Match-between-runs in ProteomicsDB

Title: Transferring identification information between experiments in ProteomicsDB using Match-between-runs

Type: MSc thesis

Category: Algorithm development, DS, DB

Programming language: [any (e.g. Python)]

Language: [English]

Prior experience: [programming skills required, no biological knowledge necessary]

Complexity/Risk: High

Contact person: Wassim Gabriel, Mathias Wilhelm

Brief background description: Match-between runs is an effective method to reduce missing values in experiments of multiple runs. The method uses the fact that even though not all peptides are subjected to fragmentation, the quantitative information for these peptides is still recorded in the MS1 spectra. We can, thus, transfer identifications to these so-called MS1 features, provided we have identified a highly similar MS1 feature in a different run. We have a much more extreme version of this problem in our large-scale resource, ProteomicsDB, where a highly heterogeneous set of experiments and runs are represented.

Literature

- Cox et al. (2014). Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction termed MaxLFQ. Molecular & cellular proteomics, 13(9), 2513-2526.
- Samaras et al. (2020). ProteomicsDB: a multi-omics and multi-organism resource for life science research. Nucleic acids research, 48(D1), D1153-D1163.

Brief description of the project: In this project, you will investigate the viability of applying Match-between-runs on the repository scale. The main challenge will be how to deal with the large heterogeneity of the data, mostly with respect to the large variability of chromatography and its low reproducibility between laboratories. Another important aspect will be to ensure the reliability of the transferred identifications using statistical models. While this is likely to be a very challenging project, showing that this idea works could revolutionize the field.

Expected result: A proof of concept for using Match-between-runs on datasets in ProteomicsDB, which could lead to a short publication in a peer-reviewed journal.