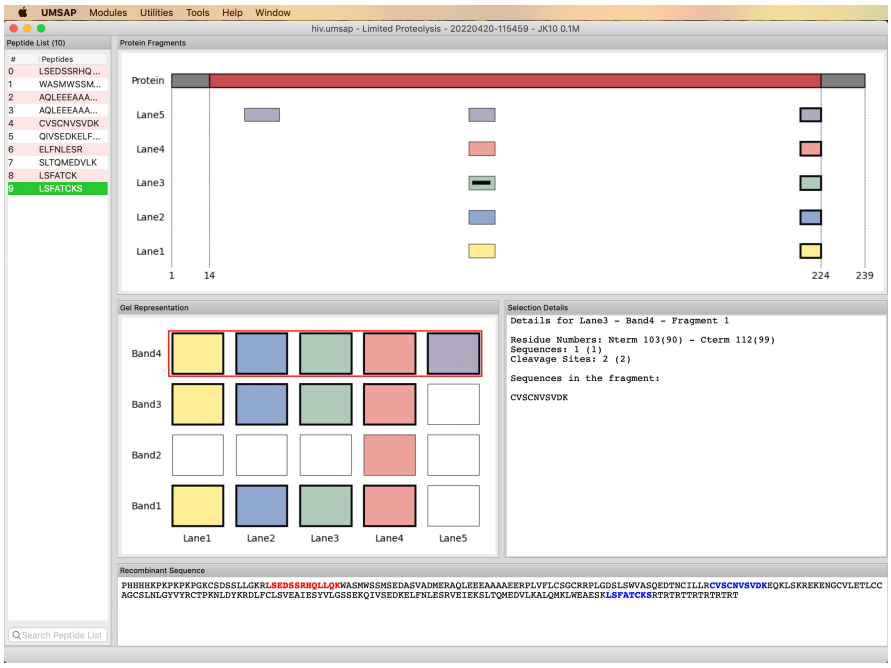

Figure 7.3: The Limited Proteolysis result window. Users can perform here the analysis of the fragments obtained in the limited proteolysis experiments.

Gel Representation and all fragments in region Protein Fragments where the selected peptide was found. In addition, region Recombinant Sequence will show the selected peptide in blue. The search box at the bottom allows searching for a sequence in the list of FP.

Region Gel Representation contains a representation of the analyzed gel. Here, each gel spot is represented with a square. White squares represent gel spot for which no peptide from the Target protein was detected with intensity values equivalent to the controls or that were not analyzed because no column number information was given when configuring the analysis. The rest of the square will be colored according to the band/lane they belong to.

There are two selection modes available for Region Gel Representation. In the Lane selection mode a left click over an empty space in the gel representation will select the closer lane. In this mode the gel spot will be colored according to the band they belong to. The selected lane will be highlighted with a red rectangle. The band selection mode works similarly, but users can select bands and the gel spot are colored according to the lane they belong to. The selection mode can be toggled through the keyboard shortcut Cmd+L or the Tools menu. In addition, the entire gel can be selected (Cmd+A) or a single gel spot can be selected with a left click. Any selection on the gel will update the content of Regions Protein Fragments, Selection Details and Recombinant Sequence.

Region Protein Fragments will display a graphical representation of the fragments found in each gel spot for the selected band/lane. The first fragment in this region represents the full length of the recombinant sequence of the Target protein. Here the central red section represents the sequence in the recombinant protein that is identical to the native

protein sequence while gray sections represent the sequences in the recombinant protein that are different to the native protein sequence. If the sequence of the native protein was not given then the fragment is shown in gray. The fragments are color coded using the same colors of the band/lane they belong to. Selecting a fragment will update the information shown in regions Selection Details and Recombinant Sequence.

Region Selection Details will show information about the selected lane/band, gel spot in region Gel Representation or the selected fragment in region Protein Fragments. The displayed information for a selected band/lane includes the number of non-empty lanes/bands, the number of fragments identified in each non-empty gel spot in the band/lane and the protein regions identified. Selecting a gel spot will display this information only for the gel spot. Selecting a fragment in region Protein Fragments will display the following information: number of cleavage sites, first and last residue number for the selected fragment and a sequence alignment of all peptides forming the fragment.

7.6 The Tools menu

The Tools menu in the window showing the results from a Limited Proteolysis analysis allows user to view any of the analyzes contained in the selected .umsap file. Users can toggle the band/lane selection mode (Cmd+L) and select all gel spot in the analysis (Cmd+A). In addition, the zoom level in the plots can be reset (Shift+Alt+Z) and an image of the plots can be created (Shift+Alt+I).

The submenu Clear Selection allows removing any selection done by the user. In particular, the entry All (Cmd+K) will remove all selections basically resetting the state of the window.

The Tools menu also allows duplicating the window (Cmd+D) for easier comparison of two or more analysis, checking the Data Preparation steps of the analysis (Cmd+P), exporting the results of the analysis to a tab separated CSV file (Cmd+E) and to export the sequence alignments (Cmd+S) between the peptides found in the analysis and the sequence of the recombinant protein.

Chapter 8

Proteome Profiling

The Proteome Profiling modules is designed to identify differentially expressed protein under various experimental conditions. A typical example is to compare the effect of two substance over protein expression using whole cell lysates.

8.1 Definitions

Before explaining in detail the interface of the module and how does the module work, let's make clear the meaning of some terms that will be used in the following paragraphs.

- *Detected protein*: any protein detected in any of the mass spectrometry experiments including the control experiments.
- *Relevant proteins*: a detected protein with a Score value above a user-defined thresholds (page 31).

8.2 The input files

The Proteome Profiling module requires only one input file. This Data file must follow the guidelines specified in Section 3.1. In short, the Data file must have a tabular format with tab separated columns and the name of the columns are expected as first row. All columns given as input in section Column numbers of Region Configuration Options of the interface (Figure 8.1) must be present in the Data file.

8.3 The interface

The tab of the Proteome Profiling module is divided in two regions (Figure 8.1).

The Data File Content region holds only a table to show the name of the columns in the selected Data File. The table will be automatically filled after selecting the file.

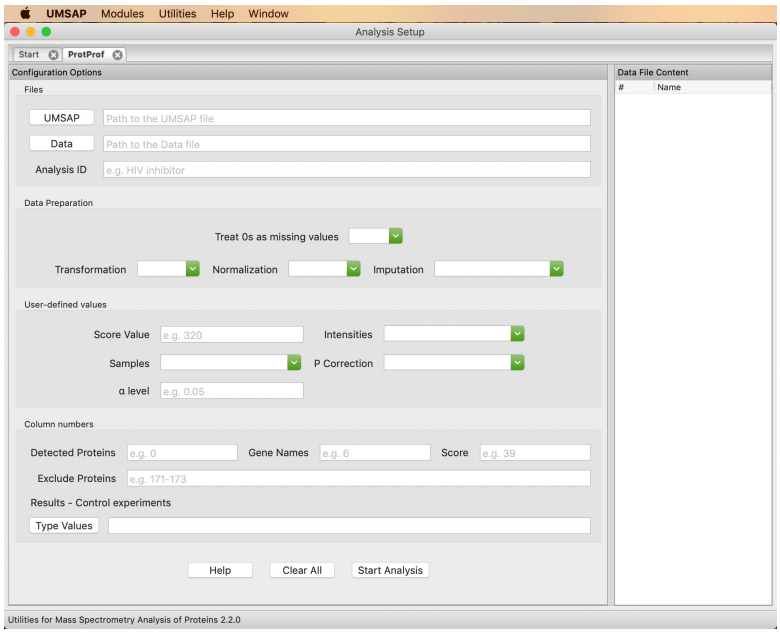

Figure 8.1: The Proteome Profiling module tab. This tab allows users to perform a proteome profiling analysis.

