Proteomics and Bioinformatics group
Departement of Medical Protein Research
VIB and Faculty of Medicine and Health Sciences, Ghent University
http://www.compomics.be

Rover manual

Niklaas Colaert

http://code.google.com/p/compomics-rover/

Contents

1 General Information **1**

1.1 Downloading and running Rover 1

1.2 Java 1

2 Starting Rover with the Rover wizard **3**

2.1 Different types of analyses 3

2.2 Starting the Rover wizard 3

2.3 Different steps in the Rover wizard 4

2.3.1 Step 1 - Selection of the information source 4

2.3.2 Step 2 - Selection of the data 5

2.3.2.1 From files 5

2.3.2.2 From ms_lims 5

2.3.3 Step 3 - Input parameters 6

2.3.3.1 Peptide identification confidence level 6

2.3.3.2 Reference set 6

2.3.3.3 Database type 6

2.3.3.4 Instrument standard deviation 7

2.3.4 Step 4 7

2.3.5 Step 5 - Data retrieval (only multiple sources approach) 8

2.3.6 Step 6 - Grouping ratios (only multiple sources approach) 8

2.3.7 Step 7 - Grouping components (only multiple sources approach) 9

2.3.8 Step 8 - Data merging 9

3 Data viewing with Rover **10**

3.1 Overview 10

3.2 Protein list panel 10

3.3 Protein bar 12

3.4 Filter 12

3.5 Ratio distribution graph 15
3.6 Protein panel 15
3.7 More info panel 17
3.8 Button panels 18
3.8.1 Top button panel 18
3.8.2 Lower button panel 18
3.9 Export panel 20

4 Statistics 22

5 Working with Rover, an example 24

6 Problems and questions 26

Chapter 1

General Information

1.1 Downloading and running Rover

The main Rover site is `http://code.google.com/p/compomics-rover/`. There, under the download section, a zip file containing the program can be downloaded. The rover program can be started by double clicking the rover-2.0.jar file after unzipping the folder.

The maximum number of *.rov* files that can be loaded at a time is limited. The size of all the *.rov* files cannot exceed the size of the virtual memory committed to Rover. Normally, the rest of the filetypes that can be loaded by Rover have smaller file sizes and the number of such files that can be loaded is thus not limited.

The virtual memory that is committed to the Rover program can be changed in the rover.properties file. This file is located in the user/.compomics/rover folder. This is not the location where the program is located. The default value is 1024 Mb. This can be changed by editing the number after *-Xmx* in the file.

1.2 Java

It is probable that a functional Java is already installed on your computer due to the widespread use of Java nowadays. If you do not know for sure you have Java 1.5 or higher installed, you can do the following simple check:

- Open a command window by `start` → `run` and enter `cmd`.
- Enter `Java -version` in the command window.

If Java is already installed, you will see something like below where x stands for the version.

```
{Java version "1.x.0_01"}
```

If your Java version is lower then 1.5 you have to upgrade Java. If you don't have Java installed, you have to make a new install. In both cases, you have to install a new Java version. The installation of Java is quite simple.

- Go to `http://java.com`
- follow the main download link and start the download
- when finished, open the installer and follow the straightforward instructions

Chapter 2

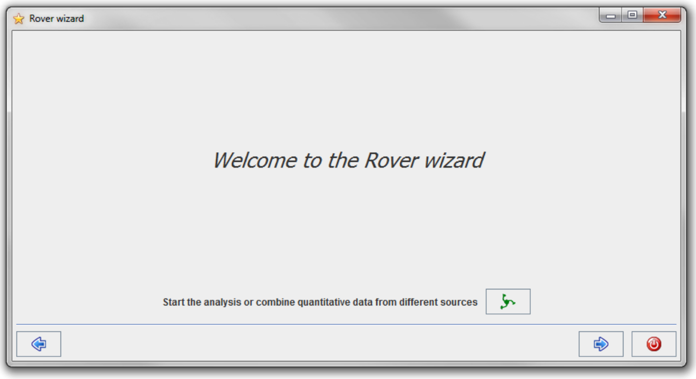
Starting Rover with the Rover wizard

2.1 Different types of analyses

Two different types of analysis can be performed in rover. The simplest one is the analysis of quantitative data. A second methods is also available where different source of quantitative data can be combined via location and scale normalization.

2.2 Starting the Rover wizard

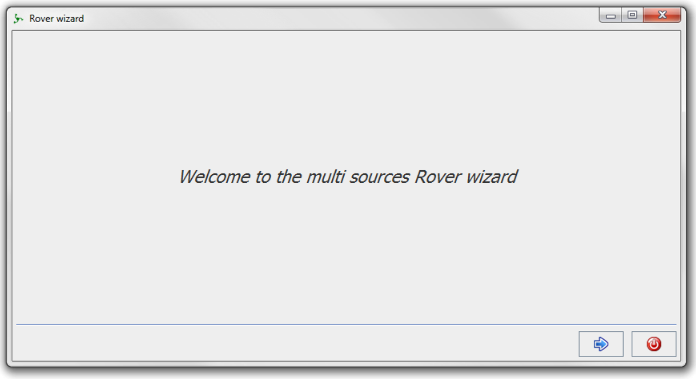
The Rover wizard can be started by double clicking the rover.jar file in the installation directory of the Rover program. When the Rover wizard is opened correctly, a welcome screen (see figure 2.1) appears. When the multiple sources approach button is clicked a different welcome screen (see figure 2.2) appears.



Welcome to the Rover wizard

Start the analysis or combine quantitative data from different sources

Figure 2.1: The Rover wizard welcome screen



Welcome to the multi sources Rover wizard

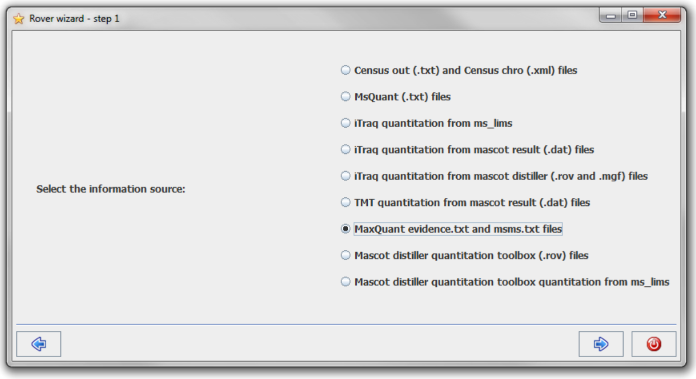
Figure 2.2: The Rover wizard welcome screen for the multiple sources approach

2.3 Different steps in the Rover wizard

The Rover wizard is a step by step tool for selecting quantitative data and starting the Rover program. This is done by clicking the next ⇨ and previous ⇦ button at the lower part of the frame.

2.3.1 Step 1 - Selection of the information source

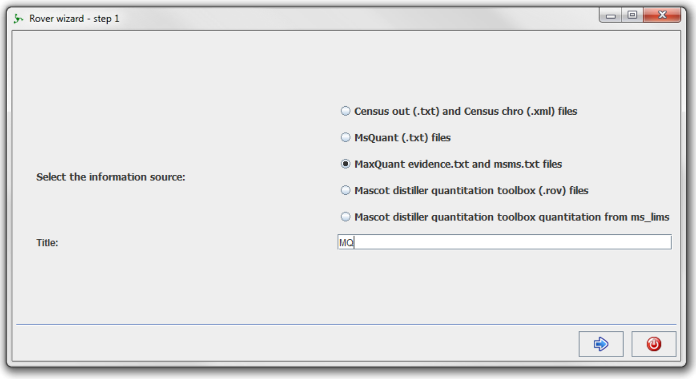
Rover can load different types of quantitative information from different sources (see figure 2.3 and 2.4 for the multiple sources approach).



