Anal. Chem. **2004,** *76,* 4193−4201

A Model for Random Sampling and Estimation of Relative Protein Abundance in Shotgun Proteomics

Hongbin Liu,[†,§,||] Rovshan G. Sadygov,[†,§] and John R. Yates, III*,[†,‡]

Department of Cell Biology, The Scripps Research Institute, La Jolla, California 92037, and Diversa, Inc., Directors Place, San Diego, California 92037

Proteomic analysis of complex protein mixtures using proteolytic digestion and liquid chromatography in combination with tandem mass spectrometry is a standard approach in biological studies. Data-dependent acquisition is used to automatically acquire tandem mass spectra of peptides eluting into the mass spectrometer. In more complicated mixtures, for example, whole cell lysates, data-dependent acquisition incompletely samples among the peptide ions present rather than acquiring tandem mass spectra for all ions available. We analyzed the sampling process and developed a statistical model to accurately predict the level of sampling expected for mixtures of a specific complexity. The model also predicts how many analyses are required for saturated sampling of a complex protein mixture. For a yeast-soluble cell lysate 10 analyses are required to reach a 95% saturation level on protein identifications based on our model. The statistical model also suggests a relationship between the level of sampling observed for a protein and the relative abundance of the protein in the mixture. We demonstrate a linear dynamic range over 2 orders of magnitude by using the number of spectra (spectral sampling) acquired for each protein.

Proteomics involves the study of complex systems of proteins and thus is highly dependent on analytical technology. Two-dimensional gel electrophoresis is a highly resolving separation technique often used to fractionate protein mixtures prior to identification of the proteins using mass spectrometry. The analyzed proteins are either homogeneous or simple mixtures of proteins and thus are straightforward to analyze. An alternate strategy, "shotgun proteomics", reduces reliance on protein fractionation and employs digestion of the protein mixtures to produce a collection of peptides analyzed by on-line mass spectrometry.[1] Two mass spectrometry strategies have emerged to analyze the peptides by mass spectrometry. Smith and colleagues use high mass accuracy mass measurements to identify peptides, generally after establishing tandem mass spectrometry identification and chromatographic retention times for peptides in the mixture.[2,3] Provided protein expression and proteolytic digestion are consistent between experiments, this method can produce high-throughput protein identifications. Eng et al. established the use of tandem mass spectrometry together with liquid chromatography to identify proteins in mixtures using tandem mass spectral database searching.[4] McCormack et al. extended this approach to the analysis of protein complexes and Link et al. used multidimensional liquid chromatography for the analysis of large complexes and whole cell lysates.[5,6] Washburn et al. improved and extended the approach for the analysis of whole yeast cells and enriched membrane fractions.[7] A subset of this approach is to reduce the complexity of the mixture by targeted enrichment of specific peptides (e.g., peptides containing Cys or His).[8−11] Shotgun proteomic analysis is now commonly used to identify proteins in biological experiments to identify protein localization, protein expression, protein complexes, and protein modifications.[12−18]

* Corresponding author. Phone: (858)784-8862. Fax: (858) 784-8883. E-mail: jyates@scripps.edu.

[†] The Scripps Research Institute.

[‡] Diversa, Inc.

[§] These two authors contributed equally to this work.

[||] Current address: Agilent Technologies, Inc., Wilmington, DE 19808.

(1) Yates, J. R., 3rd *J. Mass Spectrom.* **1998**, *33*, 1−19.

(2) Smith, R. D.; Anderson, G. A.; Lipton, M. S.; Masselon, C.; Pasa-Tolic, L.; Udseth, H.; Belov, M.; Shen, Y.; Veenstra, T. D. *Adv. Protein Chem.* **2003**, *65*, 85−131.

(3) Strittmatter, E. F.; Ferguson, P. L.; Tang, K.; Smith, R. D. *J. Am. Soc. Mass Spectrom.* **2003**, *14*, 980−991.

(4) Eng, J. K.; McCormack, A. L.; Yates, J. R., III. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976−989.

(5) McCormack, A. L.; Schieltz, D. M.; Goode, B.; Yang, S.; Barnes, G.; Drubin, D.; Yates, I., J. R. *Anal. Chem.* **1997**, *69*, 767−776.

(6) Link, A. J.; Eng, J.; Schieltz, D. M.; Carmack, E.; Mize, G. J.; Morris, D. R.; Garvik, B. M.; Yates, J. R., 3rd *Nat. Biotechnol.* **1999**, *17*, 676−682.

(7) Washburn, M. P.; Wolters, D.; Yates, J. R., 3rd *Nat. Biotechnol.* **2001**, *19*, 242−247.

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

(9) Geng, M.; Ji, J.; Regnier, F. E. *J. Chromatogr., A* **2000**, *870*, 295−313.

(10) Zhang, R.; Sioma, C. S.; Wang, S.; Regnier, F. E. *Anal.Chem.* **2001**, *73*, 5142−5149.

(11) Ficarro, S. B.; McCleland, M. L.; Stukenberg, P. T.; Burke, D. J.; Ross, M. M.; Shabanowitz, J.; Hunt, D. F.; White, F. M. *Nat. Biotechnol.* **2002**, *20*, 301−305.

(12) Schirmer, E. C.; Florens, L.; Guan, T.; Yates, J. R., 3rd; Gerace, L. *Science* **2003**, *301*, 1380−1382.

(13) Cheeseman, I. M.; Anderson, S.; Jwa, M.; Green, E. M.; Kang, J.; Yates, J. R., 3rd; Chan, C. S.; Drubin, D. G.; Barnes, G. *Cell* **2002**, *111*, 163−172.

(14) Florens, L.; Washburn, M. P.; Raine, J. D.; Anthony, R. M.; Grainger, M.; Haynes, J. D.; Moch, J. K.; Muster, N.; Sacci, J. B.; Tabb, D. L.; Witney, A. A.; Wolters, D.; Wu, Y.; Gardner, M. J.; Holder, A. A.; Sinden, R. E.; Yates, J. R.; Carucci, D. J. *Nature* **2002**, *419*, 520−526.

(15) Peng, J.; Schwartz, D.; Elias, J. E.; Thoreen, C. C.; Cheng, D.; Marsischky, G.; Roelofs, J.; Finley, D.; Gygi, S. P. *Nat. Biotechnol.* **2003**, *21*, 921−926.

(16) 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.

