A Correlation Algorithm for the Automated Quantitative Analysis of Shotgun Proteomics Data

Michael J. MacCoss,† Christine C. Wu,† Hongbin Liu, Rovshan Sadygov, and John R. Yates, III*

Department of Cell Biology, The Scripps Research Institute, La Jolla, California 92037

Quantitative shotgun proteomic analyses are facilitated using chemical tags such as ICAT and metabolic labeling strategies with stable isotopes. The rapid high-throughput production of quantitative "shotgun" proteomic data necessitates the development of software to automatically convert mass spectrometry-derived data of peptides into relative protein abundances. We describe a computer program called RelEx, which uses a least-squares regression for the calculation of the peptide ion current ratios from the mass spectrometry-derived ion chromatograms. RelEx is tolerant of poor signal-to-noise data and can automatically discard nonusable chromatograms and outlier ratios. We apply a simple correction for systematic errors that improves the accuracy of the quantitative measurement by 32 ± 4%. Our automated approach was validated using labeled mixtures composed of known molar ratios and demonstrated in a real sample by measuring the effect of osmotic stress on protein expression in *Saccharomyces cerevisiae*.

Biological research is undergoing a paradigm shift from the analysis of individual proteins and genes to global proteome-wide and genome-wide measurements. Traditional proteomic experiments have relied on the separation and visualization of proteins using either 1D or 2D gel electrophoresis; however, limitations have spurred the development of alternative "gel-free" methodologies. These alternative methods usually take a "shotgun" approach where an intact protein mixture is digested to form an even more complex peptide mixture and analyzed directly by multidimensional liquid chromatography/tandem mass spectrometry (μLC/μLC/MS/MS).[1−4] The peptide sequences are identified using software that matches the acquired MS/MS fragmentation spectra against theoretical spectra predicted from amino acid sequences of the same mass within a database.[5] This approach is an extremely effective and routine means of qualitative protein identification in mixtures, largely because the software infrastructure is well established and validated for database searching[5,6] and protein assembly.[7,8] However, the analogous infrastructure for mass spectrometry-based quantitation of protein levels remains to be developed. Here we describe an approach for the automated analysis and validation of quantitative proteomics data. Our method is completely automated, and thousands of peptide ratios can be derived and statistically evaluated in a matter of minutes.

Mass spectrometry has a long history in quantitative measurements and has been used for the quantitation of peptides in complex mixtures since the early 1980s.[9,10] Most quantitative mass spectrometry methods are based on the measurement of relative ion intensity abundances in the mass spectrometer between a compound with natural-abundance stable isotopes (unlabeled isotopomer) and a "spiked" internal standard that is structurally identical with the exception of atoms that are enriched with a "heavy" stable isotope (labeled isotopomer)—usually ^{2}H, ^{13}C, ^{15}N, or ^{18}O. A problem in proteomics is creating a quantitative internal standard for every protein in the cell. Recently, methods have been developed that incorporate stable isotopes into proteins using metabolic labeling,[11−16] enzymatic approaches,[17−19] or chemical derivatization.[20−23] After mixing the unlabeled sample with the labeled sample, the mixture is prepared and analyzed by mass

* To whom correspondence hould be addressed. E-mail: jyates@scripps.edu. Phone: 858-784-8862. Fax: 858-784-8883.

† These authors contributed equally to this work.

(1) McCormack, A. L.; Schieltz, D. M.; Goode, B.; Yang, S.; Barnes, G.; Drubin, D.; Yates, J. R., III. *Anal. Chem.* **1997**, *69*, 767−76.
(2) Link, A. J.; Eng, J. K.; Schieltz, D. M.; Carmack, E.; Mize, G. J.; Morris, D. R.; Garvik, B. M.; Yates, J. R., III. *Nat. Biotechnol.* **1999**, *17*, 676−82.
(3) Washburn, M. P.; Wolters, D.; Yates, J. R., III. *Nat. Biotechnol.* **2001**, *19*, 242−7.
(4) Peng, J.; Elias, J. E.; Thoreen, C. C.; Licklider, L. J.; Gygi, S. P. *J. Proteome Res.* **2003**, *2*, 43−50.
(5) Eng, J. K.; McCormack, A. L.; Yates, J. R., III. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976−89.
(6) Yates, J. R., III; Eng, J. K.; McCormack, A. L.; Schieltz, D. M. *Anal. Chem.* **1995**, *67*, 1426−36.
(7) Tabb, D. L.; Eng, J. K.; Yates, J. R., III. In *Proteome Research: Mass Spectrometry*; James, P., Ed.; Springer: New York, 2001.
(8) Tabb, D. L.; McDonald, W. H.; Yates, J. R., III. *J. Proteome Res.* **2002**, *1*, 21−6.
(9) Desiderio, D. M.; Yamada, S.; Tanzer, F. S.; Horton, J.; Trimble, J. *J. Chromatogr.* **1981**, *217*, 437−52.
(10) Desiderio, D. M.; Kai, M. *Biomed. Mass Spectrom.* **1983**, *10*, 471−9.
(11) Oda, Y.; Huang, K.; Cross, F. R.; Cowburn, D.; Chait, B. T. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 6591−6.
(12) Washburn, M. P.; Ulaszek, R.; Deciu, C.; Schieltz, D. M.; Yates, J. R., III. *Anal. Chem.* **2002**, *74*, 1650−7.
(13) Zhu, H.; Pan, S.; Gu, S.; Bradbury, E. M.; Chen, X. *Rapid Commun. Mass Spectrom.* **2002**, *16*, 2115−23.
(14) Krijgsveld, J.; Ketting, R. F.; Mahmoudi, T.; Johansen, J.; Artal-Sanz, M.; Verrijzer, C. P.; Plasterk, R. H.; Heck, A. J. *Nat. Biotechnol.*, in press.
(15) Foster, L. J.; De Hoog, C. L.; Mann, M. *Proc. Natl. Acad. Sci. U.S.A* **2003**, *100*, 5813−8.
(16) Ong, S. E.; Kratchmarova, I.; Mann, M. *J. Proteome Res.* **2003**, *2*, 173−81.
(17) Yao, X.; Freas, A.; Ramirez, J.; Demirev, P. A.; Fenselau, C. *Anal. Chem.* **2001**, *73*, 2836−42.
(18) Stewart, I. I.; Thomson, T.; Figeys, D. *Rapid Commun. Mass Spectrom.* **2001**, *15*, 2456−65.
(19) Reynolds, K. J.; Yao, X.; Fenselau, C. *J. Proteome Res.* **2002**, *1*, 27−33.

10.1021/ac034790h CCC: $27.50
Published on Web 00/00/0000

spectrometry. Because a peptide containing a "heavy" stable isotope should undergo the same sample preparation and measurement biases as the unlabeled peptide, changes in protein level between samples can be evaluated using the relative peptide responses in the mass spectrometer after minimizing systematic errors.[24,25]

The most error-prone step in determining the unlabeled and enriched isotopomer ratios is the assessment of the respective mass spectrometer ion intensity ratios. This ion intensity ratio is normally calculated using the areas under the ion chromatograms of the unlabeled and enriched isotopomer m/z (or m/z range). To determine the area under any peak, an integration routine must (i) decide where the peak starts and ends and (ii) decide the contribution from background on which the peak is superimposed. Automated assessment of these parameters is dependent on the chromatographic peak shape—making the process for an individual ion chromatogram highly subjective. These errors are particularly problematic in a proteome analysis because the peak shape will differ substantially between the thousands of peptide sequences identified in a μLC/μLC/MS/MS analysis, and the conditions cannot be optimized for every individual analyte. The estimate of peak onset and background requires that the peaks be well defined. However, under experimental conditions, the chromatograms may be crowded and an algorithm may not have a sufficiently large region of pure baseline to appropriately subtract the true background. Furthermore, although the desired output is the background-subtracted ion current ratio, because each of the two ion chromatograms are detected and integrated separately, the precision and accuracy of the ratio will ultimately be limited by the software's ability to handle the ion chromatogram with the poorest S/N.