Select the information source:

- ○ Census out (.txt) and Census chro (.xml) files
- ○ MsQuant (.txt) files
- ○ iTraq quantitation from ms_lims
- ○ iTraq quantitation from mascot result (.dat) files
- ○ iTraq quantitation from mascot distiller (.rov and .mgf) files
- ○ TMT quantitation from mascot result (.dat) files
- ● MaxQuant evidence.txt and msms.txt files
- ○ Mascot distiller quantitation toolbox (.rov) files
- ○ Mascot distiller quantitation toolbox quantitation from ms_lims

Figure 2.3: Step 1 of the Rover wizard



Select the information source:

○ Census out (.txt) and Census chro (.xml) files

○ MsQuant (.txt) files

● MaxQuant evidence.txt and msms.txt files

○ Mascot distiller quantitation toolbox (.rov) files

○ Mascot distiller quantitation toolbox quantitation from ms_lims

Title: MQ

Figure 2.4: Step 1 of the multiple sources Rover wizard. A title has to be given to every set.

Different types of quantitative data can be analysed and viewed by Rover.

iTRAQ™(isobaric tags for relative and absolute quantitation) data: The peptides N-terminus is covalently modified by isobaric tags. The quantitative information can be found in the MS/MS spectra. iTRAQ™data can be read from *.dat* Mascot result files and from Mascot Distiller *.rov* files. If the information comes from Mascot Distiller, a *.mgf* file must also be given to Rover[1].

(Post) metabolic labelled peptides analysed by the Mascot Distiller Quantitation toolbox: Two or more groups of proteins are metabolically labelled (ex. SILAC labelling, $^{16}0/^{18}0$ labelling, . . .). Quantitative information can be calculated by comparison of the intensities of the precursors in the MS spectra. This comparison is done by the Mascot Distiller Quantitation toolbox. The data are stored in Mascot Distiller *.rov* files. Keep in mind that these *.rov* files must be created with the Quantitation toolbox extension of Mascot Distiller and that the quantitative information and peptide identifications must be stored in the *.rov* file.

SILAC labelling analysed by MaxQuant: Double and triple SILAC labelled experiments can be analysed by MaxQuant. Two of the MaxQuant result files (msms.txt and evidence.txt) are used by Rover to extract the quantitative information.

MSQuant result files: MSQuant is a tool for quantitative proteomics/mass spectrometry and processes spectra and LC runs to find quantitative information about proteins and peptides. The quantitative export file (*.txt*) from MSQuant is needed to view MSQuant data in Rover.

[1] The name of the *.rov* file must start with the name of the *.mgf* file. This *.mgf* file is normally located in the same folder as the *.rov* file.

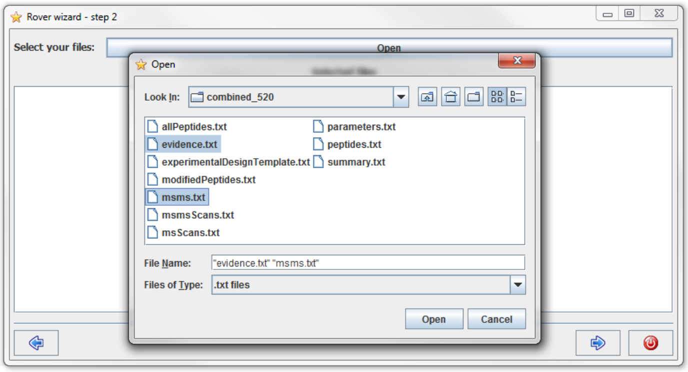
Census result files: Census uses mzXML and DTASelect data for the quantitation of peptides. The output (*.txt*) and chro (*.xml*) files are used as import by Rover.

Data can be loaded from such files or from ms_lims. Ms_lims is a mass-spectrometry focused lims database system (see `http://genesis.ugent.be/ms_lims` for more information). iTRAQ™data as well as Mascot Distiller data can be loaded from ms_lims.

2.3.2 Step 2 - Selection of the data

2.3.2.1 From files

In the second step the files with quantitative information must be selected . Click the open button (see figure 2.5) and a file chooser frame appears. Selected the files and click the open button in the file chooser. The selected files will now appear in the selected files list in the wizard frame.



Look In: combined_520

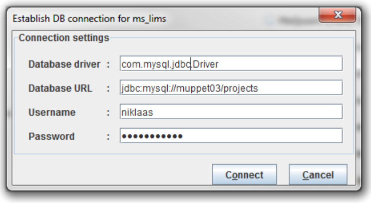
File Name: "evidence.txt" "msms.txt"

Files of Type: .txt files

Figure 2.5: Step 2 of the Rover wizard

2.3.2.2 From ms_lims

If a connection to ms_lims doesn't exist, a frame will appear (see figure 2.6) where a connection to a ms_lims database can be established.



Database driver : com.mysql.jdbc.Driver

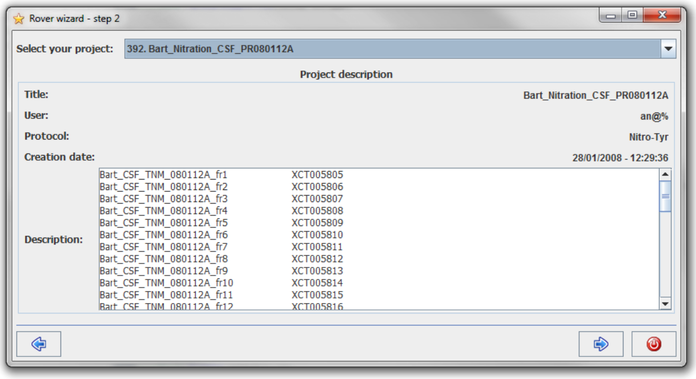
Database URL : jdbc:mysql://muppet03/projects

Username : niklaas

Password : ••••••••••••

Figure 2.6: Frame to make connection to ms_lims

If a valid connection is established to ms_lims, all the project will be loaded in a drop-down menu (see figure 2.7). If a project is selected, the user, project title, project description, ... will be loaded. The selected project will be used as data source in the following steps.



Select your project: 392. Bart_Nitration_CSF_PR080112A

Project description

Title: Bart_Nitration_CSF_PR080112A

User: an@%

Protocol: Nitro-Tyr

Creation date: 28/01/2008 - 12:29:36

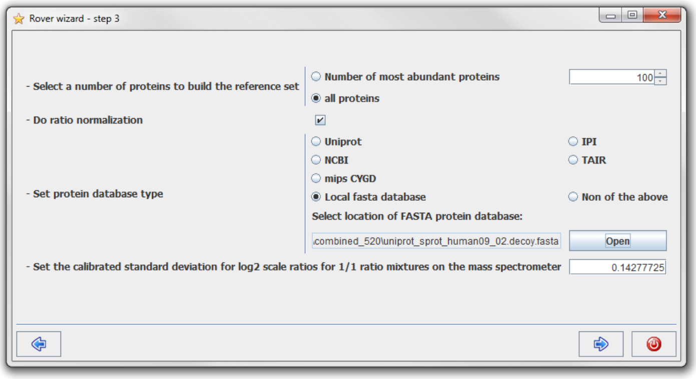
Description:

Bart_CSF_TNM_080112A_fr1	XCT005805
Bart_CSF_TNM_080112A_fr2	XCT005806
Bart_CSF_TNM_080112A_fr3	XCT005807
Bart_CSF_TNM_080112A_fr4	XCT005808
Bart_CSF_TNM_080112A_fr5	XCT005809
Bart_CSF_TNM_080112A_fr6	XCT005810
Bart_CSF_TNM_080112A_fr7	XCT005811
Bart_CSF_TNM_080112A_fr8	XCT005812
Bart_CSF_TNM_080112A_fr9	XCT005813
Bart_CSF_TNM_080112A_fr10	XCT005814
Bart_CSF_TNM_080112A_fr11	XCT005815
Bart_CSF_TNM_080112A_fr12	XCT005816

Figure 2.7: Step 2 of the Rover wizard

2.3.3 Step 3 - Input parameters

Several parameters need to be set for Rover in step 3 (see figure 2.8).



