Open Access

The Proteome Discovery Pipeline – A Data Analysis Pipeline for Mass Spectrometry-Based Differential Proteomics Discovery

Catherine P. Riley*, Erik S. Gough, Jing He, Shrinivas S. Jandhyala, Brad Kennedy, Seza Orcun, Mourad Ouzzani, Charles Buck, Ali M. Roumani and Xiang Zhang

Bindley Bioscience Center, Purdue University, W. Lafayette, IN 47907, USA

Abstract: Proteomics approaches enable interrogation of large numbers of proteins to provide a more comprehensive understanding of biological systems. High throughput proteomics typically utilizes liquid chromatography – mass spectrometry technology for data acquisition. Bioinformatic analysis tools are essential to manage and mine resulting high volume proteomics data sets. Data analysis is a current bottleneck for many proteomics researchers because complete and freely accessible already-developed systems are not available. In addition, most analysis systems require experienced bioinformatician input immediately upon data acquisition. For proteomics to achieve greatest impact in biology, data analysis must be more efficient and effective.

We present the Proteome Discovery Pipeline (PDP), a web-based analysis platform that provides proteomics data analysis without requirement for specialized hardware or input from bioinformatics specialists for initial data analyses. Functionalities of the PDP include spectrum visualization, deconvolution, alignment, normalization, statistical significance tests, and pattern recognition. The PDP provides proteomic researchers with a user-friendly web-based data analysis package that can handle multiple file formats and facilitates data analysis from multiple proteomics technology platforms. The system is flexible and extensible to enable further development. In this paper the PDP development is described and the system capabilities are illustrated through a case study of human plasma proteomics data analysis.

Keywords: Proteomic pipline, Data mining, Data analysis, Mass spectrometry.

INTRODUCTION

Mass spectrometry (MS) has emerged as the primary analytical tool for the evaluation of proteins and metabolites in complex biological systems, in part because of significant advances in the instrumentation. For example, the trend toward proteomic interrogation of biological systems with mass spectrometry was accelerated by the ability to couple MS data with other proteomic tools such as high performance liquid chromatography (HPLC) [1-3]. The use of systematic and reproducible fractionation coupled with MS enables identification and quantification of large numbers of proteins from mixtures or isolated macromolecular complexes [4-7]. This facilitates mapping protein interactions in a cellular context and analysis of biological activities. The power of high throughput global scrutiny, realized with the success of the Human Genome Project, has also been a driver for 'proteomics' approaches.

Proteomic strategies are focused on teasing out proteins that are representative of physiological changes brought on by disease [3]. These experiments commonly measure and compare the abundance of proteins from various stages in the disease. The typical procedures include digestion of the protein mixture into peptides, peptide separation using multidimensional liquid chromatography (MDLC), and finally mass spectrometry (MS) for quantification and tandem mass spectrometry MS/MS for identification of proteins from which the peptides were derived [1-3]. As for any study of complex biological systems, it is crucial that only relevant observations are identified and related to each other. Since thousands of mass spectra are generated in a single experiment, discovering the significantly changed proteins from millions of peaks detected in a typical experimental series requires intensive informatics capacity. Data analysis platforms have been recently described for LC-MS/MS proteomics [8-10]. In addition, LC-MS based proteomic quantification solutions are available including Corra [11], OpenMS [12], SuperHirn [13] and others [14-18]. However, these platforms orient around a specific type of data or limited functionalities for protein quantification. Interpretation and comprehension of the collection of mass spectra remain major challenges that involve multiple data analysis and mining steps [8, 19, 20].

An effective bioinformatic platform for MS-based proteomics must provide robust, reproducible data and should enable direct utilization by the proteomics researchers to achieve rapid and facile analyses of proteomics experimental data. In addition, it must accept input files from multiple experimental platforms and instruments and analysis tools should be integrated for proximal and efficient data analysis. Unfortunately, sample preparation, the choice of instrumentation, and configuration of the MS platform all contribute to the variation of output proteomics data. This has led to the

*Address correspondence to this author at the Bindley Bioscience Center, Purdue University, W. Lafayette, IN 47907, USA; Tel: (765)496-1756; Fax: (765)496-1518; E-mail: criley@purdue.edu

1875-0397/10 2010 Bentham Open

development of computational tools that are specific to only certain generations and/or types of instruments and MS platforms. Under the circumstances, it is often difficult to identify the appropriate software to assess proteomics data. This is further complicated by the practical issues such as data file compatibility, computational requirements, user friendliness, quality control, data visualization, *etc*. These factors present a bottleneck for progress with proteomics projects if there is a requirement that even initial data analysis must be handed-off to a dedicated bioinformatics specialist. Additionally, recent evidence suggests that the utilization of multiple proteomics tools and MS platforms can reveal different aspects of the proteome and provide independent verification of protein expression data [21, 22]. Thus, a flexible analysis platform takes on increased importance.

Mass spectral data analysis would be facilitated by a flexible and scalable pipeline infrastructure. This informatics pipeline should support the different experimental stages including sample preparation, mass spectrometry, analysis of the mass spectral data, access to external databases, and knowledge acquisition. Partial bioinformatic solutions have become available with the development of software packages such as GeneSpringMS™ [23] and OpenMS. What is not yet available however, and critically needed, is a complete pipeline that is agnostic to MS data formats and thereby able to process data collected from various MS platforms. Such a pipeline could be used to expand datasets from complex biological experiments and to enhance MS proteomics results with independent analytical analyses.

In this paper, we present the developed Proteome Discovery Pipeline (PDP) (http://proteomicsdp.org) and discuss its components that provide proteomics researchers with a web-based extensive data analysis pipeline. Using the PDP, any proteomics researcher or other user, with an internet browser, can submit mass spectral data for differential processing and application of various analyses and data mining algorithms, all with an intuitive graphical user interface. Compared with existing proteomics pipelines, PDP is able to process both stable isotope labeled and label-free proteomics profiling data. The PDP currently accepts MS data in several standard file formats. Furthermore, the utilization of web services provides flexibility for the system to adopt new data analysis packages.

MATERIALS AND METHODOLOGY

The PDP Implementation

The Proteome Discovery Pipeline (PDP) is developed to operate in a workflow based on the overall framework for mass informatics in differential proteomics (Fig. **1**), which includes: data preprocessing, protein identification, protein quantification, and analyses of protein networks. Data pre-

