Evaluation of Multidimensional Chromatography Coupled with Tandem Mass Spectrometry (LC/LC−MS/MS) for Large-Scale Protein Analysis: The Yeast Proteome

Junmin Peng,[†] Joshua E. Elias,[†] Carson C. Thoreen,[‡] Larry J. Licklider,[‡] and Steven P. Gygi*,[†,‡]

*Department of Cell Biology, and Taplin Biological Mass Spectrometry Facility, Harvard Medical School,
240 Longwood Avenue, Boston, Massachusetts 02115*

Received July 30, 2002

Highly complex protein mixtures can be directly analyzed after proteolysis by liquid chromatography coupled with tandem mass spectrometry (LC−MS/MS). In this paper, we have utilized the combination of strong cation exchange (SCX) and reversed-phase (RP) chromatography to achieve two-dimensional separation prior to MS/MS. One milligram of whole yeast protein was proteolyzed and separated by SCX chromatography (2.1 mm i.d.) with fraction collection every minute during an 80-min elution. Eighty fractions were reduced in volume and then re-injected via an autosampler in an automated fashion using a vented-column (100 μm i.d.) approach for RP-LC−MS/MS analysis. More than 162 000 MS/MS spectra were collected with 26 815 matched to yeast peptides (7537 unique peptides). A total of 1504 yeast proteins were unambiguously identified in this single analysis. We present a comparison of this experiment with a previously published yeast proteome analysis by Yates and colleagues (Washburn, M. P.; Wolters, D.; Yates, J. R., III. *Nat. Biotechnol.* **2001**, *19*, 242−7). In addition, we report an in-depth analysis of the false-positive rates associated with peptide identification using the Sequest algorithm and a reversed yeast protein database. New criteria are proposed to decrease false-positives to less than 1% and to greatly reduce the need for manual interpretation while permitting more proteins to be identified.

Keywords: proteome • tandem mass spectrometry • LC−MS/MS • vented column • Sequest criteria

Introduction

One of the greatest challenges facing researchers in the post-genomic era is to identify and quantify all expressed protein components in cells, tissues, and organisms. While two-dimensional gel electrophoresis (2DE) is a powerful technique for protein separation,[1−3] it has a number of limitations that have spawned new technologies as alternatives to 2DE.[4−12] A powerful alternative technique is the so-called MUDPIT (multidimensional analysis of proteins identification technology) pioneered by Yates and colleagues.[13−18] This technique is also termed DALPC for direct analysis of large protein complexes.[13,19,20] The acquired tandem mass spectrometry (MS/MS) data are used for database searching,[21−26] which leads to identification of peptides and proteins. At the heart of the method is the use of two-dimensional (2D) chromatography to separate a peptide mixture prior to analysis by mass spectrometry.

By utilizing peptides' unique physical properties of charge and hydrophobicity, a complex peptide mixture can be effectively resolved and concentrated prior to sequence analysis by mass spectrometry. Strong cation exchange (SCX) chromatography is generally implemented as a primary separation technique due to its potential for increased loading capacity while reversed-phase (RP) chromatography is a perfect compliment as a secondary separation technique because of its ability to remove salts and its direct compatibility with mass spectrometry through electrospray ionization.

2D chromatography can be accomplished by either an online or an offline approach.[27] For the online approach, an acidified complex peptide mixture is applied to an SCX chromatography column and discrete fractions of the adsorbed peptides are sequentially displaced directly onto the RP chromatography column using a salt step gradient. Peptides are then eluted and analyzed by tandem mass spectrometry (MS/MS). This approach can utilize as many as 15 (or more) salt "bumps" to fractionate a peptide mixture.[13,14] A principal advantage to the online technique is nearly complete automation in an unattended fashion. Essentially, the sample is loaded onto the SCX column and 24 h later the analysis is finished.

The offline approach is performed by applying the acidified complex peptide mixture to the SCX chromatography column followed by a binary gradient to high salt to elute the peptides. Fractions are typically collected every minute into a 96-well plate and reduced in volume, and then each fraction is loaded onto a RP chromatography column automatically via an autosampler and analyzed by LC−MS/MS. The offline technique has several advantages over the online technique: (i)

* To whom correspondence should be addressed. E-mail: steven_gygi@hms.harvard.edu.

[†] Department of Cell Biology.

[‡] Taplin Biological Mass Spectrometry Facility.

