Automated Statistical Analysis of Protein Abundance Ratios from Data Generated by Stable-Isotope Dilution and Tandem Mass Spectrometry

Xiao-jun Li,* Hui Zhang, Jeffrey A. Ranish, and Ruedi Aebersold

The Institute for Systems Biology, 1441 North 34th Street, Seattle, Washington 98103-8904

We describe an algorithm for the automated statistical analysis of protein abundance ratios (ASAPRatio) of proteins contained in two samples. Proteins are labeled with distinct stable-isotope tags and fragmented, and the tagged peptide fragments are separated by liquid chromatography (LC) and analyzed by electrospray ionization (ESI) tandem mass spectrometry (MS/MS). The algorithm utilizes the signals recorded for the different isotopic forms of peptides of identical sequence and numerical and statistical methods, such as Savitzky−Golay smoothing filters, statistics for weighted samples, and Dixon's test for outliers, to evaluate protein abundance ratios and their associated errors. The algorithm also provides a statistical assessment to distinguish proteins of significant abundance changes from a population of proteins of unchanged abundance. To evaluate its performance, two sets of LC-ESI-MS/MS data were analyzed by the ASAPRatio algorithm without human intervention, and the data were related to the expected and manually validated values. The utility of the ASAPRatio program was clearly demonstrated by its speed and the accuracy of the generated protein abundance ratios and by its capability to identify specific core components of the RNA polymerase II transcription complex within a high background of copurifying proteins.

Quantitative proteomics plays an increasingly important role in biological and medical research.[1−3] By systematically measuring protein abundance changes in cells, tissues, or body fluids induced by changing environmental or physiological conditions, quantitative proteomics can provide insights into the molecular changes that accompany or induce these conditions.[4,5] In addition, quantitative proteomics can be used to characterize the composition of macromolecular complexes and changes in complex composition and abundance that accompany changes in cell states.[6,7]

Protein identification and quantification are two distinct but integrated essentials in quantitative proteomics. Traditionally, proteins from different biological origins were separated by two-dimensional gel electrophoresis (2DE), and quantification was achieved by comparing the staining intensity of the spots representing the same protein in different gels. Selected protein spots were then excised, proteolyzed, and analyzed by mass spectrometry (MS) which led to protein identification.[8] Recently, a second technique was developed that is based on stable isotope tagging of proteins and automated peptide tandem mass spectrometry (MS/MS).[1] In this method, proteins contained in different samples are labeled with a distinct isotopic signature by metabolic labeling,[9−12] enzymatic reaction,[13,14] or chemical reaction.[15−17] The differentially labeled samples are then combined and concurrently processed and analyzed. In one implementation of this method, the combined labeled sample is digested, and the resulting peptide mixture is analyzed by multidimensional liquid chromatography (LC) and electrospray ionization (ESI) MS/MS. In this way, protein identification and quantification are accomplished by the identification and quantification of the corresponding sibling peptides. Importantly, in this method peptide identification and quantification are determined in a single, automated operation.

* Corresponding author. E-mail: xli@systemsbiology.org. Fax: (206)732-1299.

[1] Aebersold, R.; Mann, M. *Nature* **2003**, *422*, 198−207.

[2] Patterson, S. D.; Aebersold, R. H. *Nat. Genet.* **2003**, *33 suppl*, 311−323.

[3] Hanash, S. *Nature* **2003**, *422*, 226−232.

[4] Ideker, T.; Thorsson, V.; Ranish, J. A.; Christmas, R.; Buhler, J.; Eng, J. K.; Bumgarner, R.; Goodlett, D. R.; Aebersold, R.; Hood, L. *Science* **2001**, *292*, 929−934.

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

[6] Ranish, J. A.; Yi, E. C.; Leslie, D. M.; Purvine, S. O.; Goodlett, D. R.; Eng, J.; Aebersold, R. *Nat. Genet.* **2003**, *33*, 349−355.

[7] Blagoev, B.; Kratchmarova, I.; Ong, S. E.; Nielsen, M.; Foster, L. J.; Mann, M. *Nat. Biotechnol.* **2003**, *21*, 315−318.

[8] Shevchenko, A.; Jensen, O. N.; Podtelejnikov, A. V.; Sagliocco, F.; Wilm, M.; Vorm, O.; Mortensen, P.; Boucherie, H.; Mann, M. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 14440−14445.

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

[10] Pasa-Tolic, L.; Jensen, P. K.; Anderson, G. A.; Lipton, M. S.; Peden, K. K.; Martinovic, S.; Tolic, N.; Bruce, J. E.; Smith, R. D. *J. Am. Chem. Soc.* **1999**, *121*, 7949−7950.

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

[12] Washburn, M. P.; Ulaszek, R.; Deciu, C.; Schieltz, D. M.; Yates, J. R., III. *Anal. Chem* **2002**, *74*, 1650−1657.

[13] Mirgorodskaya, O. A.; Kozmin, Y. P.; Titov, M. I.; Korner, R.; Sonksen, C. P.; Roepstorff, P. *Rapid Commun. Mass Spectrom.* **2000**, *14*, 1226−1232.

[14] Yao, X.; Freas, A.; Ramirez, J.; Demirev, P. A.; Fenselau, C. *Anal. Chem.* **2001**, *73*, 2836−2842.

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

[16] Goodlett, D. R.; Keller, A.; Watts, J. D.; Newitt, R.; Yi, E. C.; Purvine, S.; Eng, J. K.; von Haller, P.; Aebersold, R.; Kolker, E. *Rapid Commun. Mass Spectrom.* **2001**, *15*, 1214−1221.

[17] Cagney, G.; Emili, A. *Nat. Biotechnol.* **2002**, *20*, 163−170.

10.1021/ac034633i CCC: $25.00
Published on Web 10/28/2003

In LC-ESI-MS/MS analysis, a reversed phase (RP) microcapillary LC system is normally connected online with an ESI source. Peptides separated by RP-LC are ionized via ESI upon their elution from the column, and the resulting ions are directly transferred into a mass spectrometer for analysis. ESI is achieved by a high electric potential applied to the downstream tip of the column and results in the association of peptide molecules with varying numbers of H^+ ions and therefore the formation of peptide ions at different charge states.[18] The exact distribution of a peptide in these charge states depends on factors such as peptide sequence, applied potential, sample composition, and RP buffers and is therefore difficult to accurately predict or model. To identify and quantify peptides in the same operation, a mass spectrometer is operated alternatively in MS survey and MS/MS mode. During the MS/MS mode, precursor ions within a narrow m/z window are selected and fragmented by collision-induced dissociation (CID). The peptide sequence is identified by searching the CID spectrum against a database using software such as SEQUEST[19] or Mascot.[20] During the MS survey, the ion intensity of all peptides present at a certain time point and within a selected, broad m/z range is recorded. Single-ion chromatograms for each identified peptide can be constructed from those collected MS spectra. The elution peaks of the differentially labeled peptides of identical sequence are identified and their areas calculated. The relative abundance of a peptide in different samples can then be determined by comparing the obtained areas.[5,15]

Automated LC-ESI-MS/MS systems have been used to identify and quantify thousands of peptides and proteins.[5,21] While such high throughput is necessary and crucial for systematic studies of different biological samples,[4] it also generates an overwhelming amount of data to be evaluated and validated. The commonly used manual validation of database search and quantification results is time-consuming and error-prone. In addition, criteria used for validation vary from person to person and from experiment to experiment, which makes it difficult to compare results from different groups or experiments. Recently, an advance was made to automatically assess the validity of peptide and protein identifications made by MS/MS and database searches.[22,23] However, analysis tools capable of rapidly and reliably determining relative protein abundance in isotope-tagged protein samples are still lacking.

