BMC Bioinformatics

Proceedings

An automated proteomic data analysis workflow for mass spectrometry

Ken Pendarvis[†1,4], Ranjit Kumar*[†1,2], Shane C Burgess[1,2,3,4] and Bindu Nanduri[1,2]

Address: [1]Institute for Digital Biology, Mississippi State University, Mississippi State, MS 39762, USA, [2]College of Veterinary Medicine, Mississippi State University, Mississippi State, MS 39762, USA, [3]Mississippi Agriculture and Forestry Experiment Station, Mississippi State University, Mississippi State, MS 39762, USA and [4]MSU Life Sciences and Biotechnology Institute, Mississippi State University, Mississippi State, MS 39762, USA

E-mail: Ken Pendarvis - kpendarvis@mafes.msstate.edu; Ranjit Kumar* - rkumar@cvm.msstate.edu; Shane C Burgess - burgess@cvm.msstate.edu; Bindu Nanduri - bnanduri@cvm.msstate.edu
*Corresponding author †Equal contributors

from Sixth Annual MCBIOS Conference. Transformational Bioinformatics: Delivering Value from Genomes Starkville, MS, USA 20–21 February 2009

Published: 08 October 2009

BMC Bioinformatics 2009, **10**(Suppl 11):S17 doi: 10.1186/1471-2105-10-S11-S17

This article is available from: http://www.biomedcentral.com/1471-2105/10/S11/S17

Abstract

Background: Mass spectrometry-based protein identification methods are fundamental to proteomics. Biological experiments are usually performed in replicates and proteomic analyses generate huge datasets which need to be integrated and quantitatively analyzed. The Sequest™ search algorithm is a commonly used algorithm for identifying peptides and proteins from two dimensional liquid chromatography electrospray ionization tandem mass spectrometry (2-D LC ESI MS^2) data. A number of proteomic pipelines that facilitate high throughput 'post data acquisition analysis' are described in the literature. However, these pipelines need to be updated to accommodate the rapidly evolving data analysis methods. Here, we describe a proteomic data analysis pipeline that specifically addresses two main issues pertinent to protein identification and differential expression analysis: 1) estimation of the probability of peptide and protein identifications and 2) non-parametric statistics for protein differential expression analysis. Our proteomic analysis workflow analyzes replicate datasets from a single experimental paradigm to generate a list of identified proteins with their probabilities and significant changes in protein expression using parametric and non-parametric statistics.

Results: The input for our workflow is Bioworks™ 3.2 Sequest (or a later version, including cluster) output in XML format. We use a decoy database approach to assign probability to peptide identifications. The user has the option to select "quality thresholds" on peptide identifications based on the P value. We also estimate probability for protein identification. Proteins identified with peptides at a user-specified threshold value from biological experiments are grouped as either control or treatment for further analysis in ProtQuant. ProtQuant utilizes a parametric (ANOVA) method, for calculating differences in protein expression based on the quantitative measure

ΣXcorr. Alternatively ProtQuant output can be further processed using non-parametric Monte-Carlo resampling statistics to calculate P values for differential expression. Correction for multiple testing of ANOVA and resampling P values is done using Benjamini and Hochberg's method. The results of these statistical analyses are then combined into a single output file containing a comprehensive protein list with probabilities and differential expression analysis, associated P values, and resampling statistics.

Conclusion: For biologists carrying out proteomics by mass spectrometry, our workflow facilitates automated, easy to use analyses of Bioworks (3.2 or later versions) data. All the methods used in the workflow are peer-reviewed and as such the results of our workflow are compliant with proteomic data submission guidelines to public proteomic data repositories including PRIDE. Our workflow is a necessary intermediate step that is required to link proteomics data to biological knowledge for generating testable hypotheses.

Introduction

Recent advances in genome sequencing projects have facilitated the global analysis of proteins ("proteomics") in order to study their role in health and disease. Proteomic datasets may be generated by coupling nano-flow technology with high-speed, high resolution mass spectrometers and these have generated immensely complex and very large mass spectral datasets. Analyzing these huge datasets by hand is a daunting, inefficient, and error-prone task, hence the need for an automated data analysis pipelines.

Multidimensional Protein Identification Technology (MudPIT) [1] followed by database searching is commonly used to identify proteins from a biological sample. Biological problems addressed by proteomics often include comparing two different conditions, e.g. normal versus treatment. For comparative proteomics, there is a need to determine which subset of proteins is differentially expressed (DE) at a defined statistical threshold. Sample preparation for proteomics includes total protein isolation from a target biological sample and digestion of these proteins using proteases like trypsin to generate a complex mixture of peptides that then need to be deconvoluted and analyzed by mass spectrometry. One method to reduce the complexity of peptides is separation based on their charge and hydrophobicity using two-dimensional liquid chromatography (2D-LC) before the peptides enter the mass spectrometer for MS/MS analysis.

