Multi-Q: A Fully Automated Tool for Multiplexed Protein Quantitation

Wen-Ting Lin[a†], Wei-Neng Hung[a†], Yi-Hwa Yian[a], Kun-Pin Wu[a]*, Chia-Li Han[b], Yet-Ran Chen[c],

Yu-Ju Chen[c,d], Ting-Yi Sung[a]* and Wen-Lian Hsu[a]

[a] Institute of Information Science, Academia Sinica, Taipei 115, Taiwan

[b] Department of Chemistry, National Taiwan Normal University, Taipei 116, Taiwan

[c] Institute of Chemistry, Academia Sinica, Taipei 115, Taiwan

[d] Genomics Research Center, Academia Sinica, Taipei 115, Taiwan

[*] To whom all correspondence should be addressed: kpw@iis.sinica.edu.tw, tsung@iis.sinica.edu.tw

[†] Contributed equally to this work.

Abstract

The iTRAQ labeling method combined with shotgun proteomic techniques represents a new dimension in multiplexed quantitation for relative protein expression measurement in different cell states. To expedite the analysis of vast amounts of spectral data, we present a fully automated software package, called Multi-Q, for multiplexed iTRAQ-based quantitation in protein profiling. Multi-Q is designed as a generic platform that can accommodate various input data formats from search engines and mass spectrometer manufacturers. To calculate peptide ratios, the software automatically processes iTRAQ's signature peaks, including peak detection, background subtraction, isotope correction, and normalization to remove systematic errors. Furthermore, Multi-Q allows users to define their own data-filtering thresholds based on semi-empirical values or statistical models so that the computed results of fold changes in peptide ratios are statistically significant. This feature facilitates the use of Multi-Q with

various instrument types with different dynamic ranges, which is an important aspect of iTRAQ analysis. The performance of Multi-Q is evaluated with a mixture of 10 standard proteins and human Jurkat T cells. The results are consistent with expected protein ratios and thus demonstrate the high accuracy, full automation, and high-throughput capability of Multi-Q as a large-scale quantitation proteomics tool. These features allow rapid interpretation of output from large proteomic datasets without the need for manual validation. Executable Multi-Q files are available on Windows platform at http://ms.iis.sinica.edu.tw/Multi-Q/.

Keywords

Dynamic range, iTRAQ, mass spectrometry, multiplexed protein quantitation, Multi-Q, non-degenerate peptide filtration, normalization, protein quantitation, proteomics

Introduction

In the post-genomic era, the development of technologies for systematic, qualitative, and quantitative protein analysis has become an important issue in the global analysis of biological systems. Many studies have shown that differentially expressed levels of proteins are associated with disease progression such as cancer[1-5], viral or bacterial infections[6], cardiomyopathic heart disease[7], and neurodegenerative diseases like Alzheimer's[8] or Huntington's[9]. In the search for disease-associated protein determinants, quantitation remains a vital analytical component in the study of the translational effects of regulating cellular activity and function.

Several approaches have been developed to quantify protein expression of highly complex biological samples in different states of tissues, microorganisms, or cultured eukaryotic cells. In order to analyze complex biological samples, powerful separation techniques, such as two-dimensional gel electrophoresis (2-DE)[10] or multidimensional liquid chromatography (MDLC), are required prior to

identification of differentially expressed proteins by mass spectrometry[11]. The emerging field of proteomics has evolved from the mature technology of 2-DE for protein separation and quantitation. However, despite 2-DE's unique features of rapid separation and simultaneous visualization of protein distribution as a function of molecular weight and charge, it seriously under-represents specific classes of proteins, including very acidic or basic proteins, extremely large or small proteins, and membrane proteins. Furthermore, when examining the codon bias values of identified proteins, the 2-DE approach cannot detect low abundance proteins without using additional sample enrichment techniques[12,13]. Since many proteins are present in different orders of abundance, the key to deciphering a regulatory event lies in developing more sensitive analytical tools that can comprehensively analyze the dynamic changes in proteomes associated with stimuli.

The alternative strategy, MDLC-based analysis, enhances the detection efficiency of low abundance proteins and is therefore more suitable for the global analysis of a complex biological system. The development of isotopic labeling technique followed by MDLC analysis facilitates quantitation and reduces the sample's complexity. The relative quantity of proteins can be determined by comparing the integrated peak areas of two corresponding isotope chromatographic profiles. Commonly used labeling techniques include, e.g., chemical labeling (ICAT[14], ^{18}O-labeling[15]) and metabolic labeling (SILAC[16]). These methods are generally limited to two-plex protein quantitation. Recently, multiplexed sets of iTRAQ isobaric reagents have been developed for protein quantitation[17] and also been widely used in proteomic research[18-19]. The quantitation strategy for iTRAQ-labeling of multiplexed sets of experiments is based on a set of four isobaric reagents, each of which comprises three groups: reporter, balance, and reactive groups. After cell lysis, reduction and alkylation, and protein digestion, the peptides in four states are separately labeled on N-terminals and lysine residues by the reactive group of iTRAQ. The labeled peptides are then combined and analyzed by MDLC and mass spectrometry (MS). Unlike ICAT, isotopically labeled peptide derivatives are indistinguishable in molecular weight and chromatographic separation. The four versions of isotopic iTRAQ-labeled peptides generate reporter groups with signature ions at m/z 114, 115, 116, and 117, respectively. The mass shift from the

signature ions is compensated for by the balance groups (*m/z* 31, 30, 29, 28) such that the mass of each of the four tags is identical in an MS scan. In contrast to other quantitative methods based on peptide intensity, the signature ions of the reporter group in an MS/MS spectrum are used for quantitation of protein expression levels because the intensity of each peak represents the quantity of the corresponding peptide.

The multiplexed protein quantitation capability of the iTRAQ strategy permits simultaneous comparison of samples of different cell states and disease specimens. This advantage allows us to perform a comparative study on the temporal and/or spatial distribution of proteins in cells. In addition, as the four-fold multiplexing reagents target the N terminals and the lysine side chains of tryptic peptides, more peptides of a protein can be labeled. Therefore, the protein can be identified and quantified by more peptides, thereby enhancing the confidence of protein quantitation[20]. However, the sample complexity increases substantially, which presents a challenge for data processing.

As fully automated liquid chromatography-tandem mass spectrometry (LC-MS/MS) is routinely used following the quantitative labeling strategy, multidimensional chromatography in conjunction with mass spectrometry usually generates tens of thousands of MS/MS spectra (the size of raw data usually amounts to several gigabytes). Many thousands of peptides can be collectively analyzed by multiple LC-MS/MS runs in each proteomic experiment such that hundreds, or even thousands, of proteins can be identified. Due to inevitable false-positive identification and incorrect quantitation in automatically generated results, many of them are unreliable and therefore require manual examination[21].

To reduce possible errors generated by MS/MS-based peptide quantitation methods, such as the iTRAQ strategy, we paid particular attention to the following factors in the design of our iTRAQ-labeling quantitation tool. First, iTRAQ's relatively large sample complexity inevitably includes identical peptides produced by two homologous proteins, which has been addressed by the program, Pro Group Viewer (Applied Biosystems, Foster City, CA). These degenerate (non-unique) peptides not only result in ambiguous protein identification, but also introduce protein quantitation errors. Removal of such peptides is essential in the analysis of a large dataset for protein quantitation. Second, the quality

of the MS/MS spectrum in signature peptides is particularly important for peptide quantitation. However, collision-induced dissociation (CID) conditions have to be carefully optimized in order to obtain good signal of signature ions at *m/z* 114, 115, 116, and 117, while maintaining enough ion intensity of other larger fragment ions for peptide sequence determination. The question arises: Given the limited dynamic range of detection, how can we compromise between the low S/N ratio of low abundance proteins and the saturated ion intensity of abundant proteins? Unfortunately, these characteristics of MS/MS spectra substantially reduce quantitation accuracy. Thus, the quantitation tool needs to allow users to input the threshold for dynamic range[22] filtering since an automatic analysis tool with better accuracy and confidence is essential for handling data and validation of large-scale quantitation with MS/MS signature ions.

Nowadays, i-TRACKER[23] is an available tool for iTRAQ-labeling quantitation. However, this tool is limited to peptide level quantitation and no protein level analysis is provided. Another quantitation tool is provided by Applied Biosystems, called ProQUANT (Applied Biosystems, Foster City, CA). Although the program has a peptide ratio normalization function to remove systematic errors resulting from isotope purity and the experimental process, none of the above-mentioned issues concerning the S/N filter or non-degenerate peptide differentiation is addressed by the current version. Most importantly, ProQUANT is limited to instruments developed by Applied Biosystems. A generic software package for analysis is needed, since various types of instruments have been developed by different manufacturers. Although there are other commonly used software tools such as RelEx[24] and MSQuant[25] for protein quantitation, they are applied for other isotope labeling technique other than iTRAQ for protein quantitation. However, quantitation of RelEx and MSQuant is based on MS spectra, whereas our Multi-Q deals with MS/MS spectra. Furthermore, these tools are designed for two-plexed protein quantitation and can only analyze two samples at a time.

In this paper, we present a flexible, generic, and reliable data processing software package called Multi-Q, which uses iTRAQ labeling for multiplex protein quantitation and is executable on Windows platform. As shown in Figure 1, Multi-Q provides a data converter for handling spectral raw data

generated by major mass spectrometer manufacturers, and the conversion time is much reduced. Moreover, Multi-Q can accept search results from various search engines, including SEQUEST[26], X!Tandem[27], and MASCOT[28]. After automatic filtering of non-iTRAQ labeled peptides, the ratio of all peptides with unambiguous identification is determined. To obtain protein ratios, Multi-Q filters out the ratios of degenerate peptides (i.e., peptides that match multiple proteins). In addition, we have developed user-defined filtering and statistical modules to deal with low quality MS/MS spectra, random and systematic errors in detection responses, and sample handling. The performance of the tool was evaluated on a mixture of 10 standard proteins (described in the next section) and cell lysates from Jurket T cells. Our results show that Multi-Q can achieve high quantitation accuracy without manual verification of MS/MS spectra.