10.1021/ac0498563 CCC: $27.50 © 2004 American Chemical Society
Published on Web 06/02/2004

Data acquisition using tandem mass spectrometry is computer-controlled and is generally triggered by ion abundance levels. In the separation of complex peptide mixtures co-elution is prevalent and thus the selection of ions for MS/MS is dependent on the width of the chromatographic peaks or the concentration of peptides delivered to the mass spectrometer. In very complex peptide mixtures the number of ions co-eluting can significantly exceed the number of ions for which tandem mass spectra can be acquired. Data acquisition can thus be biased against the low-abundance ion signals that are often peptides present at low levels. Washburn et al. noted in the study of yeast whole cells that abundant proteins were identified with multiple peptides and low-abundant proteins by one or two.[7] Attempts to improve sampling efficiency have employed narrow mass range scans to "gas-phase" fractionate ions or enrichment of peptides containing specific amino acid residues.[8,9,19] In the gas-phase fractionation approach the analysis needs to be repeated several times to cover a typical mass range, but potentially can improve the dynamic range of the analysis and the acquisition of low-abundance peptide ion signals.

Improving chromatographic resolution can limit the level of co-elution that occurs and this has been accomplished with some success through the use of multidimensional liquid chromatography or high-pressure liquid chromatography.[7,20−22] The chromatographic separation process employing multidimensional separations has been shown to be fairly reproducible, although the collection of tandem mass spectra and thus the protein identification process has been observed to be less reproducible for complex protein mixtures.[20] Additionally, ionization effects may play a role in the sampling of specific peptide ions. Bodnar et al. observed differences in the proteins identified depending on the mode of ionization used for peptide analysis.[23] By splitting the separation of a complex peptide mixture to send a portion of the mixture to electrospray ionization for MS/MS analysis and depositing the remaining sample on a MALDI sample plate for MALDI/MS/MS analysis, only 63% of the proteins identified overlap. In both analyses unique peptides and proteins were identified, suggesting the ionization methods exhibit different selectivities for the peptide amino acid sequence or composition. Integral to this observation is the selective nature of data-dependent data acquisition as, in general, ions are selected based on abundance levels. If an ionization technique produces ions more efficiently for certain peptide characteristics, then those ions will have a higher probability of being selected for tandem mass spectrometry.

Two issues arise in the analysis of complex mixtures. The first issue is the level of randomness observed in data-dependent acquisitions of peptide ions in complex peptide mixtures. How random is the sampling process? The second issue is related to the probability peptide ions from more abundant proteins would be selected more frequently, as observed by Washburn et al., and thus reflect protein abundance.[7] If more frequent sampling reflects actual protein abundance, how accurate is the measurement? Prior research has suggested a relationship between protein abundance and the sampling process but has expressed these measurements in different ways and provided no validation. Pang et al. and Gao et al. have suggested that peptide hits and spectral count are related to protein abundance.[21,24] Florens et al. used protein sequence coverage as a "loose" measure of protein abundance in an analysis of protein expression in *P. falciparum*.[14] The two concepts are similar but sequence coverage for a protein is dependent on the length of a protein and digestion pattern (e.g., number of Lys and Arg). Wang et al. used normalized peptide signal intensities to quantify peptide signals and thus proteins between two different samples.[25] More traditional methods determine peak areas of isotopomers using isotope dilution mass spectrometry to measure differences between two samples.[8,26,27] To better understand the relationship between sample complexity and data acquisition, we studied the acquisition of tandem mass spectra in a complex protein system. From these experiments we modeled the randomness of data acquisition and validated the use of spectral sampling as a method to measure relative abundance of proteins in complex systems. These data are compared to a recently published study of total protein expression in *S. cerevisiae*.[28]

MATERIAL AND METHODS

Materials. A micro BCA reagent kit was obtained from Pierce (Rockford, IL). SDS-PAGE molecular weight standards (low range) were obtained from Bio-Rad (Hercules, CA). Poroszyme immobilized trypsin was obtained from Applied Biosystems (Framingham, MA). Endoproteinase Lys-C was obtained from Roche Diagnostics (Indianapolis, IN). Other standard laboratory chemicals were obtained from Sigma (St. Louis, MO).

Yeast protease-deficient strain BJ5460 was grown to log phase (OD 0.9) in YPD at 30° C, then harvested, and washed by DDI water. Yeast cells were lysed by grinding a frozen cell pellet under liquid nitrogen. Proteins were extracted using digestion buffer (100 mM NH_4HCO_3, 8 M urea, pH 8.5). The soluble fraction was obtained by centrifuging total cell lysates at 21000*g* for 2 h.

MS/MS Data Collection and *Pep_Prob* Database Search. Four protein samples, 400 μg each, were prepared by mixing a yeast total cell lysate-soluble fraction with Bio-Rad SDS-PAGE low-range weight standards consisting of the following proteins: phosphorylase b, serum albumin, ovalbumin, carbonic anhydrase,

(17) Lasonder, E.; Ishihama, Y.; Andersen, J. S.; Vermunt, A. M.; Pain, A.; Sauerwein, R. W.; Eling, W. M.; Hall, N.; Waters, A. P.; Stunnenberg, H. G.; Mann, M. *Nature* **2002**, *419*, 537−542.

(18) Andersen, J. S.; Wilkinson, C. J.; Mayor, T.; Mortensen, P.; Nigg, E. A.; Mann, M. *Nature* **2003**, *426*, 570−574.

(19) Davis, M. T.; Spahr, C. S.; McGinley, M. D.; Robinson, J. H.; Bures, E. J.; Beierle, J.; Mort, J.; Yu, W.; Luethy, R.; Patterson, S. D. *Proteomics* **2001**, *1*, 108−117.

(20) Wolters, D. A.; Washburn, M. P.; Yates, J. R., 3rd *Anal. Chem.* **2001**, *73*, 5683−5690.

(21) Gao, J.; Opiteck, G. J.; Friedrichs, M. S.; Dongre, A. R.; Hefta, S. A. *J. Proteome Res.* **2003**, *2*, 643−649.

(22) Shen, Y.; Zhao, R.; Berger, S. J.; Anderson, G. A.; Rodriguez, N.; Smith, R. D. *Anal. Chem.* **2002**, *74*, 4235−4249.

(23) Bodnar, W. M.; Blackburn, R. K.; Krise, J. M.; Moseley, M. A. *J. Am. Soc. Mass Spectrom.* **2003**, *14*, 971−979.

(24) Pang, J. X.; Ginanni, N.; Dongre, A. R.; Hefta, S. A.; Opiteck, G. J. *J. Proteome Res.* **2002**, *1*, 161−169.

(25) Wang, W.; Zhou, H.; Lin, H.; Roy, S.; Shaler, T. A.; Hill, L. R.; Norton, S.; Kumar, P.; Anderle, M.; Becker, C. H. *Anal. Chem.* **2003**, *75*, 4818−4826.

