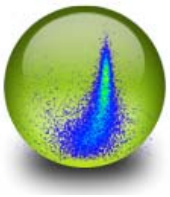
ProRata v0.1 User's Manual



Chongle Pan[1,2,4]; Guruprasad Kora[2]; David Tabb[3,5]; Hayes McDonald[1]; Dale Pelletier[3]; Greg Hurst[1]; Nagiza F. Samatova[2,*]; Robert Hettich[1,*]

[1] Chemical Sciences Division, Oak Ridge National Laboratory
[2] Computer Sciences and Mathematics Division, Oak Ridge National Laboratory
[3] Life Sciences Division, Oak Ridge National Laboratory
[4] Genome Science and Technology Graduate School, ONRL-University of Tennessee
[5] Current affiliation: Department of Biomedical Informatics, Vanderbilt University
[*] Project PI

November 10, 2005

Table of Content

Table of Contents

1. Introduction

ProRata is a computer program that automates the data processing for quantitative shotgun proteomics. Quantitative shotgun proteomics attempts to estimates the abundance of proteins in the treatment cells relative to their abundance in the reference cells. In such experiments, the reference cells are labeled with the stable isotopes via metabolic labeling, enzymatic labeling, or chemical labeling and are mixed with the treatment cells in equivalent amounts. After proteolysis, the samples are analyzed with multi-dimensional liquid chromatogram-tandem mass spectrometry. The peptides are identified with the program SEQUEST [1] and filtered with the program DTASelect [2]. ProRata takes the output from DTASelect and the mass spectral data and performs all steps necessary for estimating protein abundance ratios and their confidence interval. Those steps include: extraction of selected ion chromatograms, detection of chromatographic peaks, estimation of peptide abundance ratios, and estimation of protein abundance ratios. All intermediate results from each of the steps can be inspected with the ProRata graphical user interface.

Release note:
This is a beta release of ProRata for testing purpose. The graphical user interface and the user's manual are still under development. If you find any bug or have any suggestion, please contact Chongle Pan (panc@ornl.gov), Robert Hettich (hettichrl@ornl.gov), and Nagiza F. Samatova (samatovan@ornl.gov) . We recommend you to inform us on your installation of ProRata so that you could receive the upgrades notification.

2. ProRata Installation

ProRata v.0.1 is currently available and tested for Windows platforms including Windows 2000 and Windows XP. The computer system is recommended to have 3.0 GHz CPU and 512 MB memory. Dual monitor is desirable for better data browsing experience. To install ProRata, just run the *Setup.exe* file of ProRata and follow the setup wizard. ProRata will be added to the desktop and the start menu. You can start ProRata from both the desktop shortcut or from the start menu.

Figure 1. ProRata Setup Window.

ProRata can be completely removed from your system by uninstalling it. To uninstall ProRata, go to the start menu ProRata fold and select “unintall”.

3. Execution and Data Processing

The instructions will guide you through all the steps of the quantitative proteomics data processing with the program ProRata. To start data processing with ProRata, go to Tools → Execute ProRata from the menu. The “ProRata Data Processing” window (Figure 2) will pop up. This form will guide you through all the steps of preparing the required input files.

In the first step, select a working directory for an experiment. ProRata expects all the input data stored in a working directory and writes all output files to the working directory. The required input data is the DTASelect-filter.txt file, a directory called mzXML storing all mzXML files and a configuration file, called ProRataConfig.xml. If these inputs are found in the working directory, their corresponding fields will be pre-filled in this form. You may browse to select those inputs from elsewhere, which will then be copied to the working directory by ProRata.