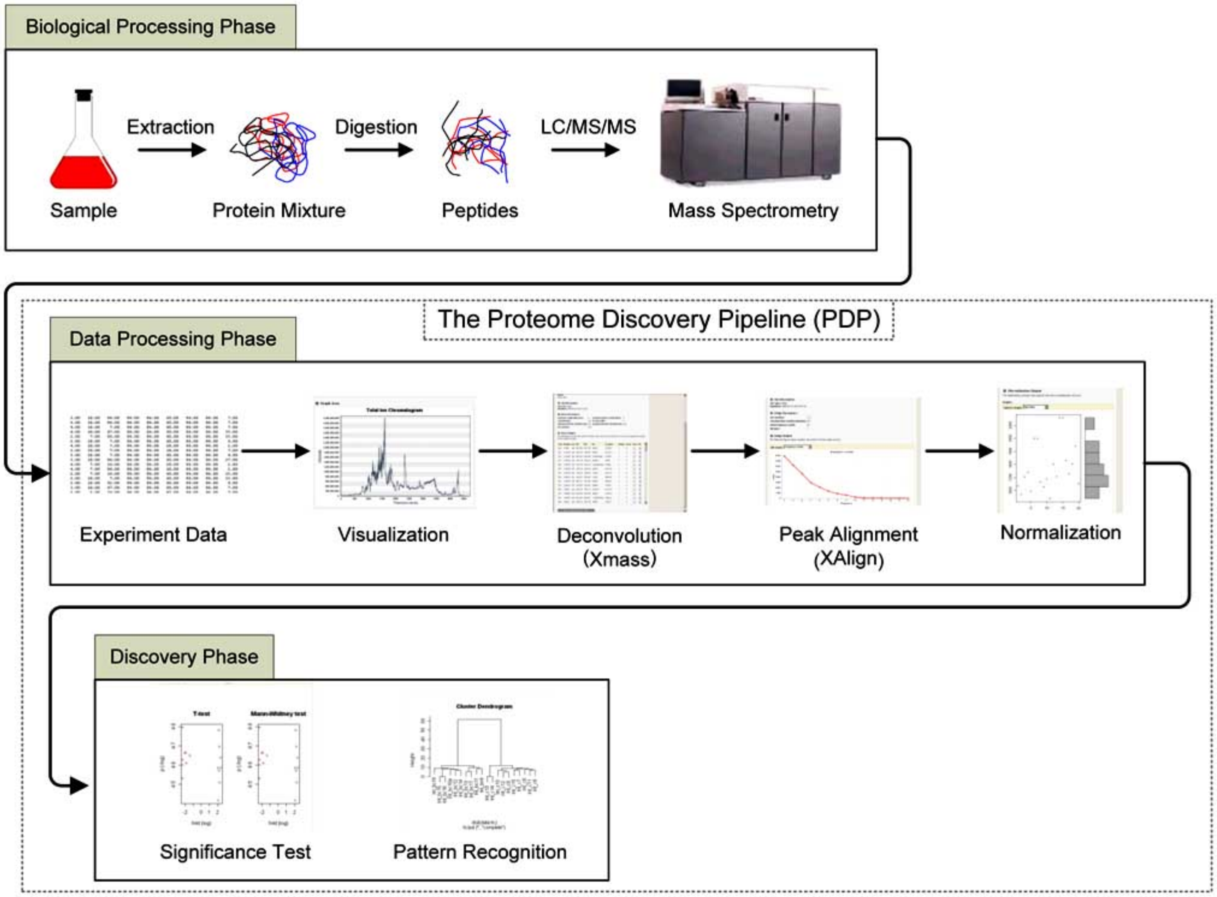

Fig. (1). Differential proteomics information flow. The Proteome Discovery Pipeline (PDP) provides all steps in the data processing and discovery phases of the workflow. The PDP enables analyses of multiple data formats and detection of quantitative differences between groups of samples.

processing includes raw mass spectra conversion, mass spectral deconvolution and peak alignment. The protein identification component identifies proteins corresponding to analyzed peptides along with a statistical significance of the identification. Peak normalization is required for protein quantification. Additional statistical significance tests discriminate differentially expressed proteins, and pattern recognition algorithms assist researchers in classification of detected and differentially expressed proteins.

The PDP provides a platform for researchers to seamlessly store, process, analyze and visualize MS datasets. The system currently supports proteomics and metabolomics datasets with the following operations: spectrum visualization, spectrum deconvolution and peak picking, peak alignment, data normalization, statistical significance testing, and pattern recognition. The PDP accepts LC-MS data from proteomics profiling experiments that do not rely upon tandem MS and therefore may capture additional data that would be lost with the tandem MS approach. The system also performs sequential data analysis routines to identify and quantify proteins, including high throughput experiments. It can process datasets in NetCDF, AgilentCsv, mzData, and mzXML formats. This ability to handle data in various formats is a distinct advantage of the PDP. Table **1** lists common MS instruments, acquisition software, raw data type for each instrument and the mzXML file converters.

Data processing and analysis tools are made available through the web-based interface and provide the required functionality for each stage. If multiple analysis tools exist for a single stage, the user is given the choice, based on the needs and nature of the experiment or personal preference. This is facilitated by wrapping analysis tools or data components as web services [24]. Web services are advertised in a central service registry and the pipeline engine queries the registry to indicate the services or tools available at any specific stage. Web services middleware enable a flexible design for the PDP, eliminating the need to change the pipeline engine as existing tools are modified, new tools are added, or the workflow is changed.

The PDP is developed with private access privileges to provide users with control and security over their data. The web-based infrastructure allows researchers to review the output generated at each step in the pipeline through electronic reports and visualization tools with interactive graphs and charts. Thus, the PDP provides easy-to-use quality control for MS proteomics data. Laboratory supervisors can manage their laboratory inputs and the lifecycle of experiments in the pipeline within their private workspace on the system, where they can re-submit data or alter input parameters for the different analysis routines.

System Components

The PDP can be broken down into two distinct components: a web-based client application and a processing server. Each component is deployed on a separate dedicated server and both use an implementation of the Simple Object Access Protocol (SOAP) [25] for communication. The client application consists of a web-based front end used to manage experiment groups and create jobs for submission to the processing server. The processing sever is responsible for data retrieval, job status queries, and executing processing algorithms on the specified data. Algorithms are advertised to the client application as web services. The processing server schedules and executes jobs based on client server requests. When processing is complete, results are stored in the user's private workspace and can also be downloaded and stored locally on the user's computer for data validation and further analyses.

Client Application

The PDP client application represents the interface through which users interact with the processing server. The client application consists of a struts-based Java Servlet application running on Apache Tomcat, a relational database for meta-data storage and a file system for job resource storage. The primary responsibilities of the client application are the management of experimental datasets, submission of data for processing, and visualization of results. The client application is also responsible for managing the data processing workflow.

The core of the PDP client is a configurable XML [26] workflow document that defines the concurrent steps of data

Table 1. Compatible Mass Spectrometer Instruments, Acquisition Software, Raw File Formats and mzXML Converters

Instruments	Software	Raw File	MZXML Converters
Agilent MSD TOF	Analyst MassHunter	.wiff .d	mzwiff supplied by vendor Trapper http://sourceforge.net/projects/sashimi/files/trapper%20%28M assHunter%20converter%29/
Agilent XCT ion trap	Chemstation	.D .baf	CmpassXport http://www.bioinformaticssolutions.com/products/peaks/support/bruker.php
Thermo LTQ	XCalibar	.RAW	ReAdW supplied by vendor
Waters LCT premier	MassLynx	.RAW	Masswolf supplied by vendor
ABI/ Maldi TOF QStar QTrap	4000 explorer Chemstation analyst	.t2d or.dat .wiff .wiff	T2dExtractor http://tools.proteomecenter.org/T2DE.php PyMsXML http://edwardslab.bmcb.georgetown.edu/software/PyMsXML.html mzwiff supplied by vendor

processing (Fig. **2**). This document controls the order in which data is processed and defines the legal flow of data from one algorithm to another. Each stage may contain multiple services and each service is associated with a job submission form and a results visualization display. The workflow document provides a mechanism that minimizes the impact of changes to the PDP, such as the addition and deployment of new analysis algorithms and services.

When a PDP user sends a job processing request to the client server, the client application consults the workflow for the services available within the corresponding stage. A wrapped SOAP request is sent from the client application to the processing server to invoke the desired service. This request includes input files and parameters for the corresponding data processing algorithm. These input parameters are packed into a XML schema file that is transferred between the two servers. Apache XMLBeans technology is used in translating the XML schema files. Once submitted, the job goes into a queue on the processing server. A background process on the client server regularly checks status of jobs submitted for processing. After the processing server completes the requested job, a wrapped SOAP response is sent back to the client. This response includes the output files and job results which can then be accessed from the web-based interface for visualization. The visualization of each job result may contain dynamically generated charts, data tabular reports, and images generated by the processing server. In addition to reviewing the data processing results online, PDP users can download the results of each specific processed job as a zip file and work locally with alternative or stand-alone data analysis tools and algorithms.

Processing Server

The processing server uses web services to advertise each stage of the PDP workflow as a service. This capability was developed using the open source software Apache Tomcat, Axis and Ant [27]. The processing server is responsible for accepting experimental data and user defined processing parameters from the client application, job scheduling, and data processing using the application and script or algorithm associated with each service. Data integrity checks are also performed to ensure data and results are not altered during transmission or processing.