Experiment

Chemicals

We purchased the following chemicals for the experiment: RPMI medium 1640 from HyClone (Logan, UT); fetal bovine serum (FBS) from Gibcobrl (Grand Island, NY); Trypsin from Promega (Madison, WI); Tris (2-carboxyethyl) phoshine (TCEP) from Fisher Scientific Corp. (Pittsburgh, PA); the BCA protein assay kit from Pierce (Rockford, IL); the methyl methanethiosulfonate (MMTS) from Applied Biosystems (Foster City, CA); and Tris, SDS, Triton X-100, trifluoroacetic acid (TFA), formic acid, and acetronitrile from Sigma-Aldrich (St. Louis, MO).

Preparation of Protein Standards

The protein standard was derived from 10 proteins in various concentrations, prepared by diluting the standard protein mixture containing 1 mg/ml of bovine serum albumin, enolase 1, trypsinogen, carbonic anhydrase, myoglobin, concanavalin A, lysozyme C, and beta-casein (all purchased from Sigma, St. Louis, MO); and alcohol dehydrogenase and ovalbumin (both purchased from Calbiochem-Novabiochem Corp., San Diego, CA).

Cell Culture and Lysis

Jurkat T cells were cultured in a RPMI medium supplemented with 10% fetal calf serum. The cells were harvested, washed and aliquoted (about $\sim 2 \times 10^7$ cells) and suspended in a lysis buffer made up of 50 mM Tris at pH 8.0, 0.1% SDS, and 0.02% Triton X-100. The individual suspensions were lysed by placing them in a sonic water bath for 30 minutes followed by subjection to three freeze (liquid N_2) and thaw cycles. Cellular debris was removed by centrifugation at 17,000 g for 45-60 minutes. The total protein concentration of each cell lysate was determined by BCA assay. The protein concentration was adjusted to approximately 1 mg/ml using 50 mM of Tris buffer containing 0.1% SDS.

Fractionation and Tryptic Digestion

Approximately 0.6 mg of the total protein was reduced with TCEP and alkylated with MMTS. SDS-PAGE was then applied to the cell lysates to fractionate proteins using 5-15% gradient gel in a Bio-Rad gel electrophoresis system (Hercules, CA). After staining with coomassie blue followed by destaining with 7% acetic acid in 25 % methanol, the molecular weight region of about 50-90 kDa was excised and further diced into 1×1 mm pieces. It was then subjected to in-gel tryptic digestion at 37 $^{\circ}$C overnight. The digested peptide mixture was extracted from gel pieces by removing the supernatant with 0.1 % TFA in 50 % (v/v) acetonitrile treated with sonication.

iTRAQ Labeling

The digested peptides were first desalted with C18 macrotrap (Michrom, Auburn, CA). The volume of eluted peptide solution was about 200 μl. The resulting peptide solution was divided into two parts: 100 μl was labeled with iTRAQ 114 reagent (Applied Biosystems' iTRAQ reagents), and the other 100 μl was labeled with iTRAQ 115 reagent according to the manufacturer's protocol.

Liquid Chromatography-Tandem Mass Spectrometry (LC-MS/MS)

The labeled peptide sample was loaded and sequentially analyzed by μLC-MS/MS. Samples were injected into a 2 cm × 100 μm capillary trap column packed in-house (Magic C_{18}, Michrom BioResources, Inc., Auburn, CA). Peptides were separated by a 10 cm × 75 μm capillary column, also packed in-house (Magic C_{18}, Michrom BioResources, Inc., Auburn, CA), and eluted with a linear gradient of 5-35% B for 60 min at ~200-300 nl/min (Buffer A: 0.1 % acetic acid in H_2O; Buffer B: 0.1 % acetic acid in acetonitrile). An HP1100 solvent delivery system (Hewlett Packard, Palo Alto, CA) was used with pre-column flow splitting, and a QSTAR Pulsar *i* mass spectrometer (Applied Biosystems, Foster City, CA) was used for all analyses. Peptide fragmentation by collision-induced dissociation was performed automatically using the information-dependent option, and the resultant MS/MS spectra were recorded.

Protein Identification and Quantitation

The MS/MS data that had not been interpreted was searched against the Celera Discovery System's human protein sequence database (version 8/3/2003) using the MASCOT MS/MS ion search algorithm (Matrix Science Inc., Boston, MA). For the MASCOT MS/MS ion search setup, the proteolytic enzyme was Trypsin, which allows 2 miscleavages. All the modifications, i.e., oxidization of methionines, iTRAQ labeling on lysines and N-terminals, and MMTS labeling of cysteines, were set to be variable, and the mass tolerance of the precursor and product ions was 0.3 Da. The data was analyzed with the Multi-Q software system.

Results and Discussion

Workflow of Multi-Q

The Multi-Q protein quantitation procedure comprises three major steps, as shown in Figure 2. Data analysis starts with conversion of mass spectrometry data files and database search results into our data formats for subsequent quantitation (Step 1). Based on the identification results and the MS/MS spectra, signature ions of the iTRAQ labeled peptides are selected, processed and quantified (Step 2). In the final

step, protein ratio determination, peptides with low identification confidence are removed. This program filters out peptide data outside the instrument's dynamic range (explained in Section 3.2) after assessment of the experimental data prior to peptide ratio normalization. Then, the normalized ratios of quantified, non-degenerate iTRAQ peptides are weighted according to their peak intensities to calculate the average protein ratios. The final protein quantitation results are exported to an output file in CSV (comma separated values) data format.

During the development of Multi-Q, four samples prepared from the mixture of 10 standard proteins (mentioned in the previous section) with a concentration ratio of 1:2:1:0.5 were analyzed. This experiment helped us to identify critical issues (e.g., dynamic range filtering in Step 3.2) in data analysis. In this experiment, the four samples were separately digested, labeled with the four isobaric reagents 1(114): 2(115): 1(116): 0.5(117), and then combined and analyzed by LC-MS/MS (Q-STAR, Applied Biosystems, USA). The raw data files thus obtained were searched against the SwissProt database using the MASCOT MS/MS ion search algorithm. All 10 proteins were confidently identified, and the search results were used as input for Multi-Q computation.

In this experiment, we observed that the resulting MS/MS spectra for each iTRAQ-labeled peptide displayed prominent y-series ions. Figure 3 shows a typical MS/MS spectrum of a triply charged peptide at *m/z* 643.44. This spectrum was used for peptide identification, and the –y and –b ions were confidently matched to the peptide of KVPQVSTPTLVEVSR (ion score: 49) from serum albumin. As shown in the insert, the cluster of peaks in the mass window *m/z* 113-118 represents the signature iTRAQ fragments used for quantitation. The four peaks showed a ratio of 1:2:1:0.5, which was consistent with the expected ratio of the standard protein mixture. Multi-Q was applied to calculate the peptide ratio from peak intensities and then calculate the protein ratio from peptide ratios. A summary of the original search results of the experiment and Multi-Q analysis output is listed in Table 1. (Complete results are listed in Supplementary Table 1, available at "Test Data" on http://ms.iis.sinica.edu.tw/Multi-Q/). To be specific, following the LC-MS/MS analysis, Multi-Q takes the following steps to analyze the input data files.

Step 1. Data Conversion and Annotation

The growing number of applications using mass spectrometry techniques has stimulated the development of different types of mass spectrometer. However, manufacturers often use different data storage formats for mass spectra acquisition. Multi-Q was therefore designed to accommodate various spectral data files from different mass spectrometers and search results from major search engines.

Step 1.1 Conversion of Spectral Data into mzXML files

In order to quantify relative protein expression levels, spectral data files from major MS manufacturers: WIFF (Applied Biosystems), RAW (Thermal Finnigan and Waters), and BAF (Bruker Daltonics) need to be converted into files of the mzXML[29] format, which is a commonly used data format for large-scale mass spectrometry-based protein quantitation. Multi-Q in particular provides a program called mzFast to convert a WIFF spectral file into an mzXML file. Furthermore, the converted mzXML file contains all necessary attributes, including MS and MS/MS peak lists, and the scan number index for further quantitation. It is worth mentioning that mzFast generates a "reduced" mzXML file to expedite the subsequent analysis. In the mzXML conversion process, a peak profiling format is used because it retains more detailed quantitation information than a peak centroid format. This optimizes signal reprocessing and facilitates manual verification. However, the size of a complete mzXML data file in the profiling mode is significantly larger than that of the original vendor's data format. To minimize the cost of data storage and the processing time, and increase the analysis throughput, the raw data is converted to a reduced mzXML file in which only MS/MS spectra containing quantitation information within a much reduced required mass range (*m/z* up to 120) are stored. In the study of the 10-standard-protein mixture, the file size (the raw data file is 13 MB) is significantly reduced from 70.5 MB for the complete mzXML profile to 1.5 MB for the reduced mzXML file. Interestingly, this function improves the system's speed substantially such that the conversion time in our experiment was

reduced from 150 minutes to 20 minutes. Hereafter, we refer to the reduced mzXML file as the mzXML file.

For RAW and BAF spectral files, since converters are already available to convert them into mzXML files, Multi-Q accepts the converted mzXML files as input. When users input their mzXML files, Multi-Q automatically determine to read"reduced"mzXML files for subsequent analysis so that the loading time can be greatly reduced.

Step 1.2 Data Annotation

In this step, Multi-Q accepts search results in different formats from MASCOT, SEQUEST, and X!Tandem and generates two data structures: PepTable and ProtTable. PepTable contains a peptide's identification confidence and degeneracy information (i.e., whether the peptide matches multiple proteins - discussed in Step 3.4); and ProtTable contains the probability of correct identification of each protein. Using these two data structures, Multi-Q can generate a protein summary list with annotated information. A user interface for visualization of the protein summary list is also provided so that users can browse the results easily.

To generate PepTable and ProtTable for MASCOT users, Multi-Q accepts output CSV files directly and uses the reported ion score as the peptide's identification confidence score. The ion score is defined as the probability that the observed database search match is a random hit[30]. In addition, for MASCOT users with WIFF spectral input files, Multi-Q provides a specific function to record a mapping of search results with the scan number index of the mzXML file. The scan number is then used for subsequent quantiation.

