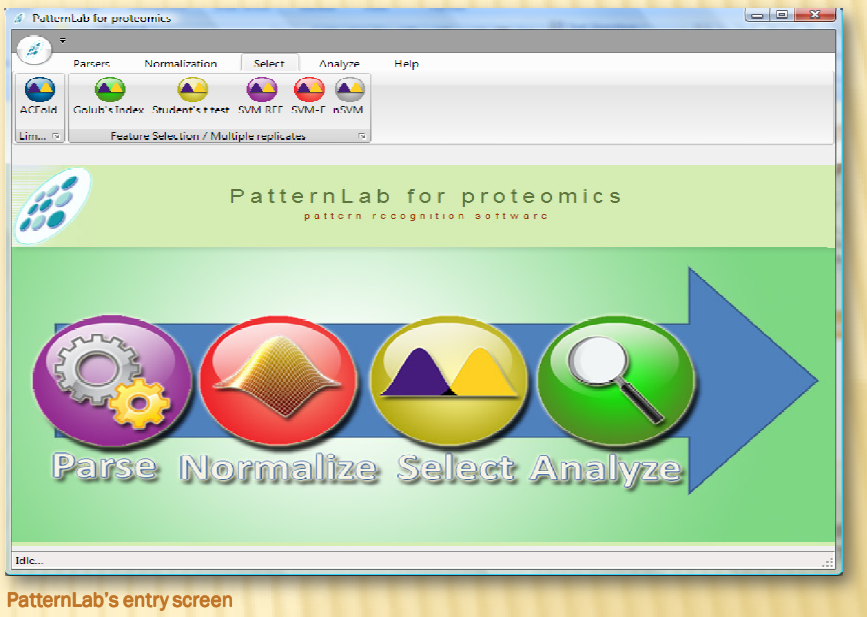
PATTERNLAB FOR PROTEOMICS: AN INTEGRATED APPROACH FOR ANALYZING SPECTRAL COUNTING DATA

Carvalho PC[1,3], Fischer JSG[2,3], Chen E[3], Barbosa VC[1], Yates JR III[3]

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

Introduction: PatternLab for Proteomics, an integrated computational environment for differential proteomics

Here we present PatternLab for proteomics, a computational environment that provides existing and new algorithms to deal with spectral counting data. Two of the new strategies are ACFold and nSVM (natural support vector machines), both for selecting differentially expressed proteins. The former combines expression fold changes, the AC test, and a theoretical false-discovery rate estimator. It stands out because it can compare data acquired with different protocols (e.g., a multi-surfactant shotgun approach). The latter, nSVM, has roots in evolutionary computing and statistical learning theory. Its usefulness was demonstrated by correctly pinpointing which and how many protein markers were spiked into yeast lysates while other widely adopted methods (e.g., the t-test and SVM-RFE) failed.

