MPA Portable Tutorial (October 2017)

1. Starting MPA Portable (GUI mode)

To launch the MPA Portable application in GUI mode under Windows, please double-click either the *mpa-portable-X.Y.jar* or *mpa-portable.bat* file. Besides the demonstrated Windows version in this tutorial, Linux users similarly may use the *mpa-portable.sh* file or launch the application from the terminal.

Note that the memory heap size can be increased by changing the RAM parameter in the *mpa-portable.bat / mpa-portable.sh* file. For example, a maximum memory usage of 1500 MB can be parameterized by the following setting: **–Xmx1500m** (Default). In case you have more memory available, feel free to assign more memory: the more, the smoother the application will run.

2. Creating projects and experiments

On startup of MPA Portable, four workflow tabs are shown on the left side of the user interface:

- Project
- Input Spectra (greyed out)
- View Results (greyed out)
- Logging

When the application is launched the *Projects* and *Experiments* tables are empty and all buttons except the *New Project* button are greyed out. In order to prepare the processing of sample files, a **project** and an **experiment** need to be created.

Please create a project by performing the following steps:

- ✓ Click the *New Project* button to open the *New Project* dialog.
- ✓ Enter a title in the *Project Name* field inside the dialog (properties can be ignored).
- ✓ Confirm by hitting the *Save* button to close the dialog.

A project may contain several experiments (e.g. different runs or search workflows). In this case, only one experiment will be created.

Please create an experiment as follows:

- ✓ Click the *Add Experiment* button to open the *New Experiment* dialog.
- ✓ Enter a title in the *Experiment Name* field inside the dialog (properties can be ignored).
- ✓ Confirm by hitting the *Save* button to close the dialog.

3. Choosing files and parameters

3.1 Spectrum file(s) selection

After hitting the *Next* button (or selecting the *Input Spectra* tab), the spectrum file can be chosen by clicking on the *Add Spectrum File(s)* button. The supported MS/MS spectrum formats is MGF (Mascot Generic Format).

- ✓ For this tutorial, please choose the **Ebendorf1000.mgf** file as input file from the provided *Dataset* folder.

In order to convert RAW spectrum data to MGF, please use the freely available ProteoWizard tool or employ the vendor software delivered with the MS instrument.

3.2 Protein database formatting

The next step involves the selection of a protein sequence database in FASTA format. Note that MPA Portable automatically preprocesses the chosen FASTA file before the search identification process. During this formatting procedure, the following steps are performed:

1. Reversing of the protein sequence database (Decoy creation)
2. Indexing of the protein sequence database (Search algorithms and MPA Portable)

- ✓ Please choose the **uniprot_sprot.fasta** file as input file from the provided *FASTA* folder. Note that processing the whole protein sequence sequence might take a while. Meanwhile, details on the progress can be found in the status bar or in the *Logging* panel.

3.3 General settings

For general settings, MPA Portable provides the following three parameters:

- Precursor ion tolerance (MS1 level; in Dalton or PPM)
- Fragment ion tolerance (MS2 level; in Dalton or PPM)
- Missed cleavages (Number of maximum missed cleavages)

- ✓ For this tutorial, please select **10 ppm** as precursor ion tolerance. Fragment ion tolerance and missed cleavages can be left as default (0.5 Da and one missed cleavage)

3.4 Search engine selection

MPA Portable supports X!Tandem (Craig *et al.* 2004), Comet (Eng *et al.* 2013) and MS-GF+ (Kim and Pevzner 2014) as database search algorithms for peptide/protein identification. These algorithms can be selected for sequential processing of the files. Internally, each of the algorithms features multiple threads (processes) to speed up the identification search *via* parallelization on multiple CPU cores.

- ✓ For this tutorial, the default selection of using X!Tandem as single search engine can be kept.

Additional parameters can be modified by clicking on the green wheel next to the respective algorithms X!Tandem, Comet and MS-GF+.

4. Identification search and results inspection

4.1 Processing data

When all respective parameters have been set, the processing can be started.