The flow rates required to separate peptides are in the nanoliter to microliter per minute range and mass spectrometers must collect data for an extended amount of time, often for many hours. The resulting data sets can contain 10s to hundreds of thousands of mass spectra, which must then be searched against a protein database to identify the peptides and thus the proteins. The protein database is in silico digested with a protease (used for sample preparation) to generate database of peptides and their theoretical spectra that can be matched with the experimental spectra collected by mass spectrometry. Several search algorithms are described in literature for database searching, including Sequest [2], MASCOT [3], and X!Tandem [4] which match experimental mass spectra to theoretical spectra derived from a protein database. Sequest is a widely used searched algorithm and our proteomics workflow is designed to analyze Sequest search results. Sequest computes a cross correlation (Xcorr) function to assess the quality of peptide spectra matches. The better the match between an experimental peptide mass spectrum and its database counterpart, the higher the Xcorr will be. Sequest also computes ΔCn, a normalized score calculated from XCorr difference between the best peptide match and the second best match. ΔCn is dependent on database size, search parameters, and sequence homologies. While both XCorr and ΔCn have been used widely in the past for filtering search results [5-8] they provide little information for distinguishing correct peptide assignments from false positives. To get the most meaningful biological data from proteomics or any high throughput experiment it is necessary to reduce the false discovery rate. Decoy database search methods for estimating probabilities for peptide identifications are described in literature [9,10]. However, open source computational tools that automate this estimation are not readily available.

Beyond the identification of peptides and proteins at acceptable statistical thresholds, for expression proteomics the end user requires computational tools for differential protein expression. Label free protein quantification methods determine relative protein abundances directly from high throughput proteomic analyses with out labeling techniques using sampling statistics like spectral counting [11], number of peptides [12], and ΣXcorr [13]. We developed ProtQuant, a java

based tool for label free quantification that uses a spectral counting method with increased specificity based on ΣXCorr. However, ProtQuant computes the statistical significance of differential protein expression using parametric statistics (ANOVA) assuming that the distribution of the control and treatment datasets closely approximates a normal distribution. However, this assumption may not be valid for shotgun proteomics due to either the biology under investigation or due to small sample sizes common to proteomic studies resulting in type I errors (i.e. increased false positive significance rate). Computer intensive distribution-free statistics offer a solution to this problem and we have applied random resampling with replacement to determine statistically significant differences in protein expression from ESI MS^2 data [14].

A recurring theme in high-throughput biology is that collecting orthogonal evidence for biology under investigation using complementary data analysis platforms could reduce the noise and identify true biological effects. For example, microarray differential expression analysis is often complemented by quantitative RT-PCR. Matching mass spectra using two different algorithms like Sequest and Mascot often generates a list of proteins that overlap but also proteins uniquely identified by each method. Likewise given enough computational resources and automated data analysis tools, biologists could evaluate differential protein expression using different statistical tests to identify a core set of differences that could represent true biological changes in expression. Furthermore, proteomic analysis workflows also require corrections for multiple testing to reduce false positive identifications of significant DE based on a single P value cutoff.

Here we describe a computational pipeline that automates the data analysis workflow from assigning probabilities for peptide identification using decoy database approach to statistical evaluation of protein DE using ANOVA and resampling statistics, with subsequent correction for multiple testing using Benjamini and Hochberg's method [15]. This integrated work flow (Figure 1) combines some of our open source software tools like ProtQuant and additional scripts to generate an output that has a list of proteins from a biological sample together with peptide and protein probabilities. Where the experimental design includes comparative proteomics, P values for protein DE adjusted for multiple testing are given for ANOVA based and resampling based (optional) methods.

To illustrate the functionality of our proteomics workflow we used the *Edwardsiella ictaluri* response to iron restriction using 2,2-dipyridyl (DP) iron chelator.