Currently, the processing server supports applications written in Java and C++ programming languages and data processing scripts written in the R language [28]. Analysis scripts and applications written for other statistical packages (such as MATLAB) can be easily integrated into this system as long as they support command line execution, provide Application Programming Interfaces (API's), or follow a service-oriented architecture.

The processing server has two modes of operation, online and batch. The online mode is responsible for job submissions, data retrieval and job status queries. The batch mode is responsible for executing processing algorithms on the specified data. As new jobs are submitted to the processing server, they are queued and executed in the order they are received. The online and batch modes operate independently ensuring uninterrupted submissions, retrieval and processing. The processing server is stateless in that it does not track the logical flow of data. The server processes data on a per job basis without reference to information regarding whether the

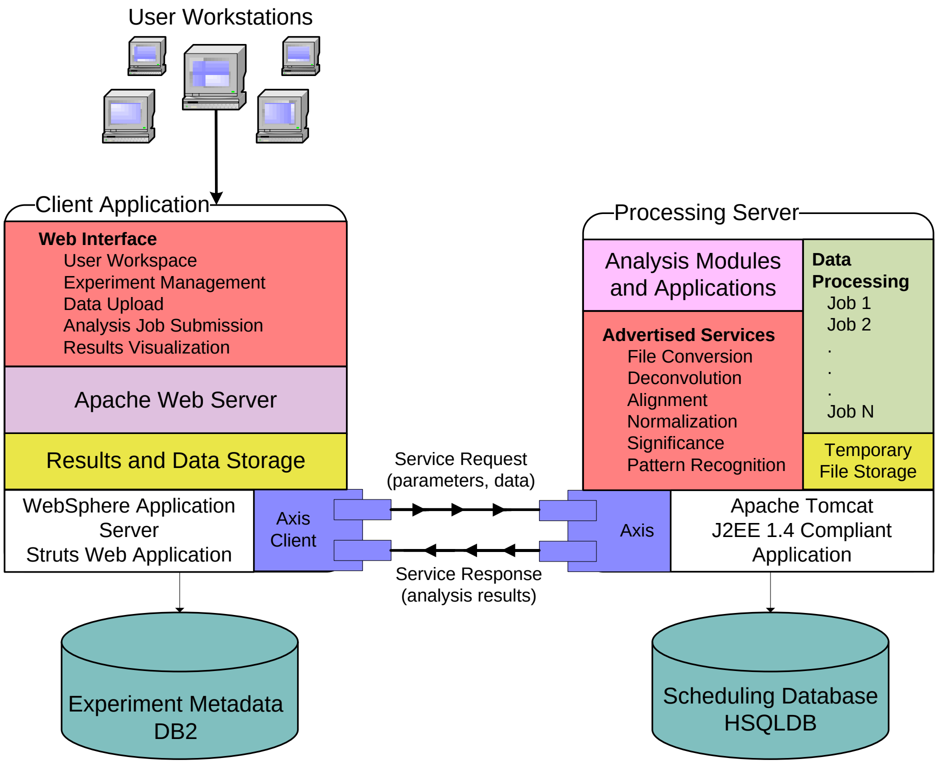

Fig. (2). The PDP architecture. Consists of two distinct components: client application (left) and processing sever (right).

output of a particular analysis module becomes the input for another module. This task of logical piping of data throughout the PDP workflow is the responsibility of the client application. This setup allows for easy transition to multiple processing servers (*e.g.* clusters) as the need for data processing grows.

Data Storage and Security

Due to the large size of biological data sets, a storage solution was implemented to provide redundant and scalable data storage. The server hosting the client application is connected to a storage area network (SAN) where datasets and results are stored. File storage is handled by a clustered, fabric-attached enterprise filer (Network Appliance FAS270), which can scale to several terabytes (TB) of storage as needed and provides fault tolerance in case of disk failure. Data stored on the SAN is only accessible by the server hosting the client application. The PDP is protected by a defense-in-depth security strategy that utilizes authentication, access control, firewalls, antivirus software, and logging to ensure the overall security of the servers as well as the integrity of file storage.

PDP Data Mining Capabilities

The PDP provides a platform for researchers to store, process, analyze and visualize LC-MS, LC-MALDI-TOF and MALDI-TOF datasets. The system supports datasets that range in size from a few kilobytes to several gigabytes. The pipeline uses a configurable workflow document that defines the logical operations permitted on the raw or transformed data. With this document, the system controls the order in which data is processed and the legal piping of data from one algorithm to another. The current version supports proteomics data with the following data operations: file format conversion, spectrum visualization, deconvolution, alignment, normalization, statistical significance testing, and pattern recognition.

Data Conversion

Significant efforts have focused on proteomics data standardization. The mzXML [29, 30] and mzData [31] are now widely accepted and various software packages convert raw instrument data into mzXML or mzData formats [32, 33]. However, not all mass spectrometer data is convertible (*e.g.*, Agilent MSD TOF data). An important feature of our pipeline is the large number of data formats that we can accommodate. Currently, the pipeline can recognize MS datasets in NetCDF, AgilentCsv, mzData, mzXML and ThermalTxt formats.

Spectrum Visualization

To enable the user to manipulate data, the PDP provides various data visualization functionalities for instrument proprietary data formats and for intermediary data. For instance, the user can view raw instrument data by choosing either the total ion chromatogram or the selected ion chromatogram. This enables researchers to directly verify data uploaded to the pipeline.

Spectrum Deconvolution

The purpose of spectral deconvolution is to differentiate signals arising from the real analyte as opposed to signals arising from contaminants or instrumental noise, and to reduce data dimensionality which will benefit downstream statistical analysis. Therefore, spectra deconvolution extracts peak information from thousands of raw mass spectra. The peak information is reported in a simple peak table. The GISTool [34] is a software package with chemical noise filtering, charge state fitting, and de-isotoping for the analysis of complex peptide samples. Overlapping peptide signals in mass spectra are deconvoluted by correlating the observed spectrum with modeled peptide isotopic peak profiles. Isotopic peak profiles for peptides are generated *in silico* from a protein database producing reference model distributions. The GISTool algorithm has been modified to enable the analysis of metabolomics data generated from a LC-MS analytical platform. This new version of the software, called XMass, is included in the PDP and reflects major improvements such as the capability of analyzing data generated from low resolution MS instruments. XMass provides for data deconvolution of overlapping mass spectral peaks, identifies doublets, and calculates the ratio of the doublets.

Peak Alignment

Ideally, the same peptide or metabolite detected on the same analytical system should produce the same signal. For example, for a peptide measured on a LC-MS system, retention time and molecular weight should be the same for this peptide in different samples. However, this may not be the case due to experimental variations. Peak alignment recognizes peaks from the same molecule occurring in different samples from the millions of peaks detected during the course of an experiment. The XAlign software implemented in PDP uses a two-step alignment approach [35]. The first step addresses systematic retention time shift by recognizing and aligning significant peaks. A significant peak refers to a peak that is present in the majority of samples and is the most intense peak in a certain m/z (mass-to-charge) and retention time range. Discrete convolution is used in the second step to align overlapped peaks.

Data Normalization

To allow multi-experiment analyses, it is important to first normalize the data for sample comparison. Normalization attempts to quantitatively filter overall peak intensity variations due to experimental errors such as systematic variable injection volumes loaded onto LC-MS. Several normalization methods have been proposed. One approach is to choose an analysis run as a reference and to sequentially normalize all others relative to this reference [36]. The intensity ratio of each aligned peak pair in reference and sample is calculated. The normalization constant for the sample being considered is then taken as the median of the ratios of intensities for all components between the sample in question and the reference sample. In another method [37], Zhu *et al.* normalized MS data by dividing the intensity at each m/z value by the average intensity of the entire spectrum. The log linear model method [38] assumes primarily multiplicative variation. The maximum likelihood and maximum *a posteriori* estimates for the parameters characterizing the multiplicative variation are derived to compute the scaling factors needed for normalization. We have implemented all of these algorithms in the PDP to allow the user to choose based on the nature of his or her data. By default, the system normalizes the data using the log linear model method for large peak sets and the constant median method for smaller sets.