Selected rows in the table can be copied (Cmd+C) and pasted (Cmd+V) to the text fields in region Configuration Options.

The Configuration Options region contains all the fields needed to configure and run the analysis.

Section Files contains two buttons and a text field. Here users select the input and output files for the analysis.

1. The button UMSAP allows users to browse the file system to select the location and name of the .umsap file. When selecting an already existing .umsap file the operating system will ask if it is ok to replace the file, the answer can be yes since UMSAP will never overwrite or replace an .umsap file, instead the new analysis will be added to the already existing file. Only .umsap files can be selected here.

2. The button Data allows users to browse the file system to select the input data file that will be used for the analysis. The Data file is expected to be a plain text file with tab separated columns and the name of the columns in the first row of the file. In addition, columns to be analyzed must contain only numbers and must be of the same length. Only .txt files can be selected here.

3. The text field Analysis ID allows users to provide an ID for the analysis to be run. The date and time of the analysis will be automatically added to the beginning of the name.

Section Data Preparation contains four dropdown boxes. Here users select how the data in the Data file should be prepared before starting the analysis.

1. The dropdown Treat 0s as missing values allows user to define how to handle 0 values

present in the Data file.

2. The dropdown Transformation allows user to select the Transformation method to be applied to the data.

3. The dropdown Normalization allows users to select the Normalization method to be applied to the data.

4. The dropdown Imputation allows user to select the Imputation method used to replace missing values in the data.

Section User-defined values contains two text fields and three dropdown box. Here users configure the Proteome Profiling analysis to be run.

1. The text field Score Value allows users to define a threshold value above which the detected proteins will be considered as relevant. The Score value is an indicator of how reliable was the detection of the protein during the MS experiments. The value given to UMSAP depends on the program generating the Data file. Only one real number equal or greater than zero will be accepted as a valid input here. A value of zero means all detected proteins will be treated as relevant proteins.

2. The dropdown Samples allows users to specify whether samples are independent or paired. For example, samples are paired when the same Petri dish is used for the control and experiments.

3. The text field α level allows users to define the significance level used for the analysis. Only a number between 0 and 1 will be accepted here.

4. The dropdown Intensities allows user to specify whether the intensity values in the Data file represent absolute or a ratio of intensities.

5. The dropdown P Correction allows user to select the correction method for the p values calculated during the analysis.

Section Column numbers contains five text fields. Here, users provide the column numbers in the Data file from where UMSAP will get the information needed to perform the analysis of the module. All columns specified in this section must be present in the Data file. Column numbers start at 0. The column numbers are shown in the table of Region Data File Content after the Data file is selected.

1. The text field Detected Proteins allows users to specify the column in the Data file containing the protein identifiers found in the Data file. Only one integer number equal or greater than zero will be accepted here.

2. The text field Gene Names allows users to specify the column in the Data file containing the gene names of the proteins found during the MS experiments. Only one integer number equal or greater than zero will be accepted here.

3. The text field Score allows users to specify the column in the Data file containing the Score values. It is in this column where the program will look for the values to be compared against the Score threshold given in section User-defined values. Only one integer number equal or greater than zero will be accepted here.

4. The text field Exclude proteins allows users to specify several columns in the Data file. Proteins found in these columns will be excluded form the analysis. The module assumes

that these columns contains numeric values and values greater than zero indicate that the respective protein must be eliminated from the analysis. Only integer numbers equal or greater than zero will be accepted here. If left empty all proteins will be considered during the analysis.

5. The text field Results - Control experiments allows users to specify the columns in the Data file containing the results of the experiments. The button Type Values calls a helper window (Figure 8.2) where users can type the information needed. Duplicate column numbers are not allowed here.

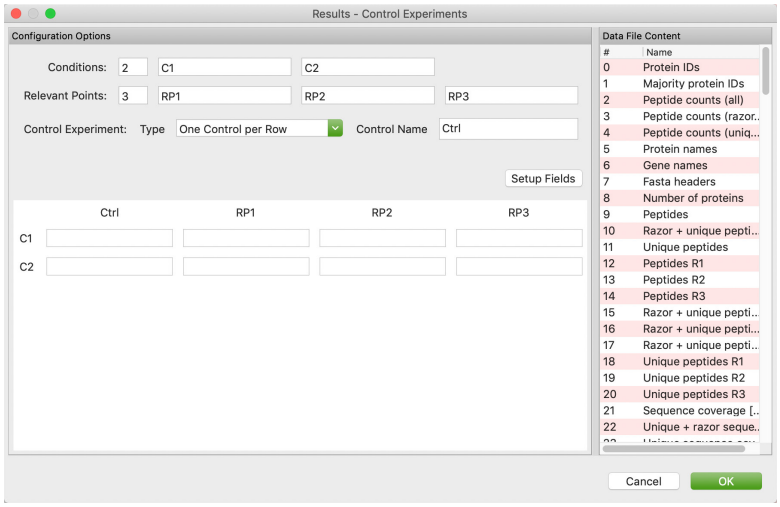

Figure 8.2: The Result - Control experiments helper window for the Proteome Profiling module. This window allows users to specify the column numbers in the Data file containing the MS results for the selected Conditions, Relevant points and control experiments.

The helper window is divided in two Regions. Region Data File Content will show the column numbers and names of the columns present in the selected Data file. Region Configuration Options has two sections. The upper section allows defining the number of conditions and relevant points analyzed, to define the kind of control experiment used as well as the label for conditions, relevant points and control experiment. The button Setup Fields allows creating the matrix of text fields in the bottom section where users type the column numbers. Each text field should contain the column numbers with the MS results for the given experiment. The values for the text fields should be positive integer numbers or a range of integers, e.g. 60–62 or left blank. Selected rows in the table can be copied (Cmd+C) and then pasted (Cmd+V) in the text fields. Duplicate column numbers are not allowed.

8.4 The analysis

First, UMSAP will check the validity of the user-provided input and then the selected Data file is read. The columns specified in section Column numbers are extracted from the Data file. All other columns present in the Data file are discarded. After this, all

steps selected in the Data Preparation section are applied to the columns specified in the text field Result - Control experiments (Chapter 6). Then, the following actions are performed.

All proteins found in the Exclude proteins columns are discarded. Also, all proteins with a Score value lower than the defined threshold are removed. The resulting data is used for the proteome profiling analysis (*2*). This includes the calculation of the fold change (FC) (Equation 8.1), the p values and the correction of the p values. Currently, the p values are calculated using a t-test.

$$FC = ave(I_{C,RP})/ave(I_{Control}) \tag{8.1}$$

8.5 The result window

The window showing the results from a Proteome Profiling analysis is divided in three regions (Figure 7.3).