For SEQUEST and X!Tandem users, Multi-Q can accept their search results as input files in HTML and XML formats, respectively. Multi-Q then uses PeptideProphet[31] and ProteinProphet[32] of TPP (Trans-Proteomic Pipeline), developed by the Institute of Systems Biology, to transform the input files into a pepXML file and a protXML file, respectively. These two files are then automatically transformed into the data structures PepTable and ProtTable. PeptideProphet contains a rescoring

algorithm that determines the correctness of a match, which represents the peptide identification confidence score. For batch analysis, the individual peak list files (e.g., MGF files) from LC-MS/MS are merged and searched as a single file before pepXML conversion. In addition to HTML and XML input files, Multi-Q can also directly accept pepXML and protXML files directly as input files to generate the desired data structures.

Step 2. Peptide Level Processing

After the data has been prepared, Multi-Q selects iTRAQ labeled peptides with confident MS/MS identification, detects signature ions, and performs automated quantitation of peptide abundance.

Step 2.1 Signature Ion Detection and Background Subtraction

An identified peptide that is not iTRAQ-labeled may still generate signals with *m/z* of 114 to 117 in an MS/MS spectrum. However, such peptides are not suitable for further quantitation because they are not signatures of iTRAQ reagents. iTRAQ-labeled peptides can be easily identified by examining PepTable. Peptide ratio determination is performed by identifying the signature peaks in an MS/MS spectrum with *m/z* of 114 to 117 and comparing their peak intensities in different cell states. Special care should be taken when differentiating between true signature peaks and noise. Before selecting signature ions, we smooth mass spectra first by the 3-point moving average method. Smoothed spectra are then followed by signature peak selection. The mass tolerance of signature peak detection is defined by users based on the mass accuracy of the instrument. In this study, for example, the peak apexes within ±0.2 *m/z* of 114, 115, 116 and 117 are selected as signature peaks based on the mass accuracy of the quadrupole time-of-flight (Q-TOF) mass spectrometer. After peak selection, background subtraction is performed. The spectrum baseline is defined as the mean of all the valleys in the smoothed curve. Note that the valleys are determined by calculating the first and second order derivatives of the curve. The peak intensity is then calculated by subtracting the baseline from the original data.

Step 2.2 Isotope Impurity Correction and Peptide Ratio Determination

The iTRAQ quantitation strategy is based on the derivatization of quartet samples with isotopically distinct tags; and the relative peak intensities of the resultant isotope clusters are determined to represent changes in peptide abundance. These synthetic tags cause variations in the true peak intensity such that each batch of iTRAQ reagents contains trace levels of isotopic impurities that must be corrected. Since the isotopic distributions of the 114, 115, 116, and 117 signature peaks interfere with each other, overrepresentation of heavy signature ions will occur. For instance, the *m/z* 115 peak is a composite peak with contributions from various isotope envelopes of *m/z* 114, 116 and 117. Using the impurity information in the "Certificate of Analysis" provided by the iTRAQ reagent manufacturer, the interference of an isotopic reagent with its two predecessors and two successors can be corrected. The real ion counts of signature peaks, denoted by *x*s, can be derived from observed ion counts, denoted by *y*s, based on the following relationship:

$$
\begin{bmatrix} y_{114} \\ y_{115} \\ y_{116} \\ y_{117} \end{bmatrix} = \begin{bmatrix} 1 & k_{115,-1} & k_{116,-2} & 0 \\ k_{114,+1} & 1 & k_{116,-1} & k_{117,-2} \\ k_{114,+2} & k_{115,+1} & 1 & k_{117,-1} \\ 0 & k_{115,+2} & k_{116,+1} & 1 \end{bmatrix} \begin{bmatrix} x_{114} \\ x_{115} \\ x_{116} \\ x_{117} \end{bmatrix} \tag{1}
$$

where $k_{a,b}$ denotes the correction factor of iTRAQ reagent *a's* effect on its *b*-th predecessor/successor signature peak. For example, $k_{115,+2}$ denotes the correction factor of iTRAQ reagent 115's interference with the signature peak at *m/z* 117. All *x*s can be obtained easily by performing Gaussian Elimination[33].

After peak detection and isotope impurity correction, Multi-Q_Peptide, the peptide quantitation program, calculates the peptide ratios according to the peak intensities of the signature peaks. The relative peptide ratio $r_{A/B}(u)$ of a peptide *u* in two different cell states, A and B, is expressed as

$$
r_{A/B}(u) = \frac{|u_A|}{|u_B|} \tag{2}
$$

where $|u_A|$ and $|u_B|$ denote the abundance (after adjustment through the above steps) of peptide *u* in cell states A and B, respectively.

Step 3. Protein Level Processing

After determining the peptide ratios, Multi-Q reconstructs the proteins' abundance ratios from the ratios of their iTRAQ labeled peptides.

Step 3.1 Filtration of Peptides with Low Identification Confidence Scores

Multi-Q filters out peptides with low identification confidence scores so that only peptides with high scores are combined for protein quantitation. In Step 1.2 of the data conversion procedure, the identification confidence score of each peptide is recorded in PepTable. Users can employ the default score thresholds from the original search engines or empirically define confidence filters for correct identification.

In our study of the mixture of 10 standard proteins, a total of 901 MS/MS spectra were generated from LC-MS/MS and 55 confidently identified proteins were returned by MASCOT (with a total score > 42). However, most of the 55 identified proteins, including the 10 standard proteins and protein impurities, were derived from redundant hits of the same proteins from different species. All the search results were input to Multi-Q_Protein, and proteins were grouped according to their shared peptides.

Step 3.2 Determination and Application of a Dynamic Range Filter

Ideally, the ratio of two peptide peak intensities should faithfully reflect the abundance ratio of the peptides in two different cell states. However, in practice, the reliability of the peak ratio depends on many experimental parameters. For protein quantitation in the proteomic scale, the variation in experimental results can be caused by the sample preparation process and the performance of the equipment used for analysis[34]. Although quantitation errors may be caused by the poor S/N ratios of low abundance peptides and overlapping of isotopic profiles, it is inevitable that errors will arise because of the limited dynamic range of the instrument used. For protein quantitation, peptides ratios outside the dynamic range must be removed prior to peptide ratio normalization.

During the experiment, we observed that the peak intensities of high-abundance peptide ions may be underestimated by mass spectrometers. The following examples illustrate how peptide quantitation errors can arise from a saturated detector's responses. Figure 4A shows the MS/MS spectrum for the triple-charged iTRAQ labeled peptide (NTDGSTDYGILQINSR from Lysozyme C, P22910) in the m/z window 113-120. The peaks' clusters have an abundance ratio of 1:2:1:0.5 in the four samples, which conforms with the expected value. In contrast, the double-labeled peptide (which is labeled on the N-terminal and the lysine residue) from the same protein, CELAAAMK, exhibits nearly identical peak intensities (illustrated in Figure 4B) because of the saturated signals in m/z 114, 115, 116 and 117. The observed ratio does not reflect the expected ratio of 1:2:1:0.5.

The saturation effect of the detector's responses is a common phenomenon, particularly in nanospray or nano-LC experiments. It occurs when the number of ions entering a mass spectrometer is very high to produce peak saturation[35]. At the first glance, it appears that the saturation effect does not impose any limitations on the MS/MS operation, because the ion counts are usually significantly lower than those in the MS operation. However, the MS/MS operation in the iTRAQ experiment may suffer from the saturation effect and therefore requires further consideration. To minimize fluctuations in the quantitation of iTRAQ labeled peptides, the parameters for collision-induced dissociation in the MS/MS mode must be optimized in order to produce signature ions with good S/N ratios. As a result, saturated signature ions in m/z 114-117 may be observed for peptides with high mass spectral responses or for high abundance peptides.

To determine the signal saturation threshold of the instrument used (Q-STAR Applied Biosystems, USA), in Figure 5, we present a quantitative analysis of the saturation effect on the ion intensity. The analysis is essential since the resulting threshold can be applied to later experiments using the same instrument. In the figure, the global linear correlation of the paired peak intensity is shown for the total number of peptides in 114 versus 115, 114 versus 116, and 114 versus 117. In the 1:1 experiment (m/z 114: m/z 116, indicated by solid triangles), the least square regression of the scatter plot reveals a linear fit with a slope of 1.19, which conforms with the expected 1:1 ratio. However, on the plot of m/z 114

versus *m/z* 115, indicated by circles, we observe a non-linear curve with a deflecting trend when the peak intensity is slightly above 1,000 counts. A deviating quadratic trend is also observed in the *m/z* 114 versus *m/z* 117 experiment indicated by squares. The curve substantially deviates from the linear fit with a slope of 0.5 for intensities over 1,000 counts. This unstable quadratic behavior becomes more pronounced in signal intensities above a certain peak intensity threshold, i.e., the detector's saturation point due to saturation of the MS detector. As a result, the quantitation ratio is either an underestimation or overestimation outside the detector's saturation point.

To filter out errors in peptide ratios caused by the detector's saturation effect, Multi-Q allows users to define the peptides' saturation point values when running the package. Only peptides with intensity values below the user-defined saturation threshold are extracted for protein quantitation. Before inputting data into the Multi-Q system, we strongly recommend that first-time users apply a standard protein test with predetermined ratios to examine the saturation point values of their mass spectrometers. The saturation effect is instrument dependent, thus, the filtering thresholds should be carefully determined.

In addition to signal saturation, minimizing peptide ratio errors caused by signal fluctuations is another important issue. The degree of fluctuation depends on the chosen instrument's stability. To demonstrate the intensity dependence of measured peptide ratios, in Figure 6, the *m/z* 116/114 signal ratios of the 10-standard-protein mixture experiment are plotted as a function of the average ion counts of the signal *m/z* 116/114. For this analysis, we chose a peptide ratio of 1:1. As Figure 6 shows, there is a significant fluctuation in the 116/114 ratio of low intensity peptides compared to that of high intensity peptides. In particular, low intensity peptides, such as those with ion counts below 100, displayed extremely wide fluctuations. The broader distribution can be attributed to the inherent intensity-based stochastic processes in ion signal measurements by mass spectrometers[36-37].

