Chapter 12
Software Development for Quantitative Proteomics Using Stable Isotope Labeling

Xin Huang and Shi-Jian Ding

Abstract Stable isotope labeling (SIL) coupled with liquid chromatography and high-resolution tandem mass spectrometry (MS) are increasingly useful for elucidation of the proteome-wide differences between multiple biological samples. Developments of more effective programs for the relative peptide/protein abundance measurements are essential for quantitative proteomic analysis. In this chapter, we present a quantification program, termed UNiquant, for analyzing quantitative proteomic data using SIL. The common steps in a quantitative proteomic software, such as MS data preprocessing, peptide identification, peptide quantification, and protein quantification, were dissected in this chapter, using UNiquant as an example. UNiquant was used to analyze the SILAC-labeled proteome mixtures with known heavy/light ratios (H/L = 1:1, 1:5, and 1:10). The pros and cons of the quantification results of UNiquant from two different MS acquisition modes, data-dependent acquisition and data-independent acquisition, were also evaluated and compared.

Keywords Software development • UNiquant • Stable isotope labeling • Mass spectrometry • Quantitative proteomics • Data-dependent acquisition • Data-independent acquisition

X. Huang, Ph.D.
Department of Pathology and Microbiology, University of Nebraska Medical Centre, Omaha, NE, USA

S.-J. Ding, Ph.D. (✉)
Department of Pathology and Microbiology, University of Nebraska Medical Centre, Omaha, NE, USA

Mass Spectrometry and Proteomics Core Facility, University of Nebraska Medical Center, Omaha, NE, USA
e-mail: dings@unmc.edu

X. Wang (ed.), *Bioinformatics of Human Proteomics*, Translational Bioinformatics 3,
DOI 10.1007/978-94-007-5811-7_12, © Springer Science+Business Media Dordrecht 2013

12.1 Introduction

Mass spectrometry-based quantitative proteomics is an emerging field capable of making a unique contribution to the understanding, prevention, and cure of human diseases (Choudhary and Mann 2010; Gstaiger and Aebersold 2009; Koomen et al. 2008). Proteomic analysis now involves larger and more reliable datasets, mostly generated using state-of-the-art mass spectrometry (MS) combined with a bottom-up (or shotgun) profiling of whole protein complements from cells, tissues, and body fluids (Mann and Kelleher 2008). Proteomics has an advantage over genomic-based assays because it offers direct examination of the molecular machinery of cell physiology, including protein expression, cell signaling, and posttranslational modifications (PTMs).

A major hurdle in quantitative proteomics is still identifying and subsequent quantifying of proteins and their expression levels in complex biological systems (Venable et al. 2004). In quantitative shotgun proteomics, proteolysis-derived peptides are commonly measured with LC-MS/MS and are used as surrogates of their parent proteins for relative quantification (Mann and Kelleher 2008; Ong and Mann 2005). In a label-free approach, proteomes under comparison are analyzed separately in standardized LC-MS/MS runs. Peptide intensities, spectra counts, and extracted ion chromatography (XIC) are used to measure the protein abundances (Fang et al. 2006; Finney et al. 2008). Alternatively, by employing stable isotope labeling (SIL), the proteomes under comparison are combined and analyzed together in one LC-MS/MS run. Comparison of the signal intensities of the same peptides and their SIL analogues yields an estimate of protein abundances (Geiger et al. 2010a; Mann 2006). In general, SIL methods minimize variability during sample processing steps and LC-MS/MS analyses and provide results with less systematic error and higher reproducibility compared to the label-free approach (Qian et al. 2010). On the other hand, absolute quantification of proteins can be obtained through the use of a stable isotope-labeled internal standard (Silva et al. 2006a).

Development of software for quantitative proteomics with SIL has made tremendous advances in this field. A number of academically developed software tools, such as ASAPRatio (Li et al. 2003), ProRata (Pan et al. 2006), RelEx (Venable et al. 2004), Xpress (Han et al. 2001), Census (Park et al. 2008), MaxQuant (Cox and Mann 2008), Vista (Bakalarski et al. 2008), WaveletQuant (Mo et al. 2010), UNiquant (Huang et al. 2011a, b), and recently IsoQuant (Liao et al. 2012), have been produced to analyze SIL-based quantitative proteomic datasets. In these experiments, information on peptide abundance is derived either from the intensity of the peptide precursor ion signal at full mass spectra or from the intensity of reporter ions after MS/MS fragmentation. The first catalog includes isotope-coded affinity tagging (ICAT), stable isotope labeling with amino acids in cell culture (SILAC), and $^{16}O/^{18}O$ labeling, while the second catalog includes tandem mass tags (TMT) and isobaric tag for relative and absolute quantitation (iTRAQ). Most of the programs target either precursor ion or reporter ion quantitation. For the precursor ion-based quantitation, low-intensity MS signals present a substantial challenge to

quantify low-abundance proteins by various programs (Bakalarski et al. 2008). Different programs adopt different strategies for distinguishing the peptide signals from the background noise. Using a data-dependent acquisition (DDA) mode, MS/MS fragmentation is performed on the most abundant precursor ions. On hybrid high-resolution mass spectrometry such as LTQ Orbitrap (Thermo Scientific, San Jose, CA), the precursor scan is performed in an Orbitrap analyzer, and the MS/MS fragmentation is usually accomplished in the linear ion trap mass analyzer. In these experiments, an LC-MS/MS data collection cycle starts with a high-accuracy, full MS survey of the all precursor ions and is followed by selecting a number of the intensive precursor ions for MS/MS fragmentation (Wilm 2009). Recently, the data-independent acquisition (DIA) strategy was developed to complement the DDA method for proteomic analysis (Venable et al. 2004). Instead of a serial selection of precursor ions for data-dependent fragmentation, the DIA approach fragments a group of co-eluting precursor ions at each given time, enabling a more unbiased detection of all LC-eluted peptides compared to the DDA method (Ramos et al. 2006; Williams et al. 2003).

Here, we are going to describe the in-house developed UNiquant software for quantitative proteomic MS data analysis with SIL. The major procedures in a quantitative software, including MS data preprocessing, detecting pairs, reading intensity, normalization, and performance and compatibility at different MS platforms, will be introduced by analyzing SILAC-labeled eukaryotic cells with known heavy-versus-light ratios.

