Anal. Chem. **2006,** *78,* 7110−7120

Robust Estimation of Peptide Abundance Ratios and Rigorous Scoring of Their Variability and Bias in Quantitative Shotgun Proteomics

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

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

The abundance ratio between the light and heavy isotopologues of an isotopically labeled peptide can be estimated from their selected ion chromatograms. However, quantitative shotgun proteomics measurements yield selected ion chromatograms at highly variable signal-to-noise ratios for tens of thousands of peptides. This challenge calls for algorithms that not only robustly estimate the abundance ratios of different peptides but also rigorously score each abundance ratio for the expected estimation bias and variability. Scoring of the abundance ratios, much like scoring of sequence assignment for tandem mass spectra by peptide identification algorithms, enables filtering of unreliable peptide quantification and use of formal statistical inference in the subsequent protein abundance ratio estimation. In this study, a parallel paired covariance algorithm is used for robust peak detection in selected ion chromatograms. A peak profile is generated for each peptide, which is a scatterplot of ion intensities measured for the two isotopologues within their chromatographic peaks. Principal component analysis of the peak profile is proposed to estimate the peptide abundance ratio and to score the estimation with the signal-to-noise ratio of the peak profile (profile signal-to-noise ratio). We demonstrate that the profile signal-to-noise ratio is inversely correlated with the variability and bias of peptide abundance ratio estimation.

In quantitative shotgun proteomics, proteolysis-derived peptides are measured with liquid chromatography−tandem mass spectrometry (LC−MS/MS) and used as surrogates of their parent proteins for relative quantification. In a label-free approach, the proteomes under comparison are analyzed separately in standardized LC−MS/MS runs. Alternatively, by employing stable isotope labeling, the proteomes under comparison are mixed together and analyzed in one LC−MS/MS run, which eliminates the variability in sample processing steps after mixing and LC−MS/MS analysis. The common stable isotope labeling methods include ^{15}N or ^{13}C metabolic labeling,[1] SILAC,[2] $H_2{}^{18}O$ digestion,[3] and ICAT.[4] Each peptide in the mixture of two isotopically labeled proteomes has two mass-different isotopic variants, the light isotopologue from one proteome and the heavy isotopologue from the other. Here, we consider algorithms for peptide relative quantification using mass-different stable isotope labeling. Note that the algorithms discussed here are not applicable to the isobaric labeling method, iTRAQ,[5] which generates specific reporter ions in tandem mass spectra for quantification.

Figure 1 illustrates the general computational procedure for estimating the abundance ratio between the light isotopologue and the heavy isotopologue of a peptide. The sequence of the peptide is identified from an MS/MS scan of one of its isotopologues (Figure 1A). The full scan that triggered this MS/MS scan is shown in Figure 1B, in which the mass spectral peaks of the two isotopologues are highlighted. Selected ion chromatograms for the two isotopologues are then extracted, and peak detection is performed to define front and back boundaries of the two isotopologues' chromatographic peaks (Figure 1C). Finally, the abundance ratio of the peptide is evaluated from the two chromatographic peaks. In this study, we developed novel algorithms for peak detection and for peptide abundance ratio evaluation.

Normally, peak detection is performed in the selected ion chromatograms. However, a large fraction of chromatographic peaks have a very low chromatographic signal-to-noise ratio (S/N) that results in incorrect assignments of their peak boundaries. We have improved the robustness of peak detection by employing a parallel paired covariance algorithm, which was developed based

* To whom correspondence should be addressed. For questions on computational methods, contact N.F.S. Phone: (865) 241-4351. E-mail: samatovan@ornl.gov. For questions on experimental methods, contact R.L.H. Phone: (865) 574-4968. E-mail: hettichrl@ornl.gov.

† Computational Biology Institute.

‡ Chemical Sciences Division.

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

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

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

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

(5) Ross, P. L.; Huang, Y. N.; Marchese, J. N.; Williamson, B.; Parker, K.; Hattan, S.; Khainovski, N.; Pillai, S.; Dey, S.; Daniels, S.; Purkayastha, S.; Juhasz, P.; Martin, S.; Bartlet-Jones, M.; He, F.; Jacobson, A.; Pappin, D. J. *Mol. Cell. Proteomics* **2004**, *3* (12), 1154−1169.

10.1021/ac0606554 CCC: $33.50
Published on Web 09/13/2006

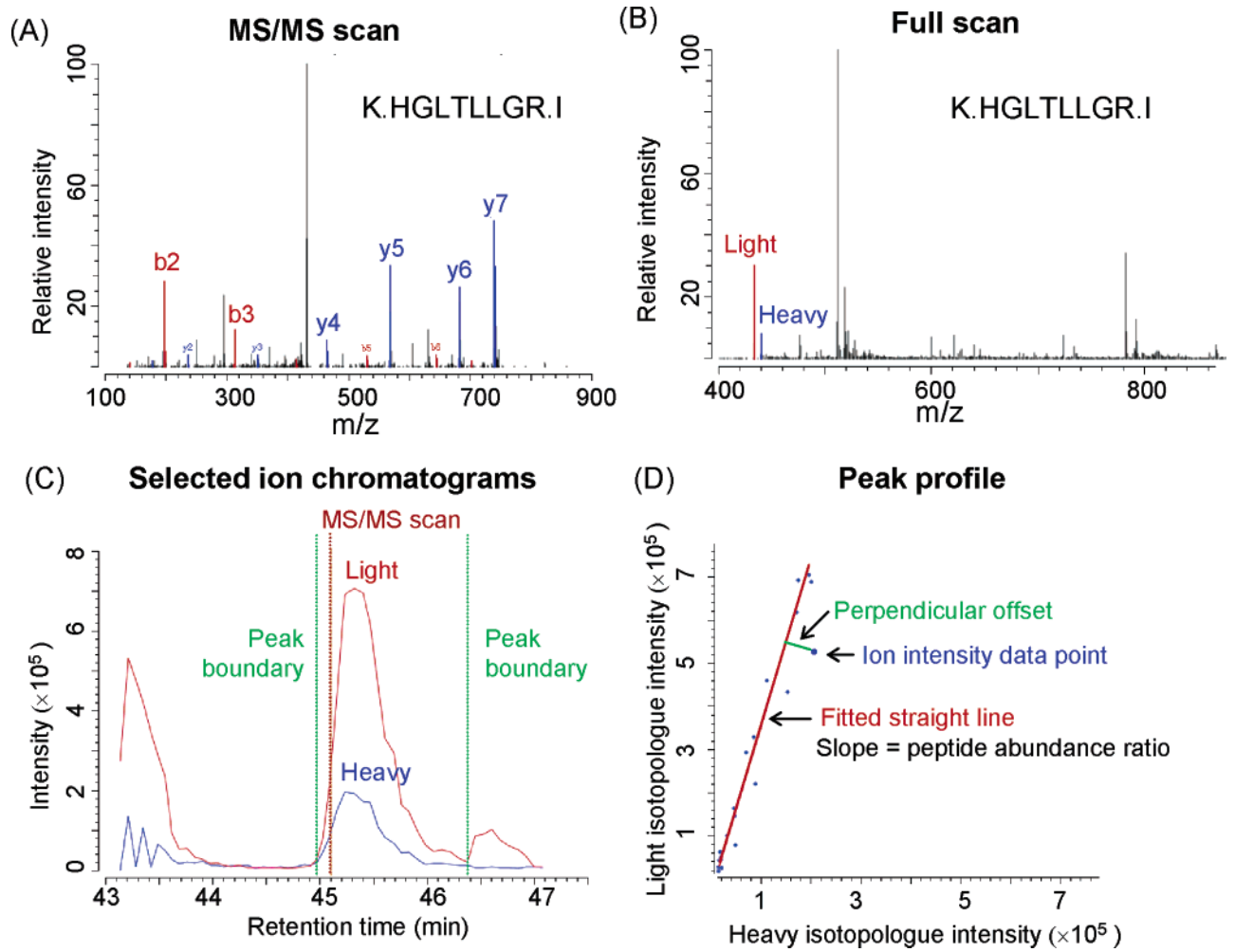

Figure 1. Estimation of peptide abundance ratios in quantitative shotgun proteomics. (A) The sequence of a peptide is identified from an MS/MS scan. (B) The peak pair in the full scan for the two isotopologues of this peptide is identified. (C) The selected ion chromatograms of the light isotopologue (red) and the heavy isotopologue (blue) are extracted from the full scans. The red vertical line indicates the MS/MS scan. The chromatographic peaks of the isotopologue pair are detected as between the two vertical green lines. The abundance ratio between the two isotopologues can be estimated as the ratio of the peak areas. (D) The abundance ratio is estimated from a peak profile. The blue data points represent the ion intensities of the two isotopologues measured in the full scans within the chromatographic peak. The red line has the minimum total squared perpendicular offset to the data points, whose slope is an abundance ratio estimator.

on a sequential paired covariance algorithm originally devised for rapid component identification from ion electropherograms.[6,7] The parallel paired covariance algorithm integrates the two selected ion chromatograms and reconstructs a covariance chromatogram of improved chromatographic S/N for peak detection.