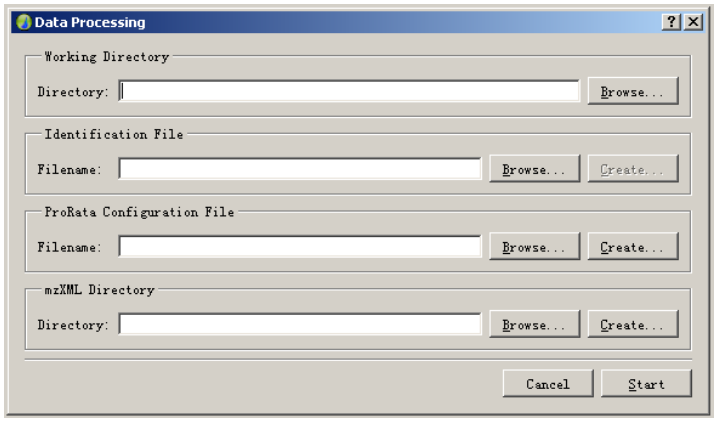


Figure 2. ProRata Data Processing Window

In the second step, select a DTASelect-filter.txt file. It is recommended to generate DTASelect-filter.txt in the following manner. Run two iterations of the SEQUEST search. Configure one to identify the labeled peptides and the other one to identify the unlabeled peptides. ProRata provides the tool to assist merging the two SEQUEST search results. Save one iteration of the SEQUEST search in the directory called "reference", "labeled" or any other meaningful one-word directory name and the other in the directory called "treatment", "unlabeled" or any other meaningful one-word directory name. Go to the Tool → Merge Directories. Add those two directories to the input directories. Browse to select an output directory. Click the "Merge" button to start merging.

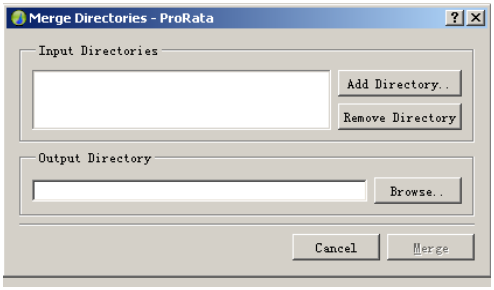


Figure 3. Merge Directories Window.

All files, sub-directories and the files in the sub-directories will be renamed to add a prefix of the input directory name and be copied to the output directory. Copying files might take a while; please, be patient if the program is not responding. Then run DTASelect in the output directory. Merging of the two SEQUEST result directories will allow DTASelect perform identification filtering and assembling with both labeled peptide IDs and unlabeled peptide IDs. It is recommended to run DTASelect with option "-t". Copy the DTASelect-filter.txt from the output directory to ProRata's working directory specified in Figure 2.

In the third step, provide a configuration file ProRataConfig.xml in Figure 2. If you have previously created a configuration file that can be re-used, you can browse to select that file and it will be copied to the working directory. You can also create ProRataConfig.xml from scratch by clicking the corresponding "Create" button in Figure 2. A dialog box will be invoked, which

has the following tabs: Denominator Isotope, Numerator Isotope, Quantification, and Chromatogram.

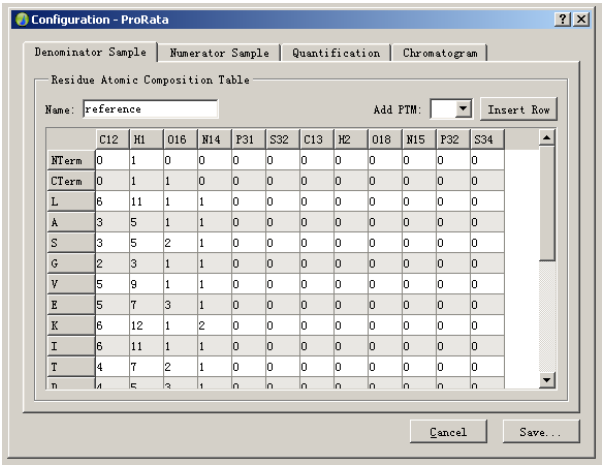


Figure 4. ProRata Configuration File Window.

The Denominator Sample tab and the Numerator Sample tab contain the residue atomic composition tables for the two isotopic forms of a peptide in the reference sample and the treatment sample. The default values are natural residues' atomic composition, which is composed of all natural isotopes. You MUST change these two tables according to your labeling technique. For example, if you are using ^{15}N metabolic labeling, change the atom number of the N15 column to the appropriate nitrogen atom number in the residues and change the atom number of the N114 column to zero. If you chemically modified the residue, please also change the residue's atomic composition according to the modification. If you searched for a PTM, please also add that PTM and give its atomic composition. You might name the two samples the way you like and these names will be used throughout the ProRata output. The denominator sample is the sample for the denominator of the abundance ratio and the numerator sample is for the numerator of the abundance ratio.

