IMPROVEMENT OF BIOFORMATIC WORKFLOW IN PROTEOMICS: DDA & DIA QUANTIFICATION

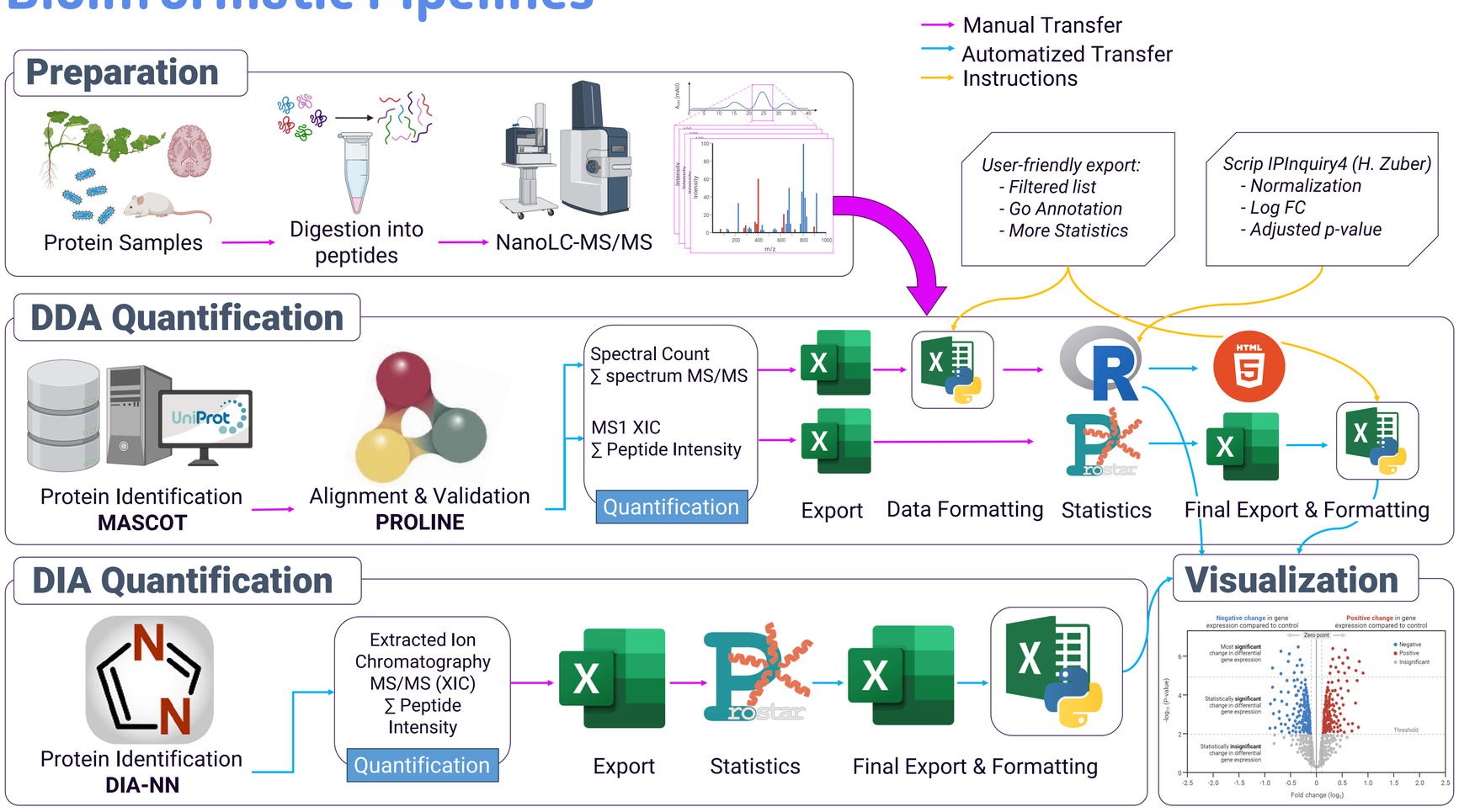
Pierre-Alexandre HO[1], Johana CHICHER[2] , Philippe HAMMANN[2]