To avoid the above-mentioned fluctuations and improve quantitation accuracy, Multi-Q determines an intensity threshold to remove peptide ratios derived from low intensity spectra. To achieve this, the

Multi-Q_Protein program analyzes the distribution of peptide ratios with a Gaussian fitting function (the insert in Fig. 7). The formula of the Gaussian curve is:

$$f(r, n_r, R, \sigma) = n_r \times e^{-(r-R)^2/(2\sigma^2)}$$ (3)

where r is a peptide ratio; n_r and σ denote the number and the standard deviation of r, respectively; and R is the mode (also the mean) of the fitted Gaussian distribution that represents the ratio of the most abundant peptides. In the insert of Figure 6, the standard deviation and mean of the fitted curve are 0.11 and 1.07, respectively. Many low abundance peptides lie in outlier groups of the fitted curve with a 99.7 % confidence interval ($\mu \pm 3\sigma$), where the three-sigma confidence interval for outlier group has been discussed in details in Kunal Aggarwal et al[18]. As they are considered deviate ratios that interfere with the determination of the true peptide ratios, they must be discarded before protein ratio reconstruction. Multi-Q_Protein calculates the intensity threshold for the low S/N ratio cut-off based on the average intensities of these outlier groups. After intensity filtration, approximately 99.7% of peptide ratios are determined to be suitable (within $\mu \pm 3\sigma$) for reconstructing protein ratios. Alternatively, the ion count filters for determining the dynamic range can be defined by users during the Multi-Q operation.

Step 3.3 Normalization of Peptide Ratios

   Appropriate normalization is essential in quantitation based on isotope labeling strategies in order to derive relative quantitation of proteins in different cell states. Most peptide ratios should be equal to 1 because, in principle, most protein expression levels in cells remain unchanged between two different cell states. As a result, most peptide ratios in an experiment should follow a normal distribution pattern with a mode (i.e., the highest value in the distribution) close to 1. In practice, however, several factors may contribute to the bias in mass spectrometry-based quantitation results. The primary noise sources in quantitation based on iTRAQ labeling are between-assay variations due to inconsistencies in protein digestion efficiency, iTRAQ labeling yield, and isotope impurities. Furthermore, between-sample variation may occur, as proteins are independently prepared and may have variable purity or concentrations. Consequently, there may be a uniform bias in favor of all peptide ratios such that most

17

ratios deviate from 1. Thus, a normalization procedure is necessary to derive the actual change in protein abundance with biological significance.

The most popular normalization method multiplies all peptide ratios by a global normalization factor, which can be based on the measured median ratios[38], external reference standard[39], or known invariant reference proteins. For high-throughput quantitative proteomics, a normal distribution should be chosen to minimize the effect of peptides with low S/N ratios on the normalization process[40].

After removing the ratios of peptides that are saturated or have low S/N intensity, Multi-Q performs a normalization procedure on the ratios within the dynamic range with the Gaussian distribution shown in Equation (3). The normalization factor, the reciprocal of R, can be used to correct the systematic bias. A detailed description of the normalization step for large-scale proteomics studies is given in the section on Quantitative Analysis of the Jurkat Cells with Multi-Q.

Step 3.4 Determination of Protein Abundance Ratios Based on Non-degenerate Peptides

Relative quantitation in the presence of degenerate peptides (i.e., peptides that match multiple proteins) due to sequence homology should be interpreted with caution[41]. Significant quantitation errors arise if a quantified peptide is not unique to its corresponding protein, i.e., it also belongs to other proteins. Consequently, we reconstruct the ratio of a protein by combining the ratios of all of its non-degenerate peptides.

Multi-Q_Protein selects non-degenerate peptides from the data structure PepTable and determines protein ratios according to their corresponding peptide ratios. In Multi-Q, the expression ratio $r_{A/B}(p)$ of a protein p in two cell states A and B is expressed as the weighted sum of all corresponding non-degenerate peptide ratios as follows:

$$r_{A/B}(p) = \sum_{u \in U} w_u r_{A/B}(u) \qquad (4)$$

where $r_{A/B}(u)$ is defined in Equation (2), U is the set containing all non-degenerate peptides of p, and w_u is the weight of the ratio $r_{A/B}(u)$ given by

18

$$w_u = \frac{(|u_A| + |u_B|)}{\sum_{v \in U}(|v_A| + |v_B|)} \qquad (5)$$

where $|u_A|$, $|u_B|$, $|v_A|$, and $|u_B|$ denote the abundance of peptides u and v in cell states A and B, respectively. Also, Multi-Q provides an additional option for calculating protein ratios by unweighted peptide ratios.

Finally, Multi-Q_Protein reports the quantitation output in the CSV data format.

Quantitative Analysis of Jurkat Cells with Multi-Q

To further evaluate Multi-Q's suitability for large-scale proteomics studies, two identical protein mixtures were prepared from human Jurkat T cells. The cell lysate was first fractionated with SDS-PAGE. One SDS-PAGE band was excised, tryptic digested, and divided into two sets, which were labeled with iTRAQ reagents 114 and 116, respectively. The expected protein expression level in both sets is 1:1. The raw data and the search results of MASCOT against the CDS (Celera Discovery System) protein database (version dated 08/13/2003) were input for Multi-Q computation.

To avoid potential false positive identification in a large database, we ran a database search against a reverse sequence database to determine the minimum score threshold for peptide identification. Based on the results of this procedure, we selected a MASCOT Ion Score of 19.5 as the filter threshold for a normal forward search to reduce possible false positives in the identification results. Using the above parameter, Multi-Q generated the partial protein identification results and iTRAQ ratios listed in Table 2 (complete result are listed in Supplementary Table 2, available at our web site), in which 48 proteins were identified from 353 isolated iTRAQ-labeled peptides. The result of this experiment yields a mean peptide ratio of 0.999±0.07.

For peptide quantitation, only the isotopic signature MS/MS patterns belonging to the above iTRAQ-labeled peptides were used to calculate the peptide ratios. We observed that approximately 7 % of the identified peptides were derived from unlabeled peptides. In protein ratio determination, degenerate peptides were excluded from the calculations. By applying user-defined intensity filters, the Multi-Q

package automatically selects peptides within the appropriate dynamic range and adequate confidence levels to normalize the peptide ratios and calculate protein ratios based on the weighted peptide intensity. Even without a normalization step, the 48 proteins quantified with a 99% confidence level yield a mean peptide ratio of 1.27 ± 0.08, which is reasonably accurate.

Normalization of Data

In the dataset of 353 filtered peptides shown in Figure 7, the distribution of the peptide abundance ratio peaks at 1.26 ± 0.01, irrespective of the size of the normalization interval. We note that the ratio is higher than the normalization factor of isotope impurity. Based on the Multi-Q procedure described in Step 3.3, the value of the normalization factor is the reciprocal of the original mode (the highest value in the distribution) in the peptide ratio distribution, in which the mode depends on the selected interval size. As shown in Figure 7(A), the distribution modes (without normal fitting) for different ratio intervals (0.1, 0.075 and 0.05) differ by 0.1. Therefore, a more stable normalization factor is required to correct the systematic errors, regardless of the selected interval size. After fitting with a normal distribution, however, the fit with different sizes of ratio intervals (0.1, 0.075 and 0.05) yields similar normalization factors for peptide ratios, as shown in Figure 7(A). The modes of the fitted curves with ratio intervals 0.10, 0.075, and 0.050 are 1.276, 1.267, and 1.253, respectively. Therefore, normalization factors based on normal fitting are relatively more stable than those calculated by methods that use mean or mode values or invariant reference proteins.

To determine the confidence level of quantitation, the distribution of the protein abundance changes after peptide-level normalization with Multi-Q_Protein was fitted with a normal distribution (Figure 7B). The protein abundance ratios from 0.7 to 1.3 were fitted into a Gaussian distribution of mean 0.999 and standard deviation 0.07 within a 95% confidence interval ($p < 0.05$). Compared with the average protein-level standard deviation (0.081) derived from the unprocessed dataset, the narrower normal distribution of protein ratios with standard deviation 0.07 clearly demonstrates better quantitation accuracy, which can be attributed to the use of intensity-weighted peptide ratios that minimize the

standard deviation. Furthermore, the results show that accurate quantitation results with statistical confidence levels can be obtained automatically; thus, time-consuming manual inspection is unnecessary.