12.2 Materials and Methods

12.2.1 Prepare SILAC Protein Mixture with Known Ratios

The human cell lines Jeko-1 and MDA-MB-231 cells were grown in either SILAC "light" (L-arginine and L-lysine) or "heavy" (L-$^{13}C_6$-arginine and L-$^{13}C_6$-lysine for Jeko-1, L-[$^{13}C_6$,$^{15}N_4$]-arginine and L-[$^{13}C_6$,$^{15}N_2$]-lysine for MDA-MB-231) medium for 2 weeks (more than five cell cycles). The heavy and light lysates were harvested mixed in three heavy/light (H/L) ratios: 1:1, 1:5, and 1:10, then followed by sample pretreatments and tryptic digestion (Huang et al. 2011a, b).

12.2.2 LC-MS/MS Analysis with DDA and DIA

In the DDA analysis, the LTQ Orbitrap mass spectrometer automatically switches between MS and MS/MS acquisition modes. In each MS cycle (about 2.5 s), a survey full-scan MS spectra (m/z 375–1,575) were acquired in the Orbitrap with resolution $R = 100{,}000$, then the most five intense ions (depending on signal intensity

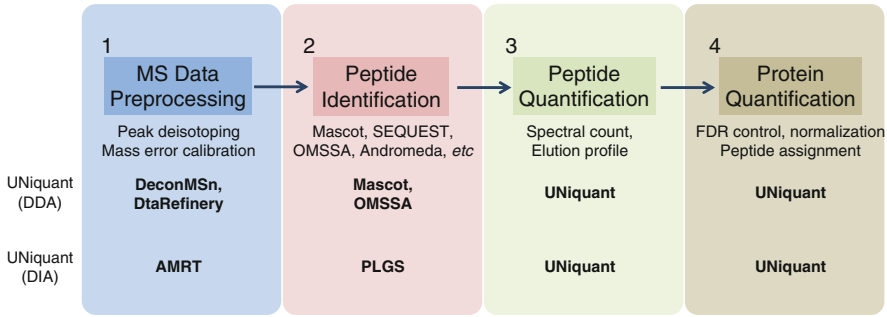


Fig. 12.1 The common steps in a quantitative proteomic software and the components in UNiquant for DDA and DIA

of survey full scan) were sequentially isolated for fragmentation in the linear ion trap using collision-induced dissociation (CID). Former target ions selected for MS/MS were dynamically excluded for 75 s. In the DIA analysis, the SYNAPT G2 mass spectrometer (Waters Co., Milford, MA) equipped with time-of-flight (TOF) analyzer was used. The Tri-Wave ion guides trap and separates precursor ions by ion mobility. Then, the CID cell was operated alternatively with low-energy and elevated energy survey of acquisitions (Bateman et al. 2002). The acquisition time in each mode was 1.0 s with an interscan delay of 0.1 s. In the low-energy mode for the survey full MS scan (*m*/*z* 300–2,000), precursor intensities were collected at constant collision energy (5 eV). In the elevated energy mode for the MS^E scan (MS/MS scan), collision energy was ramped from 15 to 40 eV during each collection cycle.

12.3 UNiquant Software for Quantitative Proteomics

12.3.1 Overview of Quantitative Proteomic Software

Protein identification and quantification are the two major components for a quantitative proteomic program. However, a quantitation software also needs other program components. Figure 12.1 shows four steps that a quantitation program usually involved.

Step 1: MS data preprocessing. Initially, the vender-specific MS raw data need to be converted to the text-formatted peak list files such as dta or mgf files or a common file format using extensible markup language (XML), such as mzXML and mzML.

Step 2: peptide identification. In this step, peptide is identified from the MS/MS peak list through a process of peptide-spectrum match (PSM) using programs such as SEQUEST (Eng et al. 1994), Mascot (Perkins et al. 1999), OMSSA (Geer et al. 2004), and Andromeda (Cox et al. 2011) to compare

the observed peak list to a protein database. Identification by these algorithms is based on a restricted database search in which MS/MS spectra are aligned with protein sequences, probably bearing a few specified PTMs attached to specific amino acids.

Step 3: peptide quantification. In label-free methods, the spectra count or the normalized XIC profiles of a peptide were measured as the intensity of identified peptide. In SIL methods, quantification programs fetch the XIC elution profiles of the heavy and light peptides from the MS raw data (or XML data) in full MS scans (SILAC, ICAT, $^{16}O/^{18}O$, etc.) according to the identified peptide sequences in the same LC-MS/MS runs. For MS/MS scan (iTRAQ, TMT)-based quantification, intensity of reporter ions is measured within the same MS/MS spectrum that a peptide was identified.

Step 4: protein quantification. The quantification results at the protein level are reported by assigning the peptide sequences to different protein IDs. To ensure the confidence of the identification and quantification results, a false discovery rate (FDR) method is used to estimate the false-positive results in the final reports. FDR should be monitored and reported (usually 0.01) at the spectrum, peptide, and proteins levels. Furthermore, peptide intensities are normalized (if needed) to reduce the deviation of quantified ratios if the heavy and light samples are not equally mixed. In a user's view, the result of a quantitative proteomic experiment is a report of thousands of proteins, with their intensities or relative intensity ratios between the heavy and light species.

12.3.2 UNiquant Software for Quantitative Proteomics with DDA

To date, most of the quantitative proteomic data were obtained using the method. In these experiments, an LC-MS/MS data collection cycle starts with a high-accuracy, full MS survey of the all precursor ions and is followed by selecting a number of precursor ions for MS/MS fragmentation (Wilm 2009). A major advantage of DDA is that the fragment ions are derived mostly from a single precursor ion, increasing the specificity of peptide identification. As shown in Fig. 12.1, UNiquant chooses the third-party academic-free softwares DeconMSn and DtaRefinery for MS data preprocessing. The DDA version of UNiquant uses the identification results of Mascot and OMSSA (open source and freely available for academic users) search engines. Finally, UNiquant program is developed for peptide and protein quantification using the outputs from first two steps.