Estimation of the peptide abundance ratios from the detected chromatographic peaks can be accomplished with several existing algorithms. A common algorithm is based on peak area. First, the selected ion chromatograms are smoothed to remove random noise. Then, the background is subtracted from the chromatograms. Finally, the area under the two chromatographic peaks is calculated by integration. The ratio between the two peak areas is considered as the abundance ratio of a peptide. The peak area algorithm has been used in quantitative proteomics programs XPRESS,[8] ASAPratio,[9] and MSQuant.[10] The accuracy of the peak area calculation highly depends on two empirical steps—chromatogram smoothing and background subtraction. MacCoss et al. argued that background subtraction is difficult to optimize for thousands of different chromatographic peaks measured in a proteomics experiment and leads to less reliable abundance ratio estimation.[11] In their program, RelEx, a correlation algorithm based on peak profiles[12] is used to calculate peptide abundance ratios.[11] A peak profile is a scatterplot of ion intensities of the two isotopologues detected in each full scan within the chromatographic peaks (Figure 1D). The correlation algorithm fits a straight line that has the minimum total squared perpendicular offset to the data points in the peak profile (Figure 1D). The slope of the fitted line is an estimator of the peptide abundance ratio.

(6) Muddiman, D. C.; Rockwood, A. L.; Gao, Q.; Severs, J. C.; Udseth, H. R.; Smith, R. D.; Proctor, A. *Anal. Chem.* **1995**, *67* (23), 4371−4375.
(7) Muddiman, D. C.; Huang, B. M.; Anderson, G. A.; Rockwood, A.; Hofstadler, S. A.; WeirLipton, M. S.; Proctor, A.; Wu, Q. Y.; Smith, R. D. *J. Chromatogr., A* **1997**, *771* (1−2), 1−7.
(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.

Existing algorithms that evaluate peptide abundance ratios do not formally "score" the abundance ratio estimates for their expected bias and variability. In quantitative shotgun proteomics, the abundance ratios for tens of thousands of identified peptides can be estimated, but with dramatically varying error. We propose a principal component analysis algorithm that not only estimates the abundance ratio of a peptide from its peak profile but also scores the estimation with a signal-to-noise ratio measure of its peak profile (profile-S/N). We show that the profile-S/N is inversely correlated with both the standard deviation and the bias of the abundance ratio estimation. Thus, the profile-S/N allows stratification of the peptide abundance ratios into those with greater or lesser estimation accuracy and precision. As a result, it becomes possible to statistically evaluate every peptide abundance ratio in the subsequent protein abundance ratio estimation.[13]

Here we describe both the parallel paired covariance algorithm for peak detection and the principal component analysis algorithm for peptide abundance ratio estimation and scoring. These two algorithms have been assembled into a computer program, termed ProRata. ProRata automates the entire data analysis pipeline for quantitative proteomics with stable isotope labeling, incorporating selected ion chromatogram extraction and peptide abundance ratio evaluation for the ultimate goal of protein abundance ratio estimation.[13] The graphical user interface of ProRata also allows manual data interrogation for result validation. ProRata is freely available at www.MSProRata.org.

MATERIALS AND METHODS

Standard Isotopically Labeled Proteome Mixture Preparation. Wild type *Rhodopseudomonas palustris* CGA0010 strain was grown anaerobically in light on defined minimal growth medium to mid-log phase at 30 °C. $(NH_4)_2SO_4$ was the only nitrogen source available for bacterial assimilation and was provided as $(^{14}NH_4)_2SO_4$ for the unlabeled culture and as $(^{15}NH_4)_2SO_4$ for the ^{15}N-labeled culture (>98 atom percentage excess, Sigma-Aldrich, St. Louis, MO). The ^{15}N-enriched nitrogen from $(NH_4)_2SO_4$ was incorporated into proteins through metabolism in the ^{15}N-labeled culture. Except for the different isotopologues of $(NH_4)_2SO_4$ in the growth medium, the two cultures were otherwise identically prepared. Cells were harvested by centrifugation and washed twice with ice-cold wash buffer (50 mM Tris-HCl buffer at pH 7.5 with 10 mM EDTA). Cells were then lysed by sonication in ice-cold wash buffer, and unbroken cells were removed with low-speed centrifugation (5000*g* for 10 min). The resulting cell lysates were fractionated by ultracentrifugation at 100000*g* for 1 h, and the supernatants from the unlabeled and ^{15}N-labeled cell lysates were labeled as the ^{14}N proteome and ^{15}N proteome, respectively. Protein concentration in each proteome was determined with Lowry's analysis.[14] Standard mixtures were prepared by mixing the two proteomes at $^{14}N/^{15}N$ ratios of 10:1, 5:1, 1:1, 1:5, and 1:10 with respect to their total protein mass.

Shotgun Proteomics Measurement. The proteins in the standard mixtures were denatured and reduced with treatment of 6 M guanidine and 10 mM dithiothreitol (DTT) (Sigma Chemical Co., St. Louis, MO) at 60 °C for 1 h. After 6-fold dilution with 50 mM Tris-HCl/10 mM $CaCl_2$ (pH 7.8), the proteins were digested at 37 °C with sequencing grade trypsin (Promega, Madison, WI). The samples were then reduced with 20 mM DTT for 1 h at 60 °C and were desalted using C18 solid-phase extraction (Sep-Pak Plus, Waters, Milford, MA). The protein digests were examined with LC−MS/MS using 12-step split-phase MudPIT.[15,16] The samples were loaded via a pressure bomb (New Objective, Woburn, MA) onto a 250-μm-i.d. front column packed with 2-cm strong cation-exchange resin (Luna, Phenomonex) and 2-cm C18 reversed-phase resin (Aqua, Phenomonex). A 100-μm-i.d. PicoFrit column (New Objective) was packed with 15-cm C18 reversed-phase resin. The front column was connected with the PicoFrit column and then placed in-line with a Surveyor quaternary HPLC (ThermoFinnigan, San Jose, CA). The composition of the aqueous solvent was 95% H_2O (Burdick & Jackson, Muskegon, MI), 5% ACN (Burdick & Jackson, Muskegon, MI), and 0.1% formic acid (EM Science, Darmstadt, Germany), and the composition of the organic solvent was 30% H_2O, 70% ACN, and 0.1% formic acid. Two-dimensional LC separation was performed with 12 salt pulses (0, 35, 50, 60, 75, 100, 125, 150, 200, 250, 300, and 500 mM ammonium acetate (Sigma Chemical Co.) in the aqueous solvent). Each salt pulse was followed by a 2-h reversed-phase gradient from 100% aqueous solvent to 50% aqueous solvent and 50% organic solvent. LC−MS/MS analysis was performed on an LTQ linear ion trap instrument (ThermoFinnigan, San Jose, CA) with dynamic exclusion enabled. Each full scan (400−1700 m/z) was followed by five data-dependent MS/MS scans at 35% normalized collision energy. All scans were averaged from two microscans.