Statistical Significance Testing

The purpose of a statistical significance test is to identify peptide or metabolite peaks that either make significant contributions to the molecular profile of a sample or distinguish a group of samples from others. Some peaks may be present in multiple sample groups but their intensity might differ between the groups. The quantitative difference indicates the situation in which a peak is present in most (or all) of the samples, but has different intensities between the groups. The standard two-sample t-test and the Wilcox-Mann-Whitney rank test [39] are implemented in PDP to compare the group differences using the 'R' statistical package [28]. The false discovery rate (FDR) based correction method [40] is applied to the significance tests. Currently, the PDP can only perform two-group significance testing using t-test in the case of parametric tests and Wilcoxon–Mann–Whitney for non-parametric tests. Other statistical tests, such as ANOVA, will be added in the upcoming version of the PDP to handle multivariate analysis.

Pattern Recognition

There are many types of pattern recognition approaches that typically fall into two main categories, supervised and unsupervised. Supervised systems require knowledge or data in which the outcome or classification is known ahead of time, so that the system can be trained to recognize and distinguish outcomes. Unsupervised systems cluster or group records without previous knowledge of outcome or classification. The most frequently used unsupervised pattern recognition approach is principal component analysis (PCA). Other unsupervised methods [39] include hierarchical clustering, k-means, and self organizing maps (SOM), linear discriminate analysis (LDA), canonical discriminate analysis (CDA). These methods have been implemented in R statistical package. Researchers can select these tools through the graphic user interface of the PDP. Supervised models, such as SVM, will be implemented in later version of the PDP for data classification.

Experimental Methods (Case Study)

Human breast cancer (n=10) and control (n=10) plasma (100 µg) samples (Asterand, Detroit, MI) were denatured with SDS (Applied Biosystems, Foster City, CA), reduced with tris-(2-carboxyethyl) phosphine (TCEP; Applied Biosystems, Foster City, CA), alkylated with methyl methanethiosulfonate (MMTS; Applied Biosystems, Foster City, CA), and digested with trypsin (5 µl of 1 µg/µl 5% w/w; Applied Biosystems, Foster City, CA). After digestion the samples were applied to a C18 microspin column (Nest Group, Southborough, MA) for buffer exchange, followed by a G25 Sephedex column (Nest Group) for desalting. The resulting peptides were dried down and resuspended in 100 µl 0.01% TFA in water. The peptides were separated on a micro system (1100 Series LC, Agilent, Santa Clara, CA). After injection of 20 µg, the peptides were concentrated on the 300SB-C18 enrichment column and washed with 5% acetonitrile (ACN); 0.01% TFA at flow rate 40 µl/min for 5 minutes. The enrichment column was switched into the micro flow path and further separated with the C-18 reversed phase ZORBAX 300SB-C18 (0.3 mm x 250 mm; Agilent, Santa Clara, CA) analytical column coupled to the electrospray ionization (ESI) source of the time of flight (TOF) mass spectrometer (MSD-TOF, Agilent, Santa Clara, CA). The column was eluted with an 80 min gradient from 5%-35% buffer B (100% acetonitrile, 0.01% TFA) at a rate of 4 µl/min, followed by a 10 minute gradient from 35%-100% buffer B. The column was re-equilibrated with an isocratic flow (5% buffer B) at 4 µl/min. Agilent's ChemStation software was used to control the system. LC-MS chromatograms were acquired in positive ion mode with a capillary voltage of 1850 V and an end plate offset of 500 V. The dry temperature was set at 275 °C. Dry gas flow was maintained at 4 L/min. Acquisition range was 350 – 2200 m/z with a scan speed of 1 scan per second. The raw data (Agilent .WIFF files) were converted to Agilent .csv format using the export function on the TOF instrument and then compressed into a zip file before uploading into PDP.

RESULTS AND DISCUSSION

Differences in gene expression that reflect physiological cellular variations are discernible at both the mRNA and protein levels. As proteins play key functional roles in cellular activities, this class of molecules is highly relevant for characterization of a biological system and as potential targets for drug treatments. The complexity of protein species and the wide dynamic range of protein concentrations within biological systems require extensive sample simplification and appropriate precise analytical methods to evaluate large numbers of proteins (*e.g.*, the LC-MS "proteomics" approach). We employ high resolution chromatography with mass spectrometry (MS) to generate data that enables us to distinguish the unique and significant features between two groups. Our method avoids loss of data due to additional duty cycle time associated with alternative tandem mass spectrometry (MS/MS) data acquisition methods. The PDP was created to handle this MS profiling data. Once the features are identified with our LC-MS protocol, targeted MS/MS is used for identification of the associated peptides and proteins with various programs, including commercial protein database search identification algorithms as discussed below.

To demonstrate the capability of PDP system, we present here a case study of proteomic evaluation of human plasma collected from 10 breast cancer patients and from 10 normal controls in an attempt to identify key protein differences. Briefly, the proteins were digested using trypsin and submitted for LC-MS for profile analysis (Agilent MSD TOF). Tandem MS protein identification was performed with ion trap LC-MS/MS (Agilent XCT plus).

The end product of our current PDP system is a list of significantly regulated LC-MS peaks and associated quality assessment information. For protein identification, we subsequently employ multiple database search algorithms such as Sequest [41], Mascot [42] and X! Tandem [43]. The association of PDP end results with protein identification results is currently done using a separate software package.

PDP Analysis

Spectrum Visualization

Each sample file was converted to an Agilent. csv file, compressed by zipping and then uploaded into the PDP. Zipping the input data decreases the time required to upload the data to the server. This zip file must only include the data files to be uploaded and not include directories or other file types. The file name was used as the sample ID to associate a specific experiment with the data file.

After uploading, each sample file was viewed using the spectral visualization component of the PDP and compared to the chromatogram generated by Agilent's qualitative software MassHunter (Fig. **3**, B and A respectively). It can be seen that the overall intensity of TIC in Fig. (**3B**) is reduced compared to the original TIC displayed in Fig. (**3A**) and the chromatogram in 2500-4000 sec interval are different. These differences were introduced by noise deduction during data conversion using the MassHunter software. We also use the spectrum visualization to quickly sift through all samples to either identify significant outlier samples or roughly identify the difference between multiple samples based on the profile of total ion chromatogram (TIC).

Spectrum Deconvolution

The converted data files (.csv) and an input parameter file (.XML) were sent to XMass for spectrum deconvolution. This process is located in the data processing server. The input parameter file contains information about file format, analyte type, instrument used, data acquisition mode, *etc*. The specific parameters used for our case study were: Agilent .csv file format, peptide analytes, TOF instrument, positive data acquisition mode, 0 molecular weight difference between paired peaks, 0 modifications for each analyte, centralized data, minimum LC peak width of 5 scans, retention