Several factors complicate peptide quantification and a straightforward connection between peptide quantification and protein quantification: (1) Multiple peptides may be identified from the same protein. (2) During multidimensional chromatographic separation, a particular peptide may be split between different fractions. (3) The same peptide may be identified multiple times in different isotopic forms or in different charge states. (4) A peptide may be misidentified or post-translationally modified. (5) In rare cases, the same peptide may elute from a RP column multiple times, a phenomena not well understood but possibly due to peptide–peptide interactions or peptide secondary structure. (6) The data quality, e.g., signal-to-noise ratio (S/N), of peptide signals may vary. Clearly, a sophisticated method is needed to take into account all of these complications before one can reliably and automatically evaluate protein quantification.

Here we report an algorithm for automated statistical analysis of protein abundance ratios (ASAPRatio) from data generated by stable-isotope dilution and MS/MS. The algorithm utilizes numerical and statistical methods, such as Savitzky–Golay smoothing filters,[24] statistics for weighted samples,[25] and Dixon's test for outliers,[26,27] to evaluate relative protein abundance ratios and their associated errors. Error analysis provides an assessment of the reliability of the quantification results, which is essential for consistent validation of large data sets from high-throughput proteomics. The ASAPRatio program further calculates protein p values that allow users to distinguish proteins of significant abundance changes from a large number of proteins of unchanged abundance. The application of the ASAPRatio tool dramatically accelerates and increases the consistency of data analysis for large-scale protein profiling experiments. The ASAPRatio program also increases the dynamic range of detectable differences in relative abundance by subtracting background noise from signal intensity in every single-ion chromatogram. This automated software increases the speed of quantitative data analysis, which is currently a major bottleneck in high-throughput quantitative proteomics.

MATERIALS AND METHODS

For all chromatographic steps, LC grade reagents were purchased from Fisher Scientific (Pittsburgh, PA). All other chemicals used in this study were purchased from Sigma (St. Louis, MO) unless specified otherwise.

Esterification of Standard Protein. Bovine serum albumin protein (50 μg) (BSA, Sigma, St. Louis, MO) was resuspended in 200 μL of buffered urea solution (8 M urea/0.4 M NH_4HCO_3, pH 8.3) and incubated at 55 °C for 30 min. The solution was diluted with 600 μL of water. Trypsin (1 μg) (Promega, Madison, WI) was added, and the sample was incubated at 37 °C overnight. The peptides were reduced by adding 8 mM TCEP (PIERCE, Rockford, IL) at room temperature for 30 min and alkylated by adding 10 mM iodoacetamide (Sigma, St. Louis, MO) at room temperature for 30 min. The peptides were purified using a C18 SepPack column (Milford, MA) according to the manufacturer's instruction. The purified peptides were split into two equal fractions (A and B) and lyophilized to dryness in a Speedvac vacuum concentrator prior to methyl esterification as described.[16,28] Briefly, 160 μL of acetyl chloride was added dropwise to 1 mL of d_0- or d_3-methanol with stirring on ice, and the reaction was continued for 5 min at

[18] Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. *Science* **1989**, *246*, 64–71.

[19] Eng, J. K.; McCormack, A. L.; Yates, J. R. III. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976–989.

[20] Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20*, 3551–3567.

[21] von Haller, P. D.; Yi, E.; Donohoe, S.; Vaughn, K.; Keller, A.; Nesvizhskii, A.; Eng, J.; Li, X.-J.; Wollscheid, B.; Goodlett, D. R.; Aebersold, R.; Watts, J. D. *Mol. Cell. Proteomics* **2003**, *2*, 428–442.

[22] Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2002**, *74*, 5383–5392.

[23] Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. *Anal. Chem* **2003**, *75*, 4646–4658.

[24] Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical recipes in C: the art of scientific computing*, 2nd ed.; Cambridge University Press: New York, 1997.

[25] Taylor, J. R. *An introduction to error analysis: the study of uncertainties in physical measurements*, 2nd ed.; University Science Press: Sausalito, CA, 1997.

[26] Dixon, W. J. *Biometrics* **1953**, *9*, 74–89.

[27] Roracher, D. B. *Anal. Chem.* **1991**, *63*, 139–146.

[28] Hunt, D. F.; Yates, J. R., III; Shabanowitz, J.; Winston, S.; Hauer, C. R. *Proc. Natl. Acad. Sci. U.S.A.* **1986**, *83*, 6233–6237.

room temperature. An amount of 380 μL of the thus freshly made d_0- and d_3-methanolic HCl solution was added to the two lyophilized tryptic albumin peptide samples and incubated for 2 h at room temperature. The reaction was stopped by lyophilization. Peptide fractions A and B were resuspended in 250 μL of 0.1% TFA and mixed to a final ratio of 1:1 (10 μL, 10 μL), 1:3 (10 μL, 30 μL), 1:10 (5 μL, 50 μL), 20:1 (50 μL, 2.5 μL), or 1:100 (1 μL, 100 μL), respectively. The mixtures were lyophilized and resuspended in 10 μL of 0.1% TFA and analyzed by a model LCQ ion-trap ESI-MS system (ThermoFinnigan, San Jose, CA).

RNA Polymerase II (Pol II) Transcription Complex. RNA Pol II transcription complexes were prepared as described previously.[6,29] Briefly, a nuclear extract from a yeast strain carrying a temperature sensitive mutation in the TATA-box-binding protein (TBPI143N) was incubated for 60 min with HIS4 promoter templates immobilized on magnetic beads (Dynal, Oslo, Norway) in the presence or absence of recombinant TBP (rTBP). Transcription buffer consisted of 20 mM HEPES buffer (pH 7.6), 100 mM potassium acetate, 5 mM magnesium acetate, and 1 mM EDTA. The templates were washed four times with transcription buffer containing 0.05% NP40 and 2.5 mM DTT, followed by one wash with transcription buffer containing 0.003% NP40. Templates were resuspended in 1 mL of PstI buffer [100 mM NaCl, 50 mM Tris-HCl, (pH 7.9) 10 mM $MgCl_2$] with 645 units of PstI (Boehringer Mannheim) and incubated for 30 min at 22 ^{o}C with agitation. The beads were concentrated with a magnet, and the supernatants were recovered. Protein samples were isotopically labeled with heavy (+rTBP) or normal (−rTBP) versions of ICAT reagent (ABI, Foster City, CA), digested with trypsin, and peptides were prepared for mass spectrometry as described.[6] Specific components of the transcription complex were distinguished from copurifying proteins by their increased abundance in the presence of TBP.

Mass Spectrometry. All samples were analyzed on an ion-trap ESI-MS (LCQ Classic, ThermoFinnigan, San Jose, CA) which was coupled to an online microcapillary RP-LC system and operated by alternating MS and MS/MS scans.[15] Peptide CID spectra were searched against suitable sequence databases, using the SEQUEST search tool.[19] The search results were filtered by SEQUEST cutoff scores using the INTERACT data organizing tool.[5] In the BSA experiment, all peptide identifications were filtered with the following SEQUEST scores: Xcorr >1.5 for the $[M + H]^+$ precursor ion, >2 for $[M + 2H]^{2+}$, and >2.5 for $[M + 3H]^{3+}$, and δCn >0.1. In the RNA Pol II complex experiment, all peptide identifications were initially filtered with the following SEQUEST scores: Xcorr >1.5 for the $[M + H]^+$ precursor ion, >1.4 for $[M + 2H]^{2+}$, and >2.0 for $[M + 3H]^{3+}$. Peptide abundance ratios were independently determined using the XPRESS quantification tool.[5] Both peptide identifications and quantifications were manually validated.[6]

Quantification Algorithm. The ASAPRatio program is written in C, and the current version runs on a Linux operating system. The ASAPRatio program does quantification after peptide sequence identification and verification are completed. Currently, it collects the following information from output files of the INTERACT data organizing tool:[5] peptide sequences, scan numbers and charge states at their identification, corresponding proteins, and experiment data files. Although the software is designed to operate automatically, it allows users the option to view, verify, and, if necessary, correct results manually through two interfaces developed in common gateway interface (CGI) programming.