12.3.3 Data Preprocessing and Peptide Identification

In UNiquant, Thermo MS raw data are converted to mgf-formatted MS/MS peak list file before Mascot search. DeconMSn (http://omics.pnl.gov/software/) was used to determine and refine the monoisotopic mass and charge state of parent ions and

to create the peak list files. Next, DtaRefinery (http://omics.pnl.gov/software/) is used to improve mass measurement errors for parent ions by modeling systematic errors based on putative peptide identifications. For the SILAC protein mixture data used in this study, Mascot search engine was used for peptide identification. To ensure the quality of the identification results, usually a "target-decoy" database search strategy (Elias and Gygi 2010) was applied by searching against a concatenated database containing the authentic protein sequences (forward database) and the reverse sequences of all proteins involved (reverse database). Then, an FDR estimator is calculated to access the confidence of the identification results. Previous studies have shown that the PSM score given by the search engines and mass accuracy of the precursors are two important parameters for discriminating the forward and reverse identifications (Ding et al. 2008).

In UNiquant coupled with Mascot as the search engine, quality of peptide identification (QPI) score is calculated by

$$\mathrm{QPI} = s \times e^{-1/2} \tag{12.1}$$

where s is Mascot peptide identification score and e is the mass error (ppm) of the precursor ions which is calculated as

$$e = 1{,}000{,}000 \times \frac{(m_{\text{observed}} - m_{\text{theoretical}})}{m_{\text{theoretical}}} \tag{12.2}$$

where m_{observed} and $m_{\text{theoretical}}$ are the observed accurate mass and theoretical mass of the precursor ions of the peptide, respectively. A Mascot score cutoff of 10 was applied for all identification results. QPI of a peptide was taken as the sum of the QPI for all MS/MS spectra that were matched to this peptide sequence. Identified peptides were sorted by a descending order of QPI values, and a cutoff was applied to ensure a total FDR < 0.01.

12.3.4 Intensity Measurement of Precursor Ions

Precursor ion intensity, measured in the high-resolution full MS, was extracted by UNiquant and used as an abundance measurement for each identified peptide. The input files for quantitation are the Thermo Xcalibur MS data (.raw) and the peptide identification output dat (Mascot) and csv (OMSSA) files. UNiquant also utilizes the search results from other search engines with text-formatted outputs containing the filtered peptide sequence, identification score, scan number, observed m/z, and charge state information. The quantitation algorithm in UNiquant was first developed for hybrid FT-MS instruments (Ding et al. 2008). Briefly, theoretical mass for a peptide (labeled or unlabeled) is calculated according the peptide sequence identified in the MS/MS spectrum and the SIL method. Then, the corresponding high-accuracy, full MS scan which derives the MS/MS spectrum is determined. A search is performed on this MS spectrum within a small range (<20 ppm)

to localize the heavy and light precursor ions. Intensities of both precursor ions are measured with a signal-to-noise (S/N) ratio above 2.0. The output of UNiquant is a tab-delimited text file which includes a list of peptides with refined *m/z*, mass error, S/N ratios, and intensities of light and heavy species.

12.3.5 Peptide and Protein Quantification

A peptide usually appears more than once in the LC-MS/MS data. The spectra count is the number of times that a peptide identified by database search. The relative abundance of each identified peptide was calculated as the sum (based on spectra counts) of peak intensities (PI) for the heavy species of the peptide divided by the sum of intensities for the light species of the peptide:

$$\text{Ratio}_{\text{H/L}} = \frac{\sum_n \text{PI}_{\text{H}}}{\sum_n \text{PI}_{\text{L}}} \tag{12.3}$$

where n is the spectra count for a specific peptide, PI_{H} is the peak intensity of the heavy species, and PI_{L} is the peak intensity of the light species. Similarly, the relative abundance of each identified protein was calculated by dividing the sum of the intensities of all peptide heavy species for the protein by the sum of the intensities of all peptide light species.

12.3.6 Post-measurement Normalization

The post-measurement normalization is needed for correcting the unequal mixing of heavy and light proteins in the quantitative proteomic experiments. In UNiquant, a locally weighted scatterplot smoothing (LOWESS) method was used to correct the H/L ratios of quantified peptides (Cleveland 1979). Briefly, LOWESS method is based on minus-add (M-A) plot of the peptide intensities for the heavy and light species:

$$M = \log_2 \frac{\left(\text{Int}_{\text{heavy}}\right)}{\left(\text{Int}_{\text{light}}\right)} \tag{12.4}$$

$$A = \frac{1}{2}\log_{10}\left(\text{Int}_{\text{heavy}} \times \text{Int}_{\text{light}}\right) \tag{12.5}$$

where $\text{Int}_{\text{heavy}}$ is the intensity of the heavy species from a quantified peptide, while $\text{Int}_{\text{light}}$ is the intensity of the corresponding light species of this peptide. M is the $\log_2$ H/L intensity ratio, and A is half of the $\log_{10}$ H × L intensity product of each quantified peptide. These *M*–*A* points were equally divided into 20 groups, based on their A-values. A linear regression line was obtained from the points in each group, and

then a fitted regression curve was obtained by connecting all the regression lines. Normalization was performed by subtracting the fitted curve from the measured $\log_2$ H/L ratio in the *M–A* plot:

$$M' = \log_2\left(\frac{\text{Int}_{\text{Heavy}}}{\text{Int}_{\text{light}}}\right) - c(A) = \log_2\left(\frac{\text{Int}_{\text{Heavy}}}{k \times \text{Int}_{\text{light}}}\right) \tag{12.6}$$

where $c(A)$ is the fitted LOWESS curve, which is a function of A. M' is the normalized log ratio of quantified peptides, which is obtained by subtracting the value of LOWESS fitting function from the measured log ratio at each value of A.

12.3.7 UNiquant Software for Quantitative Proteomics with DIA

Recent development of the DIA strategy has been introduced as a complement methodology of the DDA strategy for quantitative proteomic experiments. It has been implemented on two MS platforms: Exactive Orbitrap (Thermo Scientific) and SYNAPT G2 (Waters Co). The corresponding DIA method was named as all-ion fragmentation (Geiger et al. 2010b) and LC-MS^E technology (Silva et al. 2006b; Vissers et al. 2009), respectively. UNiquant was recently developed for analyzing the proteomic data on the LC-MS^E platform (Huang et al. 2011b). As shown in Fig. 12.1, UNiquant also covers the last two steps of peptide and protein quantification of the DIA proteomic data, while the MS raw data preprocessing and peptide identification are performed by the ProteinLynx Global Server (PLGS, Waters Co.) software. In the first step, ion detection, clustering, and retention time alignment are processed using an AMRT (accurate mass retention time) method in PLGS (Silva et al. 2005). Next, the AMRT data are searched against the Swiss-Prot protein database using a dual-pass algorithm in PLGS (Li et al. 2009).