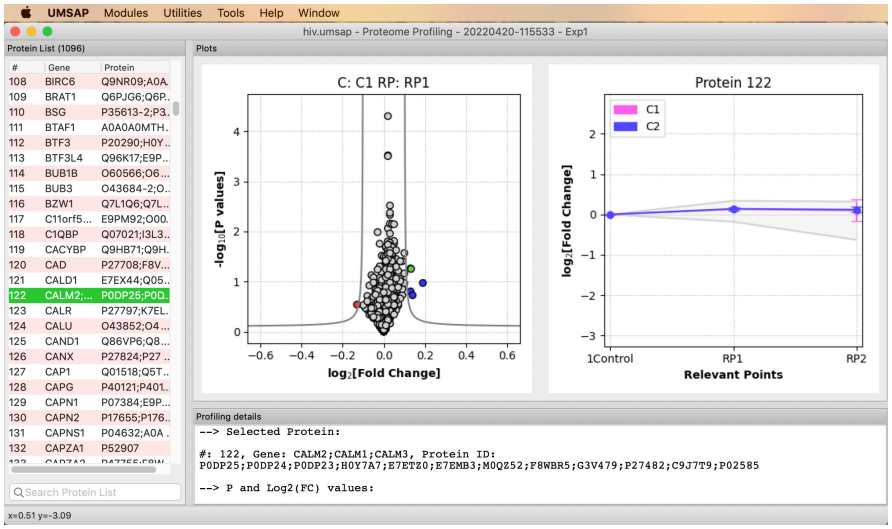

Figure 8.3: The Proteome Profiling result window. Users can perform here the analysis of the proteome profiling results.

Region Protein List contains a table with the protein IDs and Gene names that were not excluded from the analysis and whose Score values are greater than the user-defined threshold. Selecting a protein in the table will highlight the protein with a green dot in the Volcano plot and display the $log_2[FC]$ evolution along the relevant points in region Plots. In addition, information about the selected protein will be displayed in Regions Profiling details.

Region Plots contains two plots. A volcano plot showing the results for the t-test comparing the condition (C), relevant point (RP) to the corresponding control. The points in the plot can be colored by Z-score allowing to quickly identified the top up (blue) or down (red) regulated proteins or according to the hyperbolic curve cut-off (*3*).

Selecting a protein in the plot will highlight the selected protein in region Protein List and display information about it in region Profiling details, or it will add a label to the plot.

The plot of $log_2[FC]$ vs Relevant points allows to see the evolution of the FC along the Relevant points and to compare the FC across the various conditions in the analysis. The error bars in the plot represent the confident interval for the FC values calculate at the $1-\alpha$ level. The gray zone shown in the plot represent the maximum and minimum FC value for each Relevant point across all Conditions in the analysis.

Region Profiling details shows a summary of the results for the selected protein. Proteins can be selected in the table in region Protein List or in the volcano plot of region Plots. The information includes the calculated p and $log_2[FC]$ values as well as averages and standard deviations for intensities and ratios.

8.6 The Tools menu

The Tools menu in the window showing the results from a Proteome Proteolysis analysis allows user to display any of the analyses contained in the selected .umsap file.

The submenu Volcano Plot allows to control the appearance of the volcano plot. Users can change the the Condition - Relevant point displayed, add labels (Shift+A) to the points in the plot, toggle between using the left click of the mouse for selecting proteins or adding labels (Shift+P), adjusting the Color Scheme of the plot, show corrected p values, create an image of the plot (Shift+I) or reset the zoom level of the plot (Shift+Z). The entry Color Scheme allows coloring points in the Volcano plot by Z-score or the hyperbolic curve cut-off. It also allows configuring the Z-score threshold and the hyperbolic curve parameters.

The submenu FC Evolution allows toggling the display of the maximum/minimum area, to create an image of the plot (Alt+I) and to reset the zoom level of the plot (Alt+Z).

The submenu Lock Plot Scale allows setting the scale of the plot. If this option is set to No, then the scale of the plot is allowed to change every time the plots are updated, for example changing the Condition - Relevant point displayed for a given analysis. If set to Date, then the scale of the plot is fixed for all Conditions - Relevant points in one analysis and updated every time a new analysis is displayed. If set to Project, then all plots for all analysis will have the same scale.

The submenu Clear Selection allows removing any selection done by the user. In particular, the entry All (Cmd+K) will remove all selections basically resetting the state of the window.

The Tools menu also allows duplicating the window (Cmd+D) for easier comparison of two or more analysis, checking the Data Preparation steps of the analysis (Cmd+P), and exporting the results of the analysis to a tab separated CSV file (Cmd+E).

8.6.1 Filters

The proteins displayed in the Volcano plot can be filtered in order to display only proteins with a particular behavior. The submenu Filters allows managing which proteins will be displayed. Filters are applied to the current Condition and Relevant point shown in the Volcano plot. If the Condition or the Relevant point shown is changed, the new plot will show the proteins obtained after the filter was applied. This allows to follow the behavior of the filtered proteins in all Conditions and Relevant points. Any number of filters can be applied. The applied filters are shown in the bottom left corner of the window. Selecting a different analysis will reset the filters unless the option Auto Apply (Shift+Cmd+F) is selected. Also, filters can be manually reapplied with the menu entry Apply All (Shift+Cmd+A). Filter can also be removed, copied (Shift+Cmd+C) and pasted (Shift+Cmd+V) in a different window and saved (Shift+Cmd+S) and load (Shift+Cmd+L) from/to the .umsap file.

Currently, the implemented filters are:

FC Evolution

This filter allows identifying proteins whose FC evolution follows a defined pattern. For example, all proteins for which the FC is always greater than 0 or proteins that show $FC > 0$ in one condition but $FC < 0$ in others. The filter can be applied to the current selected Condition - Relevant point or considering any or all the conditions.

Hyperbolic Curve

This filter allows identifying proteins above the Hyperbolic curve cut-off. It is only applied to the current Condition - Relevant point.

Log2FC

This filter allows identifying proteins whose $log_2[FC]$ are greater or smaller than the given value. It is only applied to the current Condition - Relevant point.

P value

This filter allows identifying proteins by the calculated p value. The p value can be given in the 0 to 1 range or as a $-log10$ value. Regular or corrected P values can be used in the filter. It is only applied to the current Condition - Relevant point.

Z score

This filter allows identifying proteins by the Z score value of the FC. This is a quick way to identify, for example, the top 10 % up and down regulated proteins. It is only applied to the current Condition - Relevant point.

Chapter 9

Targeted Proteolysis

The module Targeted Proteolysis is designed to post-process the mass spectrometry data acquired during the enzymatic proteolysis of a Target protein by a single protease. In a typical experimental setup both the protease and the Target protein are mixed together under various experimental conditions and the peptides generated during the proteolysis are identified by mass spectrometry. It is expected that several control experiments and several replicates of the tested experimental conditions are performed. The main objective of the module is to identify the peptides with intensity values that are significantly different in the control experiments and the replicates of the various experimental condition tested at the chosen significance level.