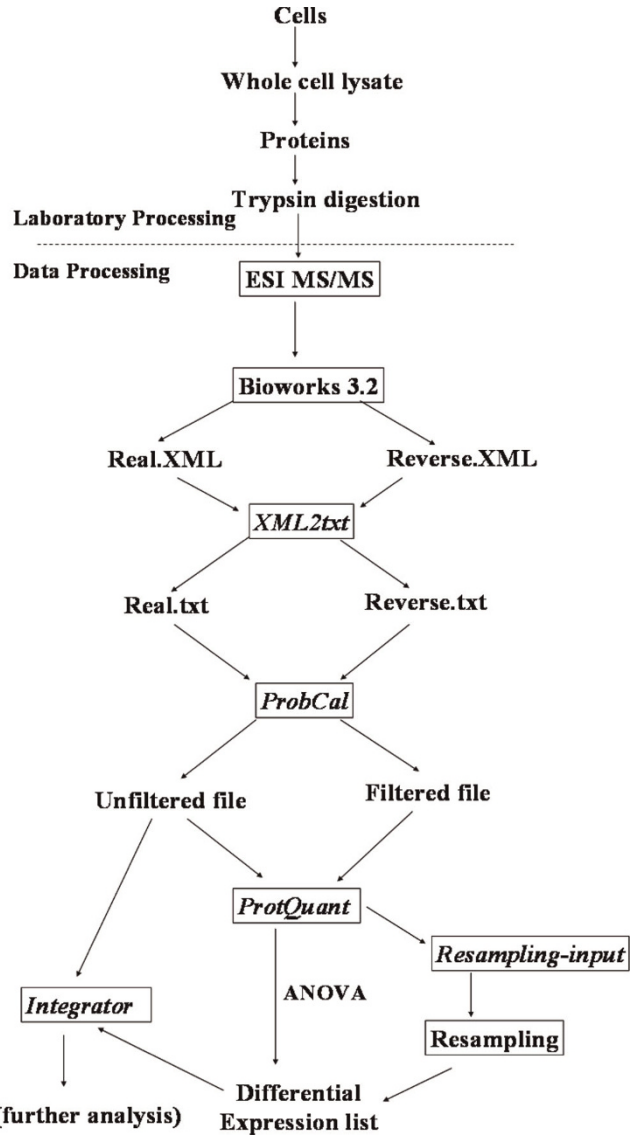

Figure 1
Proteomics data analysis workflow.

E. ictaluri cultures were grown in triplicate and outer membrane proteins were isolated. Mass spectrometry and Sequest searches with a protein and reversed-protein database were done as previously described [14]. SEQUEST results were processed using the tools and scripts described in our workflow.

Results

Our proteomics workflow starts with SEQUEST search results in XML format from Bioworks 3.2 browser for both protein and reverse database searches. We chose the XML format as a standard format for Bioworks output as it overcomes the 65536 row file size limit for some versions of Microsoft Excel spreadsheets.

When exporting Bioworks 3.2 search results, we recommend that the user does not apply any filters. However, due to the virtual memory constraints imposed by computer desktops, if exporting without filters is not practical, we suggest applying minimal filters for peptide charge state. However, the end users need to be aware that if the peptide filters are set too high, many positive matches may be lost. We created a java script named XML2TXT to quickly convert the XML output files to tab delimited text files, which are used by other scripts and can be opened in Excel/notepad for viewing.

Once the real and reverse unfiltered data files are formatted properly using XML2TXT, they can be processed by ProbCal. ProbCal is a set of PERLscripts that automate the estimation of peptide probabilities using search results from a protein and a decoy database. A t-score is obtained for each Xcorr and ΔCn pair from the reverse search results and based on this score a P value is calculated for peptides identified from protein database. The results can then be filtered using a probability cutoff, typically $p \leq 0.05$.

Individual peptide probabilities are further utilized to calculate protein probability using published methods [16]. Another subsidiary script ProbCal-filter uses the peptide probabilities to filter low quality peptides from being included in further analysis. ProbCal can be run from the command prompt, with the names of the real and reverse databases as arguments. For each real/reverse dataset pair a single tab-delimited text file is created with a column containing calculated probabilities for each protein and its associated peptides (Figure 2). We used ProbCal and ProbCal-filter to filter our *E. ictaluri* data with a peptide P value cutoff for protein identification of <0.05 (Figure 3). If differential expression is not the goal of the researcher, then the analysis is complete after ProbCal, otherwise the data is now ready to be processed by ProtQuant. Processing *E. ictaluri* datasets with ProbCal identified 3482 proteins from the normal growth condition (iron replete) and 3437 proteins from growth in the presence of iron-chelator DP at $P \leq 0.05$ for peptide probability. The probability of a protein identification being incorrect was ≤ 0.030 for all identified proteins in the control dataset and 0.038 for proteins identified in DP dataset.

	reference	consensus_score	Sf	unified_score		coverage	pI
	file	sequence	deltamass	charge	P value	xcorr	deltacn
686	NT01EI1569 conser	56.1	0.57	0	1.01E-29	27	9.2
	E_C1_00, 950	LMVTDM@PFQPLK	0.51	3	2.37E-01	1.014	0.026
	E_C1_00, 950	LMVTDM@PFQPLK	0.51	3	2.37E-01	1.014	0.026
	E_C1_20, 2587	FGRLIEYIPLSVTLGF	1.21	3	4.40E-01	1.417	0.198
	E_C1_64, 1095	LSWHSLSALMPAAF	0.71	3	2.85E-01	1.439	0.235
	E_C1_64, 1095	LSWHSLSALMPAAF	0.71	3	2.85E-01	1.439	0.235
	E_C1_00, 518	ARVVPLEGRM@AFF	1	3	1.22E-07	1.247	1.1
	E_C1_00, 1022	LGLRLPGHLPALLAG	0.61	3	9.02E-08	1.094	1.1
	E_C1_10, 647	M@TRISEVSHDDTR	0.46	2	4.34E-07	1.655	1.1
	E_C1_10, 1123	LSWHSLSALM#PAA	0.72	3	7.13E-08	1.048	1.1
	E_C1_64, 1095	LSWHSLSALMPAAF	0.71	3	1.21E-07	1.223	1.1
	E_C1_64, 1095	LSWHSLSALMPAAF	0.71	3	1.21E-07	1.223	1.1
687	NT01EI2453 DNA ir	56.1	0.36	0	4.37E-18	17.1	10.9
	E_C1_00, 948	M@GLPTALSYLAC*	1.36	3	5.11E-02	1.002	0.002
	E_C1_00, 948	M@GLPTALSYLAC*	1.36	3	5.11E-02	1.002	0.002
	E_C1_20, 378	RGLHTYWQGLR	2.3	3	6.14E-01	1.238	0.034
	E_C1_10, 683	MLLHWNSLRFPFAA	1.65	2	1.18E-02	1.271	0.581
	E_C1_15, 220	AALMVTLYALWR	2.32	3	4.23E-01	1.013	0.136
	E_C1_35, 1115	WQADVM#PRALHG	1.71	3	2.37E-01	1.006	0.21
	E_C1_10, 292	FPFAAGQR	0.93	2	4.21E-06	1.222	1.1
	E_C1_15, 836	QTLSQAYPQAWVAS	0.55	3	2.80E-08	1.652	1.1
688	NT01EI3270 protein	56.1	0.38	0	4.42E-09	53.9	6.6

