PROTEOMICS AND IMAGING TOOLS

User Guide

May 2018

Version 1.0.2

Murray Group,

Louisiana State University

Table of Contents

1. Introduction

1.1 Software overview

Proteomics and Imaging Tools (PIT) is a c# application that contains three modules to perform various tasks in a proteomics and/or mass spectrometry imaging (MSI) experiment. The first module of the software package is called PTM Parser and it can process tabulated databases downloadable from http://www.uniprot.org. The result is an output fasta file where specific post-translational modifications – which may alter the primary sequence of a protein – are included as single entries. This allows univocal identification of masses and sequences in shotgun proteomics experiments as well as MSI experiment. The second module is called MW Calculator. This module allows inclusion of the average molecular weight of a protein in a fasta file, adding it within the descriptor line of the protein. In addition, the user can decide to filter proteins based on their molecular weight.

The third module is a called Mass Filter Generator (MFG) and creates a Bruker FlexImaging file containing filters for all the mass value provided with an input file.

The fourth module is called Mass List Analyzer (MLA) and it allows the user to select a folder containing *.csv results files from Bruker flexAnalysis, creating an output file that contains all the mass values in common between the files.

The fifth module is called Peptide DB Search (PDBS) and it allows the user to match peptides to proteins contained in a fasta database. The modules uses the Protein Information Resources (PIR) line command tool to perform the blast search.

1.2 Hardware requirement

PIT has been written for .NET Framework 4.6.1 and has been tested on Windows 10 only. Nevertheless, it may be possible for the software to work on other versions of the .NET Framework.

The architecture of the program requires processing power and disk writing speed to obtain the fastest parsing performance. Slow CPUs and older HDD will affect the processing time. As a reference, a tabulated database of over 36,000 proteins will be processed in about 8 seconds with the first module using a system comprising an Intel Core i7-4510U CPU

@2.00 GHz and a 5400 rpm / 8 MB cache HDD, while a fasta file containing over 56,000 entries will be parsed with the second module in about 35 seconds using the same system. The third module will process peak lists with over 1,000 entries in about one millisecond and the fourth module displays even faster speed.

PTM Parser, MW Calculator and MGF write temporary files in the temporary folder of the computer. Depending on the result of the parsing operation, the file is either moved to the chosen output destination or deleted. Make sure your OS partition has enough memory to host the temporary file.

2. Using the software

2.1 PTM Parser

The PTM Parser module allows the user to select a tabulated file and parse it to create a fasta file that contains single entries for several post-translational modifications. Fig.1 shows the main window with additional description for some of the fields.

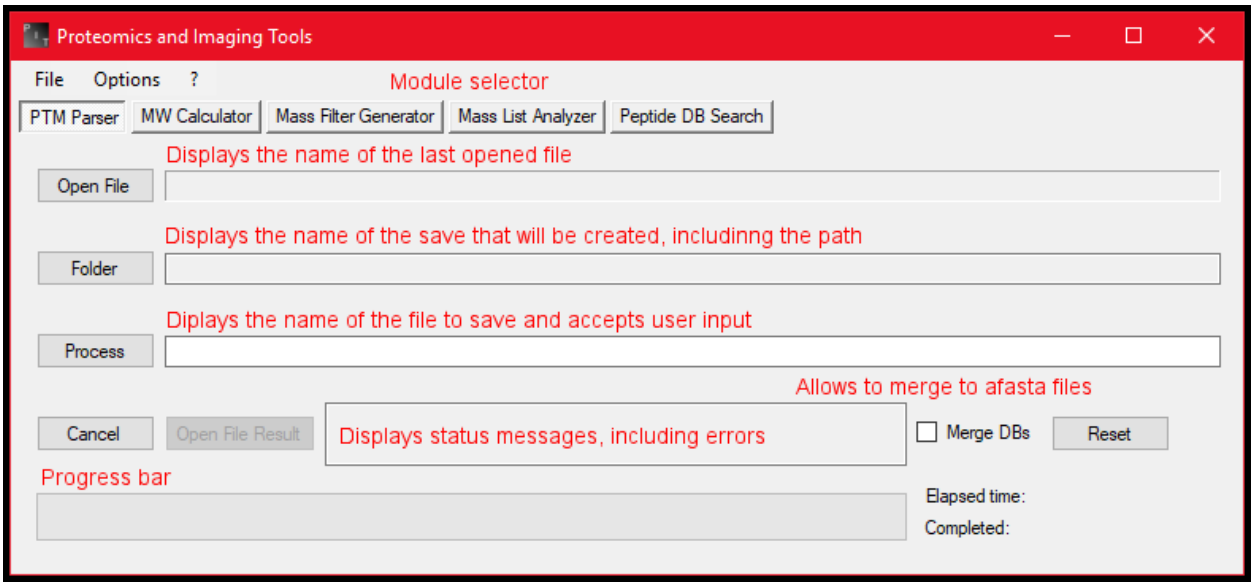

Figure 1: PIT main window with PTM Parser module selected.

