Computational mass spectrometry bioinformatics projects

Large amount of data is generated by modern mass spectrometry platforms providing deep protein, glycoprotein and metabolome profiles of biological samples used for fundamental or clinical research. We propose several bioinformatics projects in the domain of computational mass spectrometry requiring different level of programming expertise:

Starting to medium level expertise projects

- Assessment of several software to automatise glycan structure identification on label-free or labeled glycan LC-MS/MS profiles. Tools ranges from glycan identification using MSFragger tool, identification of unknow glycan structures using predicted renetion time, m/z shift based on combination of glycan building block and combinations using glycopeptidegraphms tools (see further https://bitbucket.org/glycoaddict/glycopeptidegraphms/src/master/ and https://www.beilstein-journals.org/bjoc/articles/16/253). Require python programming. The project will be assisted by Peter Horvatovich, Guinevere Langeveen-Kammeijer and Hector Franco Loponte.
- Development of python and nextflow workflow for proteomics tools (MSFRagger, DIAUmpire) for proteogenomics workflow. Assessment of data independent acquisition to identify protein variants in LC-MS/MS using proteogenomics data integration. Require python programming. The project will be assisted by Peter Horvatovich, Yanick Hagemeijer and Victor Guryev.
- Metabolite identification and quantification in untargeted metabolomics data. Require python and/or R programming. The project will be assisted by Peter Horvatovich and Hector Franco Loponte.

Expert level expertise project

The Pipelines and Systems for Threshold-Avoiding Quantification of LC-MS/MS Data (PASTAQ) is being developed to quantitatively pre-process, explore and annotate LC-MS/MS data, with a special focus on full data traceability and the quantification of low intensity signals. At its most basic level, LC-MS/MS data can be seen as a 2D image, where compounds are separated based on their chemical properties (e.g. hydrophobicity) by liquid chromatography (LC) as well as their weight by mass spectrometry (MS). These compounds appear on the image as 2D Gaussian-like peaks, and it is the pipeline's task to detect and quantify them. Many interesting computational challenges are present in this data, and we are looking for people with an interest in programming (C++, Python, R), computer science, algorithms, and/or bioinformatics. Previous knowledge of LC-MS instrumentation is not required, but it would be appreciated (if not available we will provide an introduction to the level that is necessary to understand the data structure and properties). Ideally the candidate(s) should be familiar with software version control systems (Git) and cross platform development (Windows/Mac/Linux). More details are available at https://pastaq.horvatovichlab.com/ and https://github.com/PASTAQ-MS/PASTAQ.

We have a number of self-contained projects within the PASTAQ pipeline that could be interesting candidates for a bachelor or master level project and thesis. Some of which include, but are not limited to:

- Adaptation of time alignment tool based on PASTAQ allowing to align retention time between two mzML files.
- Seamless integration of proteomics/metabolomics search engines within the pipeline. Currently we rely on external tools (e.g. SearchGUI, PeptideShaker) to generate mzIdentML files that we use to annotate the detected peaks, but in addition to that, we would like to have the option to perform the search within our own framework.
- Testing and optimisation of prototype tool performing pseudo spectra extraction to processing data obtained with data independent acquisition.
- Implementation of ion mobility dimension in PASTAQ to process LC-MS/MS from timsTOF instruments.
- Addition of more unit tests. We are using Doctest for testing a limited set of the C++ functions, but we would like to expand both the number and quality for a wider code coverage.
- Implementation of R bindings following the existing Python bindings. We use Pybind11 for the generation of Python bindings, and would like to use the API exposed from C++ to have the same functionality in the R programming language.
- Exploration of N-dimensional space partitioning techniques for quick data access (e.g. KD-trees).
- Indexing of binary data files to enable reading data only for the sections we are interested in, instead of the entire file.
- Advanced visualization using GPU powered technologies, such as OpenGL or Vulkan to process and/or visualise large LC-MS/MS images with extensive annotation.

For all projects:

We are located at A. Deusinglaan 1, 9713 AV Groningen (ERIBA building, 6th floor). The projects can be started at any time. If you are interested, feel free to contact Prof. Dr. Peter L. Horvatovich <p.l.horvatovich@rug.nl> or to discuss the scope and availability of projects.