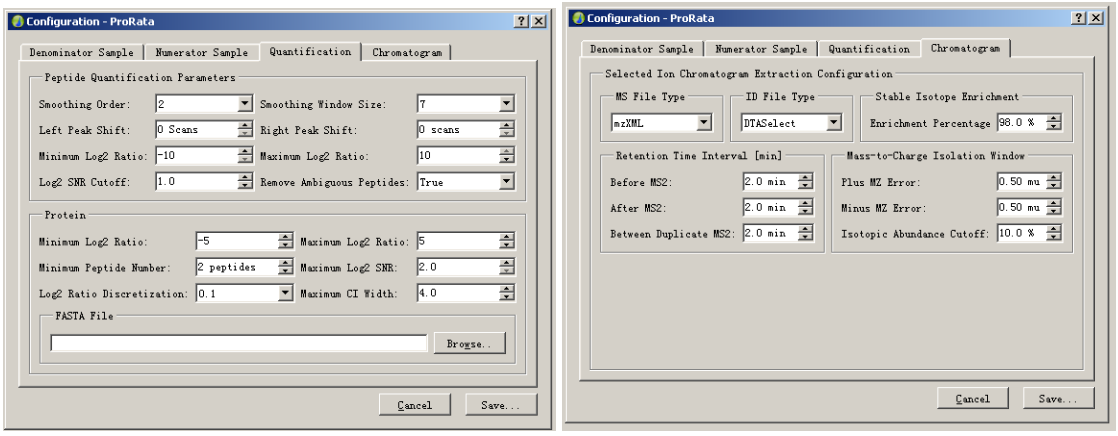


Figure 5. Quantification and Chromatogram tabs in ProRata Configuration Window.

The Quantification tab and the Chromatogram tab (Figure 5) present the parameters for the abundance ratio estimation and chromatogram extraction, respectively. The default values are

generally a good place to start. If you want to see the sequence coverage plot from a graphical user interface, please select the FASTA file used for database searching. After making all the changes, click the "Save" bottom to save the parameters to a configuration file ProRataConfig.xml in the working directory.

In the fourth step, provide an mzXML directory containing all mzXML files (Figure 6) for the experiment by clicking the corresponding "Create" button in Data Processing Window (Figure 2). If there exists a subdirectory called mzXML in the working directory, ProRata will pre-fill this field and assume that the directory has all mzXML files. If you select another directory for mzXML, ProRata will create an mzXML sub-directory and copy all mzXML files to this directory. You can also create the mzXML files from the raw files with a conversion tool provided by ProRata.

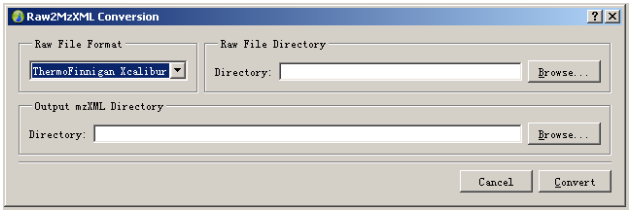


Figure 6. Raw2MzXML Conversion Window.

Select the raw file format, the directory containing all raw files, and the directory for storing the mzXML files. ProRata calls the conversion programs, ReAdW and Wolf, provided in http://sashimi.sourceforge.net/software_glossolalia.html to perform the conversion. Please wait till the conversion is finished before proceed.

After completing all four steps, click "Start" to start the data processing with ProRata. ProRata will check the validity of all inputs and prompt you to correct the invalid input. A console will appear to show the progress of data processing. After the data processing is finished, the following output will be generated in the working directory: an *xic* sub-directory containing the extracted ion chromatogram files, a quantification result file ProRata_Quantification.qpr.xml, and two tab-delimited text files ProRata_Quantification_Peptide.txt and ProRata_Quantification_Protein.txt.

You can also perform all these steps from a command line console. In the first step, create a working directory. In the second step, copy the DTASelect-filter.txt file to the working directory. Thirdly, copy the ProRataConfig.xml to the working directory. In the third step, create a mzXML sub-directory and copy all mzXML files to the mzXML directory. This will complete the preparation of the input data for ProRata and then you can start the data processing from a command line console. Open a command line console and go to the working directory. Enter "sicForma" and the xic files containing the selected ion chromatograms will be created. The xic files is stored in the xic sub-directory of the working directory. After the execution of sicForma is finished, enter "pratio" to start the peptide quantification and protein quantification. The same output files will be created.

4. Data Exploration and Visualization

ProRata provides access to both the MS data and all intermediate results used for the quantification. To start data exploration and visualization, go to File→Open in the main menu to open the ProRata_Quantification.qpr.xml file. This file MUST be in its original working directory for all the graphs to be displayed.