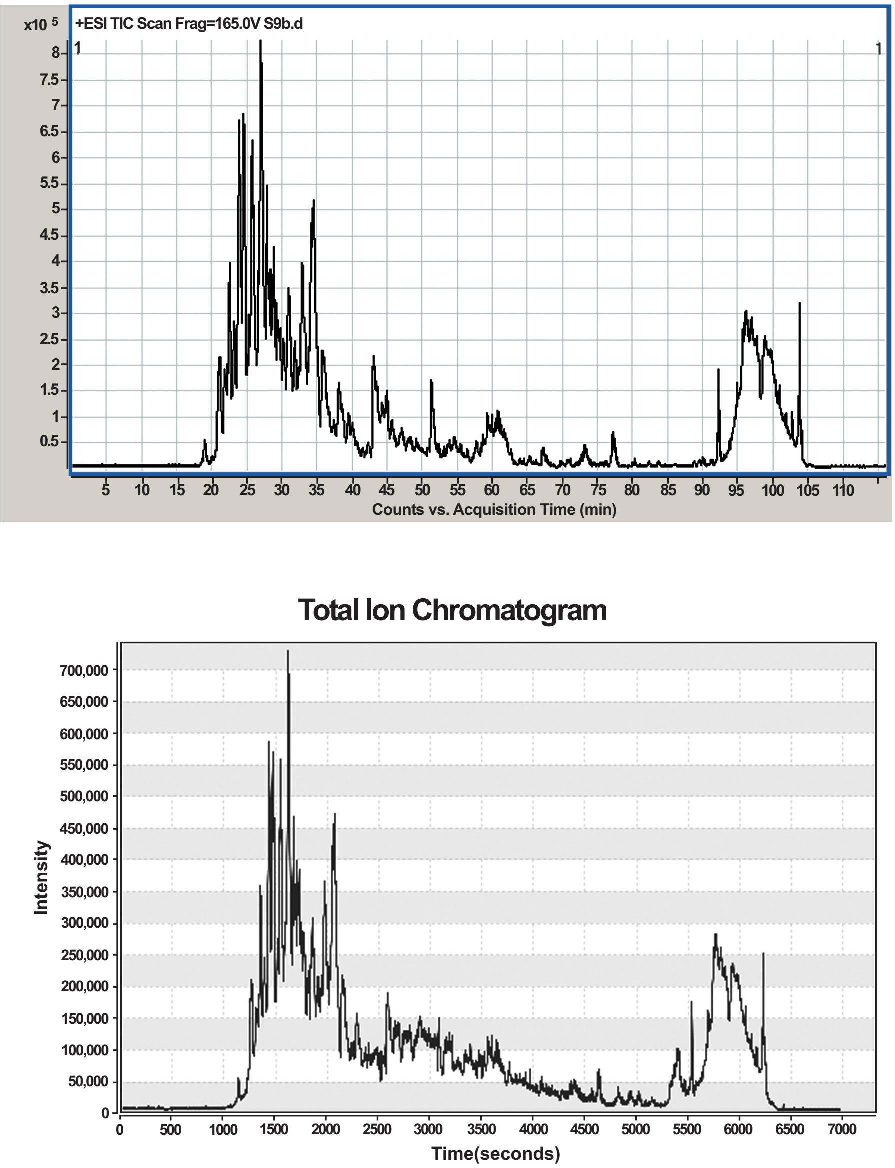

Fig. (3). Total ion chromatograms (TICs) for a single sample in the case study. **A**) TIC for sample S9b created from the Agilent. WIFF file with (Agilent) Qualitative Analysis software. **B**) TIC for the same sample created from a converted and compressed .csv file uploaded to PDP by the XMass software. The chromatograms indicate correct conversion and uploading of the data file.

time from 0 to 110 minutes, and 0.12 m/z variation between isotopes. The first four parameters inform XMass the types of input file, analyte, instrument and data acquisition mode. The fifth and the sixth parameters show that the peptide sample was not stable isotope labeled and therefore, the molecular weight difference and number of stable isotope labeled sites in each peptide are both set to zero. We used Agilent software centralized the raw data during data file conversion. This information is presented in the seventh input parameter of XMass. The minimum LC peak width and m/z variation between isotopic peaks are two important parameters in XMass. The minimum LC peak width is a threshold to determine whether a cluster of MS signals are generated by a peptide or are instrument noise. The m/z variation between isotopic peaks is used to determine whether a cluster of peaks is generated by the same peptide [34]. The XMass software generates an output file for each sample that contained the deconvoluted peaks. We detected approximately 2500 peptide isotopic peaks in each plasma sample. The peak list for each sample (.dlt file) is sent back to the client application and displayed in the "show result" window.

Peak Alignment

The .dlt files produced in the spectral deconvolution step were sent next to the XAlign algorithm [35] in the data processing server. The input parameter file in this step contains information about m/z variation, peak frequency, and retention time variation. The specific parameters used for our case study were 0.07 m/z variation, 60% peak frequency, and 0.5 min retention time variation. The peak frequency refers to the ratio of the number of samples in which a peptide peak is detected and also aligned divided by the total number of samples. This parameter, which in fact is a sample percentile threshold value, is employed to direct the processing sever to generate the selected ion chromatograms for a certain group of peptide isotopic peaks in each sample. The peak alignment service algorithm generated five data files, with the extensions .org, .qcd, qul, vmz, and .vrt, which are sent back to the client for display in the XAlign visualization window. The .org file is the alignment table that contains isotope label, charge state, m/z, retention time and peak intensities for each sample as well as the number of samples in which the peak is identified and the mean intensity for each peak. The .vmz and .vrt files contain information about the m/z and retention time variation. The .qul file contains the D value of the two dimensional Kolmgorov–Smirnov (K–S) tests between every sample pair [44], and .qcd file contains the number of peaks identified in each sample and the number of peaks aligned in each sample. All numeric information recorded in .vmz, .vrt, .qul, and .qcd files is used to assess the quality of the data prior to moving on to the statistical testing using Sprent's equation [39]. Quality control features of the alignment algorithm are described in detail in our previous publication [35]. Fig. (**4**) displays a representative graph generated by the Xalign software. A total of 965 of the detected peaks align between at least two samples, 460 peaks aligned between 10 samples, 20 peaks aligned between all 20 samples and the data quality in all samples passed the quality control tests (satisfactory). The .org file serves as the input file for the next processing step.

Data Normalization

At this point in the workflow the PDP identifies and removes systematic variations caused by sample concentration, measurement conditions, *etc*. The .org file generated from the Peak Alignment step together with an input parameter XML file are sent by the client application to the data normalization algorithms in the data processing server in a seamless manner. The XML file, generated by the client application, contains information about the number of groups, group size, *etc*. The specific input parameters for our case study were as followed: 2 groups with 10 samples in each group. The execution of the data normalization service algorithm generated four .txt data files and four .jpg graph files that are returned to the client and displayed in the "Normalization Visualization" window. The PDP automatically applies multiple normalization algorithms to the aligned data. The various methods can be visualized to determine which method is optimal for the data (Fig. **5**). In addition, the user can visualize the effect of normalization on the most variant peaks in the dataset (*i.e.*, the *extreme sample*). The log-linear normalization method was employed to process the data for

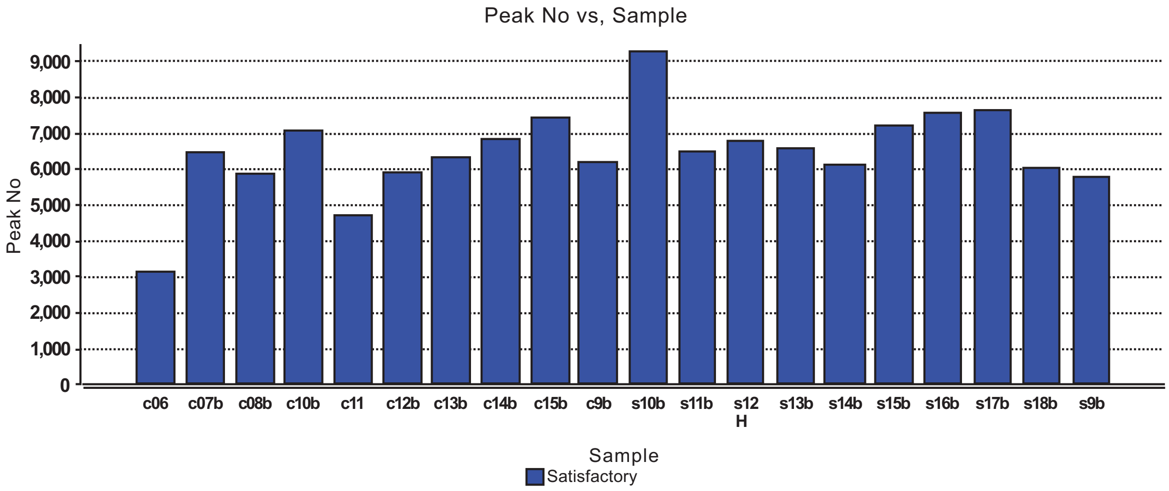

Fig. (4). Peak alignment quality control. Histogram providing the number of peaks aligned in each sample. This readout indicates samples that passed the quality control tests as satisfactory with a blue bar, and the outlier samples (none in this example) as unsatisfactory with a red bar.

