Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

Identifying differences in protein expression levels by spectral counting and feature selection

P.C. Carvalho[1], J. Hewel[2], V.C. Barbosa[1] and J.R. Yates III[2]

[1]Programa de Engenharia de Sistemas e Computação, COPPE,
Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, Brasil
[2]Department of Cell Biology, The Scripps Research Institute,
La Jolla, CA, USA

Corresponding author: P.C. Carvalho
E-mail: carvalhopc@cos.ufrj.br

Genet. Mol. Res. 7 (2): 342-356 (2008)
Received January 18, 2008
Accepted February 7, 2008
Published April 15, 2008

ABSTRACT. Spectral counting is a strategy to quantify relative protein concentrations in pre-digested protein mixtures analyzed by liquid chromatography online with tandem mass spectrometry. In the present study, we used combinations of normalization and statistical (feature selection) methods on spectral counting data to verify whether we could pinpoint which and how many proteins were differentially expressed when comparing complex protein mixtures. These combinations were evaluated on real, but controlled, experiments (yeast lysates were spiked with protein markers at different concentrations to simulate differences), which were therefore verifiable. The following normalization methods were applied: total signal, Z-normalization, hybrid normalization, and log preprocessing. The feature selection methods were: the Golub index, the Student *t*-test, a strategy based on the weighting used in a forward-support vector machine (SVM-F) model, and SVM recursive feature elimination. The results showed that Z-normalization combined with SVM-F correctly identified which and how many protein markers were added to the yeast lysates for all different concentrations. The software we used is available at http://pcarvalho.com/patternlab.

Key words: MudPIT; Feature selection; Support vector machine; Spectral counting; Feature ranking