Procedures for quantification of the LC-MS^E DIA data are similar to the procedures in the DDA approach. The AMRT files are exported to a local Microsoft Access database. Included in this output are the weight-averaged monoisotopic mass, charge state, ion drift, charge-state-reduced sum intensity, observed apex retention time, and observed start and stop time of the detected ions. Information for the identified peptides is exported to a table file as well. This contains all the theoretical and experimental properties associated with the identified precursor MS spectrum, such as the unique spectrum id, mass over charge (m/z), retention time, peptide sequence, and the identification scores. Theoretical masses of the heavy and light precursors are determined from the peptide sequence and the SIL method. The predicted precursors are used to search the AMRT database for an observed ion that matched the criteria of mass accuracy, elution time, and ion drift. The default settings were mass accuracy <5 ppm, difference in retention time <0.05 min, and difference in ion drift ≤0.5. Intensities of the SIL pair of precursors are extracted, and the heavy/light ratios were sorted and arranged with the peptide sequence and protein entry. Similar to DDA method, the relative H/L ratios of identified peptide are calculated as the

sum of the intensities for the heavy precursors divided by the sum of the intensities for the light precursors.

12.4 Results and Notes

12.4.1 Implementation of UNiquant

UNiquant is an in-house software for the quantitative proteomic data analysis with SIL. The software is developed on the platform of Microsoft .NET Framework (version 2.0). The programming languages are Microsoft VB.NET and C#. It now has two components, for analyzing the DDA and DIA data, respectively. As shown in Fig. 12.2, the DDA version of UNiquant has incorporate the softwares of DeconMSn and DtaRefinery for MS raw data deisotoping and mass error calibration, respectively. Next, the OMSSA database search engine was embedded in UNiquant as the default engine for peptide identification. UNiquant is also compatible with other engines such as Mascot. But the users need to upload the mgf files to the Mascot Daemon Server (Matrix Science) for peptide identification and obtain the output dat files for further quantification. Peptide quantification is performed in the component named "precursor search," fetching the information of precursor intensity from the MS raw data by the Xcalibur Development Kit (XDK) provided by the instrument vendor. Precursor search can be performed for individual LC-MS runs by a user-friendly "drag and drop" process or automatically performed by the UNiquant piper. Finally, intensities of the peptides from different LC-MS runs are merged and

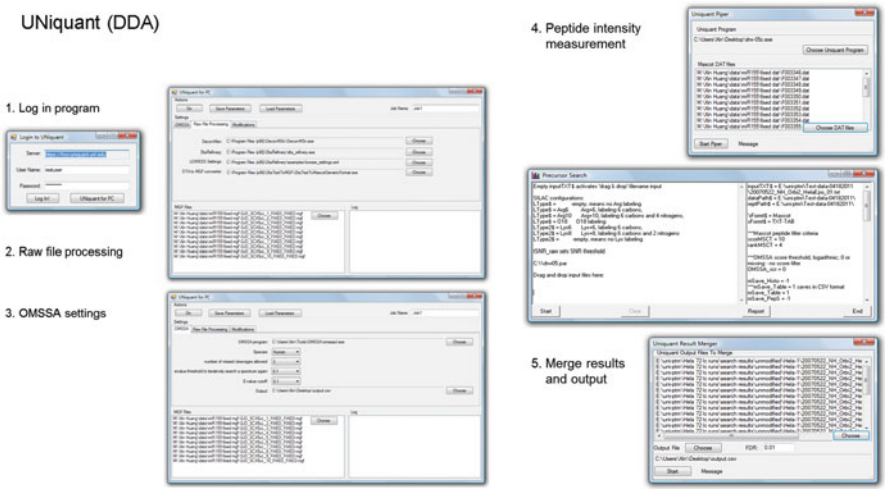


Fig. 12.2 Implementation of UNiquant for *DDA* data analysis. The program components and user interface (*UI*) in different steps are indicated as well

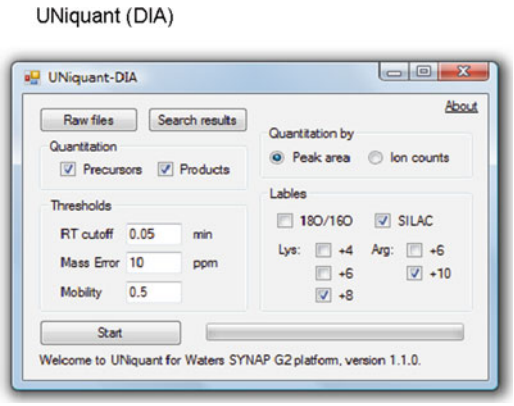


Fig. 12.3 Implementation of UNiquant for DDA data analysis

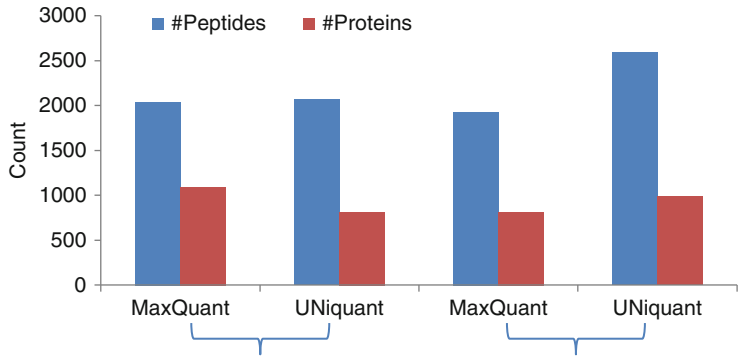
filtered by an FDR threshold and annotated by unique protein ID for protein level outputs of the quantification results.