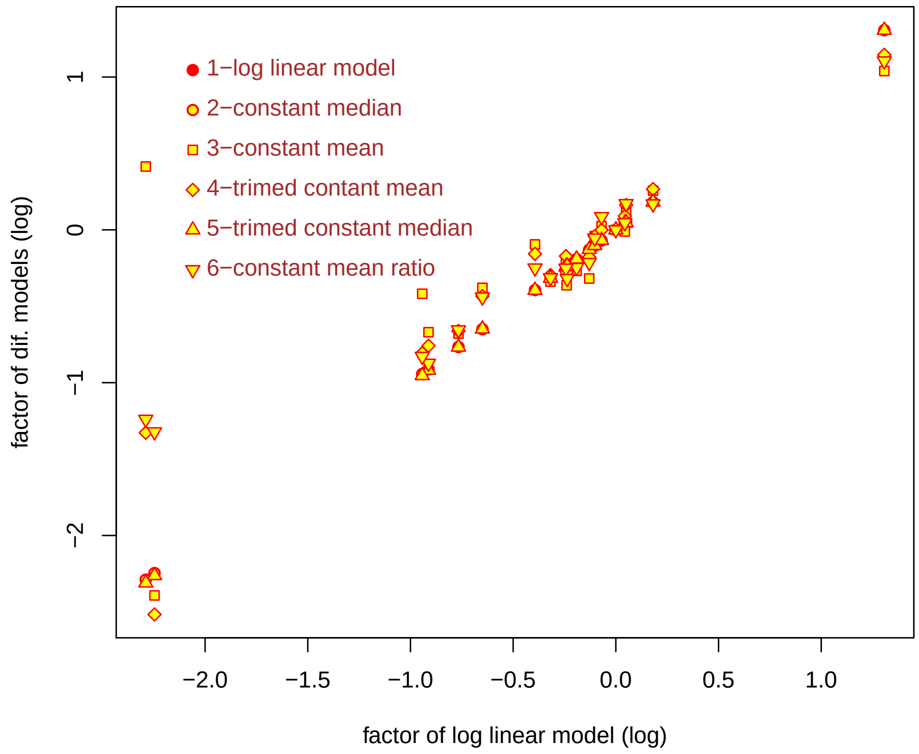

Fig. (5). Graphical output from PDP normalization. Graph showing the variation between the different normalization methods. For our case study, the log linear method was considered the most suitable choice.

our case study. The normalized peak information is saved as a .txt file on the server and used as input for the next two parallel steps, statistical significance evaluation and pattern recognition.

Statistical Significance Evaluation

PDP employs several statistical significance test methods to identify data elements that contribute to the proteomic profile of a sample or that distinguish groups of samples. Methods implemented in the current version of PDP include two-tailed t-test and Mann-Whitney tests [39]. One-way analysis of variance (ANOVA) and the Kolmogorov-Smirnov tests [39] will be incorporated with our next version. In addition, the false discovery rate (FDR) approach, as described by Benjamini & Hochberg [40], can be employed to adjust for multiple tests. The .txt file generated from the previous data normalization step and an input parameter (.XML) file containing information about the number of groups and group size are sent by the client application to the data processing server. In this particular case study, a peak was required to appear in at least 30% of all samples to be identified. To be included in the statistical analysis, the peak must have been detected in at least 65% of the samples in that group. These two parameters can be adjusted by the user. During statistical significance tests, we simply ignored the missed values. The selected statistical significance test service algorithm(s) generate three .jpg graph files and one .sig data file that contains the significance test results. Those files are returned to the client and displayed in the “T-Test Visualization” window (Fig. **6**). There are 965 aligned peptide peaks, of which 120 have been detected as significantly regulated with a p-value < 0.05. The ratio of the peak intensity from control and cancer samples indicates the degree to which the peaks are changing (fold change) and the direction of regulation (*i.e.*, up or down).

Pattern Recognition

The pipeline makes available principal component analysis (PCA), linear discriminate analysis (LDA), and canonical discriminate analysis (CDA) for data clustering. As with statistical significance evaluation, the .txt file generated from the data normalization step and an input parameter (.XML) file are concurrently sent by the client application to the pattern recognition algorithm in the data processing server. The input parameters are the same as those described in the statistical evaluation section. The PDP employs k-nearest neighbor (KNN) approach to impute missed values [39]. This pattern recognition stage of the pipeline generates graph files (.jpg) that are returned to the client application and displayed in the “Pattern Recognition” visualization window (Fig. **7**).

CONCLUSIONS

Proteomics provides for *global* and simultaneous characterization of expressed proteins. To manage these data, we developed a flexible proteomics data analysis pipeline, the Proteome Discovery Pipeline (PDP), with two distinct components: a web-based client application and a processing server. Each component is deployed on a separate server and both use the Simple Object Access Protocol (SOAP) for communication. Data processing and analysis tools are provided as web services that are invoked through the web-based interface and provide the required functionality for each stage. The current version of the PDP has several data processing functionalities including spectrum visualization,

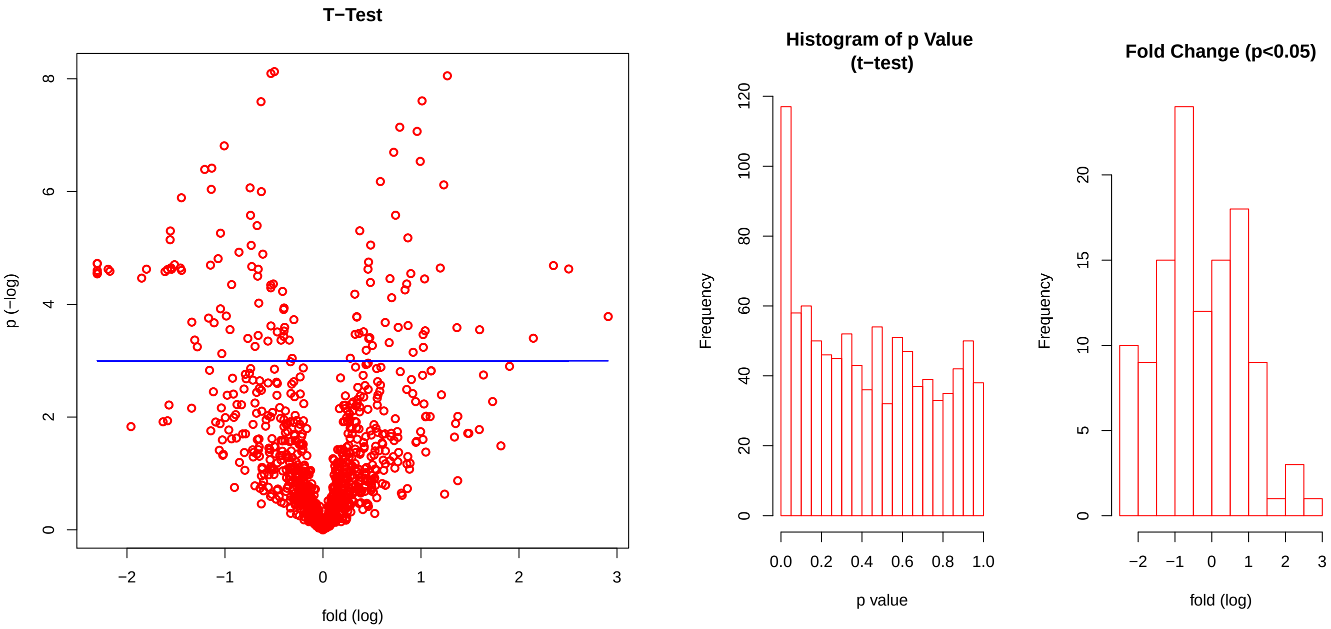

Fig. (6). Graphical output from PDP significance tests. **A**) In this *volcano plot*, peaks that are significantly differentially expressed in cancer vs. control samples are indicated above the blue line (p < 0.05). **B**) Histogram views of the number of differentially expressed peaks (*i.e.*, frequency) for both p value and fold-change differences.