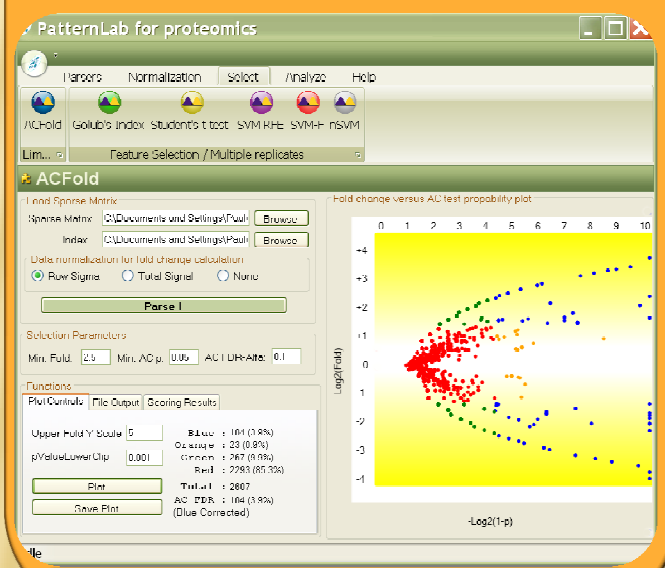


PatternLab's entry screen

Methods & Results

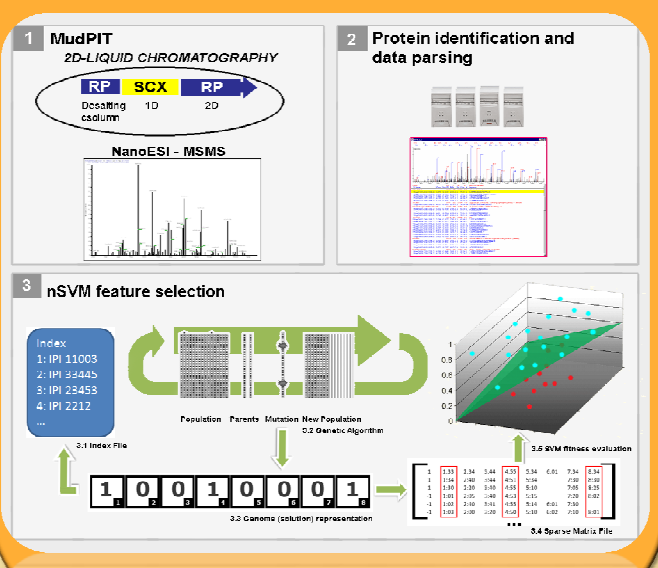
ACFOLD

The interface below displays results from real experimental data. The plot shows the distribution of identified proteins according to $\log_2$(fold change) on the ordinate (y) and $-\log_2(1-$(AC test *p*-value)) on the abscissa (x). The plot tab indicates that 104 proteins (blue dots) were differentially expressed because they satisfied both the AC test and fold-change cutoffs specified by the user. 23 proteins (orange dots) did not meet the fold-change cutoff but were indicated as statistically differentially expressed, therefore deserving a second look. 267 proteins (green dots) met the fold-change cutoff; however, the AC test indicated that this happened by chance. 2293 proteins (red dots) were pinpointed as not differentially expressed between classes because they failed both the AC test and the fold-change cutoffs. The GUI also lists an AC FDR indicating that all blue dots satisfy the established user-selected FDR of 0.1.



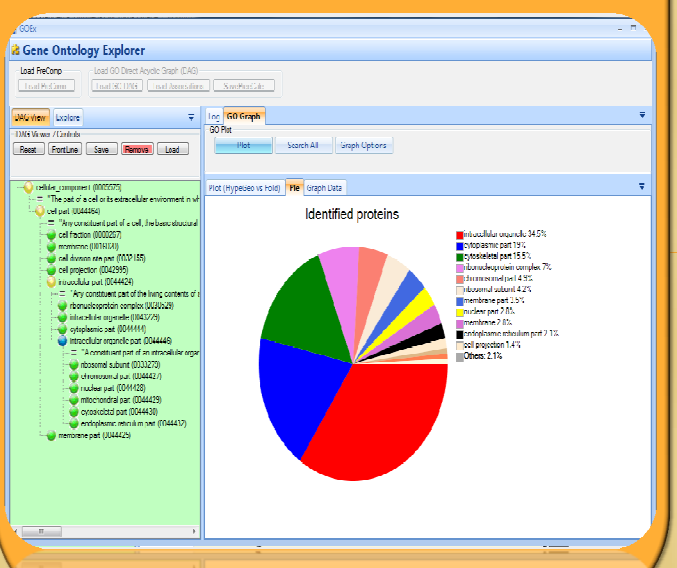
NSVM

MudPIT is applied to acquire mass spectrometry data from a biological system in different states (1). The data are subsequently identified by SEQUEST and filtered by DTASelect (2). nSVM is applied to pinpoint differences in the protein expression profiles by combining a genetic algorithm with support vector machines (3.2). Each individual's genome is an array of bits (3.3) that corresponds to a set of proteins (3.1 and 3.2) that will be selected from the dataset (3.4) to be evaluated as a solution (3.5) according to their spectral counts. nSVM was verified by correctly pinpointing spiked proteins in a yeast lysate while other methods (e.g., t-test, SVM-RFE) failed.



GO EXPLORER

GO Explorer (GOEx) is an integrated tool to help analyze proteomic spectral counting data by leveraging the gene ontology and considering protein fold changes with over-representation statistics.
Its GUI is found below. The user can analyze the differentially expressed protein reported from one of PatternLab's methods according to any of the GOEx study modes: iDAG-driven (interactive Direct Acyclic Graph), specialist-driven, and automatically driven. A pie chart showing the distribution of the identified proteins as mapped onto selected cellular component GO terms is displayed on the right. The level of specificity was chosen according to the iDAG in the left panel. Other analysis tools and graphing options are also available.



Conclusions

All methods have a carefully crafted graphical user interface and are integrated in one computational environment. To help gain biological insight, PatternLab also leverages the gene ontology database. Differently than other tools, it adds protein fold changes to the over-representation statistics. All these features considered, PatternLab is a new, simple to use, and powerful tool for analyzing shotgun proteomic spectral counting data. It is available at http://pcarvalho.com/patternlab