The KNIME workflow PD-2-ProLiC

Quick reference guide

1. General remarks

This is a quick reference guide for the usage of the PD-2-ProLiC workflow, which can be utilized within the software Konstanz Information Miner Analytics Platform (KNIME (1), http://www.knime.org/, Hyperlink checked at January 13, 2018). PD-2-ProLiC is created for the conversion of result files obtained from the Proteome Discoverer software (Thermo Fisher Scientific, Waltham, MA) into the ProLiC formatted input format. ProLiC is an open source software for the comparison of a potentially arbitrary number of files that give proteomics identification results. Such results are usually given as a list of accession numbers, which are obtained from proteomics databases. ProLiC supports different comparison methods, namely either an *accession based*, *protein sequence based* or *peptide based comparison*. The term *peptide based comparison* refers to a comparison of sets of identified peptides that are assigned to different proteins.

Here, the usage of PD-2-ProLiC is briefly described. We will not go into details, but the necessary information is given that enables full control over the conversion process.

1.1 Conventions

Locations of nodes within the KNIME node repository, labels of dialog box elements and menu options are highlighted in *italics* using `Courier` font.

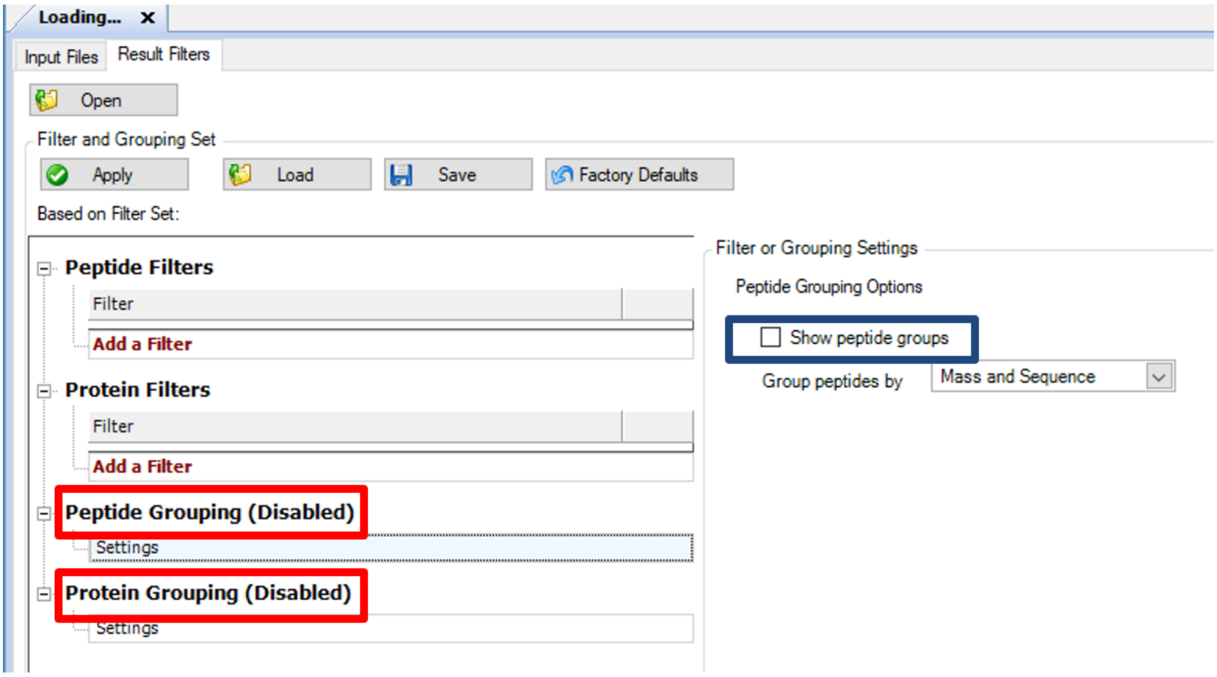
2. Exporting Proteome Discoverer result files

This section deals with exporting of Proteome Discoverer result files. These are the input files for the PD-2-ProLiC workflow.

The procedure is exemplarily described for Proteome Discoverer version 1.4 result files. For more details or if you use a different Proteome Discoverer version, please refer to the appropriate Proteome Discoverer manual.

These manuals are available from the Thermo Fisher Scientific website (https://www.thermofisher.com/de/de/home.html, Hyperlink checked at February 13, 2018).