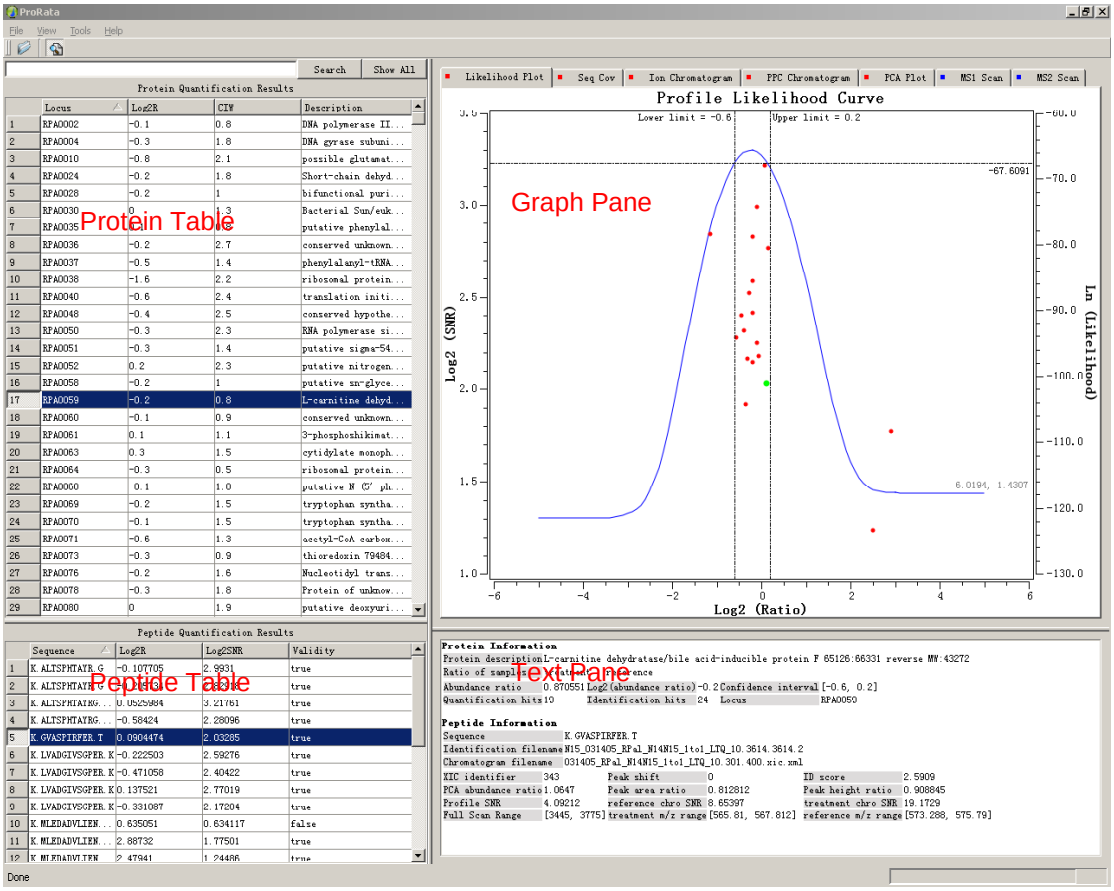


Figure 7. The ProRata Graphical User Interface for Data Exploration and Visualization.

The ProRata graphical user interface consists of four main parts: protein table, peptide table, graph pane, and text pane (Figure 7). The graph pane contains seven graphs that are displayed in the tabbed pages. The seven graphs are Likelihood Plot, Sequence Coverage Plot, Ion Chromatogram, PPC Chromatogram, PCA Plot, MS1 Scan, and MS2 Scan.

ProRata organizes the data into a hierarchy of three levels, which are the protein level, the peptide level, and the MS level. The protein table summarizes and provides the entry point to the information at the protein level, which gives the locus (in the "Locus" column), the log2 abundance ratio (in the "LogR" column), the confidence interval width (in the "CIW" column) and the description for all quantified proteins (in the "Description" column). Generally the smaller is the confidence interval width, the more precise a protein is quantified. To examine a particular protein from the protein table, just select its row from the protein table and the following information will be presented: a) all peptides from this protein will be listed in the Peptide Quantification Results table; b) the likelihood plot and sequence coverage plot of this

protein will be displayed; c) more detailed textual information about this protein will be displayed in the protein information section of the text pane. Navigation of the protein table can be assisted by Searching and Sorting of the protein table. The locus column and the description protein table can be searched by the exact phrases entered in the search pane. Only the matched proteins will be listed in the protein table as a result of search. Clicking the "Show all" button will bring back the entire protein list. The proteins can be sorted by clicking the column names.