Peptide and Protein Identification. All MS/MS scans were searched with the SEQUEST program[17] against an *R. palustris* protein sequence database.[18] The light isotopologues of peptides were identified using normal amino acid masses in the SEQUEST parameter file, and the heavy isotopologues were identified using ^{15}N-labeled amino acid masses. OUT files were converted to SQT files using UNITEMARE program (generously provided by Dr. John R. Yates' laboratory). DTASelect[19] was used to filter the peptide identifications based on Xcorr and delCN (Xcorr > 1.8 (+1), > 2.5 (+2), and > 3.5 (+3); delCN > 0.08). The peptides were assembled into proteins, retaining duplicate MS/MS scans of a peptide (DTASelect option: −t 0).

Selected Ion Chromatogram Extraction. The Xcalibur RAW files were converted into the mzXML format with the ReAdW program.[20] An mzXML parser, RAMP, was used to access the mzXML files.[21] The selected ion chromatograms were extracted

(12) 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*; Lawson, A. M., Lim, C. K., Richmond, W., Eds.; Academic Press: London, 1980.

(13) Pan, C.; Kora, G.; McDonald, W. H.; Tabb, D. L.; VerBerkmoes, N. C.; Hurst, G. B.; Pelletier, D. A.; Samatova, N. F.; Hettich, R. L. *Anal. Chem.* **2006**, *78*, 7121−7131.

(14) Lowry, O. H.; Rosebrough, N. J.; Farr, A. L.; Randall, R. J. *J. Biol. Chem.* **1951**, *193* (1), 265−275.

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

(16) 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), 7900−5.

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

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

in the following steps: (1) The peptide identifications were parsed out from DTASelect-filter.txt, including their amino acid sequences, charge states, and protein loci. (2) The m/z windows were calculated for the two isotopologues of each peptide identification. Theoretical isotope distributions for both isotopologues were calculated based on the sequence, the user-defined isotopic compositions of the atoms, and the user-defined atomic compositions of all residues and their modifications. In this study, the nitrogen atoms in the heavy isotopologues were specified to be 98%-enriched ^{15}N. The m/z windows for an isotopic distribution were configured to be its major isotopes' m/z values plus and minus the m/z tolerance; the major isotopes were specified to be the isotopes with a relative abundance of more than 10%, and the m/z tolerance was defined to be 0.5 in this study. (3) The peptide identifications with the same sequence and charge state were grouped if their MS/MS scans were acquired within a 2-min interval. Redundant identifications from a single chromatographic peak of a peptide were often found for either or both isotopologues of the peptide. (4) A pair of selected ion chromatograms was extracted for the light and heavy isotopologues of each peptide identification group. The retention time window for both selected ion chromatograms was defined as from 2 min before the first MS/MS scan to 2 min after the last MS/MS scan of the grouped peptide identifications.

Chromatographic Peak Detection. The covariance chromatogram of an isotopologue pair was reconstructed from its selected ion chromatograms with the parallel paired covariance algorithm. The parallel paired covariance at a full scan in the covariance chromatogram is the product of the background-subtracted ion intensities at that full scan in the two selected ion chromatograms:

$$C_k = (I_k^{\mathrm{L}} - I_{\mathrm{BG}}^{\mathrm{L}})(I_k^{\mathrm{H}} - I_{\mathrm{BG}}^{\mathrm{H}}), \quad d \le k \le h \qquad (1)$$

where $I_{\mathrm{BG}}^{\mathrm{L}}$ and $I_{\mathrm{BG}}^{\mathrm{H}}$ are the background ion intensities of the selected ion chromatograms for the light and heavy isotopologues, respectively; I_k^{L} and I_k^{H} are the ion intensities at full scan k in the m/z windows for the light and heavy isotopologues, respectively; and C_k is the covariance of the two isotopologues' background-subtracted intensities at full scan k. The background ion intensities $I_{\mathrm{BG}}^{\mathrm{L}}$ and $I_{\mathrm{BG}}^{\mathrm{L}}$ were defined to be the minimum ion intensities in the selected ion chromatograms for the light and heavy isotopologues, respectively. Scan d and scan h are the first and last full scans, respectively, in the selected ion chromatograms. The time series of I_k^{L} and I_k^{H} form the selected ion chromatograms for the two isotopologues, and likewise, the time series of C_k forms the covariance chromatogram of the isotopologue pair, as illustrated in Figure 2. The covariance chromatogram was then smoothed with seven-point quadratic Savitsky−Golay filter.[22] The chromatographic peak for a peptide was defined as being between scan a and scan b ($d \le a < b \le h$), which are the two local covariance minima with the smallest interval that contains all MS/MS scans matched to this peptide (Figure 2). A local covariance minimum was a scan with the lowest covariance within a seven-point symmetric window surrounding this scan in the covariance chromatogram. Scans a and b were the peak boundaries as labeled in Figure 2. The parallel paired covariance algorithm is also capable of determining the retention time shift between the two isotopologues. For this process, the two selected ion chromatograms would be shifted relative to each other scan by scan until the peak height of the peptide in the covariance chromatogram is maximized.

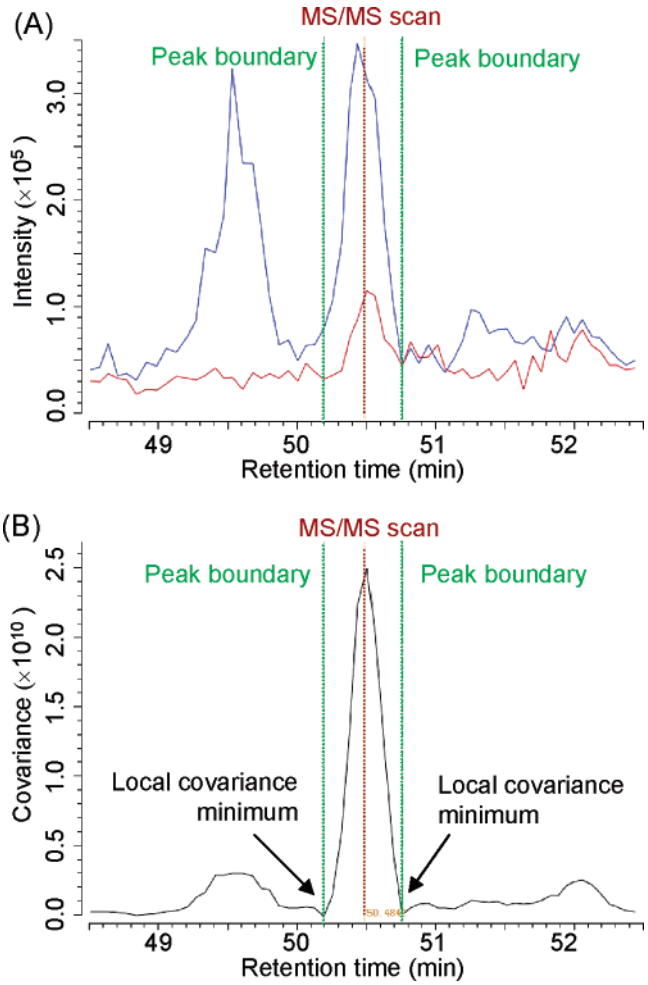

Figure 2. Selected ion chromatograms and parallel-paired covariance chromatogram. The selected ion chromatograms for the light isotopologue (red) and heavy isotopologue (blue) of a peptide are shown in part A, and the covariance chromatogram is shown in part B. The peak boundaries (vertical green lines) are determined in the covariance chromatogram and transferred to the selected ion chromatograms.

Peptide Abundance Ratio Estimation and Scoring. The peptide abundance ratios and the profile-S/Ns were estimated with principal component analysis of the peak profiles. Note that the background subtraction was only performed for peak detection. The peak profiles were constructed from the originally extracted selected ion chromatograms. Here, we present the equations used in the estimation and their derivation from a set of definitions and assumptions.

(20) Pedrioli, P. G. A.; Eng, J. K.; Hubley, R.; Vogelzang, M.; Deutsch, E. W.; Raught, B.; Pratt, B.; Nilsson, E.; Angeletti, R. H.; Apweiler, R.; Cheung, K.; Costello, C. E.; Hermjakob, H.; Huang, S.; Julian, R. K.; Kapp, E.; McComb, M. E.; Oliver, S. G.; Omenn, G.; Paton, N. W.; Simpson, R.; Smith, R.; Taylor, C. F.; Zhu, W. M.; Aebersold, R. *Nat. Biotechnol.* **2004**, *22* (11), 1459−1466.

(21) http://sashimi.sourceforge.net/software_glossolalia.html.

(22) Press: W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical Recipes in* C++, 2nd ed.; Cambridge University Press: Cambridge, UK, 2002; pp 655−660.

The detected ion intensities, I_i^L and I_i^H, of the light and heavy isotopologues at full scan i are composed of their true signal (denoted by S_i^L and S_i^H, respectively) corrupted by random noise (denoted by N_i^L and N_i^H, respectively) and superimposed on the backgrounds (denoted by B^L and B^H, respectively):[12]

$$\begin{cases} I_i^L = S_i^L + N_i^L + B^L \\ I_i^H = S_i^H + N_i^H + B^H, \end{cases} \quad a \le i \le b \tag{2}$$

where a and b are the chromatographic peak boundaries. Let us assume that (i) the backgrounds, B^L and B^H, hold constant across full scans; (ii) the random noises, N_i^L and N_i^H, have zero-mean; and (iii) the ratio between the two signals, S_i^L and S_i^H, is constant across scans and defines the peptide abundance ratio, R. The third assumption, expressed as eq 3, is based on the exact coelution of the two isotopologues:

$$R = S_i^L / S_i^H, \quad a \le i \le b \tag{3}$$

The constant background, B^L and B^H, can be eliminated from our consideration by centering the intensities and the signals on their means. Let I'^L_i and I'^H_i denote the mean-centered intensities and let S'^L_i and S'^H_i denote the mean-centered signals. Then eqs 2 and 3 can be transformed, respectively, to

$$\begin{cases} I'^L_i = S'^L_i + N_i^L \\ I'^H_i = S'^H_i + N_i^H, \end{cases} \quad a \le i \le b \tag{4}$$

$$R = S'^L_i / S'^H_i \tag{5}$$

This transformation is the reason why the peak profile-based algorithms can obviate the background subtraction step and, therefore, eliminate the probable error in this step.

Principal component analysis is generally applied to a set of vectors. Let us define the following vectors: the ion intensity vector $\mathbf{I}_i = (I_i^H, I_i^L)$, the signal vector $\mathbf{S}_i = (S'^H_i, S'^L_i)$, and the noise vector $\mathbf{N}_i = (N_i^H, N_i^L)$. Therefore, eqs 4 and 5 can be transformed to a vector form:

$$\mathbf{I}_i = \mathbf{S}_i + \mathbf{N}_i \tag{6}$$

$$R = \tan(\theta_{\mathbf{S}i}) \tag{7}$$

where $\theta_{\mathbf{S}_i}$ is the direction angle of the signal vector $\mathbf{S}_i$ (Figure 3). The vectors $\mathbf{I}_i$ are known from the measurement, but the vectors $\mathbf{S}_i$ and $\mathbf{N}_i$ are unknown and need to be determined for calculating the abundance ratio, R, and the profile signal-to-noise ratio. Obviously, $\mathbf{S}_i$ and $\mathbf{N}_i$ cannot be solved analytically from eq 6. Instead, determining $\mathbf{S}_i$ and $\mathbf{N}_i$ is formulated as an optimization problem of finding vectors $\mathbf{S}_i$ and $\mathbf{N}_i$ such that the variance of the norm of the noise vectors, $\sigma^2(|\mathbf{N}_i|)$, is minimized. This optimization problem can be solved by principal component analysis of the peak profile, as shown in eq 8,

$$\begin{cases} \mathbf{S}_i = (\mathbf{I}_i \cdot \mathbf{P}_1)\mathbf{P}_1 \\ \mathbf{N}_i = (\mathbf{I}_i \cdot \mathbf{P}_2)\mathbf{P}_2 \end{cases} \tag{8}$$

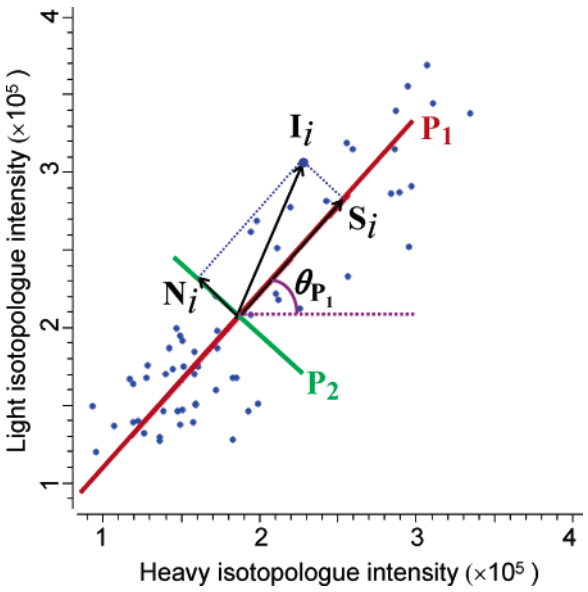

Figure 3. Estimation of peptide abundance ratio with the principal component analysis algorithm. Two principal components are represented with two lines, the red line for the first principal component $\mathbf{P}_1$ and the green line for the second principal component $\mathbf{P}_2$. The ratio between the lengths of the two lines is plotted to be equal to the profile-S/N, which captures how elliptical the ensemble of the data points in the peak profile is. The ion intensity vector for each data point ($\mathbf{I}_i$) can be decomposed to the signal vector ($\mathbf{S}_i$, the projection of $\mathbf{I}_i$ on $\mathbf{P}_1$) and the noise vector ($\mathbf{N}_i$, the projection of $\mathbf{I}_i$ on $\mathbf{P}_2$). The slope of $\mathbf{P}_1$ (tan($\theta_{\mathbf{P}_1}$)) is an estimator of the peptide abundance ratio.

where $\mathbf{P}_1$ and $\mathbf{P}_2$ are the corresponding first and second principal components of the intensity vectors $\mathbf{I}_i$. This means that the direction of all signal vectors is the direction of the first principal component and the length of a signal vector is the dot product between the intensity vector and the first principal component. The direction and length of the noise vectors are determined, likewise, with the second principal component. Geometrically, a signal vector and a noise vector are the projections of their intensity vector on the first principal component and the second principal component, respectively as illustrated in Figure 3. Principal component analysis of the intensity vectors $\mathbf{I}_i$ in the peak profile calculates the principal components $\mathbf{P}_1$ and $\mathbf{P}_2$ and their associated eigenvalues λ_1 and λ_2. The principal components and eigenvalues provide the estimators for the peptide abundance ratios and profile-S/Ns, as described below.

The abundance ratio is the tangent of the direction angle of the first principal component:

$$R = \tan(\theta_{\mathbf{P}_1}) \tag{9}$$

The abundance ratio estimated with principal component analysis is exactly the same as the abundance ratio estimated with linear correlation. This is because the direction of the first principal component is exactly the same as the direction of the straight line with minimum total squared perpendicular offset,[23] both of which are estimators of the peptide abundance ratio.

Let us define the signal-to-noise ratio for the ion intensity vectors in the peak profile as the ratio between the standard

(23) Jolliffe, I. T. *Principal component analysis*; Springer series in statistics; Springer-Verlag: New York, 2002: pp 34−36.

Table 1. Peptide Quantification Results from the Six Standard Mixture Data Sets

standard mixtures		peptide counts				log-ratio		log-profile-S/N	
$^{14}N/^{15}N$	log-ratio	^{14}N ID	^{15}N ID	SIC[a]	quantified	median	AAD[b]	median	AAD[b]
1:1a	0.0	13 766	11 665	17 574	11 919	−0.17	0.67	2.17	0.58
1:1b	0.0	13 975	12 230	18 472	12 958	−0.16	0.69	2.15	0.56
5:1	2.3	23 527	5 122	23 256	14 583	1.66	1.04	2.29	0.73
1:5	−2.3	5 676	18 037	18 855	12 453	−1.99	0.94	2.44	0.78
10:1	3.3	24 257	2 725	22 770	12 914	2.20	1.37	2.31	0.85
1:10	−3.3	3 167	21 396	20 945	12 910	−2.80	1.39	2.46	0.91
average				20 312	12 956		1.02	2.30	0.74

[a] SIC, selected ion chromatogram. [b] AAD, absolute average deviation from the median.

deviation of the length of the signal vectors and that of the length of the noise vectors:

$$\mathrm{S/N}_{\mathrm{profile}} \equiv \frac{\sigma(|\mathbf{S}_i|)}{\sigma(|\mathbf{N}_i|)} \tag{10}$$

We refer to this signal-to-noise ratio as the profile signal-to-noise ratio to distinguish it from the chromatographic S/N, since the profile-S/N is based on the peak profile, while the chromatographic S/N is based on the ion chromatogram. The first eigenvalue, λ_1, is the variance of the projection of the intensity vectors on the first principal component, which is the variance of the length of the signal vectors. Likewise, the second eigenvalue, λ_2, for the second principal component is the variance of the length of the noise vectors. The profile-S/N is calculated as the square root of the ratio between λ_1 and λ_2:

$$\mathrm{S/N}_{\mathrm{profile}} = \sqrt{\frac{\lambda_1}{\lambda_2}} \tag{11}$$

The calculation of the profile-S/N from the eigenvalues is the key feature of the principal component analysis algorithm that distinguishes it from the linear correlation algorithm.

For comparison, the peptide abundance ratios were also estimated with the peak area algorithm. To calculate the peak area, selected ion chromatograms were smoothed with a seven-point quadratic Savitsky−Golay filter.[22] The background of a selected ion chromatogram was set to be the straight line connecting the two ion intensities at the peak boundaries. The peak area is the total of the background-subtracted intensities of a peak.

RESULTS AND DISCUSSION

In this study, standard mixtures of isotopically labeled proteomes were used to test the proposed algorithms. Abundance ratios between the light and heavy isotopologues of all peptides in the standard mixtures were expected to be approximately the same as the mixing ratio of the ^{14}N proteome and the ^{15}N proteome. This does assume that the protein abundance profiles should be the same for the two proteomes extracted from the cells grown identically. Six data sets were acquired from standard mixtures, including 1:1a, 1:1b, 5:1, 1:5, 10:1, and 1:10. The 1:1a and 1:1b data sets are from duplicate measurements of the 1:1 standard mixture. Results of peptide quantification are included as Supporting Information Tables S1−S6.

Ion Chromatogram Extraction with Reorganized Peptide Identifications. Both the light and heavy isotopologues of peptides were considered when searching the MS/MS scans (Table 1). We consider every chromatographic peak as an independent measurement of the peptide abundance ratio. If a peptide is identified in multiple chromatographic peaks, all identifications are retained and used for extracting the selected ion chromatograms. However, if a peptide is identified in the same charge state at multiple retention time points across a single chromatographic peak, the different identifications are combined and used to extract a single selected ion chromatogram pair.

The m/z windows for extracting ion chromatograms were calculated from the isotopic distributions of a peptide's isotopologues. The heavy isotopologues had a theoretical isotopic distribution skewed by the incomplete enrichment of the heavy stable isotope. Our ion chromatogram extraction algorithm has the capability of handling mass spectral data of varying resolution. Normally, ion chromatograms are extracted from a single m/z window for an ion species. To allow high-resolution ion chromatogram extraction, an m/z window is opened for each major isotope in the isotopic distribution. The width of the m/z windows can be configured to fit the measurement resolution of the mass spectrometer. In this way, the background noise between two isotopes can be notched out, if a high-resolution mass spectrometer is used. In this study, as a linear ion trap instrument was used, the mass tolerance was set to be ±0.5 Da and the m/z windows for individual isotopes were merged into one m/z window per isotopologue. Our algorithm can also be configured to extract ion chromatograms for other isotope labeling techniques, such as SILAC, ICAT, and $H_2{}^{18}O$ digestion.

Chromatographic Peak Detection with Parallel Paired Covariance. The selected ion chromatograms of a peptide were extracted for a user-defined retention time window around the MS/MS scans of the peptide. Then, the exact retention time boundaries of the two isotopologues' chromatographic peaks were determined by performing peak detection in the covariance chromatogram. The covariance chromatogram was constructed by combining the two selected ion chromatograms (Figure 2). The two coeluting peaks in selected ion chromatograms are represented by one greatly enhanced peak in the covariance chromatogram, whereas the noise peaks appearing in only one selected ion chromatogram are suppressed in the covariance chromatogram (Figure 2). This indicates that the parallel paired covariance algorithm multiplies the signal of the two chromatographic peaks and effectively reduces the uncorrelated noise in

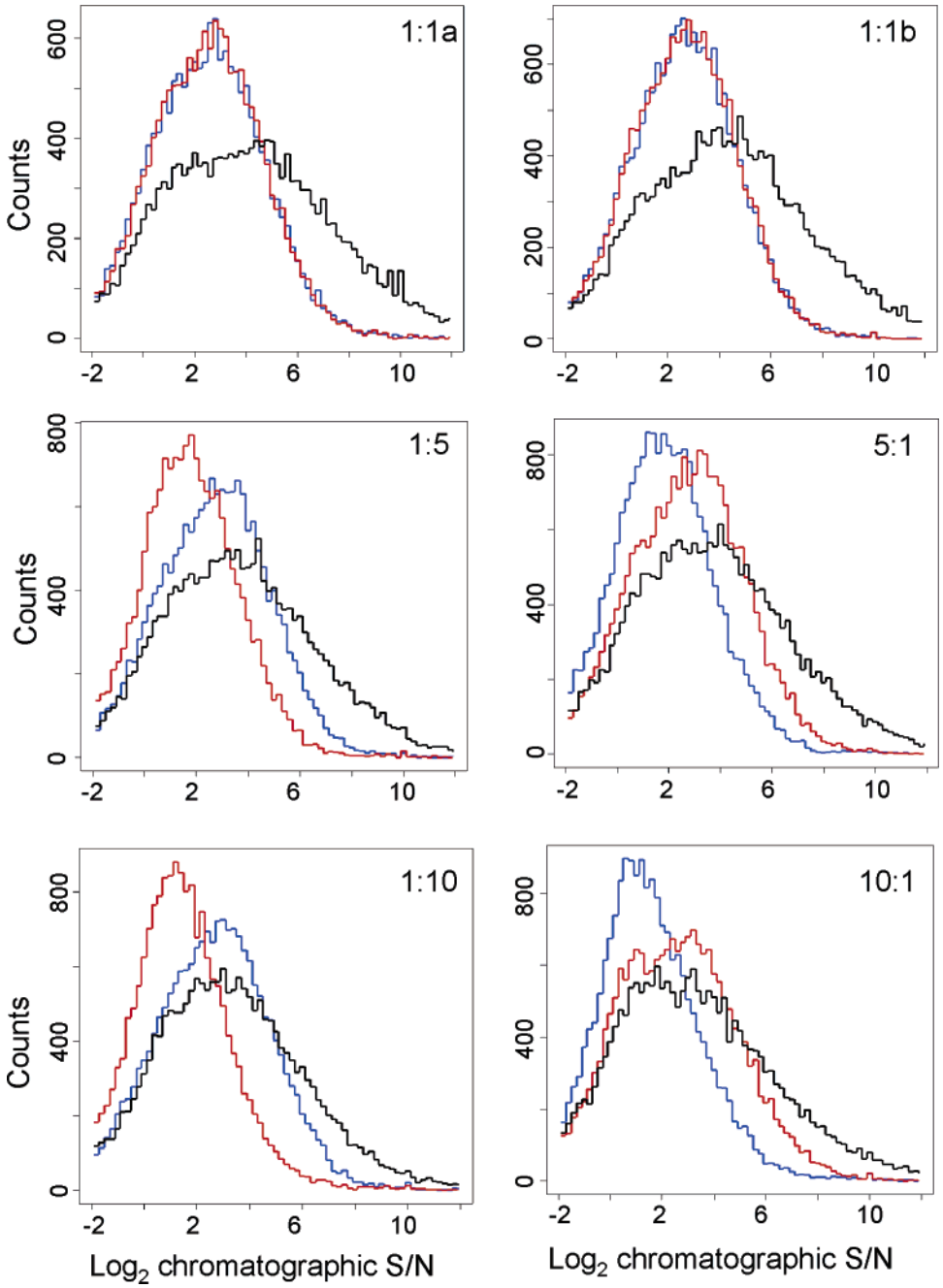

Figure 4. Distribution of $\log_2$ chromatographic S/N for the six standard mixture data sets. Histograms of $\log_2$ chromatographic S/N are shown for the two selected ion chromatograms (blue for the light isotopologue and red for the heavy isotopologue) and their covariance chromatogram (black). The covariance chromatogram has a higher average signal-to-noise ratio than either of the two selected ion chromatograms. The mixing ratio between the ^{14}N proteome and the ^{15}N proteome for a standard mixture is shown in each histogram.

the selected ion chromatograms. As a result, the signal-to-noise ratio of the covariance chromatogram is greater than either one of the selected ion chromatograms, and the peak representing the elution of the isotopologue pair can be detected with greater accuracy.

A practical advantage of using the parallel paired covariance algorithm is obviation of the need for peak detection in *both* selected ion chromatograms. In Figure 2, the peak detection for the light isotopologues (in the red chromatogram) is very difficult due to the low peak height and high noise fluctuation. Peak detection was found to be virtually impossible for the less abundant isotopologue of many peptides in the 1:10 and 10:1 standard mixtures. An alternative method is to determine peak boundaries only in the selected ion chromatogram of the more abundant isotopologue, but this method ignores the signal of the other isotopologue and requires knowledge of which isotopologue is more abundant prior to estimating the abundance ratio between the two isotopologues.

As the accuracy of the peak detection is directly related to the signal-to-noise ratio of the chromatogram, we constructed the histograms of the signal-to-noise ratio for the covariance chromatograms (the black histogram) and the selected ion chromatograms (the blue and red histograms) (Figure 4). Virtually all peptides have an improved signal-to-noise ratio in the covariance

chromatogram as compared to in the selected ion chromatogram of the more abundant isotopologue. The degree of improvement is related to the signal-to-noise ratio of the less abundant isotopologue. In accord with general chromatography protocols, a signal-to-noise ratio of 3 or greater is generally required for a chromatographic peak to be accurately defined in its chromatogram. More peptides exceed this signal-to-noise ratio threshold with their covariance chromatogram than with their individual selected ion chromatogram; therefore, more peptides can have correctly assigned peak boundaries by using the parallel paired covariance algorithm.

The parallel paired covariance algorithm is based on the assumption of the coelution of the two isotopologues. While this is generally true for most peptides labeled with ^{13}C, ^{18}O, and ^{15}N, the peptides labeled with ^{2}H can show a retention time shift between the two isotopologues. The retention time shift can be computationally offset by shifting one selected ion chromatogram relative to the other one. However, this extra computational step can add additional probable error to the peptide quantification process.

Evaluation of the Peptide Abundance Ratio Estimation Accuracy. The peptide abundance ratios were then estimated by principal component analysis of the peak profile (Figure 3). As the isotopologues begin eluting off the column, the trace of the data points starts from close to the origin and moves upward and to the right. After the two isotopologues reach the top of their chromatographic peaks, the trace of the data points regresses back toward the origin. Ideally, the trace of the data points should form a straight line whose slope is the peptide abundance ratio. However, the random noise component of the ion intensities will make the trace "wiggle" along the straight line. Essentially, the purpose of principal component analysis is to separate the noise component and the signal component of the ion intensities.

The accuracy of estimating peptide abundance ratios was benchmarked with the six standard mixture data sets (Supporting Information Tables S1−S6). The principal component analysis algorithm was compared with a more commonly used algorithm based on peak area calculation. On average, in each data set, the selected ion chromatograms were extracted for ~20 000 peptides, and after filtering the peptides with a profile-S/N cutoff of 2.0, the total number of quantified peptides was ~13 000 (Table 1). The filtering effectively removes the peptides that cannot be reliably quantified by either algorithm. This profile-S/N cutoff is discussed in the next section.

The estimated abundance ratios were transformed to logarithm base-2, abbreviated as log-ratio, and histograms of the log-ratios estimated with the principal component analysis algorithm and the peak area ratio algorithm were constructed (Figure 5). Although all peptides in a standard mixture should have the same abundance ratio as the mixing ratio, the log-ratio distributions spread around a center, which can be attributed to the random error of the estimation with both methods and, probably to a lesser extent, to the biological variability of the ^{14}N and ^{15}N cultures. For the four standard mixtures of uneven mixing ratios, their log-ratio distributions have a center slightly shifted toward zero from the log mixing ratio and have a heavier shoulder in the side toward zero. This suggests a systematic error of the estimation with both methods. However, the log-ratio distributions from the principal component analysis method (the blue histogram) are located closer to the log mixing ratio with less spread around the center and a lighter shoulder in the side toward zero, which indicates the less random and systematic errors in the log-ratio estimation with the principal component analysis algorithm.

Two features of the principal component analysis algorithm, which are also shared with the correlation algorithm, can contribute to the improved peptide abundance ratio estimation.[11,24,25] The first feature is the obviation of background subtraction by assuming a constant background within the chromatographic peak. In the peak area method, however, the backgrounds of the two chromatograms have to be subtracted from their peak area. Since the automatic routine for background estimation can be error prone for many peptides, the errors in the background estimation translate directly into errors of abundance ratio estimation. The second feature is a built-in mechanism for removing random noise. The signal component and the random noise component are separated into the first principal component and the second principal component, respectively, and the first principal component is used to estimate the abundance ratio. The peak area method can only rely on chromatogram smoothing to remove the random noise. As the selection of a routine and its parameters for chromatogram smoothing is fairly subjective, it would be difficult to obtain optimized chromatogram smoothing across thousands of selected ion chromatograms with varying peak shapes.

Compared with the peak area method, a disadvantage of the peak profile-based algorithms is their limited applicability to the isotopologue pairs with retention time shift. Although the retention time shift can be offset computationally, the offset can be incorrectly estimated, which might lead to the error in the abundance ratio estimation.

Scoring of Peptide Abundance Ratios for Estimation Variability and Bias. The principal component analysis algorithm scores each estimated peptide abundance ratio with a profile-S/N. Note that profile S/N is a signal-to-noise ratio measure of the peak profile, which is different from chromatographic S/N. Chromatographic S/N directly impacts peak area calculation accuracy, and therefore, estimation accuracy of peak area ratio for a peptide should be related to the two chromatographic S/Ns for the light and heavy isotopologues. On the other hand, the principal component analysis algorithm estimates peptide abundance ratios from peak profiles, and therefore, its estimation variability and bias are expected to be directly related to profile-S/N.

The peptides were separated into bins by the logarithm base-2 of their profile-S/N (log-profile-S/N). The bins were evenly spaced by 0.1 units between log-profile-S/Ns of 0 and 6. The log-ratio distributions were constructed in all log-profile-S/N bins. To present the series of the log-ratio distributions, a two-dimensional heat map histogram was plotted for each standard mixture data set (Figure 6). Each horizontal band of the two-dimensional histogram represents a log-ratio distribution, in which the peptide count is color-coded. The histograms show that the log-ratio distribution changes in a consistent manner with the log-profile-S/N. At high log-profile-S/N region, the estimated log-ratios tightly

(24) Thorne, G. C. G.; Simon J. *Biomed. Environ. Mass Spectrom.* **1986**, *13* (11), 605−609.
(25) Thorne, G. C. G.; Simon, J.; Payne, Peter A. *Biomed. Mass Spectrom.* **1984**, *11* (8), 415−420.

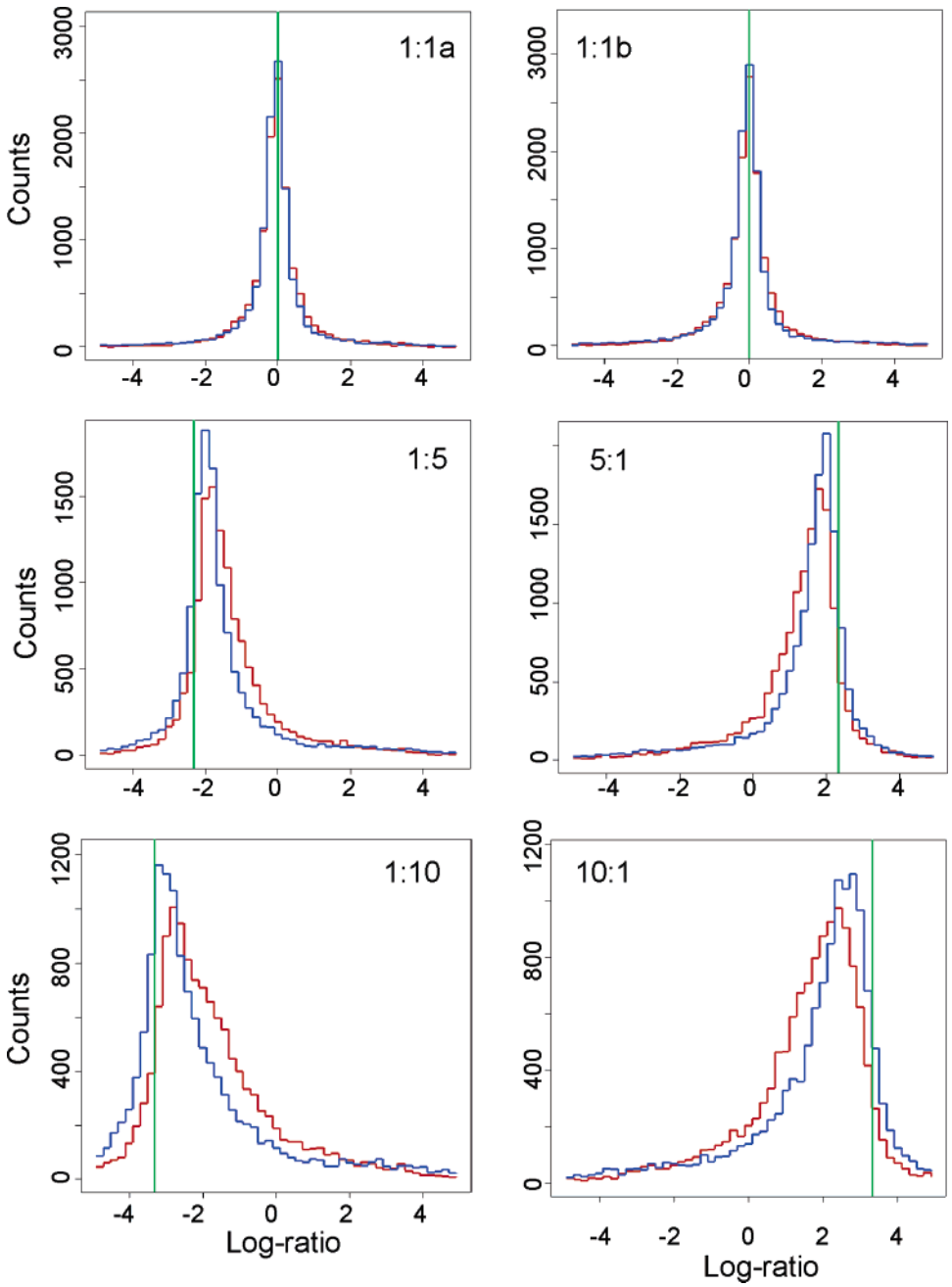

Figure 5. Distribution of peptide log-ratio estimates for the six standard mixture data sets. The histograms of the log-ratios estimated with the principal component analysis algorithm (blue histograms) and the peak area algorithm (red histograms) are shown for all standard mixture data sets. Only the peptides with profile-S/Ns greater than two are considered. The $\log_2$ mixing ratio is marked with a green vertical line in each histogram.

cluster close to the $\log_2$ mixing ratios (the dashed red lines in Figure 6), indicating accurate and precise log-ratio estimation for peptides with high log-profile-S/N. As the log-profile-S/N decreases, the spread of the log-ratio distribution increases, which suggests the elevating variability of log-ratio estimation.

When the log-profile-S/N decreases below a threshold, the log-ratio distribution gradually regresses away from the log mixing ratio and approaches the log-ratio of zero. This shows the higher bias in the log-ratio estimation for peptides with lower log-profile-S/N. Also note that the log-profile-S/N threshold for the onset of log-ratio estimation bias is higher in the standard mixtures with larger $\log_2$ mixing ratios. This supports the previous observation that the abundance ratio estimation is often biased for low-concentration peptides with large abundance difference between their isotopologues.[26] The less abundant isotopologue often receives a higher percentage of ion intensity from the noise than the more abundant isotopologue. As a result, the abundance ratio estimate becomes biased toward 1:1. For peptides with log-profile-S/Ns close to zero (i.e., both isotopologues are "buried" in the background noise), the abundance ratios will most likely be estimated as 1:1, regardless of their true value.

The variability of the log-ratio estimation can be measured with the standard deviation of the log-ratio distribution. The standard

(26) Ong, S. E.; Kratchmarova, I.; Mann, M. *J. Proteome Res.* **2003**, *2* (2), 173–181.

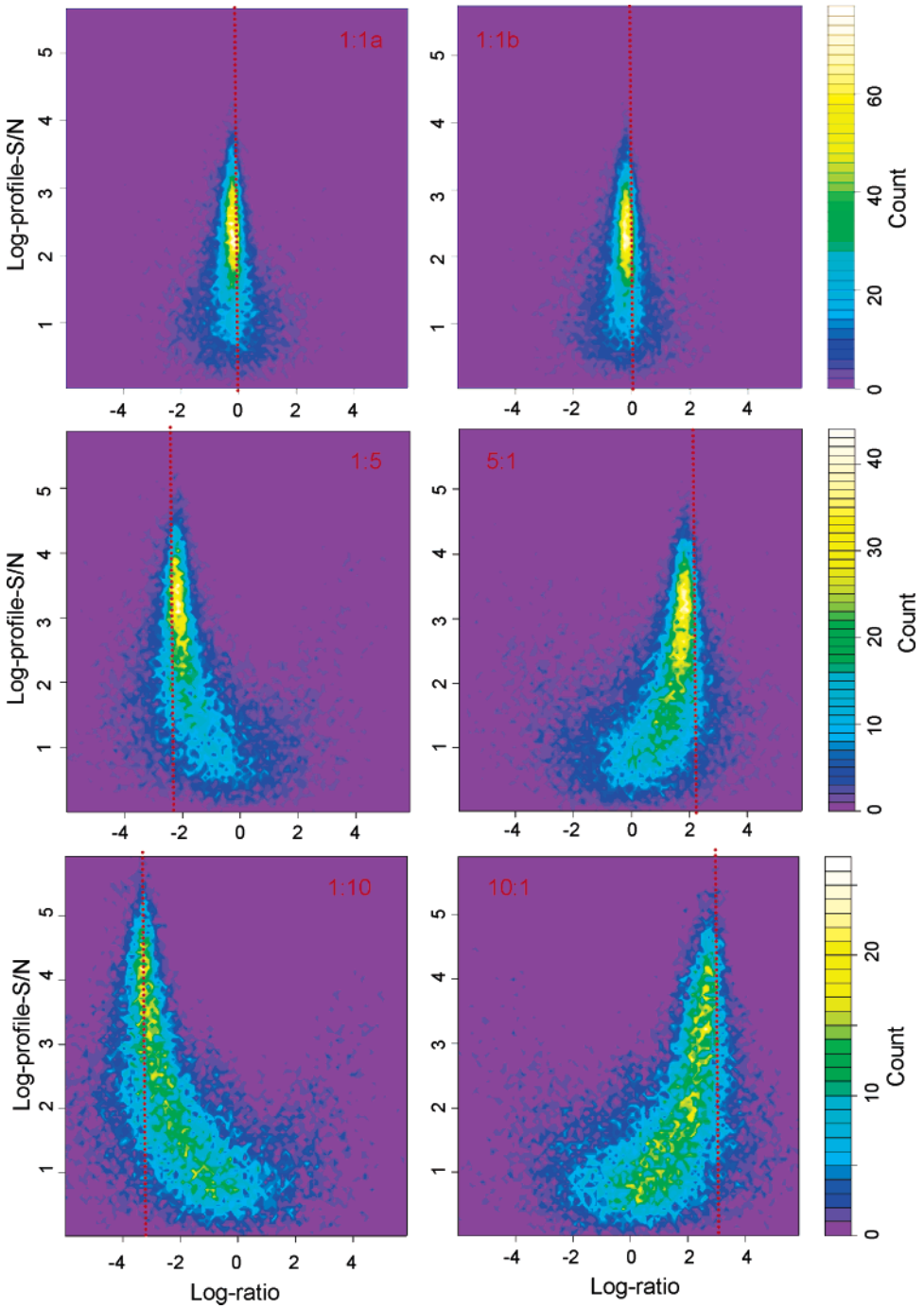

Figure 6. Two-dimensional heat map histograms of log-ratio and log-profile-S/N for the six standard mixture data sets. The color scales for the frequency of peptides are shown in the right of the histograms. The $\log_2$ mixing ratios are marked with the red dotted lines. The two-dimensional heat map histograms show the log-ratio distributions at different log-profile-S/N levels. Log-profile-S/N is inversely related to the spread of the log-ratio distribution and its deviation from the $\log_2$ mixing ratio.

deviations were plotted against log-profile-S/N (Figure 7A). There is an apparent inverse linear correlation between the standard deviation and the log-profile-S/N. A linear regression model was constructed between the log-profile-S/N and the standard deviation:

$$\sigma = 1.2 - 0.2V$$

where σ is the standard deviation and V is the log-profile-S/N. The coefficient of determination of the linear regression model was 0.766. The majority of residuals in the linear regression model arise from the stratification of the data points by the mixing ratios.

The log-ratio estimation bias can be quantified with the absolute values of the averages of the log-ratio distributions, which were plotted against the log-profile-S/N (Figure 7B). At the high log-profile-S/N range, the absolute averages are close to, but slightly below, the absolute value of the $\log_2$ mixing ratios of the data sets. As the log-profile-S/N decreases, the average regresses back to zero. A zero-intercept straight line was fit into the data points on the track of regressing to zero, and in conjunction with

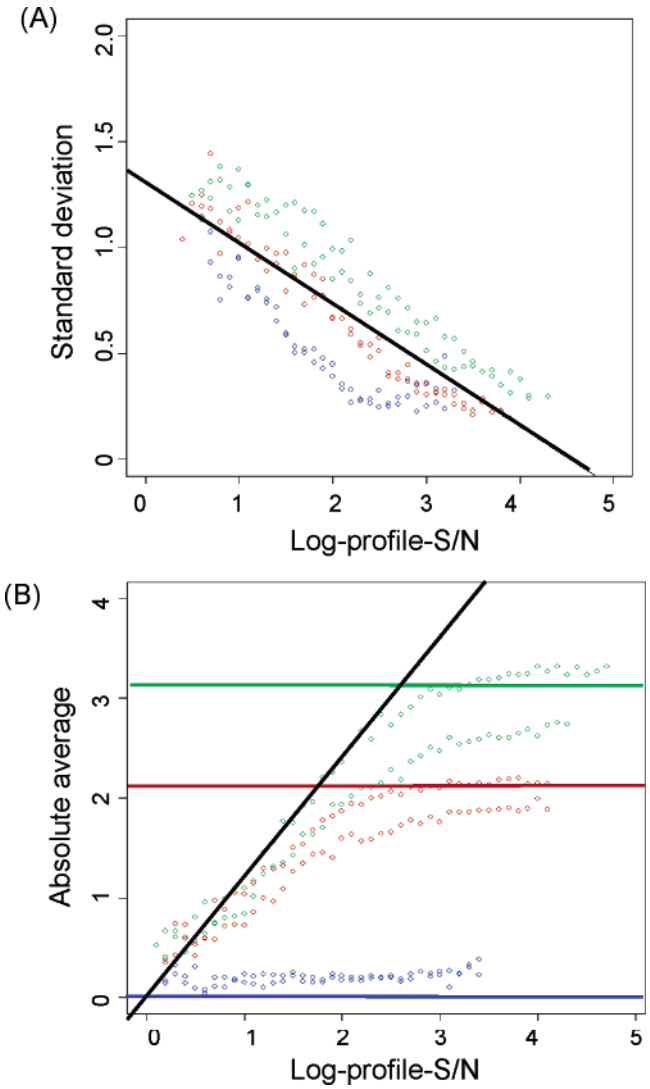

Figure 7. Linear models for the standard deviation and absolute average of the log-ratio distribution. Parts A and B show the standard deviations and absolute average of the log-ratio distributions at different log-profile-S/N levels, respectively. The data points from the two 1:1 data sets are shown in blue, those from the 5:1 and 1:5 data sets in red, and those from the 10:1 and 1:10 data sets in green. In part A, the standard deviations are modeled with a linear model of log-profile-S/N (the black straight line). In part B, the biased absolute averages are also modeled with a linear model of log-profile-S/N (the black straight line). At the high log-profile-S/N region, the absolute averages are largely consistent with the $\log_2$ mixing ratios (marked with the horizontal lines).

the largely unbiased average at the high log-profile-S/N region, a linear regression model can be obtained:

$$|\mu| = \begin{cases} 1.2 \cdot V, & 1.2 \cdot V < |H| \\ |H|, & 1.2 \cdot V \geq |H| \end{cases}$$

where $|\mu|$ is the absolute value of the average, V is the log-profile-S/N, and $|H|$ is the absolute value of the log mixing ratio or the true log-ratio. The linear regression model of the average shows that, when log-profile-S/N is zero, the average is also zero; as the log-ratio increases, the average increases proportionally until it reaches the true log-ratio and levels off afterward.

The bias and the variability of the abundance ratio estimation are indispensable for making the statistical inference on the estimated abundance ratio. In isotope dilution mass spectrometry, the variability and bias of the abundance ratio estimation are determined experimentally by replicate measurements of the sample and calibration with standard solutions.[27] These two experimental routines are of limited practicality in quantitative proteomics, where thousands of peptides are quantified in each experiment. These two linear regression models shown in Figure 7 enable the prediction of the variability and the bias of the abundance ratio estimation from the profile-S/N for quantitative proteomic experiments. This paves the way for statistically evaluating every peptide abundance ratio using its predicted estimation variability and bias in protein abundance ratio estimation process.[13]

CONCLUSIONS

We presented two algorithms for peptide quantification in quantitative shotgun proteomics. The parallel paired covariance algorithm was developed to improve the accuracy of assigning peak boundaries of the chromatographic peaks from a peptide. This algorithm integrates the two selected ion chromatograms into one covariance chromatogram. We showed that covariance chromatograms generally have a better signal-to-noise ratio than either selected ion chromatogram and result in better peak detection accuracy. We then used a principal component analysis algorithm to estimate the peptide abundance ratios from peak profiles. The estimation accuracy was shown to be better than when using a peak area algorithm. More importantly, for each peptide abundance ratio estimate, the principal component analysis algorithm provides a signal-to-noise ratio measure of the peak profile, called profile signal-to-noise ratio. The profile signal-to-noise ratio is inversely correlated with the variability and bias of the peptide abundance ratio estimation.

ACKNOWLEDGMENT

The computational and experimental work was supported by the DOE Genomics:GTL grants DOE LAB-04-32 and DOE DE-FG02-01ER63241, respectively, from 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.

AC0606554

(27) Sargent, M.; Harte, R.; Harrington, C. *Guidelines for achieving high accuracy in isotope dilution mass spectrometry (IDMS)*; Royal Society of Chemistry: Cambridge, UK, 2002.