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

(27) Washburn, M. P.; Koller, A.; Oshiro, G.; Ulaszek, R. R.; Plouffe, D.; Deciu, C.; Winzeler, E.; Yates, J. R., 3rd *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 3107−3112.

(28) Ghaemmaghami, S.; Huh, W.-K.; Bower, K.; Howson, R. W.; Belle, A.; Dephoure, N.; O'Shea, E. K.; Weissman, J. S. *Nature* **2003**, *425*, 737−741.

trypsin inhibitor, and lysozyme. Protein markers were added at relative levels of 25%, 2.5%, 1.25%, and 0.25% of the total weight of the final mixtures (400 μg). Each sample was sequentially digested, under the same conditions, with Lys-C and trypsin.[7] Approximately 70 μg of digested peptide mixture was loaded on to a biphasic (strong cation exchange/reversed phase) capillary column and washed with a buffer that contains 5% acetonitrile, 0.1% formic acid, and 95% DDI water. Two-dimensional liquid chromatography (2DLC) separation and tandem mass spectrometry conditions as described by Washburn et al. were used for the analysis.[7] In particular, the flow rate at the tip of the biphasic column was ~300 nL/min when the mobile phase composition was 95% H_2O, 5% acetonitrile, and 0.1% formic acid. The ion trap mass spectrometer, Finnigan LCQ Deca (Thermo Electron, Woburn, MA) was set to the data-dependent acquisition mode with dynamic exclusion turned on. One MS survey scan was followed by four MS/MS scans. The target value for MS was 1×10^8 and for MS/MS was 7×10^7. Maximum ion injection time was set to 100 ms.

Three parallel 12-step MudPIT experiments were carried out for each protein sample to generate the MS/MS data sets.

All data sets were searched using a modified version of the *Pep_Prob* algorithm.[29] The MS/MS data sets were searched against a database combining yeast and human protein sequences to which the sequences of phosphorylase b, serum albumin, ovalbumin, carbonic anhydrase, trypsin inhibitor, lysozyme, and some common protein contaminants (e.g., keratin) were added.

RESULTS AND DISCUSSION

Link et al. and Washburn et al. observed that data acquisition in complex peptide mixtures is not comprehensive and the process is unable to collect tandem mass spectra from all eluting peptides.[6,7] While this observation seems obvious, the impact on protein identification in complex mixtures has not been thoroughly studied. A bias toward acquisition of more abundant peptide ions frequently leads to the identification of low-abundance proteins by one or two peptides. When the analysis is repeated, peptide ions from low-abundance proteins are often missed because (1) those ions are masked from the data-dependent acquisition process by more abundant ions, (2) the ions elute at the wrong point in the data-dependent acquisition cycle (e.g., during MS/MS), and (3) chromatographic elution times are too short. Thus, the acquisition of tandem mass spectra appears to be a "semi-random" process, and the amount of randomness observed is related to sample complexity, separation resolution, ionization efficiency, ion suppression, scanning speed of the mass spectrometer, and dynamic exclusion efficiency.

To study this process, we performed 9 LC/LC/MS/MS experiments under the same instrumental conditions on the soluble fraction of a yeast cell lysate digested with trypsin. Conditions of sample complexity, separation resolution, and instrumental and data acquisition were all held constant. The collected tandem mass spectra were searched through a combined human/yeast sequence database using *Pep_Prob* modified to calculate cross-correlation scores.[29] Peptides scoring with cross-correlation scores greater than 2.0 (+1), 2.5 (+2), or 3.8 (+3) or a probability of 90% or greater with Lys or Arg at the C-terminus or Lys or Arg preceding the N-terminal residue were retained. Protein identifications were compared between each analysis, tracking the cumulative totals. A total of 1751 proteins were identified in the cumulative data set. Of these, 620 (35.4%) were found in every analysis and 420 (24%) were found in only one of the analyses. A total of 1331 (76%) proteins were found in at least two or more analyses. Ghaemmaghami et al. have shown that ~4500 proteins are expressed in log-phase growing cells by using western blots to detect tandem affinity purification (TAP) tags inserted into each yeast gene.[28] It is not expected that we would observe all yeast proteins in these analyses since our analysis encompassed only those proteins soluble in the lysis buffer. In particular, our analysis will be deficient in membrane and membrane-associated proteins.

This experiment clearly defines the limits of reproducibility of the LC/LC/MS/MS experiment for the analysis of complex peptide mixtures. It is obvious not all proteins identified in the first analysis are observed again in a second experiment, but it is also apparent that new proteins are identified with each new experiment. Replicate analyses have been used by Wu et al. and Lipton et al. as a means to increase the number of proteins identified when analyzing complex peptide mixtures.[30,31] The ability to identify new proteins with each new analysis suggests a measure of random sampling must be occurring that results in the acquisition of new peptide ions with each analysis. To understand the process, we developed and tested a statistical model for the analysis of complex peptide mixtures using the above data set to define how many experiments are necessary to produce a relatively complete analysis, e.g., saturation.

A statistical model previously used to determine the number of samplings required to measure the percentage of mRNA at particular abundance levels was used.[32]

We grouped yeast proteins based on their abundance levels (as determined by Ghaemmaghami et al.) into five bins.[28] The groups have 80, 556, 2750, 2040, and 52 proteins at the abundance levels of 5×10^5, 10^5, 1.5×10^4, 2×10^3, and 10, respectively. We are interested in determining the expected number of different protein species identified at an abundance level L after S experiments. If we denote the overall number of different protein species at the abundance level L as n_L and the number of all different protein species N, then the expectation value, K, is

$$K = n_L*(1 - (1 - L/N)^S) \qquad (1)$$

This relationship can be used to determine the number of experiments necessary to sample proteins at a certain abundance level in proteomics experiments. Thus, in a single 12-step LC/LC/MS/MS experiment the number of peptides that pass our quality filters is around 4400. The mean value of the peptide count of all identified proteins from such experiments is two. If we

(29) Sadygov, R.; J. R. Yates, I. *Anal. Chem.* **2003**, *75*, 3792−3798.

(30) Wu, C. C.; MacCoss, M. J.; Howell, K. E.; Yates, J. R. *Nat. Biotechnol.* **2003**, *21*, 532−538.

(31) Lipton, M. S.; Pasa-Tolic, L.; Anderson, G. A.; Anderson, D. J.; Auberry, D. L.; Battista, J. R.; Daly, M. J.; Fredrickson, J.; Hixson, K. K.; Kostandarithes, H.; Masselon, C.; Markillie, L. M.; Moore, R. J.; Romine, M. F.; Shen, Y.; Stritmatter, E.; Tolic, N.; Udseth, H. R.; Venkateswaran, A.; Wong, K. K.; Zhao, R.; Smith, R. D. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 11049−11054.