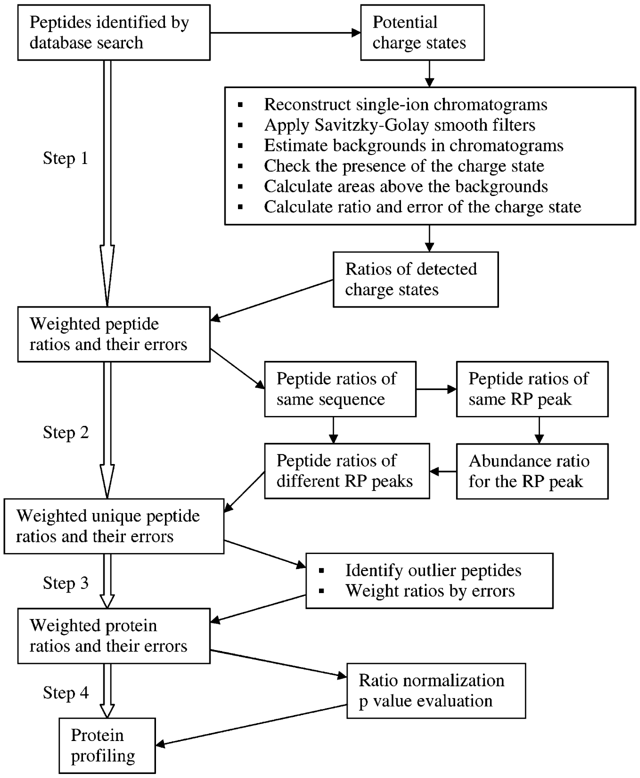

Peptides identified by database search
Potential charge states
Reconstruct single-ion chromatograms
Apply Savitzky-Golay smooth filters
Estimate backgrounds in chromatograms
Check the presence of the charge state
Calculate areas above the backgrounds
Calculate ratio and error of the charge state
Step 1
Ratios of detected charge states
Weighted peptide ratios and their errors
Peptide ratios of same sequence
Peptide ratios of same RP peak
Step 2
Peptide ratios of different RP peaks
Abundance ratio for the RP peak
Weighted unique peptide ratios and their errors
Identify outlier peptides
Weight ratios by errors
Step 3
Weighted protein ratios and their errors
Ratio normalization
p value evaluation
Step 4
Protein profiling

Figure 1. Flowchart of procedure to determine protein quantification and profiling. Here, the block arrows outline the four main steps of the procedure described in the text. The solid arrows indicate the details of the steps.

An interface allows users to specify parameters in their experiment (see Supporting Information). For example, the software allows users to either select from a list or define with text strings any number of modified amino acids. The software also allows users to specify amino acids that are isotopic partners. To account for the mobility shift between some isotopically labeled pairs during RP-LC separation, e.g., the separation of ^{1}H and ^{2}H labeled peptide pairs, the software offers users an option to specify the elution order of the two partners. These features provide the flexibility required for the analysis of data generated from peptides labeled with multiple and diverse isotopic tags.

Procedure To Determine Protein Quantification and Profiling. The procedure used in the ASAPRatio program to determine protein quantification and profiling is schematically depicted as a flow diagram in Figure 1. It includes the following four steps.

Step 1: Evaluation of a Peptide Abundance Ratio for Each Peptide Identified by MS/MS and Database Searching. The ASAPRatio program reconstructs a raw single-ion chromatogram by summing all ion intensities within an m/z range covering the first three theoretical isotopic peaks of the peptide and over the chromatographic elution period of the peptide. It then applies the Savitzky−Golay smooth filtering method[24] to obtain a smoothed chromatogram. The peptide elution peak is identified along with the corresponding peak center and peak width from the smoothed

[29] Ranish, J. A.; Yudkovsky, N.; Hahn, S. *Genes Dev.* **1999**, *13*, 49−63.

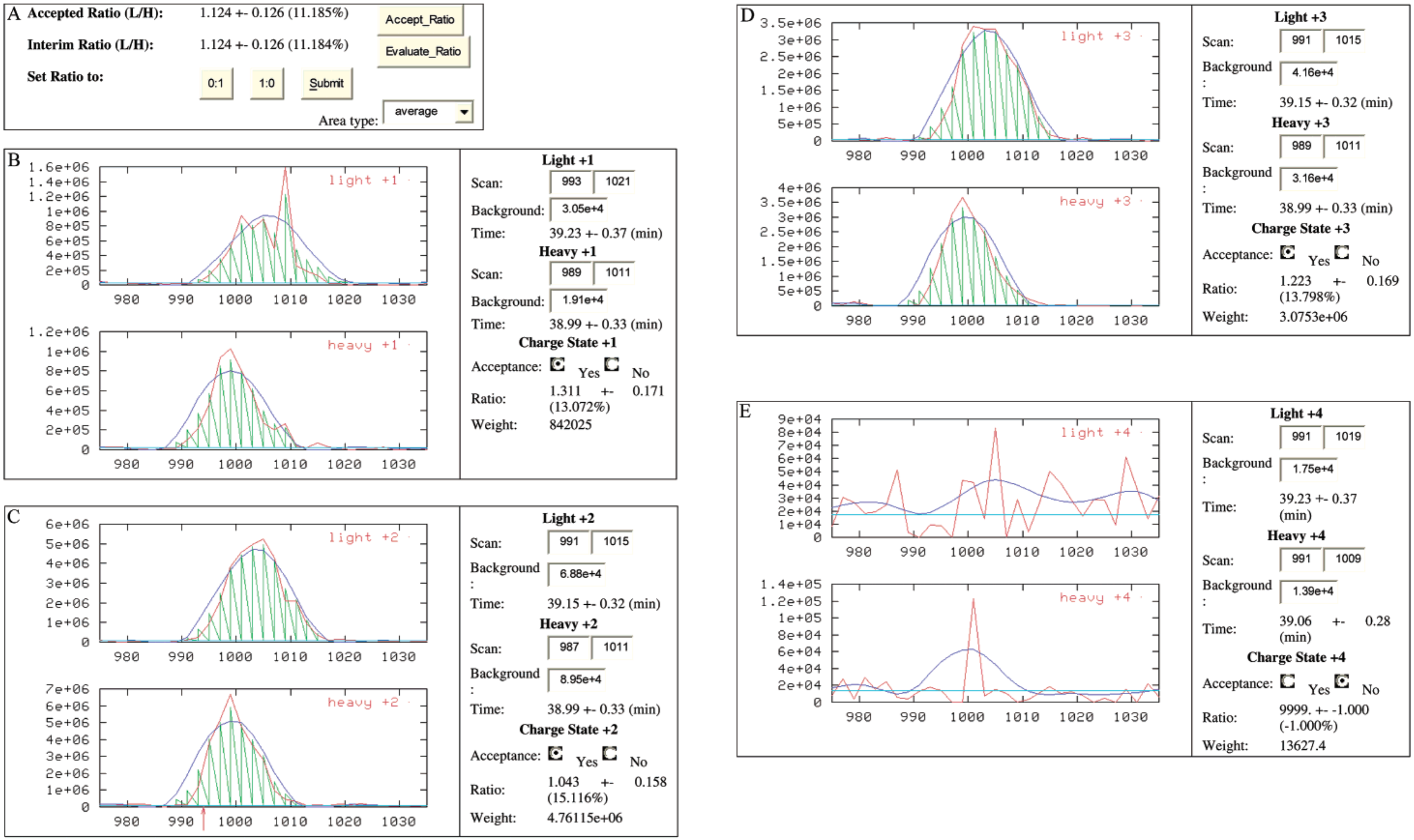
chromatogram. Signals outside the peptide elution peak are assumed to originate from background noise, and their average is accepted as an estimate for the background level in the smoothed chromatogram. A criterion that the apex of the peptide elution peak be at least twice the background level is used to determine whether an acceptable peptide signal was detected. If the elution peak is accepted, its area is calculated from the average of the raw and the smoothed chromatograms, from which the background was subtracted. The error in chromatographic signals is estimated by the signal difference of the raw and the smoothed chromatograms, from which the area error is calculated by using the standard method in error analysis.[25] If the peak is not accepted, the area is set to zero. This process is repeated for the peptide's isotopic partner. Although the partner is not necessarily identified directly by MS/MS and database searching, its m/z values can be determined easily from those of the identified peptide. Similarly, the partner's elution time can be estimated from that of the identified peptide. If a mobility shift is expected between the two partners (e.g., for $^1H/^2H$ labeled peptide pairs), the partner's estimated elution time is shifted accordingly by half of the elution peak width of the identified peptide. The direction of the shift can be easily determined since the isotopic form of the identified peptide is known. Following the same process for the identified peptide, an elution peak can be identified for the partner. In this way, the elution peaks of the two partners always overlap, but the shift between their peak centers is not restricted. If both the peptide and its isotopic partner have acceptable elution peaks, an abundance ratio is calculated as the ratio of the two corresponding elution peak areas, which are calculated from the averages of the raw and the smoothed chromatograms. The ratio error is propagated from the area errors. If one or both of the peak areas were set to zero, the abundance ratio is set to 1:0 or 0:1 or denoted "unquantifiable".

