Optimization of strategies for proteomic data analysis from the velvetbean caterpillar Anticarsia gemmatalis as a non-canonical organism model

Pedro Henrique Bastos de Oliveira Nascimento[1,2], Michele Rodrigues Martins[2], Luís Felipe Costa Ramos[2], Fábio César Sousa Nogueira[2], Danielle Maria Perpétua de Oliveira[2], Magno Rodrigues Junqueira[2]

[1.] IFRJ, Instituto Federal do Rio de Janeiro, R. Sen. Furtado, 121/125 - Maracanã, Rio de Janeiro - RJ, 20270-021;
[2.] UFRJ, Universidade Federal do Rio de Janeiro, Av. Athos da Silveira Ramos, 149 - Bloco A, 2 andar - Cidade Universitária, Rio de Janeiro - RJ, 21941-909;

Proteomics is the branch of biology that comprehensively analyzes protein sets and their expression in a biological sample. Although proteomics is an essential field of research, there needs to be more scientific investigation for some analysis, which creates a pressing need for more data, hindering progress for different investigations. For that reason, *A. gemmatalis* is considered a non-canonical organism in biology. To bridge this gap, we will employ various bioinformatics analysis software with unique capabilities to explore the protein characteristics of *A. gemmatalis* thoroughly. Additionally, we will utilize the *De novo* sequencing technique, allowing us to sequence proteins and identify new or modified peptides not present in any database. Proteomics was performed with protein extraction and tryptic digestion via the S-trap protocol and TMT-10 plex, ending with analysis by nanoLC MS/MS Q-exactive plus. UniProt database of phylogenetically similar species and the transcriptome obtained from the literature were utilized. The methods chosen included different software, Proteome Discoverer 3.1 (PD), MaxQuant 2.4.9.0 (MQ), Novor.cloud, and PEAKS Studio 11, to compare identification and quantification related to RAW files. The unique peptide abundances of unique peptides, protein groups (Master protein filtering), and PSMs (peptide spectrum matches) were used. The analysis involving PD and MQ demonstrated that the protein groups found with PD were significantly higher than with MQ, with the quantities being 2.949 (PD) and 744 (MQ) for study 1 (S1), which is related to the transcriptome; 2.681 (PD) and 596 (MQ) for study 2 (S2), associated with UniProt database. Venn diagrams demonstrated that in S1, only 8 protein groups from MQ were exclusive, and in S2, 183 were also exclusive for MQ. Moreover, with PEAKS and Novor software, the data analysis followed the same premise as PD and MQ but focused on peptide sequences. Comparing the identified sequences, we obtained: 161.905 (PEAKS) and 64.377 (Novor) in S1; 160.151 (PEAKS) and 62.563 (Novor) in S2. Peptide spectrum matches (PSMs) were massed from PD and PEAKS and their performances was compared: 152.730 (S1 from PD), 138.853 (S2 from PD); 161.905 (S1 from PEAKS, ALC>50), 160.151 (S2 from PEAKS, ALC>50). This last comparison strongly suggests that the peptide identification potential from recent algorithms and software that includes De novo sequencing is remarkably high, opening up new and exciting possibilities for the analysis of non-canonical organisms. Software directly associated with a company, consequently paid software like PD and PEAKS, demonstrates superior quantitative identification, while free software such as MQ and Novor excels in specific peptide sequence recognition. The differences in performance can be attributed to various factors, including algorithms, optimizations, data processing, and customization in paid software. These factors contribute to the superior quantitative identification of paid software, while the focus on specific peptide sequence recognition in free software is also due to the use of different algorithms and optimizations.

Agradecimentos: We express our deepest gratitude to the following scientific funding agencies, FAPERJ and CAPES, and for the utmost support from the Translational Mass Spectrometry and Neuroproteomics Laboratory (LEMTEN). Your generous support has been instrumental in advancing our research, enabling us to explore new frontiers and achieve meaningful outcomes. Thank you for your unwavering commitment to the scientific community.