9.1 Definitions

Before explaining in detail the interface and how does the module work, let's make clear the meaning of the term Filtered peptide in the context of the Targeted proteolysis module:

- *Filtered peptide*: a relevant peptide with a significantly different behavior in the control and a given experiment at the chosen significance level.

The rest of the definitions in Section 7.1 are valid for this module too.

9.2 The input files

The Targeted Proteolysis module requires a Data file containing the detected peptides and a sequence file containing the amino acid sequence of the recombinant protein used in the study. Both files must follow the guidelines specified in Section 3.1. In short, the Data file must have a tabular format with tab separated columns and the name of the columns are expected as first row. The Sequence file is expected to contain at least one sequence and to be FASTA formatted. If more than one sequence is found in

the Sequence file the first sequence will be taken as the sequence of the Recombinant protein and the second sequence will be taken as the sequence of the Native protein. All other sequences are discarded.

9.3 The interface

The tab of the Targeted Proteolysis module is divided in two regions (Figure 9.1).

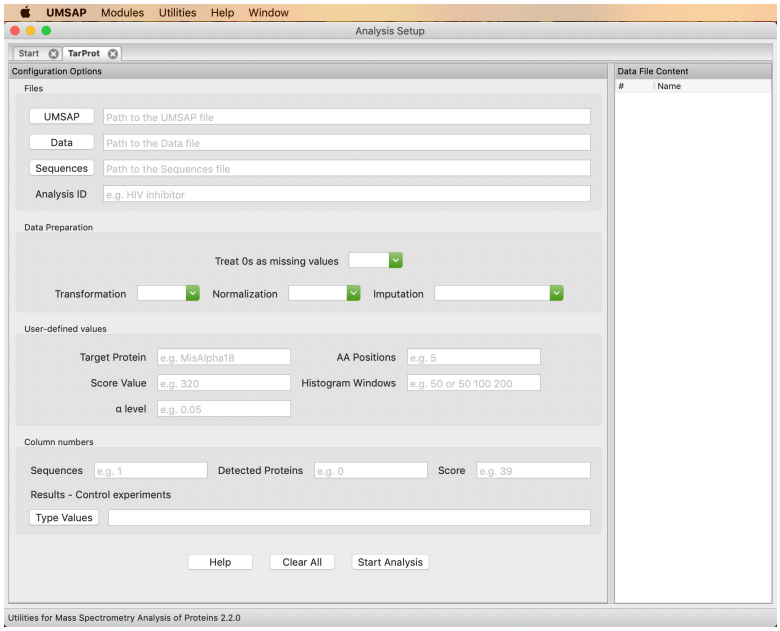

Figure 9.1: The Targeted Proteolysis module tab. This tab allows users to perform the analysis of the results obtained during an enzymatic proteolysis experiment.

The Data File Content region holds only a table to show the name of the columns in the selected Data File. The table will be automatically filled after selecting the file. Selected rows in the table can be copied (Cmd+C) and pasted (Cmd+V) to the text fields in region Configuration Options.

The Configuration Options region contains all the fields needed to configure and run the analysis.

Section Files contains three buttons and a text field. Here users select the input and output files for the analysis.

1. The button UMSAP allows users to browse the file system to select the location and name of the .umsap file. When selecting an already existing .umsap file the operating system will ask if it is ok to replace the file, the answer can be yes since UMSAP will never overwrite or replace an .umsap file, instead the new analysis will be added to the already existing file. Only .umsap files can be selected here.

2. The button Data allows users to browse the file system to select the input data file that will be used for the analysis. The Data file is expected to be a plain text file with

tab separated columns and the name of the columns in the first row of the file. In addition, columns to be analyzed must contain only numbers and must be of the same length. Only .txt files can be selected here.

3. The button Sequences allows users to browse the file system to select the FASTA file containing the sequence of the Recombinant protein and the Native protein. The FASTA file must contain at least one sequence.

4. The text field Analysis ID allows users to provide an ID for the analysis to be run. The date and time of the analysis will be automatically added to the beginning of the name.

Section Data Preparation contains four dropdown boxes. Here users select how the data in the Data file should be prepared before starting the analysis.

1. The dropdown Treat 0s as missing values allows user to define how to handle 0 values present in the Data file.

2. The dropdown Transformation allows user to select the Transformation method to be applied to the data.

3. The dropdown Normalization allows users to select the Normalization method to be applied to the data.

4. The dropdown Imputation allows user to select the Imputation method used to replace missing values in the data.

Section User-defined values contains five text fields Here users configure the Targeted Proteolysis analysis to be run.

1. The text field Target Protein allows users to specify the protein of interest. Users may type here any unique protein identifier present in the Data file. The search for the Target Protein is case-sensitive, meaning that eFeB is not the same as efeb.

2. The text field Score Value allows users to define a threshold value above which the detected peptides will be considered as relevant. The Score Value is an indicator of how reliable was the detection of the peptide during the MS experiments. The value given to UMSAP depends on the program generating the Data file. Only one real number equal or greater than zero will be accepted here. A value of zero means all detected peptides belonging to the Target Protein will be treated as relevant peptides.

3. The text field α level allows users to define the significance level used for the analysis. Only a number between 0 and 1 will be accepted here.

4. The text field AA Positions allows users to define the number of positions to be considered during the amino acid (AA) distribution calculation (subsection 9.6.1). Only one integer number greater than zero will be accepted here. If left empty the calculation will not be performed.

5. The text field Histogram Windows allows users to define the size of the windows for the Histogram analysis (subsection 9.6.3). Only integer numbers equal or greater than zero will be accepted here. In addition, the values must be organized from smaller to bigger values. Users may specify a fix histogram window size by given just one integer number greater than zero. In this case the histogram will have even spaced

windows with the width specified by Histogram Windows. If more than one number is provided here, then windows with the customs width will be created. For example, the input 50 100 will create only one window including cleavage sites between residues 50 to 99. The input 150 100 150 will create three windows including cleavages sites between residues 1 to 49, 50 to 99 and 100 to 149. Duplicate values are not allowed. If left empty the histogram will not be created.

Section Column numbers contains four text fields. Here, users provide the column numbers in the Data file from where UMSAP will get the information needed to perform the analysis of the module. All columns specified in this section must be present in the Data file. Column numbers start at 0. The column numbers are shown in the table of Region Data File Content after the Data file is selected.

1. The text field Sequences allows users to specify the column in the Data file containing the sequences of the peptides identified in the MS experiments. Only one integer number equal or greater than zero will be accepted here.

2. The text field Detected Proteins allows users to specify the column in the Data file containing the unique protein identifier for the proteins detected in the MS experiments. It is in this column where the program will look for the Target Protein value given in Section User-defined values. It is important that in this column the Target Protein value is used to refer to only one protein. Only one integer number equal or greater than zero will be accepted here.

3. The text field Score allows users to specify the column in the Data file containing the Score values. It is in this column where the program will look for the values to be compared against the Score threshold given in Section User-defined values.

4. The text field Results - Control experiments allows users to specify the columns in the Data file containing the results of the control and experiments. The button Type Values call a helper window (Figure 9.2) where users can type the information needed.