(32) Ewens, W. J.; Grant, G. R. *Statistical Methods in Bioinformatics*; Springer-Verlag: New York, 2002.

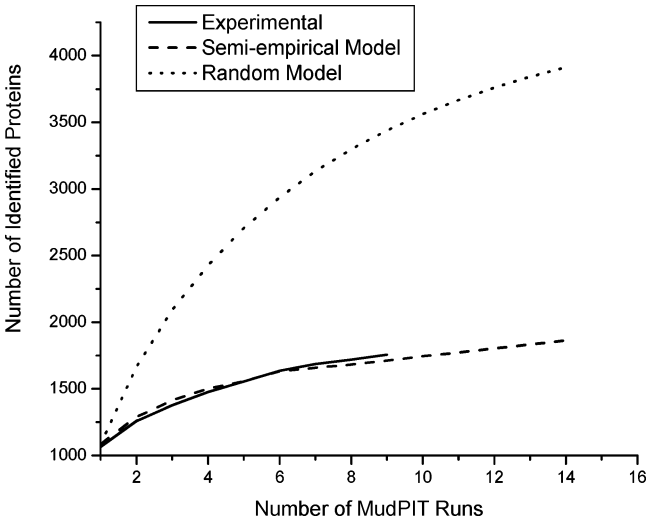

Figure 1. The cumulative number of proteins predicted to be identified by a random model is shown by the dotted line. The dashed line shows the number of proteins predicted to be identified by the statistical model (formula 1). The actual cumulative number of proteins identified in nine MudPIT analyses is shown by the solid line.

require that two peptides match a protein for the latter to be considered a true match, this provides a rough estimate of ~2000 protein identification tries per LC/LC/MS/MS experiment. It should be noted that while the formula (1) is expected to hold true for applications with other types of experiments and instruments, the number of proteins sampled will likely be dependent on the type of instrument used. Around 2000 proteins sampled per MudPIT experiment is an empirical value based on our experience and a fit to the single MudPIT results.

By using protein abundance data from yeast, we calculate with the above formula the number of expected different proteins to be 1084. This number is in good agreement with the experimentally observed value of 1064. An estimate on the number of protein identification tries for the combined LC/LC/MS/MS experiments is obtained from the ratio of peptides that filter out (from the combined LC/LC/MS/MS experiments) to the corresponding number of peptides from a single LC/LC/MS/MS experiment, 4400. If the LC/LC/MS/MS experiments were purely random and independent then for each additional experiment, we would add 2000 protein identifications attempts (as stated above this number is our empirical estimate). Both of the theoretical curves along with the experimental one are shown in Figure 1. On the basis of the 9 LC/LC/MS/MS experiments and the observed peptide identifications, we can extend the theoretical curves out to 9 LC/LC/MS/MS experiments and predict the number of protein identifications that we can expect to be observed. After 9 experiments the model predicts 1712 proteins and we observe 1751. The difference in the number of proteins predicted by the model for 15 and 20 analyses is 60, and thus there would be a minimal increase in the number of new proteins identified by 6 or 11 additional analyses. An immediate application of this statistical model is the ability to predict the number of analyses required to achieve saturation in protein identifications for a protein mixture. The statistical model also accurately predicts the number of proteins that will be identified in an LC/LC/MS/MS experiment. For organisms where protein abundance levels are unknown their values can be approximated through the use of codon bias or modeled after the distributions observed in yeast.

Table 1. Percentage of Proteins Experimentally Identified from 1, 3, 6, and 9 Combined MudPIT Runs Relative to the Results Obtained by Ghaemmaghami[28] at Different Abundance Levels, or Protein Copy Numbers per Cell

protein copy number per cell[a]	number of proteins[a]	% proteins identified from a different number of combined MudPIT runs			
		1	**3**	**6**	**9**
$>1 \times 10^5$	80	96.3	97.5	97.5	97.5
$(1 \times 10^4)-(1 \times 10^5)$	536	71.6	79.3	83.4	85.6
$(1 \times 10^3)-(1 \times 10^4)$	2184	20.8	24.4	30.6	34.9
100−1000	1036	6.5	9.7	13.2	15.5
<100	32	0	3.1	6.3	6.3

[a] Results obtained by Ghaemmaghami et al.[28]

Tightly linked to the sampling of peptides in complex mixtures is the abundance level of the original proteins. Ghaemmaghami et al. recently measured the expression level of all proteins in *S. cerevisiae* and this information can be used to evaluate the sampling process for protein identification as a function of protein abundance their measurement of protein abundance.[28] Table 1 shows the percentage of proteins identified at each abundance level, from one, three, six, and nine combined LC/LC/MS/MS experiments. The results clearly show that proteins with higher abundance levels had a higher probability of being identified. For example, 74 out of 80 proteins at expression levels higher than 10^5 copies per cell were identified in every 1 of the 9 analyses (2 of the 6 missing proteins were never observed), but only 15.3% of the total proteins expressed below 10^3 copies per cell were identified in the 9 combined LC/LC/MS/MS runs**.** Clearly, the sampling process favors the identification of high-abundance proteins.

If the number of proteins identified in each LC/LC/MS/MS experiment is plotted against each experiment, we observe that most proteins are identified in either 1 experiment or in all 9 experiments. This trend is shown in Figure 2. A large proportion of all proteins are observed in all experiments, an indication the process is reproducible for highly abundant proteins. Formula 1 predicts that 80 proteins at the highest abundance level, 5×10^5, will be observed in all 9 analyses. Of 556 proteins at the second abundance level, 10^5, ~380 will be observed in all analyses. The biggest increase in the number of newly identified proteins is predicted to come from the third abundance level, 1.5×10^4. The number of proteins increases from 527 in one analysis to 989 in the 9 combined analyses. On a percentage basis the largest increase occurs at the 10^5 level. If after 1 analysis we observe 74% of the proteins expected at this abundance level, then after 9 analyses 94% are predicted to be sampled. The number of identified proteins at the third abundance level increases by 15% after 9 analyses. The number of identified proteins from the fourth abundance level, 2×10^3, increases from 57 in 1 analysis to 117 in 9. The model indicates the dynamic range of LC/LC/MS/MS is insufficient to sample effectively at the lowest abundance level of 10 copies per cell.