- Select a number of proteins to build the reference set ○ Number of most abundant proteins 100 ● all proteins

- Do ratio normalization ☑

- Set protein database type ○ Uniprot ○ IPI ○ NCBI ○ TAIR ○ mips CYGD ● Local fasta database ○ Non of the above

Select location of FASTA protein database: combined_520\uniprot_sprot_human09_02.decoy.fasta

- Set the calibrated standard deviation for log2 scale ratios for 1/1 ratio mixtures on the mass spectrometer 0.14277725

Figure 2.8: Step 3 of the Rover wizard

2.3.3.1 Peptide identification confidence level

The Mascot peptide identification confidence level (default is 0.99) only needs to be set if data are loaded from files and not from ms_lims.

2.3.3.2 Reference set

Rover will create a reference set of proteins. This set will be used to compare protein ratio means. For more information on this reference set, see chapter 4.

Two different option exist for creating the reference set. The first option to build the reference set is to use all ratios linked to all proteins. The second option is to use the ratios linked to a specific number (20-300) of the most abundant (based on the number of identified spectra) proteins.

Rover can be set to take all ratios (valid and invalid) from the selected proteins in the reference set or it can be set to take only the valid ratios. By default rover will take only the valid ratios.

2.3.3.3 Database type

Mascot will use a *.fasta* or *.dat* (protein) database on the Mascot server to identify MS/MS spectra. Since Mascot and other file providers don't store protein sequences in their result files, Rover downloads the protein sequences. There are different types of databases and protein accessions supported. Also a local *.fasta* protein database can be selected. The option "Non of the above" has to be selected if a database was created with user-defined accessions, or if a database was downloaded from an unsupported source. The protein sequences will not be downloaded for this option. As a result, the protein bar (see 3.3) will not be visualised.

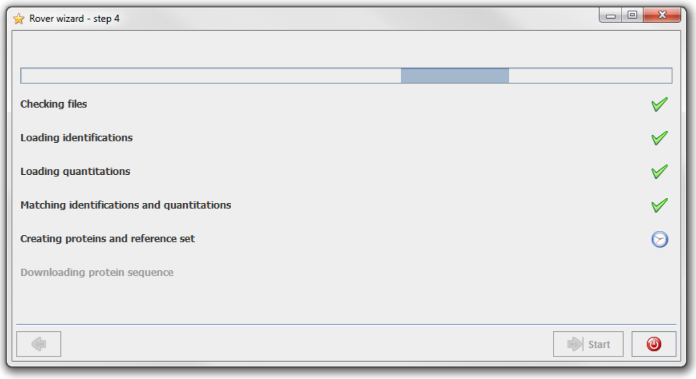
2.3.3.4 Instrument standard deviation

Rover will use statistics to calculate the significance of deviation of a ratio value when compared to the reference set. The program will not only use the reference set to create a standard deviation, it will also use the calibrated standard deviation for $\log_2$ scale ratios for $\frac{1}{1}$ ratio mixtures meaesered on a mass spectrometer. More on statistics can be found in chapter 4.

A default value of 0.14277725 is given. This was calculated for an Waters Q-TOF Premier in our laboratory.

2.3.4 Step 4

Step 4 is the final step for the single source method. The Rover wizard will start collecting the data after the start button ➡ is clicked (see figure 2.9). Different steps are performed before everything is loaded and visualized (see 3). During the loading process a window will appear (see figure 2.15) where the not regulated component and the expected ratio median must be set for every ratio.



Rover wizard - step 4

Checking files ✓

Loading identifications ✓

Loading quantitations ✓

Matching identifications and quantitations ✓

Creating proteins and reference set

Downloading protein sequence

Figure 2.9: Step 4 of the Rover wizard

- If the loaded data come from files, these files will be checked if they exist and if they have the correct quantitative information
- The peptide identifications will be loaded for every file
- The quantitative information will be loaded for every file
- The peptide identifications will be linked to the quantitative information for every file
- Different peptides are grouped to proteins and a reference set is created.
- The protein sequences will be downloaded if possible.

If the multiple sources approach is used a second source (see figure 2.10) can be added and this takes the wizard back to step 1 (data type selection step).

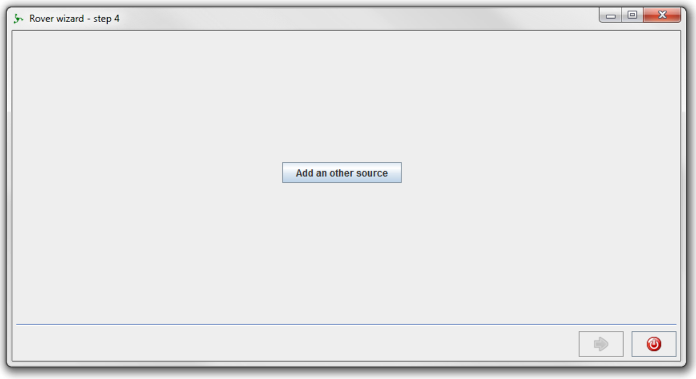
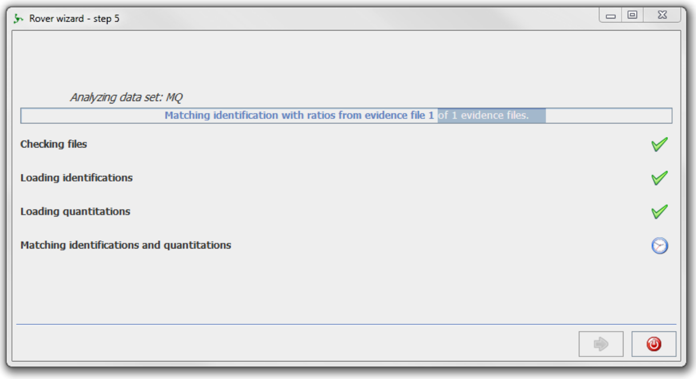


Figure 2.10: Step 4 of the multiple sources Rover wizard

2.3.5 Step 5 - Data retrieval (only multiple sources approach)

The Rover wizard will collect the data from the different sources (see figure 2.11).



Rover wizard - step 5

Analyzing data set: MQ

Matching identification with ratios from evidence file 1 of 1 evidence files.

Checking files ✓

Loading identifications ✓

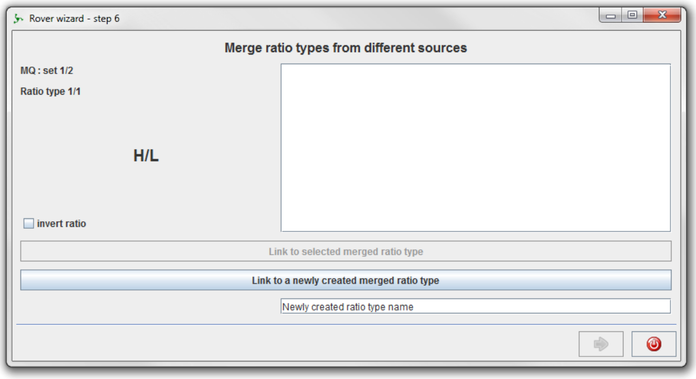
Loading quantitations ✓

Matching identifications and quantitations

Figure 2.11: Step 5 of the Rover wizard

2.3.6 Step 6 - Grouping ratios (only multiple sources approach)

The different ratios from the different sources must be linked to each other. This step is necessary because the correct ratios from different source must be merged. The common ratio name must be set by filling in the text box (see figure 2.12). The ratios from the other sources can now be linked to this common ratio by selecting it in the list (see figure 2.13). Ratios can in this step be selected for inversion. This could be necessary if L/H ratios must be merged with H/L ratios.