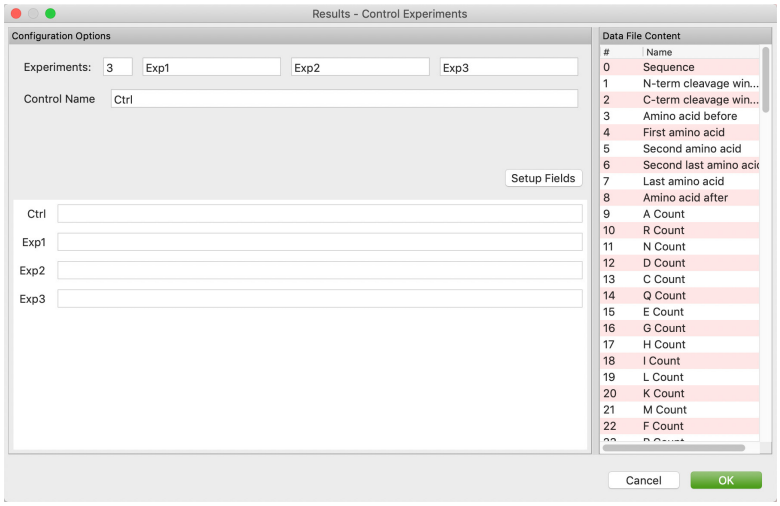

Figure 9.2: The Result - Control experiments helper window for the Targeted Proteolysis module. This window allows users to specify the column numbers in the Data file containing the MS results for the selected experiments and control.

The helper window is divided in two Regions. Region Data File Content will show the column numbers and names of the columns present in the selected Data file. Region Configuration Options has two sections. The upper section allows defining the number of experiments performed as well as the label for the control and experiments. The button Setup Fields creates the corresponding text fields in the bottom section to type the column numbers. Each text field should contain the column numbers with the MS results for the given experiment. The values for the text fields should be positive integer numbers or a range of integers, e.g. 60–62. Selected rows in the table can be copied (Cmd+C) and then pasted (Cmd+V) in the text fields. Duplicate column numbers are not allowed.

9.4 The analysis

First, UMSAP will check the validity of the user-provided input and then the selected Data file is read. The columns specified in section Column numbers are extracted from the Data file. All other columns present in the Data file are discarded. After this, all steps selected in the Data Preparation section are applied to the columns specified in the text field Result - Control experiments (Chapter 6). Then, the following actions are performed.

All rows in the Data file containing peptides that do not belong to the Target protein are removed. Then, all rows containing peptides from the Target protein but with Scores values lower than the user-defined Score threshold are removed. These steps leave only relevant peptides, this means peptides with a Score value higher than the user defined threshold that belong to the Target protein. For each one of these relevant peptides a test for Homogeneity of Regression (*4*) is performed.

The test, done to identify relevant peptides with different intensities in the control and a given experiment, is performed in three steps. First, the intensity values for the replicates in the control and in a given experiment are organized in two data sets as indicated in Figure 9.3. Each data set consist of two points. For the control data set the intensity values in the replicates of the control experiment are allocated to both points. For the experiment data set the intensity of the replicates in the control experiment are allocated to the first point and the intensity values of the replicates in the given experiment are allocated to the second point. The second step is to find the slope of the straight line best fitting each data set. The third step is to test the Homogeneity of Regression. Peptides that fail this test are included in the list of filtered peptides (FP) because the slopes of the straight lines fitting the data sets are significantly different at the chosen significance level. The fact that the slopes are different implies that the peptide is found in an increased concentration in the given experiment than in the control experiment. Only positive slopes are considered. Peptides that past this test are not included in the list of FP for the given experiment.

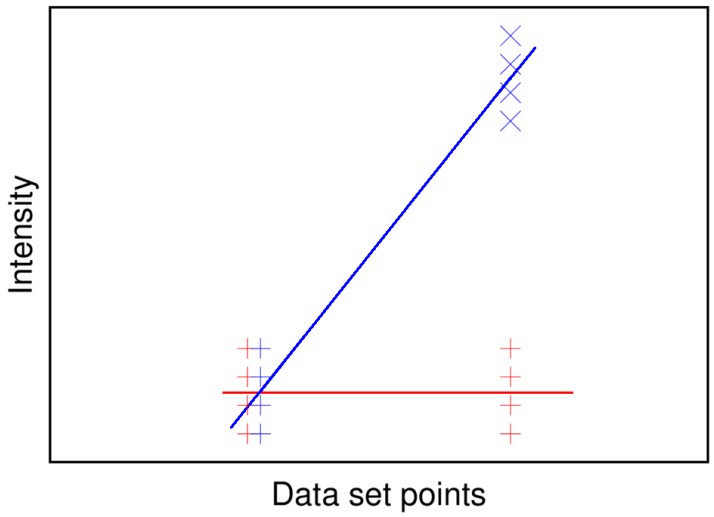

Figure 9.3: Data organization prior to the Homogeneity of Regression test. Two data sets with two points each are created, one data set for the control (red) and one for a given experiment (blue). The intensity data in the replicates of the control is used in both data points for the control (+) and in the first data point of the given experiment. The intensity data in the replicates of the given experiment is used for the second point of the data set for the given experiment (x). After this, the best fitting line for each data set is found and the slopes of the lines are compared in a test for Homogeneity of Regression.

9.5 The result window

The window showing the results from a Targeted Proteolysis analysis is divided in four regions (Figure 9.4).

Region Peptide List contains a table with all FP detected during the analysis. Selecting a peptide in the table will highlight with a thick black border all fragments in region Protein Fragments where the selected peptide was found. In addition, region Intensity will show the behavior of the intensity values in the control and experiments for the selected peptide. The search box at the bottom allows searching for a sequence in the list of FP.

Region Protein Fragments will display a graphical representation of the fragments found in each experiment. The first fragment in this region represents the full length of the recombinant sequence of the Target protein. Here the central red section represents the sequence in the recombinant protein that is identical to the native protein sequence while gray sections represent the sequences in the recombinant protein that are different to the native protein sequence. If the sequence of the native protein was not given then the fragment is shown in gray. Selecting a fragment will update the information shown in regions Selection Details.

Region Intensity will display a plot of Intensities vs experiment label. The plot will display the data for a single peptide selected in the table of region Peptide List. The

