

PeptideIdentificationModule

This pdf file describes and depicts the input method for both CSV and fasta files for this tool. The input method can look complicated but is actually quite easy to create. First, create a main text file that will be used as the input for the -proteinPeptides argument. This main text file should consist of one or more pathways to other text files. These other text files should contain the pathways to the protein-peptides.csv file of each sample. See figure 1 for a more detailed overview. In this example, the 1D50CM.txt is the main text file. This text file contains a file path to a second text file called w_Individual_mzid.txt. The second text file contains the file path to the protein-peptides.csv files. It is possible to add another text file with data to the main text file to process multiple datasets.

When creating the text file containing the file path to the CSV files, please make sure that each file path has the same dataset and sample type name (Otherwise errors will occur!). For this input file the dataset name is 1D50CM and the sample type name is Individual.

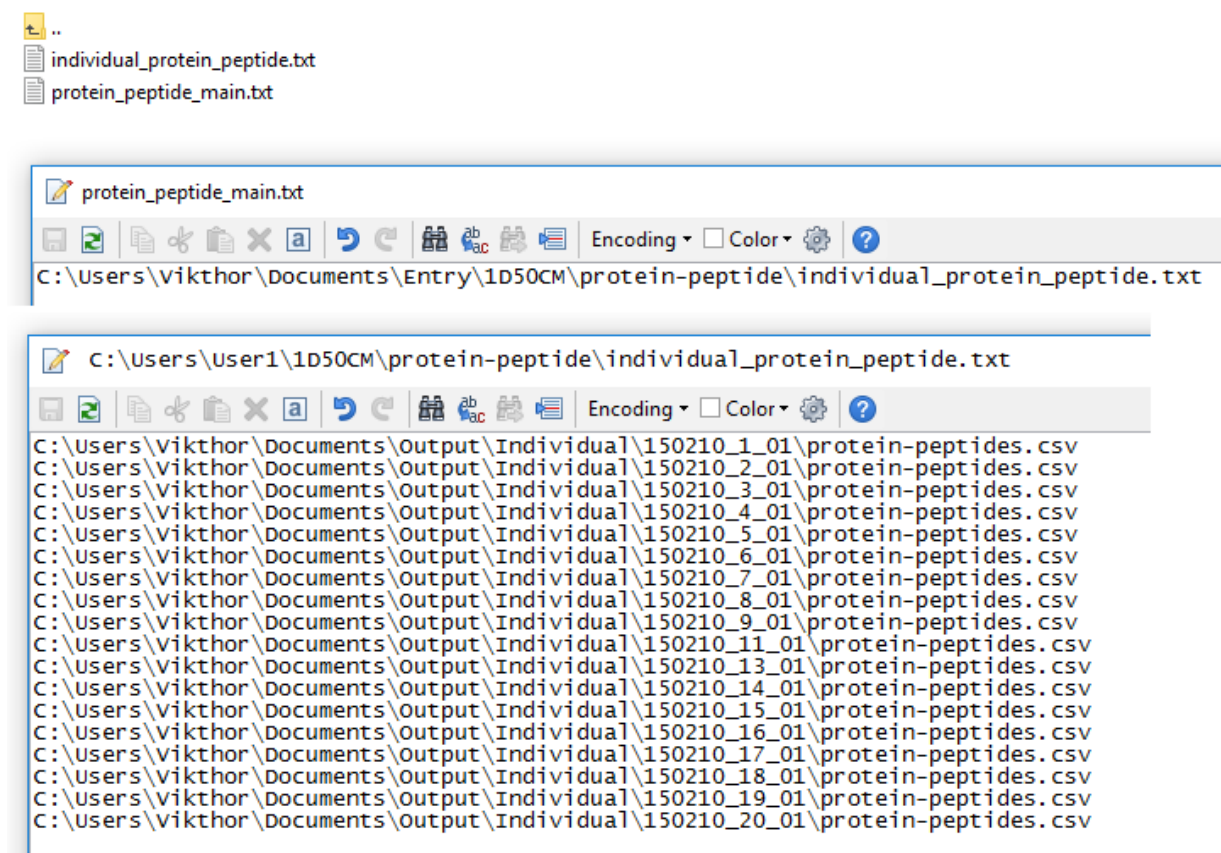
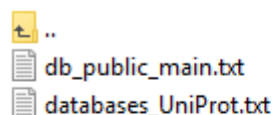


Figure 1. protein-peptides.csv format input.

