Meta4P

MetaProteins-Peptides-PSMs Parser

User guide

0. Start the application

Double-click on the Meta4P.exe file to start the application. The first time the program is launched, the Windows OS will ask for the user's explicit approval to execute the application. Click on "More info" to proceed with security checks and then on "Run anyway".

1. Input type

Based on the analysis of interest, select the input type by clicking on the corresponding button ("**Proteins**", "**Peptides**" or "**PSMs**").

Alternatively, if you have previously downloaded Meta4P outputs and just want to rename and/or reorder sample column headers, click on "**Rename/reorder sample columns**" to go directly to the last Meta4P window.

2.1. Proteins

In this step, protein identification and quantification data can be retrieved from an input file in .xlsx format and filtered based on the user's requirements. Structure and column headers of the input file must comply with those of the corresponding template file (*Proteins_input_template.xlsx*, based on the "Proteins" output table generated by *Proteome Discoverer*) provided along with this user guide. If the input file does not contain all the required columns, an error message will be shown.

Click on "**Upload input file**" to select and upload the protein identification list (including abundance columns). Once the file is uploaded, the window is populated with all the filtering options based on the file content.

Proteins can be filtered based on:

- ***Statistical confidence of protein identification ("Confidence")***: only proteins with the selected level(s) of confidence (low, medium and/or high) are kept in the output, based on the checkbox(es) checked.
- ***Protein description text ("Protein Description")***: to select only those proteins which contain a specific text (e.g., a protein name or an organism name) in their "Description" column, type the text of interest in the textbox (be aware that the filter is case sensitive) and click on "**Add**". Multiple texts can be typed and added sequentially; in this case, the user can choose between two boolean operators, "**And**" and "**Or**", to determine whether all the texts added or only one of them must be present in the string, respectively, so that a protein passes the filter.
- ***Number/percentage of valid values ("Valid values threshold")***: only proteins with a number of valid values (i.e., non-missing values) greater than or equal to the selected threshold pass the filter. The user can choose between indicating absolute or percentage values by selecting the corresponding option in the drop-down menu. Leaving the default value (0) means that all proteins pass the filter. The total number of samples in the dataset is shown in brackets under the textbox.
- ***"Marker"***: marker names included in the "Marked as" column of the input file, usually indicating which of the protein database(s) used for identification contained that protein sequence, are retrieved by the software and shown next to their respective checkboxes; only proteins annotated with the checked marker names are kept in the output.

Furthermore, the user can choose between the following visualization and calculation options:

- ***Master proteins only***: if selected, only proteins designated as "Master Protein" – i.e., a protein identified by a set of peptides that are not included (all together) in any other protein group – are kept in the output.

- ***Select normalized abundances***: if selected, normalized abundance values will be reported in the table; this option is available only when normalized abundance values are included in the input file.
- ***Re-normalize abundances after filtering***: if selected, once the chosen filters are applied and the filtered protein list is obtained, (non-normalized) abundance values measured for a protein in a sample are divided by the total protein abundance measured in that sample and multiplied by 10^{10}.
- ***Replace missing values with 0***: if selected, missing values (empty cells) are replaced by 0; this selection will be applied to all the following output tables.

At the end, the filtered table can be downloaded as a Microsoft Excel file (.xlsx format) by clicking on "**Download filtered table**".

Click on "**Next step**" to go to the "**Taxonomic annotation**" window.

2.2. Peptides

In this step, peptide identification and quantification data can be retrieved from an input file in .xlsx format and filtered based on the user's requirements. Structure and column headers of the input file must comply with those of the corresponding template file (*Peptides_input_template.xlsx*, based on the "Peptide Groups" output table generated by *Proteome Discoverer*) provided along with this user guide. If the input file does not contain all the required columns, an error message will be shown.

Click on "**Upload input file**" to select and upload the peptide identification list (including abundance columns). Once the file is uploaded, the window is populated with all the filtering options based on the file content.