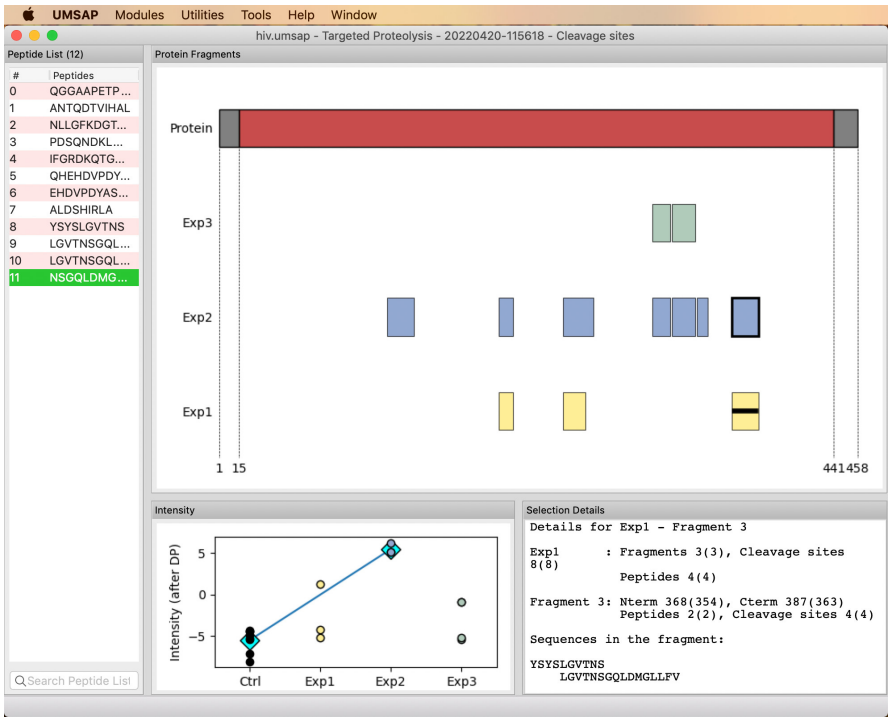

Figure 9.4: The Fragment analysis window. Users can perform here the analysis of the fragments obtained in the enzymatic proteolysis experiments.

intensity values in the plot are the ones obtained after applying the Data Preparation workflow specified for the analysis. The values for replicates are shown as circles. The average for a given experiment is shown as bigger cyan diamonds. The blue lines only connect the control experiment and experiments showing intensity values significantly different at the chosen significance level.

Region Selection Details will display detailed information about a selected fragment in Region Protein Fragments. In this Region regular numbers refer to the Recombinant protein sequence while number between parenthesis refers to the Native sequence. The Experiment section in the text contain information about the experiment in which the fragment was identified. The information includes the total number of fragments, the total number of sequences and the total number of unique cleavage sites identified in the experiment. The Fragment section in the text gives similar information about the selected fragment and also includes the first and last residue numbers in the fragment. The rest of the lines show a sequence alignment of all the peptides in the fragment.

9.6 The Tools menu

The Tools menu in the window showing the results from a Targeted Proteolysis analysis allows user to view any of the analyses contained in the selected .umsap file. The submenus Fragments and Intensities allow resetting the zoom level and create an image of the corresponding plots.

The submenu Clear Selection allows removing any selection done by the user. In particular, the entry All (Cmd+K) will remove all selections basically resetting the state of the window.

The Tools menu also allows duplicating the window (Cmd+D) for easier comparison of two or more analysis, checking the Data Preparation steps of the analysis (Cmd+P), exporting the results of the analysis to a tab separated CSV file (Cmd+E) and to export the sequence alignments (Cmd+S) between the peptides found in the analysis and the sequence of the recombinant protein. In addition, the zoom level in the plots can be reset (Shift+Alt+Z) and an image of the plots can be created (Shift+Alt+I).

The submenu Further Analysis give access to other analysis that can be performed with the information obtained with the Targeted Proteolysis analysis.

9.6.1 AA Distribution

The AA Distribution analysis allows user calculating the AA distribution around the detected cleavage sites in a Targeted Proteolysis analysis. The submenu AA Distribution gives access to all AA distribution analysis associated with the currently selected Targeted Proteolysis analysis. The submenu is one of the entries in submenu Further Analysis of the Tools menu of the result window.

The entry New AA Analysis in submenu AA Distribution allows creating a new analysis. The only input needed from the user is the number of AA positions around the cleavage sites to be considered.

The analysis

First, UMSAP will check the validity of the user-provided input. For each FP found in the current Targeted Proteolysis analysis, the sequence around the N and C terminal ends of the peptide is analyzed up to the user-provided number of AA. For the N terminus of the peptide the identity of residues in positions Pn to $P1$ is inferred from the sequence of the recombinant protein. The same is done for positions $P1'$ to Pn' at the C terminus of the peptide. If the N or C terminus of a peptide is the first or last residue of the recombinant protein under study the N or C terminus is excluded from the analysis. For each position the number of times that each AA appears at a given position are counted. Finally, the absolute numbers of AA appearances for each position are converted to percent taking the total for each position as the sum of all counted AA in the position.

In addition, UMSAP tests whether the obtained AA distribution is significantly different to the expected AA distribution from the proteolysis of the Target protein by a totally non-selective protease. The first step is to generate an AA distribution with the same number of positions defined by the user with the Position parameter. This distribution is generated assuming that all peptidic bonds in the recombinant protein may be cleaved by the protease with equal probability and that all peptidic bonds will be cleaved. Here, we are also assuming that all products of cleaving all peptidic bonds will be detected in the MS experiment. Then, UMSAP compares each position in both distributions using a χ^2 test with the same significance level given for the Targeted Proteolysis analysis.

In order to be able to perform the χ^2 test, AAs are pooled together in the same groups as described below for the color code used in the output.

The output

The window showing the results from an AA Distribution analysis contains a single plot (Figure 9.5). The plot can show the results for each experiment in the analysis or the results for each position in each experiment.

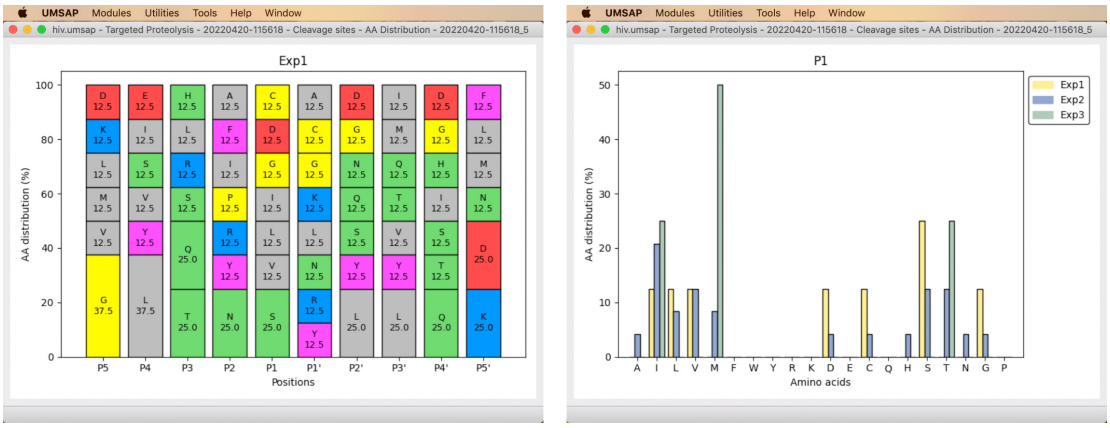

Figure 9.5: The AA Distribution result window. This window allows to visualize the results from an AA Distribution analysis.

The results for each experiment is presented as a bar graph of the AA distribution in which each bar represents a position. AAs are color coded with positively charged AAs (R and K) in blue, negatively charged AAs (D and E) in red, polar AAs (S, T, N, H, C and Q) in green, non-polar AAs (A, V, I, L and M) in gray, aromatic AAs (F, Y and W) in pink and Gly and Pro in yellow. AAs with an occurrence higher than 10 % are labeled with the one-letter code for the AA and the percentage value. For example, in Figure 9.5 the value of 25.0 % obtained for S in position $P1$ means that S was found in position $P1$ in the 25.0 % of the total cleavage sites detected.

