PhD project

Starting date: 1 October 2024; **Duration:** 3 years

The DynLib database: Development of computational tools for the structural characterization of plant specialized metabolites based on spectral metadata

Context : Plant tissues contain thousands of distinct compounds and, today, the vast majority of these plant metabolites remain unknown. Untargeted metabolomics experiments aim to characterize as many metabolite structures as possible, relying on the molecules' fragmentation spectra generated by Collision Induced Dissociation (CID) in a mass spectrometer. Determining the structure of a compound given its CID spectrum remains, however, a major challenge in metabolomics. Today, the reference method for computational identification of unknown metabolites models a fragmentation tree that best explains a CID spectrum and uses this fragmentation tree to predict the presence or absence of a series of structural features using an artificial intelligence-based approach. A relatively low success rate of this approach can be attributed, at least partially, to shortcomings in the modelled fragmentation trees. It is clear that the input of complementary information from different types of CID spectra, generated with different ionization or fragmentation parameters, or on different types of instruments (Quadrupole-Time of flight (QTOF) or Ion Trap (IT) mass spectrometers), can contribute to computing fragmentation trees that structurally make more sense, and are in better agreement with expert knowledge about Collision Induced Dissociation. We refer to this complementary spectral information, 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.

PhD Project : The objective of this PhD project is to develop tools to construct and expand a database for mass spectral metadata, the DynLib database, and to exploit these synergistic metadata to computationally predict metabolite structural features. The contribution of this research will serve metabolomics research in the field of plant science, as well as in the fields of human health and nutrition.

The first part of the thesis will consist of the generation of a mass spectral database, called DynLib, in which all relevant fragmentation data from previous and future metabolomics experiments, generated on different types of instruments and with different ionization or fragmentation parameters, can be automatically imported (Desmet et al., 2021). Together, these data will be referred to as spectral metadata. Functionalities to import data into the database, align experiments, annotate spectra, create metabolome networks, and query the database, will be developed to seamlessly integrate with the RforMassSpectrometry package ecosystem (Rainer et al., 2020).

In a second part of the thesis, the aim is to improve the performance and accuracy of molecular fragmentation trees modelling, by introducing constraints learned from spectral metadata in the newly developed database. This problem will be studied through the lens of constraint programming. Fragmentation tree problems will be modeled using well-known paradigms, such as Maximum Satisfiability (Max-SAT), and solved through dedicated powerful algorithms. A thorough experimental evaluation and comparison with other methods in the literature will be conducted. Fragmentation trees will be important for the pairwise comparison of fragmentation spectra within the DynLib database, and for the development of a computational prediction method for metabolite structures, based on spectral metadata from the DynLib database.

In a third part of the thesis, the newly developed tools and algorithms will be compared to existing reference software, and used to characterize the specialized metabolome of flax and wheat, as a part of ongoing plant physiological studies in the hosting lab.

Consortium: The PhD project will be conducted, for a period of 27 months, in the BIOPI Laboratory (**Lab 1**; UMRt BioEcoAgro, UPJV, Amiens, France), under the supervision of Dr. R. Dauwe, in close collaboration with the GOC Team (Graphes, Optimisation et Contraintes) of the MIS Laboratory (UPJV, Amiens, France), and, for a period of 9 months, in the Computational Metabolomics Team (**Lab 2**; Eurac Research, Institute for Biomedicine, Bolzano, Italy), under the supervision of Dr. Johannes Rainer.

	Year 1				Year 2				Year 3			
1/ LC-MS/MS and LC-MS^n data generation	▨	▨										
2/ Generation of the DynLib database and associated tools	▨	▨	▨	▨	▨	▨						
3/ Fragmentation Trees by AI approach						▨	▨	▨	▨	▨	▨	
4/ Evaluation, characterisation of flax and wheat metabolomes							▨	▨	▨	▨	▨	
5/ Valorisation (software, articles, thesis)						▨	▨	▨	▨	▨	▨	▨
Lab	1	1	2	2	2	1	1	1	1	1	1	1

Prerequisites :

- Background in bioinformatics/computer science/data analytics or in life science
- Good programming skills (R, Python, C/C++)
- Proficiency in English
- Strong scientific and technical curiosity
- Knowledge or interest in metabolism, biochemistry and mass spectrometry are a plus.

Contact:
To apply, please send your CV, motivation letter, and two references (name, title and e-mail) to rebecca.dauwe@u-picardie.fr and johannes.rainer@eurac.edu

Application :
www.adum.fr
Application deadline : 06 May 2024

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.

K. Scheubert, F. Hufsky, and S. Böcker. Multiple Mass Spectrometry Fragmentation Trees Revisited: Boosting Performance and Quality. In: Brown D, Morgenstern B, eds. Algorithms in Bioinformatics. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014.