Figure 2
Sample ProbCal output showing protein and peptide probabilities.

	reference file	consensus_score sequence	Sf deltamass	unified_score charge	coverage xcorr	pI deltacn	weight sp	accession rsp	ions	count	tic
685	NT01EI0006 conserv	56.1	0.38	0	22.9	6.2	47594	NT01EI0006			
	E_C1_10, 315	WALAFIGDTLDLDG	2.34	3	1.664	0.003	571.7	2	19/56	0	9609
686	NT01EI1569 conserv	56.1	0.57	0	27	9.2	60539.8	NT01EI1569			
	E_C1_00, 518	ARVVPLEGRM@AF	1	3	1.247	1.1	140.3	46	14/80	0	8402
	E_C1_00, 1022	LGLRLPGHLPALLA(	0.61	3	1.094	1.1	216.8	1	22/136	0	8605
	E_C1_10, 647	M@TRISEVSHDDTF	0.46	2	1.655	1.1	336.5	2	12/24	0	5506
	E_C1_10, 1123	LSWHSLSALM#PAA	0.72	3	1.048	1.1	180.3	27	18/148	0	12114
	E_C1_64, 1095	LSWHSLSALMPAAF	0.71	3	1.223	1.1	188.1	6	20/144	1	9037
	E_C1_64, 1095	LSWHSLSALMPAAF	0.71	3	1.223	1.1	188.1	6	20/144	1	0
687	NT01EI2453 DNA int	56.1	0.36	0	17.1	10.9	82687.4	NT01EI2453			
	E_C1_10, 683	MLLHWNSLRFPFAA	1.65	2	1.271	0.581	172.4	4	9/1	0	8496
	E_C1_10, 292	FPFAAGQR	0.93	2	1.222	1.1	180.9	2	8/14	0	2662
	E_C1_15, 836	QTLSQAYPQAWVA	0.55	3	1.652	1.1	553.9	2	23/88	0	7918
688	NT01EI3270 protein /	56.1	0.38	0	53.9	6.6	21457.8	NT01EI3270			
	E_C1_15, 982	QDGAIQLNESLCVG	0.45	3	1.242	1.1	162.4	8	20/128	0	7793
	E_C1_40, 2667	FVIADPRLC*IGC*NT	0.03	3	1.05	1.1	114.6	3	17/128	0	6437
689	NT01EI0815 conserv	56.1	0.28	0	13.6	9.5	42247.6	NT01EI0815			
	E_C1_10, 342	EDM@SM@QAIRQ'	1.82	3	1.61	0	191	11	14/64	1	6340
	E_C1_10, 342	EDM@SM@QAIRQ'	1.82	3	1.61	0	191	11	14/64	1	0
	E_C1_25, 903	YAVCGLLVALMAG(	1.05	3	1.148	1.1	212.1	5	19/100	0	7960
	E_C1_90, 531	GRWEKYAVC*GLL\	1.5	3	1.01	1.1	328.1	8	18/84	0	10586
690	NT01EI1819 peptida	56.1	0.29	0	25.7	5.4	44583.6	NT01EI1819			
	E_C1_10, 751	QGNTPHGEIRVAFT	1.75	3	1.162	1.1	69.2	81	14/80	0	3959

Figure 3
***E. ictaluri* data after being filtered by ProbCal-filter with P < 0.05**.

The next step in our proteomics workflow after filtering the initial search results is ProtQuant. ProtQuant, written in Java, is installed using a self extracting executable file downloadable form our AgBase website http://agbase.msstate.edu/. ProtQuant has a graphical user interface (Figure 4) for choosing control and treatment files and accepts the output files directly from ProbCal to perform DE analysis by ANOVA. To fill in the "missing Xcorr values", ProtQuant also requires the corresponding original unfiltered Sequest XML output for each dataset that is analyzed [13]. The built-in XML conversion tool in ProtQuant converts XML files to .txt files.