The results of the χ^2 test are given in the color of the name of the position. A green color represents that the obtained distribution in the position is significantly different to a no selectivity distribution at the level of significance given for Targeted Proteolysis analysis. A red color represents that the distributions are not significantly different at the level of significance given for the Targeted Proteolysis analysis. Finally, a black color indicates that the number of expected values below 5 was higher than the 20 % threshold recommended by Yates et al. and the test was not performed (*5*).

If the mouse pointer is placed on top of the bars, then information related to the bar and the AA will be shown in the status bar at the bottom of the window. The information includes the Position (Pos), the amino acid (AA), how many times does the AA appear in the position as a percent of the total AA count for the given position (%) and the absolute number (Abs) and how many times does the AA appear in the sequence of the recombinant protein (InSeq).

The results for the analyzed positions allows comparing the AA distribution in one

position across all experiments. This is also a bar representation in which each bar represents an experiment and each position an AA. Placing the mouse pointer over a bar shows information about it. The information includes the AA (AA), the experiment (Exp), how many times does the AA appear in the position for the given experiment as a percent of the total AA count found at the position for the given experiment (%) and the absolute number (Abs) and how many times does the AA appear in the sequence of the recombinant protein (InSeq).

The Tools menu

The Tools menu in the window allows user changing the displayed experiment or selecting the position for which results in the experiments should be compared. In addition, users may select to duplicate the window (Cmd+D) for easier comparison of the results, to export the data shown in the window (Cmd+E), save a figure of the plot (Cmd+I) and to reset zoom level of the plot (Cmd+Z).

9.6.2 Cleavage Evolution

9.6.3 Cleavage Histograms

The Cleavage Histogram analysis allows creating histograms of the identified cleavage sites using the residue numbers of the Target protein as the definition of the windows in the histograms. The submenu Cleavage Histograms gives access to all histograms associated with the currently selected Targeted Proteolysis analysis. The submenu is one of the entries in submenu Further Analysis of the Tools menu of the result window.

The entry New Histograms in submenu Cleavage Histograms allows creating a new histogram. The only input needed from the user is the definition of the windows for the histograms. Only integer numbers equal or greater than zero will be accepted as input. In addition, the values must be organized from smaller to bigger values. Users may specify a fix histogram window size by given just one integer number greater than zero. In this case the histogram will have even spaced windows with the specified width.If more than one number is provided here, then windows with the customs width will be created. For example, the input 50 100 will create only one window including cleavage sites between residues 50 to 99. The input 150 100 150 will create three windows including cleavages sites between residues 1 to 49, 50 to 99 and 100 to 149. Duplicate values are not allowed.

The analysis

First, UMSAP will check the validity of the user-provided input. After this, the windows of the histograms will be created and for each experiment in the currently selected analysis the detected cleavage sites will be assigned to the corresponding windows. Histograms are created for the residue numbers in the recombinant protein and for the residue numbers in the native protein if the native sequence was given when configuring the Targeted Proteolysis analysis.

The output

The Histograms window will display a bar plot with the selected histogram (Figure 9.6).

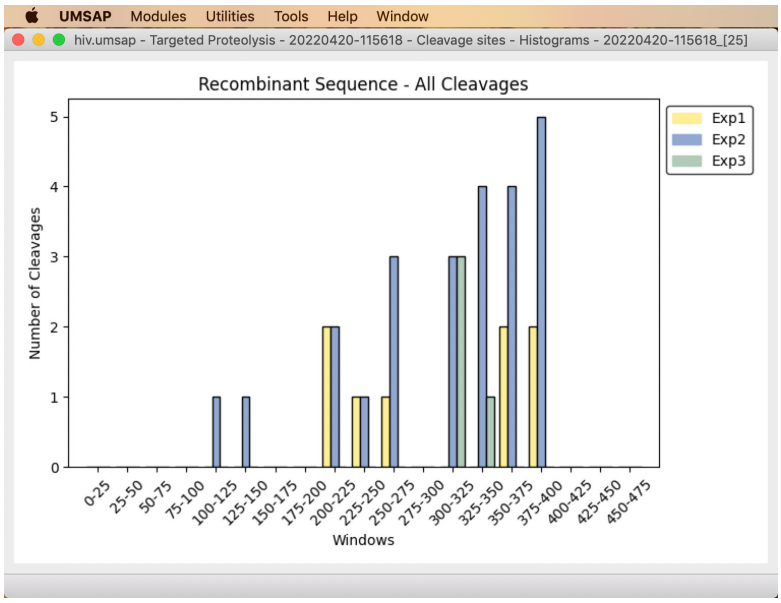

Figure 9.6: The Histograms result window. This window allows to visualize the histograms generated for the selected Targeted Proteolysis analysis.

In the histogram, experiments are shown in the order specified when configuring the analysis. Placing the mouse over the plot will display information at the bottom of the window. The information displayed includes the selected window (Win), the experiment represented by the bar (Exp) and the number of cleavages (Cleavages).

The Tools menu

The Tools menu allows showing the results for the recombinant or native sequence and showing only unique cleavages or the total count of the detected cleavages. In addition, users may select to duplicate the window (Cmd+D) for easier comparison of the results, to export the data shown in the window (Cmd+E), save a figure of the plot (Cmd+I) and to reset zoom level of the plot (Cmd+Z).

9.6.4 Cleavage per Residue

The Cleavages per residue utility calculates the absolute number of cleavages detected in the MS experiments for each residue in the recombinant protein under study. The peptides used to identify the cleavage sites are the FP contained in the .tarprot file used as input for the calculation (see page 36). How to generate the .tarprot file is discussed in Chapter 9. The FP list is a non redundant list.

The interface

The Cleavages per residue utility does not have a window since there are no options to specify. When the utility is selected users will be asked to select a .tarprot file and then users must select the output file. That is all.

The analysis

First, UMSAP will check the validity of the user provided input. After this the list of FP will be generated from the .tarprot file. Then, UMSAP will count how many times each residue in the protein under study appears at the C terminus of a FP or at the $N-1$ position of a FP (N is the N terminus of the FP). The cleavages per residue value for the first and last residue of the protein under study is of course zero. This is done for every experiment in the .tarprot file and also taking into account the results for all experiments. Finally, UMSAP will take the absolute number of cleavages per residue and will normalize the values to bring them in the 0 to 1 range. Both the absolute and normalized values are written to the output file. After the analysis is done the results will be automatically loaded and displayed in a new window (Figure 9.7).

The output

The output file from the Cleavages per residue utility will be shown as a simple number of cleavages vs residue number plot (Figure 9.7). Residues with cleavages per residue higher than one third of the maximum number of cleavages per residue will be highlighted with an asterisk (*). Placing the mouse pointer inside the plot will display the residue number and the number of cleavages in the status bar at the bottom of the window. When two data sets are plotted simultaneously, the number of cleavages are given in the same order shown by the legend in the window. The gray vertical lines enclose the native residues.