As a first step for exporting, please load a Proteome Discoverer result file (i.e. a file with the file extension .msf) within the Proteome Discoverer software**. It is important that protein and peptide grouping is disabled** (Figure 1), because later on ProLiC applies its own grouping approach. To this end, disable the checkboxes for both `Show peptide groups` and `Enable protein grouping` within the dialog shown in figure 1. Afterwards, grouping within the Proteome Discoverer software is switched off.



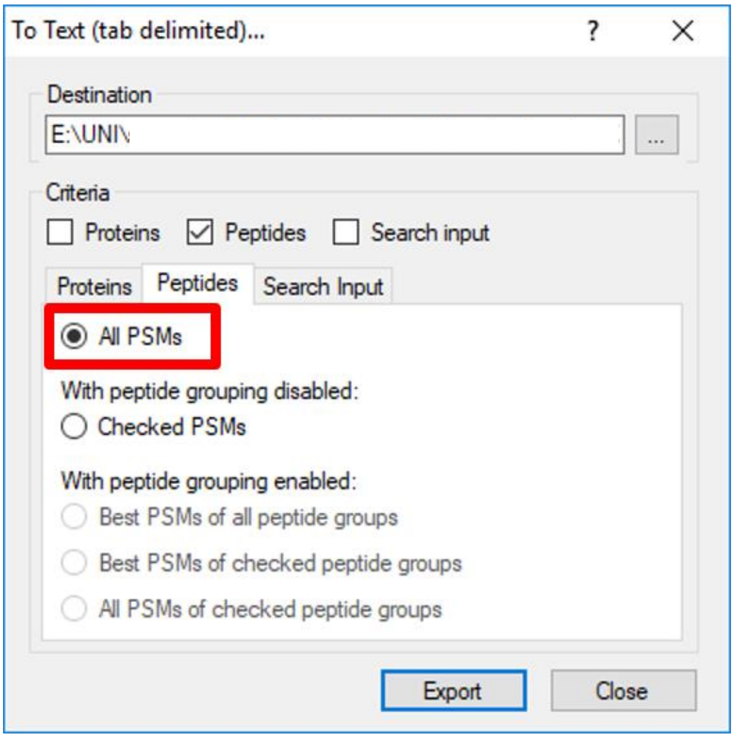
☐ Show peptide groups

Group peptides by: Mass and Sequence

Figure 1: Screenshot of the Proteome Discoverer ‘Result Filters’ dialog. Red rectangles highlight correct display for a successful deactivation of both protein and peptide grouping. The blue rectangle indicates the checkbox that switches peptide grouping on and off.

After Proteome Discoverer has finished the analysis, the results can be exported via the menu: `Choose File – Export – To Text (tab delimited)…`.

In the appearing dialog (Figure 2), please ensure that the option `All PSMs` is checked. After clicking the `Export` button, files with the suffix *_pms.txt* should appear in the specified folder. These files are input for the PD-2-ProLiC workflow.



To Text (tab delimited)...

Destination: E:\UNI\

Criteria: ☐ Proteins ☑ Peptides ☐ Search input

● All PSMs

With peptide grouping disabled:

○ Checked PSMs

With peptide grouping enabled:

○ Best PSMs of all peptide groups

○ Best PSMs of checked peptide groups

○ All PSMs of checked peptide groups

Figure 2: Screenshot of the Proteome Discoverer ‘To Text (tab delimited)…’ dialog. The required selection is highlighted with the red rectangle.

3. Workflow installation

Because PD-2-ProLiC is a KNIME workflow and uses capabilities of the R software environment, you must install both the KNIME Analytics Platform (1), which is downloadable from https://www.knime.com/downloads (Hyperlink checked at January 13, 2018) and R ((2), downloadable from https://cran.r-project.org/mirrors.html, Hyperlink checked at January 13, 2018) in a version that is appropriate for your operating system.

Because several R packages are used within the PD-2-ProLiC workflow, please install the following packages within R:

- Use install.packages("Rserve") in order to install the package **Rserve** from CRAN (Comprehensive R Archive Network).

