Interactive Statistical Resource for Proteomic Analyses

Bobbie-Jo Webb-Robertson, Lisa Bramer, Markus Kobold, Jeffrey Jensen, Kelly Stratton, Mathew Thomas, Amanda White

ABSTRACT

One of the major challenges to the discovery of robust biomarker candidates from global shotgun mass spectrometry (MS)-based proteomic experiments is processing these large and complex datasets in a manner that accounts for the underlying statistical hypothesis of interest. We believe for the field of clinical proteomics to successfully identify new prognostics or diagnostics of disease or response to therapies requires not only high quality data with respect to the instrument, but also high quality statistical analysis of the data. This project proposes new informatics technology in the form of a robust, interactive web-based software environment (P-Mart) that will enable biomedical and biological scientists to perform in-depth analyses of global proteomics data from the point of quality assessment and normalization of raw inferred abundances (e.g., peak area) to the identification of protein biomarkers.

P-MART CAPABILITIES

• Datasets collected and released by the Clinical Proteomics Tumor Analysis Consortium (CPTAC) available through https://cptac-data-portal.georgetown.edu/cptacPublic/ at the peptide, gene or protein level. P-Mart allows for the selection of parameter of interest for evaluation (e.g., clinical stage, tumor residual disease) – Figure 1.

• Modules for various analyses are available and user can select workflow or have P-Mart select based on data type – Figure 2.

• A series of quality control processes are available. Figure 3 shows both sample outlier sample identification [1] and normalization [2] as a visualization that enables users to make statistical-based decisions about the data quality control processing.

• A visualization tool called Trelliscope [3] is utilized to sort through the large number of potential biomarker candidates. Figure 4 shows Trelliscope being used to sort proteins by coverage and statistical p-value.

• Figure 5 shows the first plot based on this criteria of 996 plots. Clicking on the link brings you to the NCBI Protein database.

• Trelliscope can also be used to search for genes of interest (e.g., BRAF) and the visualization of the data associated with the proteomics data can be visualized along with meta-data, such as p-value – Figure 6.

UNDERLYING STATISTICAL HYPOTHESIS OF INTEREST

The plot on the left evaluates each sample as a potential outlier based on its data distribution. The user can modify the stringency of the test as appropriate to the analysis of interest. The plot on the right evaluates a large number of normalizations with a heatmap that indicates the most robust approaches as lighter colors.

REFERENCES


ABOUT

Pacific Northwest National Laboratory

The Pacific Northwest National Laboratory, located in southeastern Washington State, is a U.S. Department of Energy Office of Science laboratory that solves complex problems in energy, national security, and the environment, and advances scientific frontiers in the chemical, biological, materials, environmental, and computational sciences. The Laboratory employs nearly 5,000 staff members, has an annual budget in excess of $1 billion, and has been managed by Ohio-based Battelle since 1965.

For more information on the science you see here, please contact: Bobbie-Jo Webb-Robertson bj@pnnl.gov

http://pmart.azurewebsites.net/pmart/index.html

www.pnnl.gov