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.

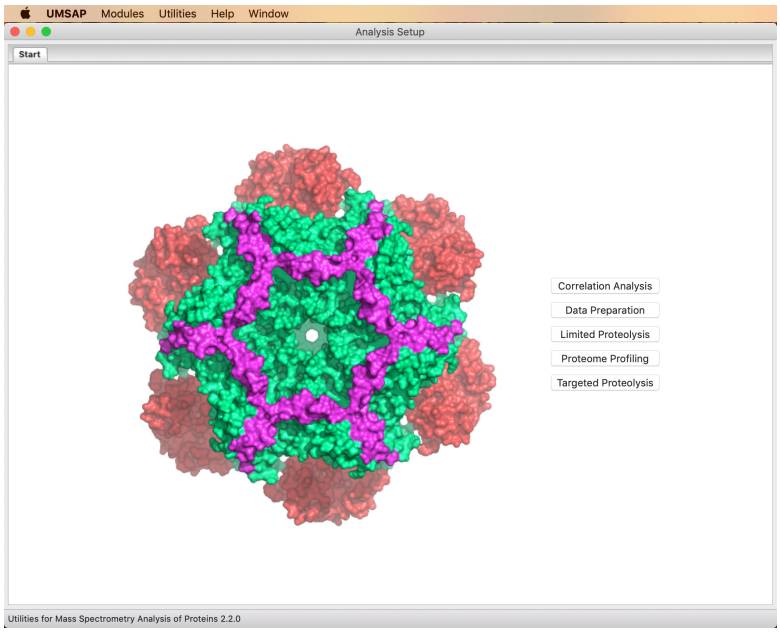

Figure 3.1: The main window of UMSAP. From this window users can access all the available Modules and Utilities.

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.

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.

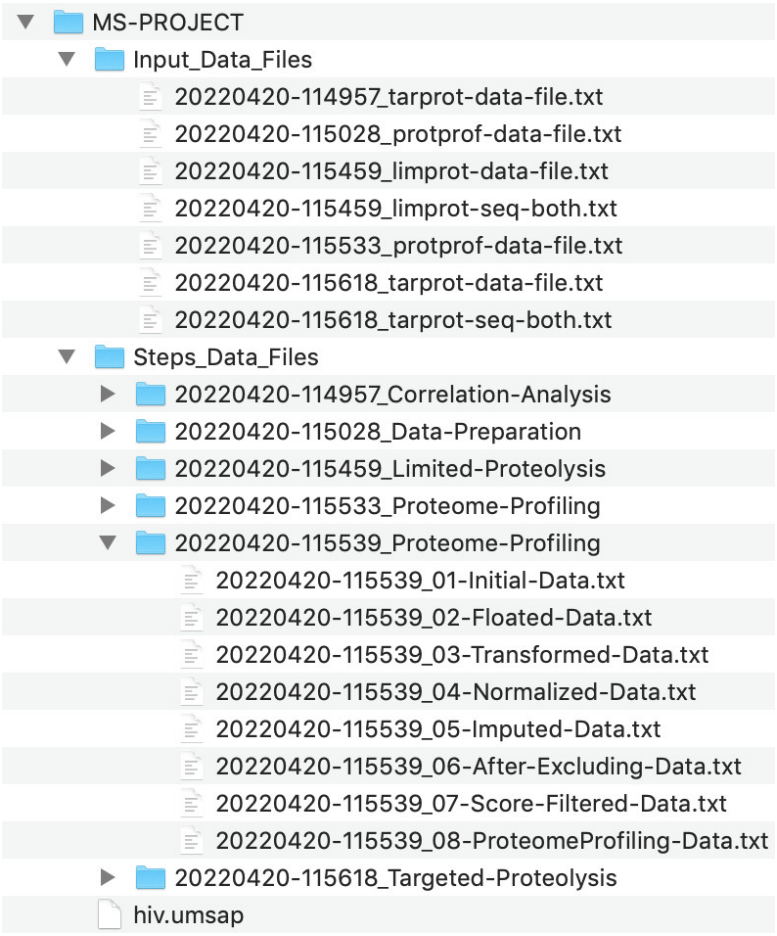

Figure 3.2: Structure of the output generated by UMSAP. Results are saved in the Steps_Data_Files folder. The .umsap file allows managing and visualizing the results.

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+P	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+P	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.