Rover wizard - step 6

Merge ratio types from different sources

MQ : set 1/2

Ratio type 1/1

H/L

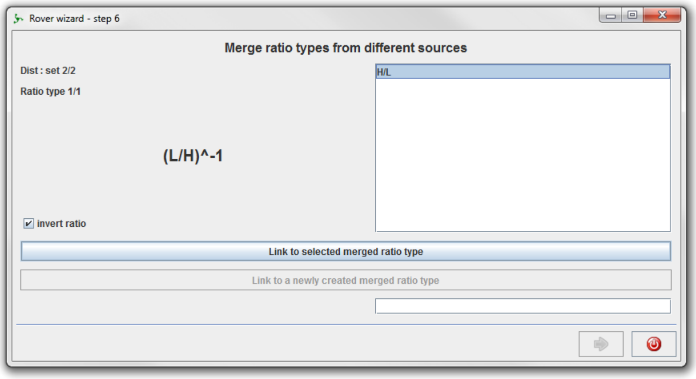
□ invert ratio

Link to selected merged ratio type

Link to a newly created merged ratio type

Newly created ratio type name

Figure 2.12: Step 6.1 of the Rover wizard



Rover wizard - step 6
Merge ratio types from different sources
Link to selected merged ratio type
Link to a newly created merged ratio type

Dist : set 2/2

Ratio type 1/1

(L/H)^-1

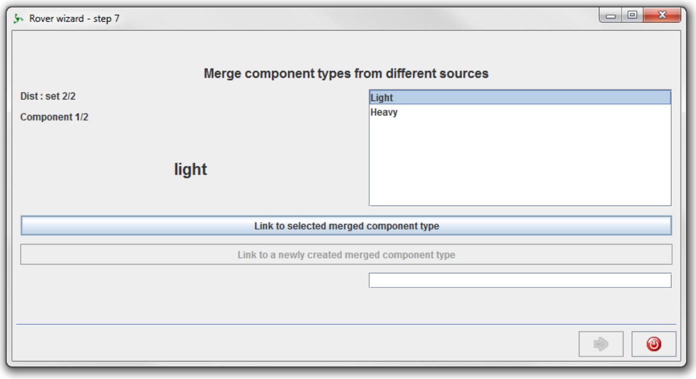
☑ invert ratio

H/L

Figure 2.13: Step 6.2 of the Rover wizard

2.3.7 Step 7 - Grouping components (only multiple sources approach)

Just like in step 6, the different components from the different sources must be merged (see figure 2.14). When the different components are linked a window will appear (see figure 2.15) where the the not regulated component and the expected ratio median must be set for every ratio from every source.



Rover wizard - step 7
Merge component types from different sources
Link to selected merged component type
Link to a newly created merged component type

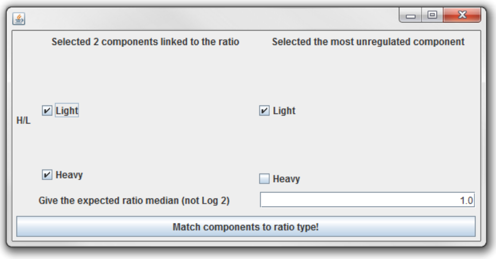
Dist : set 2/2

Component 1/2

light

Light
Heavy

Figure 2.14: Step 7.1 of the Rover wizard



Selected 2 components linked to the ratio
Selected the most unregulated component
Match components to ratio type!

H/L ☑ Light ☑ Light

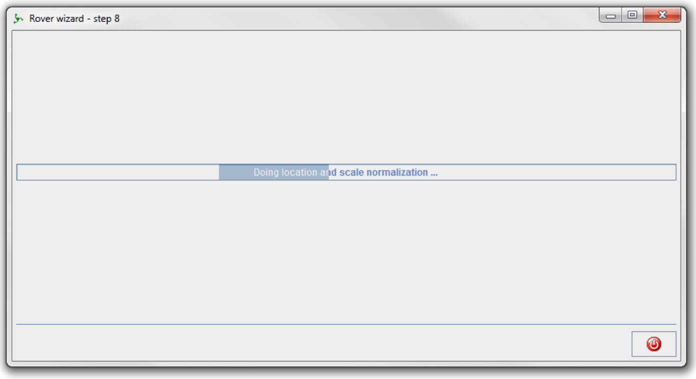
☑ Heavy □ Heavy

Give the expected ratio median (not Log 2) 1.0

Figure 2.15: Step 7.2 of the Rover wizard

2.3.8 Step 8 - Data merging

The data from the different sources will be merged. This can only be done after a location and scale normalization process.



Doing location and scale normalization ...

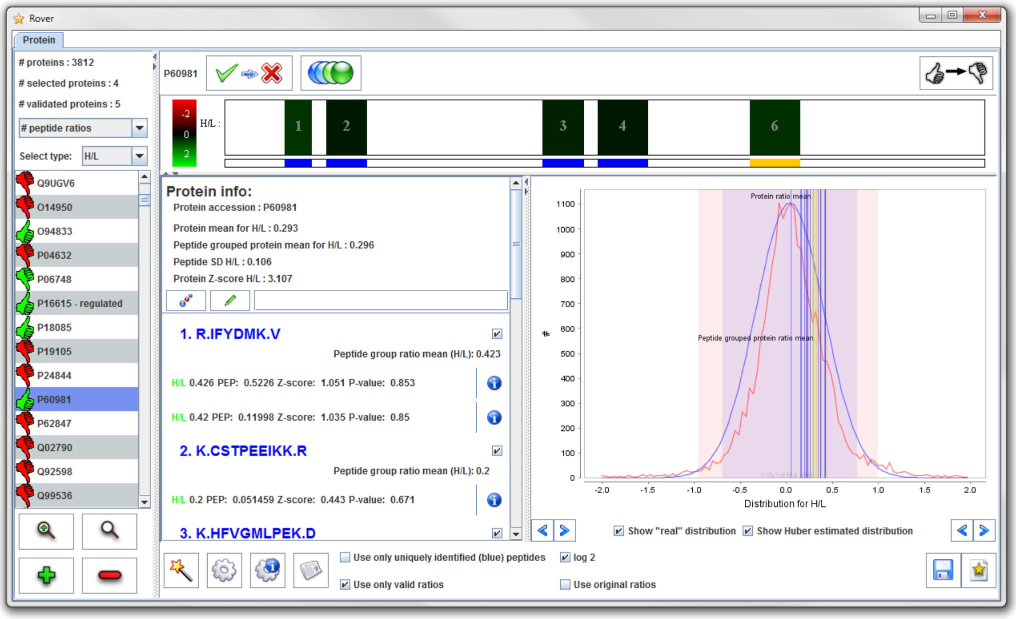
Figure 2.16: Step 8 of the Rover wizard

Chapter 3

Data viewing with Rover

3.1 Overview

After all the different steps in the Rover wizard are executed the main Rover frame will be shown. This frame (figure 3.1) consists of different panels, which will be explained in the following sections. These visualize the data in different ways and from different angles, making it easier for the user to validate a (regulated) protein. To tackle the problem of peptides pointing to multiple proteins in a data search space (protein inference), Rover uses color labels for individual peptides: peptides unique for a protein accession are labelled blue, peptides belonging to multiple protein accessions are labelled orange and Occam's razor peptides are labelled red in every panel. In this way, Rover clearly indicates which peptides are shared between different protein accessions and are thus prone to influence protein ratios in a unwanted way.



