PepExplorer: a similarity-driven tool for analyzing \textit{de novo} sequencing results

Felipe V Leprevost$^1$, Richard H Valente$^2$, Diogo L Borges$^1$, Jonas Perales$^2$, Rafael Melani$^3$, John R Yates III$^4$, Valmir C Barbosa$^5$, Magno Junqueira$^3$, Paulo C Carvalho$^{1*}$

1. Laboratory for Proteomics and Protein Engineering, Carlos Chagas Institute, Fiocruz, Paraná, Brazil.
2. Laboratory for Toxinology, Oswaldo Cruz Institute, Fiocruz, Rio de Janeiro, Brazil.
3. Proteomics Unit, Rio de Janeiro Proteomics Network, Department of Biochemistry, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil.
4. Department of Chemical Physiology, The Scripps Research Institute, La Jolla, CA, United States.
5. Systems Engineering and Computer Science Program, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

*Corresponding author: paulo@pcarvalho.com

Peptide Spectrum Matching (PSM) is the current gold standard for protein identification by mass-spectrometry-based proteomics. PSM compares experimental mass spectra against theoretical spectra generated from a sequence database to perform identification; sequences not found in the database are not identified. The alternative approach, \textit{de novo} sequencing, can infer a peptide sequence directly from a mass spectrum; yet, interpreting long lists of peptide sequences resulting from large-scale experiments is not trivial. With this as motivation, we present PepExplorer, a computational tool that relies on rigorous pattern recognition approaches to assemble a list of homologue proteins based on \textit{de novo} sequencing coupled to sequence alignment, thereby enabling biological interpretation of the data. PepExplorer can read the output of various widely adopted \textit{de novo} sequencing tools and converge to a list of proteins under a global False-Discovery Rate (FDR). To this end, it employs a radial-basis-function neural network that considers precursor charge states, \textit{de novo} sequencing scores, peptide lengths, and alignment scores to select similar protein candidates from a target-decoy database usually obtained from “close” species. Alignments are performed using a modified Smith-Waterman algorithm tailored for the task at hand. We have verified the effectiveness of our approach on a reference set of identifications generated by ProLuCID when searching for \textit{Pyrococcus furiosus} mass spectra on the corresponding NCBI Refseq database. We then modified the sequence database by swapping amino acids until ProLuCID was no longer capable of identifying any proteins. By searching the mass spectra using PepExplorer on the modified database, we have been able to recover most of the identifications under a 1\% FDR. Finally, we have employed PepExplorer to disclose, for the first time, a comprehensive proteomic assessment of the \textit{Bothropos jararaca} plasma. PepExplorer is integrated into the PatternLab for Proteomics environment, which makes available various tools for downstream data analysis, including resources for quantitative and differential proteomics.

Acknowledgements: FAPERJ, CNPq, CAPES, PDTIS