The user interface of UNiquant for DIA is shown in Fig. 12.3. Here, the UNiquant program reads the deconvoluted MS raw files (AMRT) and peptide identification results (with fixed FDR confidence) from the Waters PLGS software. UNiquant outputs all the peak lists into a local Microsoft Access database, and searching of the heavy and light precursors was performed by the SQL queries with Microsoft Office Development in Visual Studio (ODVS) components. Matching of the SIL heavy and light precursor ions is performed based on the similarity of retention time (default setting, <0.05 min) and ion mobility (<0.5) and the accurate mass (<10 ppm) of difference between the heavy and light precursors. Finally, the DIA version of UNiquant summarizes the quantification results and output the protein level H/L ratios.

12.4.2 Analysis of the SILAC Proteome Mixture with Known H/L Ratios

We analyzed the SILAC-labeled proteome digests with known H/L ratios (H/L = 1:1 and 1:10). For identification of the peptides/proteins, we used the same database search engine (Mascot), with identical searching parameters, and searched the data against the same proteome database (IPI version 3.52). Using the same FDR cutoff of 0.01, the number of peptide pairs and proteins being identified by each program is shown in Fig. 12.4. UNiquant and MaxQuant identified nearly equal numbers of peptide pairs in the H/L = 1:1 mixture data. For the H/L = 1:10 proteome data, UNiquant identified 34 % more peptide pairs and proteins than MaxQuant. However, the number of quantified proteins is similar for these two programs.

Before normalizing of the quantification results, the median log ratio of peptides quantified by UNiquant was generally equal to the true value of the log ratio in each proteome mixtures (Fig. 12.4). In the H/L = 1:1 proteome mixture, the median log ratio of peptides quantified by UNiquant was −0.029, which is closer to the true log



H/L =	1:1 (true log-ratio = 0)			1:10 (truelog-ratio = -1)		
Peptides	N	Median	SD	N	Median	SD
MaxQuant	2038	-0.113	0.096	1929	-1.229	0.189
UNiquant	2063	-0.029	0.176	2591	-0.968	0.393

Fig. 12.4 Peptide and proteins quantified by MaxQuant and UNiquant for the standard SILAC mixtures (H/L = 1:1 and 1:10). The true values of $\log_{10}$ H/L ratios are indicated, and the statistics of quantified peptides for each mixture were tabulated in the following table

ratio = 0, compared to −0.113 obtained by MaxQuant. Similarly, the median ratio log ratios quantified by UNiquant and MaxQuant are −0.968 and −1.229, respectively, for the H/L = 1:10 proteome mixture (true log ratio = −1). The frequency of the log ratios quantified by both UNiquant and MaxQuant is generally Gaussian distributed in all mixture data, but with different variances. The standard deviation of the log ratios quantified by MaxQuant is lower than the log ratios quantified by UNiquant.

Different programs provided complementary results of quantified proteins. UNiquant and MaxQuant chose different strategies for SIL-pair detection. By the scenario of DDA, the selected isotopic peaks for MS/MS fragmentation can be derived from either light or heavy peptides. UNiquant does not detect SIL pair of peptide before identification. After database search, theoretical masses for both the heavy and light peptides were determined, and intensities were calculated based on the confident identifications. This strategy was also applied by other programs such as Vista and IsoQuant. In contrast, MaxQuant uses an alternative strategy for peak pair detection, one which identifies pairs of light and heavy peptides from the MS data prior to peptide identification (Cox and Mann 2008). Advantages of the strategy in MaxQuant are that the resulting peak list is much cleaner than the peak list from the original raw data, the peaks have a high S/N ratio, and co-eluting peptides can be readily identified. However, this strategy may result in some loss of pairs, especially in the case of peptide pairs with low-intensity or high-noise background. Such as the case of H/L = 1:10 data, MaxQuant quantified more

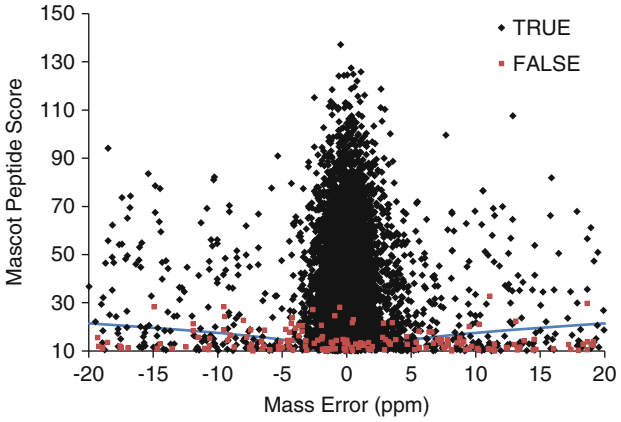


Fig. 12.5 The distribution of Mascot peptide identification score and mass error in the H/L = 1:1 mixture. Peptides from the forward database were labeled as *black points*, and peptides from backward database were labeled as *red points*. The *blue curve* indicates the QPI cutoff. *Points* under the cutoff were removed from the quantitation

peptides than MaxQuant but increases the variance of the quantified results as the compensation.

Furthermore, the way for calculating FDR is slightly different between UNiquant and MaxQuant. MaxQuant corrects the mass precision of precursors and used a posterior error probability based on the peptide *P*-score distribution by different categories of peptide length to set the cutoff of FDR (Cox and Mann 2008; Olsen and Mann 2004). UNiquant used QPI (Eq. 12.1), an estimator involving the peptide identification score and mass error of the precursor ions. Figure 12.5 shows the distribution of the Mascot peptide score versus the mass error of precursor ions for our SILAC datasets. The false peptides have lower peptide score and higher mass errors compared to the true peptides. The blue line in Fig. 12.4 shows the QPI cutoff in this dataset to remove all the low-confidence identifications.

12.4.3 Post-measurement Normalization

We plotted the quantification results of standard SILAC mixture data with DDA in Fig. 12.6. Before normalization, the identified and quantified peptides from all three proteome mixtures with known ratios (H/L = 1:1, 1:5, and 1:10) are plotted by their $\log_2$ (H/L) intensity ratios versus the $\log_{10}$ (H × L) intensity products (Fig. 12.6a–c). The $\log_2$ (H/L) ratios of quantified peptides show a comet-like distribution from the three mixtures. The region of low-abundance peptides generally has higher variance $\log_2$ (H/L) ratios compared to the high-abundance peptides. In H/L = 1:5 and 1:10 mixtures, the data in low-intensity region tends to a log ratio of 0, whereas they should have a value of −2.32 for the 1:5 data and a value of −3.32 for the 1:10 data.