Rover

Protein

proteins : 3812

selected proteins : 4

validated proteins : 5

peptide ratios

Select type: H/L

Q9UGV6
O14950
O94833
P04632
P06748
P16615 - regulated
P18085
P19105
P24844
P60981
P62847
Q02790
Q92598
Q99536

P60981

Protein info:
Protein accession : P60981

Protein mean for H/L : 0.293

Peptide grouped protein mean for H/L : 0.296

Peptide SD H/L : 0.106

Protein Z-score H/L : 3.107

1. R.IFYDMK.V ☑

Peptide group ratio mean (H/L): 0.423

H/L 0.426 PEP: 0.5226 Z-score: 1.051 P-value: 0.853

H/L 0.42 PEP: 0.11998 Z-score: 1.035 P-value: 0.85

2. K.CSTPEEIKK.R ☑

Peptide group ratio mean (H/L): 0.2

H/L 0.2 PEP: 0.051459 Z-score: 0.443 P-value: 0.671

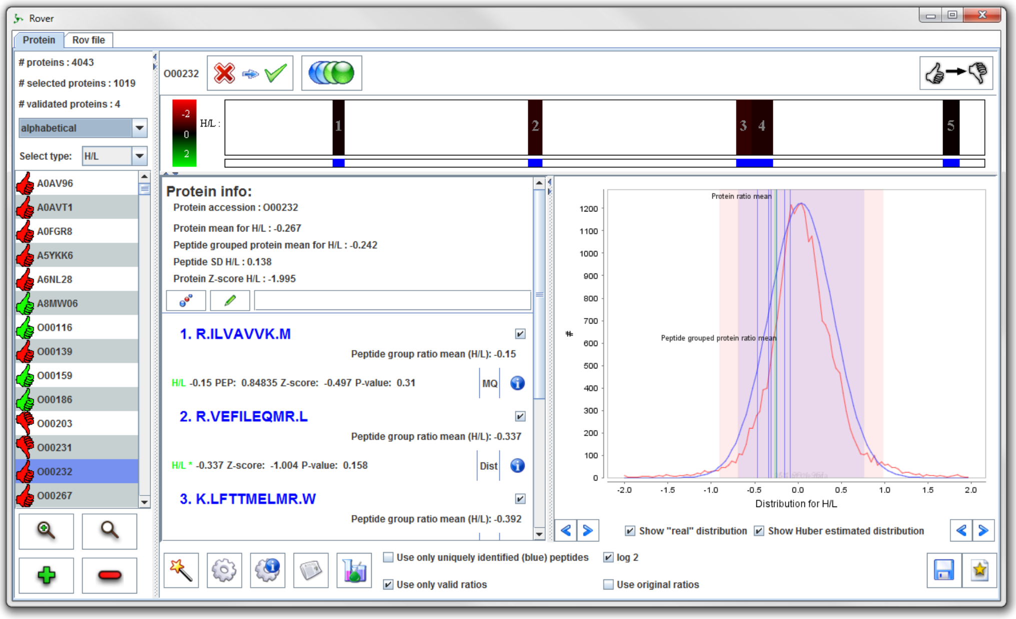
3. K.HFVGMLPEK.D ☑

☑ Show "real" distribution ☑ Show Huber estimated distribution

☐ Use only uniquely identified (blue) peptides ☑ log 2

☑ Use only valid ratios ☐ Use original ratios

Figure 3.1: The main Rover frame



Rover

Protein Rov file

proteins : 4043

selected proteins : 1019

validated proteins : 4

alphabetical

Select type: H/L

A0AV96
A0AVT1
A0FGR8
A5YKK6
A6NL28
A8MW06
O00116
O00139
O00159
O00186
O00203
O00231
O00232
O00267

O00232

Protein info:
Protein accession : O00232

Protein mean for H/L : -0.267

Peptide grouped protein mean for H/L : -0.242

Peptide SD H/L : 0.138

Protein Z-score H/L : -1.995

1. R.ILVAVVK.M ☑

Peptide group ratio mean (H/L): -0.15

H/L -0.15 PEP: 0.84835 Z-score: -0.497 P-value: 0.31 MQ

2. R.VEFILEQMR.L ☑

Peptide group ratio mean (H/L): -0.337

H/L * -0.337 Z-score: -1.004 P-value: 0.158 Dist

3. K.LFTTMELMR.W ☑

Peptide group ratio mean (H/L): -0.392

☑ Show "real" distribution ☑ Show Huber estimated distribution

☐ Use only uniquely identified (blue) peptides ☑ log 2

☑ Use only valid ratios ☐ Use original ratios

Figure 3.2: The main multiple sources Rover frame

3.2 Protein list panel

A list with all the proteins (identified by their protein accessions) that were found in the input data is displayed on the left side of the main Rover frame (see figure 3.1). When a filter (see section 3.4) is applied, only the filtered proteins will be displayed. Four different icons display the status of each protein. This protein status indicates if a user selected an interesting protein (selected status) and if a user indicated that a protein has been validated (validated status).

- 👎 The protein is *not selected* and *not validated*.
- 👍 The protein is *selected* and *not validated*.
- 👎 The protein is *not selected* and *validated*.
- 👍 The protein is *selected* and *validated*.

At the top of the protein list panel three numbers indicate the total number of proteins, the number of selected proteins and the number of validated proteins. Also the proteins can be sorted in different order by choosing a sorting type and ratio just below these numbers.

At the bottom of the protein list panel are four buttons.

- 🔍 When this button is clicked *all the proteins* will be displayed in the protein list.
- 🔍 When this button is clicked *the selected proteins* will be displayed in the protein list.
- ➕ When this button is clicked all the *proteins in the protein list* will be *selected*.
- ➖ When this button is clicked all the *proteins in the protein list* will be *unselected*.

3.3 Protein bar

The protein bar (see figure 3.3), which is located at the top of the main Rover frame (see figure 3.1), visualizes the protein coverage and the peptide ratios. Rover will only show the peptides and ratios that are used for the calculation of the protein mean. Thus, if Rover is in the "*show only valid peptides*" mode, only the peptides with a valid peptide ratio will be displayed in the protein bar.

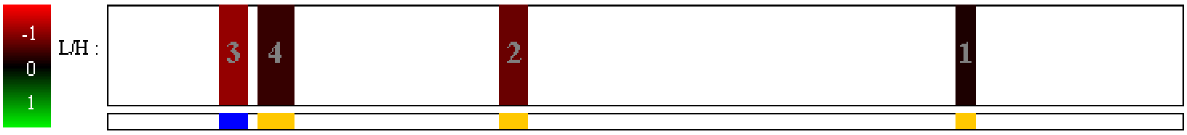


Figure 3.3: The protein bar

A color gradient box on the left side of this panel shows a gradient from red to green between the minimum and maximum ratio values. This minimum and maximum can be changed by changing the width of the ratio distribution graph (see section 3.5).