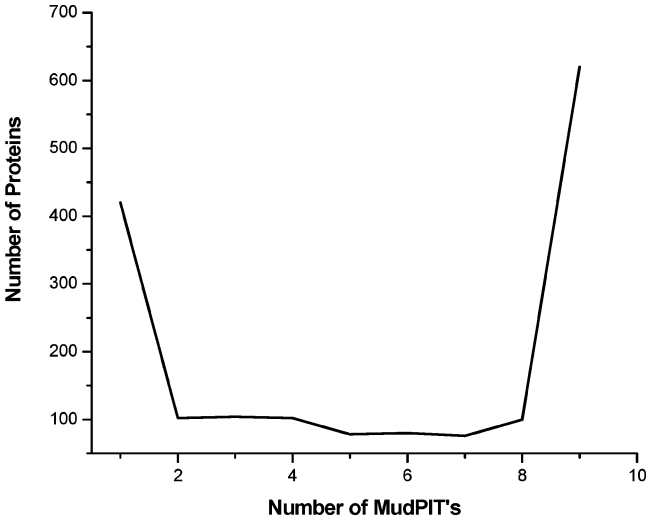

Figure 2. The number of unique proteins identified in each of the nine analyses is plotted by the solid line. Most of the proteins are identified in either one experiment or in all nine experiments.

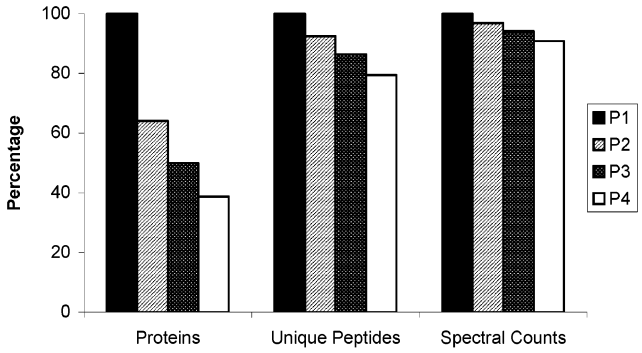

Figure 3. The number of proteins, unique peptides, and spectral counts obtained in a representative 12-step MudPIT experiment decreased, at different rates, as the number of unique peptides required for protein identification increased. P1, P2, P3, and P4 denote proteins identified by at least 1, 2, 3, or 4 peptides in the protein identification process. The values obtained at P1 were set at 100% for their respective groups.

A Model for Semiquantitative Estimation of Protein Abundance in Complex Mixtures Using MudPIT Results. Our model for random sampling suggests the probability of a protein being identified is directly related to a protein's abundance. More abundant proteins have a higher probability of identification and thus are observed more reproducibly with the acquisition of spectra for more peptides. Shotgun proteomic strategies use peptides for both identification and quantification of proteins, but require the use of stable isotope internal standards.[8,26] The statistical model for sampling suggests that when the number of spectra acquired as one of its parameters is used, an accurate prediction for the number of proteins that can be identified could be obtained. To examine the elements of the identification process which might be a reliable indicator of abundance, we used a set of the yeast LC/LC/MS/MS data to determine which features of the analysis are the most variable. We applied the same parameters to filter peptide matches. Peptides scoring with 2.0 (+1), 2.5 (+2), or 3.8 (+3) with Lys or Arg at the C-terminus or Lys or Arg proceeding the N-terminal residue were retained, but we required a different number of unique peptides for protein identification (e.g., 1−4 peptides). When more peptides were required for protein identification, the number of proteins identified dropped substantially from 1049 when one peptide was required to 406 (a 61% decrease) when four peptides were required (Figure 3). Quite surprisingly, however, the number of unique peptides identified and especially the number of spectral copies of peptides identified did not decrease as significantly as the number of proteins identified. Unique peptides identified decreased from 4769 to 3788 (a 21% decrease) and spectral copies from 12625 to 11462 (a 9% decrease). These results suggest that the number of spectra acquired from a complex peptide mixture is the most invariant property of the analysis.

The redundancy of peptide ion acquisition for abundant peptides decreases the efficiency of data acquisition for peptide ions present at low abundance in a predictable manner. Sampling statistics, thus, predict it may be possible to measure the relative abundance of proteins in mixtures without the use of internal standards. The hypothesis is that the more abundant a peptide ion is in a mixture, the higher the rate of sampling during the course of a shotgun proteomics experiment. This hypothesis is also supported by data published by Washburn et al.,[7] Pang et al.,[24] and Gao et al.[21] Velculescu et al. used a sampling approach, serial analysis of gene expression, to quantitate gene expression in pancreas and yeast. [33, 34]

Validation Experiments: Spectral Count Is a Good Measure of Relative Abundance. To prove this hypothesis for relative measurement of protein abundance in complex mixtures, a yeast-soluble protein mixture was spiked with specific quantities of 6 different proteins (SDS-PAGE low molecular weight standards). Four different mixtures were created with decreasing amounts of the six proteins, but the same quantity of total protein. Each protein mixture was independently digested using trypsin and then analyzed three times using LC/LC/MS/MS. Figure 4A shows the relationship between the number of spectral copies and the relative abundance of the six protein markers. All six proteins showed a uniform linear correlation between spectral copy number and their abundance from 4.17% to 0.0417% of total protein content representing a two order linear dynamic range. R^2 values were between 0.9967 and 0.9995 for the linear correlation between spectral count and the amount of each protein markers added to the yeast cell lysate.

Such a large linear range has not been observed in other quantitative proteomics methods[35,36] where the linear dynamic range has been explicitly measured. We do not expect there was a particular bias associated with the proteins used in the study. They covered a wide range of molecular weights from 16 to 100 kDa, and had different sequences and chemical and physical characteristics. It is quite striking that the number of tandem mass spectra collected from peptides of these six proteins displayed near perfect linearity with respect to concentration, which is in agreement with our statistical model. The ability to measure the

(33) Velculescu, V. E.; Zhang, L.; Vogelstein, B.; Kinzler, K. W. *Science* **1995**, *270*, 484−487.
(34) Velculescu, V. E.; Zhang, L.; Zhou, W.; Vogelstein, J.; Basrai, M. A.; Bassett, D. E., Jr.; Hieter, P.; Vogelstein, B.; Kinzler, K. W. *Cell* **1997**, *88*, 243−251.
(35) Ong, S. E.; Blagoev, B.; Kratchmarova, I.; Kristensen, D. B.; Steen, H.; Pandey, A.; Mann, M. *Mol. Cell Proteomics* **2002**, *1*, 376−386.
(36) Washburn, M. P.; Ulaszek, R.; Deciu, C.; Schieltz, D. M.; Yates, J. R., 3rd *Anal. Chem.* **2002**, *74*, 1650−1657.

