Anal. Chem. **2006,** *78,* 7121−7131

ProRata: A Quantitative Proteomics Program for Accurate Protein Abundance Ratio Estimation with Confidence Interval Evaluation

Chongle Pan,[†,‡,§,||] Guruprasad Kora,[‡,||] W. Hayes McDonald,[†] David L. Tabb,[⊥,#] Nathan C. VerBerkmoes,[†] Gregory B. Hurst,[†] Dale A. Pelletier,[⊥] Nagiza F. Samatova,[*,‡,||] and Robert L. Hettich[*,†]

Chemical Sciences Division, Computational Biology Institute, Computer Science and Mathematics Division, Life Sciences Division, Oak Ridge National Laboratory, Genome Science and Technology Graduate School, Oak Ridge National Laboratory—University of Tennessee, Oak Ridge, Tennessee 37830

A profile likelihood algorithm is proposed for quantitative shotgun proteomics to infer the abundance ratios of proteins from the abundance ratios of isotopically labeled peptides derived from proteolysis. Previously, we have shown that the estimation variability and bias of peptide abundance ratios can be predicted from their profile signal-to-noise ratios. Given multiple quantified peptides for a protein, the profile likelihood algorithm probabilistically weighs the peptide abundance ratios by their inferred estimation variability, accounts for their expected estimation bias, and suppresses contribution from outliers. This algorithm yields maximum likelihood point estimation and profile likelihood confidence interval estimation of protein abundance ratios. This point estimator is more accurate than an estimator based on the average of peptide abundance ratios. The confidence interval estimation provides an "error bar" for each protein abundance ratio that reflects its estimation precision and statistical uncertainty. The accuracy of the point estimation and the precision and confidence level of the interval estimation were benchmarked with standard mixtures of isotopically labeled proteomes. The profile likelihood algorithm was integrated into a quantitative proteomics program, called ProRata, freely available at www.MSProRata.org.

Organisms often respond to environmental or physiological stimuli by adjusting the type and abundance of proteins in their cells. Measurement of the relative abundances of proteins in treatment cells subjected to stimuli, compared to that in reference cells, provides valuable insights about protein function and regulation. Quantitative shotgun proteomics has recently emerged as a high-throughput technique for measuring the relative abundances of thousands of proteins between two cellular conditions.[1] The reference and treatment proteomes are labeled with different stable isotope tags[2−5] and then mixed together in equivalent amounts. In such a proteome mixture, each protein has two mass-different isotopic variants: the light isotopologue and the heavy isotopologue,[6] or a protein isotopologue pair. Finally, the proteome mixture is digested and then analyzed with liquid chromatography−tandem mass spectrometry (LC−MS/MS).[7] The proteolysis turns each protein isotopologue pair into multiple peptide isotopologue pairs. Each of the peptide isotopologue pairs is expected to have the same abundance ratio as the protein isotopologue pair. Although multiplication of isotopologue pairs in the mixture by proteolysis increases the complexity of the sample for LC−MS/MS analysis, it provides multiple indirect measurements of a protein's abundance ratio, derived from the abundance ratios of its peptide isotopologue pairs.

To evaluate protein abundance ratios from quantitative proteomics measurements, two types of statistical estimation should be employed: point estimation and interval estimation. The point estimation gives an abundance ratio for every quantified protein, which "best" approximates the true abundance ratio. Unfortunately, the point estimation provides no information about protein quantification precision, which can significantly vary across different proteins. Generally, a protein should have better quantification precision if it has more proteolytic peptides quantified from mass spectral data of higher signal-to-noise ratio. It is misleading in quantitative proteomics to treat all proteins' abundance ratios identically, regardless of their estimation precision.

The interval estimation complements the point estimation by providing confidence intervals for protein abundance ratios. If 90%

[*] To whom correspondence should be addressed. For questions on computational methods, (phone) (865) 241-4351, (e-mail) samatovan@ornl.gov. For questions on experimental methods, (phone) (865) 574-4968, (e-mail) hettichrl@ornl.gov.

[†] Chemical Sciences Division.

[‡] Computational Biology Institute.

[§] Genome Science and Technology Graduate School.

[||] Computer Science and Mathematics Division.

[⊥] Life Sciences Division.

[#] Current address: Department of Biomedical Informatics, Vanderbilt University, Nashville, TN 37232.

[1] Ong, S. E.; Mann, M. *Nat. Chem. Biol.* **2005**, *1* (5), 252−262.

[2] Oda, Y.; Huang, K.; Cross, F. R.; Cowburn, D.; Chait, B. T. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96* (12), 6591−6596.

[3] Ong, S. E.; Blagoev, B.; Kratchmarova, I.; Kristensen, D. B.; Steen, H.; Pandey, A.; Mann, M. *Mol. Cell. Proteomics* **2002**, *1* (5), 376−386.

[4] Gygi, S. P.; Rist, B.; Gerber, S. A.; Turecek, F.; Gelb, M. H.; Aebersold, R. *Nat. Biotechnol.* **1999**, *17* (10), 994−999.

[5] Yao, X. D.; Freas, A.; Ramirez, J.; Demirev, P. A.; Fenselau, C. *Anal. Chem.* **2001**, *73* (13), 2836−2842.

[6] Muller, P. *Pure Appl. Chem.* **1994**, *66* (5), 1132.

[7] Link, A. J.; Eng, J.; Schieltz, D. M.; Carmack, E.; Mize, G. J.; Morris, D. R.; Garvik, B. M.; Yates, J. R. *Nat. Biotechnol.* **1999**, *17* (7), 676−682.

10.1021/ac060654b CCC: $33.50 © 2006 American Chemical Society
Published on Web 09/13/2006

of quantified proteins have confidence intervals that contain their true abundance ratios, then confidence intervals are estimated at a 90% confidence level. The confidence level for the interval estimation in quantitative proteomics is analogous to the true positive rate for protein identification in qualitative proteomics. More importantly, at a given confidence level, the confidence interval intuitively reflects the quantification precision for each protein as an "error bar" of the abundance ratio estimate.

In quantitative shotgun proteomics, each protein's abundance ratio is estimated by "combining" multiple peptide abundance ratios measured with LC−MS/MS. Several computer programs have been developed for the point estimation of protein abundance ratios, including XPRESS,[8] ASAPratio,[9] MSQuant,[10] and RelEx.[11] The first three programs calculate peptide abundance ratios from ratios of peak areas in selected ion chromatograms. RelEx, on the other hand, estimates peptide abundance ratios using a linear correlation algorithm, which reduces the peptide abundance ratio estimation error. The average of peptide abundance ratios is then used by RelEx, XPRESS, and MSQuant to estimate protein abundance ratios. To improve the accuracy of protein abundance ratio estimation, ASAPratio weighs the peptide abundance ratios using their peak areas and uses the weighted average to estimate the protein abundance ratios. In these programs, the standard deviation of the peptide abundance ratios is used as a measure of the variability of the protein abundance ratio estimation. However, without assuming the normality of the peptide abundance ratio distribution, the standard deviation is not directly related to the confidence interval of the protein abundance ratio.

In our previous study, we scored every peptide's abundance ratio with a profile signal-to-noise ratio for the peptide's mass spectral data.[12] It was observed that the estimation variability and bias of peptide abundance ratios in $\log_2$ scale (peptide log-ratios) were linearly correlated with profile signal-to-noise ratios in $\log_2$ scale (log-profile-S/Ns). This was illustrated with a standard mixture of isotopically labeled proteomes, in which all peptides were expected to have an abundance ratio of 1:5 or a log-ratio of −2.3. The two-dimensional heat map histogram of peptide log-ratio versus log-profile-S/N shows a comet-like distribution from this standard mixture (Figure 1). In the high log-profile-S/N region, the peptide log-ratio distributions are tight with a center at the expected log-ratio. As the log-profile-S/N level decreases, the horizontal spread of the log-ratio distributions increases, indicating the elevation of variability in peptide log-ratio estimation. At the same time, the distributions also deviate more from the true log-ratio, indicating the increase of bias in peptide log-ratio estimation. The change of variability and bias of the log-ratio distributions with log-profile-S/N makes it unsuitable to treat peptides with different log-profile-S/N identically and to assume normality for the aggregated peptide log-ratio distribution.

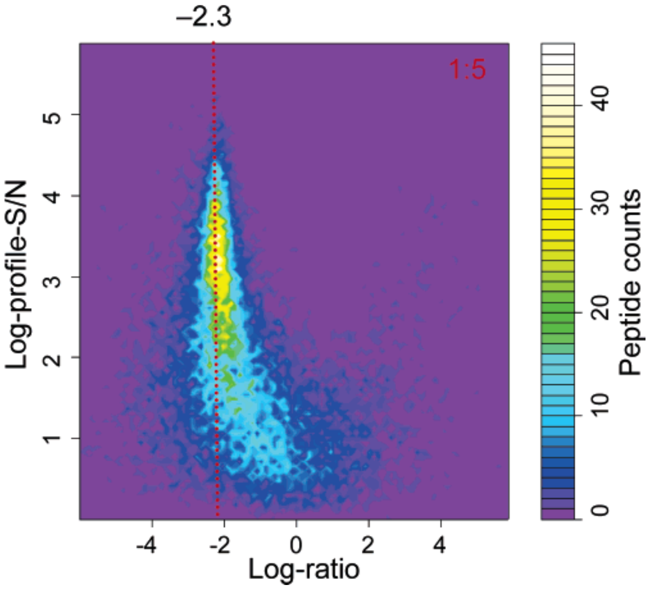


Figure 1. Two-dimensional heat map histogram of peptide abundance ratio in $\log_2$ scale (peptide log-ratio) versus profile S/N ratio in $\log_2$ scale (log-profile-S/N). The color scale (shown on the right) represents the number of peptides at a given log-ratio and log-profile-S/N location. The expected log-ratio for all peptides in this 1:5 standard mixture is −2.3, indicated with the red dotted vertical line. As the log-profile S/N level lowers, the horizontal distribution of the peptide log-ratios spreads wider and deviates more from the line of expected log-ratio.

In this study, we propose a profile likelihood algorithm that yields both maximum likelihood point estimation[13] and profile likelihood confidence interval estimation[14] of protein abundance ratios. This likelihood-based approach allows us take into account the changing estimation variability and bias of peptide abundance ratios in the process of protein abundance ratio estimation. This improves the accuracy of the point estimation and the precision and confidence level of the interval estimation, as benchmarked with standard mixtures of isotopically labeled proteomes.

The profile likelihood algorithm is part of a computer program called ProRata, which automates the entire data analysis process for quantitative shotgun proteomics. ProRata is applicable for a variety of stable-isotope labeling techniques, including ^{15}N or ^{13}C metabolic labeling,[2] SILAC,[3] ICAT,[4] and $H_2{}^{18}O$ proteolysis.[5] Figure 2 shows the data analysis flowchart of ProRata. ProRata extracts selected ion chromatograms for peptide isotopologue pairs and detects their chromatographic peaks with a parallel paired covariance algorithm.[12] Principal component analysis is then used to calculate peptide abundance ratios and profile signal-to-noise ratios. Finally, protein abundance ratios and their confidence intervals are estimated with the profile likelihood algorithm.

MATERIALS AND METHODS

Chemicals and Reagents. HPLC-grade water and acetonitrile were obtained from Burdick & Jackson (Muskegon, MI), and the 98% formic acid was from EM Science (Darmstadt, Germany). All other chemicals were purchased from Sigma-Aldrich (St. Louis, MO) unless noted otherwise.

[8] Han, D. K.; Eng, J.; Zhou, H.; Aebersold, R. *Nat. Biotechnol.* **2001**, *19* (10), 946−951.

[9] Li, X. J.; Zhang, H.; Ranish, J. A.; Aebersold, R. *Anal. Chem.* **2003**, *75* (23), 6648−6657.

[10] Schulze, W. X.; Mann, M. *J. Biol. Chem.* **2004**, *279* (11), 10756−10764.

[11] MacCoss, M. J.; Wu, C. C.; Liu, H.; Sadygov, R.; Yates, J. R., 3rd. *Anal. Chem.* **2003**, *75* (24), 6912−6921.

[12] Pan, C.; Kora, G.; Tabb, D. L.; Pelletier, D. A.; McDonald, W. H.; Hurst, G. B.; Hettich, R. L.; Samatova, N. F. *Anal. Chem.* **2006**, *78*, 7110−7120.

[13] Eliason, S. R. *Maximum Likelihood Estimation: Logic and Practice*; Sage Publications Inc.: Newbury Park, CA, 1993.

[14] Venzon, D. J. *Appl. Statistics* **1988**, *37* (1), 87−94.

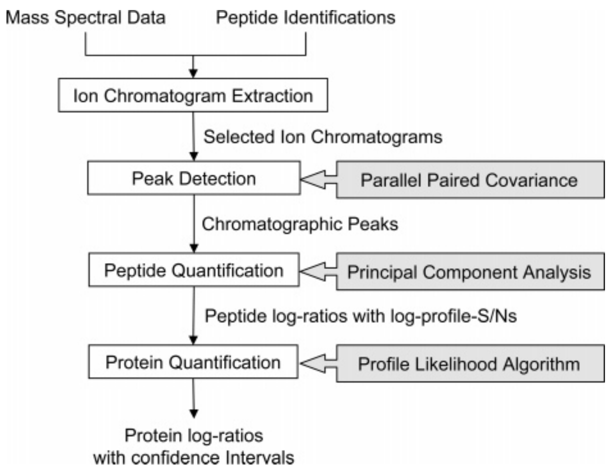

Mass Spectral Data
Peptide Identifications
Ion Chromatogram Extraction
Selected Ion Chromatograms
Peak Detection
Parallel Paired Covariance
Chromatographic Peaks
Peptide Quantification
Principal Component Analysis
Peptide log-ratios with log-profile-S/Ns
Protein Quantification
Profile Likelihood Algorithm
Protein log-ratios
with confidence Intervals

Figure 2. Data processing flowchart of ProRata. ProRata consists of four modules, shown as blocks. The algorithm used in each module is specified on the right. The data flow from one module to the next is shown with the solid arrows, starting from the input data: mass spectral data and peptide identification results.

Standard Isotopically Labeled Proteome Mixture Preparation. *Rhodopseudomonas palustris* CGA0010 strain was grown anaerobically in light at 30 °C to mid-log phase. The defined minimal growth medium supplies $(NH_4)_2SO_4$ as the only nitrogen source for bacterial growth. The unlabeled culture was grown with $(^{14}NH_4)_2SO_4$. The ^{15}N-labeled culture was grown identically with $(^{15}NH_4)_2SO_4$ (>98 atom % excess, from Sigma-Aldrich). The ^{14}N proteome and the ^{15}N proteome were prepared from the unlabeled culture and the ^{15}N-labeled culture, respectively, as described.[12,15] The total protein concentration of each proteome was quantified with Lowry's analysis. Six standard mixtures were prepared by mixing the ^{14}N proteome and the ^{15}N proteome at the ratios of 10:1, 5:1, 1:1, 1:5, and 1:10 by total protein mass. An aliquot of the ^{14}N proteome was also retained for shotgun proteomics measurement.

Shotgun Proteomics Measurements. The proteome samples were processed with the described procedure.[12] Briefly, after disulfide bond reduction and protein denaturation with 10 mM DTT and 6 M guanidine, the proteome samples were digested with sequencing grade trypsin (Promega, Madison, WI). The samples were then treated with 20 mM DTT for 1 h at 60 °C as a final reduction step. The samples were immediately desalted with Sep-Pak Plus C18 solid-phase extraction (Waters, Milford, MA) and solvent exchanged into 0.1% formic acid in water by centrifugal evaporation. The protein digests were examined with the 12-step split-phase MudPIT technique, as described previously.[16,17] Briefly, the samples were first separated by 12-step strong cation ion exchange liquid chromatography and then by 2-h continuous gradient reversed-phase liquid chromatography. Eluted peptides were electrosprayed at 2-kV distal electrospray voltage into an LTQ mass spectrometer (Thermo Finnigan, San Jose, CA). Tandem mass spectrometry analysis was performed with each full scan (400−1700 m/z) followed by five data-dependent MS/MS scans at 35% normalized collision energy. Dynamic exclusion was enabled. All scans were averaged from two microscans.

Peptide and Protein Identification. All MS/MS scans were searched in two iterations against a FASTA database containing all annotated *R. palustris* proteins[18] using the SEQUEST program.[19] In the first iteration, the unmodified amino acids were used, and in the second iteration the ^{15}N-labeled amino acids were used. The peptide identifications from the two iterations were merged. The DTASelect program[20] was used to filter the peptide identifications and to assemble the peptides into proteins using the following parameters: retain the duplicate MS/MS scans for each peptide sequence (DTASelect option, −t 0), consider fully tryptic peptides only, filter with a delCN of at least 0.08, and cross-correlation scores (Xcorrs) of at least 1.8 (+1), 2.5 (+2), and 3.5 (+3).

Ion Chromatogram Extraction. For each identified peptide, two selected ion chromatograms were extracted for the two peptide isotopologues. The m/z window for the light isotopologue was calculated from the natural isotopic envelope of the peptide. The m/z window for the heavy isotopologue was calculated by using 98%-enriched ^{15}N for all nitrogen atoms. The retention time window of the selected ion chromatograms was defined as from 2 min before the MS/MS scans to 2 min after the MS/MS scans.

Peptide Abundance Ratio Estimation. The chromatographic peaks of the peptide isotopologue pairs were detected with a parallel paired covariance algorithm. The abundance ratio and the profile S/N of a peptide were calculated by analyzing its peak profile with principal component analysis as described previously.[12] Briefly, the peak profile was constructed as a scatterplot with ion intensities of the two isotopologues as its coordinates.[21] Principal component analysis of the peak profile generated two principal components and their associated eigenvalues. The peptide abundance ratio was estimated as the slope of the first principal component. The profile S/N was calculated as the square root of the ratio between the first eigenvalue and the second eigenvalue. Peptides with a profile S/N below 2.0 were removed, due to the large estimation error of their abundance ratios. Peptides shared among multiple proteins were also discarded, because the abundance ratio of a shared peptide will be a weighted average of the abundance ratios of multiple proteins and thus cannot be used for any of the individual proteins.

Protein Abundance Ratio Estimation. Quantified peptides were assembled into proteins. Proteins with two or more quantified peptides were selected for abundance ratio estimation.

[15] VerBerkmoes, N. C.; Shah, M. B.; Lankford, P. K.; Pelletier, D. A.; Strader, M. B.; Tabb, D. L.; McDonald, W. H.; Barton, J. W.; Hurst, G. B.; Hauser, L.; Davison, B. H.; Beatty, J. T.; Harwood, C. S.; Tabita, F. R.; Hettich, R. L.; Larimer, F. W. *J. Proteome Res.* **2006**, *5* (2), 287−298.

[16] McDonald, W. H.; Ohi, R.; Miyamoto, D. T.; Mitchison, T. J.; Yates, J. R. *Int. J. Mass Spectrom.* **2002**, *219* (1), 245−251.

[17] MacCoss, M. J.; McDonald, W. H.; Saraf, A.; Sadygov, R.; Clark, J. M.; Tasto, J. J.; Gould, K. L.; Wolters, D.; Washburn, M.; Weiss, A.; Clark, J. I.; Yates, J. R., 3rd. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99* (12), 7900−7905.

[18] Larimer, F. W.; Chain, P.; Hauser, L.; Lamerdin, J.; Malfatti, S.; Do, L.; Land, M. L.; Pelletier, D. A.; Beatty, J. T.; Lang, A. S.; Tabita, F. R.; Gibson, J. L.; Hanson, T. E.; Bobst, C.; Torres, J. L.; Peres, C.; Harrison, F. H.; Gibson, J.; Harwood, C. S. *Nat. Biotechnol.* **2004**, *22* (1), 55−61.

[19] Eng, J. K.; Mccormack, A. L.; Yates, J. R., III. *J. Am. Soc. Mass Spectrom.* **1994**, *5* (11), 976−989.

[20] Tabb, D. L.; McDonald. W. H.; Yates. J. R. *J. Proteome Res.* **2002**, *1* (1), 21−26.

[21] Lawson, A. M.; Kim. C. K.; Richmond. W.; Samson. D. M.; Setchell. K. D. R.; Thomas. A. C. S. Isotope dilution mass spectrometry as a basis for accuracy in clinical chemistry. In *Current Developments in the Clinical Applications of HPLC, GC and MS*; Academic Press: London, Middlesex, UK, 1980.

For the point estimation and confidence interval estimation, the profile likelihood algorithm solves a likelihood function of protein log-ratio with a numerical method, in the absence of an analytical method. The profile likelihood algorithm has three steps:

Log-ratio enumeration step: Enumerate all protein log-ratios to be considered through discretization of a continuous log-ratio interval.

Likelihood calculation step: Calculate the likelihood for each considered log-ratio to be the true log-ratio of a protein, given the abundance ratios and profile S/Ns of multiple peptides from this protein.

Point and interval estimation step: Select a log-ratio with the maximum likelihood as the maximum likelihood estimate; select two log-ratios on a likelihood threshold as the lower and upper limits of profile likelihood confidence interval.

Below we describe the three steps in detail and the rationale for the procedures.

Log-Ratio Enumeration Step. The continuous interval of protein log-ratio, [−7.0, 7.0], is discretized at the precision of 0.1 to create a discrete set of protein log-ratios, i.e. −7.0, −6.9, −6.8, ... 6.8, 6.9, and 7.0. This set, denoted as *G*, enumerates all protein log-ratios that the profile likelihood algorithm will consider for every protein. The minimum and maximum protein log-ratios and the discretization precision are configurable in ProRata. The maximum protein log-ratio of 7 and the minimum protein log-ratio of −7 correspond to 128-fold upregulation and 128-fold downregulation in protein abundance, respectively. These maximum and minimum log-ratios sufficiently encompass the practical dynamic range of our instruments for quantification. The discretization precision of 0.1 is also the protein quantification precision that can be realistically achieved in quantitative proteomics measurements. The discretization of protein log-ratio solution space allows for an efficient and correct numerical solution of the likelihood function of protein log-ratio by brute force enumeration.

Likelihood Calculation Step. The likelihood for each protein log-ratio in *G* to be the true log-ratio of a protein is calculated. Assume that the protein has *n* peptides with log-ratios of R_1, R_2 ... R_n and log-profile-S/Ns of V_1, V_2 ... V_n. Let *H* be an arbitrary log-ratio from *G*. Then the likelihood for *H* to be the true log-ratio of a protein given the log-ratios and the log-profile-S/Ns of its peptides equals the probability of observing these peptide log-ratios given their log-profile-S/Ns and the protein log-ratio of *H*:

$$L(H|R_1,R_2 \cdots R_n,V_1,V_2 \cdots V_n) = P(R_1,R_2 \cdots R_n|V_1,V_2 \cdots V_n,H) \quad (1)$$

where $L(\,)$ is the likelihood function and $P(\,)$ is the probability function. Assume that the protein's peptides are measured independently. Then the probability of observing these *n* peptides together is the product of individual probabilities of observing each of these peptides independently:

$$P(R_1,R_2 \cdots R_n|V_1,V_2 \cdots V_n,H) = P(R_1|V_1,H) \cdots P(R_2|V_2,H) \cdots P(R_n|V_n,H) = \prod_{i=1}^{n} \mathrm{P}(R_i|V_i,H) \quad (2)$$

Theoretically, the log-ratio of a peptide is expected to be equal to the log-ratio of its protein. However, due to the changing variability and bias of peptide log-ratio estimation, the probability distribution of peptide log-ratio is modeled with a mixture model of normal distribution and uniform distribution:

$$P(R_\mathrm{i}|V_\mathrm{i},H) = 85\%\cdot P_\mathrm{normal}(R_\mathrm{i}|\mu_\mathrm{i},\sigma_\mathrm{i}) + 15\%\cdot P_\mathrm{uniform} \quad (3)$$

where $P_\mathrm{normal}(\,)$ is a normal probability function and $P_\mathrm{uniform}(\,)$ is a uniform probability function. The mixture model is employed because the uniform distribution with a weight of 15% models ~15% of outlier peptides that are not well captured by the normal distribution.

The absolute value of the mean ($|\mu_\mathrm{i}|$) and the standard deviation (σ_i) of the normal distribution in the mixture model are approximated with two linear functions of the log-profile-S/N (V_i) and the protein log-ratio (H):

$$|\mu_\mathrm{i}| = \begin{cases} 1.2\cdot V_i, & 1.2\cdot V_i < |H|, \\ |H| \quad , & 1.2\cdot V_i \geq |H|, \end{cases} \quad (4)$$

$$\sigma_\mathrm{i} = 1.2 - 0.2\cdot V_i \quad (5)$$

The sign of the mean (μ_i) is the same as that of the protein log-ratio (H). The two linear functions and their coefficients were estimated from experimental data as described previously.[12] Briefly, two-dimensional heat map histograms were constructed for six different standard mixtures of isotopically labeled proteomes as shown in Figure 1. The means and standard deviations of the peptide log-ratio distributions at different log-profile-S/N levels were calculated, from which the two linear functions were then derived. The absolute mean linear function (eq 4) captures that the bias of peptide log-ratio estimation decreases to zero with an increase of log-profile-S/N. The negative slope of the standard deviation linear function (eq 5) models the decrease of peptide log-ratio estimation variability with an increase of log-profile-S/N.

In summary, the likelihood function of protein log-ratio is constructed as

$$L(H|R_1,R_2 \cdots R_n,V_1,V_2 \cdots V_n) = \prod_{i=1}^{n} P(R_\mathrm{i}|V_\mathrm{i},H) = \prod_{i=1}^{n} (85\%\cdot P_\mathrm{normal}(R_\mathrm{i}|\mu_\mathrm{i},\sigma_\mathrm{i}) + 15\%\cdot P_\mathrm{uniform}) \quad (6)$$

The likelihood *H* is calculated for each protein log-ratio in the set *G* and transformed to the natural logarithm scale, denoted as *ln*-likelihood. The calculation result of the likelihood function can be represented graphically with a profile likelihood curve. Let the *x*-axis and the *y*-axis be log-ratio and *ln*-likelihood, respectively, and plot all considered protein log-ratios with their *ln*-likelihood as points. The profile likelihood curve is constructed by connecting the points adjacent along the log-ratio axis (the blue curves in Figure 3). For manual data analysis, the data points representing the peptides of a protein (the red points in Figure 3) are overlaid with the protein's profile likelihood curve.

Point and Interval Estimation Step. Both maximum likelihood estimate and profile likelihood confidence interval of a protein's log-ratio are estimated from the protein's profile likeli-

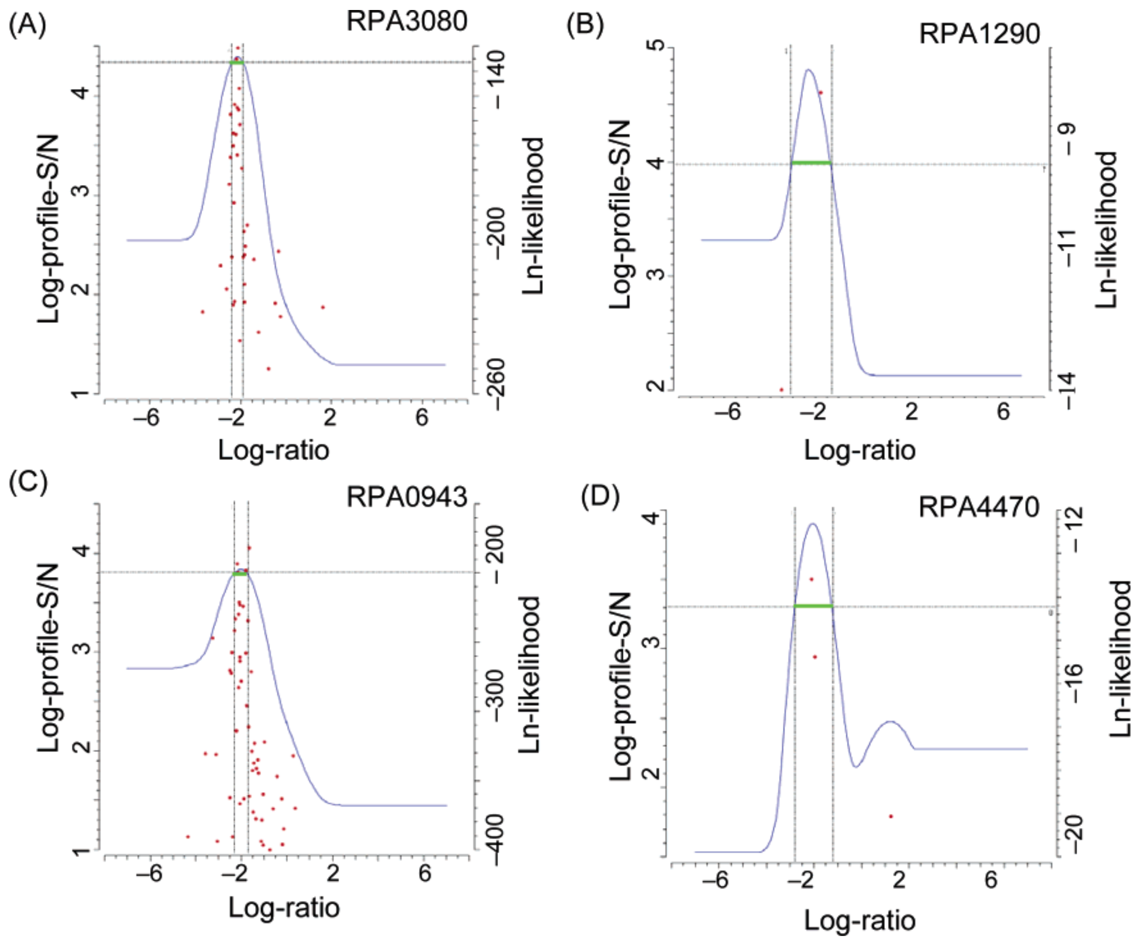


Figure 3. Estimation of protein log-ratios with profile likelihood curves. All four proteins (locus shown in the upper right corner) are expected to have a log-ratio of −2.3. Profile likelihood curves (the blue curves) plot the ln likelihood (*y*-axis on the right) for the log-ratios (*x*-axis) of proteins. Note the different ln-likelihood ranges for different proteins. Peptide data points (the red dots) represent the log-ratio (*x*-axis) and the log-profile-S/N (*y*-axis on the left) of the quantified peptides. The confidence intervals are shown as the green bars.

hood curve. The maximum likelihood estimate of protein log-ratio is the log-ratio with the maximum likelihood. The confidence interval is calculated by assuming a χ^2 distribution for the likelihood ratio test as described in the standard methodology for profile likelihood confidence interval estimation.[14] Let $L_{\max}$ be the maximum likelihood. The confidence interval with a nominal confidence level of $(1 - \alpha) \times 100\%$ includes all the log-ratios that have an *ln*-likelihood exceeding the threshold of $\ln(L_{\max}) - 0.5\cdot\chi^2_{1,\alpha}$. The *ln*-likelihood threshold is $\ln(L_{\max}) - 1.96$ for the confidence interval of 95% nominal confidence level ($0.5\cdot\chi^2_{1,\,0.05} = 1.96$). The lower and upper limits of the confidence interval are the minimum and maximum log-ratios with *ln*-likelihood exceeding the threshold, respectively. We discarded proteins with confidence intervals that are wider than 7.

Protein abundance ratios were also estimated with the RelEx program[11] for comparison purposes. The DTASelect result files and the Xcalibur data files were input to the RelEx program. Data smoothing, ratio correction, and chromatogram filtering were enabled with the default settings. The protein filter of minimum peptide number of two was also applied. RelEx was executed in two iterations, one using the light isotopologue for peak detection and the other using the heavy isotopologue. The results of the two iterations are combined. The abundance ratios of the proteins quantified in both iterations are assigned as the average of the abundance ratios estimated in the two iterations.

Software Development. ProRata was written in C++ and compiled with the MinGW g++ compiler. The graphical user interface was implemented using Qt library. ProRata used the RAMP (random access minimal parser) library to access mzXML files.[22] The histograms were constructed with R scripts.[23]

RESULTS AND DISCUSSION

The features of the profile likelihood algorithm were evaluated initially with individual proteins. Then the aggregate performance metrics were determined, including the accuracy of the point estimation as well as the confidence level and median width of the confidence interval estimation. Finally, ProRata was equipped with a graphical user interface to enable manual interrogation of the quantification result for any given protein. Note the following abbreviations used in this article: (a) log-ratio for the abundance ratio in $\log_2$ scale, (b) log-profile-S/N for the profile S/N in $\log_2$ scale, and (c) *ln*-likelihood for the natural logarithm of the likelihood for a log-ratio to be true for a protein. The $\log_2$ transformation for the abundance ratio is to treat up- and downregulation of protein abundance symmetrically and to replace a multiplication operation on the abundance ratios with the addition operation on the log-ratios.

[22] http://sashimi.sourceforge.net/software_glossolalia.html.

[23] R Development Core Team. *R: A language and environment for statistical computing*; R Foundation for Statistical Computing, Vienna, Austria, 2006.

Table 1. Summary of Protein Quantification Results from the Standard Mixtures of Isotopically Labeled Proteomes

standard mixture		protein count		log-ratio point estimation		confidence interval estimation		hypothesis testing	
^{14}N:^{15}N	log ratio	identified	quantified	median	AAD[a]	median width	confidence level (%)	significance (%)	power (%)
1:1a	0.0	1392	1117	−0.2	0.318	1.4	93	8	
1:1b	0.0	1348	1071	−0.2	0.361	1.4	92	10	
5:1	2.3	1384	1054	1.8	0.481	1.4	90		94
1:5	−2.3	1263	1024	−2.1	0.390	1.4	92		97
10:1	3.3	1475	1096	2.5	0.561	1.6	88		96
1:10	−3.3	1312	1000	−3.1	0.639	1.6	87		98
average		1362	1060		0.458	1.5	90	9	96

[a] AAD, average absolute deviation from median.

Point Estimation and Confidence Interval Estimation of Protein Abundance Ratios with Profile Likelihood Curves. The maximum likelihood point estimation and the profile likelihood confidence interval estimation of a protein's log-ratio is based on the protein's profile likelihood curve, which is constructed from the quantified peptides of the protein. Figure 3 shows profile likelihood curves (blue curves) together with peptide data points (red points) for four proteins in the 1:5 standard mixture.

The maximum likelihood point estimate is the protein log-ratio with the maximum likelihood, i.e., the log-ratio position of the highest peak in the profile likelihood curve. Conceptually, the sharper the peak is, the more precise the maximum likelihood estimate is and, therefore, the narrower the confidence interval should be. This intuition is captured by how the profile likelihood confidence interval is estimated. The confidence interval at the nominal confidence level of 95% is the log-ratio range of the curve segment above the *ln*-likelihood threshold (the horizontal lines shown in Figure 3) at 1.96 units ($0.5\cdot\chi^2_{1,\ 0.05}$) below the peak top. A profile likelihood confidence interval can be asymmetric, with different distances from the lower and upper interval limits to the point estimate.

The shape of the profile likelihood curve is determined by peptide data points in a number of ways, as illustrated in Figure 3 showing four proteins with an expected log-ratio of −2.3. First, a profile likelihood curve forms a peak at the log-ratio location with largest density of peptide data points of high log-profile-S/N. To illustrate this, Figure 3A shows the profile likelihood curve of 50S ribosomal protein L9 (locus RPA3080). This protein has many peptides with high log-profile-S/Ns and consistent log-ratios. This leads to a high and sharp profile likelihood peak in the ln-likelihood range of [−260, −140] and a narrow confidence interval of [−2.4, −1.9].

Second, the log-ratio location of a profile likelihood peak is largely determined by the peptide data points with higher log-profile S/N. Figure 3B shows a putative oxidoreductase (locus RPA1290) with only two quantified peptides. The peptide with higher log-profile-S/N has a log-ratio of −1.8, and the other peptide has a log-ratio of −3.5. The log-ratio position of the profile likelihood peak top, i.e., the maximum likelihood estimate, is −2.3, which is closer to the log-ratio of the peptide with higher log-profile S/N.

Third, the protein log-ratio estimation accounts for the log-ratio estimation bias in peptides with low log-profile-S/N. Figure 3C shows a phosphoglycerate kinase (locus RPA0943) that has a large fraction of peptide data points with poor log-profile-S/Ns and log-ratios considerably biased toward 0. The prevalence of biased peptide log-ratios in the low log-profile-S/N region is also shown in Figure 1. A simple average of all peptide log-ratios would give a biased estimation of the protein log-ratio. In contrast, the profile likelihood peak is located in the log-ratio region containing the peptides with high log-profile-S/N, which yields an unbiased estimation of protein log-ratio. The peptides with low log-profile-S/N were used to suppress the *ln*-likelihood of the protein log-ratios on the right side of the biased peptide log-ratios.

Fourth, the profile likelihood peak excludes the peptide data points that are outliers in the log-ratio axis. Figure 3D shows the profile likelihood curve for a hypothetical protein (locus RPA4470). Only three peptides are quantified, and one of them is likely to be an outlier with an erroneous log-ratio. The outlier creates a small profile likelihood peak but has no effect on the large profile likelihood peak used for protein log-ratio estimation.

In summary, the point estimate and the confidence interval of a protein log-ratio is calculated with peptide log-ratio weighting, bias suppression, and outlier exclusion. All of these are achieved using the likelihood function of protein log-ratio (eq 6). A weakness of this algorithm is the need to set the following parameters in the likelihood function: (a) the proportion between the normal and uniform distributions in the mixture model (eq 3), which sets the tolerance to outliers modeled by the uniform distribution; and (b) the parameters in the linear models for inferring the standard deviation and the mean of peptide log-ratio distributions (eqs 4 and 5), which set the relative weights and the expected biases of peptide log-ratios with their log-profile-S/Ns. These parameters were estimated from experimental data in our previous study.[12] This weakness might be alleviated in the future by improving the likelihood function of our algorithm or by employing other related algorithms from data fusion,[24] pattern recognition,[25] data mining,[26] etc.

Benchmark of Protein Abundance Ratio Estimation Performance Using Standard Mixtures of Isotopically Labeled Proteomes. The profile likelihood algorithm was tested as a part of the program ProRata using the standard mixtures. A widely recognized challenge in proteomics is the enormous dynamic

[24] Hall, D. L.; McMullen, S. A. H. *Mathematical techniques in multisensor data fusion*, 2nd ed.; Artech House: Boston, MA, 2004.

[25] Marques, J. P. *Pattern Recognition: Concepts, Methods, and Applications*; Springer: Berlin, 2001.

[26] Larose, D. T. *Data Mining Methods and Models*; Wiley-Interscience: Hoboken, NJ, 2006.

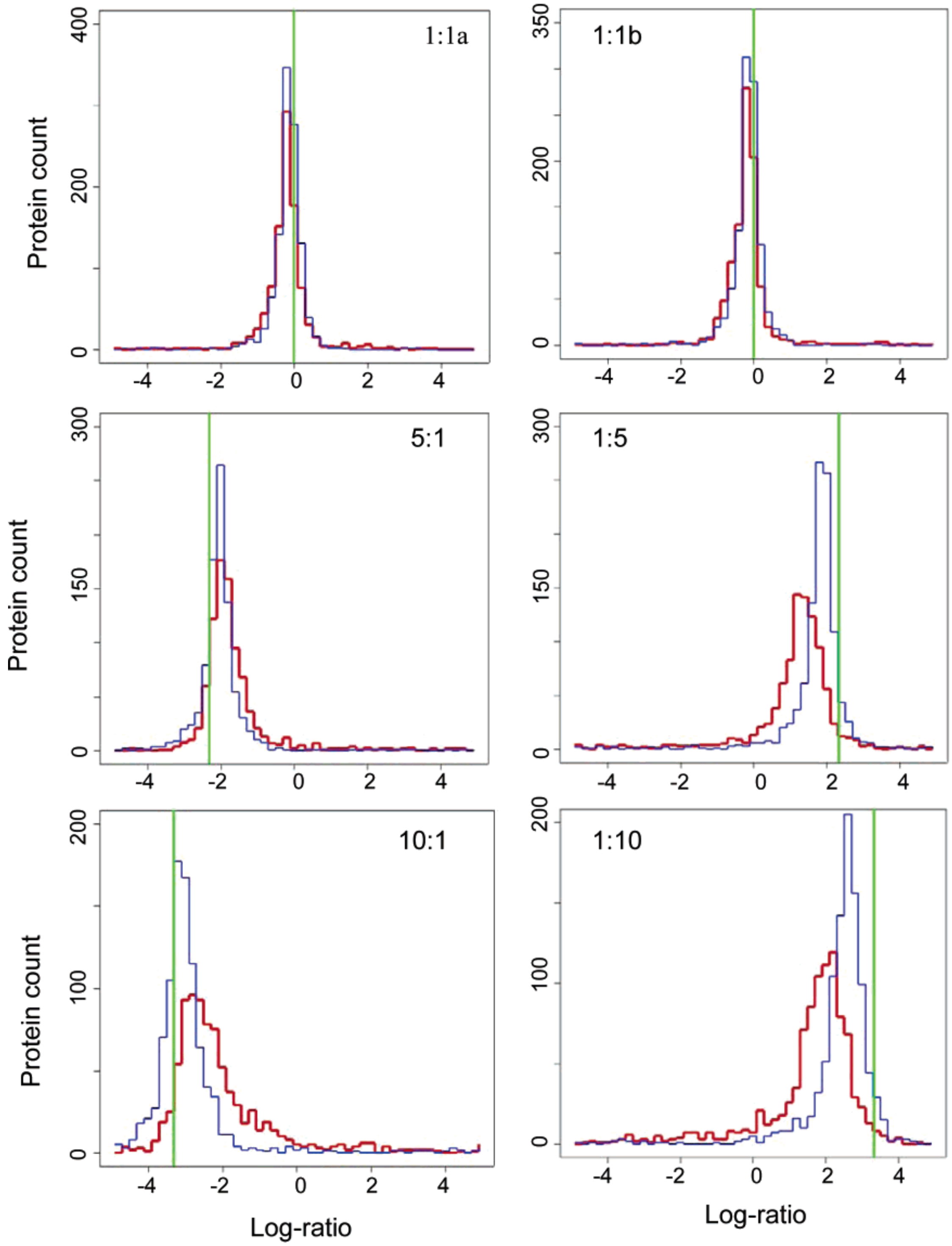


Figure 4. Comparison of protein log-ratio point estimation with RelEx[11] and ProRata. The histograms of the protein log-ratios estimated with the two programs (blue for ProRata and red for RelEx) are constructed for six standard mixtures. The mixing ratios in $\log_2$ scale are represented by the vertical green lines.

range between different proteins. Quantitative proteomics presents another type of dynamic range challenge: the potentially large abundance difference between the two isotopologues of a protein. We refer to the former dynamic range as protein dynamic range and to the latter one as isotopologue dynamic range. Various standard mixtures of metabolically labeled *R. palustris* proteomes were prepared with different mixing ratios, which represent the isotopologue dynamic range. The following five mixing ratios between the ^{14}N and ^{15}N proteomes were used in this study: 10:1, 10:1, 5:1, 1:5, and 1:1. The 1:1 mixture was analyzed in duplicate, and the two data sets are designated as 1:1a and 1:1b. The protein quantification results are presented in Supporting Information Tables S1–S6 and are summarized in Table 1.

On average, 1362 proteins were identified in a standard mixture (Table 1). Approximately 200 fewer proteins were identified from these standard mixtures than from the proteome sample before

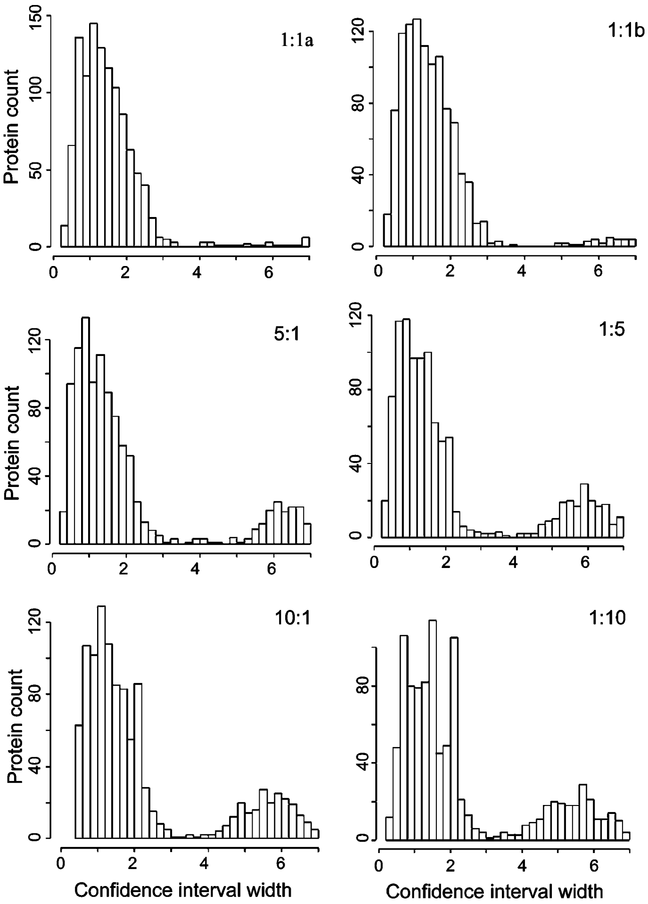


Figure 5. Histograms of the width of protein log-ratio confidence intervals. The distribution of confidence interval width reflects the varying quantification precision of proteins in a quantitative proteomics measurement.

mixing. This reduction is probably because the mixing essentially doubled the sample complexity. Full scans from the standard mixtures contain "doublet" peaks from the two isotopologues of peptides. Many MS/MS scans were targeted to different isotopologues of the same peptide rather than to new peptides. On average, 1060 proteins were quantified out of the 1362 identified proteins. Not every identified protein can be quantified, as quantification of a protein requires at least two quantified peptides with relatively high profile S/Ns and consistent log-ratios.

We compared RelEx's and ProRata's point estimation of protein abundance ratios. RelEx was used for comparison, because both RelEx and ProRata take the identification results from the DTASelect program and they employ a similar strategy for calculating peptide abundance ratios. RelEx, unlike ProRata, uses the average of peptide abundance ratios as the protein abundance ratio estimate. To evaluate the protein quantification results, the histogram of protein log-ratio estimates were constructed for each standard mixture. The protein log-ratio estimates should have the same true value in a standard mixture. Hence, the spread of the log-ratio distribution would reflect the random estimation error and the difference between the distribution center and the true log-ratio would reflect the systematic estimation error.

The protein log-ratio distributions for the two 1:1 standard mixtures are similar between ProRata (blue) and RelEx (red) (Figure 4, top). For the other mixtures, the protein log-ratio distributions from ProRata (blue) are closer to the true log-ratio (green line) and tighter than those from RelEx (red) (Figure 4, middle and bottom). This indicates that the profile likelihood algorithm gives a more accurate and precise point estimation of protein log-ratio than averaging. The median and the average absolute deviation of ProRata's point estimation are shown in Table 1. The average absolute deviation is the average difference of the point estimates from their median, which indicates the spread of the distribution.