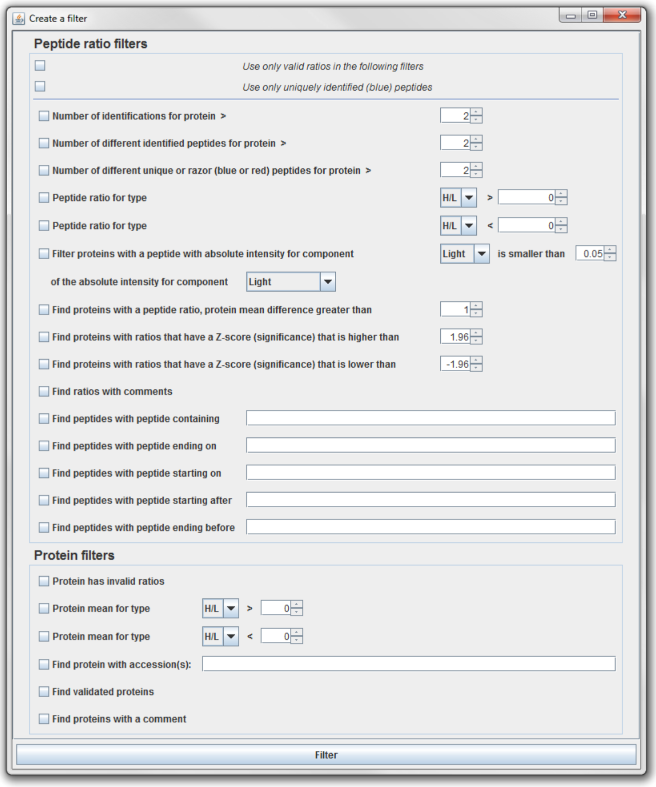
The black rectangle represents the protein. On this rectangle, the peptides with their corresponding positions in the protein are shown. The number in the peptide box corresponds to the peptide group with the same number in the protein panel (see section 3.6). The length of these peptide boxes correlates with the actual peptide length normalised to the protein length. The color of the peptide box reflects the peptide ratio (if the same peptide is identified more than once, the ratio mean will be used). If multiple ratios were calculated (ex. iTRAQ) for the peptides, multiple rectangles will be shown.

At the bottom of this panel there is a small bar with blue, orange and red boxes. The blue boxes show the position of peptides (with valid and invalid ratios) that are uniquely linked to this protein. Orange boxes reflect the position of peptides (with valid and invalid ratios) that can be linked to multiple proteins.

3.4 Filter

Not all the proteins must be validated. With the filter tool, proteins of interest or peptides with a specific ratio can be selected. After a filter step is performed, Rover will show how many proteins it has filtered and it will present these in the protein list (see section 3.2).

Different filters were created to select a specific subset of proteins. There are generally two types of filters. Filters that will filter on the peptide (ratio) level and filters that will filter on the protein level. Filters that work on the peptide level can be set to only use peptides with a valid ratio or uniquely identified peptides.



Create a filter

Peptide ratio filters

□ *Use only valid ratios in the following filters*

□ *Use only uniquely identified (blue) peptides*

□ Number of identifications for protein > 2

□ Number of different identified peptides for protein > 2

□ Number of different unique or razor (blue or red) peptides for protein > 2

□ Peptide ratio for type H/L > 0

□ Peptide ratio for type H/L < 0

□ Filter proteins with a peptide with absolute intensity for component Light is smaller than 0.05

of the absolute intensity for component Light

□ Find proteins with a peptide ratio, protein mean difference greater than 1

□ Find proteins with ratios that have a Z-score (significance) that is higher than 1.96

□ Find proteins with ratios that have a Z-score (significance) that is lower than -1.96

□ Find ratios with comments

□ Find peptides with peptide containing

□ Find peptides with peptide ending on

□ Find peptides with peptide starting on

□ Find peptides with peptide starting after

□ Find peptides with peptide ending before

Protein filters

□ Protein has invalid ratios

□ Protein mean for type H/L > 0

□ Protein mean for type H/L < 0

□ Find protein with accession(s):

□ Find validated proteins

□ Find proteins with a comment

Figure 3.4: The filter panel

Number of identifications for protein > ★

A protein will be filtered if more than ★ peptides are linked to this protein.

Number of different identified peptides for protein > ★

A protein will be filtered if more than ★ different peptides are linked to this protein.

Peptide ratio for type A > ★

A protein will be filtered if one of the peptide ratio means for A is larger than ★.

Peptide ratio for type A < ★

A protein will be filtered if one of the peptide ratio means for A is smaller than ★.

Filter proteins with a peptide with absolute intensity for component A is smaller than ★ of the absolute intensity for component B

A protein will be filtered if the absolute intensity for A is ★ times the absolute intensity of B. This filter only works if the data are from the Mascot Distiller Quantitation toolbox.

Find proteins with a peptide ratio, protein mean difference greater than: ★

A protein will be filtered if a difference between a peptide ratio and protein mean is larger than ★.

Find proteins with ratio that have a Z-score (significance) that is higher than: ★

A protein will be filtered if a significance of a peptide ratio is larger than ★. See chapter 4 for more information on the significance.

Find proteins with ratio that have a Z-score (significance) that is lower than: ★

A protein will be filtered if a significance of a peptide ratio is smaller than ★. See chapter 4 for more information on the significance.

Find ratios with comments

A protein will be filtered if one of the peptide ratios that is linked to the protein has a comment made by a Rover user.

Find peptides with peptide containing ★

A protein will be filtered if one of the peptides contains the ★ sequence.

Find peptides with peptide ending on ★

A protein will be filtered if one of the peptides ends on ★.

Find peptides with peptide starting on ★

A protein will be filtered if one of the peptides starts with ★.

Find peptides with peptide starting after ★

A protein will be filtered if one of the peptides starts after ★.

Find peptides with peptide ending before ★

A protein will be filtered if one of the peptides ends before ★.

Find peptides with N-terminal modification ★

A protein will be filtered if one the N-terminal modifications equals ★. (This filter can only be used when the data comes from Mascot Distiller)

Protein has invalid ratios

A protein will be filtered if one of the peptide ratios that is linked to the protein is invalid.

Protein mean for type A > ★

A protein will be filtered if the protein mean for A is larger than ★.

Protein mean for type ratiotype < ★

A protein will be filtered if the protein mean for ratiotype (ex. L/H) is smaller than ★.

Find protein with accession: ★

A protein will be filtered if the protein accession equals ★. Multiple accessions separated by commas are also valid as an input parameter.

Find validated proteins

A protein will be filtered if it is validated.

Find proteins with a comment

A protein will be filtered whenever it has a protein comment.

3.5 Ratio distribution graph

In the ratio distribution graph both the peptide ratios of the selected protein and the ratios in the reference set (see section 2.3.3.2) are shown. The distribution in the reference set can be visualized in two different ways.

- The first way is the “real” distribution. The ratios of the reference set are used to create a histogram. The red background spans 95% ([2.5 %; 97.5 %]) of all data.
- The second way is the “Huber” estimated distribution. The ratios of the reference set are used to calculate an average and a standard deviation (σ). These are used to create a distribution (see chapter 4 for more information). The blue background is the 95% confidence interval [-1.96 σ; 1.96 σ].

The blue, orange and red vertical lines in the graph are the peptide ratios linked to the selected protein. These are blue when the peptide is only linked to this protein, orange when the peptide is linked to multiple proteins and red when the peptide is an Occam’s razor peptide. Two green lines show the “*Protein ratio mean*” and the “*Peptide grouped protein ratio mean*”.

At the bottom of the ratio distribution graph are check buttons to choose which of the distribution presentations are visualized. There are also buttons to change the width of the graph.

3.6 Protein panel

The protein panel is located at the left side of the ratio distribution graph in the main Rover frame (see figure 3.1). The "*protein mean*" and the"*peptide grouped protein mean*" are at the top of the protein info panel. The "*peptide grouped protein mean*" is the mean of all the peptide groups ratio means (the mean ratio of all the peptides with the same identified sequence). The "*protein mean*" is the mean of all peptide ratios. Sometimes some peptide ratios will be excluded in the calculation of the protein mean. This can happen when only valid or unique peptides are used in the calculation. This usage can be set by two check boxes on the lower button panel (see 3.8.2). The "*Peptide SD*" is the standard deviation calculated for all the ratios used in the protein ratio calculation. The "*Protein Z-score*" is the Z-score calculated for the protein ratio based on the reference set median and standard deviation (σ) and the number of the ratios (n) used in the protein ratio calculation (see formula 3.1).

$$Z - score = \frac{Protein\ ratio - Reference\ set\ median}{\frac{Reference\ set\ \sigma}{\sqrt{n}}} \tag{3.1}$$

A comment can be given to a protein by typing it in the comment text field and by clicking the comment button ✏. This comment will append the name (= accession) of the protein. This name can be seen in the protein list (see 3.2) and in the top button panel (see 3.8.1).