The Tools menu

The window shows by default the absolute number of cleavages considering all FP for the recombinant protein. The Tools menu allows users to change this. Users may select to plot the results for a particular experiment or to compare two experiments. In addition, only the native sequence could be plotted or the normalized cleavages per residue values may be shown. An image of the plot can be created using the Save Plot Image entry in the Tools menu. The Reset View option restores the default appearance of the window. The Export Data option allows to save the data shown in the window to a plain text file.

9.6.5 PDB Mapping

The Cleavages to PDB Files utility maps the number of cleavages per residue found in a .tarprot file to a .pdb file containing the structure of the Target protein. The peptides used to identify the cleavage sites are the FP contained in the .tarprot file used as input for the calculation, see page 36. How to generate the .tarprot file is discussed in Chapter 9. The FP list is a non redundant list.

The interface

The Cleavages to PDB Files window is divided in three regions, see Figure 9.8.

Region 1 contains three buttons allowing users to quickly delete all provided input to generate a new .pdf file. The Clear all button will delete all user provided input. The Clear files button will delete all values in section Files of Region 2. Finally, the Clear values button will delete all values in section Values of Region 2.

Region 2 contains the fields where users provide the information needed in order to

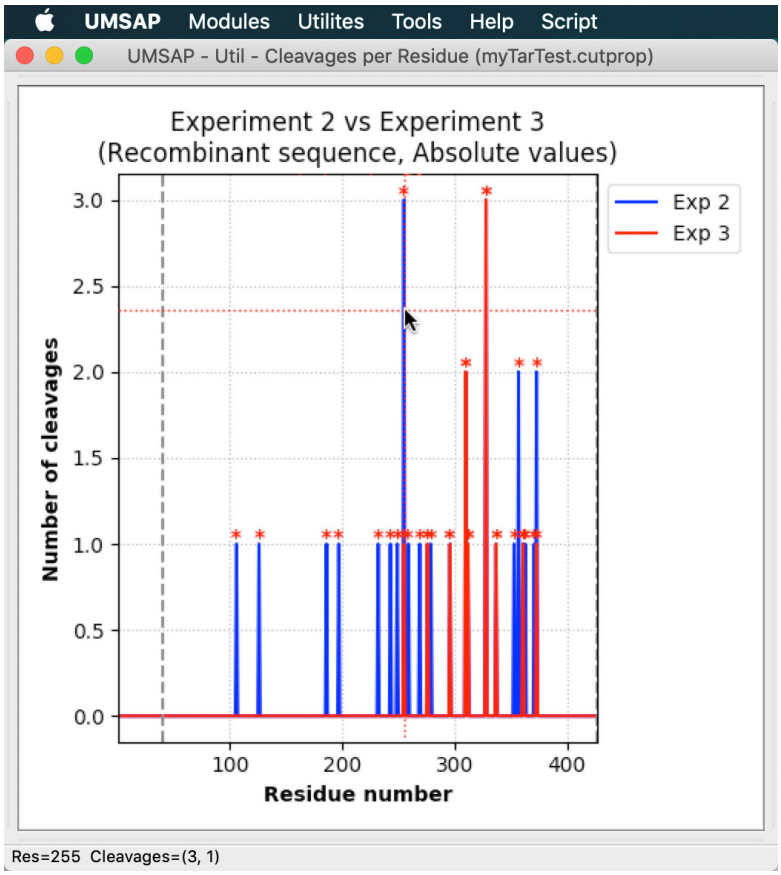

Figure 9.7: The Cleavages per Residue analysis window. This window allows to visualize the results contained in a .cutprop file.

perform the mapping of the number of cleavages. The Tarprot file button allows users to browse the file system to select the .tarprot file that will be use for the analysis. Only one .tarprot file can be provided here. The PDB file allows user to browse the file system to select a .pdb file. The .pdb file will contain the structure of the Target protein. This field can be left empty if the PDB file is to be downloaded from the PDB database. The Output folder button allows users to browse the file system to select the location of the resulting PDB folder containing the results. If left empty, then the PDB folder, resulting from the analysis, will be saved in the same directory containing the .tarprot file. If the Output folder option is left empty and the folder containing the selected .tarprot file already contains a PDB folder, then UMSAP will add the current date and time to the second to the end of the folder name in order to avoid overwriting the older PDB folder without explicit user permission.

The PDB ID field allows users to specify the chain or the segment id in the .pdb file that should be used for the mapping. Alternatively, if the PDB file field is left empty a code from the PDB database plus the chain or segment id may be given here. In this case, the pdb file will be directly downloaded from the PDB database. The expected syntax in this case is code;chain or code;segment for example, 2f4y;A or 2f4y;PROA.

Region 3 contains the Help and Start analysis buttons and a progress bar. The Help button leads to an online tutorial while the Start analysis button will start the analysis. The progress bar will give users a rough idea of the remaining processing time once the analysis is started.

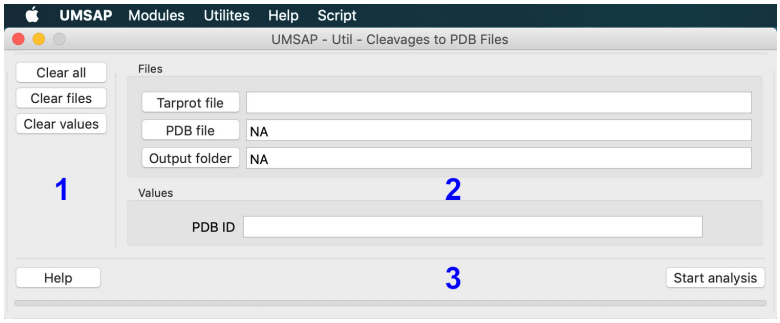

Figure 9.8: The Cleavages to PDB Files utility window. This window allows to map the detected number of cleavages per residue to a .pdb file containing the structure of the Target protein. The number of cleavages are mapped to the beta field in the .pdb.

The analysis

First, UMSAP will check the validity of the user provided input. Then, a temporal .cutprop file will be created. Details about the .cutprop files can be found in subsection 9.6.4. After this, the sequence from the .pdb file is extracted and aligned with the sequence of the recombinant protein found in the .tarprot file. Finally, the number of cleavages found in the .cutprop file are mapped to the corresponding residues in the .pdb file.

The output

The output from this utility is a series of .pdb files that will be saved in a PDB folder. Each file contains the number of cleavages mapped to the beta field of the corresponding residue in the .pdb structure. The results for each experiment and the FP list are mapped to individual files. The mapped values can be visualized by opening the .pdb files with VMD, PyMol or Chimera and coloring the structure by beta factors.

Bibliography

1. G. B. Limentani, M. C. Ringo, F. Ye, M. L. Bergquist, E. O. MCSorley, *Analytical Chemistry* **77**, 221 A–226 A, ISSN: 0003-2700 (June 2005).
2. J. T. Aguilan, K. Kulej, S. Sidoli, *Molecular Omics* **16**, 573–582, ISSN: 2515-4184 (2020).
3. W. LI, *Journal of Bioinformatics and Computational Biology* **10**, 1231003, ISSN: 0219-7200 (Dec. 2012).
4. *http://vassarstats.net/textbook/index.html*.
5. D. Yates, D. Moore, G. McCabe, *The practice of Statistics* (Freeman, New York, 1999), p. 734.