*The open file window will search for *.tab files as default, but any tab delimited file is accepted.*

The first step is to obtain a database file from Uniprot. You can find additional info about retrieving entries from Uniprot at: http://www.uniprot.org/help/retrieve_sets. The first line of the file must have headers in the following order:

1. Entry
2. Entry name
3. Protein names
4. Mass
5. Chain
6. Sequence
7. Initiator methionine
8. Peptide
9. Propeptide
10. Signal peptide
11. Transit peptide
12. Status

The name of each section is identical to those on the Uniprot database. When downloading the database file, please make sure that these headers are in the order shown in the list above. You can refer to Fig. 2 for additional help.

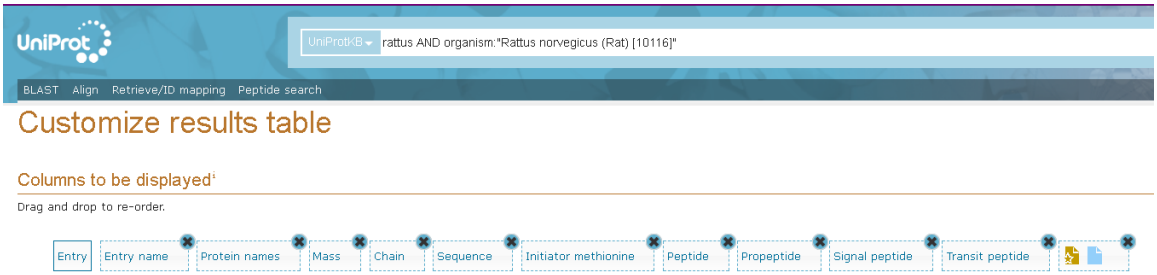

Figure 2: example of customized Uniprot results table. The symbol to at the end of the list represents the element "Status" (#13 in the list in section 2.1).

If the headers are not in the correct number, the software will display an error.

Once you have a tab-delimited file with the right headers, you can press "Open File" to load it into the program. PTM Parser automatically select the input file folder and name as default for the output file. You can change both using the "Folder" button and directly writing into the field next to the "Process" button (Fig, 1). Remember to press the enter key after you write the new file name. If you want to empty all fields, press the "Reset" button.

Once you decided on a file name and a folder, confirm that the output file path in the field next to the "folder" button is correct and press the "Process" button to start the parsing task.

You can cancel the task with the "Cancel" button.

If the output file exists, the software will ask you to change the file name or the file folder.

If you want to merge the result file with an existing .fasta database, check the "Merge DBs" box and the program will prompt for selection of the desired fasta file once it is ready to write the output file.

Once the process is completed, you can check the result file pressing the "Open File Result" button. Each PTM is easily identifiable by the suffix added after the Uniprot identifier. In addition, the number of the first and last amino acid of the original sequence is reported as well. Refer to the example section for additional details.

2.2 MW Calculator

The MW Calculator module allows the user to select a fasta file and add the average molecular weight of each protein entry to the protein descriptor line. Fig.3 shows the main window with additional description for some of the fields.