ProtQuant has several modes of operation; it can analyze replicates as pairs or as pooled replicates and generate a single output file. A simple check-box can be selected to activate the function to fill in missing Xcorr values. Once the appropriate input files are selected as either controls or treatments and the output directory is specified, clicking "Go" will start the differential expression analysis. We chose to analyze our *E. ictaluri* results from ProbCal in ProtQuant as control (iron replete) versus iron restricted (DP). The output from ProtQuant is a text file containing a list of proteins present in the combined replicates for control and treatment datasets with an ANOVA P value in the last column for protein DE (Figure 5). ProtQuant computes the statistical significance of DE for proteins using one-way ANOVA. This method requires at least 3 peptides for each protein from the combination of the control and treatment to calculate P value. Using a custom Perl script process_ProtQuant we do the correction for multiple testing based on the published method of Benjamini and

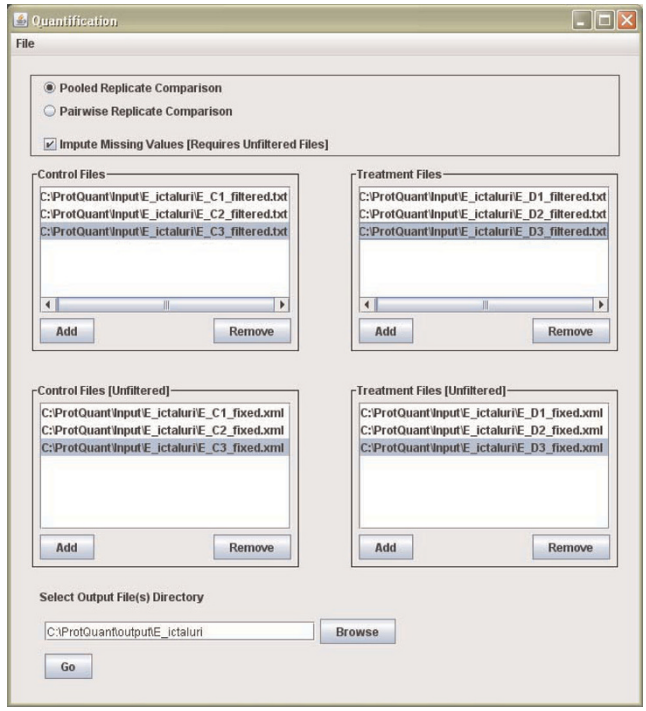

Figure 4
ProtQuant graphical user interface.

ID	Description		Control Non-Filled	Filled Sum	Num Of Peptides		Treatment Non-Filled	Filled Sum	Num Of Peptides		ANOVA
	Sequence	Charge	Sum XCORR	Xcorr	XCORR	Filled	Sum XCORR	XCORR	XCORR	Filled	P value
NT01EI0735	NT01EI0735 cell division pr		25.777	19.747	21		4.869	4.869	4		5.14E-05
	DALYLALAFGLAM	3			0	0			1.071	0	
	DWVM#GAGEADS	2			0	0			1.441	0	
	FSLSLTGGVLR	2			1.04	0			0	0	
	GLTLPLISYGGSSI	3			0	0			1.141	0	
	GLTLPLISYGGSSI	3			1.012	0			0	0	
	IDYETRQANAQAV	3			1.222	0			0	0	
	IPM@EFWQRWSN	3			1.091	0			0	0	
	IQPAEFSKLSLFC*	2			1.339	0			0	0	
	LADDPFLFAKRDA	3			1.363	0			0	0	
	LALPRPRLPHLR	3			1.2	0			0	0	
	LPHLRLPRFSLSLT	3			1.207	0			0	0	
	LWQFLAIIC*SGIFA	3			0	0			1.216	0	
	LWQFLAIIC*SGIFA	3			1.321	0			0	0	
	LWQFLAIIC*SGIFA	3			1.133	0			0	0	
	PMGVMVVLAVLLL	3			1.794	0			0	0	
	RVTSFWNPWADP	3			1.086	0			0	0	
	TLLWLTFGLAAM@	3			1.146	0			0	0	
	TLLWLTFGLAAM@	3			1.029	0			0	0	
	WSNAM#LLLSVAM	3			1.079	0			0	0	
	WSNAM#LLLSVAM	3			1.685	0			0	0	

Figure 5
Sample ProtQuant output.

Hochberg. Another perl script add_protein_prob combines ProbCal and ProtQuant outputs to generate a master output file with proteins identified from control and treatment datasets and also additional information for each protein including: numbers of peptides used for protein identification, ΣXcorr, peptide probability and protein probability, ANOVA P value, and significance after multiple testing correction.

ProtQuant analysis of the *E. ictaluri* iron replete and DP comparison identified expression of 217 proteins to be significantly increased or decreased (at Benjamini-Hochberg adjusted $p \leq 0.05$).

