A large-scale approach for pinpointing mutations / polymorphism in complex protein mixtures by mass spectrometry

Carvalho P.C.¹; León, I.R.²; Valente, R.H.³; Jensen, O.N.²; Perales J³; Fischer J.S.G¹; Yates J.R.; III⁴; Xu T⁴; Barbosa V.C.⁵; Neves-Ferreira, A.G.C.³

¹ Carlos Chagas Institute, Fiocruz, Paraná, Brazil; ² Protein Research Group, University of Southern Denmark, Odense, Denmark; ³ Laboratory of Toxinology, Fiocruz, Rio de Janeiro, Brazil; ⁴ The Scripps Research Institute, USA ⁵ Systems Engineering and Computer Science Program, UFRJ, Brazil.

Proteomic approaches have served the community as a key discovery tool for describing differentially expressed proteins in complex protein mixtures. Its three data analysis cornerstones are: peptide spectrum matching (PSM), tag searching, and de novo sequencing. The PSM associates experimental spectra to those theoretically generated from a sequence database (DB) and thus is blind to sequences not present in the DB. The tag searching computes short amino acid sequences and then returns the closest peptide sequence from the DB. De novo sequencing does not require a DB, but is unarguably the most error prone. As such, core proteomic pipelines are not well-designed for determining mutations / polymorphisms and thus, ultimately, miss crucial information in the study of diseases (e.g., cancer) and rapidly evolving species. Here, we propose a shift in the paradigm to the widely adopted proteomic strategies and demonstrate its effectiveness on a snake venom sample. Briefly, the sample was trypsinized and LC/LC/MS/MS data obtained on an Orbitrap XL. The data were searched using ProLuCID and confident PSMs were statistically obtained using the Search Engine Processor (SEPro). An automatic quality assessment was employed to fish high quality spectra not identified. These were then analyzed by a new algorithm that combines tag searching, PSM, and a novel combinatorial approach to allow sequencing hits not found in the DB. The results were assessed by SEPro to meet a 1% FDR and numerous novel sequence modifications were described.

Keywords: Mutations, Polymorphism, shotgun proteomics
Supported by: CAPES, FAPERJ, FIOCRUZ / PDTIS.