BMC Bioinformatics

Poster presentation

Proteome discovery pipeline for mass spectrometry-based proteomics

Erik Gough*[1], Cheolhwan Oh[1], Jing He[1], Catherine P Riley[1], Charles R Buck[1] and Xiang Zhang[2]

Address: [1]Bindley Bioscience Center, Purdue University, West Lafayette, IN 47907, USA and [2]Department of Chemistry, Center for Regulatory and Environment Analytical Metabolomics, University of Louisville, Louisville, KY 40292, USA

Email: Erik Gough* - goughes@purdue.edu

* Corresponding author

from UT-ORNL-KBRIN Bioinformatics Summit 2008
Cadiz, KY, USA. 28–30 March 2008

Published: 8 July 2008

BMC Bioinformatics 2008, **9**(Suppl 7):P21 doi:10.1186/1471-2105-9-S7-P21

This abstract is available from: http://www.biomedcentral.com/1471-2105/9/S7/P21

Overview

We have developed the Proteome Discovery Pipeline, a stand-alone bioinformatics platform used for LC/MS data analysis and biomarker discovery. Data is processed in a series of self-contained analytical steps using modules that are controlled by a graphical user interface. The user interface was developed in Visual C++ 6.0 and provides a multi-threaded, tabbed user interface with each tab representing a step in the analysis process. Modules included are spectrum deconvolution, alignment, normalization, significance tests and pattern recognition. Modules consist of applications developed in C++ and the R scripting language, which are called as external processes from the GUI using inputted parameters. Molecular correlation analysis can be viewed interactively using SysNet. Figure 1 shows the architecture of the Proteome Discovery Pipeline.

Spectrum deconvolution

XMass [1] uses chemical noise filtering, charge state fitting and de-isotoping for improved analysis of complex peptide samples. Overlapping peptide signals in mass spectra were deconvoluted by correlation with modeled peptide isotopic peak profiles. Isotopic peak profiles for peptides were generated *in silico* from a protein database to produce reference model distributions.

Peak alignment

XAlign [2] is a two-step alignment algorithm. The first step is to detect significant peaks that are common to all samples. In the second step, all samples are aligned to the median sample using refined m/z and retention time variation values, where pattern recognition is applied as needed.

Normalization

Several normalization methods have been developed for proteomics, including auto-scaling, reference sample, log linear model, trimmed constant mean, and average intensity.

Statistical significance tests

Several different test methods (two-tailed t-test, one-way ANOVA, Kolmogorov-Smirnov test, the Mann-Whitney test) can be used to identify data elements that make large contributions to the protein profile of a sample or that distinguish groups of samples from others.

Pattern recognition

We have implemented principal component analysis (PCA), linear discriminate analysis (LDA), canonical discriminate analysis (CDA), and clustering objects on subset of attributes (COSA) [3] as clustering methods.

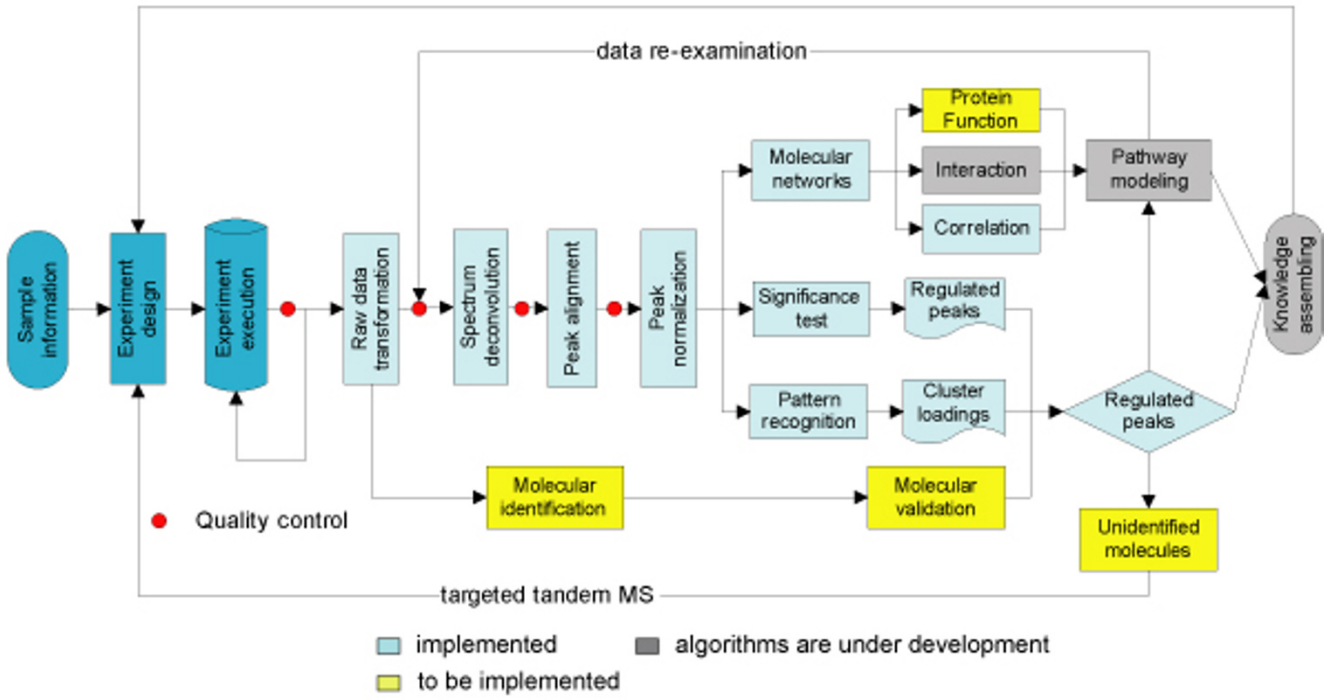

Figure 1
Architecture of the proteome discovery pipeline.

Molecular correlation

The software package, SysNet [4], is used to provide a dynamic visualization environment for molecular correlation of 'omics data. SysNet visualizes the 'omics expression data as a two-dimensional network. It features a circular layout, where molecular species are represented as nodes and all nodes are located on circles. The intermolecular correlations are represented as links, or edges, between nodes.

References

1. Zhang X, Asara J, Adamec J, Ouzzani M, Elmagarmid A: **Data preprocessing in liquid chromatography mass spectrometry based proteomics.** *Bioinformatics* 2005, **21:**4054-4059.
2. Zhang X, Hines W, Adamec J, Asara J, Naylor S, Regnier F: **An automated method for the analysis of stable isotope labeling data for proteomics.** *J Am Soc Mass Spectrom* 2005, **16:**1181-1191.
3. Friedman JH, Meulman JJ: **Clustering objects on subsets of attributes.** *J R Statist Soc B* 2004, **66(Part 4):**1-25.
4. Zhang M, Ouyang Q, Stephenson A, Kane MD, Salt DE, Prabhakar S, Burger J, Buck C, Zhang X: **Interactive analysis of 'omics molecular expression data.** *BMC Systems Biology* 2008, **2:**23.