Since ProtQuant output has the compiled information for all replicate datasets for two conditions that are being compared, we use ProtQuant output as a template for performing resampling-statistics-based DE analysis using our MATLAB script, rsProt. The first step is to reformat the ProtQuant output to remove protein entries that do not have at least three peptides in at least one dataset that is being compared using a Perl script, resampling-input. The output of resampling-input is in the required format for running rsProt in MATLAB. To run rsProt, the first line in rsProt must be modified to match the filename of the sample to be analyzed by resampling and the last line can be modified only if a specific output file name is desired. rsPRot requires a user specified number of iterations for estimating the P value. After running rsProt a text file is generated containing four columns (Figure 6). Columns one, two, three, and four contain

ID	Resamp_meandiff	Resamp_pvalue	Resamp_sign
NT01EI0001	0.0661	0.352	1
NT01EI0002	-0.1888	0.168	-1
NT01EI0003	0.3459	0.057	1
NT01EI0004	0.2712	0.022	1
NT01EI0005	-0.0691	0.439	-1
NT01EI0006	0.0187	0.43	1
NT01EI0007	0.1396	0.361	1
NT01EI0009	2.4417	0.005	1
NT01EI0010	2.1639	0	1
NT01EI0012	-0.1165	0.274	-1
NT01EI0013	0.0537	0.417	1
NT01EI0014	0.2684	0.025	1
NT01EI0015	0.1479	0.276	1
NT01EI0016	-1.2585	0.003	-1
NT01EI0017	-0.4056	0.089	-1
NT01EI0018	0.4831	0.027	1
NT01EI0020	-0.653	0.076	-1
NT01EI0021	-0.1581	0.359	-1
NT01EI0022	-0.3224	0.111	-1

Figure 6
Sample resampling output generated by rsProt.

the protein id, mean difference in the ΣXcorr, probability, and direction of differential expression relative to the control dataset, respectively (1 represents increased expression of a protein in the treatment compared to the control and -1 represents opposite trend).

The final step in our proteomics workflow that includes resampling statistics is to compile the results from ProbCal, ProtQuant, and rsProt to generate a

comprehensive list of differentially expressed proteins. A collection of custom Perl scripts entitled integrator is used to accomplish this task. Integrator is run in steps from the command prompt and produces a single text file. Step one is to run process-protquant which reduces the ProtQuant data to a list of significantly expressed proteins with ΣXcorr for the control and treatment and performs a Benjamini-Hochberg correction. Step two is to run add_protein_prob which adds the protein probabilities calculated by ProbCal to the ProtQuant results. The final step is to run add_resampling_data which adds the columns from the resampling results to the file. The compiled results contains a master list of differentially expressed proteins, the P values calculated by ProbCal indicating confidence in identification, relative expression data from ProtQuant, and resampling data indicating the probability of being wrong that the protein is differentially expressed. Figures 7 and 8 show a sample of the *E. ictaluri* data after being compiled by integrator.

Conclusion

The proteomic data analysis workflow described here for Bioworks Sequest results includes a modular design of the work flow wherein different components can be combined together to perform different analyses. The work flow can be as simple as identifying proteins at a certain probability threshold or as extensive as comparing two datasets for differential protein expression using multiple statistical methods. All the tools and scripts described here can be implemented and further modified to accommodate additional analyses design but do require basic programming skills. All the tools and scripts used are compatible with both Linux and Windows platforms.

Methods

Implementation

XML2TXT is a java program that converts an xml file into a tab delimited text file, further used by other scripts. It is implemented using Xalan-Java, which is an XSLT (XSL Transformations) processor for transforming XML documents into text document types. javax.xml.transform interface is used as java API for XML Processing (JAXP) 1.3.

Perl scripts ProbCal, ProbCal-filter and integrator require the installation of the Active Perl runtime environment available at http://www.activestate.com/activeperl/. ProbCal is the implementation of the peptide probability calculation. Individual peptide probabilities are further utilized to calculate the probability that a protein identification is incorrect. Another subsidiary script ProbCal-filter uses the peptide probabilities to filter low quality peptides from being included in further analysis.