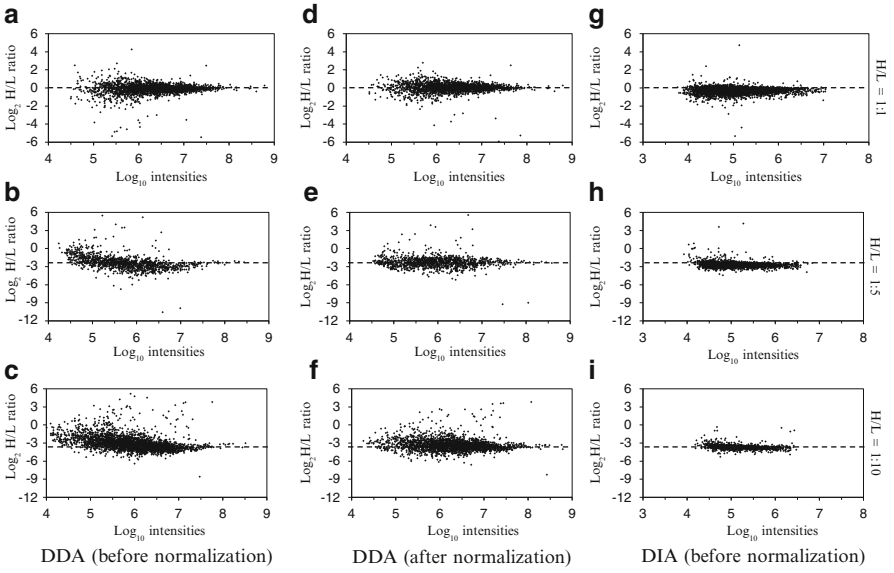


Fig. 12.6 Accuracy of quantitation for SILAC data analysis is compared between the DDA analysis on LTQ Orbitrap platform (**a**–**c**) before and (**d**–**f**) after normalization and the DIA analysis on (**g**–**i**) SYNAPT G2 platform for three proteome mixtures with H/L = 1:1 (**a**, **d**, **g**), 1:5 (**b**, **e**, **h**), and 1:10 (**c**, **f**, **i**). In each scatterplot, the quantified peptides were distributed by their $\log_2$ (H/L) intensity ratios versus $\log_{10}$ (H × L) intensity products. The true $\log_2$ (H/L) ratio is indicated as a *dashed line* for the H/L = 1:1 ($\log_2$ ratio = 0), 1:5 ($\log_2$ ratio = −2.32), and 1:10 ($\log_2$ ratio = −3.32) mixtures, respectively

The LOWESS method corrects and straightens the median ratios of quantified peptides by different categories of intensity and straightens the LOWESS regression curve into a straight line. As shown in Fig. 12.6d–f, the log ratios of peptides in the H/L = 1:1 mixture were similar between and after normalization. But the log ratios of the low-abundance peptides in the H/L = 1:5 and 1:10 mixtures are corrected to the true ratios of −2.32 (H/L = 1:5) and −3.32 (H/L = 1:10), respectively. In the contract, the log ratios of the high-abundance peptides do not change too much in these two mixtures.

In quantitative proteomic data analysis, a normalization approach is usually applied by assuming that the amounts of most proteins in the sample will be unchanged by the variable being tested. The purpose of normalization is to overcome the effects of unequal mixing of the heavy and light species during the sample preparation. Thus, the averaged heavy/light abundances of all the quantified proteins can be adjusted to one. In MaxQuant and UNiquant, the relative peptide/protein abundances before and after normalization are both provided (Cox and Mann 2008). MaxQuant uses the median-center method for normalization, but UNiquant uses the LOWESS method. However, this normalization approach may not be applied in some cases especially if specific portion of proteins are enriched, such as the phosphoproteins (by molecular function) or the nuclear proteins (by cellular components). For instance,

the use of phosphatase inhibitors will affect a broad range of cellular phosphorylation events. Therefore, the assumption of normalization is not valid if only the phosphoproteome was investigated. Furthermore, the assumption that the majority of proteins are unchanged might be incorrect when the specific treatment could affect a broad range of protein concentration, such as transcription factor and microRNA. So a quantitative proteomic solution for accurate relative protein abundance measurement is still necessary.

12.4.4 Comparison of the Quantification Results from DDA and DIA

Peptide quantitation results obtained on the SYNAPT G2 MS with DIA and on the LTQ Orbitrap MS with DDA were compared. In the SYNAPT G2 analysis, $\log_2$ (H/L) ratios of quantified peptides show a more uniform distribution for each of the three mixtures (Fig. 12.6g–i). In the H/L = 1:5 and 1:10 mixtures, the $\log_2$ (H/L) ratios are closer to the expected ratios (−2.32 and −3.32). In the H/L 1:1 mixtures, the dynamic ranges ($\log_{10}$ intensities) of both the LTQ Orbitrap data and the SYNAPT G2 data are about 4 orders of magnitude (Fig. 12.6d, g). In H/L = 1:5 and 1:10 mixtures, the dynamic range of LTQ Orbitrap data is still 4 orders of magnitude (Fig. 12.6e, f), whereas the range of the data from the SYNAPT G2 drops to 3.5 orders of magnitude in the H/L = 1:5 mixture and to 3.0 in the 1:10 mixture (Fig. 12.6h, i).

Currently, the LTQ-FT/Orbitrap MS with DDA is the major MS platform for SIL-based quantitative proteomic applications. With this platform, the MS scans are used for peptide quantitation, while MS/MS scans are used for peptide identification. Because the number of precursor ions selected for MS/MS fragmentation is fixed in the DDA mode, the total number of peptides identified for a given protein from complex proteome mixtures is relatively low due to the limited number of MS/MS spectra that can be generated for peptide identification. The DIA approach is an interesting alternative to complement DDA for SIL-based quantitative proteomic analysis. First, it provides more time for MS/MS fragmentation, making it possible to identify more peptides. Second, the high-resolution MS and MS/MS data makes quantitation possible from both precursor and product ions.