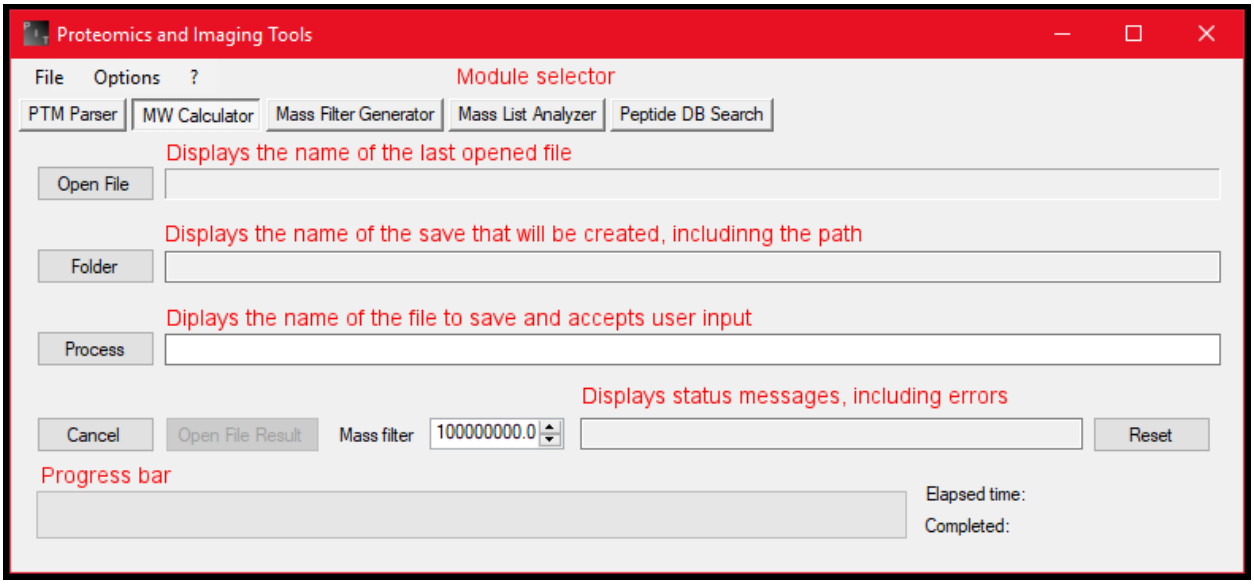

Figure 3: PIT main window with PTM Parser module selected.

*The open file window will search for *.fasta files as default. Any other file will produce an error when processing it.*

After selecting the fasta file, the fields are filled with default values. You can change the output folder and the output file name the same way you change them in the PTM Parser module. You can specify a mass filter: only the proteins below the filter will be added to the output file. The standard filter is set at 10,000,000 Daltons. If you want to empty all fields and reset the mass filter to its default value, press the "Reset" button.

The "Option" menu allows to set the default value of the mass filter to the current value of the control. This will override the default 10,000,000 Da value.

Once you decided on a file name and a folder, confirm that the output file path in the field next to the "folder" button is correct and press the "Process" button to start the parsing task. You can cancel the task with the "Cancel" button

If the output file exists, the software will ask you to change the file name or the file folder.

Once the process is completed, you can check the result file pressing the "Open File Result" button. Refer to the example section for additional details.

2.3 Mass Filter Generator

The MFG module allows the user to select a list file and create a *.mir file containing Bruker flexImaging filters. Fig.4 shows the main window with additional description for some of the fields, while Table 1 provides additional description as well as comparison with the mass filter window in flexImaging.