Making use of the fact that a specific peptide is normally observed in more than one charge state during ESI,[18] the ASAPRatio program takes an extra step to identify all potential charge states ranging from $[M + H]^+$ to $[M + 4H]^{4+}$ in which the peptide and its isotopic partner may be present. This is done by checking whether signals corresponding to the theoretical m/z values of the differentially charged ions are detected within the chromatographic window. For each observed charge state, the ASAPRatio program calculates an abundance ratio as already described. All valid abundance ratios from the different charge states are then collected, weighted by the sum of the two corresponding elution peak areas, and used to calculate a peptide abundance ratio and its standard deviation by statistical methods for weighted samples.[25] For each peptide, abundance ratios with weights less than $^1/_{10}$ of the heaviest weight are discarded from the calculation. If there are at least three abundance ratios, Dixon's test[26,27] is applied to eliminate any outliers prior to statistical analysis. The result of step one of the process is a weighted abundance ratio for each observation of an identified peptide. An example illustrating this process is shown in Figure 2.

Step 2: Evaluation of a "Unique Peptide Ratio" for Each Identified Peptide Sequence. In typical LC-ESI-MS/MS experiments, some peptide sequences are identified only once, while others are identified multiple times. Multiple identifications of the same peptide sequence may occur when a particular peptide is split between consecutive chromatographic fractions, when a peptide is detected repeatedly outside the dynamic exclusion window, or when different isotopic forms and/or different charge states of a peptide are identified. Multiple independent observations of the same peptide raise the question of how the peptide's contribution to the abundance ratio of the corresponding protein should be calculated. Since each peptide has a unique abundance ratio before LC-MS/MS analysis, a natural solution to the question is to evaluate this unique abundance ratio, which we will call the "unique peptide ratio", from all the measured peptide abundance ratios of the peptide. The ASAPRatio program determines the unique peptide ratio in two substeps: (1) Peptide abundance ratios of all peptides identified during the same RP elution peak (either in different isotopic forms or in different charge states) are first grouped together to calculate an abundance ratio for the RP peak. (2) Abundance ratios of different RP peaks (either in different chromatographic fractions or at different elution times during the same RP run) are then grouped together to calculate the unique peptide ratio. In each of these two substeps, individual ratios are weighted by the areas of the corresponding RP elution peaks in the most abundant charge states, and the weighted mean and standard deviation are accepted as the ratio and error for the group. If there is only one ratio, its value and error are passed on to the next level. If there are at least three individual ratios in the group, Dixon's test[26,27] is applied to identify outliers whose ratios are disregarded from the calculation. The result of this step of the process is a weighted unique abundance ratio for each identified peptide.

Step 3: Evaluation of Protein Abundance Ratio for Each Identified Protein. The main function of the ASAPRatio program is to evaluate an abundance ratio for each protein identified in an LC-ESI-MS/MS experiment of isotopically labeled samples. If a single peptide is identified for a protein, the corresponding unique peptide ratio and its associated error are passed on as the protein abundance ratio and its error, respectively. Frequently, however, in an experiment more than one unique peptide is identified for the same protein, and for each peptide, a unique peptide ratio is calculated. In this case, statistical methods for weighted samples are applied to calculate the protein abundance ratio and its associated standard deviation from all of its corresponding unique peptide ratios. The unique peptide ratios are weighted by their errors as in standard statistical analysis.[25] If three or more unique peptides are identified for a protein, Dixon's test[26,27] is applied to identify any outlier peptides whose unique peptide ratios are not used in the calculation of the protein abundance ratio. An interface using CGI programming is available for users to verify protein abundance ratios (see Supporting Information). The structure of the CGI generated web page details the hierarchy structure linking each protein abundance ratio with all of its corresponding unique peptide ratios, which in turn are linked with their corresponding peptide abundance ratios. The result of this step of the process is a weighted protein abundance ratio for each identified protein for which at least one peptide has been identified and quantified.

Step 4: Evaluation of the Significance of Abundance Change for Each Identified Protein. A major application of quantitative proteomics is the identification of changes of protein expression in different cell states by accurately measuring the relative abundance of a large number of proteins present in two or more



Accepted Ratio (L/H): 1.124 +- 0.126 (11.185%)

Interim Ratio (L/H): 1.124 +- 0.126 (11.184%)

Set Ratio to:

Area type: average

Light +1
Scan: 993 1021
Background: 3.05e+4
Time: 39.23 +- 0.37 (min)
Heavy +1
Scan: 989 1011
Background: 1.91e+4
Time: 38.99 +- 0.33 (min)
Charge State +1
Acceptance: ● Yes ○ No
Ratio: 1.311 +- 0.171 (13.072%)
Weight: 842025

Light +2
Scan: 991 1015
Background : 6.88e+4
Time: 39.15 +- 0.32 (min)
Heavy +2
Scan: 987 1011
Background : 8.95e+4
Time: 38.99 +- 0.33 (min)
Charge State +2
Acceptance: ● Yes ○ No
Ratio: 1.043 +- 0.158 (15.116%)
Weight: 4.76115e+06

Light +3
Scan: 991 1015
Background : 4.16e+4
Time: 39.15 +- 0.32 (min)
Heavy +3
Scan: 989 1011
Background : 3.16e+4
Time: 38.99 +- 0.33 (min)
Charge State +3
Acceptance: ● Yes ○ No
Ratio: 1.223 +- 0.169 (13.798%)
Weight: 3.0753e+06

Light +4
Scan: 991 1019
Background : 1.75e+4
Time: 39.23 +- 0.37 (min)
Heavy +4
Scan: 991 1009
Background : 1.39e+4
Time: 39.06 +- 0.28 (min)
Charge State +4
Acceptance: ○ Yes ● No
Ratio: 9999. +- -1.000 (-1.000%)
Weight: 13627.4

Figure 2. Graphic interface displaying details of peptide abundance ratio calculation. (A) Control panel for peptide abundance ratio. (B) Details of calculating light/heavy ratio for ions in the $[M + H]^{+}$ charge state. On the left are single-ion LC-MS chromatograms of the isotopically light and heavy peptide partners. Raw chromatograms are plotted in red, smoothed chromatograms in blue, areas used for calculating abundance ratio of the charge state in green, backgrounds in cyan. On the right are the start and the end scan numbers, the background, and the elution time of the isotopically light and heavy peptide partners, the acceptance, the abundance ratio, and the weight of the charge state. Users may change scan numbers, background levels, and the acceptance of the charge state. (C) Same as part B but for ions in the $[M + 2H]^{2+}$ charge state. Here, the red arrow near the bottom of the graphic indicates the isotopic form, the charge state, and the scan number at which the peptide sequence was identified. (D) Same as parts B and C but for ions in the $[M + 3H]^{3+}$ charge state. (E) Same as parts B−D but for ions in the $[M + 4H]^{4+}$ charge state. Here, the background lines stand out more distinctly, and no green lines are plotted since no acceptable peak was detected for the charge state.