- ✓ Please click on the *Start searching* button to initiate the MPA Portable search.
- ✓ To inspect the details of the processing, click on the *Logging* panel tab.

Note that the processing might take between 10-20 minutes – depending on the CPU hardware. If any error occurs, details are shown in the *Logging* panel.

To inspect the results please perform the following steps:

- ✓ Click on *the View Results* tab.
- ✓ Hit the Fetch Results button and wait for the process to finish.

4.2 Overview panel

The *Overview* panel (**Figure 1**) contains various options to explore raw or processed results:

1. The *Summary* table provides details on the number of identified spectra, peptides and proteins.
   - ✓ How many MS/MS spectra have been identified?
   - ✓ How many peptides and proteins are found?
   - ✓ What is difference between distinct and unique peptides?
2. The top-right *Chart View* panel presents taxonomic and functional information in bar and pie charts while also allowing selecting individual elements for further inspection in the bottom-right *Chart Details* table.
   - ✓ How many proteins are found for Amino-acid biosynthesis?
   - ✓ How many spectra were assigned to Methanogenesis?
   - ✓ What is the most abundant phylum at the spectrum and peptide level?
   - ✓ Why are there different quantitative units (e.g. proteins/peptides/spectra)?

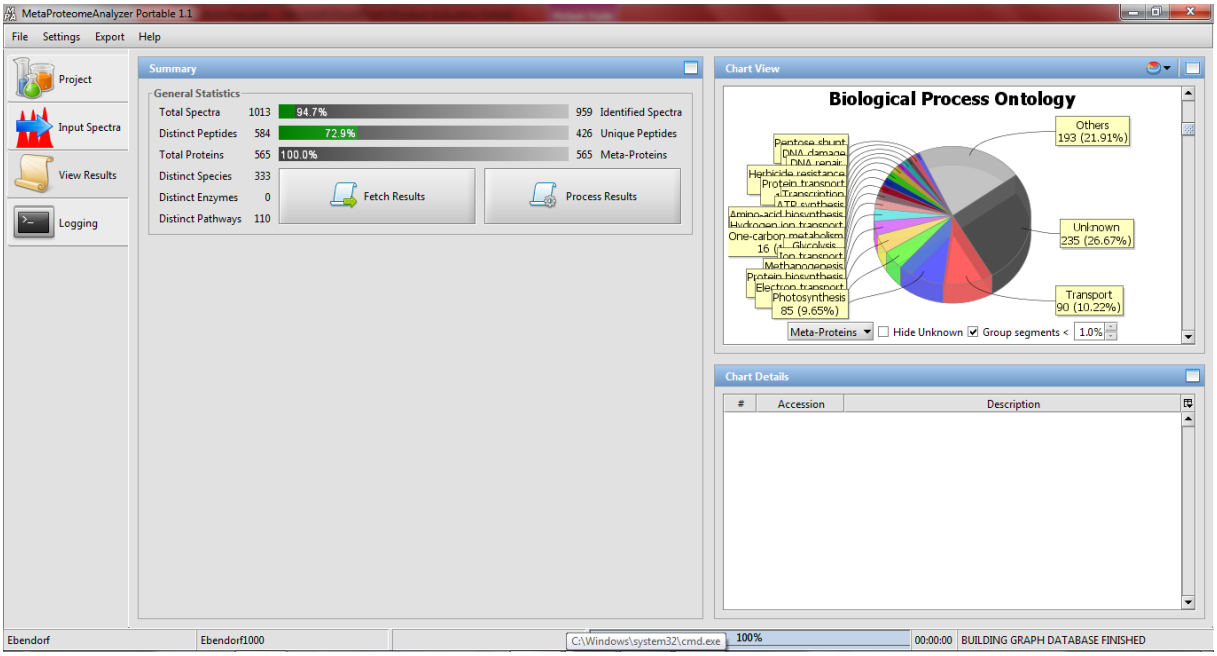

Figure 1: The Overview panel of the Database Search Results

4.3 Examining Results in Detail

The *Database Search Results* panel contains various, primarily tabular views for exploring the search result contents in detail (**Figure 2**).