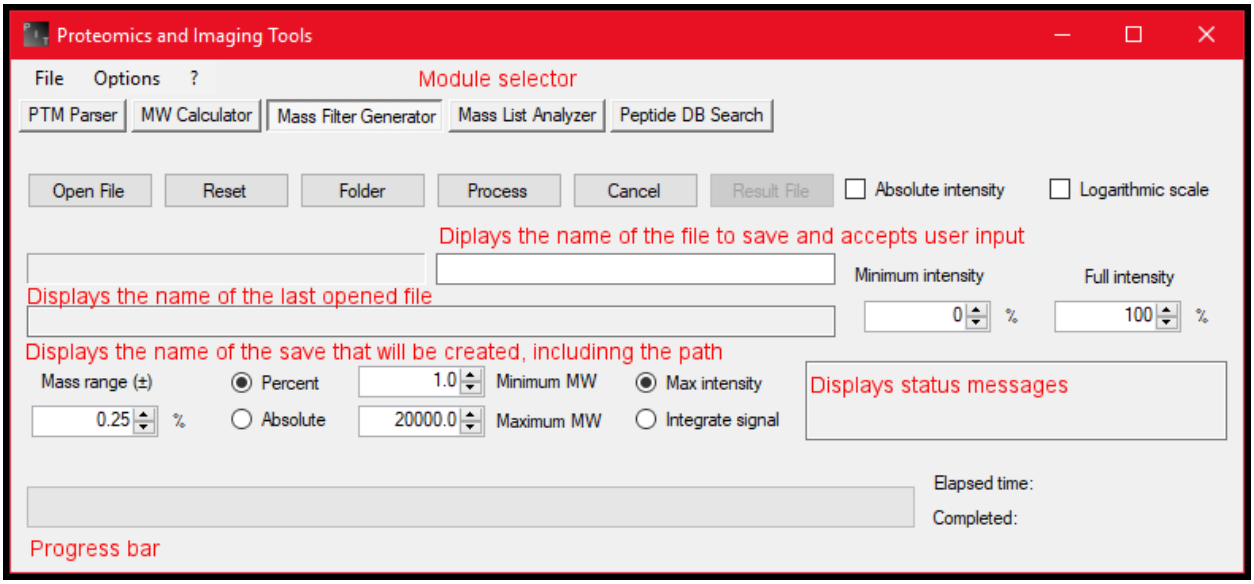

Figure 4: PIT main window with the MFG module selected.

*The open file window will search for *.list files as default, but any file containing one number per line will be accepted.*

Table 1: parameter conversion between PIT and flexImaging

PIT parameter	flexImaging parameter	Description
Absolute intensity	Use Absolute Intensity	If the checkbox is selected, the filter will use the arbitrary unit value without normalization
Logarithmic scale	Logarithmic Scale	If the checkbox is selected, the filter will be displayed applying a logarithmic scale to the intensity value instead of a linear one
Minimum intensity	Minimum Intensity	Any signal below this value will not be considered for image display
Full Intensity	Full Intensity Threshold	Signal higher than this value will be displayed with the highest value of the image intensity scale
Mass range (±)	Mass Range	Determines the mass window to consider for data points extraction.
Percent/Absolute	Range in Percent / Absolute Range	Display the mass range in percentage of the filter mass value or in absolute Da
Minimum/Maximum MW	Not present	Only mass values in within this range will be written to the output file
Max Intensity	Use Maximum Intensity in Range	The value to plot for image generation is the data point with the highest intensity in the mass range
Integrate signal	Integrate Intensity for Mass Range	The whole signal in the mass range is integrated to obtain the value to plot for image generation

After selecting the list file, the fields are filled with default values. You can change the output folder and the output file name the same way you change them with the previous modules. You can save the current values of most field and control using the "Set default values" in Options→Mass Filter Generator. If you want to empty all fields and reset all controls to their default value, press the "Reset" button.

Once you decided on a file name and a folder, confirm that the output file path in the field next to the "Minimum intensity" control is correct and press the "Process" button to start the parsing task. You can cancel the task with the "Cancel" button.

If the output file exists, the software will ask you to change the file name or the file folder.

Once the process is completed, you can check the result file pressing the "Open File Result" button. A flexImaging *.mir files can be opened with any text editor. If you want to load it into an imaging experiment, copy the file into the sequencing directory. Refer the flexImaging user manual for further details. The generated filter are all de-selected.

2.4 Mass List Analyzer

The MLA module allows the user to select a folder containing .csv files originated from Bruker flexAnalysis and create two output csv files with condensed and raw results from comparison of all the input csv files in the selected folder. Fig.5 shows the main window with additional description for some of the fields.