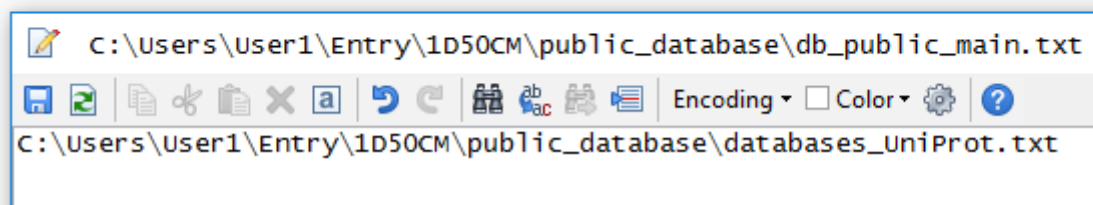
The MzIdentML input is done by using text files that contain paths to the protein-peptides.csv files. The file used for the -proteinPeptides input parameter is the protein_peptide_main.txt file. This text file should contain a pathway to the text file of each set of samples used for this data set. In this case only samples of the individual data set are present. The contents of this file should only consists of pathways to the mzid files of each sample. It is possible to add other entries to the main protein_peptide_main.txt file to analyze multiple sets of samples.

The database text files are provided as input in a similar way. The main file should contain file paths to the text files containing the protein sequence fasta files of a public database (Uniprot, Ensembl etc.). The text file for the reference database argument should contain databases that are specific to the given sample. To add a different public database simply create a new text file and add the database to this text file.

When loading in the databases take into consideration that the public databases are used to match to each sample, while the reference database uses indexing to match peptides from a sample to the given protein database. Please make sure that the order of the reference database fasta files matches the order of the protein-peptides.csv files. This ensures that the right peptide data is matched to the right protein sequence database.

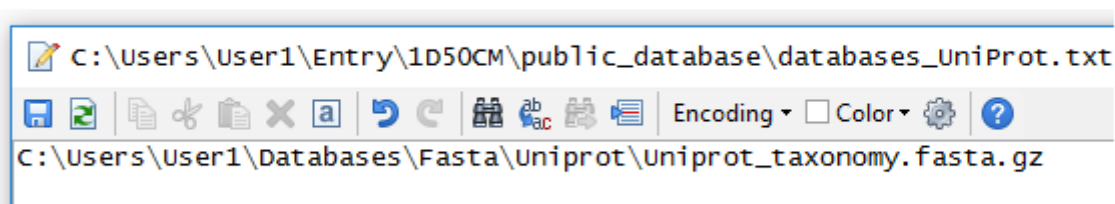


..
db_public_main.txt
databases_UniProt.txt



C:\Users\User1\Entry\1D50CM\public_database\db_public_main.txt

C:\Users\User1\Entry\1D50CM\public_database\databases_UniProt.txt



C:\Users\User1\Entry\1D50CM\public_database\databases_UniProt.txt

C:\Users\User1\Databases\Fasta\Uniprot\Uniprot_taxonomy.fasta.gz

Figure 2. Database input layout.

This picture shows the public databases map which contains the main public database text file. As seen in the picture it only contains.

..
databases_individual.txt
db_reference_main.txt

The image shows two screenshots of text editors. The top screenshot shows a file explorer with two files: 'databases_individual.txt' and 'db_reference_main.txt'. Below it, a text editor window displays the path 'C:\Users\User1\Entry\1D50CM\reference_database\db_reference_main.txt' and another path 'C:\Users\User1\Entry\1D50CM\reference_database\databases_individual.txt'. The bottom screenshot shows a text editor window with a list of database files in a FASTA format, such as 'C:\Users\Vikthor\Documents\Databases\Fasta\Individual\Control_1_06_385_database.fa'.

Figure 3. Database input layout of reference fasta files.

This figure shows the layout of the reference fasta files. As seen it is similar to the public database text file, except that it contains a fasta entry for each protein-peptides.csv entered in the protein-peptides main text file. Please make sure that the sample numbers correspond in order to get the right results.