- Use the following commands in order to install the package **UniProt.ws** from Bioconductor:

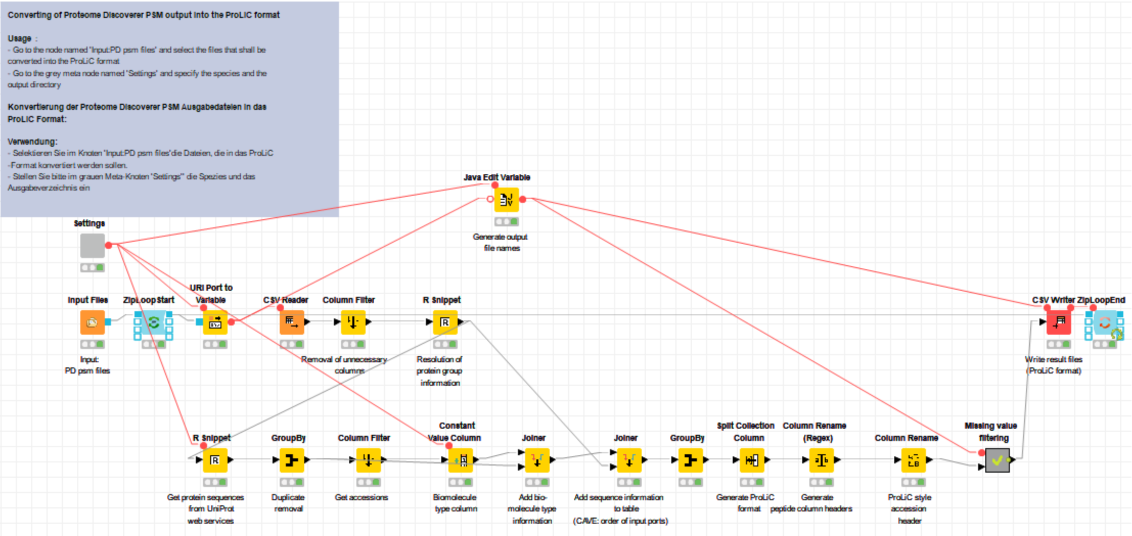
```
source("https://bioconductor.org/biocLite.R")

biocLite("UniProt.ws")
```

Afterwards connect KNIME and your local R installation: Within the KNIME software, open the preferences via the menu entry `File – Preferences`. In the appearing dialog select `KNIME – R` and specify the appropriate path in the text field named `Path to R Home`.

You are now ready to import the PD-2-ProLiC workflow into KNIME. To this end, please use the menu option `File-Import KNIME workflow` and select the file named PD-2-ProLiC.knwf, which is part of the ProLiC download.

If the import succeeded, the name of the workflow appears within the frame named ‘KNIME Explorer’, which is located top left per default. Double-click on the workflow entry to open it. Within the central frame of KNIME (i.e. the so-called workflow editor) the PD-2-ProLiC workflow appears (Fig. 3). The box located in the upper left named ‘Converting of Proteome Discoverer PSM output into the ProLiC format’ gives the most important information for utilization of the workflow.



Converting of Proteome Discoverer PSM output into the ProLiC format

Usage :
- Go to the node named "Input:PD psm files" and select the files that shall be converted into the ProLiC format
- Go to the grey meta node named "Settings" and specify the species and the output directory

Konvertierung der Proteome Discoverer PSM Ausgabedateien in das ProLiC Format:

Verwendung:
- Selektieren Sie im Knoten "Input:PD psm files"die Dateien, die in das ProLiC -Format konvertiert werden sollen.
- Stellen Sie bitte im grauen Meta-Knoten "Settings"' die Spezies und das Ausgabeverzeichnis ein

Figure 3: Screenshot of the loaded PD-2-ProLiC workflow within KNIME

4. Settings / Workflow preparations

The grey node named ‘Settings’ located on the left-hand side of the workflow editor allows customization of the workflow.

Double-click on the ‘Settings’ node and the settings dialog is shown (Figure 4). In the remainder of this paragraph, the settings are described in detail.

4.1 Settings

- Setting `Available species`

  For both sequence and a specific type of peptide based comparison ProLiC requires protein sequence information. However, the Proteome Discoverer output files do not contain such information. Therefore, the PD-2-ProLiC workflow retrieves appropriate protein sequences from the internet. Currently, this is supported for three species, namely human, mouse and rat. Users must specify the appropriate selection that fits their experiments.

- Setting `Output directory`

  The here specified directory (given as absolute path) is the location where the generated result files and the log file of the workflow are stored after execution of the PD-2-ProLiC workflow.
  This setting is important to enable the execution of the workflow in general. The originally given location (C:\tmp) is only a wildcard. Please replace this entry with a location that fits to your computer. It is also important that you have writing permissions for this directory.

- Setting `Biomolecule type selection`

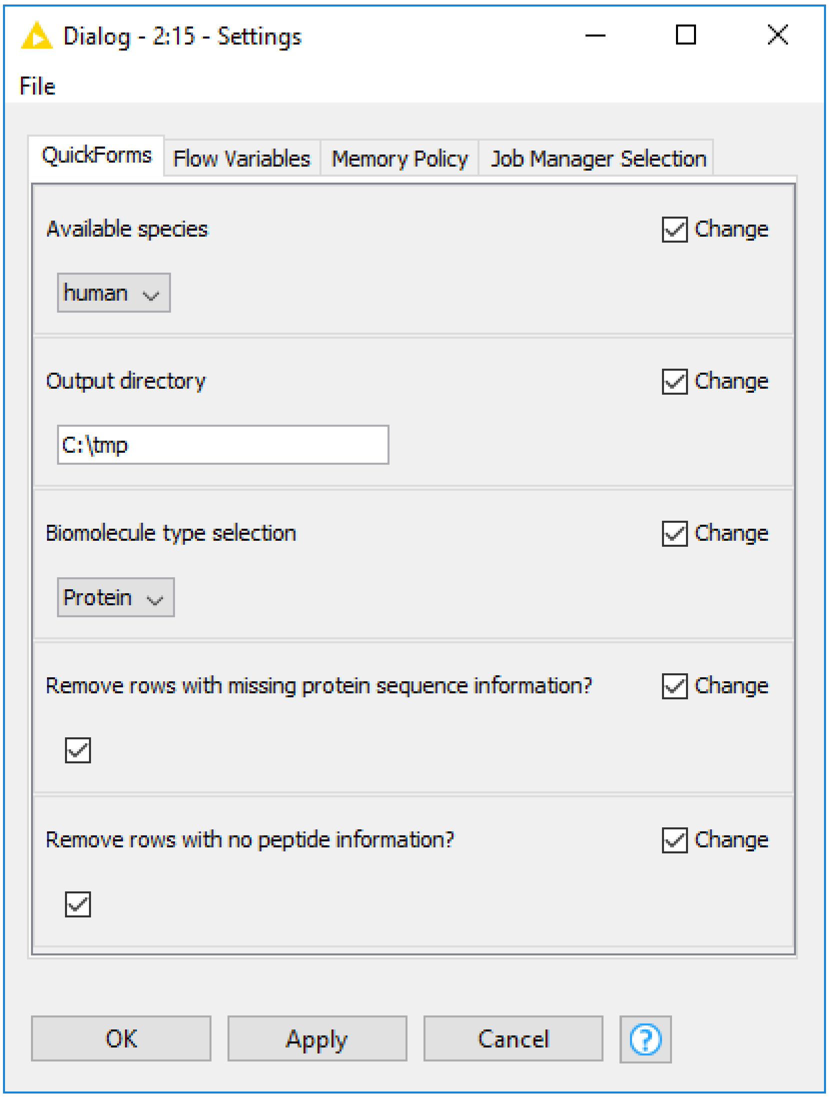
  Future versions of ProLiC will not only support comparison of proteomics results, but also enable comparison of transcriptomics experiments. This parameter is included with this application in mind. Currently only ‘Protein’ can be chosen as parameter setting.

- Setting `Remove rows with missing protein sequence information?`

If this parameter is checked, rows are removed if no protein sequence information could be retrieved from the internet.

- Setting `Remove rows with no peptide information?`

  If this parameter is checked rows are removed if no peptides (i.e. peptide sequences) are assigned for this protein.



Available species ☑ Change
human

Output directory ☑ Change
C:\tmp

Biomolecule type selection ☑ Change
Protein

Remove rows with missing protein sequence information? ☑ Change
☑

Remove rows with no peptide information? ☑ Change
☑

Figure 4: Screenshot of the dialog that allows the setting of all relevant workflow parameters. Please note that the setting C:\tmp of the parameter `Output directory` is only a wildcard. Please enter the path to a directory with sufficient writing permissions. The here specified directory receives the result files that are generated by the workflow.

5. Running the workflow

Because the protein sequence information must be retrieved from the internet, a successful execution of the workflow requires access to the internet.

