MassQLab

An Implementation of MassQL for Streamlining Common Mass Spectrometry Analysis Workflows

SpectraSpectra applies a series of queries (written in the language of MassQL) to a directory containing mass spectrometry data in mzML format. Results are tabulated and saved as images and xlsx documents.

MassQL: https://github.com/mwang87/MassQueryLanguage

Usage

1. Launch MassQLab
    - In root MassQLab directory, run "MassQLab.bat"
        - OR in MassQLab/MassQLab directory, run "MassQLab.exe"
    - Wait for GUI window to open. This may take additional time on first launch
2. Populate User Configuration fields in GUI
    - Use "Browse Directory" button to open a file browser and select a directory containing mzML files to be analyzed
    - Use "Browse File" button beside queryfile to open a file browser and select a MassQL query file (see queryfile)
    - Leave other settings alone for typical usage
3. Click "Run" to apply queries in queryfile to files in data_directory
    - Analysis progress will be displayed in console of GUI
    - Console will state "Run Complete" when finished
4. Open results in "<data_directory>/MassQLab_Output/<timestamp>"

Query File

- The query file (queryfile) is used to define MassQL queries and a name for the query.
    - MassQLab accepts the queryfile in JSON, CSV, or XLSX formats.
- An example JSON format and CSV format queryfile is included in the root MassQL directory.
    - MassQL_Queries.json
    - MassQL_Queries.csv
- A valid query in the queryfile requires a "name" and "query"
- "name" is what will be used to uniquely identify the query in the results files and images
    - The "name" should be unique from other query names in the query class
        - ie. duplication is allowed if name is used for an MS1 and MS2 query
    - Note: There may be an issue with very long query names
- "query" contains the MassQL query itself
    - queries fall into two categories, MS1 or MS2, that are handled independently in the workflow
    - see "MassQL Queries" for detailed description
- See "Query File Advanced" for additional parameters

MassQL Queries

See https://mwang87.github.io/MassQueryLanguage_Documentation/ for full documentation

1. MS1 queries will return a dataframe of total intensity of MS1 scan vs retention time for each MS1 scan that matches query
    - Peak will be fit with a gaussian and area will be determined from gaussian.
    - Peak must meet detection criteria to be considered valid
        - peak prominence > Intensity_Max / 10
        - peak height > Intensity_Average x 1.1
        - peak center > (RTMIN - (RTMAX - RTMIN))
        - peak center < (RTMAX + (RTMAX - RTMIN))
        - fwhm < (RTMAX - RTMIN)
2. MS2 queries will return a dataframe of total intensity of MS2 scan vs retention time for each MS2 scan that matches query
    - Downstream analysis will split MS2 query results for each file based on any collision energy grouping as defined in scan metadata
        - ie. HCD and CID will be split
        - ie. Collision energy 20 vs 30 vs 40 will be split
        - An MS2 group will be each combination, ie. HCD_20, CID_30, etc.
    - When more than one valid scan is returned from the MassQL query for an MS2 group, the highest intensity scan will be used for downstream analysis

MS1 query examples

- Get MS1 scans matching m/z of 207.1418 with tolerance of 2.5 ppm with retention time between 1.0 and 1.2 minutes and filter for 207.1418 peak intensity
    - `QUERY scaninfo(MS1DATA) FILTER MS1MZ=207.1418:TOLERANCEPPM=2.5 AND RTMIN=1.0 AND RTMAX=1.2`
- Get MS1 scans where a MS2 with product ion is present
    - `QUERY scaninfo(MS1DATA) WHERE MS2PROD=226.18`

MS2 query examples

- Get MS2 scans with a precursor ion matching m/z of 429.3765 and with retention time between 9.0 and 9.5 minutes and return total intensity of each scan
    - `QUERY scaninfo(MS2DATA) WHERE MS2PREC=429.3765:TOLERANCEPPM=2.5 AND RTMIN=9.0 AND RTMAX=9.5`
- Get MS2 scans with a precursor ion matching m/z of 429.3765 and with retention time between 9.0 and 9.5 minutes and return the intensity of the peak with m/z 85.0281
    - `QUERY scaninfo(MS2DATA) WHERE MS2PREC=429.3765:TOLERANCEPPM=2.5 AND RTMIN=9.0 AND RTMAX=9.5 FILTER MS2PROD=85.0281:TOLERANCEPPM=10`