Conclusion

Accurate, robust, high-throughput computation is a key issue in the analysis of differential proteomic expressions. The quantitative analysis package Multi-Q presented in this paper incorporates several useful features. First, it provides a program to convert WIFF spectral files into "reduced" mzXML files for subsequent quantitation, and the conversion time is substantially reduced. Though Multi-Q accepts mzXML files converted from RAW and BAF files directly as input, Multi-Q automatically determine to read reduced mzXML files so that the loading time can also be reduced. In addition, Multi-Q accepts search results from different search engines, including MASCOT, SEQUEST and X!Tandem, and provides a user interface for visualization of the protein summary list. Most importantly, Multi-Q allows users to define their own data-filtering thresholds based on statistical modeling or semi-empirical judgments so that false-positive fold changes can be substantially reduced. This feature allows Multi-Q to be used for data generated by various MS instrument types with different dynamic ranges. In addition to isotope impurity correction, it incorporates a normalization function to correct systematic errors in peptide ratios. Finally, since Multi-Q only assigns non-degenerate peptides for protein quantification, it improves quantitation accuracy. Our experiments demonstrate that Multi-Q yields precise and fully automatic protein quantitation. Thus, rapid interpretation of output from large proteomic datasets can be achieved without manual validation. In summary, Multi-Q is a fully automated quantitation tool suitable for large-scale proteomics studies. In the future, we will adapt this tool to multiplexed quantitative analysis for other *in vitro*/*in vivo* labeling strategies with various isotopic tags.
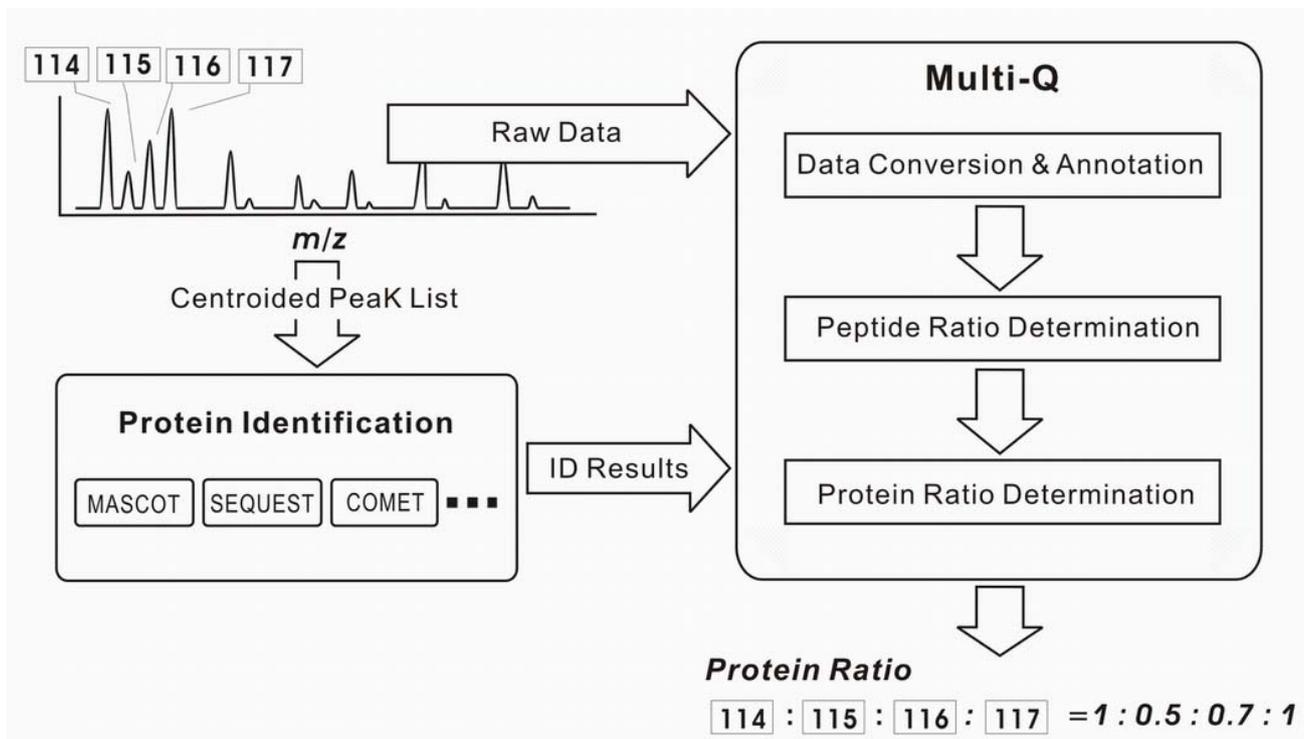
Figure 1. Multi-Q quantitation system. Multi-Q package accepts two data inputs: Raw data from mass spectra and protein identification results from search engines such as MASCOT, SEQUEST, and X!Tandem. Data then undergo data conversion, peptide ratio determination, protein ratio determination, and final protein relative abundance ratios are produced.
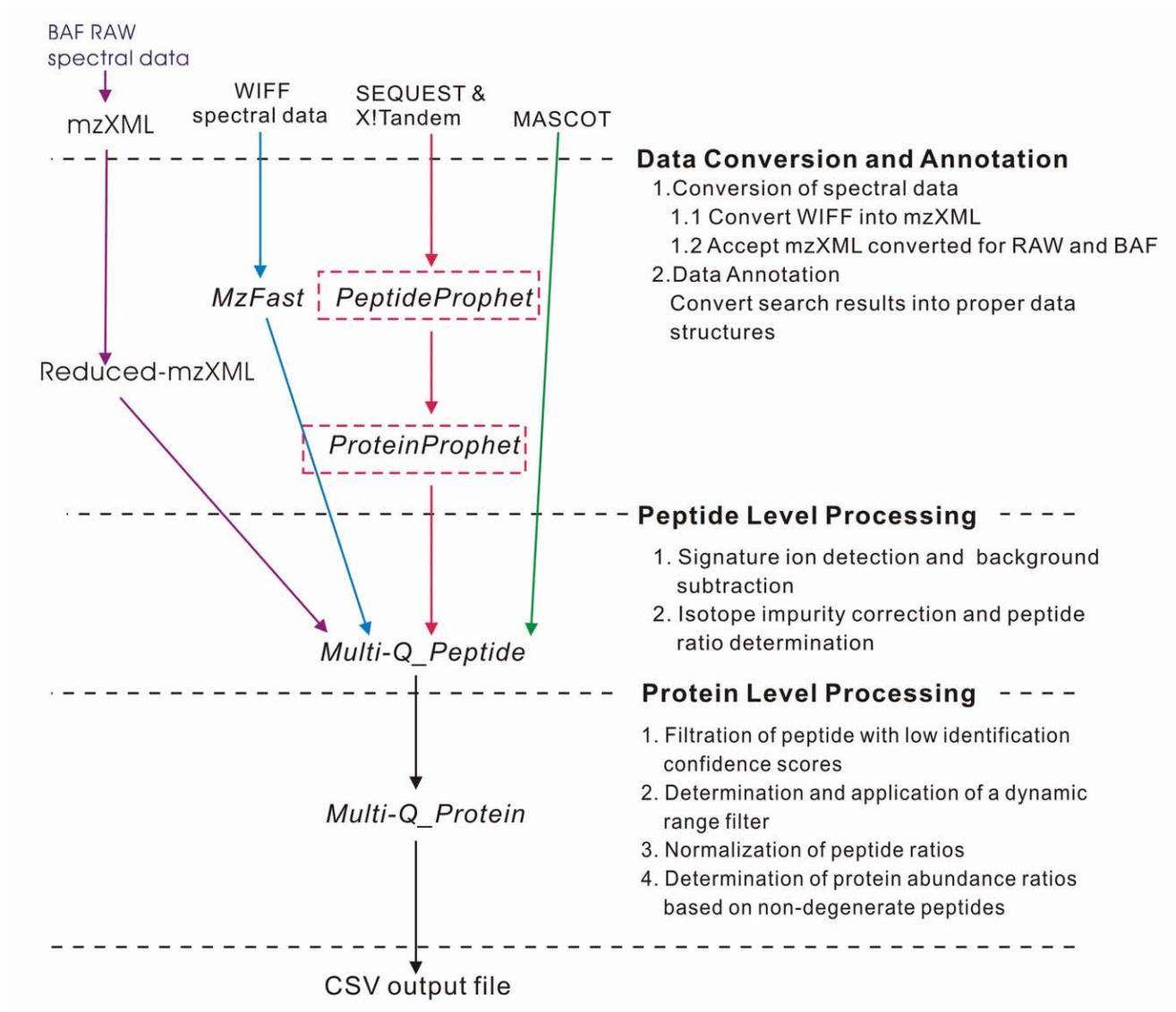
Figure 2. Flowchart of the Multi-Q package for determining protein quantitation. Here the dash line delimits three main steps of Multi-Q: Data Conversion and Annotation, Peptide Level Processing, and Protein Level Processing.