Load the workflow PD-2-ProLiC within the ‘KNIME Analytics Platform’ software. Afterwards, double-click on the ‘Settings’ node and set up the workflow. Also, double click on the node named ‘Input PD psm files’ and select the files previously exported from the Proteome Discoverer software that should be converted into the ProLiC file format. Select the node named ‘ZipLoopEnd’ on the far right of the workflow. Afterwards, the workflow can be started with each of the two execution icons highlighted in Figure 5.

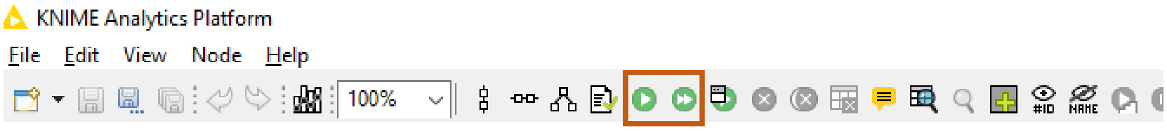
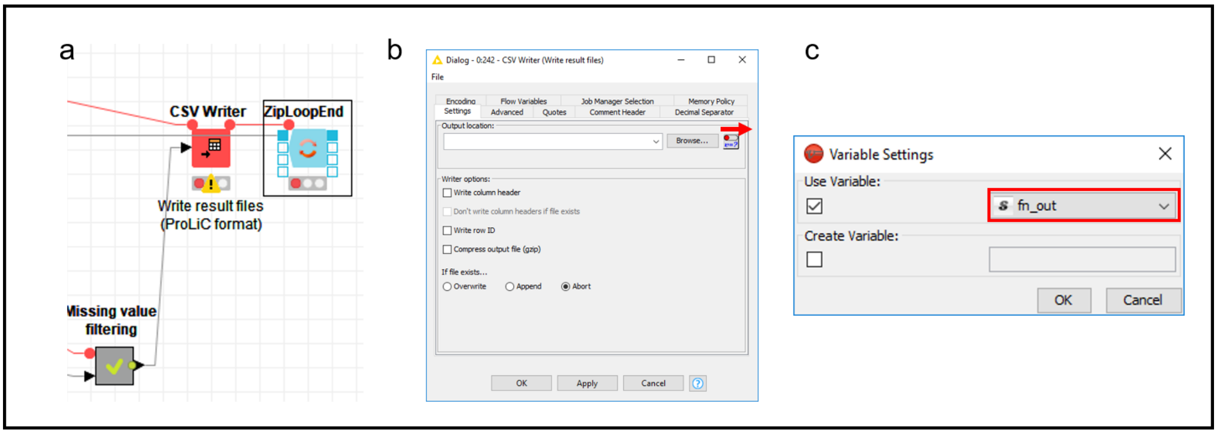


Figure 5: Detail view of the menu bar and the tool bar of the software ‘KNIME Analytics Platform’. The icons that can be used to start a KNIME workflow are highlighted with a red rectangle.

During the first run of the workflow a node named ‘Write result files (ProLiC format)’, which is located next to the final ‘ZipLoopEnd’ gives an error (Figure 6, a). In order to solve this issue, please open the dialog of this node (Figure 6, b) and draw the red dot in the dialog element labelled ‘v=?’ to the right. In the appearing dialog (Figure 6, c) please select the variable named ‘fn_out’ from the list box. Further executions of the should proceed without similar problems.



☐ Write column header
☐ Don't write column headers if file exists
☐ Write row ID
☐ Compress output file (gzip)

If file exists... ○ Overwrite ○ Append ● Abort

Use Variable: ☑ fn_out

Create Variable: ☐

Figure 6: Detail view of PD-2-ProLiC workflow elements. Figure *a* shows the error thrown by the node named ‘Write result files (ProLiC format)’. Figure *b* presents the configuration dialog of the ‘Write result files (ProLiC format)’ node. The red arrow indicates the required direction of movement of the subjacent dialog element in order to reach the subsidiary dialog shown in Figure *c*. The selection of the appropriate flow variable is highlighted by the red rectangle.

After a successful execution of the workflow all traffic lights below the nodes should have switched to green (cf. Figure 3). You will find the converted files in the folder that has been specified with the `Output directory` setting in the dialog of the 'Settings' node. These files are ready to use as input files for comparison with the ProLiC software (Please ensure that the file extension matches the settings specified in the ProLiC configuration file).

6. References

1. Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., and Wiswedel, B. (2007) KNIME: The Konstanz Information Miner. *Studies in Classification, Data Analysis, and Knowledge Organization*, Springer

2. R Core Team (2016) *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria. URL http://www.R-project.org/.