- Get MS2 scans where a product ion and neutral loss is present
    - `QUERY scaninfo(MS2DATA) WHERE MS2NL=176.0321 AND MS2PROD=85.02915`
- Get MS2 scans where a product ion matches arithmetic
    - `QUERY scaninfo(MS2DATA) WHERE MS2PROD=144+formula(CH2)`

Results

Results can be found in MassQLab_Output directory located within the defined data_directory. MassQLab_Output directory contains timestamped subdirectories.

- ms1_raw_df.csv: Raw output returned from applying MS1 MassQL queries and merged with query metadata
- ms1_analysis_df.csv: Output after peak fitting analysis of ms1_raw_df
- ms1_RT_analysis_df.csv: Statistical analysis of MS1 peak center retention time relative to retention defined in query
- ms1_traces.pdf: Document containing plots showing gaussian fit of each query applied to each file
- ms1_summary_traces.pdf: Document containing plots of all peaks for each query
- ms1_summary_areas.pdf: Document containing plots of areas of all peaks for each query
- ms1_summary_areas_inverse.pdf: Document containing plots of areas of all peaks for each file
- ms2_raw_df.csv: Raw output returned from applying MS2 MassQL queries and merged with query metadata
- ms2_analysis_df.csv: Output after peak picking analysis of ms2_raw_df
- ms2_plots.pdf: Document containing plots of intensity of each scan returned for each file and query
    - Note: lines are shown between points that share the same MS1 scan
    - Note: highest intensity scan for each condition is used for downstream analysis
- ms2_summary_plots.pdf: Document containing plots of highest intensity scan for each query
- ms2_cluster_plots_group.pdf: Document containing summary plots of each group of queries as defined by "group" parameter in queryfile if present

Query File Advanced

This section describes parameters that may be used in the queryfile in addition to "name" and "query"

Any additional parameters in queryfile will also be carried over into excel/csv outputs for ad-hoc usage

- "abundance":
    - an expected or anticipated abundance of the peak area (ms1) or intensity (MS2)
    - abundance validity will be measured with a 10% (ms1) or 20% (ms2) threshold of this value
        - abundances not within this threshold will be flagged as invalid and will not be considered in downstream analysis
        - customizing threshold will be added in future release
- "group":
    - A designation for queries that have some relationship
    - Ase the same argument for the "group" parameter in multiple queries to associate the queries with each other
    - Use case 1: Assign acyl-carnitine query and each acyl-carnite subfragment query a "

group" argument of "acyl-carnitine"

- In "ms2_cluster_plots_group" result, each group will be consolidated and visualized together

- Use case 2: Assign all phosphatidylcholine lipids a "group" argument of "PC"
  - Function will be expanded in future release to automate quantifying total group (ie. PC) content

- "KEGG":
  - Argument does not currently carry any special meaning, but this and any other parameter in the query file will be carried over into excel/csv outputs
  - The query file is a good place to define unique compound identifiers for ad-hoc usage
    - metabolites: KEGG, InChI string/key, CASRN, SMILES, IUPAC, PubChem ID
    - proteins: PDB ID, Gene name, UniProt, Ensembl, RefSeq

MSConvert

mzML files can be created from raw files on the fly if MSConvert (part of ProteoWizard software) is installed separately

https://proteowizard.sourceforge.io/download.html

You will need to identify the path to "msconvert.exe" and select it in the "msconvertexe" field in the MassQLab GUI window. You will also need to check the "convert_raw" button.

On my system, the path to mscovert.exe looks like this:
`"C:\Users\username\AppData\Local\Apps\ProteoWizard 3.0.23118.b2ed96f 64-bit\msconvert.exe"`