Development of informatics tools to store and analyse very large sets of spectral metadata from LC-MS metabolomics experiments

Master 2 project

Starting date: February 2024

Length of internship: 6 months

Localization: BIOPI unit of the UMRt BioEcoAgro (team 5), UFR Sciences, UPJV, Amiens

Contact: Rebecca Dauwe (rebecca.dauwe@u-picardie.fr)

Context : Plant tissues contain thousands of distinct compounds and, today, the vast majority of these plant metabolites remain unknown. Untargeted metabolomics experiments usually rely on LC/MS-derived data, and metabolite structures are retrieved by matching fragmentation spectra with spectral databases or by machine learning techniques using molecular structure fingerprints and molecular structure databases. However, most of the profiled plant metabolites are not present in any spectral or molecular structure database, and the elucidation of their structure relies then on the interpretation of their fragmentation spectra based on chemical principles. Clearly, structural elucidation benefits from including as much different types of mass spectral data as possible. For instance, the **complementary information of positive and negative ionization mode** spectra allows a distinction between charge-driven and charge-remote fragmentations. In addition, the spectral information gained from commonly used **Quadrupole-Time of flight (QTOF) and Ion Trap (IT) instruments is complementary**. Whereas QTOF-based MS/MS spectra offer a higher mass accuracy of the product ions and display low-mass product ions, IT-based MS^n spectra allow the relationships between the product ions to be delineated. **We refer to the spectra from complementary instruments and generated with different ionization modes, as spectral metadata**. The main bottleneck for the **exploitation of spectral metadata in machine learning algorithms** for structural elucidation, is the availability of a database, archiving both the spectral metadata of a large training set of identified or partially characterized compounds, and the spectral metadata of the unknown compounds. In this thesis project, we develop tools to construct, expand and exploit an in-house database for mass spectral metadata.

Objective : The objective of this internship will be to create an R package that can (1) import all relevant fragmentation data from metabolomics experiments, generated on different types of instruments, into an SQL database, (2) connect the spectra that belong to a same compound, and (3) analyse them simultaneously.

Tasks : To achieve this goal, the student will carry out the following steps:

- Design the database architecture in close collaboration with researchers from bioinformatics, biology, and analytical chemistry fields.
- Develop the scripts and workflow for importing the spectra of interest from raw LC-MS data into the database. The approach will be based on an existing in-house Perl script. The newly developed R scripts will be using tools and infrastructure for storing and handling mass spectrometry data that were recently developed in R (Rainer et al., 2022).
- Adapt the existing tools to aid in structural elucidation, today stored in our in-house R package RDynLib, to the newly developed infrastructure. To date, the RDynLib package includes tools to: connect and plot different spectra that are generated for the same compound but in independent

chromatograms, calculate chemical formulae for product ions and neutral losses, find complementary ions in a fragmentation spectrum, find isomers, annotate adducts, fragments and isotopes in the MS1 spectrum, search for spectral motifs, or create metabolic networks (Desmet et al., 2021).

Consortium: The Master project will take place at the UFR Sciences, UPJV, Amiens, in the BIOPI unit of the UMRt BioEcoAgro (team 5), in close collaboration with the Computational Metabolomics team at Eurac, Bolzano, Italy, and with the Computational Biology and Bioinformatics team (CBIO) at Institut De Duve, UCLouvain, Belgium.

Ideal candidate profile :
Background in bioinformatics/computer science/data analytics or in life science
A strong knowledge of R, and a strong scientific and technical curiosity are required.
Knowledge or interest in metabolism, biochemistry and mass spectrometry are a plus.

How to apply?

Please, send us by email:

• A cover letter explaining your interest in this internship (1-page max)

• A CV (max. 2 pages)

References:

Desmet S, Saeys Y, Verstaen K, Dauwe R, Kim H, Niculaes C, Fukushima A, Goeminne G, Vanholme R, Ralph J, Boerjan W, Morreel K. 2021. Maize specialized metabolome networks reveal organ-preferential mixed glycosides. *Computational and Structural Biotechnology Journal,* **19**: 1127-1144.
Rainer J, Vicini A, Salzer L, Stanstrup J, Badia JM, Neumann S, Stravs MA, Verri Hernandes V, Gatto L, Gibb S, Witting M. 2022. A Modular and Expandable Ecosystem for Metabolomics Data Annotation in R. *Metabolites,* **12**: 173.