Utilities for Mass Spectrometry Analysis of Proteins

User's Manual

Version 2.1.0

May 2020

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

Contents

List of Figures

List of Tables

Chapter 1

Utilities

Currently, there are 15 utilities. Users can access the utilities in two ways. From the main interface (**??**) users can select Utilities in the list to the right and a new window will appear with a complete list of available utilities (Figure 1.1). The alternative option is to directly select the desired utility from the menu entry, Utilities. The second approach is faster since does not require to use the Utilities window (Figure 1.1).

The utilities are organized in General Utilities and utilities that are specific for a given module. The following sections describe each one the implemented utilities.

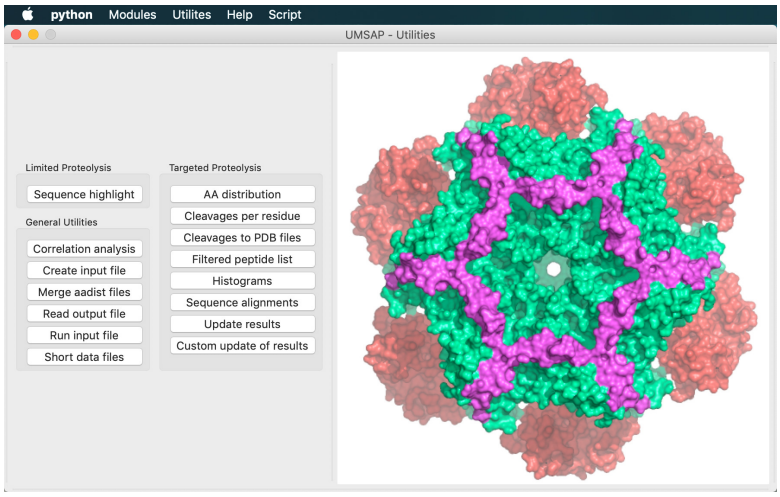

Figure 1.1: The Utilities window. From this window users can access all the available utilities.

1.1 Limited Proteolysis utilities

1.1.1 Sequence highlight

The Sequence Highlight utility allows user to highlight on the sequence of the Target protein the peptides detected in each gel lane/band considered in the .limprot file. The sequences with the highlighted peptides are saved in a .pdf file. How to generate a .limprot file is discussed in **??**.

The interface

The window of the Sequence Highlight utility is divided in three regions.

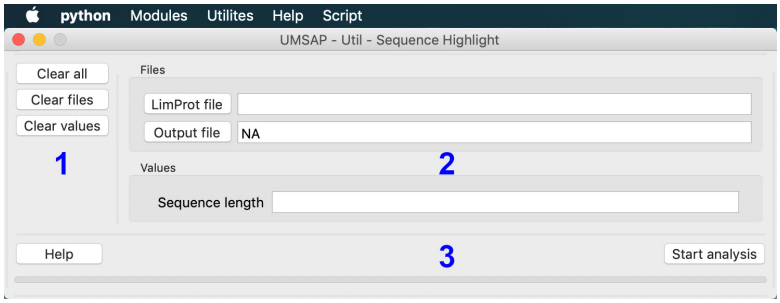

Figure 1.2: The Sequence Highlight window. This window allows to generate a .pdf file showing the location of the peptides detected in the investigated gel lane/band on the sequence of the Target protein.

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 the path to all user provided files. Finally, the Clear values button will delete all user provided numerical values.

Region 2 contains the fields where users provide the information needed in order to create the .pdf file with the highlighted sequences. The Limprot file button allows users to browse the file system to select the .limprot file that will be used for the generation of the .pdf file. Only one .limprot file can be provided here. The Output file button allows users to browse the file system to select the location and name of the .pdf file. If left empty, then the .pdf file, resulting from the analysis, will be saved in the same directory containing the .limprot file and will have the same name as the .limprot file. If the folder containing the selected .limprot file already contains a .pdf file with the same name as the selected .limprot file, then UMSAP will add the current date and time to the seconds to the end of the .pdf file name in order to avoid overwriting the older .pdf file without explicit user permission.

The Sequence length parameter allows user to define the maximum number of residues per line to be used during the creation of the .pdf file. The value here must be an integer number greater than zero.

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 generation of the .pdf file. The progress bar will give a rough estimate of the remaining time for

completing the analysis.

The analysis

First, UMSAP will check the validity of the user provided input. After this, for each gel lane/band analyzed in the .limprot file, UMSAP will print the sequence of the Target protein to the .pdf file with the peptides found in the gel lane/band highlighted.

The output

The output is a .pdf file containing a page for each gel lane/band found in the .limprot file. As described before, each page contains the sequence of the Target protein with the detected peptides highlighted in red. In addition, the residue number of the beginning and ending of the highlighted fragments are also given. If the sequence of the native protein was provided when creating the .limprot file, then the information is given for the recombinant and native protein.

The Tools menu

This utility does not have a Tool menu.

1.2 Targeted Proteolysis utilities

1.2.1 AA Distribution

The AA Distribution utility allows user to calculate the AA distribution around the detected cleavage sites using a list of FP (see page **??**). The list of FP is automatically generated from a .tarprot file. How to generate a .tarprot file is discussed in **??**. The list of FP is a non redundant list.

The interface

The window of the AA Distribution utility is divided in three regions (Figure 1.3).

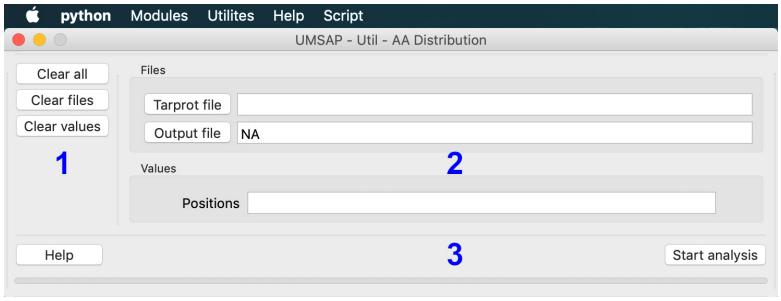

Figure 1.3: The AA Distribution window. This window allows to obtain the AA distribution around the detected cleavage sites from a .tarprot file.

Region 1 contains three buttons allowing users to quickly delete all provided input and start a new distribution analysis. The Clear all button will delete all user provided input. The Clear files button will delete the path to all user provided files. Finally, the Clear values button will delete all user provided numerical values.

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

perform the AA distribution calculation. The Tarprot file button allows users to browse the file system to select the .tarprot file that will be used for the analysis. Only one .tarprot file can be provided here. The Output file button allows users to browse the file system to select the location and name of the Output file. If left empty, then the .aadist file, resulting from the analysis, will be saved in the same directory containing the .tarprot file and will have the same name as the .tarprot file. If the folder containing the selected .tarprot file already contains a .aadist file with the same name as the selected .tarprot file, then UMSAP will add the current date and time to the seconds to the end of the .aadist file in order to avoid overwriting the older .aadist file without explicit user permission.

The parameter Positions indicates the number of positions around the cleavage sites to be analyzed. The value here must be an integer number greater than zero.

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 a rough estimate of the remaining time for completing the analysis.

The analysis

First, UMSAP will check the validity of the user provided input. After this, a list of FP is created using the .tarprot file. For each FP, the sequence around the N and C terminal ends of the peptide is analyzed up to the user provided number of Positions. 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 contained in the .tarprot file. 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 significance level found in the .tarprot file. In order to be able to perform the χ^2 test, AAs are pooled together in the same groups as described for the color code used in the output.

The output

The output from the AA distribution calculation is a file with .aadist extension. The file will be automatically loaded and a graphical representation of the results will be shown (Figure 1.4). There are two graphical representations. The first representation shows

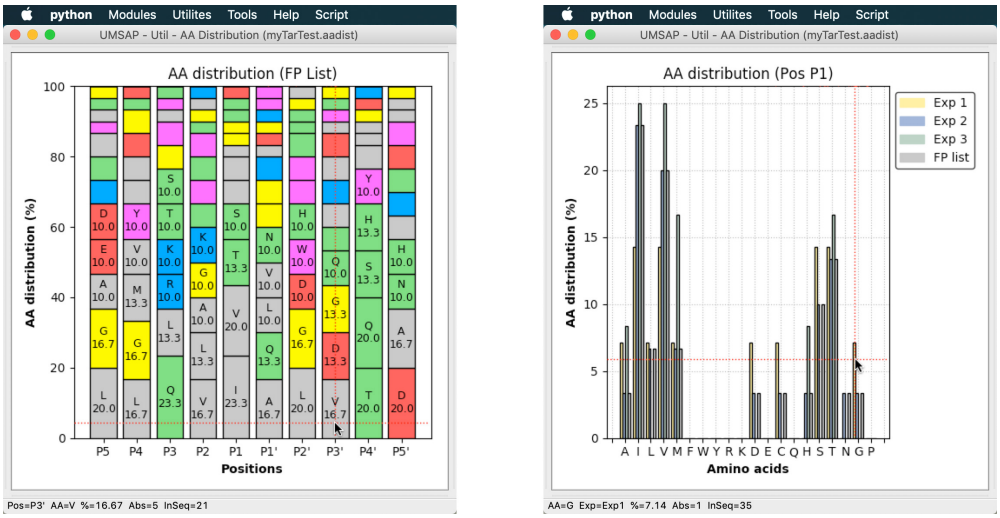

Figure 1.4: The AA Distribution analysis window. This window allows to visualize the results contained in a .aadist file.

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 1.4 the value of 16.7 % obtained for A in position $P1'$ means that A was found in position $P1'$ in the 16.7 % 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 found in the .tarprot file. A red color represents that the distributions are not significantly different at the level of significance found in the .tarprot file. 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 (**Yates1999**).

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 appears 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 appears in the sequence of the recombinant protein (InSeq).

The second representation allows to compare 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 appears 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 appears in the sequence of the recombinant protein (InSeq).

The Tools menu

By default, the AA distribution for all AA in the FP list will be shown when the .aadist file is loaded. The Tools menu in the window allows user to change the displayed experiment or to select the position for which results in the experiments should be compared. In addition, users may select to save a figure of the plot and to reset the view.

1.2.2 Cleavages 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 **??**). How to generate the .tarprot file is discussed in **??**. 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 1.5).

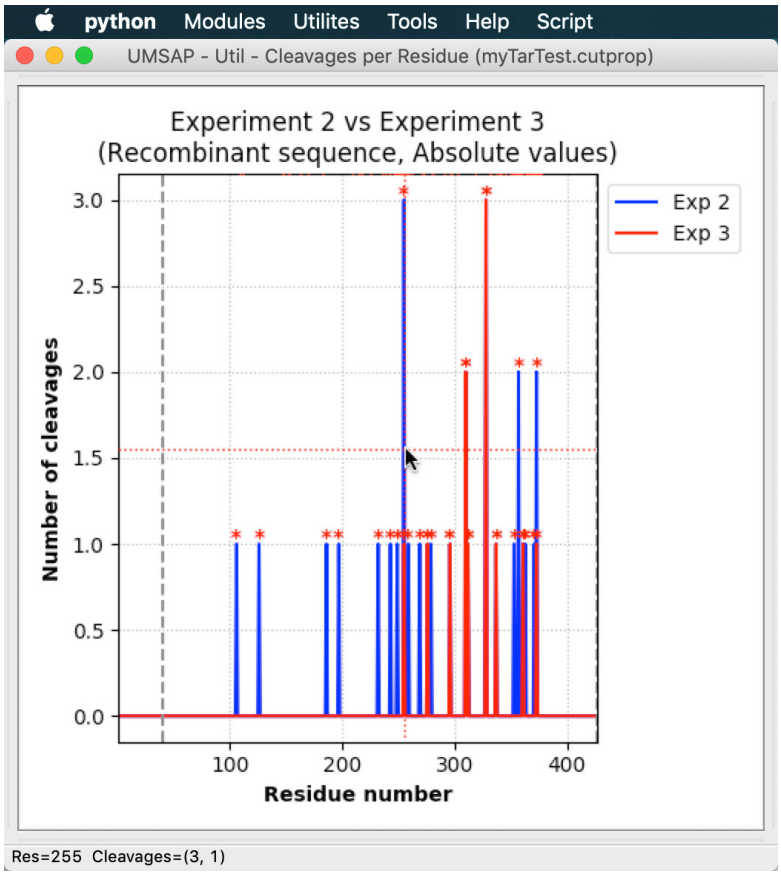

Figure 1.5: The Cleavages per Residue analysis window. This window allows to visualize the results contained in a .cutprop file.

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 1.5). 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.

1.2.3 Cleavages to PDB Files

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 **??**). How to generate the .tarprot file is discussed in **??**. The FP list is a non redundant list.

The interface

The Cleavages to PDB Files window is divided in three regions (Figure 1.6).

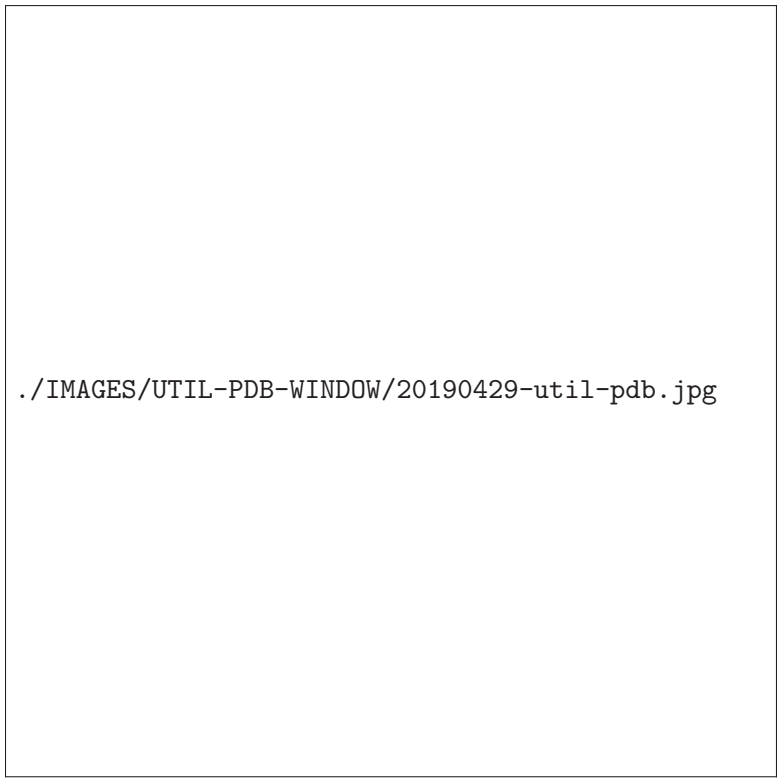

Figure 1.6: The Cleavages to PDB Files window. This window allows to map the detected number of cleavages per residue to a pdb file containing the structure of the Target protein. The number of cleavages are mapped to the beta field of the pdb.

Region 1 contains three buttons allowing users to quickly delete all provided input and start a new distribution analysis. The Clear all button will delete all user provided input. The Clear files button will delete the path to all user provided files. Finally, the Clear values button will delete all user provided numerical values.

Region 2 contains the fields where users provide the information needed in order to 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 data base. 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 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 data base. 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 buttons and a progress bar. The Help button leads to an online tutorial while the Start button will start the analysis. The progress bar will give a rough estimate of the remaining time for completing the analysis.

The analysis

First, UMSAP will check the validity of the user provided input. Then, a temporal .cutprop file will be created. 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.

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.

1.2.4 Create Input File

The Create Input File utility allows user to read a .tarprot file and to create an input file. How to generate the .tarprot file is discussed in **??**. The input file is used to configure a module without users having to type in all the information required by the module. Thus, if a second analysis of a data file is needed users can quickly load the input file into UMSAP, apply the required modifications in the window of the module and Start the analysis without having to type in or modify the options that will be the same between the old and new analysis. If created by hand, the input file can be used to run a module avoiding typing the information into the window of the module. The extension .uscr is reserved for the input files. Each .uscr file can contain information for configuring one module for one analysis.

The interface

The Create Input File 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. Then the .tarprot file will be read in and the configuration values will be extracted and saved in the .uscr file. The utility is able to process .tarprot files from previous versions of UMSAP.

The output

The input file has a simple format in which each line has a keyword and an argument. The keyword and the argument are separated by a semicolon (;). The following is an example for the format of the .uscr file and the keywords available for the Targeted Proteolysis module:

module; TarProt
Data file; /Users/bravo/BORRAR-GUI/DATA/Mod-Enz-Dig-data-ms.txt
Sequence (rec); /Users/bravo/BORRAR-GUI/DATA/Mod-Enz-Dig-data-seq.txt
Sequence (nat); P31545
PDB file; NA
Output folder; /Users/bravo/Desktop/t
Target protein; efeB
Score value; 50
Data normalization; Log2
a-value; 0.05
Positions; 5
Sequence length; 100
Histogram windows; 50
PDB ID; 2yf4:A
Sequences; 0
Detected proteins; 38
Score; 44
Columns to extract; 0 1 2 3 4-10
Control experiments; 98-105
Results; 109-111, 112 113 114, 115-117 120

The menu entry Script/Run Input File allows to load the input file into UMSAP.

1.2.5 Filtered Peptide List

The Filtered Peptide List utility allows users to read a .tarprot file and to create a file containing the FP list. How to generate the .tarprot file is discussed in **??**. The FP list is a non redundant list of all peptides identified during the Targeted Proteolysis analysis. The extension .filtpept is reserved for a file containing the FP list.

The interface

The Filtered Peptide List 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. Then the .tarprot file will be read in and all FP will be saved in the output file.

The output

The .filtpept file has a tabular format in which the first row contains the name of the columns and columns are tab separated. It is a plain text file that can be read with any text editor.

1.2.6 Histograms

The Histograms utility allows to create histograms of the identified cleavage sites using the residue numbers of the target protein as the definition of the windows in the histograms. Histograms are created from a .tarprot file. How to generate the .tarprot file is discussed in **??**. Only FP are used to create the histograms, see page **??**. The list of FP is a non redundant list.

The interface

The Histograms window is divided in three regions (Figure 1.7).

Region 1 contains three buttons allowing users to quickly delete all provided input and start a new calculation. The Clear all button will delete all user provided input. The Clear files button will delete the path to all user provided files. Finally, the Clear values button will delete all user provided numerical values.

Region 2 contains the fields where users provide the information needed in order to create the histograms. The Tarprot file button allows users to browse the file system to select the .tarprot file that will be used for the analysis. Only one .tarprot file can be selected here. The Histograms utility generates two files that will be saved in a folder named Histograms. The Output folder button allows users to browse the file system to select the location of the Histograms folder. If no Output folder is selected the folder Histograms will be created in the same directory as the selected .tarprot file. If an older Histograms folder exists in the selected Output folder, UMSAP will create a new Histograms folder with the date and time added to the end of the name in order to avoid overwriting any file.

The parameter Windows lengths allows to define the length of the windows in the histograms. Values here are expected to be integers greater than zero. A single value will results in equally spaced windows covering the entire length of the recombinant protein under study. Several values will result in custom sized windows. This may be useful if users want to define windows matching a structure related property of the target protein e.g. secondary structure. In this case, the values must be space separated and organized from lower to higher values. For example, the input 150 100 200 220 will create four windows covering residues 1 to 49, 50 to 99, 100 to 199 and 200 to 219.

Region 3 contains the Help and Start buttons and a progress bar. The Help button leads to an online tutorial while the Start button will start the analysis. The progress

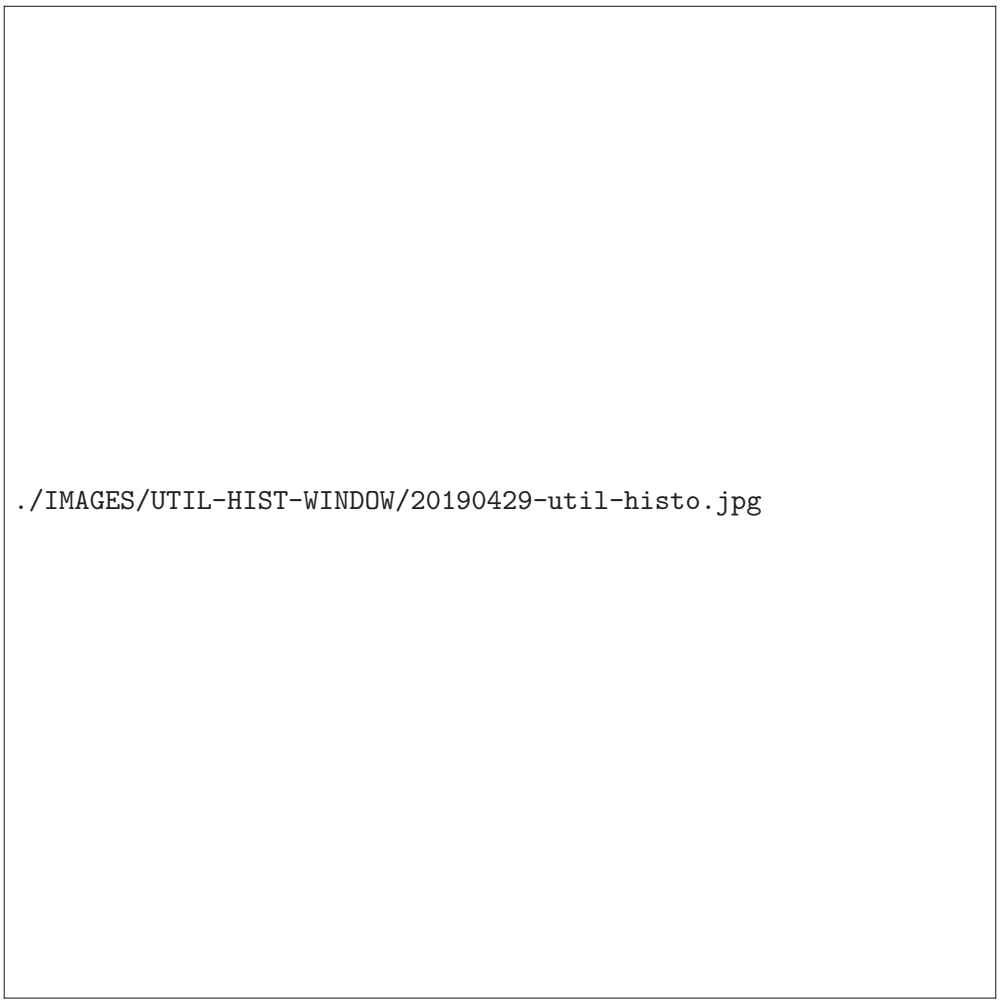

Figure 1.7: The Histograms window. This window allows to create histograms of the identified cleavage sites using the residue numbers of the target protein as the definition of the windows in the histograms.

bar will give a rough estimate of the remaining time for completing the analysis.

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 .tarprot file the detected cleavage sites will be assigned to the corresponding windows. Cleavage sites are only counted once per experiment, independently of how many peptides share the same cleavage site. In addition, the total number of cleavage sites identified considering the results of all experiments is also calculated. Since most of the time the protein under study is a recombinant protein containing purification tags or only a region of the native protein, histograms are created for the residue numbers in the recombinant protein and for the residue numbers in the native protein. For this to be possible the sequence of the native protein must have been provided during the creation of the .tarprot file. In the last case cleavage sites outside the native sequence contained in the recombinant

protein are discarded and the residue numbers used for the definition of the windows and cleavages sites are the residue numbers of the native sequence. Both files will be saved inside a Histograms folder in the user specified location in Output folder. After the analysis is done the file containing the histograms for the recombinant sequence will be automatically loaded and the results shown in a new window (Figure 1.8).

The output

The Histograms analysis window will display the results contained in a .hist file (Figure 1.8). The extension .hist is reserved for the histograms files. In the histogram, experiments are shown in the order specified when creating the .tarprot file. In addition, the values for the histograms considering the results from all experiments (all FP) will be displayed as the last bar and colored in gray. 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 unique cleavages (Cleavages).

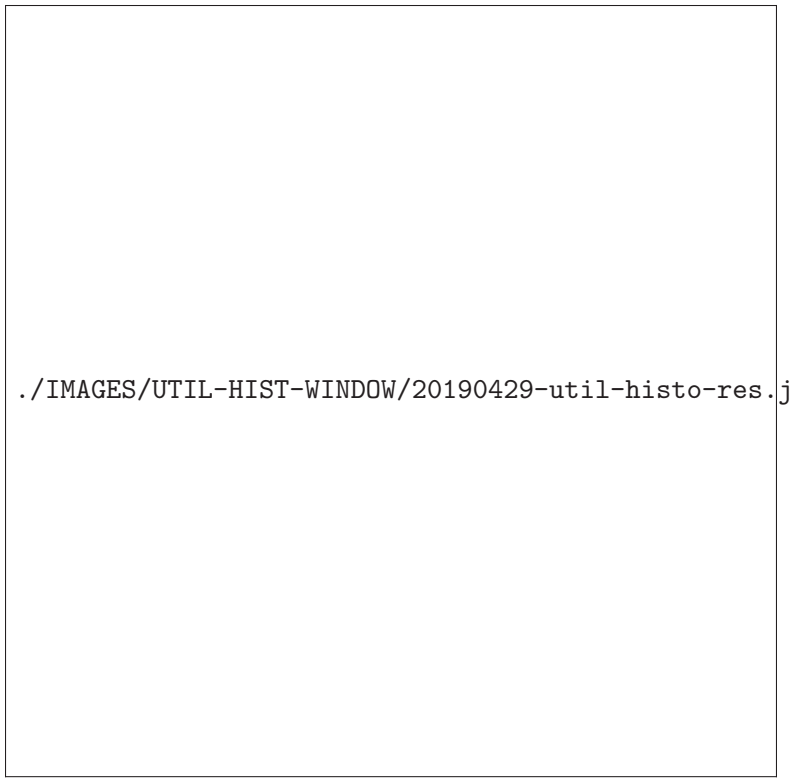

Figure 1.8: The Histograms analysis window. This window allows to visualize the results contained in a .hist file.

The Tools menu

The Tools menu allows to save an image of the plot.

1.2.7 Sequence Alignments

The Sequence Alignments utility generates sequence alignments between the FP for each experiment and the sequence of the recombinant protein. The list of FP is generated from a .tarprot file (see page **??**). The list of FP is a non redundant list. How to generate the .tarprot file is discussed in **??**.

The interface

The Sequence Alignments window is divided in three regions (Figure 1.9).

Region 1 contains three buttons allowing user to quickly delete all provided input and start a new calculation. The Clear all button will delete all user provided input. The Clear files button will delete the path to all user provided files. Finally, the Clear values button will delete all user provided numerical values.

Region 2 contains the fields where users provide the information needed in order to generate the alignments. The Tarprot file button allows users to browse the file system to select the .tarprot file that will be used for the analysis. Only one .tarprot file can be provided here. The Sequence alignments utility generates multiple files that will be saved in a folder named Sequences. The Output folder button allows users to browse the file system to select a location for the output folder Sequences. If the Output folder option is left empty, the output folder Sequences will be created in the same directory as the .tarprot file. If there is a Sequence folder in the selected Output folder, then UMSAP will create a new Sequences folder with the date and time to the seconds added to the end of the name in order to avoid overwriting any file.

The parameter Sequence length allows to define the maximum number of residues per line in the short version of the sequence alignment files. The value here is expected to be an integer greater than zero.

Region 3 contains the Help and Start buttons and a progress bar. The Help button leads to an online tutorial while the Start button will start the analysis. The progress bar will give a rough estimate of the remaining time for completing the analysis.

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 and UMSAP will generate the sequence alignments.

The output

The output of the Sequence alignments utility is composed of several files that will be saved inside a folder named Sequences. Each file will be a plain text file containing an alignment. Alignments will be generated for each experiment and for the entire FP list. The sequences of the FP in each file will be N-terminally organized. Files for the recombinant and native sequences are generated. In addition, files containing one sequence per line or the specified maximum number of residues per line are also created. The sequence alignment files can be viewed with any text editor since they are just plain text files.

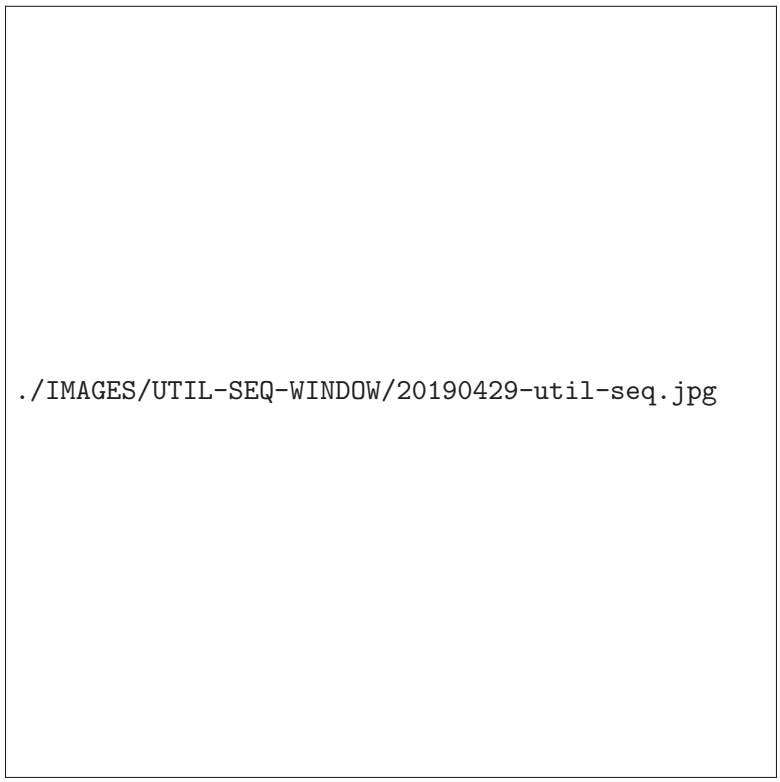

Figure 1.9: The Sequence Alignments window. This window allows to generate sequence alignment files between the FP of each experiment and the sequence of the recombinant protein under study.

1.2.8 Short Data Files

The Short Data Files utilities allows users to create short versions of the Data file used to create a .tarprot file. How to generate the .tarprot file is discussed in **??**.

The interface

The Short Data Files window is divided in three regions (Figure 1.10).

Region 1 contains three buttons allowing user to quickly delete all provided input and start a new analysis. The Clear all button will delete all user provided input. The Clear files button will delete the path to all user provided files. Finally, the Clear values button will delete all user provided numerical values.

Region 2 contains the fields where users provide the information needed in order to generate the short data files. The Tarprot file button allows users to browse the file system to select the .tarprot file that will be used for the analysis. Only one .tarprot file can be provided here. The Data file button allows users to browse the file system to select the data file that will be used for the analysis. The Short Data File utility generates multiple files that will be saved in a folder named Data. The Output folder button allows users to browse the file system to select a location for the output folder Data. If the Output folder option is left empty, the output folder Data will be created

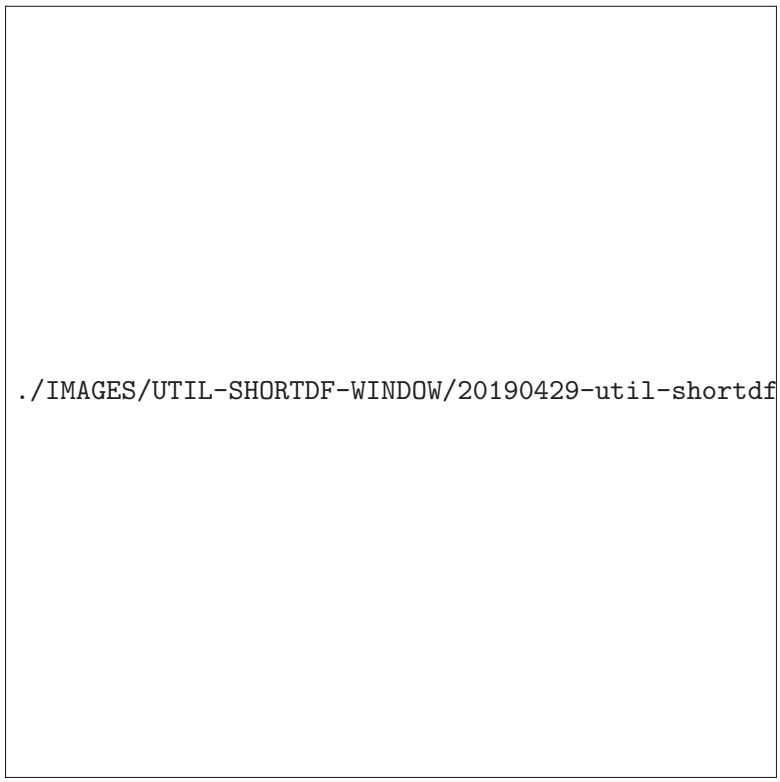

Figure 1.10: The Short Data File window. This window allows to generate smaller versions of the Data file used to generate a .tarprot file.

in the same directory as the .tarprot file. If there is a Data folder in the selected Output folder, then UMSAP will create a new Data folder with the date and time to the seconds added to the end of the name in order to avoid overwriting any file.

The parameter Columns to extract allows to define which columns from the data file will be extracted to the short data files. The values here are expected to be integers greater than zero.

Region 3 contains the Help and Start buttons and a progress bar. The Help button leads to an online tutorial while the Start button will start the analysis. The progress bar will give a rough estimate of the remaining time for completing the analysis.

The analysis

First, UMSAP will check the validity of the user provided input. If only a .tarprot file is given, then the data file location will be read from the .tarprot file. After this, the short data files will be written to the Data folder.

The output

The output consist of three files that will be saved in a Data folder. Assuming that the name of the target protein is efeB, the name of the files will be:
all-columns-all-efeB-records.txt
selected-columns-all-efeB-records.txt

selected-columns-relevant-efeB-records.txt

These files are just shorter versions of the data file containing only relevant information about the target protein. They are plain text files with a tabular format. The first row contains the name of the columns and columns are tab separated. The file all-columns-all-efeB-records.txt contains the same number of columns as the data file but only the rows for the target protein. The file selected-columns-all-efeB-records.txt contains only rows for the target protein and the columns specified in Columns to extract. The selected-columns-relevant-efeB-records.txt file is similar to the previous file but contains only the relevant peptides of the target protein. These files can be viewed with any text editor.

1.2.9 Update Results

As discussed in **??**, UMSAP can only read the .tarprot file from previous versions. The Update Results utility offers a way to quickly generate files for the optional analyses allowed in the Targeted Proteolysis module that are compatible with the current version of UMSAP.

The interface

The Update Results 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 folder. That is all.

The analysis

UMSAP will read the .tarprot file and will perform the optional analysis specified in the .tarprot file. This will result in the creation of up to date files for the optional analyses specified in the .tarprot file. The up to date files can be viewed with the current version of UMSAP.

The output

UMSAP will generate the files discussed in the previous sections as required by the specified optional analyses found in the given .tarprot file. All generated files will be saved in a TarProtUpdate folder created inside the specified Output folder. If there is already a TarProtUpdate folder in the Output folder, then the current date and time to the seconds will be add to the folder name in order to avoid overwriting previous files.

1.2.10 Custom Update of Results

The Custom Update of Results utility is similar to the Update Results utility because they both allows to use a .tarprot file from an older version of UMSAP to generate files for the optional analyses available in the Targeted Proteolysis module that are compatible with the current version of UMSAP. The main difference is that with Custom Update of Results a custom update can be done.

The interface

The Custom Update of Results utility does not have a window. When the utility is

selected users will be asked to select a .tarprot file and then the interface for the Targeted Proteolysis module is created and the information found in the selected .tarprot file is used to fill the fields in the interface of the module, see **??**.

The analysis

After the interface for the Targeted Proteolysis module is created and filled with the information found in the selected .tarprot file, users may modify the read values or add information to perform optional analyses that were not performed with the previous versions of UMSAP, see **??** for more details.

The output

The output generated depends on the options given to the Targeted Proteolysis module, see **??** for details.

1.3 General Utilities

1.3.1 Correlation Analysis

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

The interface

The Correlation Analysis window is divided in four regions (Figure 1.11).

Region 1 contains three buttons allowing users to quickly delete all provided input and start a new calculation. The Clear all button will delete all user provided input and will reset the state of the list boxes in Region 3. The Clear files button will delete the path to the user provided files and will reset the state of the list boxes in Region 3. Finally, the Clear values button will delete all user provided values.

Region 2 contains the fields where users provide the information needed in order to calculate the correlation between the data. The Data file button allows users to browse the file system to select the 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 provided here. The Output file button allows users to browse the file system to select the location and name of the output file. If left empty, the name of the output file will be the same as the data file and will be saved in the same folder containing the data file.

The Data normalization list allows users to select a normalization algorithm to be performed before the correlation analysis. Currently, the possible options are a Log_2 normalization or no normalization. The list will we expanded in the near future.

Region 3 contains two list boxes and a button. The list box to the left will display the names of the columns present in the data file, once the data file is selected. Loading of the column names is automatically done after selecting the data file using the Data file button in Region 1 or pressing the Enter key while the text box has the focus of the

keyboard. Columns in the left list box can not be deleted, except in the case of loading a different data file or using the Clear input or Clear all buttons in Region 1. The Add columns button in the middle of Region 3 will add the selected columns in the left list box to the right list box. The columns will be added to the right list box in the same order as they are selected from the left list box.

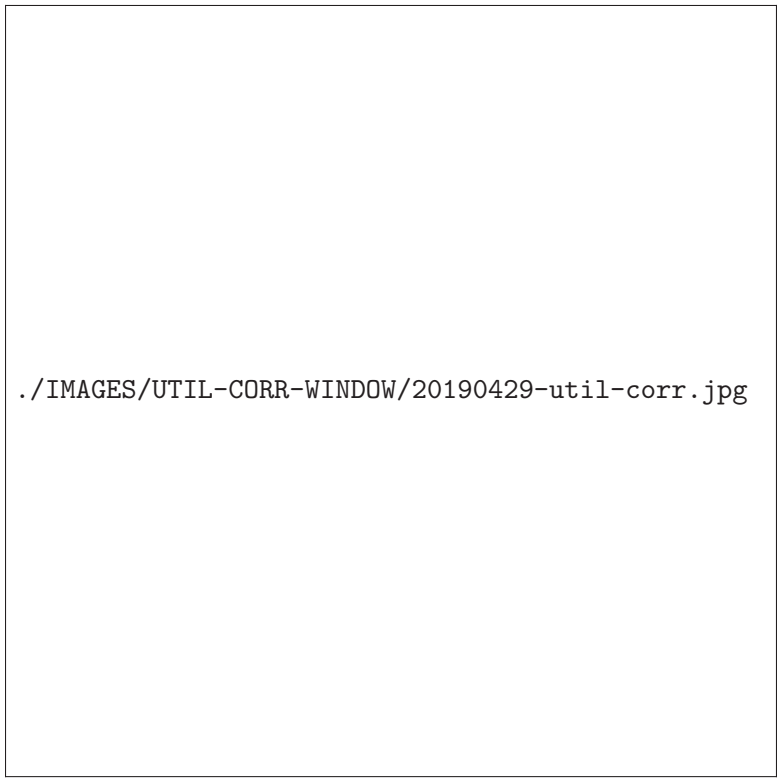

Figure 1.11: The Correlation Analysis window. This windows allows to perform a correlation analysis of the data contained in a given data file.

The list box to the right of Region 3 contains the columns for which the correlation analysis will be performed. Correlation between all columns in the right list box will be performed. The order of the rows and columns in the resulting matrix containing the correlation coefficients will be the same as the order of the columns shown in the right list box. Therefore, users are advised to fill the right list box in such a way that replicates of the same experiment are consecutive to each other in the right list box. Columns in the right list box can be deleted by selecting the columns and then using the right mouse button over the right list box. Columns in the right list box will be unique, meaning that a column can only be added once.

Region 4 contains the Help and Start buttons and a progress bar. The Help button leads to an online tutorial while the Start button will start the analysis. The progress bar will give a rough estimate of the remaining time for completing the analysis.

The analysis

First, UMSAP will check the validity of the user provided input. Then, columns in the

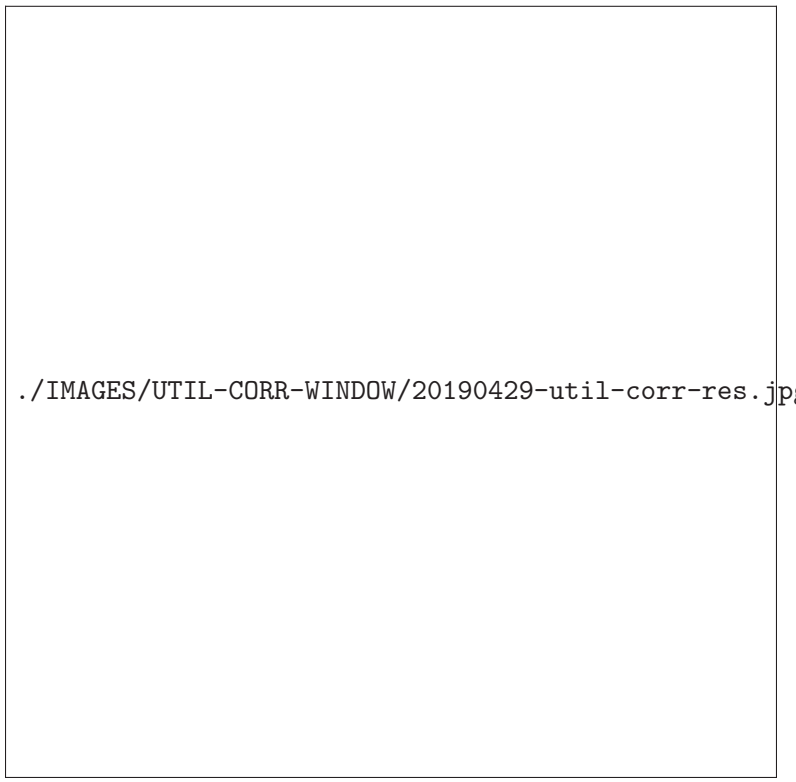

Figure 1.12: The Correlation Analysis result window. The correlation coefficients are shown in a color coded matrix. Values between −1 to 0 are shown in shades of red, 0 is shown in white and values between 0 to 1 in shades of blue. NA values are shown in green.

right list box 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, a Pearson correlation (**Pearson1895**) analysis will be performed for each possible pair of columns according to Equation 1.1. If the use of Equation 1.1 leads to a division by zero, then the corresponding coefficient is set to NA. After the analysis is done the results will be automatically loaded and displayed in a new window (Figure 1.12).

$$c_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}} \tag{1.1}$$

The output

The extension .corr is reserved for a file containing the output from a correlation analysis. The results in a .corr file will be shown as a color coded matrix (Figure 1.12). 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 numbers used to calculate the correlation. Information about a specific matrix element can be obtain by simply putting the mouse

pointer over the matrix element.

The Tools menu

The Tools menu in the configuration window of the correlation analysis (Figure 1.11) allows users to empty the right list box and also to export the path to the selected data file and the number of the columns in the right list box to the Targeted Proteolysis module. If there are no selected entries in the right list box, all of them will be exported. If some entries in the right list box are selected, then only the selected entries will be exported. The exported data is used to fill the Data file and Results entries of the Targeted Proteolysis module. In the case of the Results entry, users have to add the coma (,) separating the replicates from different experiments in the appropriate positions.

The Tools menu in the window showing the results in a .corr file (Figure 1.12) offers the same export functionality plus the option to create an image of the plot.

1.3.2 Merge aadist Files

The .aadist files contain an AA distribution analysis as described in subsection 1.2.1. The Merge aadist Files utility allows to merge several .aadist files into a single file.

The interface

The Merge aadist Files window is divided in four regions (Figure 1.13).

Figure 1.13: The Merge .aadist Files window. This window allows users to merge several .aadist files in a single file.

Region 1 contains two buttons allowing users to quickly delete all provided input and start a new calculation. The Clear all button will delete all user provided input and will reset the state of the list box in Region 3. The Clear files button will reset the state of the list box in Region 3 and delete all user provided paths to files.

Region 2 contains the fields where users provide the information needed in order to merge the .aadist files. The aadist files button allows users to browse the file system to select multiple .aadist file from a folder. Only .aadist files can be selected here. Once the files are selected the complete path to the files will be displayed in the list box in Region 3. The Output file button allows users to browse the file system to select the location and name of the output file.

Region 3 contains a list box showing all selected .aadist files. Files can only be added one time to the list box. The order of the files in the list box is meaningless. Selected files can be deleted from the list box by pressing the right mouse button over the list box.

Region 4 contains the Help and Start buttons and a progress bar. The Help button leads to an online tutorial while the Start button will start the analysis. The progress bar will give a rough estimate of the remaining time for completing the analysis.

The analysis

First, UMSAP will check the validity of the user provided input. Then, the number of positions and experiments in each .aadist files are checked. If they do not match, the merging is aborted. Finally, UMSAP check that all AA distributions were originated from the same sequence and abort the task at hand if they do not. After this, files are merged. For merging the files, UMSAP adds the number of times each amino acids appears in a given position for each file. When all files have been merged the results will be automatically loaded and displayed in a new window (Figure 1.4). The significance level for the merged file is the highest value found in all files that were merged.

The output

The Merge aadist Files utility generates a .aadist file. This file can be visualized in the same way as the output from the AA distribution utility, see Figure 1.4 for more details.

1.3.3 Read Output File

The Read Output File utility simply loads an output file generated by UMSAP. After selecting this option from the Utility window (Figure 1.1) or the Utilities menu entry, a dialog box will be presented allowing users to select some of the output files generated by UMSAP. Currently, only .aadist, .corr, .cutprop, .tarprot and .hist files can be selected. After selecting the file, the appropriate window showing the graphical representation of the file will be created.