[1]Master student in ERAMUS MUNDUS Chemoinformatics+ track In Silico Drug Design, Universities of Strasbourg, Milano & Paris Cité
[2]Engineer Plateforme Protéomique Strasbourg-Esplanade, IBMC, CNRS UAR1589

Summary

Proteomics, as the other Omics technics, plays an important role in biology, biotechnology and pharmaceutical research strategies. Classical approach consists in bottom-up identification of proteins via MS/MS analysis with a DDA quantification method. Nevertheless, this technic can be improved in term of identification and the reproducibility of quantification. Recently, the new DIA approach is more and more implemented in research laboratories. This change of paradigm increase proteomic depth and a better reproducibility. However, these two methods require several time-consuming steps in the bioinformatic processing. To improve the workflow, automatization of excel report via a python script has been done in this work included some basic statistics and data visualization.

Bioinformatic Pipelines

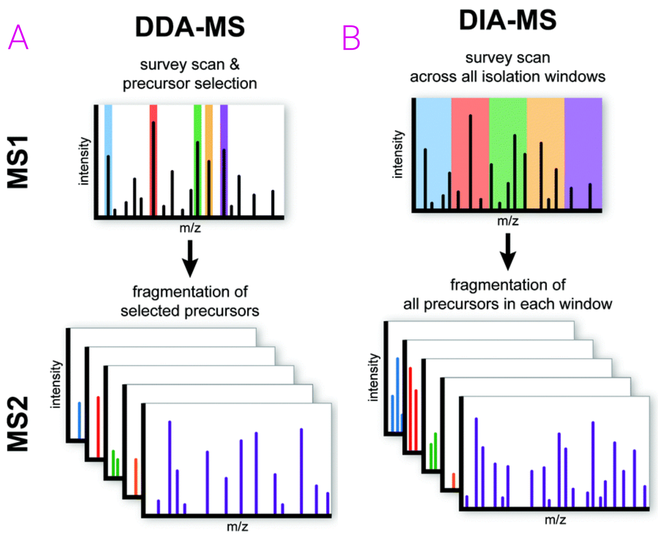


Materials & Methods

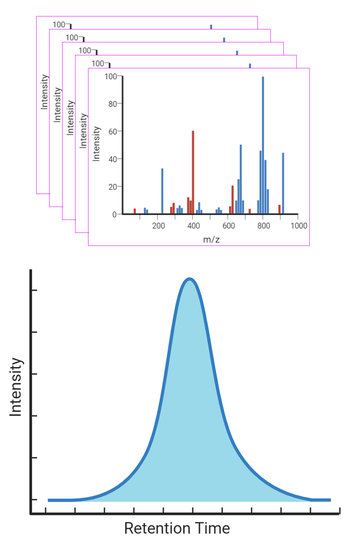
Purified pollen proteins were precipitated in 0.1M ammonium acetate in methanol then reduced, alkylated and digested overnight in trypsin (Promega). Peptide mixtures were analyzed by nanoLC-MS/MS on a reversed phase nanoElute 2 coupled to a TIMS-TOF Pro 2 mass spectrometer (Bruker Daltonik Gmbh) using a DDA or a DIA strategies. Peptides were separated on the integrated emitter column IonOpticks Aurora Elite. A range of 100–1700 m/z in full MS was applied for the DDA-PASEF acquisition. For the 10 MS/MS PASEF, collision energy was ramped with increasing mobility (20 to 59 eV). A dynamic exclusion time of 20 s was applied during the peak selection process. For DIA acquisitions, full MS data were acquired in a range of 100–1700 m/z and windows ranged from 400 m/z to 1200 m/z with 26Th isolation windows and were acquired with ramp times of 100ms.

A.thaliana proteins were identified via comparison to the TAIR v10 protein database, analyzed using Mascot algorithm (v2.8) with a decoy strategy. Data were imported into Proline v2.0 software proteins were validated on Mascot pretty rank equal to 1, a Mascot score threshold set at 25 and 1% FDR on both peptide spectrum matches (PSM score) and protein sets (Protein Set score) and MS1 XIC were used to quantify each protein. The same TAIR v10 .fasta file was used for DIA analysis with DIA-NN with in a database-free configuration.
For both analyses, Prostar 1.34.5 was used for the statistical analyses of the intensities. Partially observed values (POV) were imputed with structured least square adaptative regression (SLSA) method while values missing in the entire condition (MEC) were imputed with det quantile 1%.

DDA vs DIA

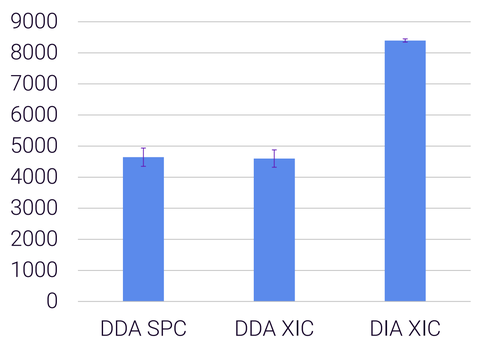


SpC vs XIC

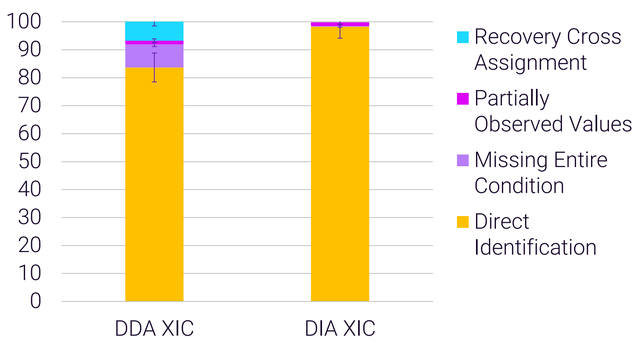


DDA for *Data Dependent Acquisition,* consists of scanning all the **precursor peptide ions** during the first acquisition (**MS1**) then **fragmenting the x most abundant** ones in a second acquisition (**MS2**) (A). The identification of proteins is performed by comparison to a specific database and the quantification via **spectral count analysis** (SpC) which counts the number of spectra obtained for a protein (C), or via **extracted ion chromatogram** (XIC) which integrates the peak of the protein in the MS1 (D). This method allows better tracking of the protein but generates more **missing values** which must be imputed. In contrast, DIA for *Data Independent Acquisition*, **fragments everything in a selected m/z window** (B). This has two advantages (figure below): A wider depth of analysis & fewer missing values. However, this technic needs more computational power for the deconvolution of spectra and is making spectacular progress thanks to **deep learning** approaches.

Number of quantified protein



Percentage of imputed values

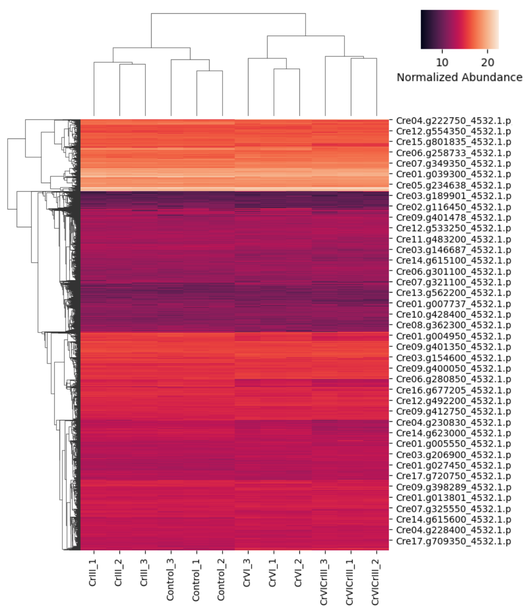


Statistics & Visualizations

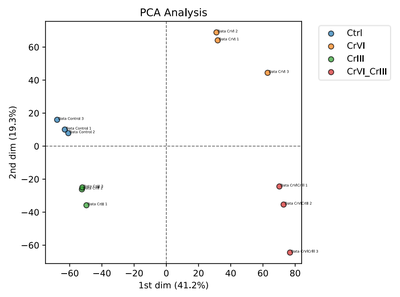
Interactive graph (html format) created with Plotly python library

Global Overview

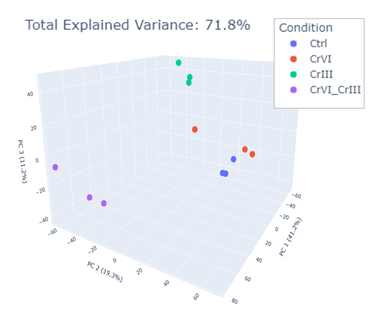
Heatmap



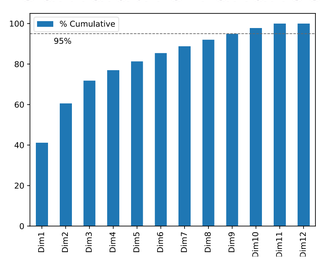
PCA



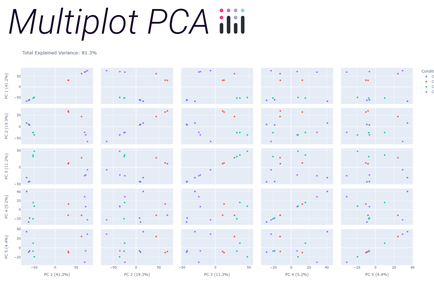
3D PCA



Cumulative Variance

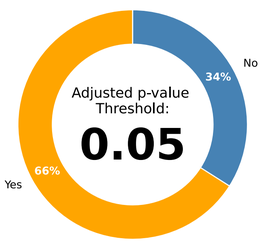


Multiplot PCA



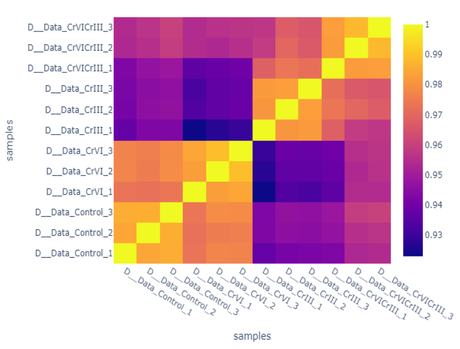
Anova

Ring Chart



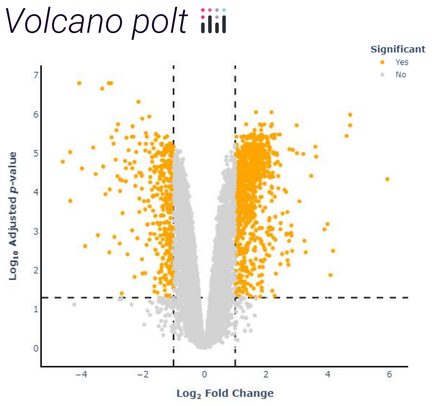
66% of protein impact significatively the different conditions tested

Correlation Matrix



2 by 2 Comparison

Volcano polt



Seaborn package was used for the heatmap (*clustermap*), scikit-learn for the PCA analysis, scipy for the anova (*f_oneway*) & statsmodel for the **Benjamini-Hochberg adjusted *p*-valvue** (*multipletests*). The Fold Change for the 2 by 2 comparison was calculated with the Prostart software using the a **Limma test** after exhaustive map alignment followed by a median ratio normalization of the intensities.

Acknowledgments

This project was developed in the context of a non mandatory internship during the Erasmus Mundus Joint Master Chemoinformatics+ funded by the European Union. This work was done in the Plateforme Proteomique Strasbourg-Esplanade from IBMC. I would like to thanks Mr. HAMMANN Philipe and Mrs. CHICHER Johana for welcoming me into their department, for their explanation and introducing me to the world of Proteomics. I am grateful to Pr. MARCOU Gilles for his advice in statics methods. I acknowledge Mr. V Normant who has authorized me to use his data for this poster.

Reference

K. Lukas, Paul H. Huang. « Data-Independent Acquisition Mass Spectrometry (DIA-MS) for Proteomic Applications in Oncology ». *Molecular Omics* 17, nº 1 (2021): 29-42

Poster Author Information

Pierre-Alexandre HO
M1 Chemoinformatics+
ISDD
hpierrealex@yahoo.fr