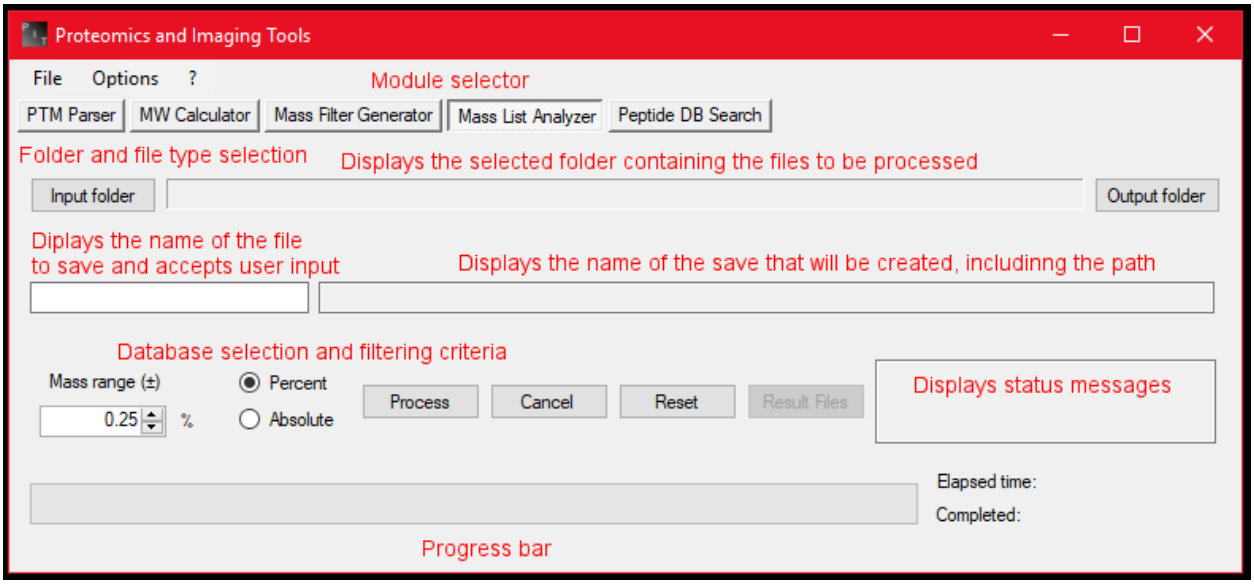

Figure 5: PIT main window with the MLA module selected.

The select folder window does not show any file. Double click a folder to open it and use the button at the bottom of the window to select the folder containing the csv files.

After selecting the input folder, the fields are filled with default values. You can change the output folder using the "Output folder" button on the right, which will update the output path but will not change the input path. The mass range field allows you to specify a window tolerance for peak matching either in percentage or in absolute Daltons. You can set the current value of this field as default clicking on Options→Mass List Analyzer → Set default mass range. If you want to empty all fields and reset all controls to their default value, press the "Reset" button.

Once you select an input folder and an output folder, confirm that the output file path is correct and press the "Process" button to start the parsing task. You can cancel the task with the "Cancel" button.

If the output files exist, the software will ask you to change the file name or the file folder.

Once the process is completed, you can check the result file pressing the "Result Files" button, which will display a drop-down menu that contains a button for each of the two output files. The

button “Open raw data” will open the .csv file that contains all the parsed peaks. If a cell contains more than a value (*e.g.* in the case of a peak mass found in more than one file), each number is separated by the character “/”. The button “Open output file” will open a .csv file where values for each peaks are condensed. Values of peaks with multiple matches are reported as mean while the cell next to them displays the standard deviation. Refer to section 3.4 for a practical example.

If your input folder contains .csv files with format other than the one from export of flexAnalysis results, the program will generate an error.

2.5 Peptide DB Search

The PDBS module allows the user to select a folder containing either CSV output files from ProteinLynx Global Server (PLGS, Waters Corporation) or *.list files of peptide sequences. The modules allows to choose any *.fasta database in the input folder. Preprocessing criteria include filtering of peptides based on minimum # of amino acids per peptide and minimum # of peptides per protein. Figure 6 shows the main window with additional description for some of the fields.

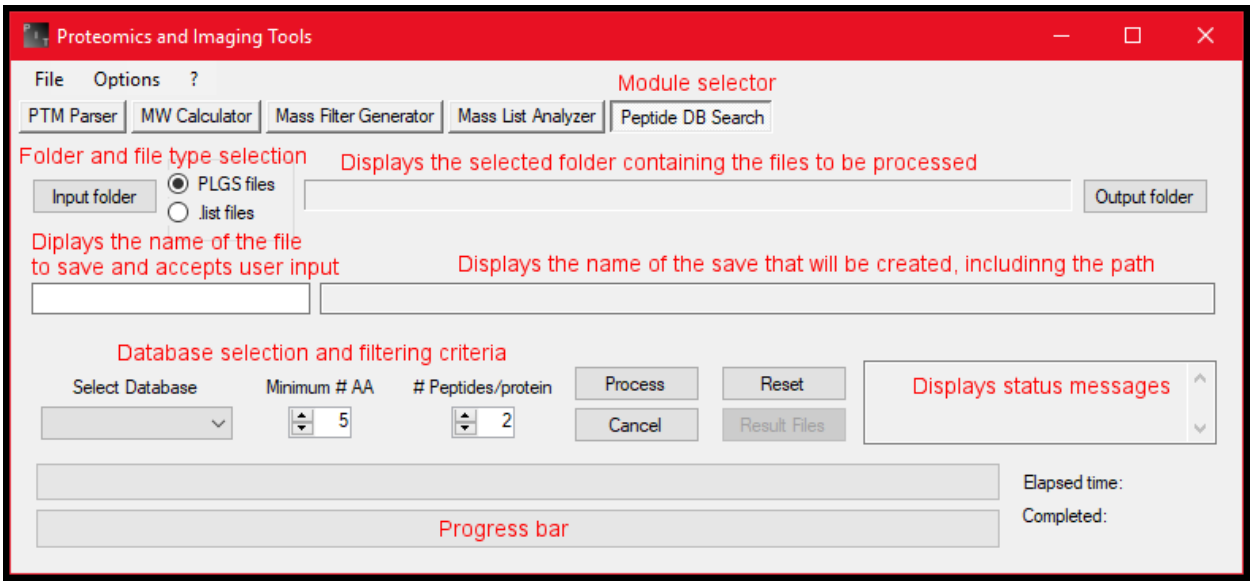

Figure 6: PIT main window with the PDBS module selected.

The input folder window does not show any file. Double click a folder to open it and use the button at the bottom of the window to select the folder containing the csv files.

After selecting the input folder, the fields are filled with default values. You can change the output folder using the "Output folder" button on the right, which will update the output path but will not change the input path. Depending on the file type selection, PDBS will parse either CSV file or *.list files. Only PLGS output CSV files will be accepted and any other CSV file will result in an error.

If you want to search peptide lists from other sources, use .list files (i.e. a text file containing one-letter code sequences – one sequence per line – and saved with the ".list" extension).

Any *.fasta file present in the output will be displayed in the dropdown menu for the DB selection. Once the process is completed, you can check the result file pressing the "Result Files" button

The software will show a csv file with a summary of the search parameter used and a list of protein matches including the number of peptides matched, the extended name of the protein, their database ID, their mass and the sequence of each matching peptide.

*If the entries in your database do not have a specified mass, all the field will be shifted to the left. You can add masses to any *.fasta database using PIT module 2, MW Calculator.*

After each search, a log file will generated with the content of the status window.

3. Examples

In the installation directory of PIT, you can find a folder containing example files, which you can use to test the software.

3.1 Using the PTM Parser module

The PTM Parser module can be tested using the example file "Example tab delimited database.tab". This file contains 166 protein entries and the output file will produce a fasta file with 238 entries. Each entry will have the name of the PTM modification and the starting and ending position of the modification in the original sequence embedded in the Uniprot identifier. The entry Q07008 has a chain (from 1744 to 2531) and a signal peptide (from 1 to 18). The 2 identifiers in the fasta files are Q07008_CHAIN_1744_2531 and Q07008_SIGNAL_1_18.

3.2 Using the MW Calculator module

The example file to test this module is "Example fasta file.fasta", which is actually the output file of the previous module example. If you do not modify the mass filter, the module will add the MW to each of the 238 entries of the fasta file. For example, Q07008_CHAIN_1744_2531 has an average molecular weight of 84788.2 Da while Q07008_SIGNAL_1_18 has a molecular weight of 1919.5 Da.

3.3 Using the Mass Filter Generator module

The example for this module is "Example peaklist.list", which contains 238 mass values. With the initial default parameters, processing this file will generate a *.mir filter list containing a filter for each of the 238 mass values.

3.4 Using Mass List Analyzer

The example for this module is the folder "Example MLA folder", which contains four csv files with 32 peaks each. The mass of each entry has been modified by subtraction of 1 Da, addition of 1 Da or addition of 0.3 Da to simulate similar spectra. Processing of these files with a mass range of 0.5 Da will produce an output file with 93 entries, 31 of which with 2 occurrences because two

of the four input files had peak mass within the specified tolerance. The corresponding raw data file will display two values in each cells for these 31 peaks.

3.5 Using Peptide DB Search

The example for this module is the folder “Example PDBS”, which contains two database with 2 proteins (one with and one without protein masses) and a peptide list file with 31 entries. The software will create the database for the command line tool and then search the peptide list against it. Using a minimum #AA equal to 5 and two peptides per protein as filtering criteria, the result file will have two proteins, one with 3 matches and one with 24 matches. You can repeat the search using the database without the protein masses to see how the result file changes.

4. Contacts

If you want to know more about PIT, report a bug, request a modification or a feature implementation, please contact Dr. Fabrizio Donnarumma at fabrizio@lsu.edu.