Peptides can be filtered based on:

- ***Statistical confidence of peptide identification ("Confidence")***: only peptides with the selected level(s) of confidence (low, medium and/or high) are kept in the output, based on the checkbox(es) checked.
- ***Master Protein description text ("Master Protein Description")***: to select only those peptides which belong to a Master Protein containing a specific text in its description (e.g., a protein name or an organism name), type the text of interest in the textbox (be aware that the filter is case sensitive) and click on "**Add**". Multiple texts can be typed and added sequentially; in this case, the user can choose between two boolean operators, "**And**" and "**Or**", to determine whether all the texts added or only one of them must be present in the string, respectively, so that the peptide passes the filter.
- ***Number/percentage of valid values ("Valid values threshold")***: only peptides with a number of valid values (i.e., non-missing values) greater than or equal to the selected threshold pass the filter. The user can choose between indicating absolute or percentage values by selecting the corresponding option in the drop-down menu. Leaving the default value (0) means that all peptides pass the filter. The total number of samples in the dataset is shown in brackets under the textbox.
- ***Quantification information ("Quan info")***: only peptides belonging to the selected categories are kept in the output (by default, the "**shared**" peptides checkbox is selected, while peptides annotated as "no quan values" or "not reliable" are not selected).
- ***"Marker"***: marker names included in the "Marked as" column of the input file, usually indicating which of the protein database(s) used for identification contained that peptide sequence, are retrieved by the software and shown next to their respective checkboxes; only peptides annotated with the checked marker names are kept in the output.

Furthermore, the user can choose between the following visualization and calculation options:

- ***Show Protein Accessions***: if selected, the "Protein Accessions" column (i.e., the column indicating the Accession number of all the protein entries matching with a peptide, including non-master proteins) is included in the filtered table; this option is available only when this column is present in the input file.

- ***Select normalized abundances***: if selected, normalized abundance values will be reported in the table; this option is available only when normalized abundance values are included in the input file.
- ***Re-normalize abundances after filtering***: if selected, once the chosen filters are applied and the filtered peptide list is obtained, (non-normalized) abundance values measured for a peptide in a sample are divided by the total peptide abundance measured in that sample and multiplied by 10^{10}; this option cannot be chosen while the "select normalized abundance" option is selected.
- ***Replace missing values with 0***: if selected, missing values (empty cells) are replaced by 0; this selection will be applied to all the following output tables.

At the end, the filtered table can be downloaded as a Microsoft Excel file (.xlsx format) by clicking on "**Download filtered table**".

Click on "**Next step**" to go to the "**Taxonomic annotation**" window.

2.3. PSMs

In this step, PSM identification data can be retrieved from an input file in .xlsx format and parsed to obtain PSM counts per sample for each peptide sequence identified. Structure and column headers of the input file must comply with those of the corresponding template file (*PSMs_input_template.xlsx*, based on the "PSMs" output table generated by *Proteome Discoverer*) provided along with this user guide. If the input file does not contain all the required columns, an error message will be shown.

Click on "**Upload input file**" to select and upload the PSM list. Once the file is uploaded, the window is populated with all the filtering options based on the file content.

PSMs can be filtered based on:

- ***Statistical confidence of peptide identification ("Confidence")***: only PSMs with the selected level of confidence (low, medium and/or high) are kept in the list, based on the checkbox(es) checked.
- ***Master Protein description text ("Master Protein Description")***: to select only PSMs that belong to a Master Protein containing a specific text in its description (e.g., a protein name or an organism name), type the text of interest in the textbox (be aware that the filter is case sensitive) and click on "**Add**". Multiple texts can be typed and added sequentially. The user can also choose between two boolean operators, "**And**" and "**Or**", to determine whether all the texts or only one of them need to be present in the string, respectively, so that the PSM passes the filter.
- ***"Marker"***: marker names included in the "Marked as" column of the input file, usually indicating which of the protein database(s) used for identification contained that peptide sequence, are retrieved by the software and shown next to their respective checkboxes; only PSMs annotated with the checked marker names are kept in the output.

As a further option, the user can choose if visualizing missing values as an empty cell or as a zero value (in the latter case, check the "**Replace missing values with 0**" checkbox); this selection will be applied to all the following output tables.

At the end, Meta4P calculates the number of PSMs detected per sample and reports the corresponding values (next to the corresponding peptide sequence) in a tabular format. The filtered table can be downloaded as a Microsoft Excel file (.xlsx format) by clicking on "**Download filtered table**".

Click on "**Next step**" to go to the "**Taxonomic annotation**" window.

3. Taxonomic annotation

In this step, taxonomic annotations can be retrieved from an input file (typically a *Unipept* output, upon conversion in .xlsx format), parsed and included in the previously generated output table (proteins, peptides or PSMs). Structure and column headers of the input file must comply with those of the corresponding template files (*taxonomic_annotation_template-proteins.xlsx* when analyzing proteins or *taxonomic_annotation_template-peptides.xlsx* when analyzing peptides or PSMs) provided along with this user guide. If the input file does not contain all the required columns, an error message will be shown.