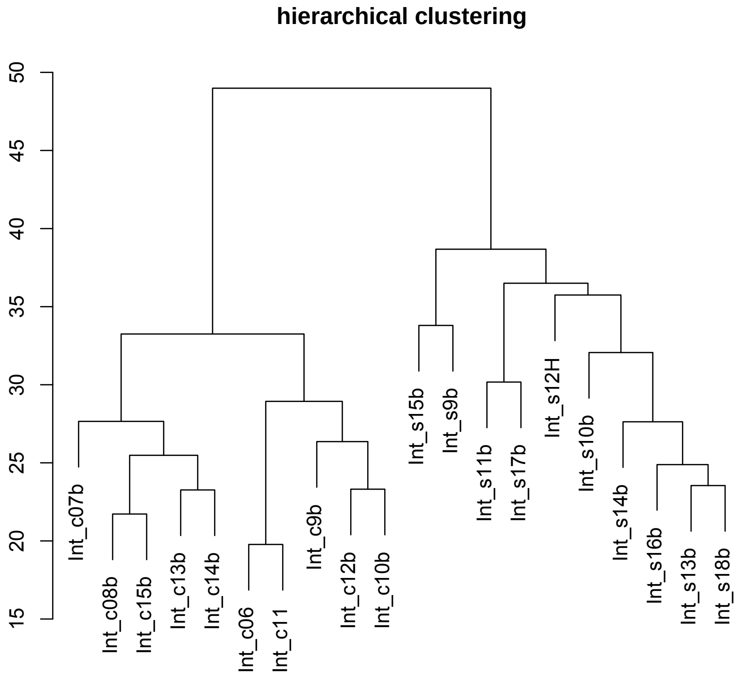

Fig. (7). Pattern recognition in PDP. Samples were subjected to the hierarchical clustering function. The major branches distinguish control samples (left side, samples C06 to C15) from cancer samples (right side, S09 to S18).

deconvolution, alignment, normalization, statistical significance tests, and pattern recognition. Key advantages of PDP include the capability of processing experimental data generated by either label free and/or stable isotope labeling proteomics such as ICAT, GIST or SILAC experiments, the ability to accept multiple experimental data formats, and the flexibility of incorporating multiple algorithms for each data analysis step using web services, which provide the user the option to select the appropriate algorithm based on the nature of the experimental data. Additionally, the PDP provides the platform and capabilities to easily navigate back and forth between the different analysis steps in the workflow. Finally, the PDP facilitates high throughput experiments with high throughput/performance computational capacity due to its scalable distributed architecture as web client and processing servers. This extensible system has been successfully utilized for several protein biomarker discovery projects.

The use of web services in the PDP enables a flexible design and eliminates the need to change the pipeline engine as existing tools are modified, new tools are added, or the workflow is changed. We will provide services to integrate different bioinformatics tools into the PDP if these tools become highly demanded by multiple proteomics laboratories.

ACKNOWLEDGEMENTS

Funding for this project was provided by the Bindley Bioscience Center, eEnterprise Center and the Cyber Center, all at Purdue University's Discovery Park. Extramural funding was provided from NIH awards: 5U24CA126480 (NCI) and 5R33DK070290 (NIDDK).

REFERENCES

[1] Domon, B.; Aebersold, R. Mass spectrometry and protein analysis. *Science*, **2006**, *312* (5771), 212-7.

[2] Lambert, J.P.; Ethier, M.; Smith, J.C.; Figeys, D. Proteomics: from gel based to gel free. *Anal. Chem.*, **2005**, *77* (12), 3771-87.

[3] Good, D.M.; Thongboonkerd, V.; Novak, J.; Bascands, J.L.; Schanstra, J.P.; Coon, J.J.; Dominiczak, A.; Mischak, H. Body fluid proteomics for biomarker discovery: lessons from the past hold the key to success in the future. *J. Proteome Res.*, **2007**, *6* (12), 4549-55.

[4] Asara, J.M.; Zhang, X.; Zheng, B.; Christofk, H.H.; Wu, N.; Cantley, L.C. In-Gel Stable-Isotope Labeling (ISIL): a strategy for mass spectrometry-based relative quantification. *J. Proteome Res.*, **2006**, *5* (1), 155-63.

[5] Qiu, R.; Zhang, X.; Regnier, F.E. A method for the identification of glycoproteins from human serum by a combination of lectin affinity chromatography along with anion exchange and Cu-IMAC selection of tryptic peptides. *J. Chromatogr. B. Anal. Technol. Biomed. Life Sci.*, **2007**, *845* (1), 143-50.

[6] Wang, S.; Zhang, X.; Regnier, F.E. Quantitative proteomics strategy involving the selection of peptides containing both cysteine and histidine from tryptic digests of cell lysates. *J. Chromatogr. A.*, **2002**, *949* (1-2), 153-62.

[7] Link, A.J.; Eng, J.; Schieltz, D.M.; Carmack, E.; Mize, G.J.; Morris, D.R.; Garvik, B.M.; Yates, J.R. 3rd, Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.*, **1999**, *17* (7), 676-82.

[8] Nesvizhskii, A.I.; Vitek, O.; Aebersold, R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods*, **2007**, *4* (10), 787-97.

[9] Hartler, J.; Thallinger, G.G.; Stocker, G.; Sturn, A.; Burkard, T.R.; Korner, E.; Rader, R.; Schmidt, A.; Mechtler, K.; Trajanoski, Z. MASPECTRAS: a platform for management and analysis of proteomics LC-MS/MS data. *BMC Bioinformatics*, **2007**, *8*, 197.

[10] Rauch, A.; Bellew, M.; Eng, J.; Fitzgibbon, M.; Holzman, T.; Hussey, P.; Igra, M.; Maclean, B.; Lin, C.W.; Detter, A.; Fang, R.; Faca, V.; Gafken, P.; Zhang, H.; Whiteaker, J.; States, D.; Hanash, S.; Paulovich, A.; McIntosh, M.W. Computational Proteomics Analysis System (CPAS): an extensible, open-source analytic system for evaluating and publishing proteomic data and high throughput biological experiments. *J. Proteome Res.*, **2006**, *5* (1), 112-21.

[11] Brusniak, M.Y.; Bodenmiller, B.; Campbell, D.; Cooke, K.; Eddes, J.; Garbutt, A.; Lau, H.; Letarte, S.; Mueller, L.N.; Sharma, V.; Vitek, O.; Zhang, N.; Aebersold, R.; Watts, J.D. Corra: Computational framework and tools for LC-MS discovery and targeted mass spectrometry-based proteomics. *BMC Bioinformatics*, **2008**, *9*, 542.

[12] Kohlbacher, O.; Reinert, K.; Gropl, C.; Lange, E.; Pfeifer, N.; Schulz-Trieglaff, O.; Sturm, M. TOPP--the OpenMS proteomics pipeline. *Bioinformatics*, **2007**, *23* (2), e191-7.

[13] Mueller, L.N.; Rinner, O.; Schmidt, A.; Letarte, S.; Bodenmiller, B.; Brusniak, M.Y.; Vitek, O.; Aebersold, R.; Muller, M. SuperHirn - a novel tool for high resolution LC-MS-based peptide protein profiling. *Proteomics*, **2007**, *7* (19), 3470-80.

