**WebSearch and FileProcessing User Guide**

*Prerequisites*

WebSearch and FileProcessing are Python programs that require Python to be installed on your computer. Python version 2.7 minimum is required.

The programs can be run through Terminal (Mac OSX/Linux) and Command line operation (Windows). Refer to the ‘Installing Python’ guide on GitHub (https://github.com/cjbrasher/LipidFinder) to ensure you have all the required libraries and dependencies installed for Python.

*File installation*

Download the program files from the GitHub repository and save them all to the same folder.

|  |  |  |
| --- | --- | --- |
| **File** | **Purpose** | **Required by** |
| getFile.py | module to read in input file name | WebSearch |
| lipidhomeWebSearch.py | module to search LipidHome database | WebSearch |
| lipidMapsWebSearch.py | module to search LIPID MAPS database | WebSearch |
| hmdbWebSearch.py | module to search HMDB database | WebSearch |
| webSearch.py | module to run the WebSearch program | WebSearch |
| hmdb\_parsed\_file.csv | downloaded version of HMDB database used in processing | WebSearch |
| websearch\_parameters.csv | parameters that are used by WebSearch program | WebSearch |
| fileProcessing.py | module to run the FileProcessing program | FileProcessing |
| selectData.py | module to filter and clean up data | FileProcessing |
| processingSteps.py | module to read in input file names and merge files | FileProcessing |
| categories\_map.csv | file of lipid category standardisation | FileProcessing |

Note: .pyc files will be automatically generated by python once you run a program for the first time.

**WebSearch**

*Input file format*

WebSearch is carried out after PeakFilter (and Amalgamator) has processed the file(s).

The input file to WebSearch is the output file from PeakFilter. It can be a modified file, but it must have the following columns (order is not important, but spelling and case is): MZ, Time, Polarity. It must be saved to the same folder as the program files (above).

*Parameter settings*

A number of parameters that are used by the WebSearch program can be changed in the websearch\_parameters.csv file. The first three are the *m/z* tolerance values that are used in the search query to each of the databases. The next parameter is the delta ppm value that is used as a cutoff value during the FileProcessing execution – any values outside of this range are automatically deleted from the data. It is calculated as follows:

The final parameter is for the search criteria: whether to search for computational-only data (COM), curated-only data (CUR) or both (ALL). NOTE: the only database that allows a restricted search is LIPID MAPS.

* Selecting CUR means that only the LIPID MAPS database is searched and only for curated data.
* Selecting COM means LIPID MAPS is only searched for computational data and HMDB and LipidHome are searched without any restriction, as this option does not apply to them.
* Selecting ALL means that LIPID MAPS is searched for both computational and

curated data and HMDB and LipidHome are searched without any restriction, as this option does not apply to them.

*Starting WebSearch*

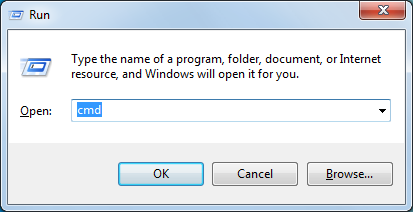
*Option 1: Terminal on Mac OSX or Linux*

1. Open Terminal
2. Type *cd path\_to\_folder* to go to the folder containing the program files and the input file(s) (change '*path\_to\_folder*'to the name of your folder). To run the program type: *python webSearch.py* at the prompt and press *Enter*. Then enter the name of the input file (e.g. *testMasses.csv*) and press *Enter*.

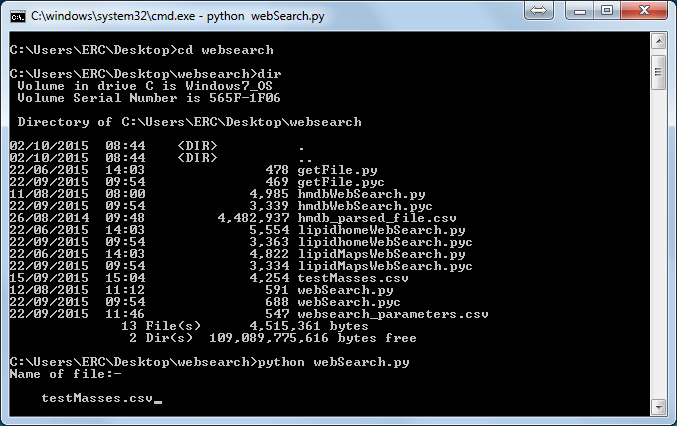


*Option 2: Command line on Windows*

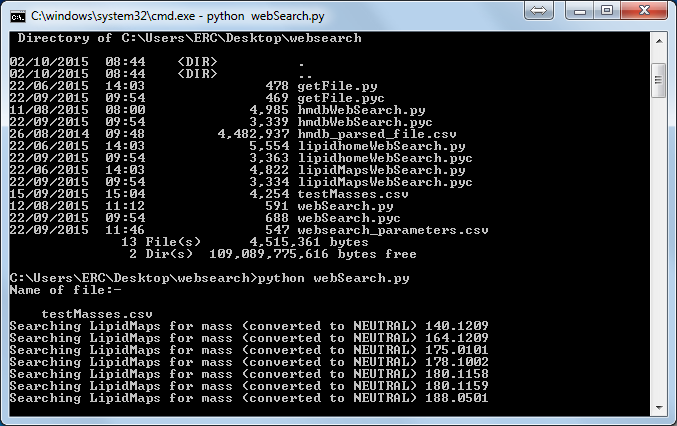
1. Press Win+R to launch the Command Prompt to use the Run window. Type cmd or cmd.exe and press Enter or click/tap OK.