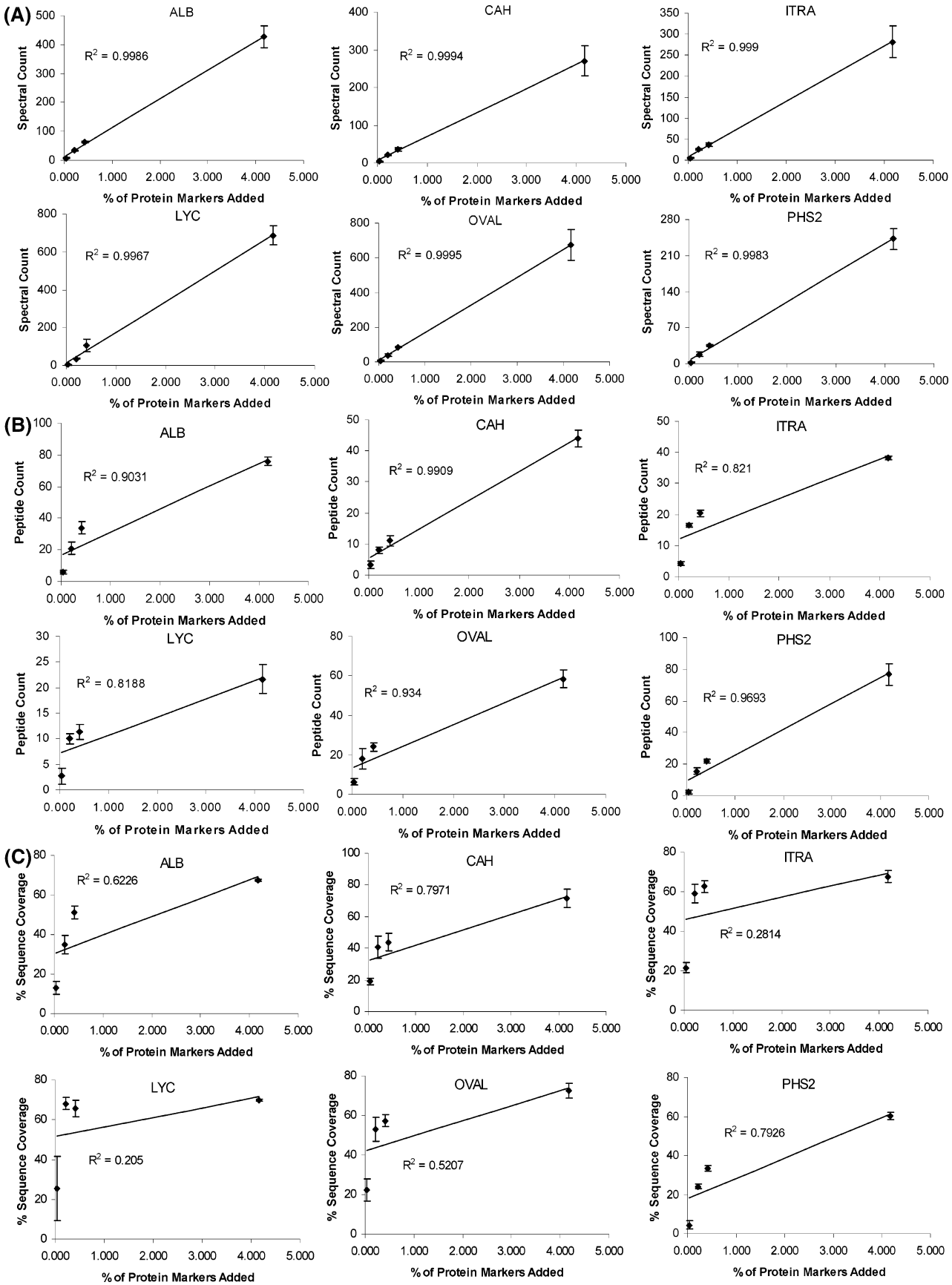

Figure 4. Correlation between spectral count, peptide count, and sequence coverage with protein abundance level. In each experiment six protein markers were added at different percentages, from 0.0417% to 4.17%, representing different protein abundance levels. For all protein markers, a linear correlation was observed between spectral count and relative protein abundance (i.e., % of protein markers added) (A), but not between peptide count and relative protein abundance (B), and sequence coverage and relative protein abundance (C). Abbreviations used here are as follows: CAH, carbonic anhydrase 2; ALB, albumin; ITRA, trypsin inhibitor; LYC, lysozyme; OVAL, ovalbumin; PHS2, phosphorylase b.

Table 2. Summary of Shotgun Proteomics Results for the Six Protein Markers[a]

25% of Total[g]

	experiment 1			experiment 2			experiment 3				
locus ID	peptide count[b]	spectral count[c]	sequence coverage[d]	peptide count[b]	spectral count[c]	sequence coverage[d]	peptide count[b]	spectral count[c]	sequence coverage[d]	protein length[e]	MW[f]
ALBU	79	432	67.2	74	387	68	75	464	67.1	607	69294
CAH2	42	240	65.3	47	315	73.4	43	258	76.4	259	28981
ITRA	39	243	68.1	38	319	64.4	38	283	70.4	216	24005
LYC	25	635	70.1	20	694	69.4	20	733	70.1	147	16239
OVAL	63	652	73.5	54	600	68.6	58	773	75.6	385	42750
PHS2	75	263	58.4	84	222	61.8	71	243	60.5	842	97158

2.5% of Total[g]

	experiment 1			experiment 2			experiment 3				
locus ID	peptide count[b]	spectral count[c]	sequence coverage[d]	peptide count[b]	spectral count[c]	sequence coverage[d]	peptide count[b]	spectral count[c]	sequence coverage[d]	protein length[e]	MW[f]
ALBU	30	65	47.3	37	63	52.2	35	59	53.9	607	69294
CAH2	10	31	38.6	13	35	49.8	10	40	43.2	259	28981
ITRA	19	35	64.4	21	35	59.3	21	42	64.4	216	24005
LYC	10	136	61.9	13	72	70.1	11	109	64.6	147	16239
OVAL	22	83	55.1	26	83	60.8	24	87	56.1	385	42750
PHS2	21	35	32.9	23	33	34.9	21	37	32.2	842	97158

1.25% of Total[g]

	experiment 1			experiment 2			experiment 3				
locus ID	peptide count[b]	spectral count[c]	sequence coverage[d]	peptide count[b]	spectral count[c]	sequence coverage[d]	peptide count[b]	spectral count[c]	sequence coverage[d]	protein length[e]	MW[f]
ALBU	17	33	29.7	24	35	38.1	22	33	36.7	607	69294
CAH2	7	21	32.4	8	20	44	9	22	45.2	259	28981
ITRA	17	28	55.6	17	23	64.4	16	25	56.9	216	24005
LYC	9	35	64.6	10	38	70.1	11	34	69.4	147	16239
OVAL	13	32	48.8	23	46	60	18	39	50.1	385	42750
PHS2	15	20	24.3	14	14	22.8	18	21	25.1	842	97158