[14] de Groot, J.C.; Fiers, M.W.; van Ham, R.C.; America, A.H. Post alignment clustering procedure for comparative quantitative proteomics LC-MS data. *Proteomics*, **2008**, *8* (1), 32-6.

[15] Mueller, L.N.; Brusniak, M.Y.; Mani, D.R.; Aebersold, R. An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data. *J. Proteome Res.*, **2008**, *7* (1), 51-61.

[16] Braisted, J.C.; Kuntumalla, S.; Vogel, C.; Marcotte, E.M.; Rodrigues, A.R.; Wang, R.; Huang, S.T.; Ferlanti, E.S.; Saeed, A.I.; Fleischmann, R.D.; Peterson, S.N.; Pieper, R. The APEX Quantitative Proteomics Tool: generating protein quantitation estimates from LC-MS/MS proteomics results. *BMC Bioinformatics*, **2008**, *9*, 529.

[17] Saito, A.; Nagasaki, M.; Oyama, M.; Kozuka-Hata, H.; Semba, K.; Sugano, S.; Yamamoto, T.; Miyano, S. AYUMS: an algorithm for completely automatic quantitation based on LC-MS/MS proteome data and its application to the analysis of signal transduction. *BMC Bioinformatics*, **2007**, *8*, 15.

[18] Monroe, M.E.; Shaw, J.L.; Daly, D.S.; Adkins, J.N.; Smith, R.D. MASIC: a software program for fast quantitation and flexible visualization of chromatographic profiles from detected LC-MS(/MS) features. *Comput. Biol. Chem.*, **2008**, *32* (3), 215-7.

[19] Polpitiya, A.D.; Qian, W.J.; Jaitly, N.; Petyuk, V.A.; Adkins, J.N.; Camp, D.G., 2nd; Anderson, G.A.; Smith, R.D. DAnTE: a statistical tool for quantitative analysis of -omics data. *Bioinformatics*, **2008**, *24* (13), 1556-8.

[20] Thiele H, G.J.; Hufnagel P.; Korting G.; Bluggel M. Managing proteomics data: from generation and data warehousing to central data repository. *J. Proteomics Bioinform.*, **2008**, *1* (9), 485-507.

[21] Elias, J.E.; Haas, W.; Faherty, B.K.; Gygi, S.P. Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nat. Methods*, **2005**, *2* (9), 667-75.

[22] Kapp, E.A.; Schutz, F.; Connolly, L.M.; Chakel, J.A.; Meza, J.E.; Miller, C.A.; Fenyo, D.; Eng, J.K.; Adkins, J.N.; Omenn, G.S.; Simpson, R.J. An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis. *Proteomics*, **2005**, *5* (13), 3475-90.

[23] Sana, T.R.; Waddell, K.; Fischer, S.M. A sample extraction and chromatographic strategy for increasing LC/MS detection coverage of the erythrocyte metabolome. *J. Chromatogr. B. Anal. Technol. Biomed. Life Sci.*, **2008**, *871* (2), 314-21.

[24] W3C. http://www.w3.org/2002/ws (accessed October 16, **2009**).

[25] Gudgin, M.; Hadley, M.; Mendelsohn, N.; Moreau, J.-J.; Nielsen, H.F.; Karmarkar, A.; Lafon, Y. SOAP Version 1.2 Part 1: Messaging Framework (Second Edition), W3C Recommendation. http://www.w3.org/TR/2007/REC-soap12-part1-20070427/ (accessed October 16, **2009**).

[26] Bray, T.; Paoli, J.; Sperberg-McQueen, C.M.; Eve Maler, S.M.; Yergeau, F. Extensible Markup Language (XML) 1.0 (Fifth Edition), W3C Recommendation. http://www.w3.org/TR/2008/REC-xml-20081126/ (accessed October 16, **2009**).

[27] The Apache Software Foundation. http:/www.apache.org (accessed October 16, **2009**).

[28] The R Project for Statistical Computing. http://www.r-project.org (accessed October 16, **2009**).

[29] Lin, S.M.; Zhu, L.; Winter, A.Q.; Sasinowski, M.; Kibbe, W.A. What is mzXML good for? *Expert Rev. Proteom.*, **2005**, *2* (6), 839-45.

[30] Pedrioli, P.G.; Eng, J.K.; Hubley, R.; Vogelzang, M.; Deutsch, E.W.; Raught, B.; Pratt, B.; Nilsson, E.; Angeletti, R.H.; Apweiler, R.; Cheung, K.; Costello, C.E.; Hermjakob, H.; Huang, S.; Julian, R.K.; Kapp, E.; McComb, M.E.; Oliver, S.G.; Omenn, G.; Paton, N.W.; Simpson, R.; Smith, R.; Taylor, C.F.; Zhu, W.; Aebersold, R. A common open representation of mass spectrometry data and its application to proteomics research. *Nat. Biotechnol.*, **2004**, *22* (11), 1459-66.

[31] mzData. http://www.psidev.info/index.php?q=node/80#mzdata (accessed October 16, **2009**).

[32] mzXML Converters. http://tools.proteomecenter.org/wiki/index.php?title=Formats:mzXML (accessed October 16, **2009**).

[33] CompassXport. http://www.brukerdaltonics.com (accessed October 16, **2009**).

[34] Zhang, X.; Hines, W.; Adamec, J.; Asara, J.M.; Naylor, S.; Regnier, F.E. An automated method for the analysis of stable isotope labeling data in proteomics. *J. Am. Soc. Mass Spectrom.*, **2005**, *16* (7), 1181-91.

[35] Zhang, X.; Asara, J.M.; Adamec, J.; Ouzzani, M.; Elmagarmid, A.K. Data pre-processing in liquid chromatography-mass

spectrometry-based proteomics. *Bioinformatics*, **2005**, *21* (21), 4054-9.

[36] Wang, W.; Zhou, H.; Lin, H.; Roy, S.; Shaler, T.A.; Hill, L.R.; Norton, S.; Kumar, P.; Anderle, M.; Becker, C.H. Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Anal. Chem.*, **2003**, *75* (18), 4818-26.

[37] Zhu, W.; Wang, X.; Ma, Y.; Rao, M.; Glimm, J.; Kovach, J.S. Detection of cancer-specific markers amid massive mass spectral data. *Proc. Natl. Acad. Sci. USA*, **2003**, *100* (25), 14666-71.

[38] Hartemink A.J.; G.D.; Jaakola T.S.; Young R.A. Maximum likelihood estimation of optimal scaling factors for expression array normalization. *Proc SPIE*, **2001**, *4266*, 132-140.

[39] Sheskin, D. *Handbook of Parametric and Nanoparametric Statistical Procedures*. Chapman & Hall/CRC, Washington DC: **2000**.

[40] Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B.*, **1995**, (57), 289-300.

[41] Eng, J.; McCormack, A.; Yates, J.R. 3rd, An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom*, **1994**, *20*, 3551-3567.

[42] Perkins, D.N.; Pappin, D.J.; Creasy, D.M.; Cottrell, J.S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **1999**, *20* (18), 3551-67.

[43] Craig, R.; Beavis, R.C. A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid. Commun. Mass. Spectrom.*, **2003**, *17* (20), 2310-6.

[44] Press, W.H.; Teukolsky, S.A.; Vettering, W.T.; Flannery, B.P. *Numercial Receipes in C++, The Art of Scientific Computing, 2nd ed*. Cambridge University Press, Cambridge, UK: **2002**; pp. 650-654.

Received: November 30, 2009 Revised: February 25, 2010 Accepted: March 02, 2010

© Riley *et al*.; Licensee *Bentham Open*.

This is an open access article licensed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/) which permits unrestricted, non-commercial use, distribution and reproduction in any medium, provided the work is properly cited.