samples.[1] In the absence of suitable software tools, protein abundance ratios are typically used to identify differentially expressed proteins without considering the effect of the confidence level of the protein abundance ratios.[5–7] The ASAPRatio program adopts a more sophisticated statistical approach to improve this analysis. It is valid if the expression level of a large number of identified proteins does not change between the two cell states. In this case, a distribution of the logarithm (base 10) of all unique peptide ratios in an LC-ESI-MS/MS experiment is first generated. Assuming the dominant peak in the distribution is attributed to proteins of unchanged abundance, the ASAPRatio program fits the peak automatically with a normal distribution

$$n(r;A,r_0,\sigma) = A \times \exp[-(\log(r/r_0))^2/2\sigma^2] \qquad (1)$$

Here r_0 is the most likely abundance ratio of a protein of unchanged abundance. The validity of using a normal distribution to model the data is justified by the following consideration: Due to natural variation in labeling efficiency, the logarithm (base 10) of unique peptide ratios belonging to proteins of unchanged abundance spreads into a distribution. On the basis of the central limit theorem,[25] this distribution is approximately normal as long as the data set is large. The r_0 value can be used to normalize protein abundance ratios to correct for any systematic errors introduced during sample handling.[21] For a protein abundance ratio r_P and its error Δr_P, the normalized ratio is given by $\hat{r}_P = r_P/r_0$, and its associated error is given by $\Delta\hat{r}_P = \hat{r}_P \sqrt{(\Delta r_P/r_P)^2+(\Delta r_0/r_0)^2}$, where Δr_0 is the fitting error of r_0. The probability of the protein not changing in abundance is described statistically by the p value, which is given by

$$p = \mathrm{erfc}[|\log(r_P/r_0)|/\sqrt{2((\Delta\log r_P)^2 + (\Delta\log r_0)^2 + \sigma^2)}] \qquad (2)$$

where erfc(x) is the complementary error function,[24] $\Delta\log_{10} r_P = 0.4343\Delta r_P/r_P$ and $\Delta\log r_0 = 0.4343\Delta r_0/r_0$. This formula is derived from the normal distribution in eq 1. Its accuracy in describing a particular protein depends on how well the normal distribution in eq 1 fits the overall data of unchanged proteins. Certain p values can be chosen as significant in assessing protein abundance changes. Besides the protein abundance ratio, the error of the protein abundance ratio and the background distribution all affect the p value. A protein with a large ratio may still be considered as not significantly distinguished from the background if its ratio error is too large or if the background distribution is too wide. The evaluation of p values provides a reliable method for assessing the significance of protein abundance changes in large-scale protein profiling experiments and for making data in different quantitative data sets transparently comparable. The result of this final step of the process is a calculated significance of abundance change for each identified protein.

RESULTS

Application of the ASAPRatio Program in Data Analysis of the BSA Experiment. To evaluate the performance of the ASAPRatio program for determining the abundance ratios of isotopically labeled proteins using LC-ESI-MS/MS data, two equal aliquots of tryptic peptides from BSA were labeled separately via either d_0- or d_3-methyl esterification and mixed together in different proportions to generate samples of abundance ratios of 1:1, 1:3, 1:10, 20:1, and 1:100, respectively. The samples were analyzed by LC-MS/MS, and the data were processed through the ASAPRatio software tool. As in most LC-ESI-MS/MS experiments, some peptides of BSA were identified only once, while others were identified multiple times. For example, peptide FKDLGEEHFK was identified six times in the 1:1 ratio sample. Closer examination of the data revealed that all six identifications were generated from the same chromatographic peak. Four of the identifications were identified as the light partner, two in the $[M + H]^+$ charge state, one in $[M + 2H]^{2+}$, and another in $[M + 3H]^{3+}$, while the other two identifications were from the heavy partner, one in the $[M + 2H]^{2+}$ charge state and another in $[M + 3H]^{3+}$, respectively. Independent of the isotopic form or the charge state in which an identification of the peptide was made, the ASAPRatio program first calculated abundance ratios from all observed charge states and then used these abundance ratios to calculate a peptide abundance ratio for each of the six identifications. Since all six identifications were generated from the same chromatographic peak, they all led to the same abundance ratios from three observed charge states, more specifically, 1.43 ± 0.36 from the $[M + H]^+$ charge state, 0.77 ± 0.22 from $[M + 2H]^{2+}$, and 0.90 ± 0.24 from $[M + 3H]^{3+}$. It is hence no surprise that all six identifications produced the same peptide abundance ratio of 1.02 ± 0.24, which was also their unique peptide ratio since the peptide had only one chromatographic peak. The situation was different for peptide DAFLGSFLYEYSR which was identified five times in the 1:1 ratio sample. Four of the identifications were from a peptide with a retention time of 39 min, two each from the light and heavy partners each one being identified in the $[M + 2H]^{2+}$ and the $[M + 3H]^{3+}$ charge states. These had the same peptide abundance ratios of 1.12 ± 0.13. The remaining identification was from a heavy labeled peptide in the $[M + 2H]^{2+}$ charge state at a retention time of 42 min. The calculated peptide abundance ratio was 0.90 ± 0.14. These two chromatographic peaks were well distinguished from their background noise and clearly separated in time. In this case, the peptide DAFLGSFLYEYSR was detected at different elution times from the RP column. This phenomenon is not uncommon and is usually explained by the formation of a different secondary structure under certain solvent conditions or the interaction with other peptides. A unique peptide ratio of 1.04 ± 0.18 was calculated for the peptide from the abundance ratios of the two chromatographic peaks.

All unique peptide ratios obtained from the BSA experiment are plotted in Figure 3. The peptides from the nominally 1:1, 1:3, 1:10, 20:1, and 1:100 sample mixtures are shown in Figure 3A–E, respectively. The data clearly demonstrate the power of statistical analysis. Different unique peptide ratios of peptides with the same nominal abundance ratio show a typical dispersion. Most unique peptide ratios in each sample overlapped with each other within one standard deviation. The corresponding protein abundance ratio was at the center of the population, and the standard deviation of this value was generally smaller than that of the individual unique peptide ratios. Even the presence of a few outlier data points did not affect the quality of the final result.