1. Type *cd path\_to\_folder* to go to the folder containing the program files and the input file(s) (change '*path\_to\_folder*'to the name of your folder). To run the program type: *python webSearch.py* at the prompt and press *Enter*. Then enter the name of the input file (e.g. *testMasses.csv*) and press *Enter*.



The program will search for the masses listed in the input file on each of the website databases. You will see the masses scroll by on the screen, as the program, searches.



If there is a problem connecting to a particular website, then this database will be skipped and any results up to that point will be saved. It can take a long time for this program to finish, but you can follow the progress on screen as each *m/z* value is searched. Once it has finished running there will be 2 output files from each database search; one containing the search results and one listing the masses that were not found in that database. The files will be date and time stamped.

**FileProcessing**

*Input file(s) format*

FileProcessing is carried out after WebSearch. The inputs to FileProcessing are the (result) output files from WebSearch. The files should not be altered in any way; otherwise the program will not be able to read them correctly. This program also requires the original file that you used with the WebSearch program.

The program will merge the results files from the WebSearch into one file. It then calculates the delta ppm value and removes any records when this value is outside of the cutoff specified in the websearch\_parameters.csv file. Next, the program standardizes the category names based on the mappings in the categories\_map.csv file. Adduct names are also standardized as in table below. All other adducts are removed.

|  |  |
| --- | --- |
| **Original adduct ion** | **Standardized name** |
| M-H | [M-H]- |
| M+H | [M+H]+ |
| M+Na | [M+Na]+ |
| M+NH4 | [M+NH4]+ |
| M+CH3COO | [M+CH3COO]- |

The merged file is then saved as output (and date/time stamped). Additionally, 3 further files are generated:

*Category\_counts\_date-timestamp.csv* is generated from the merged file. Unlike the merged file, it has just one result per *m/z*/retention time combination. This is based on the CATEGORY column – whichever category has the highest count for a particular *m/z*/retention time combination, this is selected as the overall category for that *m/z*/retention time combination and the CATEGORY column is updated to reflect this. The count column lists the number of this category that were found for this *m/z*/retention time combination. This file should have exactly the same number of rows as the original search file that was used for the WebSearch.

*DB\_mass\_counts\_data-timestamp.csv* lists each of the *m/z* values and how many were found in each of the databases that were searched. Only *m/z* values that had search results in at least one database are included.

*DB\_mass\_presence\_absence\_data-timestamp.csv* lists each of the *m/z* values and whether they were found in each of the databases that were searched. 1 indicates the *m/z* value was found, 0 indicates it wasn't. Only *m/z* values that had search results in at least one database are included.

*Starting FileProcessing*

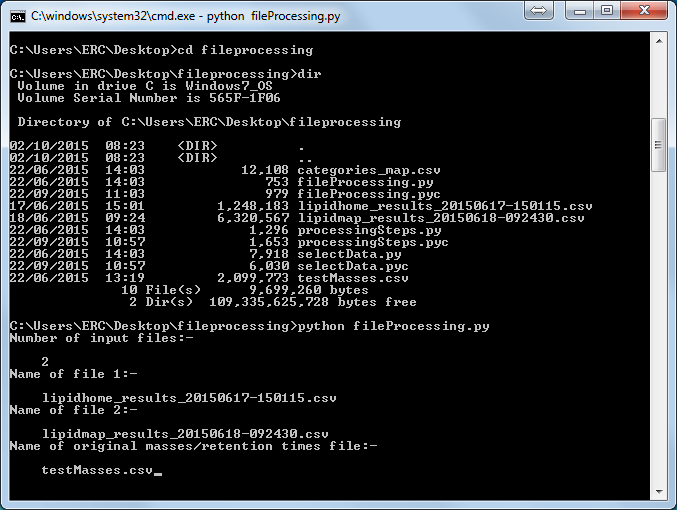
*Option 1: Terminal on Mac OSX or Linux*

1. Open Terminal.app.
2. Type *cd path\_to\_folder* to go to the folder containing the program files and the input file(s) (change '*path\_to\_folder*'to the name of your folder). To run the program type: *python fileProcessing.py* at the prompt and press *Enter*. You will asked to enter the number of input files to be processing: type in the number and press *Enter*. Then enter the name of each results file from WebSearch when prompted (e.g. *lipidhome\_results\_20150617-150115.csv*) and press *Enter*. You will also be asked to enter the name of the original file that you used with the WebSearch program (e.g. *testMasses.csv*).



*Option 2: Command line on Windows*

1. Press Win+R to launch the Command Prompt to use the Run window, as before. Type cmd or cmd.exe and press Enter or click/tap OK.
2. Type *cd path\_to\_folder* to go to the folder containing the program files and the input file(s) (change '*path\_to\_folder*'to the name of your folder). To run the program type: *python fileProcessing.py* at the prompt and press *Enter*. Then enter the name of each results file from WebSearch when prompted (e.g. *lipidhome\_results\_20150617-150115.csv*) and press *Enter*. You will also be asked to enter the name of the original file that you used with the WebSearch program (e.g. *testMasses.csv*).



This is a very quick program (seconds, rather than minutes to finish). There are 4 output files generated and they will be date and time stamped.