MASS-SUITE

Project Sponsor: Edward Kolodziej (University of Washington, Center of Urban water)
Project Advisor: Edward Kolodziej, David Beck (University of Washington)
Project Team: Ximin Hu (Civil & Environmental Engineering), Derek Mar (Materials Science & Engineering), Nozomi Suzuki (Materials Science & Engineering), Bowei Zhang (Materials Science & Engineering)

Project Background:

Mass-suite is a python based open source package that designed to utilize High Resolution Mass Spectrometry (HRMS) data for water quality assessment. The analysis of HRMS data for water quality assessment is still in its infancy, with many basic aspects of data reduction, analysis, and interpretation still lightly developed. Here, our package will allow users flexible and various options to process the HRMS data: from basic functions to advanced data analysis, such as dilution rate prediction and source tracking studies that are not currently covered by current software. Furthermore, mass-suite is developed in a modularized concept so that the user can use different combinations of parts of our code to accomplish their tasks. By providing this package, we hope to open a new space for HRMS data analysis, resulting in more rapid and detailed research in this area.

Overview:

Peak picking & alignment

- Import data from .mzml file
- Detect and integrate peaks, assessing by trained model
- Alignment across different samples

HRMS data analysis – clustering &modeling

- Clustering of different chemical features based on machine learning algorithms
- Noise removal, cluster labeling
- Modeling & prediction based on dilution series samples
- Quick source tracking tool

Visualization & database searching

- Basic plots for HRMS data
- Interactive plot upon user selection
- Online database search for advanced analysis

Key strengths:

- Prior to modeling, the signals from MS are sorted, and aligned for data cleaning and clustering
- Using the clustered data, models can be created to accomplish the goal of source and dilution tracking
- Users have flexible options through out the whole workflow
- Most of the analysis functions equipped with a result report for validation

Example Output:

Peak picking & alignment:

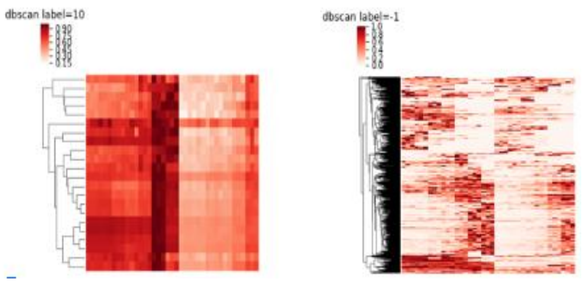
- Filters out peaks that arise from noise
- Align similar RT/mz values to create cohesive dataset

	Average m/z	Average RT (min)	Average sn	Average score	20181114_CoulterCreek_2	20181114_Crescent-Creek-May_2	20181114_SR518-Cal_1000mL_dil-to-160mL_3
0	100.111603	5.590000	inf	0.10	745722.56250	0.0000	0.000000
1	100.111671	5.470000	inf	0.40	0.00000	658076.8375	0.000000
2	100.111801	5.340000	1.649	0.60	0.00000	0.0000	136595.515625
3	107.070297	13.760000	inf	1.00	0.00000	0.0000	0.000000
4	114.091904	3.530000	inf	0.60	0.00000	0.0000	0.000000
...	...	...	...	...	...	...	...
5379	1350.883789	19.264000	inf	0.48	407645.21875	0.0000	396962.562500
5380	1350.884521	19.382999	inf	0.60	411676.21875	0.0000	0.000000
5381	1392.412842	20.740000	1.070	0.60	154861.71875	0.0000	0.000000
5382	1393.415649	20.700001	1.062	0.60	0.00000	245449.8125	0.000000
5383	1393.416748	20.799999	1.049	1.00	0.00000	0.0000	0.000000

```
Please input the mzml file path:D:/UW/mzmltest

Please input the noise threshold for ms1 spectrum:5000

please define if enable the peak score(Y/N):N
Reading mzml files...
100%|██████████| 1/1 [00:03<00:00,  3.93s/it]
['ex_1.mzML']
Batch read finished!
Processing peak list...
Processing 1 out of  1 file
Generating mz list...
Finding peaks...
100%|██████████| 2819/2819 [03:58<00:00, 11.84it/s]
Peak processing finished!
Dataframe created!

please define the rt error:0.5

please define the mz error:0.015

export path:D:/UW/mzmltest/

export name:test1.csv
Process completed!
Initial reference built
Alignment beginning..
100%|██████████| 3012/3012 [00:08<00:00, 338.20it/s]
```

Clustering:

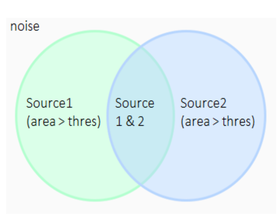
- Different chemicals' 'intensity' behaves differently across dilutions
- We need to cluster chemicals into different groups to predict the dilution levels
- We can see resulting clusters in the images above

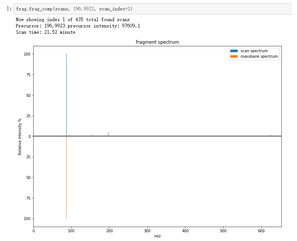


Dilution Prediction and Source Tracking:

- Once the chemicals are clustered, the dilution can be predicted by training a model based on example data
- As shown in the workflow above, users can choose a best fitting model
- Source identification is currently simple and will be later expanded to consider various modeling techniques
- And more..

...	Coulter Cv	Crescent	Crescent Cv	Miller	Miller Cv	Swan	Swan Cv	SR520-Cal-in-DI_1000mL	SR520-Cal-in-DI_1000mL Cv	source
...	0.209046	3615.166667	0.253152	42648.000000	0.714844	3911.333333	0.330629	27219.904762	1.114922	NA
...	0.249619	2882.000000	0.513198	163117.333333	0.139863	1728.833333	0.555111	10014.333333	1.284517	Miller
...	0.183676	3561.500000	0.193846	170958.333333	0.124775	1735.000000	0.332350	20995.619048	1.277956	Miller
...	1.270866	860.000000	0.480707	88572.333333	0.546327	892.166667	0.589734	358528.095238	1.236113	Miller,SR520-Cal-in-DI_1000mL
...	0.000000	64.166667	1.119582	17132.333333	1.001384	224.333333	0.480879	9811.190476	1.095273	NA





Mass-Suite Features and Advantages:

- **General:**
  - First package on python which provide full workflow from data import to alignment result
  - Light memory usage make it possible to be run on personal laptops
  - Remote data process option
  - Modularized setup – always possible for more
  - Open source – free to go!
- **Peak picking:**
  - Pymzml import enables faster speed
  - Peak score to enhance data quality
- **Alignment:**
  - High efficiency
  - Flexible settings for users
- **Data analysis:**
  - Visualization result for assess the analysis quality
  - Open end algorithm options to fit different needs
  - Handy function to boost up efficiency

Future Developments:

- Validation of workflow using more data
- Understanding chemical commonalities behind clusters
- GUI and/or webtools for supporting non-technical users
- Dynamically adapt and update program to fill growing needs

Contact Information:

github:
https://github.com/XiminHu/mass-suite
email:
xhu66@uw.edu

Acknowledgements:
A big shout to Professor Edward Kolodziej, Professor David Beck, Katherine Peter, and the Center of Urban Water for all their help during the development of this package.

Data Reference: Application of Nontarget High Resolution Mass Spectrometry Data to Quantitative Source Apportionment, Katherine T. Peter, Christopher Wu, Zhenyu Tian, and Edward P. Kolodziej, *Environmental Science & Technology* **2019** *53* (21), 12257-12268, DOI: 10.1021/acs.est.9b04481