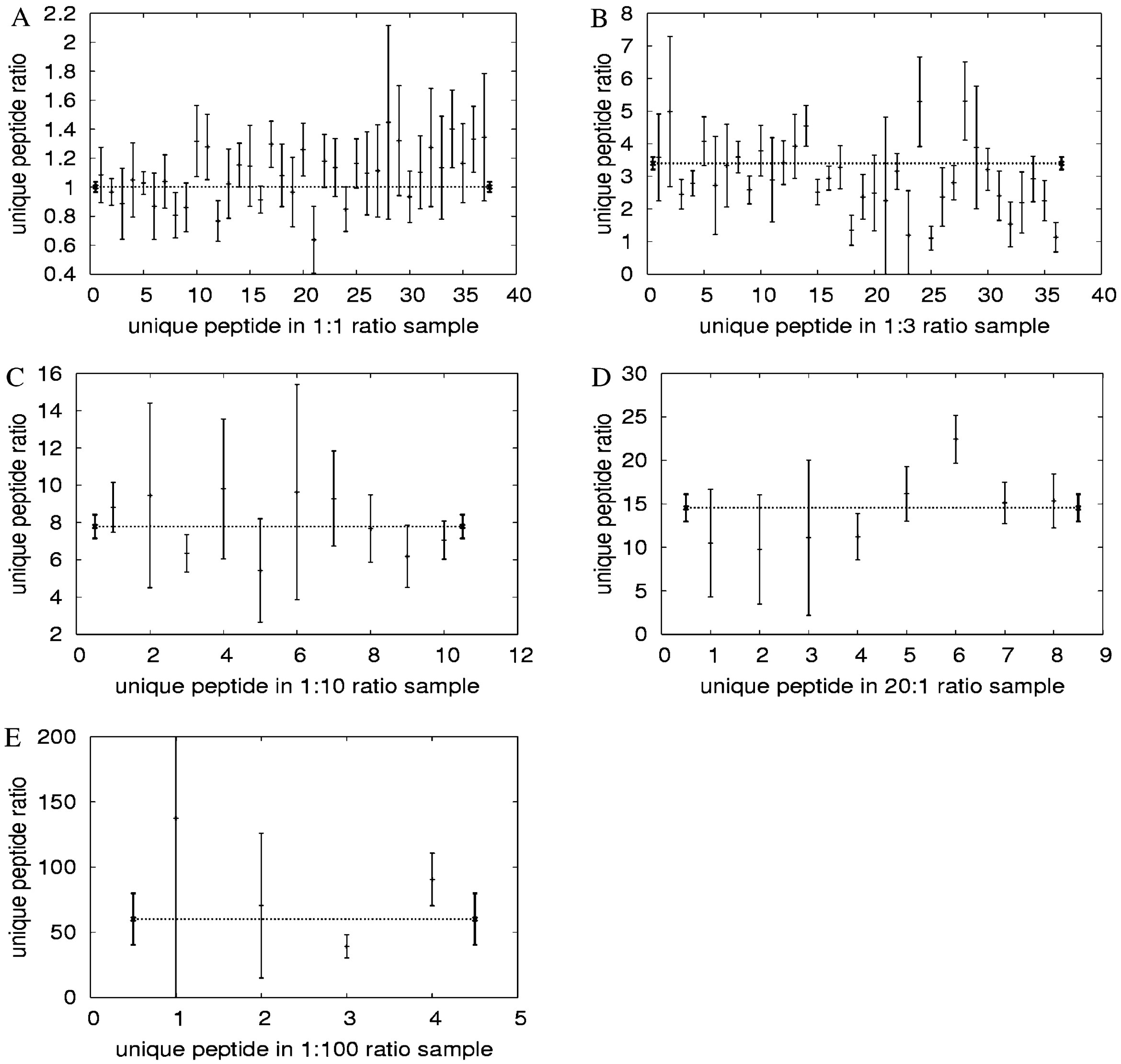


Figure 3. Unique peptide ratios and their errors in the BSA experiment. The expected BSA light/heavy abundance ratios are (A) 1:1, (B) 1:3, (C) 1:10, (D) 20:1, and (E) 1:100. The ratios are plotted as light/heavy in parts A and D, but as heavy/light in parts B, C, and E. The horizontal dashed lines are the obtained protein abundance ratios with their error bars attached at the two ends.

Table 1. Results on BSA Abundance Ratios Determined by the ASAPRatio Program[a]

format[b]	expected	ratio ± error	relative error[c] (%)	unique peptides
L/H	1	1.00 ± 0.04	0	37
H/L	3	3.40 ± 0.19	13	36
H/L	10	7.79 ± 0.63	22	10
L/H	20	14.54 ± 1.56	27	8
H/L	100	60.14 ± 19.80	40	4

[a] All peptides were filtered with Xcorr (1.5 for $[M + H]^+$; 2 for $[M + 2H]^{2+}$; 2.5 for $[M + 3H]^{3+}$) and δCn (0.1). [b] L/H stands for light/heavy ratio while H/L for heavy/light. [c] Between expected and obtained ratios.

The protein abundance ratios obtained by the ASAPRatio program for the five BSA samples are summarized in Table 1. The relative error between measured and expected values ranged from 0% for the 1:1 ratio sample to 40% for the 1:100 ratio sample, with an average error of 20%. This accuracy is reasonable considering the dynamic range of sample concentrations involved. The relative error increased monotonically with the relative abundance of the two isotopic partners. Several factors may have contributed to the increase of relative errors: (1) The MS signal for the more abundant isotopic partner could be somewhat suppressed. This may reflect a property of the detector used to understate the size of signals approaching detector saturation. (2) It was possible that background noise was not removed completely with the Savitzky−Golay smooth filtering method.[30] As a result, the MS signal for the less abundant isotopic partner was overestimated. Note that the ASAPRatio program obtained a reasonable result even when the light/heavy ratio was 1:100. In comparison, more than 10-fold differences in protein abundance changes have been rarely reported as reliable in the literature of MS technology.[5−7,12,15] The program's capability of measuring a 100-fold difference between samples is due in part to the subtraction of background noise from single-ion chromatograms.

Six peptides were identified as outliers by the ASAPRatio program for the BSA samples. Three of them were misidentified

[30] Perrin, C.; Walczak, B.; Massart, D. L. *Anal. Chem.* **2001**, *73*, 4903−4917.

peptides. These included peptides KLKECCDKPLLEK and KDLGEEHFK from the 1:1 ratio sample, and peptide KDLGEEHFK from the 1:100 ratio sample. For example, peptide KLKECCDKPLLEK was identified only once in the light isotopic form and in the $[M + 2H]^{2+}$ charge state. It had SEQUEST scores Xcorr = 2.3621 and δCn = 0.163, two miscleavage sites, and a unique peptide ratio of 1:0. The remaining three peptides did not have valid unique peptide ratios due to very noisy chromatograms. These included peptide FVEVTK from the 1:3 ratio sample, peptide LGEYGFQNALIVR from the 20:1 ratio sample, and peptide VHKECCHGDLLECADDR from the 1:100 ratio sample. For example, peptide VHKECCHGDLLECADDR was identified twice in the heavy isotopic form and in the $[M + 2H]^{2+}$ (with Xcorr = 2.0318 and δCn = 0.411) and $[M + 3H]^{3+}$ (with Xcorr = 2.3665 and δCn = 0.265) charge states. It had a unique peptide ratio of 0:1 since the signals in the chromatogram of its light isotopic form were near noise level. By eliminating outlier peptides from the evaluation of protein abundance ratios, the ASAPRatio program is resilient against misidentifications and data of poor quality. These data indicate that the ASAPRatio program accurately determines abundance ratios over the full dynamic range of the mass spectrometer used to generate the data.

Application of the ASAPRatio Program for the Analysis of the RNA Pol II Transcription Complex. In this experiment, RNA Pol II transcription complexes and a control sample mainly consisting of proteins nonspecifically binding to the DNA template were isolated from a nuclear extract via DNA affinity chromatography using immobilized promoter DNA templates. The samples were labeled with the isotopically light and heavy ICAT reagent, respectively, and analyzed by LC-ESI-MS/MS as described.[6] The abundance ratios were used to identify the proteins that were specifically enriched in the sample containing the functional complex compared to the control and thus to differentiate between specific components of the core RNA Pol II transcription complex and a background of copurifying proteins.[6] Prior to the development of the ASAPRatio program, the data were analyzed by the XPRESS tool and manually verified, a task that consumed about a week. In contrast, the same data were analyzed by the ASAPRatio program in about 15 min without human manipulation. Among the 1932 peptides identified from the database search, hundreds were determined to be unquantifiable by manual interpretation of the XPRESS output. In contrast, only 13 peptides were determined unquantifiable by the ASAPRatio program. The heavy/light ratios of 723 peptides quantified by both methods are plotted against each other in Figure 4A. Among the 576 peptides for which the XPRESS ratios were between 0.5 and 2, the average difference between the XPRESS ratios and their corresponding ASAPRatio ratios was 13%. About 80% of the 576 manually validated XPRESS ratios agreed within 20% with the corresponding ASAPRatio ratios. Furthermore, about 77% of the 576 XPRESS ratios were within one standard deviation of the ASAPRatio ratios. The distribution of z scores, defined as $z = (r_A - r_X)/\Delta r$ in which r_X is the peptide abundance ratio determined by the XPRESS tool and r_A and Δr are the corresponding peptide abundance ratio and its standard deviation determined by the ASAPRatio program, was plotted in Figure 4B. Clearly the majority of data had $|z| < 1$. Collectively, these data indicate that the ASAPRatio program is capable of generating reliable peptide abundance ratios automatically. Among the 9 peptides for which the XPRESS ratios were less than 0.5, 4 of the ASAPRatio ratios were smaller than the corresponding XPRESS ratios. Among the 138 peptides for which the XPRESS ratios were greater than 2, about 75% of the ASAPRatio ratios were larger than the corresponding XPRESS ratios, see the inset of Figure 4A. The largest and smallest ratios determined by the XPRESS tool were 50.00 and 0.45, respectively. In comparison, the corresponding values determined by the ASAPRatio program were 157.06 and 0.21, respectively. This increase of dynamic range provided by the ASAPRatio program was due to the fact that it subtracted background in the single-ion chromatograms.