ID	Description	Control_sumXcorr	Control_pep_number	Treatment_sumXcorr	Treatment_pep_number	ANOVA_Pvalue
NT01EI0167	NT01EI0167 translation	411.887	163	726.164	246	2.20E-16
NT01EI1665	NT01EI1665 aldehyde-a	355.759	127	758.263	248	2.20E-16
NT01EI1908	NT01EI1908 hemin degr	10.705	8	378.754	131	2.20E-16
NT01EI1909	NT01EI1909 TonB-depe	15.906	13	307.831	105	2.20E-16
NT01EI2738	NT01EI2738 conserved	161.371	59	36.456	17	2.20E-16
NT01EI3596	NT01EI3596 elongation	405.85498	159	719.584	243	2.20E-16
NT01EI1641	NT01EI1641 glyceraldeh	141.748	65	341.91	119	3.48E-16
NT01EI3188	NT01EI3188 hypothetica	0	0	6.671	6	5.17E-14
NT01EI0216	NT01EI0216 Bacterial e	9.909	8	101.464005	38	8.23E-14
NT01EI3191	NT01EI3191 globin dom	4.682	4	68.279	24	1.09E-11
NT01EI0443	NT01EI0443 conserved	12.286	11	0	0	1.45E-11
NT01EI2695	NT01EI2695 phosphate	58.11	36	192.761	76	2.93E-11
NT01EI3420	NT01EI3420 conserved	6.266	5	0	0	1.78E-10
NT01EI3091	NT01EI3091 ferritin and	26.353	11	115.989	36	2.46E-10
NT01EI1651	NT01EI1651 conserved	10.326	8	0	0	3.52E-10
NT01EI0831	NT01EI0831 flavodoxin/	0	0	22.501999	8	8.11E-10
NT01EI2788	NT01EI2788 conserved	9.566	6	0	0	1.41E-09
NT01EI3368	NT01EI3368 phosphogly	41.305	19	118.078995	49	1.53E-09
NT01EI0758	NT01EI0758 pyruvate d	87.47	41	193.992	88	3.46E-09
NT01EI3355	NT01EI3355 2-polypren	30.734	24	4.818	4	3.50E-09
NT01EI1027	NT01EI1027 conserved	5.734	5	0	0	8.29E-09
NT01EI0393	NT01EI0393 hypothetica	0	0	3.203	3	1.45E-08
NT01EI3051	NT01EI3051 transposas	4.402	4	0	0	3.42E-08
NT01EI2796	NT01EI2796 transposon	7.2079997	6	0	0	4.04E-08

Figure 7
Combined output from integrator showing ProtQuant data.

Benjamini_corrected Pvlaue	Significant?	Protein_prob_control	Protein_prob_treatment	Resamp_meandiff	Resamp_pvalue	Resamp_sign
6.99E-13	Sig	0	0	1.1055	0	1
3.49E-13	Sig	0	0	1.3754	0	1
2.33E-13	Sig	9.95E-30	0	2.6783	0	1
1.75E-13	Sig	7.81E-59	0	2.5445	0	1
1.40E-13	Sig	0	2.66E-81	-1.8975	0	-1
1.16E-13	Sig	0	0	1.1055	0	1
1.58E-13	Sig	0	0	1.4193	0	1
2.05E-11	Sig	1.14E-01	8.55E-27	1.1118	0	1
2.90E-11	Sig	1.24E-37	4.88E-263	1.974	0	1
3.47E-09	Sig	6.33E-24	2.15E-143	2.3955	0	1
4.17E-09	Sig	1.17E-45	1.03E-03	-1.1222	0	-1
7.74E-09	Sig	3.21E-206	6.91E-323	1.3607	0	1
4.35E-08	Sig	3.23E-32	2.46E-03	-1.2532	0.001	-1
5.58E-08	Sig	1.56E-94	9.55E-276	2.1026	0	1
7.46E-08	Sig	2.00E-26	1.19E-02	-1.3167	0	-1
1.61E-07	Sig		1.63E-38	2.8128	0	1
2.63E-07	Sig	8.40E-38	5.24E-01	-1.5943	0.001	-1
2.69E-07	Sig	6.40E-96	1.23E-305	1.3381	0	1
5.78E-07	Sig	1.05E-149	0	0.9477	0	1
5.56E-07	Sig	4.02E-108	1.28E-27	-0.9963	0	-1
1.25E-06	Sig	6.80E-42	9.59E-02	-1.141	0.001	-1
2.10E-06	Sig	1.90E-01	3.00E-21	1.0677	0.015	1
4.72E-06	Sig	2.13E-26	9.85E-03	-1.1005	0	-1
5.34E-06	Sig	2.65E-39	1.28E-05	-1.1888	0.001	-1

Figure 8
Combined output from integrator showing Benjamini-Hochberg correction, protein probabilities and resampling data.

ProtQuant is implemented in Java 5 for platform independence. A self-installing executable for Windows has been generated using Macrovision InstallShield. An instruction for installing and using the tool in a Linux environment is also available. ANOVA analysis is done using a library from the R statistical package http://www.rproject.org/. Because of the size of the datasets that ProtQuant must handle, MySQL is used for data storage and efficient data manipulation. ProtQuant uses the file extension of input files to determine the format. ProtQuant includes a custom built parser for XML files. rsProt, for resampling, requires the installation of MatLab, available for purchase from MathWorks at http://www.mathworks.com/products/matlab/.