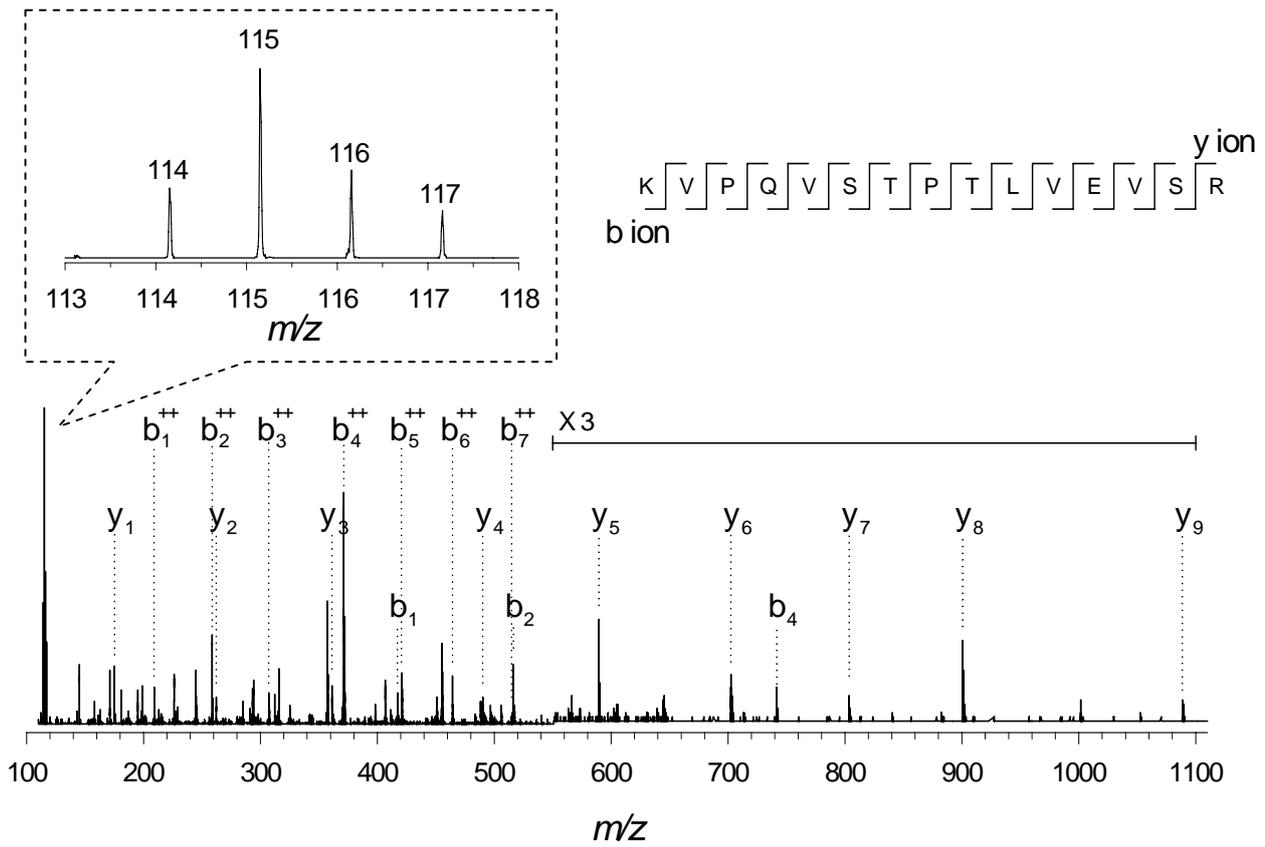
Figure 3. MS/MS spectrum of peptide KVPQVSTPTLVEVSR from a tryptic digest mixture of 10 standard proteins in a 1:2:1:0.5 ratio. Zoomed-in figure shows the signature peaks (*m/z* 114, 115, 116, and 117).
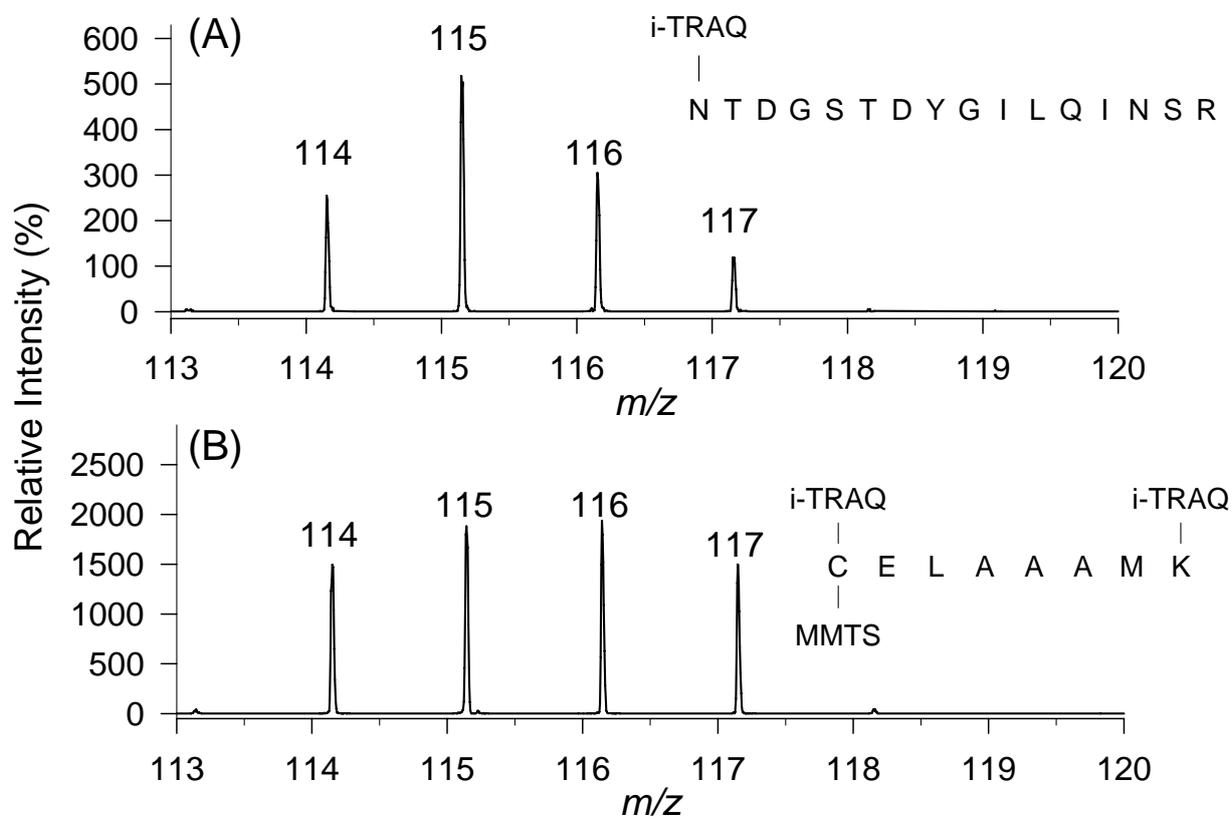
Figure 4. MS/MS spectra of iTRAQ labeled peptides with different signature peak abundance in *m/z* 113-120 regions: (A) The average signature peak intensity reveals clearly 1:2:1:0.5, in agreement with the expected value of NTDGSTDYGILQINSR from Lysozyme C, P22910; (B) Due to signal saturation effect, the average signature peak intensity of CELAAAML, also from Lysozyme C, P22910, shows nearly identical peak intensities.
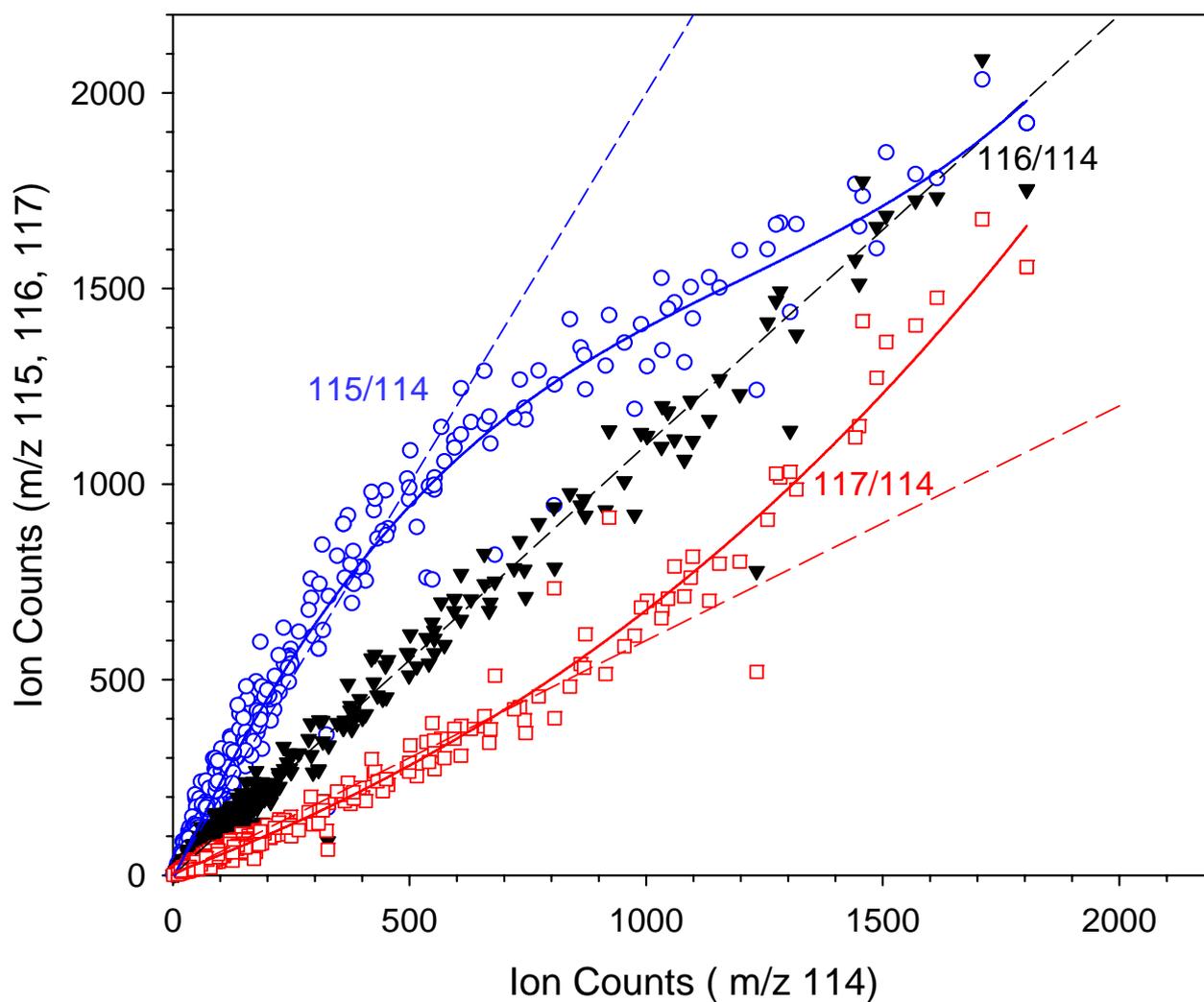
Figure 5. The distribution of peak intensities of peptides with two different labels are shown by circles for 115/114 (with expected ratio 2), triangles for 116/114 (with expected ratio 1), and rectangles for 117/114 (with expected ratio 0.5). The dashed lines represent the linear regression of unsaturated ion counts, while the solid curves represent fitting curves for ion counts of all peptides. Deflecting trends toward 1:1 ratio (x=y) are observed when the peak intensity is over approximately 1000 counts
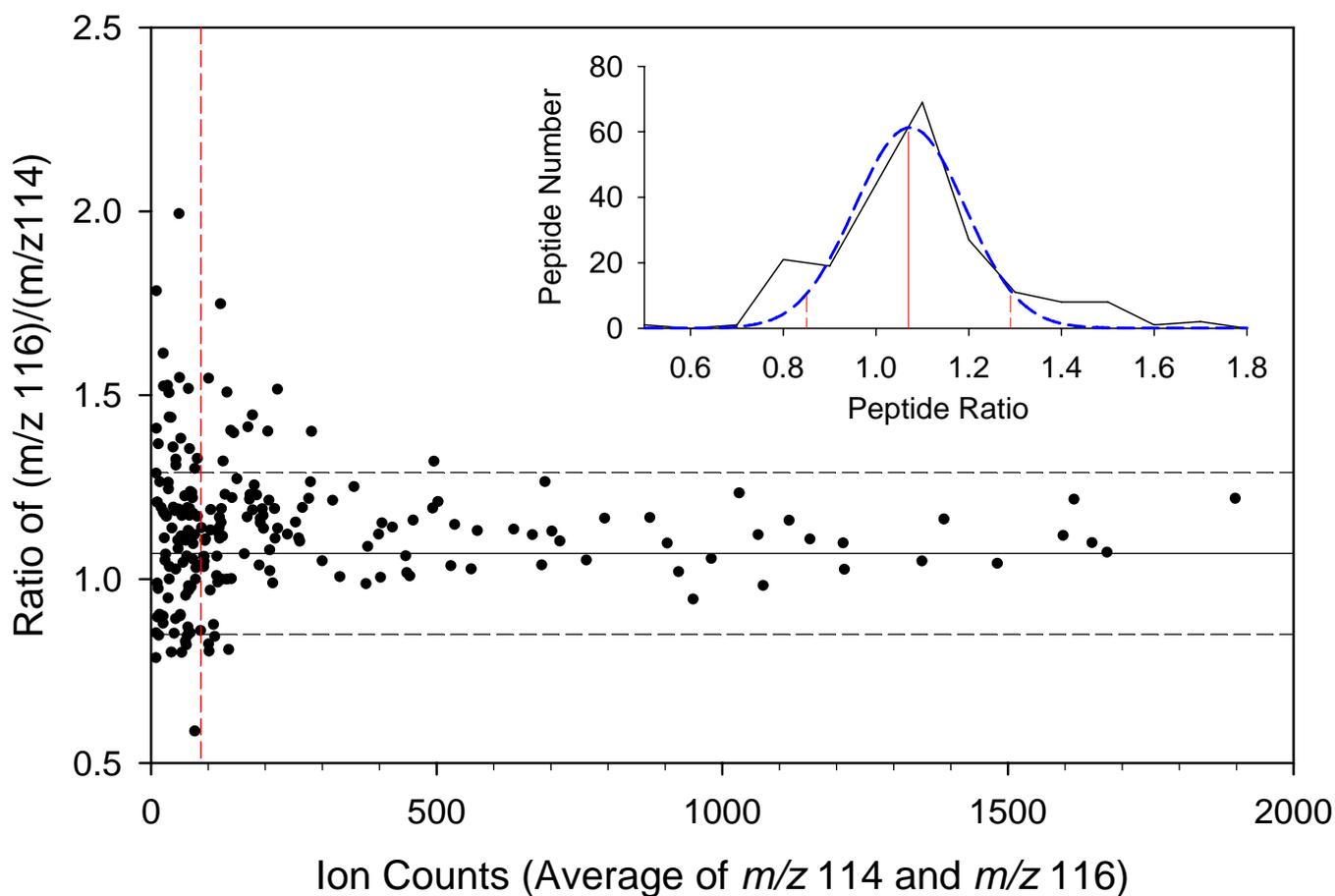
Figure 6. Effect of ion count on measured peptide ratio. The 116/114 ratios from the 10-standard-protein mixture are plotted as a function of ion counts. Fluctuation of peptide ratio is greater in low ion counts than that in high ion counts. *Inset* shows the distribution of peptide ratios, where the solid line is the original data distribution, and the dashed line is the Gaussian fitting curve over the original data. The ratios belonging to the outlier groups (95% confidence interval) of the fitting curves are considered as deviated ratios.
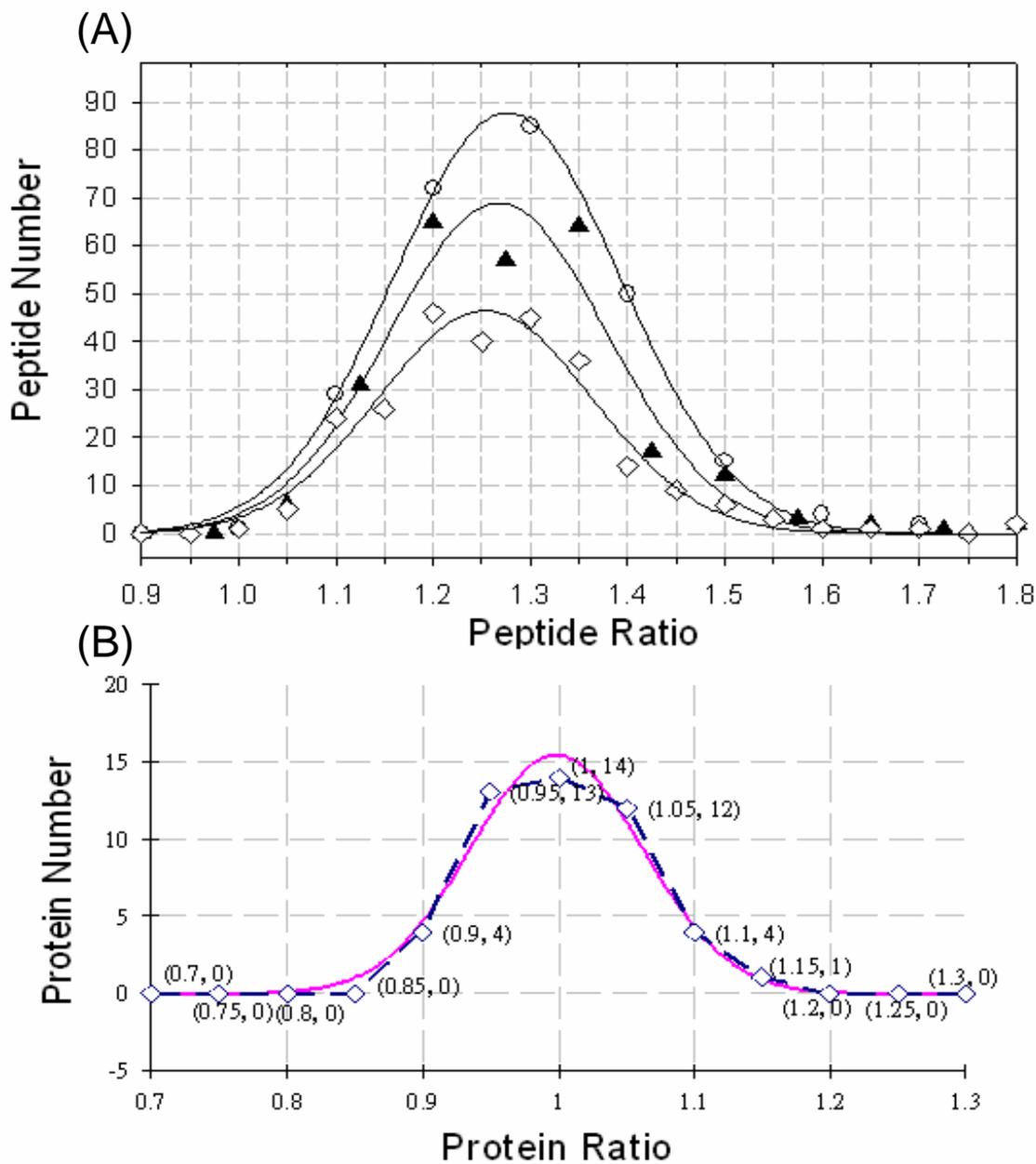
Figure 7. Distribution of peptide ratios in the Jurkat cells experiment. (A) Effect of granularity of peptide ratios: The original distributions of peptide ratios with intervals 0.1, 0.075, and 0.05 are designated by circles, triangles, and diamonds, respectively. Depending on interval size, the mode of original distributions of peptide ratio varies from 1.2 to 1.3. However, after Gaussian fitting, the peak apices are 1.276, 1.267, and 1.253.