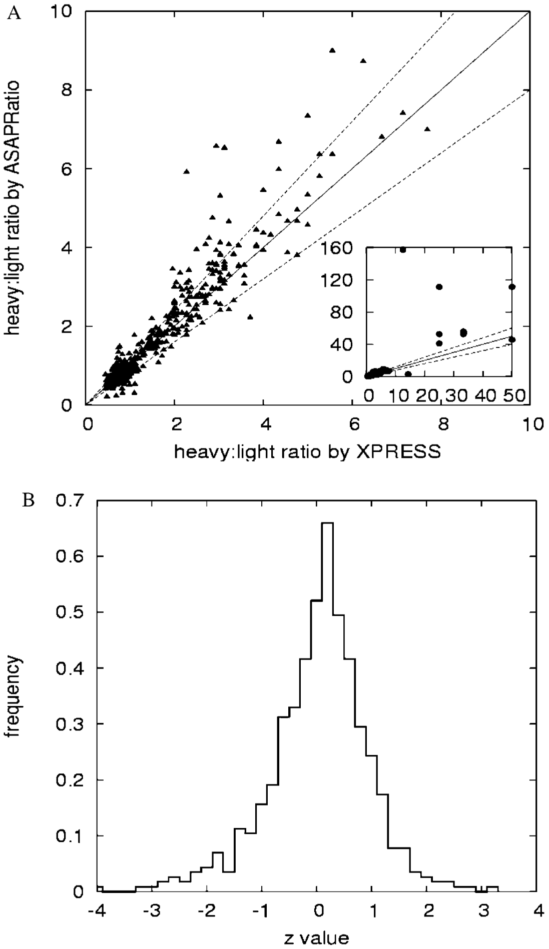


Figure 4. (A) Comparison of peptide heavy/light abundance ratios obtained by the XPRESS tool and manual validation and by the ASAPRatio program for data in RNA Pol II transcription complex experiment. The solid line ($x = y$) is plotted as a guide for full agreement while the two dashed lines indicate a range of 20% difference. The inset is same as the main graph except that it represents the whole data set. (B) Distribution of z score for the same data as in part A but with XPRESS ratios between 0.5 and 2.

Among all identified peptides in this experiment, 1857 peptides had acceptable chromatographic peaks in both the light and the heavy isotopic forms. By monitoring the presence of charge states ranging from $[M + H]^{+}$ to $[M + 4H]^{4+}$ for these 1857 peptides, frequently multiple charge states per peptide were detected. The results from the analysis of the charge state distribution are summarized in Table 2. The left panel illustrates the presence of the charge state distribution of the 1857 peptides identified in the

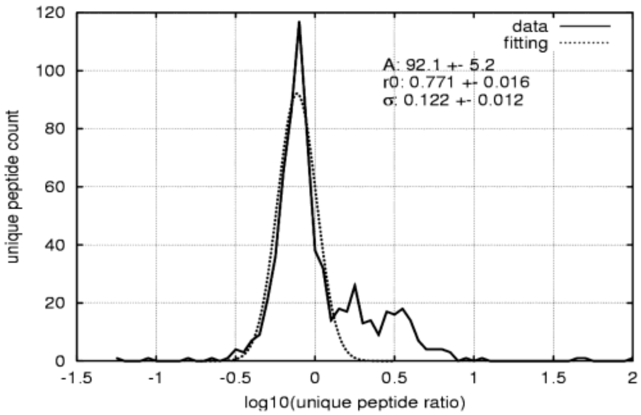


Figure 5. Distribution of heavy/light unique peptide ratios in RNA Pol II transcription complex experiment. The solid line is original data, and the dotted line is the fitting curve of a normal distribution over the main peak.

Table 2. Charge State Distribution of 1857 Peptides Identified in the Pol II Transcription Complex Experiment

number of peptides in individual charge states			number of peptides in multiple charge states simultaneously		
charge state	peptide no.	peptide %	state no.	peptide no.	peptide %
$[M + H]^{+}$	530	28.5	1	272	14.6
$[M + 2H]^{2+}$	1649	88.8	2	1086	58.5
$[M + 3H]^{3+}$	1477	79.5	3	494	26.6
$[M + 4H]^{4+}$	290	15.6	4	5	0.3
total	3946	212.5	total	1857	100

experiment. Most peptides were present in the $[M + 2H]^{2+}$ (88.8%) or $[M + 3H]^{3+}$ (79.5%) charge state. Since a peptide is frequently present in more than one charge state, the sum of the recorded events is greater than 1857, the number of peptides identified. On average, a peptide was present in 2.1 different charge states. The right panel shows the presence of the 1857 peptides in various number of different charge states. Most peptides (58.5%) were present in two different charge states while a small fraction (14.6%) was present in one charge state only. Among the abundance ratios calculated by the ASAPRatio program for the different charge states of the same peptide, about 50% differed by less than 20% while about 20% differed by more than 50%. Clearly most abundance ratios of the same peptide but from different charge states were close to each other, but a certain degree of disagreement also existed, which contributed uncertainties to peptide abundance ratios. By evaluating abundance ratios from different charge states and using them to calculate peptide abundance ratios, the ASAPRatio program provides a more reliable method to evaluate peptide abundance ratios than the conventional approach of evaluating peptide abundance ratios from each identified charge state independently.

Table 3. Proteins with a Significant Increase in Heavy/Light Abundance Ratio in RNA Pol II Transcription Complex Experiment

ORF	gene	D8/D0[a]	error[a]	*p* value	Pol II transcription
YKL058W	TOA2	67.961	18.582	2.84×10^{-27}	core[b]
YPL082C	MOT1	6.729	0.549	4.79×10^{-11}	core
YOR194C	TOA1	8.479	1.468	6.77×10^{-11}	core
YPR086W	SUA7	6.766	0.739	1.57×10^{-10}	core
YOL135C	MED7	5.069	0.407	2.06×10^{-08}	core
YLR071C	RGR1	4.748	0.474	1.35×10^{-07}	core
YGL151W	NUT1	4.534	0.318	1.36×10^{-07}	core
YNL236W	SIN4	4.217	0.298	5.28×10^{-07}	core
YGR005C	TFG2	5.219	1.189	4.31×10^{-06}	core
YOR151C	RPB2	3.777	0.303	4.40×10^{-06}	core
YGR186W	TFG1	6.234	1.780	4.41×10^{-06}	core
YDL140C	RPO21	3.838	0.381	5.26×10^{-06}	core
YIL021W	RPB3	3.808	0.362	5.38×10^{-06}	core
YPR025C	CCL1	4.366	0.730	5.58×10^{-06}	core
YBR253W	SRB6	4.607	0.906	7.33×10^{-06}	core
YHR041C	SRB2	3.562	0.332	1.48×10^{-05}	core
YKL028W	TFA1	3.883	0.568	1.58×10^{-05}	core
YDL108W	KIN28	4.619	1.043	1.94×10^{-05}	core
YDR443C	SSN2	4.965	1.281	2.40×10^{-05}	core
YGL070C	RPB9	3.603	0.547	5.22×10^{-05}	core
YDL005C	MED2	5.057	1.469	5.56×10^{-05}	core
YJL140W	RPB4	4.601	1.195	6.03×10^{-05}	core
YOR174W	MED4	4.240	1.016	8.28×10^{-05}	core
YDR460W	TFB3	3.703	0.748	1.40×10^{-04}	core
YPR070W	MED1	4.312	1.228	2.41×10^{-04}	core
YCR081W	SRB8	3.153	0.624	7.68×10^{-04}	core
YER171W	RAD3	2.568	0.180	1.01×10^{-03}	core
YKR062W	TFA2	3.934	1.268	1.29×10^{-03}	core
YDR404C	RPB7	3.225	0.767	1.37×10^{-03}	core
YOL005C	RPB11	3.255	0.793	1.41×10^{-03}	core
YPL038W	MET31	2.566	0.350	2.34×10^{-03}	role[c]
YIL143C	SSL2	2.751	0.525	2.70×10^{-03}	core
YMR005W	TAF48/MPT1	2.340	0.259	4.51×10^{-03}	core
YCR042C	TSM1	2.260	0.170	4.66×10^{-03}	core
YFR031C	SMC2	2.245	0.224	6.19×10^{-03}	unknown[d]
YOL051W	GAL11	3.170	1.023	6.76×10^{-03}	core
YPR056W	TFB4	2.347	0.371	7.65×10^{-03}	core
YLR005W	SSL1	2.234	0.280	8.44×10^{-03}	core
YER155C	BEM2	2.300	0.382	1.01×10^{-02}	unknown
YKR001C	TOF2	6.004	4.008	1.32×10^{-02}	unknown
YGR274C	TAF145	2.017	0.192	1.70×10^{-02}	core
YGL112C	TAF60	2.416	0.593	1.74×10^{-02}	core
YDR080W	VPS41	2.152	0.372	1.92×10^{-02}	unknown
YBR198C	TAF90	1.953	0.193	2.33×10^{-02}	core
YPL122C	TFB2	2.366	0.627	2.49×10^{-02}	core
YDR145W	TAF61	2.166	0.454	2.65×10^{-02}	core
YER148W	SPT15	37.865	62.569	3.01×10^{-02}	core
YPL011C	TAF47	1.825	0.257	5.38×10^{-02}	core
YDR311W	TFB1	2.164	0.625	5.43×10^{-02}	core
YLR442C	SIR3	2.649	1.153	5.93×10^{-02}	role
YER022W	SRB4	2.278	0.768	5.95×10^{-02}	core
YER133W	GLC7	2.127	0.617	6.03×10^{-02}	unknown
YDR079C-A	TFB5	1.868	0.363	6.57×10^{-02}	core
YPL133C	RDS2	2.108	0.651	7.28×10^{-02}	role
YMR236W	TAF17	2.232	0.851	8.87×10^{-02}	core
YLR055C	SPT8	2.195	0.825	9.26×10^{-02}	role
YER164W	CHD1	2.244	0.880	9.28×10^{-02}	role
YMR227C	TAF7	1.906	0.774	1.90×10^{-01}	core
YML015C	TAF11	1.439	0.201	2.42×10^{-01}	core
YLR399C	BDF1	1.836	0.992	3.17×10^{-01}	core