Dynamic range for protein quantitation is one of the key features of quantitative proteomic analysis. Using the DIA method, we identified more proteins in the 1:5 and 1:10 mixtures compared to the 1:1 mixture; however, the number of SIL-peptide pairs and the dynamic range of protein intensities decreased in the 1:5 and 1:10 mixtures. The decrease in peptide pairs and dynamic range of protein intensities is mainly due to the loss of low-intensity heavy peptides and to saturation of high-intensity light peptides. That occurred because the peak detection algorithm used a cutoff for acceptable peaks that was based on peak intensity and signal-to-noise ratio. These excluded small peaks which impacted the number of heavy peptides observed but ensured that the selected peaks were actually peptides. Additionally,

the dynamic range of protein intensities was consistent (about 4 orders of magnitude) in the DDA analysis of the three complex proteome mixtures with different ratios.

Accurate quantitation of protein abundance is an essential task for MS instruments and its associated data analysis tools. Overall, the SYNAPT G2 with DIA approach showed better quantitation accuracy and reliability than the LTQ Orbitrap with DDA analysis presumably due to the fundamental difference between these two mass analyzers (Pan et al. 2006; Bakalarski et al. 2008). In a TOF analyzer such as the SYNAPT G2, the signal intensity comes from direct ion counting, and many spectra are accumulated up to 10,000 specs per second. Each TOF spectrum usually has a small dynamic range, and a collection of multiple spectra can increase it. If the TOF analyzer is optimized for high sensitivity such as in the case of this study, the SYNAPT G2 instrument gives correct intensity measurements for low-intensity ions but saturated readings for high-intensity ions. Thus, the very high-intensity ions are discriminated against in the final results because the saturated ions increase internal errors of both measured intensity and mass accuracy. In the Orbitrap analyzer, signal intensities are obtained by Fourier transformation of an ion signal induced on the detection electrodes of the Orbitrap cell (Hu et al. 2005). Just one ion signal spectrum is sufficient to obtain a full m/z spectrum with a high dynamic range in ion intensities. In addition to these signal detection differences, the front part ion optics for the two instruments is distinct. The SYNAPT G2 uses a stack ring ion guide, while the LTQ Orbitrap uses a linear quadrupole. These differences produce different ion intensity scales. Together, these differences may explain the difference in the quantification results obtained by the SYNAPT G2 and LTQ Orbitrap platforms.

References

Bakalarski CE, Elias JE, Villen J, Haas W, Gerber SA, Everley PA, Gygi SP. The impact of peptide abundance and dynamic range on stable-isotope-based quantitative proteomic analyses. J Proteome Res. 2008;7:4756–65.

Bateman RH, Carruthers R, Hoyes JB, Jones C, Langridge JI, Millar A, Vissers JP. A novel precursor ion discovery method on a hybrid quadrupole orthogonal acceleration time-of-flight (Q-TOF) mass spectrometer for studying protein phosphorylation. J Am Soc Mass Spectrom. 2002;13:792–803.

Choudhary C, Mann M. Decoding signalling networks by mass spectrometry-based proteomics. Nat Rev Mol Cell Biol. 2010;11:427–39.

Cleveland WS. Robust locally weighted regression and smoothing scatterplots. J Am Statist Assoc. 1979;74:829–36.

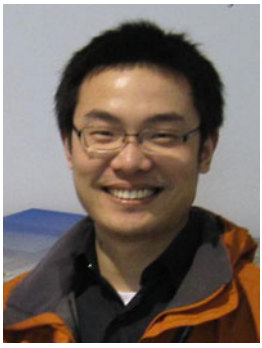
Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol. 2008;26:1367–72.

Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M. Andromeda: a peptide search engine integrated into the MaxQuant environment. J Proteome Res. 2011;10:1794–805.

Ding SJ, Wang Y, Jacobs JM, Qian WJ, Yang F, Tolmachev AV, Du X, Wang W, Moore RJ, Monroe ME, Purvine SO, Waters K, Heibeck TH, Adkins JN, Camp 2nd DG, Klemke RL, Smith RD. Quantitative phosphoproteome analysis of lysophosphatidic acid induced chemotaxis applying