0.25% of Total[g]

	experiment 1			experiment 2			experiment 3				
locus ID	peptide count[b]	spectral count[c]	sequence coverage[d]	peptide count[b]	spectral count[c]	sequence coverage[d]	peptide count[b]	spectral count[c]	sequence coverage[d]	protein length[e]	MW[f]
ALBU	5	8	11.2	5	5	10.7	7	8	16.6	607	69294
CAH2	4	4	20.1	2	4	16.6	4	6	20.1	259	28981
ITRA	4	5	20.4	5	6	24.5	4	4	19.9	216	24005
LYC	1	1	8.2	3	4	28.6	4	5	40.1	147	16239
OVAL	6	6	22.6	5	5	16.6	8	9	27.5	385	42750
PHS2	1	1	1.9	3	3	5.7	3	3	5.5	842	97158

[a] Peptide count, spectral count, and sequence coverage of the six protein markers were obtained in yeast cell lysate samples spiked in with different percentages (i.e., 25%, 2.5%, 1.25%, and 0.25%, by weight) of a 6-component protein marker mixture. Three parallel MudPIT experiments were conducted for each percentage, or abundance level, of protein markers. Abbreviations used: CAH2, bovine carbonic anhydrase 2; ALBU, bovine serum albumin; ITRA, soybean trypsin inhibitor; LYC, chicken lysozyme; OVAL, chicken ovalbumin; PHS2, rabbit phosphorylase b. [b] Peptide count is the nonredundant number of peptides identified from a protein. [c] Spectral count is the total number of spectra identified from a protein. [d] Sequence coverage is the percentage of a protein's amino acid sequence covered by peptides identified. [e] Protein length is the number of amino acids in a protein. [f] MW is predicted molecular weight of a protein. [g] The weight percentage of protein markers added to a yeast cell lysate sample.

relative amount of proteins in a mixture without the use of stable isotope labels offers a simplistic method to measure the relative abundance of proteins in a mixture, and potentially an approach to measure large expression changes of the same protein between different samples. In contrast to the use of spectral copy number to derive relative abundance data for proteins, % sequence coverage and number of peptides identified per protein did not show as a good of a linear correlation as spectral count (Figure 4B,C).

Stable isotope labeling methods excel at measuring small changes between proteins.[8,26] Detecting large differences has been more problematic as the linear dynamic range for detecting differences between isotopomers has been small (~10/1) where the linear dynamic range has been explicitly measured.[35,36] The data shown in Figure 4 that illustrates this approach can detect differences of a factor of 2 (i.e., 0.208% vs 0.417%) and can detect larger differences (4.17% vs 0.0417%). All six protein markers were identified in all 12 LC/LC/MS/MS experiments over this 10^2 concentration range. The reproducibility of the spectral count of these six known proteins is good (Table 2). We also determined the reproducibility and variability of spectral sampling for the 600 proteins identified in each of the 9 experiments (see Supporting

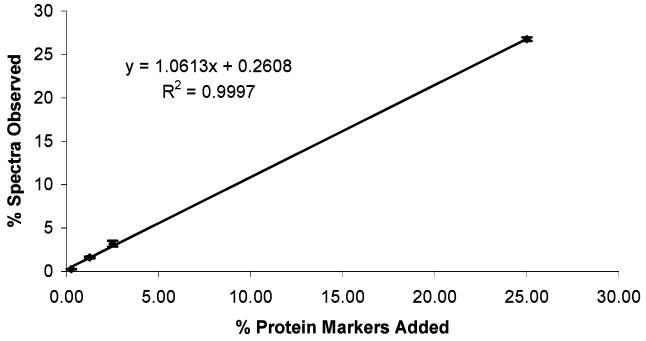

Figure 5. Overall correlation between spectral sampling frequency, expressed as the percentage of combined spectra observed from all six protein markers, and the weight percentage of protein makers added. The overall correlation is linear, with a R^2 value of 0.9997 and a slope very close to 1.

Information, Table 1). These results suggest the reproducibility of spectral sampling across many experiments is good and the linear dynamic range exceeds that of stable isotope labeling.

Toward Relative Abundance Measurements between Different Proteins. Our semiquantitative model shows that under the conditions of our experiment the predicted protein abundance is directly proportional to its relative abundance. This model was validated by measuring known quantities of specific proteins in a soluble yeast cell lysate and comparing the number of spectra sampled as a function of amount. We conclude from these results that if all proteins in a mixture are analyzed using the same conditions, we will be able to measure an accurate relative abundance by counting the number of spectral copies obtained in the LC/LC/MS/MS experiments. We do not believe the validation experiment represents a unique case as the six proteins are of different molecular weights and sequence. The differences in spectral sampling for each protein are small enough that when the concentration difference is large, that is, larger than five times, we can reliably tell the abundance difference by the number of collected spectra. Using spectral sampling to measure the difference between proteins suggests a practical and simple method to measure differences between different proteins in a mixture. Most stable isotope dilution methods measure differences between the same protein by comparison to an internal standard. The AQUA technique can measure absolute differences between proteins but it is not a global measurement technique at this point in time.[37] Measuring relative changes between different proteins would have value to determine how proteins are changing in cells relative to one another.

To further substantiate the relationship between protein concentration and spectral copy number, we analyzed all six protein markers as a group. The percentage of the spectra for all six proteins as a function of total number of spectra that passed the DTASelect filter was calculated. These data were plotted against the known percentage weight of these six proteins (Figure 5). There are two important features in this figure: first, there is a near perfect linear correlation over 2 orders of magnitude of the concentration range; an R^2 value of 0.9997 was observed, reconfirming the use of spectral count as a measure of protein abundance. Second, the slope of the line regression is 1, indicating

(37) Gerber, S. A.; Rush, J.; Stemman, O.; Kirschner, M. W.; Gygi, S. P. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 6940−6945.