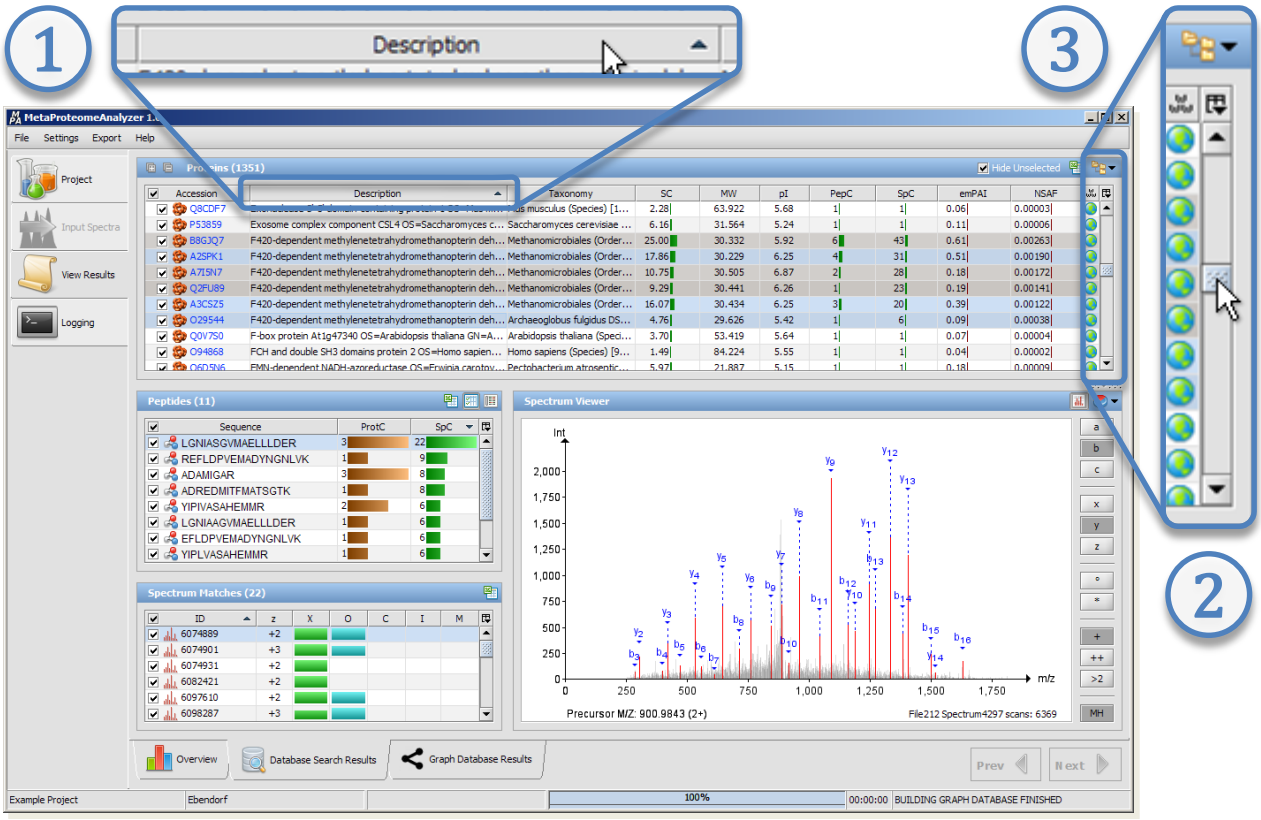

Figure 2: The Basic View of the Database Search Results panel

1. ✓ click the table header labeled *Description* in the *Proteins* table to sort the table contents in alphabetical order

2. ✓ scroll down in the *Proteins* table and locate protein entries with the description *F420-dependent methylenetetrahydromethanopterin dehydrogenase*
   - ✓ How many proteins are found for this entry?
   - ✓ Why do we find multiple entries for a particular protein?