Click on "**Upload annotation**" to select and upload the taxonomic annotation file. Columns containing the main taxonomic annotations (LCA, superkingdom, kingdom, phylum, class, order, family, genus, species) will be retrieved and added to the table generated in the previous step.
Once the file is uploaded, the output table can be downloaded as a Microsoft Excel file (.xlsx format) by clicking on "**Download annotated table**".

Click on "**Next step**" to go to the "**Functional annotation**" window.
If no taxonomic annotation data are available, the user can directly skip this step by clicking on "**Skip step**".

4. Functional annotation

In this step, protein functional annotations can be retrieved from an input file (typically an *eggNOG-mapper* .xlsx output), parsed and included in the previously generated output table (proteins, peptides or PSMs). Structure and column headers of the input file must comply with those of the corresponding template file (*functional_annotation_template.xlsx*, based on protein accession numbers and containing comment lines from *eggNOG-mapper*) provided along with this user guide. If the input file does not contain all the required columns, an error message will be shown.

Click on "**Upload annotation**" to select and upload the functional annotation file. Columns containing the main functional levels provided by the input file (i.e., COG category, GO category, EC number, CAZy code, as well as KEGG KO, Pathway, Module and Reaction annotations) will be retrieved and added to the table generated in the previous step.

As an option, the user can choose to retrieve and include in the table (as supplementary columns) the annotation names provided by the KEGG database for all KEGG categories, by clicking on "**Retrieve KEGG name**". As this information is retrieved from the KEGG website, a working internet connection is needed in order that this operation is performed. In case a protein has multiple functional annotations, their names will be separated by a comma as well as their codes (except for the "KEGG Module" annotation names, for which a vertical bar is used to avoid mistakes). In case a peptide is associated to multiple Master Proteins (usually separated by a semicolon), codes and names of functional annotations assigned to different Master Proteins are consistently separated by a semicolon (except for the "KEGG Reaction" annotation names, for which a vertical bar is used to avoid mistakes).

Once the file is uploaded, the output table can be downloaded as a Microsoft Excel file (.xlsx format) by clicking on "**Download annotated table**".

Click on "**Next step**" to go to the "**Data aggregation**" window.
If no functional annotation data are available, the user can directly skip this step by clicking on "**Skip step**".

5. Data aggregation

In this step, abundance data can be aggregated based on taxonomic, functional and/or taxon-specific functional annotations; in other words, the abundances of all proteins/peptides/PSMs sharing the same annotation are summed for each sample.

Taxonomic and functional levels of interest can be selected, if available, by checking their respective checkbox. In addition, taxon-specific functions can be customized by combining a taxonomic level (drop-down menu on the left) with a functional level (drop-down menu on the right) and clicking on "**Add**".
At the end, the list(s) of all annotations (taxa, functions and/or taxon-specific functions) belonging to the selected annotation level(s), together with their aggregated abundance values, can be downloaded as Microsoft Excel file (.xlsx format) by clicking on "**Download tables**".

As an option, the user can choose to retrieve and include in the table(s) (as a supplementary column) the annotation names provided by the KEGG database for all the selected KEGG categories, by checking the "**Retrieve KEGG name**" checkbox.

As a further option, by checking the corresponding checkbox, a supplementary table (for each output table selected) can be generated containing the **feature-related peptide (or protein) counts**, i.e., the number of peptides (or proteins) per sample for which an abundance value was measured (contributing to the summed abundance showed as aggregated value in the main table).

Click on "**Next step**" to go to the "**Rename/reorder sample columns**" window.

6. Rename/reorder sample columns

In this step, the user can customize name and order of sample columns, based on a conversion file.

To create a conversion file ("**Create conversion file**" section), upload one of the output tables generated by Meta4P in the previous steps; in particular, sample column headers must not have been modified in any way by the user and must correspond to the original headers of the Proteome Discoverer output. Based on this input, the software creates a conversion file, i.e., a Microsoft Excel file (.xlsx format) containing a first column with the current list of samples (based on the original input file, one sample per row) and a second column to be completed by the user with the customized sample names once the conversion file has been downloaded. If useful, the sample order in the first column can also be changed and Meta4P will change the column order in the output tables accordingly.

Once the conversion file has been filled in, upload it in the "**Load conversion file**" section.

Then, select which of the output table(s) generated and downloaded in the previous steps need(s) to be subjected to renaming/reordering of sample columns ("**Select file(s) to edit**" section).

Finally, click on "**Rename/reorder sample columns**" to complete the rename/reorder process.