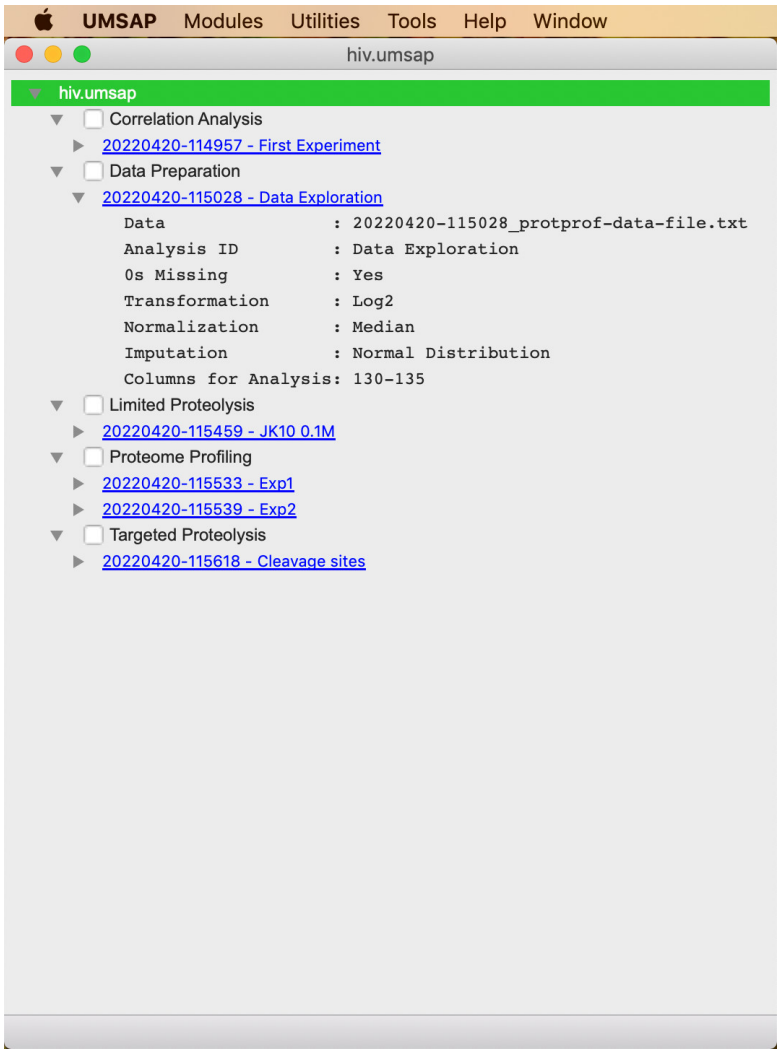

Figure 4.1: The UMSAP Control window. The content of the selected .umsap file is shown in alphabetically order. The window allows managing the content of the .umsap file and to visualize the results of the analysis in the file.

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).

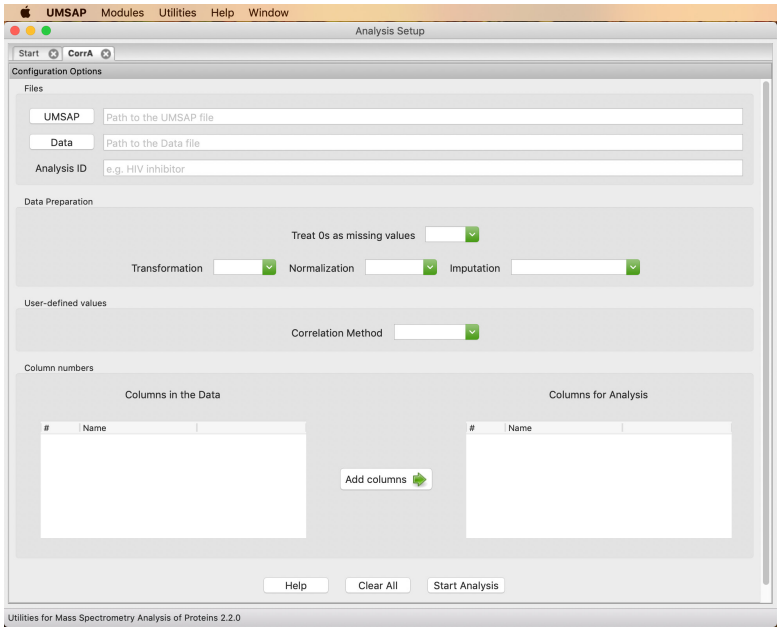

Figure 5.1: The Correlation Analysis tab. This tab allows to perform a correlation analysis of the data contained in a given Data file.

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+P 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 output

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).

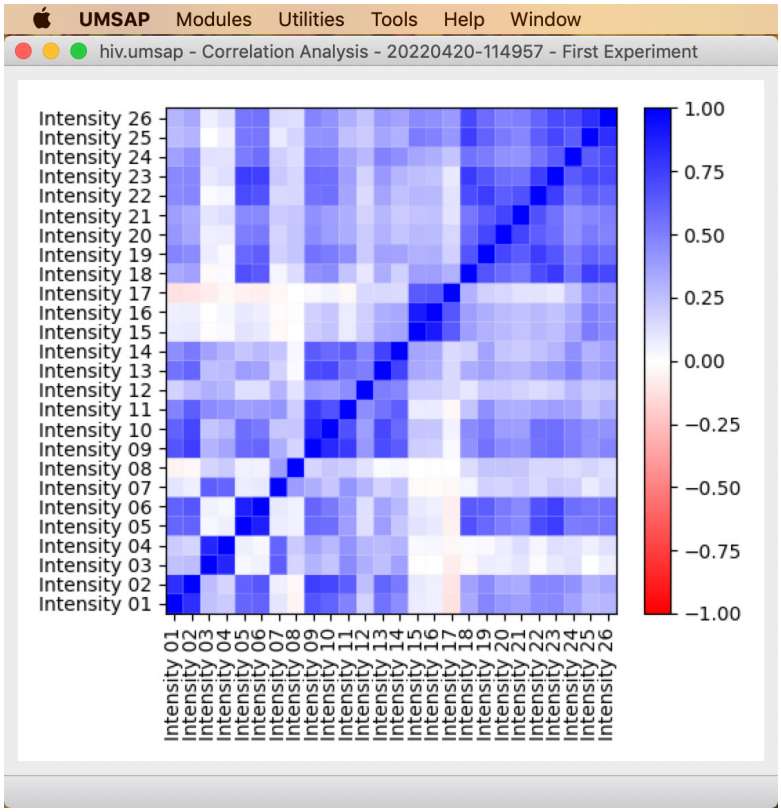

Figure 5.2: The Correlation Analysis result window. The correlation coefficients are shown as a color coded matrix. Values between −1 to 0 are shown in shades of red, 0 is shown in white and values between 0 to 1 in shades of blue. NA values are shown in green.

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).

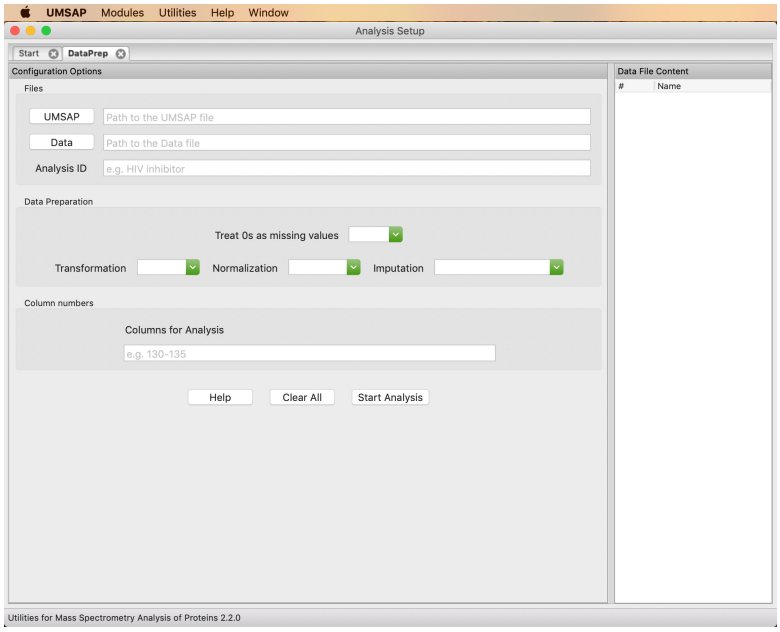

Figure 6.1: The Data Preparation tab. This tab allows to perform a statistical exploration of the data contained in a given Data file.

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+P) 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 (Figure 6.2).

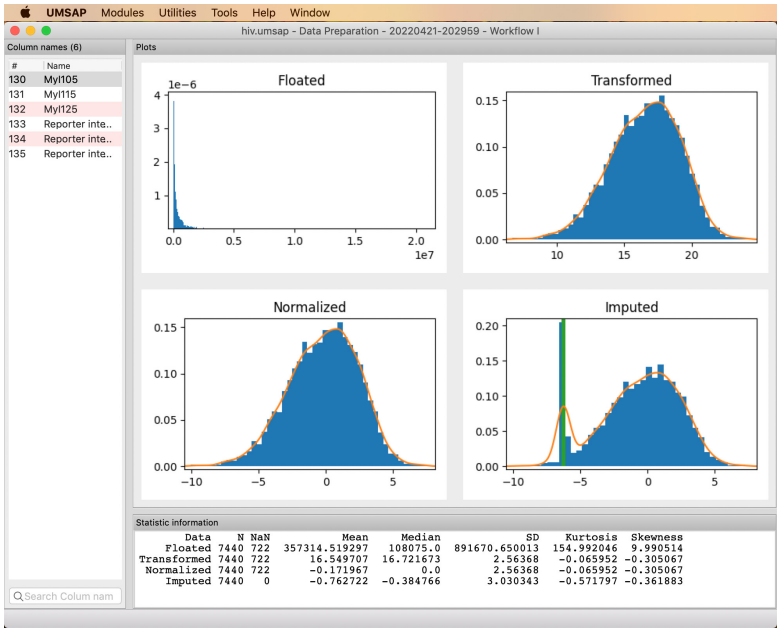

Figure 6.2: The Data Preparation result window. Histograms for the initial, transformed, normalized and imputed data are shown.

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).