Next, the protein results need to be processed further, e.g. for grouping protein identifications (meta-protein generation):

- ✓ click on the *Process Results* button in the *Overview* panel
- ✓ in the following dialog, leave the default values and hit the *OK* button
- ✓ Play around with the settings for the meta-protein generation and observe how the parameters affect the absolute amount of meta-proteins (see *Summary* panel)

The generation of the meta-proteins takes a few moments. Eventually, the generated meta-proteins can be inspected in the *Meta-Protein* View in the *Database Search Results* panel.

The tabular protein views in the Database Search Results Panel provide additional information for each listed protein in the table columns which may be sorted or hidden individually via right-click context menu on the column header. In addition, an option exists for numerical columns in hierarchical views to select an aggregate function to determine values for category rows.

Note that some columns are hidden by default and may be restored for each table individually using the button in the top right corner of the table.

Certain elements in the Accession column of the Proteins table contain links to external resources such as KEGG Pathway maps in the Pathway View or UniProt Knowledgebase entries in all views.

Additional useful web resource links are listed in a context menu available by clicking the respective button. Note that internet access is required for all external web resources.

4.4 Examining the Meta-Protein View

The *Meta-Protein View* (**Figure 3**) features a specialized table containing proteins grouped under meta-proteins defined according to the settings chosen in step 3.2.

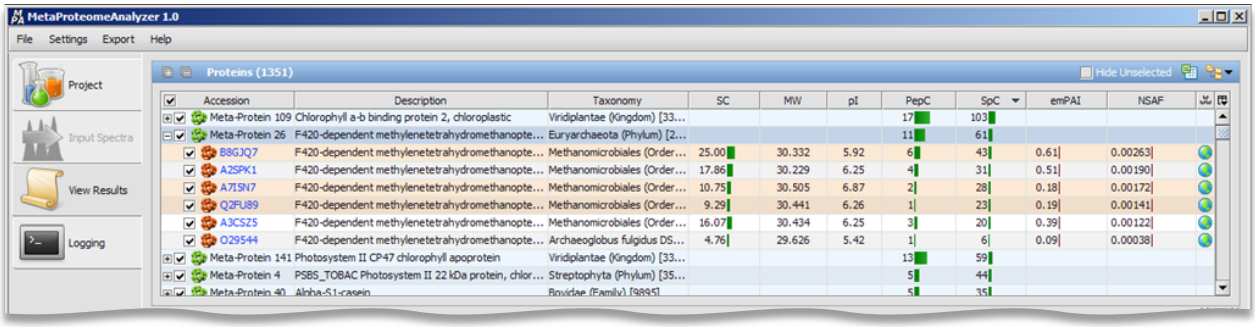

Figure 3: The Meta-Protein View of the Database Search Results Panel

- ✓ Please locate the meta-protein labeled *MTD_METB6 F420-dependent methylenetetrahydromethanopterin dehydrogenase*
- ✓ Expand its table entry by clicking the ⊞ icon you should see all six instances previously found in the *Basic View* grouped under this meta-protein.
- ✓ Select the meta-protein in the top protein table and click on *Coverage View* icon in the upper right corner of the *Peptides* panel (middle)
- ✓ How many peptides are found for this protein group aka. meta-protein?

Besides the Meta-Protein View various other hierarchical views can be selected each of which categorizes proteins in different fashion. For instance, all six instances of the example protein F420-dependent methylenetetrahydromethanopterin dehydrogenase are also grouped together in the Enzyme View (see 4.5 Enzyme View).

For even more in-depth data analysis the application's Graph Database Results panel's capabilities may be employed. Here you can either select from a list of pre-defined queries or define your own to create a custom hierarchical tabular view on the results. This way you are able to tailor the output to specific biological questions you seek to answer.

4.5 Taxonomic View

Taxonomic assignments are shown up to the species/strain level. Checkboxes enable/disable respective identifications in the other views.