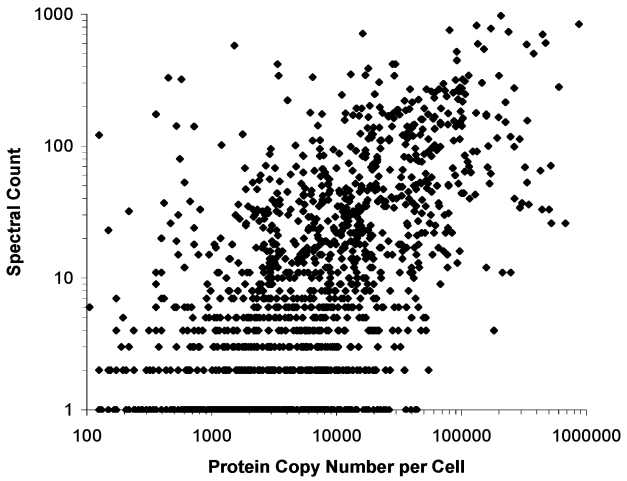

Figure 6. Plot of protein abundance in yeast obtained using our spectral count method vs recently published results from a large-scale yeast protein expression study (Ghaemmaghami et al.[28]).

Table 3. Proteins That Were Predicted To Express at High-Abundance Levels Based on High Spectral Counts Observed, but Were Identified as Low-Abundance Proteins in a Different Study by Ghaemmaghami et al.[28]

locus ID	spectral count	protein length	MW	protein copy number per cell	protein name[a]
YPL131W	108	297	33715		RPL5
YEL034W	162	157	17114		HYP2
YFL039C	117	375	41690		ACT1
YGL123W	102	254	27450		RPS2
YJR123W	170	225	25039		RPS5
YLR340W	114	312	33717		RPP0
YJL138C	111	395	44697		TIF2
YPR080W	529	458	50033	377	TEF1
YOR063W	125	387	43758	450	RPL3
YBR118W	524	458	50033	450	TEF2
YJR145C	125	261	29410	573	RPS4A
YHR174W	3344	437	46914	2610	ENO2
YJL191W	152	138	14650	3370	RPS14B
YNL302C	131	144	15891	3440	RPS19B

[a] The names of proteins are their standard names according to the Saccharomyces Genome Database (http://www.yeastgenome.org).

that the spectral copy percentage is equal to the weight percentage of this group of six proteins in a background of yeast proteins.

Many factors could affect spectral sampling of a protein obtained from a LC/LC/MS/MS analysis of complex protein mixtures, such as proteolytic digestion efficiency, peptide separation, peptide ionization efficiency, co-eluting peptides, randomness generated from data-dependent MS/MS acquisition and dynamic exclusion. Spectral sampling is a secondary measurement of the process, however, and many factors that affect the acquisition of spectra may average out and allow it to accurately reflect protein abundance in the mixtures. This is especially true when estimating the abundance of a group of proteins where a large number of peptides with different properties are available. If two mixtures are to be compared, the same experimental conditions need to be used to minimize systematic errors.

The data collected for yeast-soluble proteins were analyzed to determine the spectral copy number for all proteins in the mixture. It is clear from this analysis proteins known to be highly abundant

Table 4. Proteins That Were Predicted To Express at Low-Abundance Levels Based on Low Spectral Counts Observed, but Were Identified as High-Abundance Proteins a Study by Ghaemmaghami et al.[28] [a]

locus ID	spectral count	protein length	MW	protein copy number per cell	protein name
YLR110C	12	133	13058	158000	CCW12
YBR010W	11	136	15356	213000	HHT1
YCR024C−A	0	40	4501	124000	PMP1
YDR033W	4	320	36190	182000	MRH1
YKL096W−A	0	92	8911	1590000	CWP2
YNL031C	11	136	15356	248000	HHT2

[a] Most of these proteins have relatively low molecular weights and therefore may be under-represented in the mass spectrometry data set.

in the yeast proteome have a larger number of spectral copies. Ghaemmaghami et al. recently measured the abundance of all yeast proteins expressed in log-phase growing cells by using western blot measurements of epitope tagged ORFs. We plotted our data versus the cell copy numbers calculated by Ghaemmaghami et al. (Figure 6). The overall trend of the data agrees fairly well with a linear correlation of 0.58 calculated by the Pearson correlation. There are some obvious outliers suggesting the insertion of the tandem affinity purification tag may have disrupted normal expression of the gene. For example, 14 proteins in our study had 10 times more spectra than the mean number for all proteins identified (within the top 15% of spectral count rank), but are observed in the Ghaemmaghami study at <4000 copies/cell or whose expression was not observed at all (Table 3). Some of these proteins have been observed in other studies at high expression levels.[34,38,39] Six proteins observed by Ghaemmaghami et al. at high copy number/cell (larger than 1×10^5) were then observed in our data set at low spectral count level or not at all (Table 4). Five of these proteins are short and thus do not produce peptides of the optimum size for mass spectrometry after tryptic digestion. This situation reflects a potential limitation to all mass spectrometry methods.

(38) Futcher, B.; Latter, G. I.; Monardo, P.; McLaughlin, C. S.; Garrels, J. I. *Mol. Cell Biol.* **1999**, *19*, 7357−7368.
(39) Yan, J. X.; Sanchez, J. C.; Tonella, L.; Williams, K. L.; Hochstrasser, D. F. *Electrophoresis* **1999**, *20*, 738−742.

CONCLUSION

The use of tandem mass spectrometry to identify proteins in mixtures is becoming a widely used method. We have shown the sampling process is predictable and agrees well with a statistical model that allows the number of experiments required to reach a reasonable level of completeness in an analysis to be predicted. Furthermore, this model accurately predicts that higher abundance proteins are sampled more frequently, but that greater coverage of lower abundance proteins can be achieved by increasing the number of experiments performed on a sample. Spectral sampling is shown to accurately reflect relative abundance with a linear correlation over a 2 order of magnitude linear dynamic range. Spectral sampling is less accurate at measuring small differences between proteins but was shown to be very accurate at measuring large changes between proteins. We also expect this method to provide an accurate measure of differences between different samples, allowing a measure of how proteins are changing overall. We have also determined a reasonable correlation between spectral sampling and a recent measurement of the abundance of yeast proteins in log-phase growing cells. A higher level of correlation may exist but our study used a different strain of yeast and log growth cells in contrast to the Ghaemmaghami et al. study.[28] Furthermore, the insertion of a C-terminal epitope tag into yeast ORFs may have disrupted the expression or turnover of some proteins also contributing to poorer correlation between our measurements and protein levels measured by Ghaemmaghami et al.[28]

ACKNOWLEDGMENT

This research was supported by the Office of Naval Research N00014-00-1-0421 and NIH Grants RR11823-08 and 5R01 MH067880. We thank present and former members of the Yates laboratory for helpful discussions.

SUPPORTING INFORMATION AVAILABLE

Additional data in tabular form. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review January 23, 2004. Accepted April 23, 2004.

AC0498563