MS/MS spectra that identify the same peptide are grouped under this peptide sequence in peptide groups. If the peptide sequence can be matched on the protein sequence the amino acid before and after the peptide will also be given before and after the identified peptide sequence (seperated by a "."). The colour of the peptide sequence is orange if this peptide is linked to multiple proteins, blue when the peptide is only be linked to the selected protein and red when the peptide is an Occam's razor peptide. When the mouse is hovered over an orange and red peptides a small box with the accessions of the proteins to which it is linked, appears.

Whenever the peptide sequence from one peptide group is clicked it will show or hide (depending on the visualization status) all the ratio groups that are linked to this peptide group. All the ratio groups can be hidden by clicking ⁘ or can be revealed again by clicking ⁚.

Every peptide identification is linked to a ratio group. A ratio group has more than one ratio, if more than one ratio type is calculated (ex. L/H, M/H and L/M). At the right side of this ratio group there is a "*more information*" ⓘ button. When this button is clicked, the more info panel is shown (see 3.7).

If the ratio is invalid, the ratio value is in red and the ratio quality is shown (only with Mascot Distiller Quantitation toolbox data). If the Mascot Distiller Quantitation toolbox has reasons to set the ratio invalid it will be shown there. The ratio is green if it's a valid ratio. The Z-score is the number of standard deviations that corresponds with the difference between the ratio and the reference set ratio mean (see section 2.3.3.2 and chapter 4) The P-value is calculated with this Z-score. The ratio comment (if any) will be shown below the ratio in italics.

If Rover is run in the multiple sources mode, the title of the source will be indicated left to the "*more information*" ⓘ button. If a ratio is inverted in the combination process the name will of the ratio will be appended by * (see figure 3.2 for an example).

If one of the peptides groups is not to be used to calculate a protein mean it can be deselected by clicking the check box next to the peptide sequence. This will also hide the ratios from the distribution graph (see 3.5) and the peptide box from the protein bar (see 3.3).

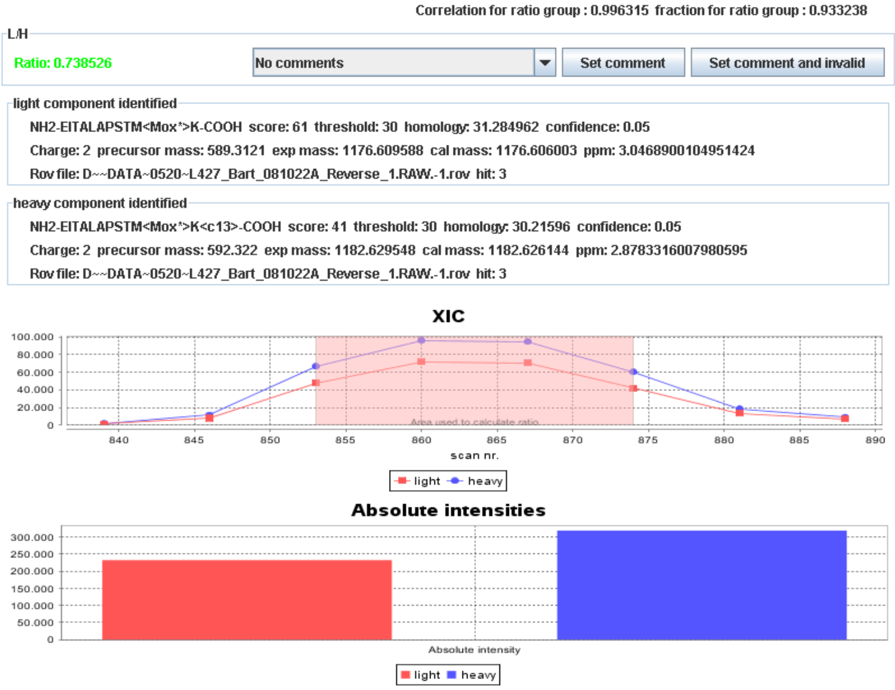
3.7 More info panel

In this panel (see figure 3.5) the ratio can be commented and set (in)valid. If the data are loaded from ms_lims, this comment and valid change will be directly stored in ms_lims. Otherwise, this can be stored in a *.ROVER* file (see section 3.9).

More information concerning the peptide identification (score, mass, modified peptide, . . .) is shown on this panel.

If Mascot Distiller Quantititation toolbox data is loaded, the correlation and fraction for this peptide group is given on the top of the panel. An XIC graph and a bar chart with absolute intensities is also shown. The pink area on the XIC graph shows the scans that the Quantitation toolbox used to calculate the ratio.

If MaxQuant data is loaded a bar chart with absolute intensities is visualized.



Correlation for ratio group : 0.996315 fraction for ratio group : 0.933238

L/H

Ratio: 0.738526

No comments

light component identified

NH2-EITALAPSTM<Mox*>K-COOH score: 61 threshold: 30 homology: 31.284962 confidence: 0.05

Charge: 2 precursor mass: 589.3121 exp mass: 1176.609588 cal mass: 1176.606003 ppm: 3.0468900104951424

Rov file: D~~DATA~0520~L427_Bart_081022A_Reverse_1.RAW.-1.rov hit: 3

heavy component identified

NH2-EITALAPSTM<Mox*>K<c13>-COOH score: 41 threshold: 30 homology: 30.21596 confidence: 0.05

Charge: 2 precursor mass: 592.322 exp mass: 1182.629548 cal mass: 1182.626144 ppm: 2.8783316007980595

Rov file: D~~DATA~0520~L427_Bart_081022A_Reverse_1.RAW.-1.rov hit: 3

Figure 3.5: This is an example of a “more info” panel that displays data loaded from Mascot Distiller Quantitation toolbox files.

3.8 Button panels

3.8.1 Top button panel

The buttons on the top panel relate to the selected protein

- ✗ → ✓ When this button is pushed the protein will be given a *validated* status.
- ✓ → ✗ When this button is pushed the protein will be given a *not validated* status.
- 🌐 When this button is pushed the default internet browser will be opened to the website (if available) describing this protein.
- ●●● When this button is pushed the possible *isoforms* (proteins with peptides that can be linked to two or more proteins) will be loaded in the protein list (see section 3.2).
- 👎→👍 When this button is pushed the protein will be *selected*.
- 👍→👎 When this button is pushed the protein will be *deselected*.

3.8.2 Lower button panel

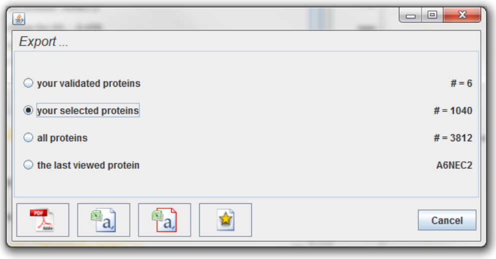
The buttons at the lower end have general functions.

- When this button is pushed the filter panel appears (see section 3.4).
- When this button is pushed a small panel appears where the protein reference set can be adapted (see section 2.3.3.2).
- When this button is pushed a small panel appears with information concerning the protein reference set. The number of proteins and ratios used in the reference set as well as the standard deviation and mean calculated with the reference set will be listed in this panel. The exact limits of the 95% confidence interval [-1.96 σ; 1.96 σ] and the data limits [2.5%, 97.5%] are given for every ratio type.
- When this button is pushed a small panel appears with a log. In this log you can see how many proteins were filtered, the protein that was validated, deselected, ...
- When this button is pushed (only available if the loaded data are Mascot Distiller Quantitation toolbox data) all the ratio groups of the selected protein will be grouped by *.rov* file and *hit* and will be put in a list on a second tab. When one ratio group is selected the "*more information panel*" is shown (see 3.7). In this tab, the ratio groups with invalid ratios can be selected by clicking the "*show only invalid ratios*" check box at the bottom of the tab.
- When this button is pushed the save/export panel appears (see 3.9).
- When this button is pushed selected proteins and information on peptide ratios, ratio comments, . . . will be loaded from a *.ROVER* file.
- A check box controls the state ($\log_2$ or normal value peptide ratios) of the calculated and visualized ratio.
- A check box controls the state of the "*Use only uniquely identified (blue) peptides*".
- A check box controls the state of the "*use only valid peptide ratios*".
- A check box controls the state of the "*use original ratios*". If this is checked the original ratios will be used in the calculations and not the location and scale normalized ones. This is only available in the multiple sources approach.

3.9 Export panel

Different types of export can be choosen on the export panel (see figure 3.6). Also, the user can choose which proteins will be exported. *All* proteins, the *selected* proteins, the *validated* proteins or the *previously viewed* protein can be choosen.

- When this button is pushed, the protein bar (see 3.3), the ratio distribution graph (see 3.5) and the information in the protein panel (see 3.6) will be written in a *.PDF* file for every protein selected for export.
- When this button is pushed information for every ratio group from every protein selected for export will be written in an excel viewable *.CSV* file.
- When this button is pushed information for every protein selected for export will be written in an excel viewable *.CSV* file.
- When this button is pushed a *.ROVER* file will be written. In a *.ROVER* file the user adjusted data will be stored[1]. Five types of user adjustable data exist.
    - The ratio status can be set valid or invalid.
    - A comment can be given to a ratio
    - A protein can be set selected and validated
    - Peptide groups (peptide with the same peptide identification) can be excluded in the protein mean calculation.
    - A comment can be given to a protein.



Export ...

○ your validated proteins # = 6

● your selected proteins # = 1040

○ all proteins # = 3812

○ the last viewed protein A6NEC2

Figure 3.6: The export panel

[1] If quantitative data from ms_lims are loaded, the ratio status and ratio comment will be stored in the database.

Chapter 4

Statistics

Statistics are used to compare and review the peptide ratios from the selected proteins with the peptide ratios from a reference set. This reference set is a number (20-300) of the most abundant proteins or are all proteins found in the data (see section 2.3.3.2).

Robust statistics is used to correct for the effect of outliers on the average and standard deviation (σ). Not the mean but the median is used as the average $\log_2$ peptide ratio value for the reference set. The standard deviation is used as measure of the variability of the $\log_2$ peptide ratios in the reference set.

The standard deviation is calculated by a Huber's M-estimation. The $\log_2$ peptide ratios will be transformed by a process called winsorisation. In this process, the $\log_2$ peptide ratios with a value (see formula 4.1 and 4.2) will be changed in 4.3 and 4.4 respectively.

$$\log_2(peptide\ ratio) < median - 1.5\ \sigma\ (of\ the\ previous\ winsorization) \tag{4.1}$$

$$\log_2(peptide\ ratio) > median + 1.5\ \sigma\ (of\ the\ previous\ winsorization) \tag{4.2}$$

$$median - 1.5\ \sigma\ (of\ the\ previous\ winsorization) \tag{4.3}$$

$$median + 1.5\ \sigma\ (of\ the\ previous\ winsorization) \tag{4.4}$$

After each winsorisation the standard deviation will be calculated and is 1.134 times the standard deviation of the winsorised data. If the difference between two consecutive standard deviation is smaller than 0.000001 the winsorisation cycle stops and the the standard deviation (σ_{call}) will be 1.134 times the standard deviation of the last winsorised data.

The instrument standard deviation (σ_{instr}, see section 2.3.3.4) is also used to calculate the final/corrected standard deviation (σ_{corr}). This instrument standard deviation is the measured standard deviation for $\log_2$ scale ratios of peptides mixed in equal amounts on the mass spectrometer used for your project. The corrected standard deviation will be calculated with formula 4.5

$$\sigma_{corr} = \sqrt{\sigma_{call}^2 + \sigma_{instr}^2} \tag{4.5}$$

A significance or Z-score can be calculated for a specific peptide ratio by using formula 4.6. The P value is calculated with formula 4.7.

$$Z = \frac{(\log_2(peptide\ ratio) - median)}{\sigma_{corr}} \tag{4.6}$$

$$P = 1 - \mid erf(\frac{Z}{\sqrt{2}}) \mid \tag{4.7}$$

The $\log_2$ values of the peptide ratios of the reference set are normally distributed and the median typically centers around zero in an experiment in which equal amounts of peptides were loaded. The normal unchanged value of the peptide ratios are not normally distributed and the median is typically ± 1.0. These two distributions can be seen in figure 4.1.

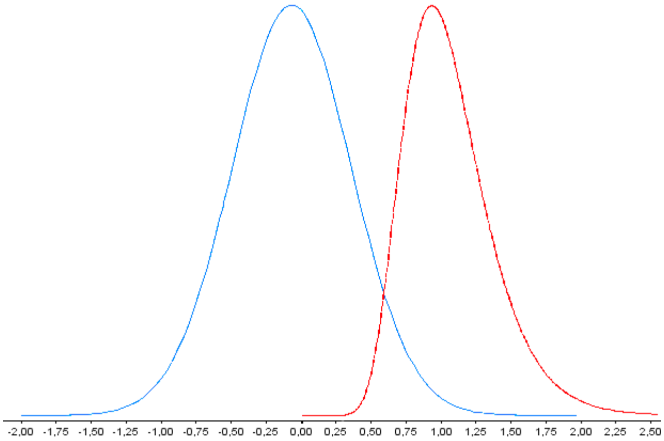


Figure 4.1: The $\log_2$ distribution of the peptide ratios in the reference set is in blue and is normally distributed. The untransformed peptide ratio distribution is in red and is not normally distributed and skewed to the right.

Chapter 5

Working with Rover, an example

Open the Rover wizard and load the data as described in chapter 2.

In a typical quantitative proteomics experiments the aim of the study is the identification of regulated proteins. We will use two filters for the selection of those regulated proteins. With the first filter we will try to find proteins with a protein mean for L/H that is larger than 1.6. The second filter will filter proteins with a protein mean for L/H that is smaller than 0.5. The boundaries of these filters are derived from the ratio distribution graph or can be found in the panel that appears after ℹ is clicked. Every time a filter is applied, Rover warns the user how many proteins were filtered and put these in the protein list (see 3.2). These are "interesting proteins", and that is why we will select all these proteins for validation by clicking ✚. This must be done after each filtering step.

At this point we have a group of selected proteins. We can view these by clicking 🔍. Now it's time to validated every protein. After you have decided to keep the protein or discard it, you have to click the validated button ✖→✓. If you want to deselect the protein afterwards, you have to click the deselect button 👍→👎.

After the validation we will save the validated proteins and peptide (ratios) in a *.ROVER* file. With this *.ROVER* file, the analysis can be reloaded. The validated or selected proteins can be exported to a *.PDF* or *.CSV* file using the export panel (see section 3.6).

Chapter 6

Problems and questions

A google discussion group (`http://groups.google.com/group/rover_quantviewer`) was created for problems and questions. Also, if you have a request (a new filter, ...) you can post a message on the google discussion group.

Issues can also be posted on the issue page of the Rover program (`http://code.google.com/p/compomics-rover/issues/list`). It would be helpful if the issue is well described and if the log file (rover-log4j/log) could be attached the the message (the log file is located in the user/.compomics/rover folder).