[a] Normalized. [b] Core components of Pol II transcription. [c] With known role in Pol II transcription. [d] With no previously known role in Pol II transcription.

In this experiment, core components of the Pol II transcription complex were enriched in the sample containing the functional complex compared to the control sample and therefore had larger protein heavy/light abundance ratios than copurifying proteins.[6] To demonstrate the capability of the ASAPRatio program in large-scale protein profiling, the distribution of the logarithm (base 10) of all heavy/light unique peptide ratios in the experiment was fitted with a normal distribution as described in eq 1 with fitting parameters $A = 92.1$, $r_0 = 0.771$, $\sigma = 0.122$, and $\Delta r_0 = 0.016$ (see Figure 5). Protein abundance ratios and their standard deviations were normalized accordingly, and *p* values were evaluated for all quantifiable proteins, using eq 2. Among 262 quantifiable proteins, 57 passed the filtering of $\hat{r}_P > 1$ and $p < 0.1$ (see Table 3), where $\hat{r}_P$ is the normalized protein abundance ratio. Among these 57 proteins, 47 were known components of the core Pol II transcription complex, 5 had a known role in Pol II transcription, and the remaining 5 had no previously known role in Pol II transcription. There are a total of 60 known core Pol II components that were potentially identifiable; 47 (78%) of them were selected by the filtering of $\hat{r}_P > 1$ and $p < 0.1$. One core component (SSN2) was previously considered to be unquantifiable[6] but was clearly identified by the ASAPRatio program. The smallest *p* value among the 5 proteins with no previously known role in Pol II transcription was ranked 35 in Table 3 in which the identified proteins are listed in order of increasing *p* values. Three known components of the core Pol II transcription complex that did not pass the filtering of $\hat{r}_P > 1$ and $p < 0.1$ are also included in Table 3. Two of these (TAF7 and BDF1) had very large errors in their protein abundance ratios while the third (TAF11) did not have a large protein abundance ratio. By evaluating protein *p* values to specify the significance of protein abundance changes, the ASAPRatio program clearly has the capability of promptly identifying biologically interesting proteins from a very large background.

DISCUSSION

We describe here an algorithm for automated protein abundance analysis and protein profiling based on stable-isotope labeling and LC-ESI-MS/MS data. Two sets of data were analyzed by the software ASAPRatio without human intervention. As indicated by the analysis of BSA esterification data, the ASAPRatio program is capable of obtaining reliable protein abundance ratios over a large dynamic range. As demonstrated in the analysis of RNA Pol II transcription complex data, the ASAPRatio program provides an effective statistical tool for distinguishing proteins of significant abundance changes from proteins of unchanged abundance.

Several features are unique to the ASAPRatio program: (1) It evaluates protein *p* values to provide a statistical criterion in distinguishing proteins of significant abundance changes from the typically large population of proteins of unchanged abundance. This feature is essential for analyzing the data from large-scale protein profiling experiments. (2) The ASAPRatio program evaluates the associated standard deviation of each obtained abundance ratio. The standard deviation provides an assessment of the reliability of the ratio. (3) The ASAPRatio program calculates a peptide abundance ratio not only from the charge state of the peptide that is identified but also from all other charge states for which a signal is detected. The presence of most peptides in multiple charge states makes statistical analysis on protein abundance ratio possible even when only a single peptide is identified for a protein. (4) The ASAPRatio program connects the evaluation of protein abundance ratios with the experimental procedure generating the data. To avoid the situation in which a protein abundance ratio is biased toward the ratios from its most frequently identified peptides, the ASAPRatio program calculates intermediate unique peptide ratios before calculating the protein abundance ratio from those unique peptide ratios. (5) The ASAPRatio program subtracts background noise from MS signals when calculating areas of single-ion chromatograms. This has the advantage of increasing the accuracy and the dynamic range of detectable changes in peptide abundance ratios. (6) The ASAPRatio program is capable of analyzing LC-ESI-MS/MS data generated by many types of isotopic tags. However, the mass difference between isotopic tags should be large enough so that MS signals of isotopic peptide partners do not overlap. Normally a mass difference less than 6 Da between isotopic peptide partners may cause errors in the evaluation of the peptide abundance ratio and a reduction of the dynamic range. (7) The ASAPRatio program uses Dixon's test to identify outlier data points and eliminate them from the evaluation process and hence limit the affects of a few "bad" data points. The ability to reliably identify outlier peptides may be biologically significant because such peptides may not only be caused by misidentification or low quality data but, in fact, may indicate the presence of post-translational modification, differential splicing, or other mRNA or protein processing events. On the basis of these unique features, the ASAPRatio program greatly reduces the burden of manually analyzing large-scale LC-ESI-MS/MS data sets, a bottleneck in current MS-based quantitative proteomics experiments. The current version of the ASAPRatio program runs on a Linux operating systems and will be available to the public upon request. A new version for the Windows operating system is also planned and will be freely distributed under an open source license. More information on the ASAPRatio program is available at http://www.proteomecenter.org/software.php.

ACKNOWLEDGMENT

We appreciate many fruitful discussions with Mr. Jimmy Eng and Dr. David Goodlett. We thank Dr. Julian Watts, Dr. Wei Yan, Ms. Rose Boyle, Dr. Dan Martin, and Ms. Sharon Chen for their comments on the software and the manuscript. This project has been funded in part with federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, under Contract N01-HV-28179 and by a sponsored research agreement from Oxford GlycoSciences.

SUPPORTING INFORMATION AVAILABLE

Brief discussion of CGI generated interface for the ASAPRatio program and CGI generated interface for the evaluation of protein abundance ratio. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review June 10, 2003. Accepted September 17, 2003.

AC034633I