- ✓ Please deselect the checkbox for Eukaryota to filter out eukaryotic proteins
- ✓ Have a look at the other views (Basic view / Meta-Protein View etc.) to observe the changes when filtering out certain taxa.
- ✓ Deselect the *Hide Unselected* checkbox in the top right corner of the protein view to inspect the entries that have been removed from the other views.
- ✓ Which taxa are listed for *F420-dependent methylenetetrahydromethanopterin dehydrogenase*?

Methyl-coenzyme M reductase is a key enzyme for the production of methane in biogas plants. Next, we try to spot this protein in the taxonomic view.

- ✓ Filter out all superkingdom entries except for *Archaea*. Expand the *Archaea* superkingdom.
- ✓ Which phyla and species are found for *Methyl-coenzyme M reductase*?

4.5 Enzyme View

Enzyme Commission (EC) identifiers characterize protein identifications by classification keywords/terms (e.g. EC 3.X.X.X = Hydrolases). In the following, we try to find out to which EC number *F420-dependent methylenetetrahydromethanopterin dehydrogenase* was assigned.

- ✓ Expand the tree view in the *Enyzme* View (by clicking on the *Expand All* button) in the upper left corner and try to find the F420-MTMHO proteins.
- ✓ Below which EC identifier is the protein listed?
- ✓ What are the first (EC X.-.-.-) and the second (EC X.Y.-.-) EC category?

4.6 Pathway View

The previous proteins belong to the pathway of methane production in biogas plants. Therefore, we now have at a look at this pathway route in more detail.

- ✓ In the *Pathway View*, go to Metabolism → Energy metabolism → Methane metabolism
- ✓ Which identifier can be found for this pathway?
- ✓ Besides *F420-dependent methylenetetrahydromethanopterin dehydrogenase,* which proteins are found below this pathway?
- ✓ Click on the methane metabolism identifier to open up a KEGG pathway mapping inside a web browser. What can you find there? Which purpose fulfil the arrows?

5. Optional: Command line application

5.1 Windows (CLI mode)

Here is a minimum working example of the command line interface for the Windows operating system. *X*, *Y* and *Z* have to be replaced by the actual version of the MPA Portable software and *my folder* by the folder containing the desired files:

```
java -cp mpa-portable-X.Y.Z.jar de.mpa.cli.CmdLineInterface
-Xmx4000m
-spectrum_files C:\my_folder\spectrum_file.mgf
-database C:\my_folder\uniprot_sprot.fasta
-missed_cleav 1
-prec_tol 10ppm
-frag_tol 0.5Da
-output_folder C:\my_folder\output
```

For sake of readability, the input parameters are split over multiple lines. When using the command line, however, all parameters should be included as single line.

Note that the memory heap size can be increased by changing the -Xmx parameter. Here, a maximum memory usage of 4GB is specified as follows: ***–Xmx4000m****.*

5.2 Linux (CLI mode)

Here is another example of the command line interface for the Linux operating system featuring all optional parameters explicitly. In this setup, X!Tandem, Comet and MS-GF+ are employed (using 8 threads) for protein identification, iterative searching is turned off, an FDR threshold of 1% is applied and proteins are grouped based on the meta-protein rule of requiring a single shared peptide. Both taxonomy and cluster rule are turned off.

```
java -cp mpa-portable-X.Y.Z.jar de.mpa.cli.CmdLineInterface
-Xmx4000m
-spectrum_files /home/my_folder/spectrum_file.mgf
-database /home/my_folder/uniprot_sprot.fasta
-missed_cleav 1
-prec_tol 10ppm
-frag_tol 0.5Da
-xtandem 1
-comet 1
-msgf 1
-iterative_search 0
-fdr_threshold 0.01
-generate_metaproteins 1
-peptide_rule 0
-cluster_rule -1
-taxonomy_rule -1
-threads 8
-output_folder /home/my_folder/output/
```

Note that the memory heap size can be increased by changing the -Xmx parameter. Here, a maximum memory usage of 4GB is specified as follows: ***–Xmx4000m****.*