All graphs can be detached from the graph pane by right-clicking and selecting "Detach" from the popup menu (Figure 8). The detached graph becomes a floating window that can be moved and resized. Closing a detached graph window will put that graph back into the graph pane. All graphs can also be exported to the PNG images. Select a graph, invoke the popup menu and select "Export..". Then give a filename to save the PNG image file.

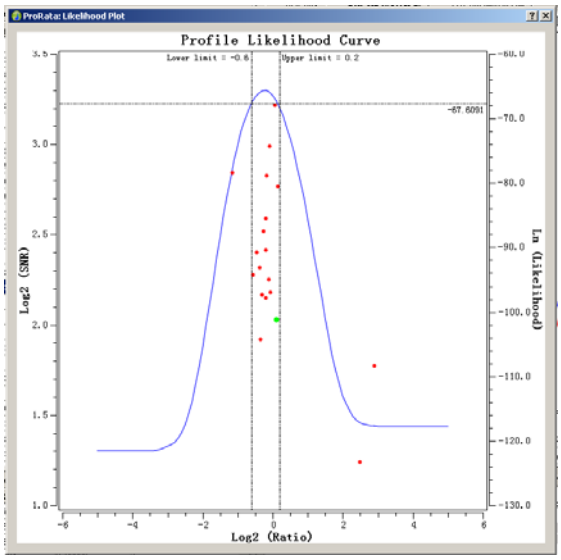


Figure 8. A Detached Likelihood Plot.

The Likelihood plot (Figure 8) informs how the abundance ratio of the selected protein is estimated from its peptides. The red dots represent quantified peptides of the protein. The x axis (bottom axis) is the log2 abundance ratio of the peptides and the y axis (left axis) is the log2 signal to noise ratio of the peptides. The blue curve represents the likelihood profile curve. Its x axis (bottom axis) is the log2 abundance ratio of the protein and its y axis (right axis) is the *ln* of the likelihood of the abundance ratio. The peak top of the likelihood profile curve is the maximum likelihood estimation of the protein. The horizontal line is the likelihood cutoff for the 95% confidence interval and the two vertical lines are the upper and lower limits of the confidence interval. The peptide quantification dots relate to the protein likelihood curve in the following way: a peptide with log2 abundance ratio of *a* will bump the likelihood curve up at *a* and the higher log2 eigenvalue this peptide has, the higher the bump will be. A peak in the likelihood curve is calculated from multiple peptides with close log2 abundance ratios and the peak top is determined from the relative amplitude of those peptides' log2 eigenvalue ratios

After selecting a protein, its peptides can then be examined. The Peptide Quantification Results table summarizes and provides the entry point to the information at the peptide level. The peptide table gives the sequence (in the "Sequence" column), the log2 abundance ratio (in the "Log2R"

column), the log2 profile signal-to-noise ratio (in the “log2SNR” column), and the quantification validity for all the peptides (in the “Validity” column) from the selected protein. To access the information about a peptide, just select this peptide and the following information will be available for display from the graph panel: a) the ion chromatograph, b) the PPC chromatogram, and c) the PCA plot. Also the more detailed textual information will be shown in the peptide information section of the text pane. Peptide table can also be sorted by clicking the appropriate column name.

A peptide is measured in multiple MS1 and MS2 scans, and the ion chromatogram provides the entry point to each individual MS scan. Left-clicking the ion chromatogram will select a retention time. Then the MS1 scan acquired at that time will be displayed and, if an MS2 scan of this peptide is acquired around that time, that MS2 scan will also be displayed (Figure 9). The m/z regions in the MS1 scan are highlighted with blue and red for the m/z isolation windows of the two isotopologues. In the MS2 scan, the singly charged y ions and b ions are also highlighted with blue and red, respectively. The MS1 scan can be displayed sequentially scan by scan by scrolling the scroll wheel over the ion chromatogram. Both mass spectra can be zoomed in by drawing a rectangular in the spectrum. The mass spectra can be zoomed out step by step by clicking the scroll wheel or be zoomed out to entire m/z range by right-clicking.

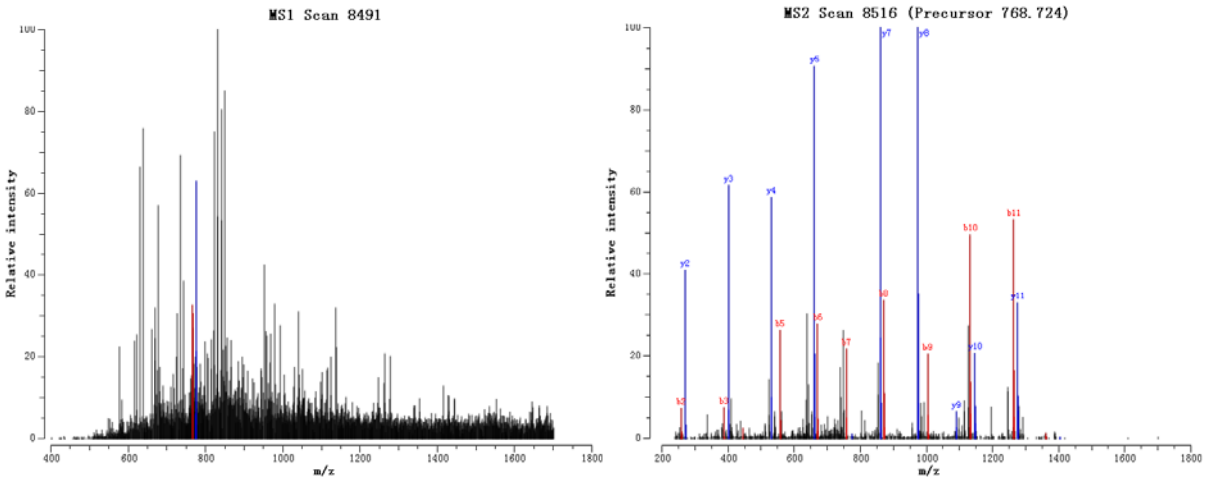


Figure 9. The Exported PNG Images of MS1 Scan and MS2 Scan.

Next, the fields in the text pane will be explained. The text pane consists of two sections: the protein information section and the peptide information section. If a protein or a peptide is selected, the corresponding sections will be updated to display the information for that selected protein or peptide. The protein description is shown in the “Protein description” field again, as most protein’s description cannot be displayed in full within the description column of the protein table. The numerator and the denominator of the abundance ratio are shown in the “Abundance ratio” field. The abundance ratios in the linear scale and in the log2 scale are shown in the “Abundance ratio” and the “Log2(abundance ratio)” field, respectively. The confident interval is shown in the “Confidence interval” field. The “Identification hits” indicates how many times peptides from this protein are identified after filtering. The “Quantification hits” show how many of those identification hits give good peptide abundance ratio estimates.

The peptide “Identification filename” is the dta files for the identification hits in this chromatogram. The “Chromatogram filename” and the “XIC identifier” are the *xic* filename and the identifier for the chromatogram. The “Peak shift” is the number of MS1 scans that the ion chromatogram has been selected. The “ID score” is the identification score for the MS2 scan. If there are multiple MS2 scans, the maximum ID score is shown. The PCA abundance ratio, Peak area ratio, and Peak height ratio is the peptide abundance ratio estimated with the PCA, peak area, and peak height, respectively. The “Profile SNR ratio” field shows the profile signal-to-noise ratio in the PCA abundance ratio estimation. The chromatographic signal-to-noise ratio (chro SNR) is calculated for both selected ion chromatograms (reference and treatment) as the ratio of the peak top to the baseline. The Full Scan Range shows the first and last MS1 scans in the selected ion chromatogram. The m/z ranges of the two ion chromatograms are also shown.

5. Data formats

1. The configuration file ProRataConfig.xml

2. The ion chromatogram files .xic

3. The quantitative proteomics result file .qpr.xml

4. Peptide tab-delimited table

5. Protein tab-delimited table

6. Feedback

7. Acknowledgement

This work is performed as part of the [OBMS funding ack should go here] and the Scientific Data Management Center (http://sdmcenter.lbl.gov) under the Department of Energy's Scientific Discovery through Advanced Computing program (http://www.scidac.org). Oak Ridge National Laboratory is managed by UT-Battelle for the LLC U.S. D.O.E. under Contract No. DE-AC05-00OR22725.

8. References

1. Eng, J.K., Mccormack, A.L., and Yates, J.R., *An Approach to Correlate Tandem Mass-Spectral Data of Peptides with Amino-Acid-Sequences in a Protein Database.* Journal of the American Society for Mass Spectrometry, 1994. **5**(11): p. 976-989.
2. Tabb, D.L., McDonald, W.H., and Yates, J.R., *DTASelect and contrast: Tools for assembling and comparing protein identifications from shotgun proteomics.* Journal of Proteome Research, 2002. **1**(1): p. 21-26.