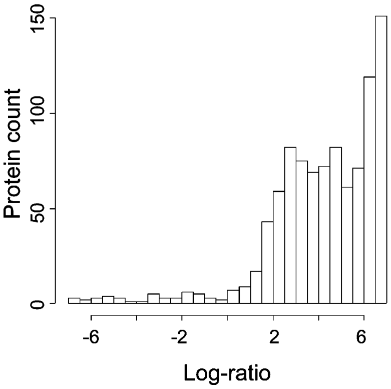
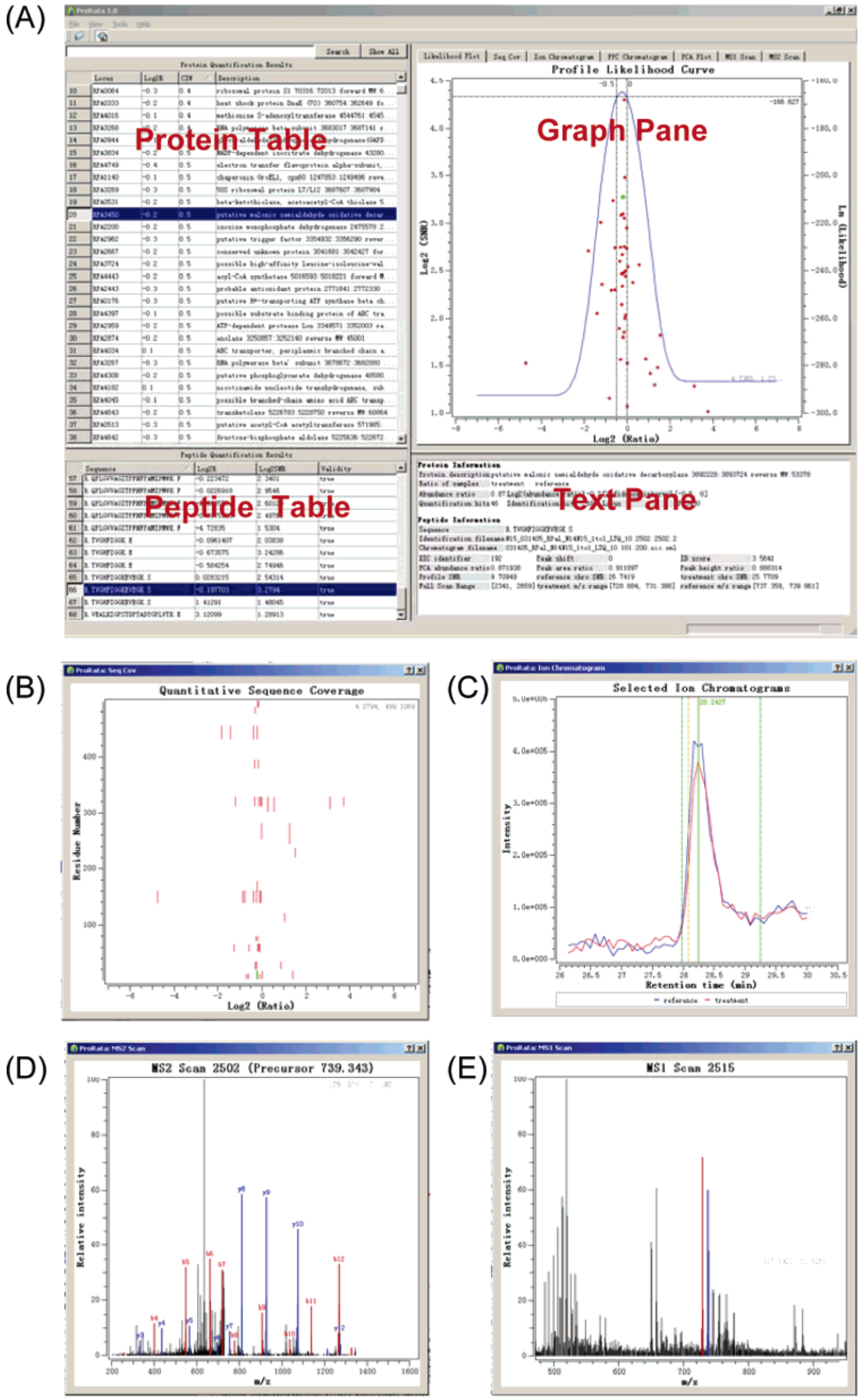


Figure 6. Histogram of the log-ratio estimates for proteins with extremely large abundance change. All proteins are expected to have a very large abundance ratio between the light isotopologue and the nonexistent heavy isotopologue. 26% of the estimated protein abundance ratios are greater than 64:1 and 90% are greater than 4:1.

The profile likelihood confidence intervals were also estimated for the quantified proteins. The confidence interval width for the majority of proteins was distributed between 0 and 3 (Figure 5). We observed that confidence intervals are generally smaller for high-abundance proteins, such as ribosomal proteins, than for low-abundance proteins, such as DNA polymerase proteins. In the 5:1, 1:5, 1:10, and 10:1 standard mixtures, there was a distinct, small distribution of confidence intervals that are wider than 4. This distribution largely stems from highly asymmetric confidence intervals that have a lower limit extending to the minimum log-ratio of −7.0 or an upper limit to the maximum log-ratio of 7.0. For example, consider a dehydrogenase protein (locus RPA4259) in the 10:1 standard mixture. The point estimate for this protein's log-ratio was 2.1, and the confidence interval was [1.0, 7.0]. The profile likelihood algorithm only determined that the log-ratio is greater than 1.0 and extended the upper limit to the maximum log-ratio, 7.0, which gave rise to a wide confidence interval.

The confidence level was benchmarked as the percentage of the true confidence intervals. In a standard mixture, a confidence interval was determined to be true if it contained the median of the protein log-ratio point estimates. On average, 955 proteins out of the 1060 quantified proteins in a standard mixture had true confidence intervals. This means that, although the confidence interval estimation has a nominal confidence level of 95%, only an average confidence level of 90% was obtained in the standard mixtures (Table 1). The decrease of the observed confidence level from the nominal one is probably because the peptide log-ratio probability model (eqs 3−5) is only an approximation to the true distribution. The confidence level of the interval estimation can be increased at the expense of widening the confidence intervals. This can be achieved by increasing the value of α in the *ln*-likelihood threshold, $\ln(L_{\max}) - 0.5 \cdot \chi^2_{1,\alpha}$, which lowers the ln-



Protein Quantification Results

	Locus	LogR	CIW	Description
10	RPA3064	−0.3	0.4	ribosomal protein S1 70316:72013 forward MW:6...
11	RPA0333	−0.2	0.4	heat shock protein DnaK (70) 360754:362649 fo...
12	RPA4016	−0.1	0.4	methionine S-adenosyltransferase 4544761:4545...
13	RPA3268	−0.2	0.4	RNA polymerase beta subunit 3683017:3687141 r...
14	RPA2944	[illegible]	[illegible]	[illegible] dehydrogenase (GAPD...
15	RPA3834	−0.2	0.5	NADP-dependent isocitrate dehydrogenase 43280...
16	RPA4749	−0.4	0.5	electron transfer flavoprotein alpha-subunit,...
17	RPA1140	−0.1	0.5	chaperonin GroEL1, cpn60 1247053:1249496 reve...
18	RPA3269	−0.3	0.5	50S ribosomal protein L7/L12 3687607:3687984 ...
19	RPA0531	−0.2	0.5	beta-ketothiolase, acetoacetyl-CoA thiolase 5...
20	RPA3450	−0.2	0.5	putative malonic semialdehyde oxidative decar...
21	RPA2200	−0.2	0.5	inosine monophosphate dehydrogenase 2475578:2...
22	RPA2962	−0.3	0.5	putative trigger factor 3354832:3356290 rever...
23	RPA2667	−0.2	0.5	conserved unknown protein 3041681:3042427 for...
24	RPA3724	−0.2	0.5	possible high-affinity leucine-isoleucine-val...
25	RPA4443	−0.2	0.5	acyl-CoA synthetase 5016593:5018221 forward M...
26	RPA2443	−0.3	0.5	probable antioxidant protein 2771641:2772330 ...
27	RPA0176	−0.3	0.5	putative H+-transporting ATP synthase beta ch...
28	RPA4397	−0.1	0.5	possible substrate binding protein of ABC tra...
29	RPA2959	−0.2	0.5	ATP-dependent protease Lon 3349571:3352003 re...
30	RPA2874	−0.2	0.5	enolase 3250857:3252140 reverse MW:45001
31	RPA4034	0.1	0.5	ABC transporter, periplasmic branched chain a...
32	RPA3267	−0.3	0.5	RNA polymerase beta' subunit 3678672:3682980 ...
33	RPA4308	−0.2	0.5	putative phosphoglycerate dehydrogenase 48580...
34	RPA4182	0.1	0.5	nicotinamide nucleotide transhydrogenase, sub...
35	RPA4045	−0.1	0.5	possible branched-chain amino acid ABC transp...
36	RPA4643	−0.2	0.5	transketolase 5226783:5228750 reverse MW:68064
37	RPA0513	−0.3	0.5	putative acetyl-CoA acetyltransferase 571985:...
38	RPA4642	−0.3	0.5	fructose-bisphosphate aldolase 5225636:522672...

Peptide Quantification Results

	Sequence	LogR	LogSNR	Validity
57	R.QFLGVVAGITPPNPFAMIPMWK.F	−0.223472	2.3401	true
58	R.QFLGVVAGITPPNPFAMIPMWK.F	−0.0226918	2.9546	true
59	R.QFLGVVAGITPPNP[illegible]	[illegible]	2.601[illegible]	[illegible]
60	R.QFLGVVAGITPPNP[illegible]	[illegible]	2.49[illegible]	[illegible]
61	R.QFLGVVAGITPPNPFAMIPMWK.F	−4.72835	1.5304	true
62	R.TVGHFIGGK.E	−0.0961407	2.03838	true
63	R.TVGHFIGGK.E	−0.673575	3.24286	true
64	R.TVGHFIGGK.E	−0.584254	2.74946	true
65	R.TVGHFIGGKEVEGK.S	0.0283215	2.54314	true
66	R.TVGHFIGGKEVEGK.S	−0.197703	3.2794	true
67	R.TVGHFIGGKEVEGK.S	1.41291	1.48045	true
68	R.VEALEIGPSTDPTADTGPLVTK.E	3.12099	3.28913	true

Protein Information
Protein description putative malonic semialdehyde oxidative decarboxylase 3092228:3093724 reverse MW:53378
Ratio of samples treatment : reference
Abundance ratio 0.87 Log2(abundance ratio) -0.2 Confidence interval [-0.4, 0]
Quantification hits 45 Identification hits [illegible]

Peptide Information
Sequence R.TVGHFIGGKEVEGK.S
Identification filename R15_031405_RPal_R14R15_1to1_LTQ_10.2502.2502.2
Chromatogram filename 031405_RPal_R14R15_1to1_LTQ_10.101.200.xic.xml
XIC identifier 192 Peak shift 0 ID score 3.5641
PCA abundance ratio 0.871938 Peak area ratio 0.911897 Peak height ratio 0.888314
Profile SNR 9.70949 reference chro SNR 26.7419 treatment chro SNR 25.7709
Full Scan Range [2341, 2659] treatment m/z range [728.884, 731.388] reference m/z range [737.359, 739.861]

Figure 7. Graphical user interface of ProRata. The main window of ProRata has four panes: Protein Table, Peptide Table, Graph Pane, and Text Pane (A). The Graph Pane contains two protein plots (sequence coverage plot (B) and profile likelihood curve plot), three peptide plots (selected ion chromatograms (C), parallel paired covariance chromatogram, and principal component analysis of peak profile), and two types of mass spectra (MS/MS scans (D) and full scans (E)). In the sequence coverage plot of a protein, a peptide is represented with a vertical segment indicating its log-ratio and its location on the protein sequence (B).

likelihood threshold (the horizontal dashed line in the profile likelihood curves shown in Figure 3).

Confidence interval estimation enables hypothesis testing on the abundance change of a protein. The hypothesis testing can be used to filter the quantified proteins and select those with significant abundance change for further examination. Since most of the proteins in a proteome are not affected by a treatment, the null hypothesis is that there is no statistically significant difference in the abundance of a protein between two proteomes. The alternative hypothesis is that there is such a difference. The null hypothesis can be rejected for a protein if the protein's confidence interval does not include zero.

Two performance characteristics of a hypothesis testing method are its power and significance. A post hoc significance test was performed using the two 1:1 standard mixtures, in which the null hypothesis should hold for all proteins (Table 1). The average significance was 9%, which means that 9% of the proteins with no abundance change are falsely asserted to have significant change. A post hoc power analysis was then performed using the standard mixtures with the other mixing ratios. The average power was 96%, which means that 96% of the proteins with abundance change are correctly identified. The abundance change of those proteins with accepted alternative hypothesis is only statistically significant, and the proteins should then be selected by the biological significance of their abundance change.

Testing of Abundance Ratio Estimation for Proteins with Extremely Large Abundance Change. The performance of the profile likelihood algorithm is sensitive to the isotopologue dynamic range. As the abundance difference between the two isotopologues increases among different standard mixtures, the average absolute deviation of the log-ratio point estimation increases and the confidence level of the interval estimation drops (Table 1). As the performance decrease is not very significant, we believe that the isotopologue dynamic range can reach 10-fold abundance difference with the LTQ-MS instrument and the ProRata program.

However, real-world biological samples might have proteins with extremely large abundance change, such as present in one proteome and absent in the other. We tested ProRata with an unlabeled proteome sample (Supporting Information Table S7). In this case, all proteins only have the light isotopologue, and their log-ratios between the two isotopologues are expected to be infinity. Ideally, all protein log-ratios estimated by ProRata should appear at the log-ratio of 7. A total of 961 proteins were quantified. The histogram of their log-ratio estimates is shown in Figure 6. About 250 proteins have a log-ratio estimate next to the maximum log-ratio and the log-ratio estimates for most other proteins are evenly distributed between 2.0 and 6.0. The underestimation of protein log-ratios is derived from the underestimation of peptide log-ratios, which stems from noise fluctuations that falsely define the abundance of the nonexistent heavy isotopologue.

Confidence interval estimation and hypothesis testing were also tested. The confidence intervals for half of the quantified proteins have an upper limit at the maximum log-ratio, which means only the lower bound can be estimated for those protein log-ratios. In this unlabeled proteome sample, the alternative hypothesis should be accepted for all proteins because of their change from present to absent. We found it to be the case for 97% of the quantified proteins. This percentage agrees with the observed power of the hypothesis testing in other standard mixtures. Therefore, although many of these proteins with very large abundance change have considerable error in their log-ratio point estimates, most of them can be correctly identified to have significant abundance change.

Manual Inspection of Protein Abundance Ratio Estimation through ProRata Graphical User Interface. Point estimation and confidence interval estimation of protein log-ratios were performed with ProRata automatically. However, there were a small percentage of proteins with spurious estimation results. One effective way of reducing the uncertainty is to manually validate the proteins of interest. ProRata is equipped with a graphical user interface to enable interactive data interrogation and facilitate manual result validation.

ProRata's graphical user interface has four panes in its main window, including a Protein Table, a Peptide Table, a Text Pane, and a Graph Pane (Figure 7A). It was designed to give users a hierarchical view of their proteomics measurements. All quantified proteins are listed in the Protein Table. When a protein of interest is selected from the Protein Table, its profile likelihood curve and sequence coverage (Figure 7B) are displayed in the Graph Pane and its peptides are listed in the Peptide Table. Then a peptide from this protein can be selected to show its selected ion chromatograms (Figure 7C) and MS/MS scans (Figure 7D). Furthermore, the full scan at a retention time point in the selected ion chromatograms can be viewed with the mass spectral peaks for the two isotopologues highlighted (Figure 7E).

We have examined the proteins with erroneous point estimation, false confidence interval estimation, or both. These proteins usually have less than three reliably quantified peptides. The reliably quantified peptides can generally be ascertained by inspecting their MS/MS scans, full scans, selected ion chromatograms, and peak profiles. Therefore, manual validation of the automated estimation results can provide an additional safeguard in quantitative proteomics against yielding false information.

CONCLUSIONS

In this study, we applied maximum likelihood point estimation and profile likelihood confidence interval estimation for protein abundance ratio evaluation in quantitative shotgun proteomics with a profile likelihood algorithm. This algorithm is able to weigh peptide abundance ratios by their estimation variability, account for peptide abundance ratio estimation bias, and suppress contribution from outliers. The algorithm was tested with standard mixtures of isotopically labeled proteomes at various mixing ratios. We demonstrated that the point estimation accuracy was improved using maximum likelihood estimation. The confidence intervals were estimated at the observed confidence level of 90%. With confidence interval estimation, hypothesis testing was performed on protein abundance change, which was benchmarked to have a significance of 9% and a power of 96%. The profile likelihood algorithm was also tested with an unlabeled proteome sample to show its ability to analyze proteins with extremely large abundance change. The profile likelihood algorithm was built into a computer program, ProRata, which automates the entire data analysis procedure for quantitative shotgun proteomics. ProRata's graphical user interface allows for manual validation of protein quantification results.

ACKNOWLEDGMENT

The computational and experimental work was supported by the DOE Genomics:GTL program of the U.S. Department of Energy (Office of Biological and Environmental Research, Office of Science) and by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory (ORNL). Judson Hervey is gratefully acknowledged for helpful discussions and software testing. The work of G.K. was funded by the Scientific Data Management Center (http://sdmcenter.lbl.gov) under the Department of Energy's Scientific Discovery through Advanced Computing program (http://www.scidac.org). Oak Ridge National Laboratory is managed by the University of Tennessee—Battelle, L.L.C. for the Department of Energy under contract DOE-AC05-00OR22725.

SUPPORTING INFORMATION AVAILABLE

Additional information as noted in text. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review April 7, 2006. Accepted July 24, 2006.

AC060654B