A NOVEL STRATEGY FOR DIFFERENTIAL PROTEOMICS BY SPECTRAL COUNTING AND STATISTICAL LEARNING

Carvalho P.C.[1], **Fischer J.S.G.**[2], **Domont G.B.**[2], **Carvalho M.G.C.**[3], **Hewel J.**[4], **Yates J.R. III**[4], **Barbosa V.C.**[1]

[1]Systems Engineering and Computer Science Program, Federal University of Rio de Janeiro, Brazil. [2]Systems Biology Laboratory, IQ, Federal University of Rio de Janeiro, Brazil. [3]Control of Gene Expression Laboratory, IBCCF, Federal University of Rio de Janeiro, Brazil. [4]Department of Cell Biology, The Scripps Research Institute, La Jolla, California, USA.

A goal of proteomics is to distinguish between various states of a system to identify protein expression differences. Multidimensional chromatography coupled online with mass spectrometry (LC/LC/MS/MS) has become a reference method for high throughput protein identification in complex mixtures. By using the numbers of tandem mass spectra obtained for each protein or "spectral counting" as a surrogate for protein abundance in a mixture, Liu *et al*. demonstrated the use of LC/LC/MS/MS to obtain semi-quantitative data on mixtures. However, two issues remained open: how to normalize spectral counting data and how to identify the differences between samples. We developed a new strategy, having roots on statistical learning theory, genetic algorithms and support vector machines, to indicate which and how many proteins are differentially expressed when comparing LC/LC/MS/MS data. The strategy was validated by correctly pinpointing which and how many protein markers were spiked into yeast lysates.

Acknowledgements: CNPq, FAPERJ BBP grant, National Institutes of Health (P41 RR11823-10, 5R01 MH067880, U19 AI063603-02) and the Genesis Molecular Biology Laboratory.

Key words: MuDPIT, pattern recognition, machine learning, proteomics, bioinformatics