Because of the success of correlation routines for qualitative identification of peptides from tandem mass spectral data of all quality,[5,26–28] we have adapted a least-squares correlation[29–31] to calculate the background-subtracted intensity ratio between two ion chromatograms in a single step. The quality of the ratio measurement is independent of the chromatographic peak shape and, thus, unlike some peak integration algorithms does not require the definition of a peak shape function.[32] Furthermore, because both ion chromatograms are handled simultaneously, the peak detection only needs to be performed on the ion chromatogram of the most intense isotopomer. In this report, we demonstrate that our approach minimizes errors resulting from mass spectrometry ion current ratio measurements when used for large-scale proteomic studies and demonstrate the application of this algorithm with the measurement of changes in relative protein levels in *Saccharomyces cerevisiae* in response to osmotic stress.

EXPERIMENTAL SECTION

Materials. An Isotope Coded Affinity Tag Kit and Porozyme bulk immobilized trypsin were obtained from Applied Biosystems (Foster City, CA). HPLC-grade solvents were purchased from J.T. Baker (Phillipsburg, NJ). Monomeric avidin affinity columns were obtained from Pierce (Rockford, IL). Endoproteinase Lys-C was purchased from Roche Diagnostics (Indianapolis, IN). All remaining laboratory reagents were purchased from Sigma-Aldrich (St. Louis, MO) unless noted otherwise.

Preparation of Yeast Samples. *S. cerevisiae* (strain S288C) were cultured in minimal growth media (YNB/5% glucose/0.5% $(NH_4)_2SO_4$) and grown to log-phase growth (0.5−1.0 OD_{600}/mL) at 30 °C prior to use in each of the following experiments. For ^{15}N-labeled yeast, cells were cultured in minimal growth media using $(^{15}NH_4)_2SO_4$ (>98 atom percent excess, ape) instead of $(NH_4)_2SO_4$. Unless noted otherwise, all yeast samples were lysed and proteins digested to peptides as described previously.[3,33]

(1) For Samples with Known Unlabeled/^{15}N-Labeled Protein Ratios. Yeast cultured in unlabeled minimal growth media were mixed with yeast cultured in ^{15}N-enriched media in approximate ratios of 1:4, 2:4, 3:4, and 4:4 as determined by OD_{600}/mL. The mixture of yeast cells was collected by pelleting.

(2) For Partially Labeled Yeast Samples. Yeast were inoculated into minimal growth media (YNB/5% glucose/0.5% $(NH_4)_2SO_4$) where the total $(NH_4)_2SO_4$ in the media was replaced with 70% $(^{15}NH_4)_2SO_4$/30% $(NH_4)_2SO_4$, 80% $(^{15}NH_4)_2SO_4$/20% $(NH_4)_2SO_4$, or 90% $(^{15}NH_4)_2SO_4$/10% $(NH_4)_2SO_4$ and cultured to log-phase growth. Yeast cells from partially enriched media were collected by pelleting.

(3) For Analysis of Osmotically Stressed Yeast Samples. Yeast were cultured separately in minimal growth media to mid-log phase (0.5 OD_{600}/mL). A sample was collected for the 0-min time point. The media was then exchanged by pelleting the cells and resuspending them in an equal volume of YNB/5% glucose/0.5% $(NH_4)_2SO_4$/5% NaCl prewarmed at 30 °C. Cells were incubated for 40 min, and a sample was collected for the 0- and 40-min time points. Both the 0- and 40-min samples were mixed with yeast cultured in ^{15}N-enriched media at a 1:1 ratio as determined by OD_{600}/mL. The mixed yeast cells were collected by pelleting.

(4) Sample Preparation Using Isotope Coded Affinity Tags (ICAT). Yeast (strain BJ5460) were grown in YPD media and lysed according to the protocol described by Washburn et al.[3] Two equal aliquots of yeast soluble proteins (500 μg) were reduced and labeled with unlabeled and $^2H_8$-enriched ICAT reagents, respectively, according to the manufacturer's instructions. After the labeling reaction was complete, the two differential labeled protein samples were combined and digested serially first with endoproteinase Lys-C and then by immobilized trypsin as de-

(20) Gygi, S. P.; Rist, B.; Gerber, S. A.; Turecek, F.; Gelb, M. H.; Aebersold, R. *Nat. Biotechnol.* **2000**, *17*, 994−9.
(21) Chakraborty, A.; Regnier, F. E. *J. Chromatogr., A* **2002**, *949*, 173−84.
(22) Griffin, T. J.; Lock, C. M.; Li, X. J.; Patel, A.; Chervetsova, I.; Lee, H.; Wright, M. E.; Ranish, J. A.; Chen, S. S.; Aebersold, R. *Anal. Chem.* **2003**, *75*, 867−74.
(23) Griffin, T. J.; Gygi, S. P.; Ideker, T.; Rist, B.; Eng, J.; Hood, L.; Aebersold, R. *Mol. Cell Proteomics* **2002**, *1*, 323−33.
(24) Matthews, D. E.; Hayes, J. M. *Anal. Chem.* **1976**, *48*, 1375−82.
(25) Colby, B. N.; McCaman, M. W. *Biomed. Mass Spectrom.* **1979**, *6*, 225−30.
(26) Sadygov, R. G.; Eng, J. K.; Durr, E.; Saraf, A.; McDonald, W. H.; MacCoss, M. J.; Yates, J. R., III. *J. Proteome Res.* **2003**, *1*, 211−5.
(27) Yates, J. R., III; Morgan S. F.; Gatlin, C. L.; Griffin, P. R.; Eng, J. K. *Anal. Chem.* **1998**, *70*, 3557−67.
(28) MacCoss, M. J.; Wu, C. C.; Yates, J. R., III. *Anal. Chem.* **2002**, *74*, 5593−9.
(29) Thorne, G. C.; Gaskell, S. J. *Biomed. Environ. Mass Spectrom.* **1986**, *13*, 605−9.
(30) Thorne, G. C.; Gaskell, S. J.; Payne, P. A. *Biomed. Mass Spectrom.* **1984**, *11*, 415−20.
(31) Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical Recipes in C*, 2nd ed.; Cambridge University Press: New York, 1992.
(32) Goodman, K. J.; Brenna, J. T. *Anal. Chem.* **1994**, *66*, 1294−301.
(33) Wu, C. C.; MacCoss, M. J.; Howell, K. E.; Yates, J. R. *Nat. Biotechnol.* **2003**, *21*, 532−8.

scribed.[3] The ICAT reagent-labeled peptides were enriched by affinity purification using a monomeric avidin column as recommended by the manufacturer (Pierce).

Multidimensional Protein Identification Technology (MudPIT). The protein digests were loaded directly onto a fused-silica capillary column (a 100-μm i.d.) packed with 7 cm of 5-μm Aqua C18 material (Phenomenex, Ventura, CA) at the tip, 3 cm of 5-μm Partisphere strong cation exchanger (Whatman, Clifton, NJ), an additional 3 cm of 5-μm Aqua C18 material as described previously.[28] After loading the peptide digests, the column was placed in-line with an Surveyor quaternary HPLC (ThermoFinnigan, San Jose, CA) and analyzed using a 12-step separation described previously.[33] As peptides eluted from the microcapillary column, they were electrosprayed directly into an LCQ-Deca mass spectrometer (ThermoFinnigan) with the application of a distal 2-kV spray voltage. A cycle of one full-scan mass spectrum (400−1400 m/z) followed by three data-dependent MS/MS spectra at a 35% normalized collision energy was repeated continuously throughout each step of the multidimensional separation. The application of all mass spectrometer scan functions and HPLC solvent gradients were controlled by the Xcaliber data system.

The acquired tandem mass spectra were searched against a fasta database containing the yeast open reading frames and the human RefSeq protein sequences using a parallelized implementation of SEQUEST-NORM[28] with no enzyme specificity selected in the parameter file. The program DTASelect[8] was used to filter the peptide identifications and assemble the peptides into proteins. Peptide identifications matching human protein sequences, normalized Xcorr < 0.3, and ΔCn < 0.1 were omitted from the final output list.

Ion Chromatogram Extraction. For each peptide exceeding the DTASelect criteria, ion chromatograms were extracted from the Xcaliber data file for the m/z range surrounding the unlabeled and ^{15}N-enriched peptide isotope distributions. The program EXTRACT-CHRO was written in the C programming language and compiled using the GNU GCC compiler under the Linux operating system. Each identified peptide sequence was obtained by parsing the DTASelect-filter.txt output file. These peptide sequences were used to calculate the elemental composition and predict the isotope distribution[34] for each unlabeled and ^{15}N-enriched peptide sequence—taking into account the ^{15}N atomic enrichment of the precursor and the resolution of the mass analyzer. Using this information, chromatograms were extracted for 100 MS scans (precursor scans) surrounding the MS/MS spectrum that identified the peptide for the m/z range of the predicted unlabeled and ^{15}N-enriched peptide isotope distributions. The respective ion chromatograms were stored in a tab-delimited file with a .chr extension. Additional options within EXTRACT-CHRO exist for extracting ion chromatograms from other quantitative proteomics labeling techniques including, but not limited to, ICAT,[20] SILAC,[35] derivatization with acetyl *N*-hydroxysuccinimide,[36] and methyl esterification.[37,38]

Calculation of Ion Current Ratios and Estimation of Protein Ratios. Each pair of ion chromatograms extracted from the Xcaliber data file was analyzed using a computer program called RelEx (*Rel*ative *Ex*pression). RelEx performed all aspects of the quantitative peak detection, peptide ratio calculations, and estimation of the protein ratio. The program was written in Visual Basic 6.0 and incorporates a graphical user interface to facilitate the manual validation of all aspects of the ion current ratio calculation. The software is available from the authors for individual use and evaluation through an Institutional Software Transfer Agreement (see http://fields.scripps.edu/relex for details).

RelEx performs a series of five sequential steps on each chr file written by EXTRACT-CHRO within a selected directory. (1) Both the unlabeled and labeled isotopomer chromatograms are smoothed using a modified seven-point quadratic Savitsky−Golay filter.[31,39,40] (2) A peak detection algorithm identifies peaks within the ion chromatogram exceeding a predetermined threshold. (3) The peak with a retention time nearest to the tandem mass spectrum that identified the peptide using SEQUEST is chosen for the calculation of the isotopomer ratio. (4) A linear least-squares correlation with errors in both coordinates[41,42] is used to calculate the slope between the data points of the unlabeled and labeled ion chromatograms. (5) The slope of the regression analysis or simply the background-subtracted ion current ratio is stored along with the peptide sequence and protein locus pending the correlation coefficient (r) exceeding a threshold defined in the parameter file.

After calculating the peptide ion current ratio, these data are used to estimate the relative protein level. First, the peptide ratios are sorted by protein locus and outliers are omitted using a Dixon's Q-test.[43] The remaining peptide ratios are used to estimate the protein mean and standard deviation. Differences between mean protein ratios were established using t-tests. All output is sorted by statistical significance and displayed as a PNG image using the program matrix2png.[44]

RESULTS AND DISCUSSION

An overview of the described quantitative approach is shown in Figure 1. A sample is metabolically labeled with ^{15}N-enriched media to use as an internal standard for all subsequent relative abundance measurements. This internal standard is then added in equal amounts to any number of samples (different time points, conditions, etc.) for relative quantitative measurements. The internal standard is added at the beginning of the sample preparation to ensure that the major errors in the analysis of the sample and the internal standard covary and do not affect the accuracy of the ratio measurement. The samples are analyzed by MudPIT, where the peptide sequence is qualitatively identified from the MS/MS spectrum and the peptide ion current ratio is calculated using ion chromatograms extracted from the precursor

(34) Kubinyi, H. *Anal. Chim. Acta* **1991**, *247*, 107−19.
(35) Ong, S. E.; Blagoev, B.; Kratchmarova, I.; Kristensen, D. B.; Steen, H.; Pandey, A.; Mann, M. *Mol. Cell Proteomics* **2002**, *1*, 376−86.
(36) Regnier, F. E.; Riggs, L.; Zhang, R.; Xiong, L.; Liu, P.; Chakraborty, A.; Seeley, E.; Sioma, C.; Thompson, R. A. *J. Mass Spectrom.* **2002**, *37*, 133−45.
(37) Ficarro, S.; Chertihin, O.; Westbrook, V. A.; White, F.; Jayes, F.; Kalab, P.; Marto, J. A.; Shabanowitz, J.; Herr, J. C.; Hunt, D. F.; Visconti, P. E. *J. Biol. Chem.* **2003**, *278*, 11579−89.
(38) Goodlett, D. R.; Keller, A.; Watts, J. D.; Newitt, R.; Yi, E. C.; Purvine, S.; Eng, J. K.; Haller, Pv; Aebersold, R.; Kolker, E. *Rapid Commun. Mass Spectrom.* **2001**, *15*, 1214−21.
(39) Gorry, P. A. *Anal. Chem.* **1990**, *62*, 570−3.
(40) Savitzky, A.; Golay, M. J. E. *Anal. Chem.* **1964**, *36*, 1627−39.
(41) Reed, B. C. *Am. J. Phys.* **1989**, *57*, 642−6.
(42) Reed, B. C. *Am. J. Phys.* **1990**, *58*, 189.
(43) Rorabacher, D. B. *Anal. Chem.* **1991**, *63*, 139−46.
(44) Pavlidis, P.; Noble, W. S. *Bioinformatics* **2003**, *19*, 295−6.

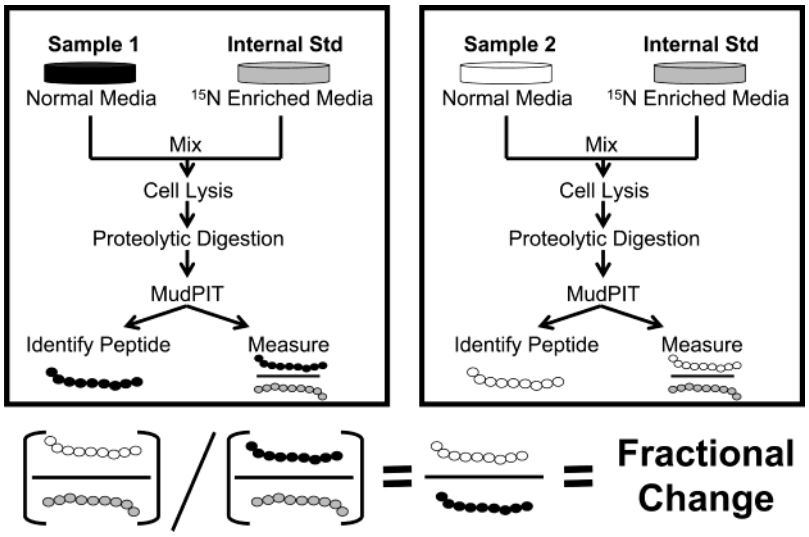

Figure 1. General scheme for quantitative proteomics using metabolic labeling. Cells grown in media enriched in ^{15}N are used as an internal standard for all quantitative measurements. This internal standard is mixed with cells from different conditions as early during the sample preparation as possible so that any protein losses during the cell lysis, digestion, and measurement will be accounted for by their respective ^{15}N-labeled protein. Changes in protein level are expressed relative to another sample to minimize systematic errors.

scans. The ion current ratios of two analyses are then used to estimate the relative change in protein level. Because both analyses contain the same amount of labeled internal standard, the internal standard intensities cancel along with any systematic errors, resulting in the fractional change between the two samples.

Ion Chromatogram Extraction, Peak Detection, and Ion Current Ratio Calculation. The first step in handling quantitative data from a MudPIT experiment is the extraction of the ion chromatograms for the m/z range surrounding the unlabeled and labeled isotopomers. The ion chromatogram extraction program, EXTRACT-CHRO, uses the peptide sequence derived from the SEQUEST database search to calculate the elemental composition and predict the isotope distribution (Figure 2A−D). The m/z range surrounding the predicted isotope distribution is then used for the extraction of mass chromatograms. After extraction of the ion chromatograms, the program RelEx applies a Savitsky−Golay filter and detects the scan region encompassing the most intense isotopomer (Figure 2E). Knowing the approximate location of the eluting peak, RelEx then correlates this region of the two chromatograms against each other using a least-squares regression (Figure 2F). Thus, in a single calculation, the slope of the regression provides an accurate measure of the background-subtracted ion intensity ratio,[29,30] the intercept represents the ratio of the two backgrounds, and the correlation coefficient indicates the quality of the match between the two elution profiles.

Panels A−C of Figure 3 illustrate RelEx's ion current ratio measurement for selected peptides from yeast cells with a median unlabeled to ^{15}N-enriched ratio of 2.4:1. In the left chromatogram window, the blue profile is from the unlabeled peptide m/z and the red profile is the ^{15}N-enriched peptide m/z. The right correlation window displays the least-squares regression of the two chromatograms from within the yellow area of the chromatogram window. RelEx is unaffected by the peak shape and can efficiently obtain a slope (ion current ratio) from isotopomer chromatogram pairs with signal greater than background. Unlike any other program for the calculation of peak area ratios from mass spectrometry data, RelEx produces a quantitative measure of the ratio quality. Because both the unlabeled and labeled isotopomers of the same peptide should display similar peak shapes, the correlation coefficient is used as a measure of the overall measurement quality. Isotopomer chromatogram pairs that correlate poorly are often the result of coeluting isobaric interferences (example shown in Figure 3C), false positive peptide identifications, or if one or both of the peak profiles is indistinguishable from the background.

Because the enrichment of the stable isotope-labeled atoms will always be <100%, the true atom percent excess (ape) of the labeled atoms must be accounted for in the prediction of the labeled isotopomer's isotope distribution (Figure 2B). As the enrichment of individual isotope-labeled atoms decreases, the isotope distribution of the peptide becomes broader and shifts to lower m/z. Although the intensity of an individual isotope peak will be affected greatly by the enrichment of the stable isotope-labeled material, the sum of the intensity of the entire isotope distribution will not. Accounting for the incomplete enrichment of the labeled isotopomer ensures an accurate estimate of the m/z range of the extracted ion chromatogram and enables the stable isotope-labeled material from incompletely enriched precursor material to be used as internal standards in proteomics. Figure 4 shows the ion chromatograms of a tryptic peptide from 1:1 mixtures of unlabeled and ^{15}N-labeled *S. cerevisiae* grown using different enrichments of $(^{15}NH_4)_2SO_4$ in the growth media. The peptide, VINDAFGIEEGLMTTVHSLTATQK, derived from the protein glyceraldehyde-3-phosphate dehydrogenase (TDH3) produces a background-subtracted ion current ratio of 1.039, 1.083, and 1.004 in growth media of 90, 80, and 70 ape ^{15}N, respectively. As expected, by summing the entire isotope envelope, the resulting quantitative peptide ion current ratios are unaffected by the enrichment of the material used to label the internal standard. Although peptides with broader isotope distributions often have lower S/N, this approach is tolerant of any enrichment of labeled atoms providing (1) the atomic enrichment of the labeled atoms is known, (2) the protein has completely equilibrated with the enriched atoms from which it is synthesized, and (3) the labeled

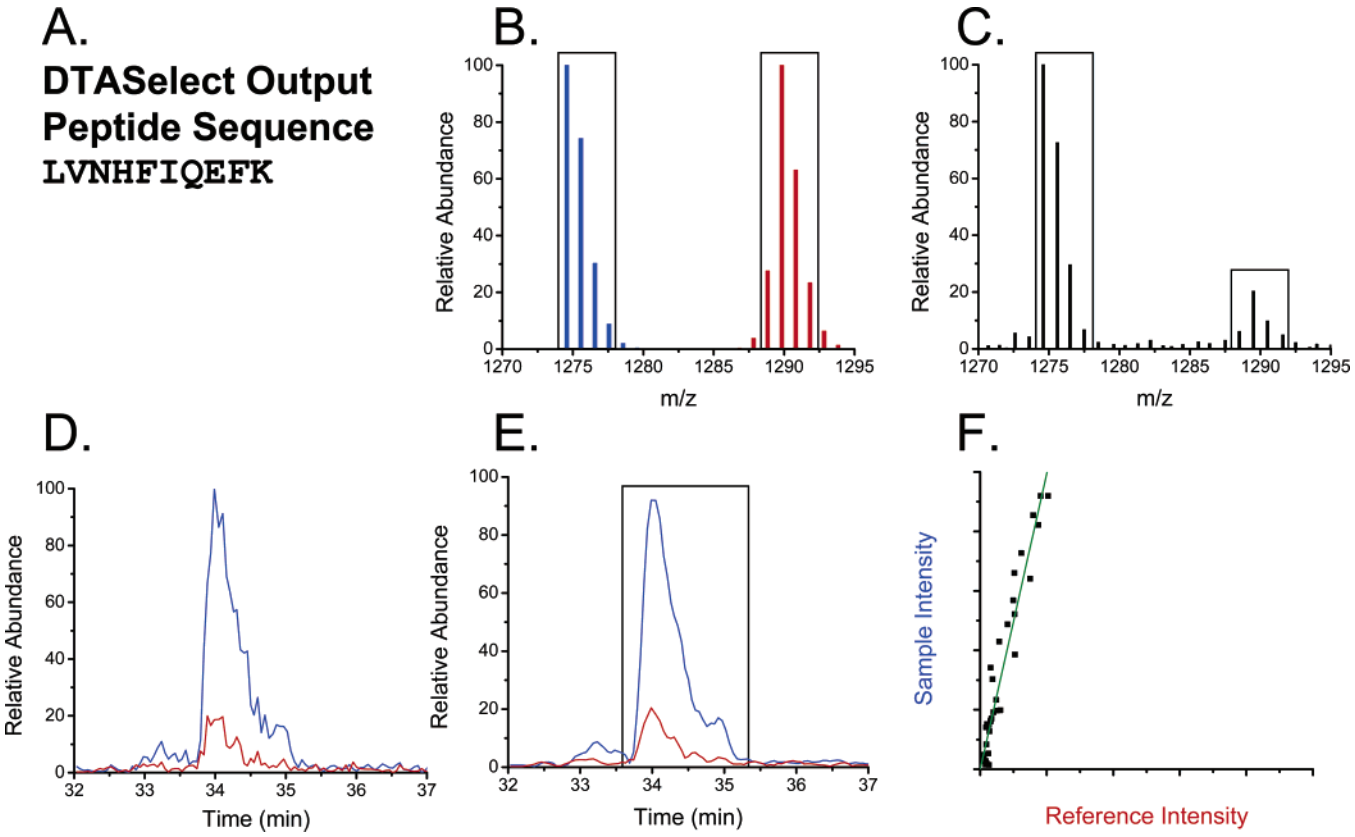

Figure 2. Calculation of peptide ion current ratios from samples containing a stable isotope-labeled internal standard. (A) Peptide sequences are identified from the tandem mass spectra acquired during a data-dependent μLC/μLC/MS/MS analysis using SEQUEST.[5] (B) The peptide isotope distributions are predicted for the unlabeled (blue) and labeled (red) peptide sequence. These isotope distributions are used to estimate the *m*/*z* range for the extraction of ion chromatograms (boxes). (C) The intensity is summed for the unlabeled and labeled *m*/*z* range (boxes) from the MS scans surrounding the MS/MS spectrum identifying the peptide sequence. (D) The intensity data from the individual MS spectra are used to produce unlabeled and labeled ion chromatograms. (E) Chromatograms are smoothed using a seven-point Savitsky−Golay filter and the approximate peak is located. (F) The background-subtracted ion current ratio is obtained from the slope of a least-squares regression between the two ion chromatograms.

and unlabeled isotope distributions are separated and do not contribute in intensity to one another.

Analysis of Peptide Ion Current Ratios Using Labels Containing Deuterium-Enriched Atoms. The chromatographic separation of protium (^{1}H)- and deuterium (^{2}H)-labeled isotopomer pairs is significantly greater than any of the other of the commonly used stable isotope labels.[45,46] Zhang et al. studied this isotope effect with respect to proteomics data[47,48] and reported large errors if the isotopomer ratio was sampled at a single point during the entire elution profile. If the ratio is sampled at the beginning of the peak elution, it is enriched in the deuterated species. Likewise, if the ratio is sampled at the tailing end of the peak, it is enriched in the nondeuterated species. However, when the area under the entire ion chromatogram profile is used in the ratio calculation (instead of a single spectrum), the chromatographic separation of unlabeled/labeled isotopomers results in minimal quantitative errors.[49,50]

Our correlation approach is also capable of handling peptide pairs that have been labeled using deuterium (i.e., ICAT) and are fractionated chromatographically. One might anticipate that this regression approach would be incapable of handling these data because the separation of the different isotopic species would result in poor correlation between two isotopomer ion chromatograms. A feature has been added to RelEx to offset the two ion chromatograms scan by scan until the correlation coefficient reaches a maximum. This approach is similar to one used by the program SEQUEST to estimate and subtract the background during the cross-correlation of two spectra.[5] However, whereas SEQUEST uses the offset to estimate the background and assumes that the maximum score occurs at a zero offset, RelEx uses the offset to find the correlation of two chromatograms when the maximum score does not occur at a zero offset. Figure 5 illustrates the effect of shifting the ion chromatograms to improve the correlation for an ICAT-derivatized peptide VNLDTDC$_{ICAT}$-QYAYLTGIR. Without shifting the chromatograms (Figure 5A), the correlation was poor (r = 0.4804) and an elliptical ratio response is observed in the correlation window as the per scan ratio changes during the elution of the partially resolved isoto-

(45) Bentley, R.; Saha, N. C.; Sweeley, C. C. *Anal. Chem.* **1965**, *37*, 1118−22.
(46) Klein, P. D. In *Advances in Chromatography*; Giddings, J. C., Keller, R. A., Eds.; Marcel Dekker: New York, 1966.
(47) Zhang, R.; Sioma, C. S.; Thompson, R. A.; Xiong, L.; Regnier, F. E. *Anal. Chem.* **2002**, *74*, 3662−9.
(48) Zhang, R.; Sioma, C. S.; Wang, S.; Regnier, F. E. *Anal. Chem.* **2001**, *73*, 5142−9.
(49) Lindeman, L. P.; Annis, J. L. *Anal. Chem.* **1960**, *32*, 1742−9.
(50) Sweeley, C. C.; Elliott, W. H.; Fries, I.; Ryhage, R. *Anal. Chem.* **1966**, *38*, 1549−53.

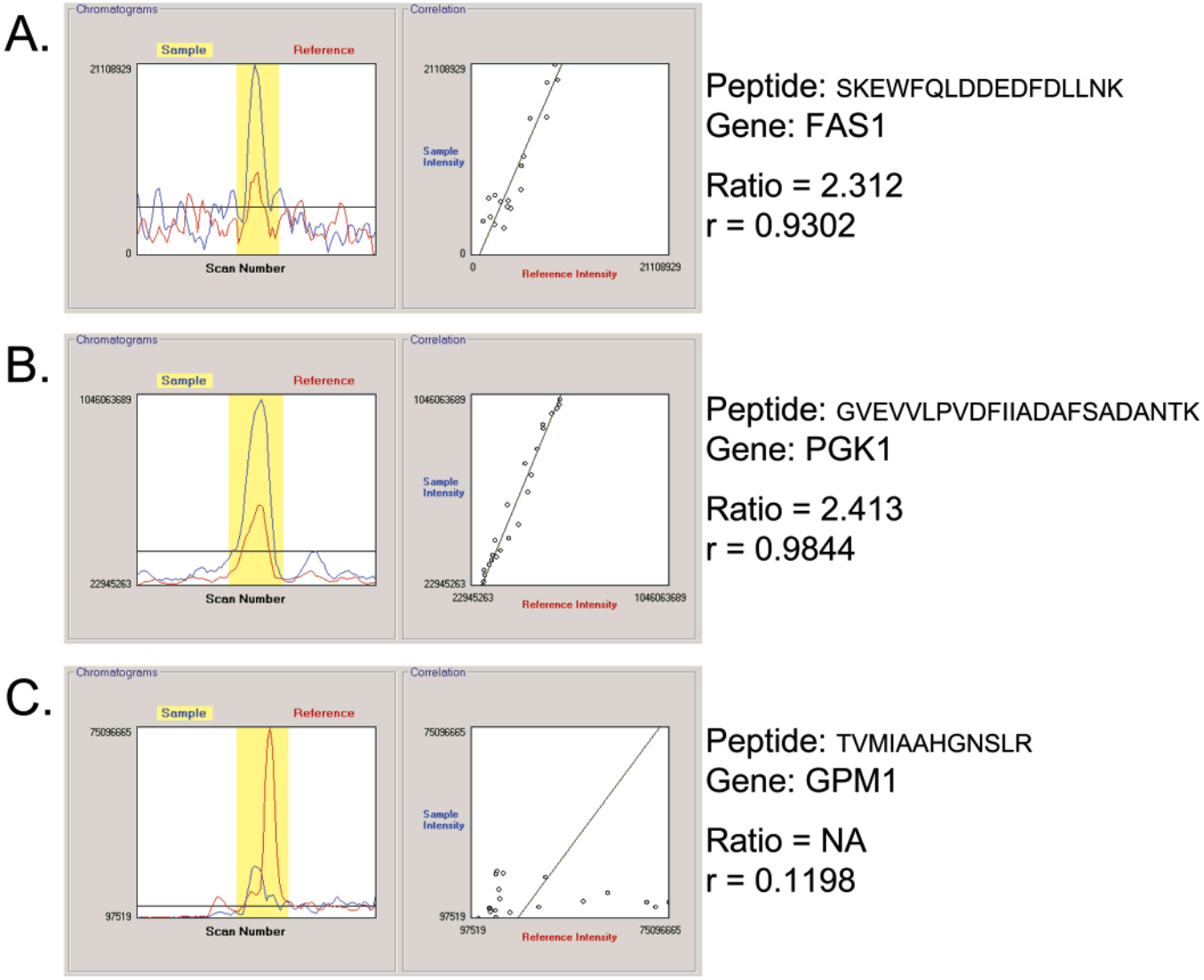

Figure 3. Display of peptide ion chromatograms and least-squares regression in RelEx's graphical user interface. RelEx is able to effectively detect the peaks and calculate the ion current ratios for chromatograms of varying S/N (A and B). The correlation coefficient provides a quantitative measure useful in eliminating chromatograms with isobaric interferences (C).

pomers. However, after shifting the deuterium-labeled chromatogram backward by two scans to maximize the correlation, the elliptical ratio response was minimized and the correlation coefficient improved to $r = 0.9879$ (Figure 5B).

Normalization of Mass Spectrometry-Derived Peptide Ion Current Ratios. Just as microarray Cy5/Cy3 ratios will not always directly approximate mRNA ratios, mass spectrometry-derived ion current ratios of peptides will not necessarily represent relative protein ratios. Because of this limitation, the first transformation of mRNA microarray data is the normalization of the fluorescence response ratios so that meaningful biological comparisons can be made.[51] Microarray data must be normalized for a number of reasons including but not limited to the following: (1) differences in labeling and detection efficiencies, (2) unequal starting material between the two samples, and (3) systematic biases in the measured expression levels.[51,52] These reasons for normalization are equally, if not more prevalent in mass spectrometry-derived quantitative proteomics data.

The most common way of normalizing these data has been to apply a constant factor to all the loci based on the overall mean intensity ratio,[51] the median ratio,[53] invariant reference genes (e.g., housekeeping genes), or external reference standards.[54] However, this approach is limited if there is a systematic dependence of the measured ratio on intensity as suggested for low S/N mass spectrometry data by Ong et al.[16] Scatterplots of the ratio versus intensity (known as an "$R-I$" plot')[52] can reveal intensity-dependent dependences in the measurement of the ratio.[51,52] Figure 6A shows a plot of the $\log_2(^{14}N/^{15}N)$ versus the log $(S/N_{14N}S/N_{15N})$ for 3677 peptide ratios (from 893 loci) calculated using RelEx from a unlabeled yeast sample taken after 40 min of NaCl osmotic stress and a ^{15}N-enriched yeast internal standard grown under normal conditions. Unlike the large systematic biases

(51) Quackenbush, J. *Nat. Genet.* **2002**, *32 Suppl*, 496−501.
(52) Yang, I. V.; Chen, E.; Hasseman, J. P.; Liang, W.; Frank, B. C.; Wang, S.; Sharov, V.; Saeed, A. I.; White, J.; Li, J.; Lee, N. H.; Yeatman, T. J.; Quackenbush, J. *Genome Biol.* **2002**, *3*, research 0062.
(53) Yang, Y. H.; Dudoit, S.; Luu, P.; Lin, D. M.; Peng, V.; Ngai, J.; Speed, T. P. *Nucleic Acids Res.* **2002**, *30*, e15.
(54) Eickhoff, B.; Korn, B.; Schick, M.; Poustka, A.; van der, Bosch J. *Nucleic Acids Res.* **1999**, *27*, e33.

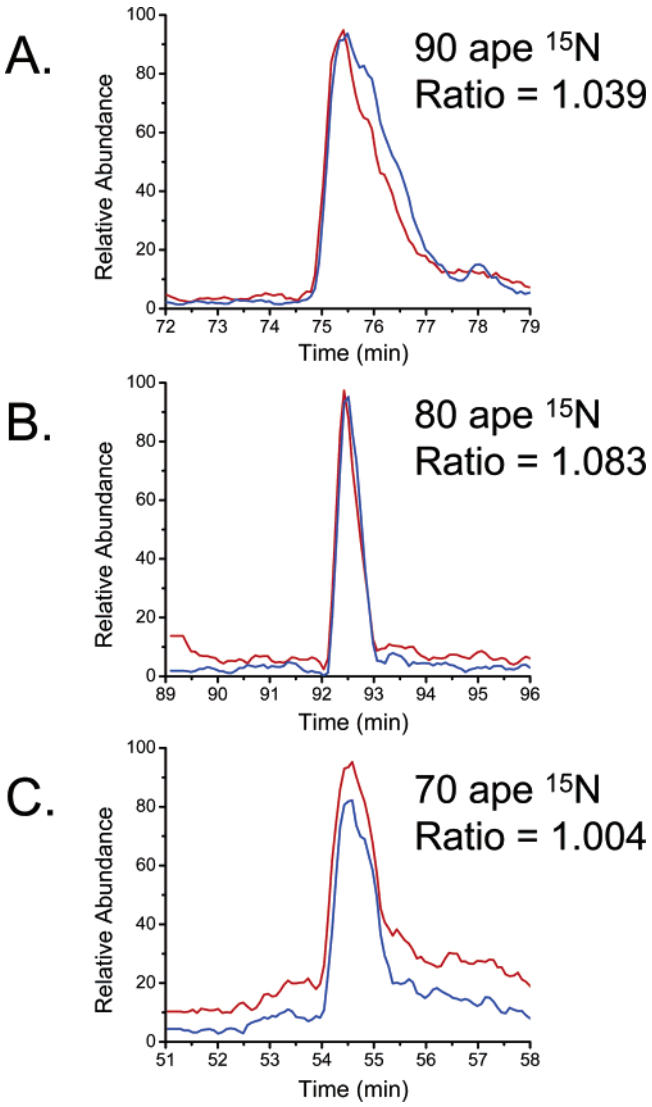

Figure 4. Measurement of peptide levels using *S. cerevisiae* grown in different enrichments of $(^{15}NH_4)_2SO_4$ in the growth media. Using the described approach, the quantitative accuracy is not affected by the enrichment of the material used to produce the internal standard. The peptide, VINDAFGIEEGLMTTVHSLTATQK, derived from the protein glyceraldehyde-3-phosphate dehydrogenase produces a background-subtracted ion current ratio of 1.039, 1.083, and 1.004 in growth media of 90, 80, and 70 ape ^{15}N, respectively.

that are often observed from unnormalized mRNA microarray data,[53] there does not appear to be any obvious systematic intensity-based dependency in our peptide ratio data. However, we do see a broader distribution of ratio values (variance) with the low S/N data than with the high S/N results. This broader distribution is the result of intensity-based stochastic processes inherent in any ion current ratio measurement by mass spectrometry.[55]

Because minimal intensity-dependent biases were observed in the ratios over different intensities, a global normalization scheme applied over all intensities was chosen instead of a locally weighted normalization.[52,53,56] Figure 6B shows the frequency versus ratio plot before (solid line) and after normalization (dashed line) against the median ratio (0.839). The distribution is near Gaussian with an $r^2 = 0.973$. Normalizing the data in this fashion assumes that the number of peptides that increase equals the number of peptides that decrease. This assumption is acceptable given basic mass balance, equal amounts of protein between samples, and that the measured proteins are a representative sampling of the total expressed genome. A preferred normalization would be against the peptide ratios from a known nonvariant protein. However, in this experiment we chose to normalize to the median because under osmotic stress many "housekeeping" and cytoskeletal genes routinely used for normalization of gene expression data have known altered expression.[57,58]

Systematic Errors in the Estimate of Protein Ratios. Yeast cells were grown in either unlabeled or ^{15}N-enriched media and then mixed in known unlabeled/^{15}N-enriched ratios to evaluate the accuracy and precision of automated isotopomer ratio measurements with RelEx. Each peptide's ratio was calculated and normalized, and the mean peptide ratio for the protein was calculated following the removal of outliers using a Dixon's test. Figure 7A shows the measured mean peptide ratio (±SE) for three selected proteins at four different protein ratios in the context of a whole cell lysate. The solid, dashed, and dotted response curves are from the proteins encoded by the genes TSA1 (slope, 1.335 ± 0.015), SSA1 (slope, 1.004 ± 0.018), and ADH1 (slope, 0.661 ± 0.045), respectively. The protein ratios were estimated from the measured mean ratio for all the proteins within the mixture. Although the linearity of the quantitative protein measurement is excellent ($r^2 > 0.99$), selected proteins in the analysis have a response factor or slope that is not equal to unity (Figure 7A).

The observation of systematic errors is a well-established phenomenon in isotope dilution measurements,[24,59] yet has been largely ignored in the quantitative proteomic analyses to date. Traditionally, standard samples that contain known ratios of labeled and unlabeled material are measured and plotted as a function of the known ratios to calibrate the mass spectrometer response for the analyte.[60] Obviously the production of a calibration curve with known ratios for thousands of analytes profiled in a proteomic experiment is prohibitive. However, the overall scheme shown in Figure 1 compensates for inaccuracies in the ratio by dividing the ratio of a protein from one sample by the ratio of the same protein in a control sample. Because both samples contain the same internal standard, any inaccuracies derived from systematic errors should be present in both analyses and cancel—appropriately adjusting the final fractional response ratio. The calculation of fractional response from the ratio of two ratios is the equivalent of dividing each individual ratio by the slope from the standard curve.

Figure 7B demonstrates the improvement in accuracy of the protein ratio measurement after correction for systematic errors. The shaded and unshaded bars are the average ratios of 35

(55) MacCoss, M. J.; Toth, M. J.; Matthews, D. E. *Anal. Chem.* **2001**, *73*, 2976−84.
(56) Cleveland, W. S. *J. Am. Stat. Assoc.* **1979**, *74*, 829−36.
(57) Causton, H. C.; Ren, B.; Koh, S. S.; Harbison, C. T.; Kanin, E.; Jennings, E. G.; Lee, T. I.; True, H. L.; Lander, E. S.; Young, R. A. *Mol. Biol. Cell* **2001**, *12*, 323−37.
(58) Yuzyuk, T.; Foehr, M.; Amberg, D. C. *Mol. Biol. Cell* **2002**, *13*, 2869−80.
(59) Matthews, D. E. In *Amino Acid/Protein Metabolism in Health and Disease: Nutritional Implications*; Tessari, P., Soeters, P. B., Pittoni, G., Tiengo, A., Eds.; Smith-Gordon: London, 1997; Chapter 3.
(60) Watson, J. T. In *Methods in Enzymology*; McCloskey, J. A., Ed.; Academic Press: San Diego, 1990; Chapter 4.

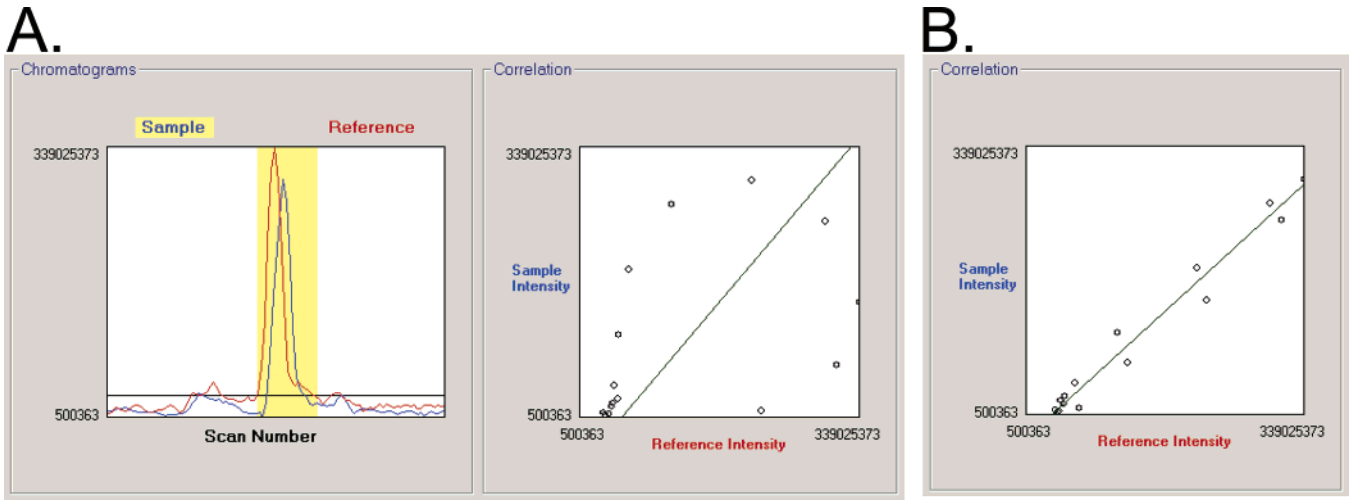

Figure 5. Calculation of ion current ratios using ion chromatograms from $^2H_0$- and $^2H_8$-labeled ICAT peptide pairs. (A) Chromatograms are shown for the ICAT derivatived peptide VNLDTDC$_{ICAT}$QYAYLTGIR from yeast proteins mixed in a 1:1 mole ratio. The peptide containing the deuterium labels (red chromatogramm) elutes prior to the identical peptide containing only natural-abundance isotopes (blue chromatogram) using reversed-phase chromatography. This chromatographic shift causes a large elliptical ratio response and thus poor correlation ($r = 0.4804$) because the two ion chromatograms are offset. (B) RelEx can shift the two chromatograms across each other a fixed scan range until the correlation coefficient reaches a maximum. After shifting the $^2H_8$-labeled peptide back two scans, the correlation coefficient improves to $r = 0.9879$.

proteins (±SD) before and after correction, respectively, using an external sample. Although all 35 proteins have the same unlabeled/^{15}N-labeled mole ratios, the measured mass spectrometry ion current ratios are not the same. As expected, the mean ratios of the 35 proteins show no significant change after correction for mass spectrometer response. However, the deviation of the measured ratio from the true mole ratio (standard deviation) is reduced after correction for systematic errors with the exception of the 2.4:1 sample ($p = 0.06$). This simple minimization of systematic errors using a ratio of two ratios improved the relative quantitative accuracy of protein measurements by 32 ± 4%.

Measurement of Changes in *S. cerevisiae* Protein Level in Response to Hyperosmotic Stress. Yeast grown to mid-log phase were subjected to 5% NaCl for 40 min and an aliquot of cells at time 0 min (control) and time 40 min were collected. Each aliquot was mixed immediately with equal ODs of ^{15}N-labeled cells and analyzed as described in the Experimental Section. The ion current ratios were calculated using RelEx, and differences in protein level were estimated as described above.

The calculation of peptide ratios and estimate of protein ratios using RelEx is completely automated. After the removal of peptide chromatograms that correlated poorly or had insufficient signal intensity, RelEx calculated 635 and 992 protein ratios from the 0- and 40-min samples, respectively. The intersection of these two data sets allowed the measurement of fractional change for 323 total proteins. The significance of the change between the two time points was estimated for these proteins, and the list was sorted by *p*-value.

The change in response to osmotic stress is shown in Figure 8 for the 25 proteins with the greatest statistical significance. Because the scale for the fractional change is different for repressed and induced expression (e.g., 0.1−0.5 versus 10.0−2.0), all repressed values are expressed as $-(1/R_i)$ and induced values are expressed simply as R_i (where R_i is the mean peptide ratio for protein i). This conversion ensures that values ⟨⟩ 1 are

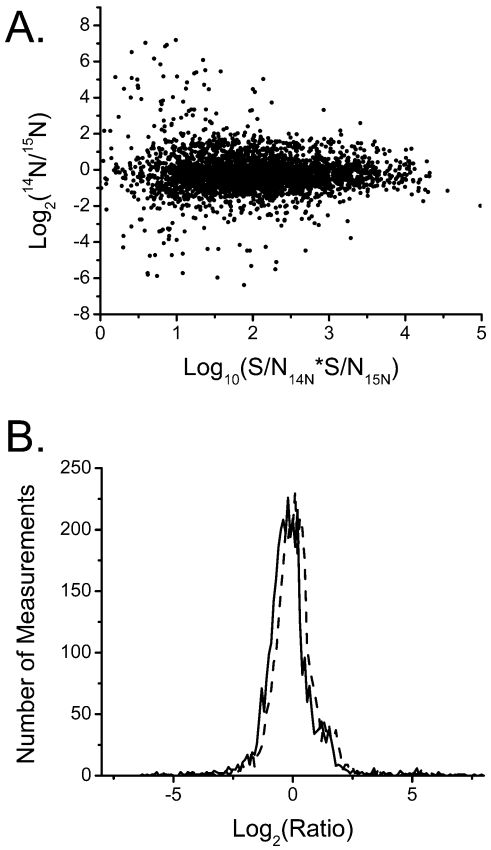

Figure 6. Normalization of peptide ion current ratios. (A) RelEx-derived ion current ratios were plotted versus intensity for 3677 different peptide measurements, and no intensity-based bias toward either the labeled or unlabeled signal is observed. However, as predicted by Poisson statistics, there is a greater variance at low intensity relative to high intensity. Because only stochastic error was observed, a global normalization factor was used to adjust the median change of all peptide ion current ratios to zero. (B) The solid line represents the distribution of peptide ratios prior to correction, and the dashed line represents the peptide ratios after application of the normalization factor.

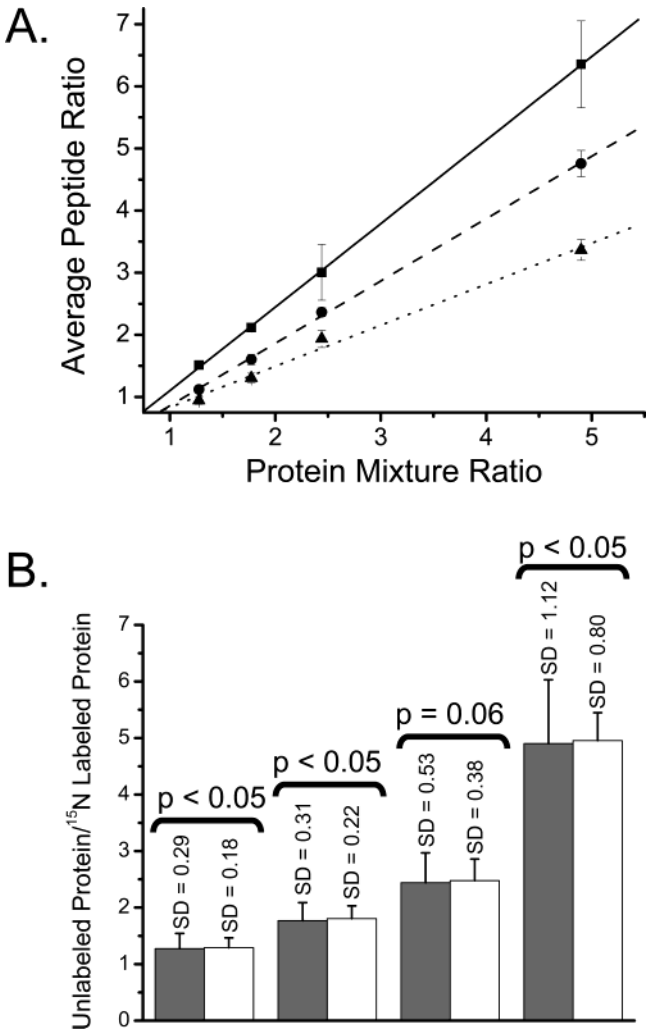

Figure 7. Systematic errors in the estimate of protein ratios. (A) The measured mean peptide ratio (±SE) for three selected proteins from a whole cell lysate are plotted at four different protein mole ratios. The solid, dashed, and dotted response curves are from the proteins encoded by the genes TSA1 (slope, 1.335 ± 0.015), SSA1 (slope, 1.004 ± 0.018), and ADH1 (slope, 0.661 ± 0.045), respectively. Although the linearity of the quantitative protein measurement is excellent (r^2 > 0.99), selected proteins in the analysis display systematic errors and have a response factor or slope that is not equal to unity. (B) The systematic errors, if uncorrected, will degrade the accuracy of the quantitative measurement. The figure represents the mean ± SD for 35 proteins from *S. cerevisiae* cells mixed in 4 different unlabeled/^{15}N-enriched ratios before (shaded) and after (unshaded) correction for systematic errors. After correction relative to an external sample, the number of proteins that approximate the mole ratio improves (i.e., the standard deviation about the mean decreases).

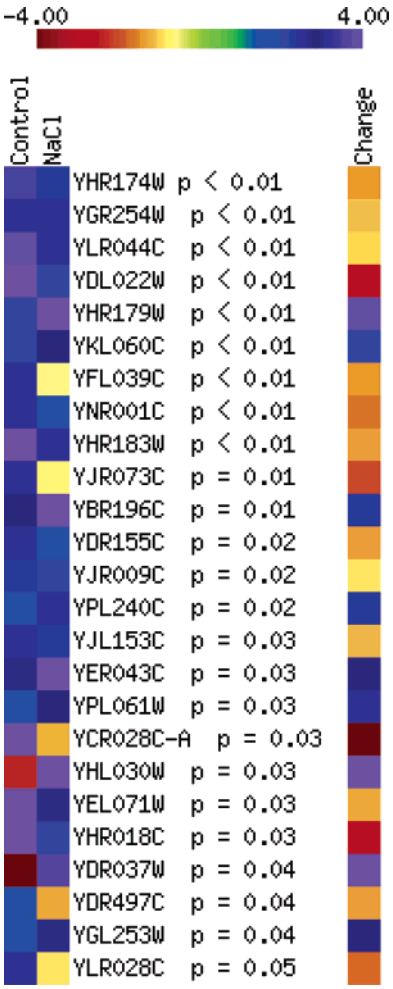

Figure 8. Measurement of the effect of NaCl osmotic stress on relative protein level. The change in protein level was calculated and sorted by *p*-value. The display uses a "full-spectrum" color scheme, where red, orange, and yellow represent repressed proteins, green represents proteins with no change, and blue, indigo, and violet represent induced proteins. The first two columns are from the uncorrected protein ratios at time 0 and time 40 min, respectively. The rightmost column is the change at time 40 min after correction relative to time 0. Of the 296 proteins quantified (ratios measured at both time 0 and time 40 min), 25 had $p \leq 0.05$ and 41 had $p \leq 0.10$.

presented on the same scale and effectively converts the fractional change to a "factor change". A full-spectrum color scheme (red-orange-yellow-green-blue-indigo-violet) was adopted to display the changes – where red, orange, and yellow represent repressed proteins and blue, indigo, and violet represent induced proteins. There were 25 proteins that have a significant change ($p \leq 0.05$) and a total of 42 proteins with a $p \leq 0.10$.

The ability to establish a difference between protein ratios is heavily dependent on the number of peptide ion current ratio measurements used to establish the protein ratio, and therefore, the proteins with the largest factor change are not always the most significant. This point is important because most quantitative proteomics studies to date have considered changes exceeding a selected value to be significant.[61,62] Furthermore, these data emphasize the importance of methodologies that improve protein sequence coverage[33,63] for quantitative measurements.

CONCLUSION

Developments in stable isotope-labeling strategies have initiated a multitude of quantitative proteomic applications. The conversion of mass spectrometry-derived data of peptides to relative protein abundances is tedious and subjective because these analyses are often filtered and calculated manually. We

(61) Blagoev, B.; Kratchmarova, I.; Ong, S. E.; Nielsen, M.; Foster, L. J.; Mann, M. *Nat. Biotechnol.* **2003**, *21*, 315–8.
(62) Ranish, J. A.; Yi, E. C.; Leslie, D. M.; Purvine, S. O.; Goodlett, D. R.; Eng, J.; Aebersold, R. *Nat. Genet.* **2003**, *33*, 349–55.
(63) 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., III. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 7900–5.

described a computer program that uses a least-squares regression for the automated conversion of mass spectrometry-derived peptide data (ion chromatograms) to relative protein abundances. This approach is tolerant of poor S/N data and can automatically discard nonintegratable chromatograms and outlier ratios without any user intervention. Systematic errors are minimized using a ratio of two ratios, and differences between samples are assessed using *t*-tests. Our approach was validated using complex mixtures with known mole ratio and demonstrated in a real system by measuring the affect of NaCl osmotic stress on protein level in *S. cerevisiae*. This software is broadly compatible with all of the stable isotope labeled approaches reported.

ACKNOWLEDGMENT

The authors gratefully acknowledge financial support from National Institutes of Health grants RR11823-08 (to J.R.Y.), F32DK59731 (to M.J.M.), F32AI54333 (to C.C.W.), and Office of Naval Research grant N00014-00-1-0421 (to J.R.Y.). We appreciate helpful discussions with Dwight Matthews and Steve Shinebarger at the University of Vermont and Claire Delahunty, Hayes McDonald, David Tabb, and John Venable of the Yates' laboratory.

Received for review July 13, 2003. Accepted September 30, 2003.

AC034790H