dual-step (18)O labeling coupled with immobilized metal-ion affinity chromatography. J Proteome Res. 2008;7:4215–24.
Elias JE, Gygi SP. Target-decoy search strategy for mass spectrometry-based proteomics. Methods Mol Biol. 2010;604:55–71.
Eng JK, Mccormack AL, Yates JR. An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database. J Am Soc Mass Spectr. 1994;5:976–89.
Fang R, Elias DA, Monroe ME, Shen Y, McIntosh M, Wang P, Goddard CD, Callister SJ, Moore RJ, Gorby YA, Adkins JN, Fredrickson JK, Lipton MS, Smith RD. Differential label-free quantitative proteomic analysis of *Shewanella oneidensis* cultured under aerobic and suboxic conditions by accurate mass and time tag approach. Mol Cell Proteomics. 2006;5:714–25.
Finney GL, Blackler AR, Hoopmann MR, Canterbury JD, Wu CC, MacCoss MJ. Label-free comparative analysis of proteomics mixtures using chromatographic alignment of high-resolution muLC-MS data. Anal Chem. 2008;80:961–71.
Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH. Open mass spectrometry search algorithm. J Proteome Res. 2004;3:958–64.
Geiger T, Cox J, Ostasiewicz P, Wisniewski JR, Mann M. Super-SILAC mix for quantitative proteomics of human tumor tissue. Nat Methods. 2010a;7:383–5.
Geiger T, Cox J, Mann M. Proteomics on an Orbitrap benchtop mass spectrometer using all-ion fragmentation. Mol Cell Proteomics. 2010b;9:2252–61.
Gstaiger M, Aebersold R. Applying mass spectrometry-based proteomics to genetics, genomics and network biology. Nat Rev Genet. 2009;10:617–27.
Han DK, Eng J, Zhou H, Aebersold R. Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. Nat Biotechnol. 2001;19:946–51.
Hu Q, Noll RJ, Li H, Makarov A, Hardman M, Graham Cooks R. The Orbitrap: a new mass spectrometer. J Mass Spectrom. 2005;40:430–43.
Huang X, Tolmachev AV, Shen Y, Liu M, Huang L, Zhang Z, Anderson GA, Smith RD, Chan WC, Hinrichs SH, Fu K, Ding SJ. UNiquant, a program for quantitative proteomics analysis using stable isotope labeling. J Proteome Res. 2011a;10:1228–37.
Huang X, Liu M, Nold MJ, Tian C, Fu K, Zheng J, Geromanos SJ, Ding SJ. Software for quantitative proteomic analysis using stable isotope labeling and data independent acquisition. Anal Chem. 2011b;83:6971–9.
Koomen JM, Haura EB, Bepler G, Sutphen R, Remily-Wood ER, Benson K, Hussein M, Hazlehurst LA, Yeatman TJ, Hildreth LT, Sellers TA, Jacobsen PB, Fenstermacher DA, Dalton WS. Proteomic contributions to personalized cancer care. Mol Cell Proteomics. 2008;7:1780–94.
Li XJ, Zhang H, Ranish JA, Aebersold R. Automated statistical analysis of protein abundance ratios from data generated by stable-isotope dilution and tandem mass spectrometry. Anal Chem. 2003;75:6648–57.
Li GZ, Vissers JP, Silva JC, Golick D, Gorenstein MV, Geromanos SJ. Database searching and accounting of multiplexed precursor and product ion spectra from the data independent analysis of simple and complex peptide mixtures. Proteomics. 2009;9:1696–719.
Liao Z, Wan Y, Thomas SN, Yang AJ. IsoQuant: a software tool for stable isotope labeling by amino acids in cell culture-based mass spectrometry quantitation. Anal Chem. 2012;84: 4535–43.
Mann M. Functional and quantitative proteomics using SILAC. Nat Rev Mol Cell Biol. 2006;7:952–8.
Mann M, Kelleher NL. Precision proteomics: the case for high resolution and high mass accuracy. Proc Natl Acad Sci U S A. 2008;105:18132–8.
Mo F, Mo Q, Chen Y, Goodlett DR, Hood L, Omenn GS, Li S, Lin B. WaveletQuant, an improved quantification software based on wavelet signal threshold de-noising for labeled quantitative proteomic analysis. BMC Bioinformatics. 2010;11:219.

Olsen JV, Mann M. Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation. Proc Natl Acad Sci U S A. 2004;101:13417–22.
Ong SE, Mann M. Mass spectrometry-based proteomics turns quantitative. Nat Chem Biol. 2005;1:252–62.
Pan C, Kora G, McDonald WH, Tabb DL, VerBerkmoes NC, Hurst GB, Pelletier DA, Samatova NF, Hettich RL. ProRata: a quantitative proteomics program for accurate protein abundance ratio estimation with confidence interval evaluation. Anal Chem. 2006;78:7121–31.
Park SK, Venable JD, Xu T, Yates 3rd JR. A quantitative analysis software tool for mass spectrometry-based proteomics. Nat Methods. 2008;5:319–22.
Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis. 1999;20:3551–67.
Qian WJ, Petritis BO, Kaushal A, Finnerty CC, Jeschke MG, Monroe ME, Moore RJ, Schepmoes AA, Xiao W, Moldawer LL, Davis RW, Tompkins RG, Herndon DN, Camp DG, Smith RD. Plasma proteome response to severe burn injury revealed by (18)O-labeled "universal" reference-based quantitative proteomics. J Proteome Res. 2010;9:4779–89.
Ramos AA, Yang H, Rosen LE, Yao X. Tandem parallel fragmentation of peptides for mass spectrometry. Anal Chem. 2006;78:6391–7.
Silva JC, Denny R, Dorschel CA, Gorenstein M, Kass IJ, Li GZ, McKenna T, Nold MJ, Richardson K, Young P, Geromanos S. Quantitative proteomic analysis by accurate mass retention time pairs. Anal Chem. 2005;77:2187–200.
Silva JC, Denny R, Dorschel C, Gorenstein MV, Li GZ, Richardson K, Wall D, Geromanos SJ. Simultaneous qualitative and quantitative analysis of the *Escherichia coli* proteome: a sweet tale. Mol Cell Proteomics. 2006a;5:589–607.
Silva JC, Gorenstein MV, Li GZ, Vissers JP, Geromanos SJ. Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. Mol Cell Proteomics. 2006b;5:144–56.
Venable JD, Dong MQ, Wohlschlegel J, Dillin A, Yates JR. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. Nat Methods. 2004;1:39–45.
Vissers JP, Pons S, Hulin A, Tissier R, Berdeaux A, Connolly JB, Langridge JI, Geromanos SJ, Ghaleh B. The use of proteome similarity for the qualitative and quantitative profiling of rep-erfused myocardium. J Chromatogr B Analyt Technol Biomed Life Sci. 2009;877:1317–26.
Williams JD, Flanagan M, Lopez L, Fischer S, Miller LA. Using accurate mass electrospray ionization-time-of-flight mass spectrometry with in-source collision-induced dissociation to sequence peptide mixtures. J Chromatogr A. 2003;1020:11–26.
Wilm M. Quantitative proteomics in biological research. Proteomics. 2009;9:4590–605.



Xin Huang, Ph.D., Nebraska, USA Xin Huang is the graduate student at the Department of Pathology and Microbiology at University of Nebraska Medical Center. He earned a BE and MS from Zhejiang University, and a PhD from University of Nebraska Medical Center.

Shi-Jian Ding, Ph.D., Director, Assistant Professor, USA
Shi-Jian Ding is the technical director of the Mass Spectrometry and Proteomics Core Facility at University of Nebraska Medical Center and an assistant professor at the Department of Pathology and Microbiology at University of Nebraska Medical Center. Prior to moving to Nebraska, Dr. Ding was a postdoctoral fellow at Pacific Northwest National Laboratory (2004–2007). His research interests center on development and application of mass spectrometry-based proteomic approaches for answering biological questions.

He has over 20 publications including *Proceedings of the National Academy of Sciences*, *Molecular and Cellular Proteomics*, *Analytical Chemistry*, *Journal of Proteome Research*, and *Proteomics*. He earned a BS from Lanzhou University and a Ph.D. from Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences.