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 (**??**). 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.

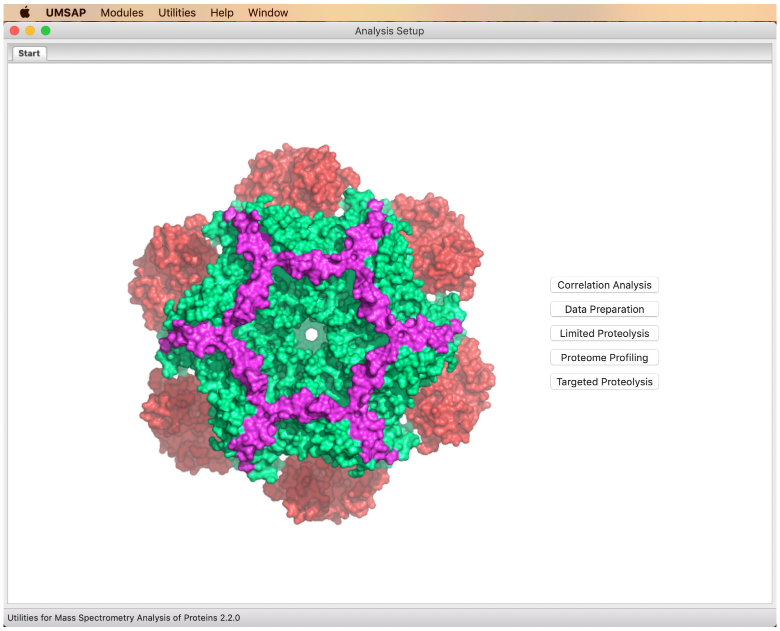


Figure 3.1: The main window of UMSAP. From this window users can access all the available Modules and Utilities.

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.

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

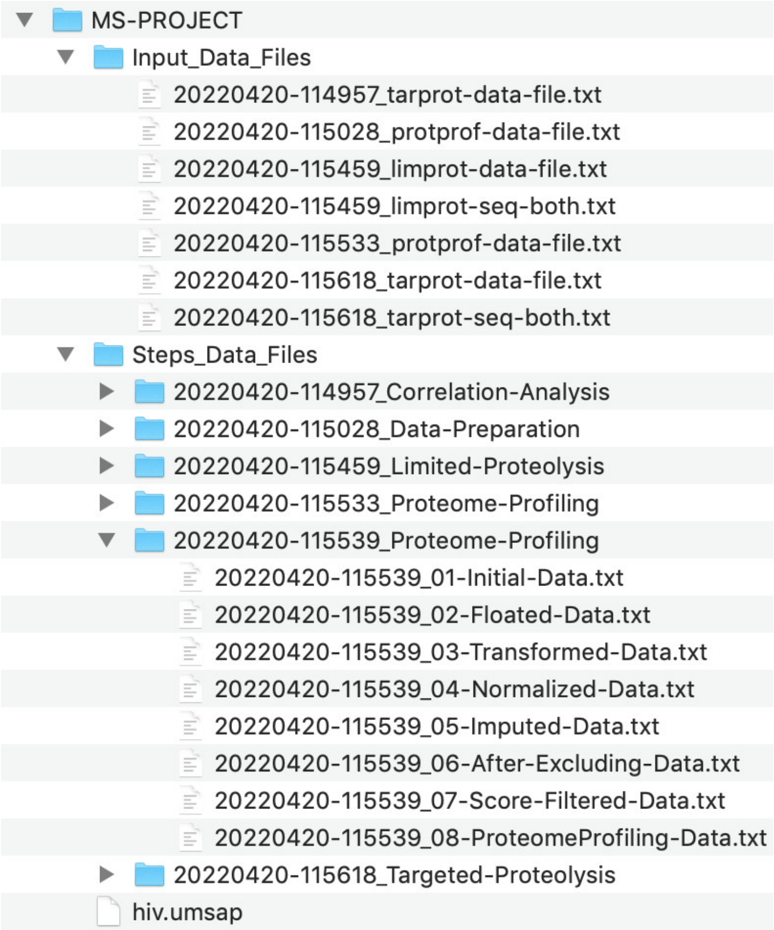
3.2 The output files

Results generated by UMSAP will be saved in two folders and a file with extension .umsap (**??**). 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 (**??**). 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.



```
▼ MS-PROJECT
    ▼ Input_Data_Files
          20220420-114957_tarprot-data-file.txt
          20220420-115028_protprof-data-file.txt
          20220420-115459_limprot-data-file.txt
          20220420-115459_limprot-seq-both.txt
          20220420-115533_protprof-data-file.txt
          20220420-115618_tarprot-data-file.txt
          20220420-115618_tarprot-seq-both.txt
    ▼ Steps_Data_Files
        ▶ 20220420-114957_Correlation-Analysis
        ▶ 20220420-115028_Data-Preparation
        ▶ 20220420-115459_Limited-Proteolysis
        ▶ 20220420-115533_Proteome-Profiling
        ▼ 20220420-115539_Proteome-Profiling
              20220420-115539_01-Initial-Data.txt
              20220420-115539_02-Floated-Data.txt
              20220420-115539_03-Transformed-Data.txt
              20220420-115539_04-Normalized-Data.txt
              20220420-115539_05-Imputed-Data.txt
              20220420-115539_06-After-Excluding-Data.txt
              20220420-115539_07-Score-Filtered-Data.txt
              20220420-115539_08-ProteomeProfiling-Data.txt
        ▶ 20220420-115618_Targeted-Proteolysis
      hiv.umsap
```

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 **??**.

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 (**??**).

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 (**??**) 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.

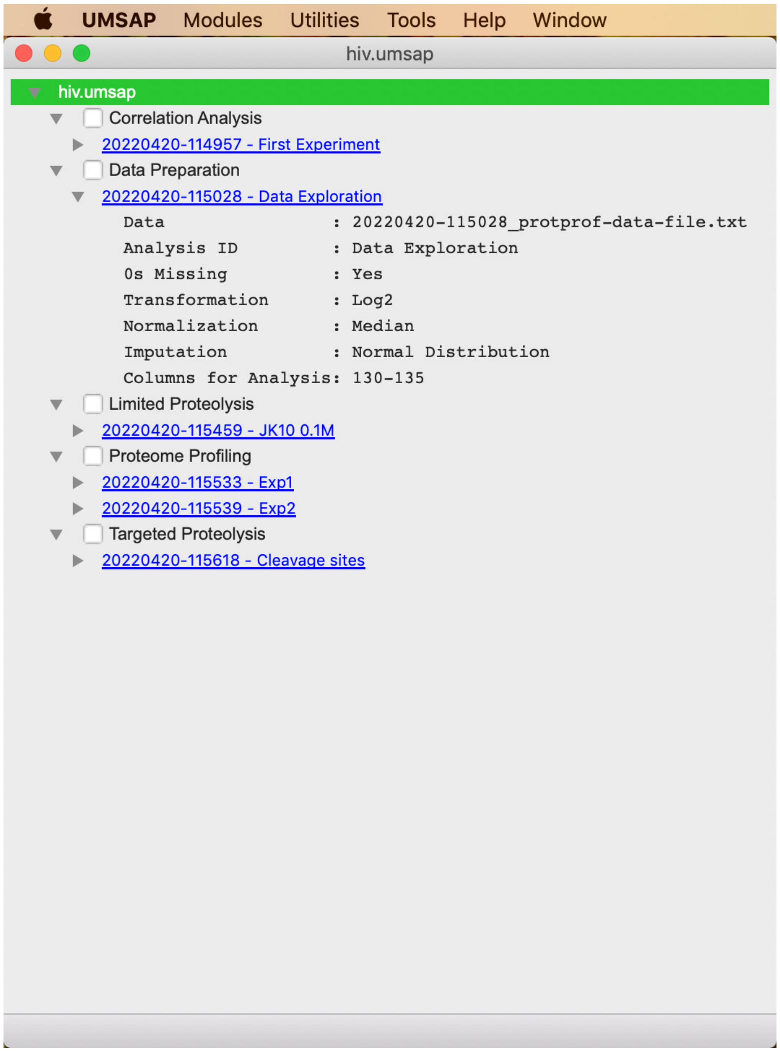
4.2 The Tools menu

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.



```
▼ hiv.umsap
    ▼ □ Correlation Analysis
        ▶ 20220420-114957 - First Experiment
    ▼ □ Data Preparation
        ▼ 20220420-115028 - Data Exploration
            Data                 : 20220420-115028_protprof-data-file.txt
            Analysis ID          : Data Exploration
            0s Missing           : Yes
            Transformation       : Log2
            Normalization        : Median
            Imputation           : Normal Distribution
            Columns for Analysis: 130-135
    ▼ □ Limited Proteolysis
        ▶ 20220420-115459 - JK10 0.1M
    ▼ □ Proteome Profiling
        ▶ 20220420-115533 - Exp1
        ▶ 20220420-115539 - Exp2
    ▼ □ Targeted Proteolysis
        ▶ 20220420-115618 - Cleavage sites
```

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 (**??**).

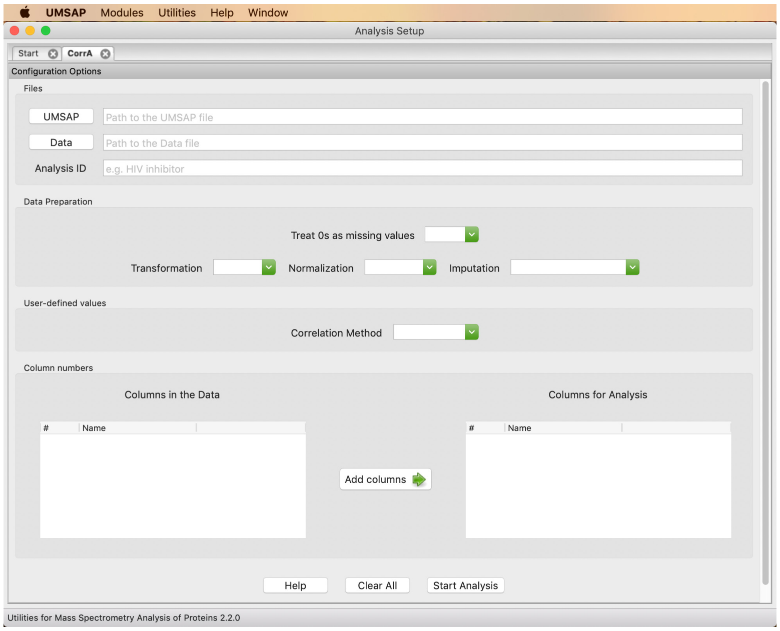


Figure 5.1: The Correlation Analysis tab. This tab allows to perform a correlation analysis of the data contained in a given Data file.

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 (**??**). 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 output

The correlation coefficients resulting from a Correlation Analysis will be shown as a color coded matrix (**??**). 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).

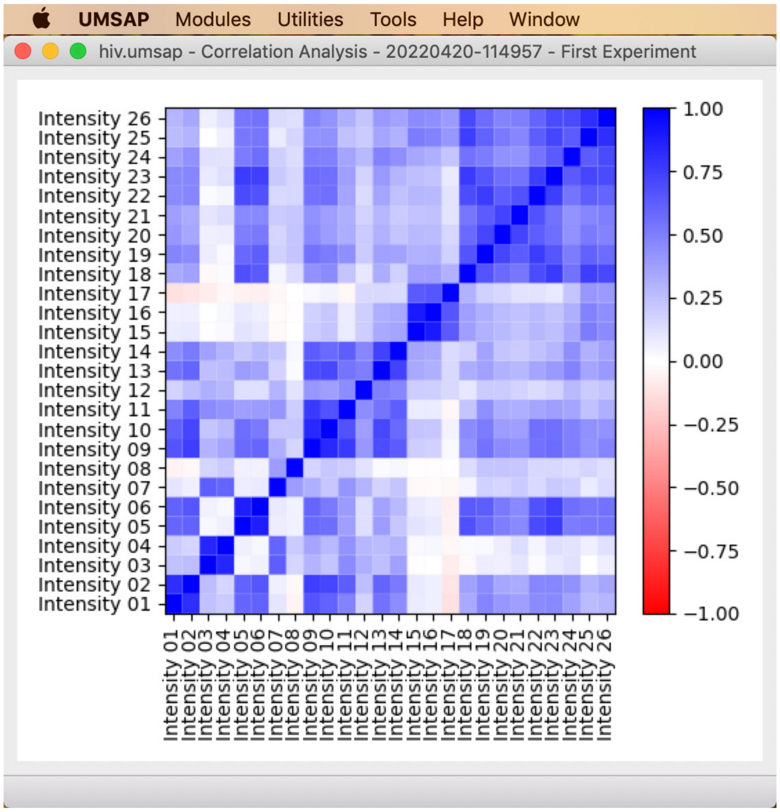


Figure 5.2: The Correlation Analysis result window. The correlation coefficients are shown as a color coded matrix. Values between −1 to 0 are shown in shades of red, 0 is shown in white and values between 0 to 1 in shades of blue. NA values are shown in green.

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 (**??**).

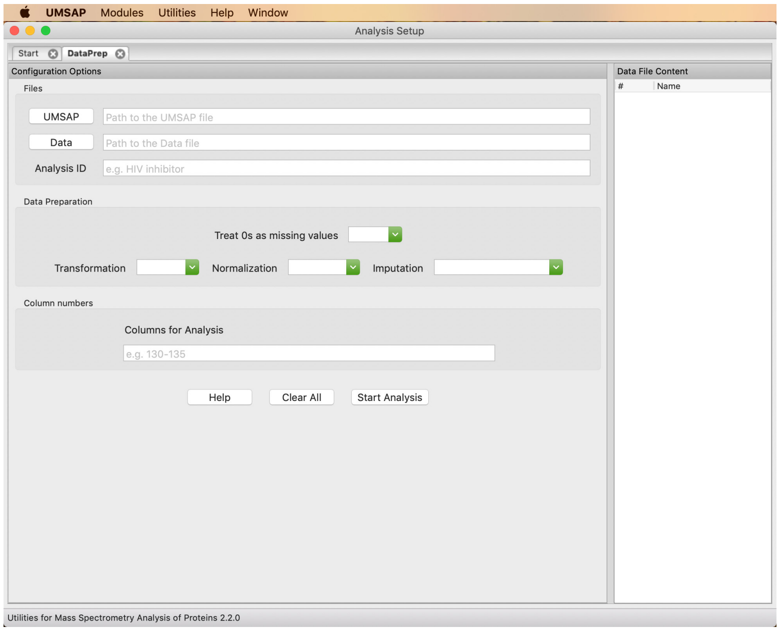


Figure 6.1: The Data Preparation tab. This tab allows to perform a statistical exploration of the data contained in a given Data file.

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

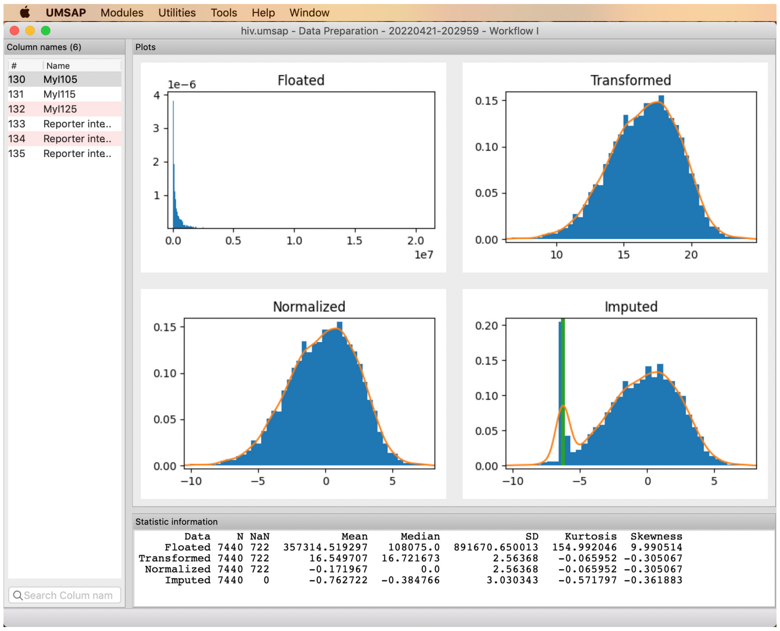
First, UMSAP will check the validity of the user provided input. After this, the selected Data File is read and the following steps are taken:

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 output

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



#	Name
130	Myl105
131	Myl115
132	Myl125
133	Reporter inte..
134	Reporter inte..
135	Reporter inte..

```
       Data    N NaN          Mean     Median            SD   Kurtosis  Skewness
    Floated 7440 722 357314.519297   108075.0 891670.650013 154.992046  9.990514
Transformed 7440 722     16.549707  16.721673       2.56368  -0.065952 -0.305067
 Normalized 7440 722     -0.171967        0.0       2.56368  -0.065952 -0.305067
    Imputed 7440   0     -0.762722  -0.384766      3.030343  -0.571797 -0.361883
```

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

The Limited Proteolysis module

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(**Limentani2005**) 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 **??**.
- *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 ??. 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 (??).

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.

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

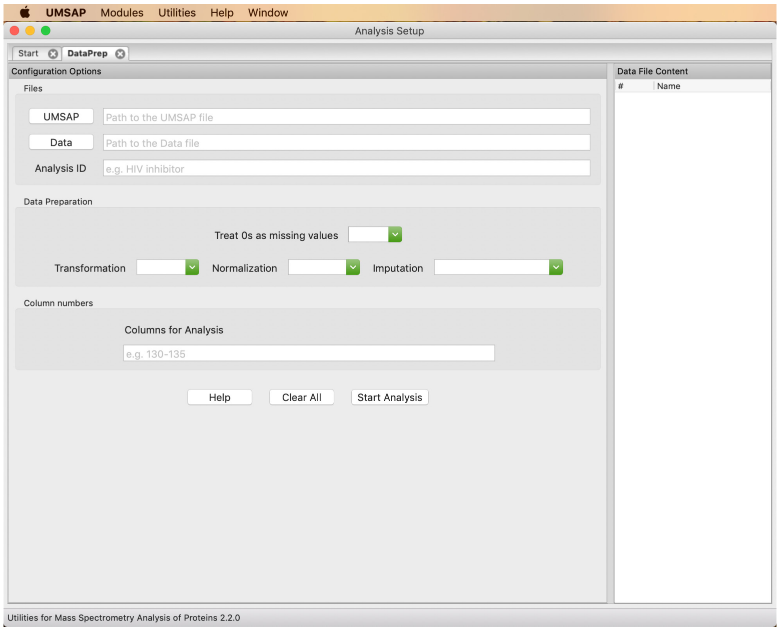


Figure 7.1: The Limited Proteolysis module tab. This tab allows users to perform the analysis of the results obtained during a two steps enzymatic proteolysis experiment where the products from the first limited digestions are separated using SDS-PAGE electrophoresis.

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 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(**Limentani2005**) performed to identify peptides in the selected gel spots with equivalent intensity values to the control spots (**??**). 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.

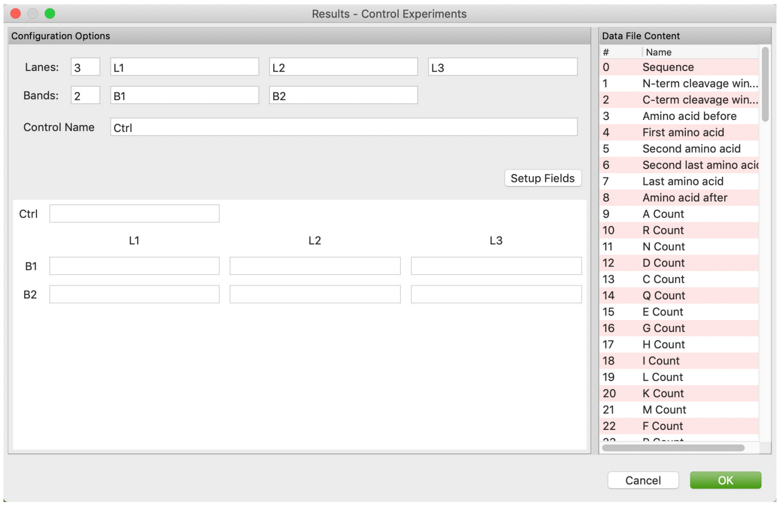
1. The text field Sequences allows users to specify the column in the Data file containing the sequences of the peptides identified in the MS experiments. Only one integer number equal or greater than zero will be accepted here.

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 (**??**) where users can type the information needed.

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



Lanes: 3 L1 L2 L3

Bands: 2 B1 B2

Control Name Ctrl

	L1	L2	L3
B1			
B2			

#	Name
0	Sequence
1	N-term cleavage win...
2	C-term cleavage win...
3	Amino acid before
4	First amino acid
5	Second amino acid
6	Second last amino acic
7	Last amino acid
8	Amino acid after
9	A Count
10	R Count
11	N Count
12	D Count
13	C Count
14	Q Count
15	E Count
16	G Count
17	H Count
18	I Count
19	L Count
20	K Count
21	M Count
22	F Count
23	P Count

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.

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. Then, the Data file is processed as follow. 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 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 (**Limentani2005**).

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 Values of Region 2 of the main window 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 **??** and then the acceptance criterion is calculated with **??**. Finally, the confidence interval for the mean difference for the gel spot and the control is calculated with **??** 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 Region 2 of the module's interface 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 to large and then the equivalence test is not able to detect the peptides with intensity values in the experiments lower than in the control.

If the sequence of the native protein is given the module 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 the module 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.

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

7.5 The output files

All the output generated by the Limited Proteolysis module will be contained in a LimProt folder created inside the selected Output folder. If the Output folder field in section *Files* of Region 2 of the interface is left empty, then the LimProt folder will be created in the directory containing the Data file. If the selected Output folder already contains a LimProt folder, then the current date and time to the second will be added to the name in order to avoid overwriting files from previous analyses. By default the LimProt folder will contain two files with extensions .limprot and .uscr and a Data_Steps folder. The name of these files is provided with the Output name field in section *Files* of Region 2 of the interface. Depending on the user provided input extra folders and files will be created inside the LimProt folder (**??**). For the rest of this chapter we will assume that the user provided name for the Output folder was *t*, the Output name was myLimTest, the Target protein was *Mis18alpha* and all optional analyses were performed.

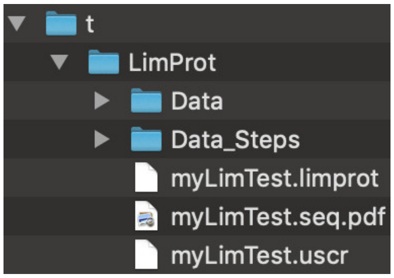
Information regarding the content and use of the .uscr file can be found in **??**.

If the parameter Sequence length is different than NA, then the file myLimTest.seq.pdf will be created. The file contains the sequence of the Target protein with the sequence of the FP highlighted. More details are given in **??**.

If the parameter Columns to extract is different than NA, then the folder Data will be created. The folder will contain several files that are shorter versions of the Data file. These files will contain only information regarding the Target protein. More details are given in **??**.

The folder Data_Steps contains a step by step account of all the calculations performed so users can check the accuracy of the calculation or perform further analysis. The files inside Data_Steps are plain text file with tab separated columns. The first line contains the name of the columns in the file.

A file containing a list of FP can also be generated as indicated in **??**.



```
▼ t
    ▼ LimProt
        ▶ Data
        ▶ Data_Steps
          myLimTest.limprot
          myLimTest.seq.pdf
          myLimTest.uscr
```

Figure 7.3: The structure of the Output folder from the Limited Proteolysis module. The folder Data and the file myLimTest.seq.pdf will be created only if requested.

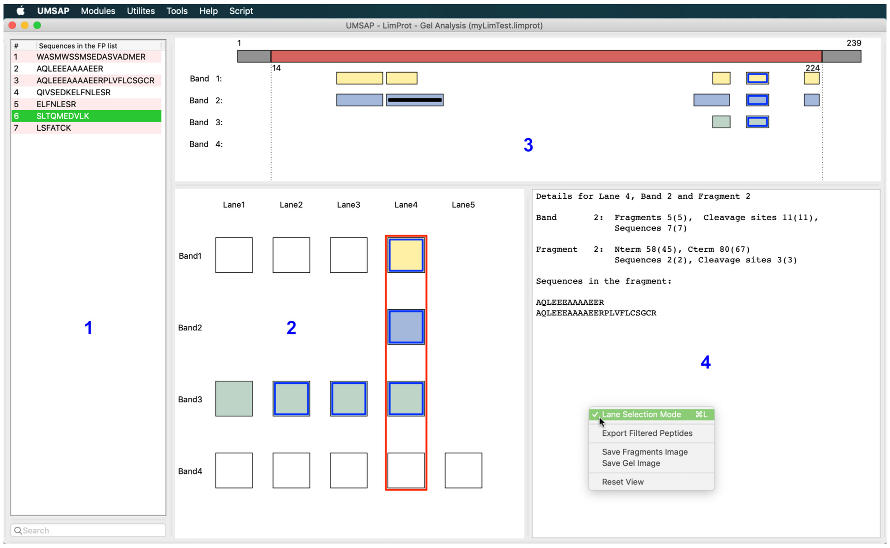
7.6 Visualizing the output files

The main output of the module is the .limprot file. This file contains the list of FP, all parameters values and all the information needed to visualize the results with UMSAP. After creating the file at the end of the analysis, the Limited Proteolysis module will automatically load the file and create a windows to display the results, see **??**.

The Gel analysis window is divided in four Regions.

Region 1 contains a list of all FP contained in the .limprot file being shown. The search box at the bottom allows to search for a sequence in the list of FP.

Region 2 contains a representation of the analyzed gel. Here, each gel spot is represented with square. When a square is not shown this means that the corresponding gel spot was not analyzed. Empty squares represent gel spot where no peptide from the Target



#	Sequences in the FP list
1	WASMWSSMSEDASVADMER
2	AQLEEEAAAAEER
3	AQLEEEAAAAEERPLVFLCSGCR
4	QIVSEDKELFNLESR
5	ELFNLESR
6	SLTQMEDVLK
7	LSFATCK

```
Details for Lane 4, Band 2 and Fragment 2

Band       2:  Fragments 5(5),  Cleavage sites 11(11),
               Sequences 7(7)

Fragment   2:  Nterm 58(45), Cterm 80(67)
               Sequences 2(2), Cleavage sites 3(3)

Sequences in the fragment:

AQLEEEAAAAEER
AQLEEEAAAAEERPLVFLCSGCR
```

Figure 7.4: The Gel Analysis window. Users can performed here the analysis of the fragments obtained in the limited proteolysis experiments.

protein was detected with intensity values equivalent to the controls. The rest of the square will be colored according to the band they belong to or the lane. There are two selection modes available for Region 2. In the Lane selection mode selecting one of the Gel spot will also select the entire lane containing the selected gel spot. 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 right mouse button, the Tools menu or the keyboard shortcut Ctrl/Cmd+L can be used to change the selection mode. In the Band selection mode, selecting a gel spot will highlight the band containing the gel spot and the gel spots are colored by lane. Selecting a band or a lane in Region 2 will display information about the band/lane in Regions 3 and 4.

Region 3 will display a graphical representation of the fragments found in each gel spot for the selected band or lane. The first fragment in this region represent 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 peptide from the list box in Region 1 will highlight with a blue border the gel spots in Region 2 where the peptide is found. If a lane/band in Region 2 is already selected, then the fragments shown in Region 3 that contains the selected peptide in the list box will also be highlighted with a blue border.

Region 4 will show information about the selected lane/band or gel spot in Region 2 or the selected fragment in Region 3. 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 3 will display in Region 4 the following information: number of cleavage sites and fragments, first and last residue number for the selected fragment and a sequence alignment of all peptides forming the fragment.

7.6.1 The Tools menu

The Tools menu for the window allows changing the selection mode in Region 2, export the data shown in the window (see **??**), save an image of Region 2 or 3 and to reset the view of the window.

The file containing the exported list of FP will have a tabular format with tab delimited columns. Each row of the file will contain a FP. The columns in the file contain information about the N and C residue numbers, the sequence and Score of the FP. In addition, the results of the equivalence test for each gel spot is given as 0 or 1 value, with 0 meaning not equivalent. The file will have a .txt extension and can be viewed with a simple text editor or Excel.