E. ictaluri *Proteomics*

E. ictaluri cultures were grown in triplicate in BHI (iron replete) and BHI with 100 M dipyridyl (iron restriction). Outer membrane proteins were isolated by sodium N-lauroylsarcosinate (SLS) extraction [17]. Protein concentrations were determined using the Plus one 2D quant kit following the manufacturer's protocol (Amersham Biosciences, Piscataway, NJ). Trypsin digestion proteins and analysis of tryptic peptides by 2-D LC ESI MS/MS were conducted as described previously [14]. For protein identification all searches were done using TurboSEQUEST™ (Bioworks Browser 3.2, ThermoElectron). Mass spectra and tandem mass spectra were searched against an in silico trypsin-digested *E. ictaluri* protein database (3786 proteins). Cysteine carboxyamidomethylation and methionine single and double oxidation were included in the search criteria. For decoy searches a reversed version of the protein database was generated using the reverse database function in Bioworks 3.2. The reversed database was also in silico trypsin digested and used for searches with tandem mass spectra exactly as described for the protein database. Bioworks results were exported in XML format for proteomic analysis workflow described here.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

BN and SCB developed/implemented the methods for proteomic analysis. RK wrote the scripts for joining different components of the workflow. KP conducted proteomic data analysis of *E. ictaluri* raw mass spec data (including Bioworks 3.2 searches) and did extensive testing of the analysis workflow. KP wrote the draft of the manuscript. All authors contributed to writing the

manuscript and have read and approved the final manuscript.

Acknowledgements

This project was partially supported by a grant from the National Research Initiative of the USDA Cooperative State Research, Education and Extension Service grant number #2006-35600-17688 and National Science Foundation (EPS-0556308-06040293). We acknowledge Dr. Mark L Lawrence for providing *E. ictaluri* proteomic datasets. We acknowledge Tibor Pechan of the Life Sciences and Biosciences Technology Institute, Mississippi State University for running the mass spectrometer. The Life Sciences and Biotechnology Institute provided salary support for Ken Pendarvis.

This article has been published as part of *BMC Bioinformatics* Volume 10 Supplement 11, 2009: Proceedings of the Sixth Annual MCBIOS Conference. Transformational Bioinformatics: Delivering Value from Genomes. The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2105/10?issue=S11.

References

1. Wolters DA, Washburn MP and Yates JR 3rd: **An automated multidimensional protein identification technology for shotgun proteomics.** *Anal Chem* 2001, **73(23):**5683–90.
2. Yates JR 3rd, *et al*: **Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database.** *Anal Chem* 1995, **67(8):**1426–36.
3. Perkins DN, *et al*: **Probability-based protein identification by searching sequence databases using mass spectrometry data.** *Electrophoresis* 1999, **20(18):**3551–67.
4. Craig R and Beavis RC: **TANDEM: matching proteins with tandem mass spectra.** *Bioinformatics* 2004, **20(9):**1466–7.
5. Lee SR, *et al*: **Bovine viral diarrhea virus infection affects the expression of proteins related to professional antigen presentation in bovine monocytes.** *Biochim Biophys Acta* 2009, **1794(1):**14–22.
6. Lee SR, *et al*: **Differential detergent fractionation for non-electrophoretic bovine peripheral blood monocyte proteomics reveals proteins involved in professional antigen presentation.** *Dev Comp Immunol* 2006, **30(11):**1070–83.
7. Nanduri B, *et al*: **Effects of subminimum inhibitory concentrations of antibiotics on the Pasteurella multocida proteome.** *J Proteome Res* 2006, **5(3):**572–80.
8. Nanduri B, *et al*: **Proteomic analysis using an unfinished bacterial genome: the effects of subminimum inhibitory concentrations of antibiotics on Mannheimia haemolytica virulence factor expression.** *Proteomics* 2005, **5(18):**4852–63.
9. Choi H and Nesvizhskii AI: **Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics.** *J Proteome Res* 2008, **7(1):**254–65.
10. Choi H and Nesvizhskii AI: **False discovery rates and related statistical concepts in mass spectrometry-based proteomics.** *J Proteome Res* 2008, **7(1):**47–50.
11. Liu H, Sadygov RG and Yates JR 3rd: **A model for random sampling and estimation of relative protein abundance in shotgun proteomics.** *Anal Chem* 2004, **76(14):**4193–201.
12. Gao J, *et al*: **Changes in the protein expression of yeast as a function of carbon source.** *J Proteome Res* 2003, **2(6):**643–9.
13. Bridges SM, *et al*: **ProtQuant: a tool for the label-free quantification of MudPIT proteomics data.** *BMC Bioinformatics* 2007, **8(Suppl 7):**S24.
14. Nanduri B, *et al*: **Quantitative analysis of Streptococcus pneumoniae TIGR4 response to in vitro iron restriction by 2-D LC ESI MS/MS.** *Proteomics* 2008, **8(10):**2104–14.
15. Benjamini Y and Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 1995, **57(1):**289–300.
16. Lopez-Ferrer D, *et al*: **Statistical model for large-scale peptide identification in databases from tandem mass spectra using SEQUEST.** *Anal Chem* 2004, **76(23):**6853–60.
17. Williams ML, Azadi P and Lawrence ML: **Comparison of Cellular and Extracellular Products Expressed by Virulent and Attenuated Strains of Edwardsiella ictaluri.** *Journal of Aquatic Animal Health* 2003, **15:**264–273.