(B) The distribution of protein abundance ratios after peptide-level normalization: The dashed lines are the original distribution of protein ratios, while the solid line is the curve of protein ratio distribution after normalization.

Table 1. Summary of protein identification and quantification results in 10 standard proteins

Accession Number	Protein Name	ID Prob. [a]	Peptide Number	114:114	114:115 [b]	114:116	114:117
(P00330)	Alcohol dehydrogenase I	1.00	8	1	2.11±0.25	1.04±0.14	0.55
(Q9P4C2)	Alcohol dehydrogenase II	1.00	4	1	2.40±0.18	1.19	0.56
(P33409)	Alpha-S2 casein precursor	1.00	1	1	2.36	1.26	0.63
(Q9TSI0)	Beta casein precursor	1.00	4	1	2.22±0.13	1.27±0.21	0.65±0.02
(P00921)	Carbonic anhydrase II	1.00	11	1	2.27±0.43	1.19±0.19	0.53±0.04
(P02866)	Concanavalin A precursor	1.00	3	1	2.26	1.02	0.51
(P00004)	Cytochrome c	1.00	7	1	2.48±0.24	1.20±0.01	0.55±0.07
(P00924)	Enolase 1	1.00	17	1	2.20±0.21	1.07±0.10	0.55±0.04
(P02668)	Kappa casein precursor	1.00	3	1	2.35±0.28	1.22±0.21	0.56
(P22910)	Lysozyme C	1.00	3	1	2.02	1.11	0.43
(P00701)	Lysozyme C precursor	1.00	1	1	2.2	1.16	0.61
(P02188)	Myoglobin	1.00	8	1	2.61±0.49	1.34±0.39	0.59±0.03
(P02769)	Serum albumin precursor	1.00	20	1	1.99±0.29	1.14±0.07	0.56±0.04
(P00760)	Trypsinogen	1.00	8	1	2.14±0.44	1.14±0.22	0.57±0.12

a)   Protein probability obtained from ProteinProphet

b)   Protein ratio without uncertainty interval is calculated by one peptide.

Table 2. Partial summary of protein identification and quantification results in Jurkat cell.

Accession Number	Protein Name	ID Prob.[a]	Mascot Score	Peptide Number	114:116[b]	S.D.
(P08238)	Heat shock protein HSP 90-beta	1.00	806	9	0.96	0.10
(P11142)	Heat shock cognate 71 kDa protein	1.00	737	8	0.95	0.07
(P07900)	Heat shock protein HSP 90-alpha	1.00	580	22	0.95	0.07
(P13796)	L-plastin	1.00	569	22	0.99	0.09
(P11021)	78 kDa glucose-regulated protein precursor	1.00	384	11	1.00	0.06
(P19338)	Nucleolin	1.00	379	13	0.98	0.10
(P04264)	Keratin, type II cytoskeletal 1	1.00	338	11	1.04	0.15
(P08133)	Annexin A6	1.00	330	8	1.04	0.05
(P31939)	Bifunctional purine biosynthesis protein PURH	1.00	329	8	0.96	0.07
(P13639)	Elongation factor 2	1.00	283	6	0.99	0.18
(P26038)	Moesin	1.00	280	3	1.01	0.08
(P14625)	Endoplasmin precursor	1.00	240	7	0.93	0.08
(P13010)	ATP-dependent DNA helicase II, 80 kDa subunit	1.00	237	6	0.92	0.09
(P35908)	Keratin, type II cytoskeletal 2 epidermal	1.00	233	5	0.94	0.07
(P49368)	T-complex protein 1, gamma subunit	1.00	225	4	1.04	0.08
(P38646)	Stress-70 protein, mitochondrial precursor	1.00	202	6	1.01	0.04
(P12956)	ATP-dependent DNA helicase II, 70 kDa subunit	1.00	198	9	1.00	0.11
(P29401)	Transketolase	1.00	183	4	0.99	0.14
(P33993)	DNA replication licensing factor MCM7	1.00	161	3	0.93	0.03
(Q14974)	Importin beta-1 subunit	1.00	158	4	1.05	0.18
(P55072)	Transitional endoplasmic reticulum ATPase	1.00	137	5	1.01	0.12

a) Protein probability obtained from ProteinProphet

b) ratio of *m/z* 114: *m/z* 116 after normalization

Acknowledgement

References

1.   Ornstein, D. K.; Gillespie, J. W.; Paweletz, C. P.; Duray, P. H.; Herring, J.; Vocke, C. D.; Topalian, S. L.; Bostwick, D. G.; Linehan, W. M.; Petricoin, E. F., III; Emmert-Buck, M. R., Proteomic analysis of laser capture microdissected human prostate cancer and in vitro prostate cell lines. *Electrophoresis* **2000**, *21,* (11), 2235-42.

2.   Jones, M. B.; Krutzsch, H.; Shu, H.; Zhao, Y.; Liotta, L. A.; Kohn, E. C.; Petricoin, E. F., III, Proteomic analysis and identification of new biomarkers and therapeutic targets for invasive ovarian cancer. *Proteomics* **2002**, *2,* (1), 76-84.

3.   Zhou, G.; Li, H.; DeCamp, D.; Chen, S.; Shu, H.; Gong, Y.; Flaig, M.; Gillespie, J. W.; Hu, N.; Taylor, P. R.; Emmert-Buck, M. R.; Liotta, L. A.; Petricoin, E. F., III; Zhao, Y., 2D differential in-gel electrophoresis for the identification of esophageal scans cell cancer-specific protein markers. *Mol Cell Proteomics* **2002,** *1,* (2), 117-24.

4.   Alaiya, A. A.; Franzen, B.; Moberger, B.; Silfversward, C.; Linder, S.; Auer, G., Two-dimensional gel analysis of protein expression in ovarian tumors shows a low degree of intratumoral heterogeneity. *Electrophoresis* **1999**, *20,* (4-5), 1039-46.

5.   Franzen, B.; Linder, S.; Alaiya, A. A.; Eriksson, E.; Fujioka, K.; Bergman, A. C.; Jornvall, H.; Auer, G., Analysis of polypeptide expression in benign and malignant human breast lesions. *Electrophoresis* **1997,** *18,* (3-4), 582-7.

6. Bini, L.; Magi, B.; Marzocchi, B.; Cellesi, C.; Berti, B.; Raggiaschi, R.; Rossolini, A.; Pallini, V., Two-dimensional electrophoretic patterns of acute-phase human serum proteins in the course of bacterial and viral diseases. *Electrophoresis* **1996**, *17,* (3), 612-6.

7. Scheler, C.; Li, X. P.; Salnikow, J.; Dunn, M. J.; Jungblut, P. R., Comparison of two-dimensional electrophoresis patterns of heat shock protein Hsp27 species in normal and cardiomyopathic hearts. *Electrophoresis* **1999,** *20,* (18), 3623-8.

8. Davidsson, P.; Westman-Brinkmalm, A.; Nilsson, C. L.; Lindbjer, M.; Paulson, L.; Andreasen, N.; Sjogren, M.; Blennow, K., Proteome analysis of cerebrospinal fluid proteins in Alzheimer patients. *Neuroreport* **2002,** *13,* (5), 611-5.

9. Zabel, C.; Chamrad, D. C.; Priller, J.; Woodman, B.; Meyer, H. E.; Bates, G. P.; Klose, J., Alterations in the mouse and human proteome caused by Huntington's disease. *Mol Cell Proteomics* **2002,** *1,* (5), 366-75.

10. Klose, J., Large-gel 2-D electrophoresis. *Methods Mol Biol* **1999,** *112,* 147-72.

11. O'Farrell, P. H., High resolution two-dimensional electrophoresis of proteins. *J Biol Chem* **1975,** *250,* (10), 4007-21.

12. Gygi, S. P., and Aebersold, R., *Proteomics: A Trends Guide.* 2000, 32.

13. Gygi, S. P.; Corthals, G. L.; Zhang, Y.; Rochon, Y.; Aebersold, R., Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc Natl Acad Sci U S A* **2000,** *97,* (17), 9390-5.

14. Gygi, S. P.; Han, D. K.; Gingras, A. C.; Sonenberg, N.; Aebersold, R., Protein analysis by mass spectrometry and sequence database searching: tools for cancer research in the post-genomic era. *Electrophoresis* **1999,** *20,* (2), 310-9.

15. Desiderio, D. M.; Kai, M, *Biomed. Mass Spectrom*, **1983**, *10*, 471

16. Ong, S.E.; Blagoev, B.; Kratchmarova, I.; Kristensen, D.B.; Steen, H.; Pandey, A.; Mann, M., Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* **2002,** *1,* (5), 376-86.

17. Ross, P.L.; Huang, Y.N.; Marchese, J.N.; Williamson, B.; Parker, K.; Hattan, S.; Khainovski, N.; Pillai, S.; Dey, S.; Daniels, S.; Purkayastha, S.; Juhasz, P.; Martin, S.; Bartlet-Jones, M.; He, F.; Jacobson, A.; Pappin, D. J., Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics* **2004,** 3, (12), 1154-69.

18. Aggarwal K.; Choe L.H.; Lee K.H., Quantitative analysis of protein expression using amine-specific isobaric tags in Escherichia coli cells expressing rhsA elements. *Proteomics* **2005,** *5,* 2297-2308.

19. Hardt, M.; Witkowska, H.E.; Webb, S.; Thomas, L.R.; Dixon, S.E.; Hall, S.C.; Fisher, S. J., Assessing the effects of diurnal variation on the composition of human parotid saliva: quantitative analysis of native peptides using iTRAQ reagents. *Anal Chem* **2005,** *77*, (15), 4947-54.

20. Callister, S.J.; Barry R.C.; Adkins, J.N.; Johnson, E.T.; Qian, W., Webb-Robertson B.M., Smith R.D., and Lipton M.S., Normalization Approaches for Removing Systematic Biases Associated with Mass Spectrometry and Label-Free Proteomics. *J. of Proteome Research*, **2006**, *5,* 277-286

21. Kristensen, D.B.; Brond, J. C.; Nielsen, P.A.; Andersen, J.R.; Sorensen, O.T.; Jorgensen, V.; Budin, K.; Matthiesen, J.; Veno, P.; Jespersen, H.M.; Ahrens, C.H.; Schandorff, S.; Ruhoff, P.T.; Wisniewski, J.R.; Bennett, K.L.; Podtelejnikov, A.V., Experimental Peptide Identification Repository (EPIR): an integrated peptide-centric platform for validation and mining of tandem mass spectrometry data. *Mol Cell Proteomics* **2004,** *3,* (10), 1023-38.

22. Skoog, D. A.; Holler, F. J.; Nieman, T.A., *Principle of Instrumental Analysis*, 5th Edition; Brooks Cole Publishing, 1997.

23. Shadforth, I.P.; Dunkley, T.P.; Lilley, K.S.; Bessant, C., i-Tracker: for quantitative proteomics using iTRAQ. *BMC Genomics* **2005,** *6,* 145.

24. MacCoss, M. J.; Wu, C. C.; Liu, H.; Sadygov, R.; Yates, J. R., III, A correlation algorithm for the automated quantitative analysis of shotgun proteomics data. *Anal Chem* **2003,** *75,* (24), 6912-21.

25. Schulze, W. X.; Mann, M., A novel proteomic screen for peptide-protein interactions. *J Biol Chem* **2004,** *279,* (11), 10756-64.

26. Eng, J.K.; McCormack, A.L.; Yates, J.R., III, An approach to correlate tandem mass spectral data of peptides with amino acid sequences in protein database. *J. Am. Soc. Mass Spectrom.* **1994,** *5,* 976-989

27. 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,* 2310-2316

28. Perkins, D.N.; Pappin, D.J.C.; Creasy, D.M.; Cottrell, J.S., Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 1999, *20,* 3551-3567

29. 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.

30. http://www.matrixscience.com

31. Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R., Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* **2002,** 74, (20), 5383-92.

32. Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R., A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* **2003,** *75,* (17), 4646-58.

33. Golub, G.H.; Van Loan, C.F., *Matrix Computations,* 3rd edition; The Johns Hopkins University Press: USA, 1996.

34. Anderle, M.; Roy, S.; Lin, H.; Becker, C.; Joho, K., Quantifying reproducibility for differential proteomics: noise analysis for protein liquid chromatography-mass spectrometry of human serum. *Bioinformatics* **2004,** *20,* (18), 3575-82.

35. Chernushevich, I. V.; Loboda, A. V.; Thomson, B. A., An introduction to quadrupole-time-of-flight mass spectrometry. *J Mass Spectrom* **2001,** *36,* (8), 849-65.

36. MacCoss, M. J.; Wu, C. C.; Liu, H.; Sadygov, R.; Yates, J. R., 3rd, A correlation algorithm for the automated quantitative analysis of shotgun proteomics data. *Anal Chem* **2003,** *75,* (24), 6912-21.

37. MacCoss, M. J.; Toth, M. J.; Matthews, D. E., Evaluation and optimization of ion-current ratio measurements by selected-ion-monitoring mass spectrometry. *Anal Chem* **2001,** *73,* (13), 2976-84.

38. Yang, Y. H.; Dudoit, S.; Luu, P.; Lin, D. M.; Peng, V.; Ngai, J.; Speed, T. P., Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res* **2002,** *30,* (4), e15.

39. Ravin, N. V.; Ravin, V. K., Use of a linear multicopy vector based on the mini-replicon of temperate coliphage N15 for cloning DNA with abnormal secondary structures. *Nucleic Acids Res* **1999,** *27,* (17), e13.

40. Li, X.J.; Zhang, H.; Ranish, J.A; Aebersold, R., Automated statistical analysis of protein abundance ratios from data generated by stable-isotope dilution and tandem mass spectrometry. *Anal. Chem.* **2003,** *75,* 6648-6657

41. Nesvizhskii, A. I.; Aebersold, R., Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics* **2005,** 4, (10), 1419-40.