10.1021/pr025556v CCC: $25.00

Published on Web 10/15/2002

peptide separation is superior using a linear gradient instead of a step-gradient; (ii) ideally, SCX chromatography is performed with >20% acetonitrile (AcN) in the buffers to linearize peptides,[28] yet only 5−10% AcN can be tolerated with the online approach; (iii) more peptide fractions can be collected with the offline approach (e.g., 96 vs 15); and (iv) user discretion as to which fractions are to be analyzed is available in the offline approach, and interesting fractions can be re-visited (reanalyzed).

We report here the use of an off-line LC/LC−MS/MS strategy to catalog the expressed proteins of the yeast *S. cerevisiae*. One milligram of total protein was proteolyzed and analyzed by LC/LC−MS/MS. A large dataset was generated that allowed us to estimate false-positive rates during database searching using the Sequest program and led to the establishment of new criteria to reduce the false-positive rate to less than 1%.

Experimental Section

Apparatus. An LCQ Deca ion trap mass spectrometer was purchased from Thermo Finnigan (San Jose, CA). Accessories for capillary liquid chromatography were from Upchurch Scientific Inc. (Oak Harbor, WA). Capillary tubing containing a borosilicate frit (Integrafrit) was from New Objective Inc. (Cambridge, MA). Magic C18AQ resin (5 μm, 200 Å) was from Michrom BioResources (Auburn, CA).

Reagents. Sequencing grade modified trypsin was purchased from Promega (Madison, WI). High purity acetonitrile (AcN), methanol, and acetic acid (HOAc) were from VWR (Bridgeport, NJ), and heptafluorobutyric acid (HFBA) was from Pierce (Rockford, IL).

Procedure. Preparation and Digestion of Total Cell Lysate from *S. cerevisiae*. Strain S288C was grown to log phase in YPD medium. The cells were collected by centrifugation and lysed with glass beads in lysis buffer (20 mM Tris, pH 8.0, 10 mM NaF, 10 mM NaCl, 0.1% deoxycholic acid, 0.5 mM EDTA, and protease inhibitor cocktail from Roche). The protein concentration was measured by a Bio-rad assay using bovine serum albumin (BSA) as standard. The lysate containing 1 mg of proteins was denatured by adding urea to 8 M and SDS to 0.1%, reduced with 10 mM DTT at 37 °C for 1 h, and alkylated with 50 mM iodoacetamide in the dark for 30 min. Dialysis and dilution were used to reduce urea to 1 M and SDS to 0.025% as described.[27]A 20 μg sample of trypsin (1:50) was added to digest the proteins at 37 °C overnight.

Two-Dimensional Liquid Chromatography with Tandem Mass Spectrometry. The tryptic peptides were acidified with TFA to 0.1%, supplemented with AcN to 25%, and then loaded onto a 2.1 mm × 20 cm polysulfoethyl A column (Poly LC Inc., Columbia, MD) connected to a Beckman Gold HPLC. Fractions were collected every minute during an 80-min gradient from 5% to 35% solvent B (solvent A: 5 mM phosphate buffer, 25% AcN, pH 3.0; solvent B: the same as A with 350 mM KCl; flow rate: 0.2 mL/min). All collected fractions ($n = 80$, 0.2 mL each) were reduced in volume to ~0.1 mL to remove AcN by vacuum centrifugation and then analyzed by RP-LC−MS/MS in a completely automated fashion using a 100 μm i.d. vented column as described.[29] We loaded as many as 29 fractions onto a single V-column. In other work, we used a single V-column to analyze more than 100 samples (data not shown). Peptides were eluted for each analysis during a gradient from 10% to 30% buffer B (buffer A: 0.4% acetic acid, 0.005% HFBA, 5% AcN; buffer B: the same as A except 95% AcN; flow rate: ~300 nL/min). In general, the amount of sample loaded was adjusted according to its peptide concentration: about 50% of the total volume for fractions with OD_{214} < 1.2 and 25% for all others (OD_{214} at least 1.2, fractions 14−18). Gradient length was largely dependent on the peptide concentration estimated by its OD_{214} absorbance: 60 min (fractions 1−8 and 67−80, $OD_{214} < 0.2$), 90 min (fractions 42−66, OD_{214} between 0.2 and 0.5), 120 min (fractions 28−41, OD_{214} between 0.2 and 0.5), or 150 min (fractions 9−27, $OD_{214} > 0.5$). Peptide ions were detected in a survey scan from 400 to 1700 amu (3 μscans) followed by five data-dependent MS/MS scans (5 μscans each, isolation width 3 amu, 35% normalized collision energy, dynamic exclusion for 3 min) in a completely automated fashion on an LCQ-DECA ion trap mass spectrometer (Thermo Finnigan, San Jose, CA).

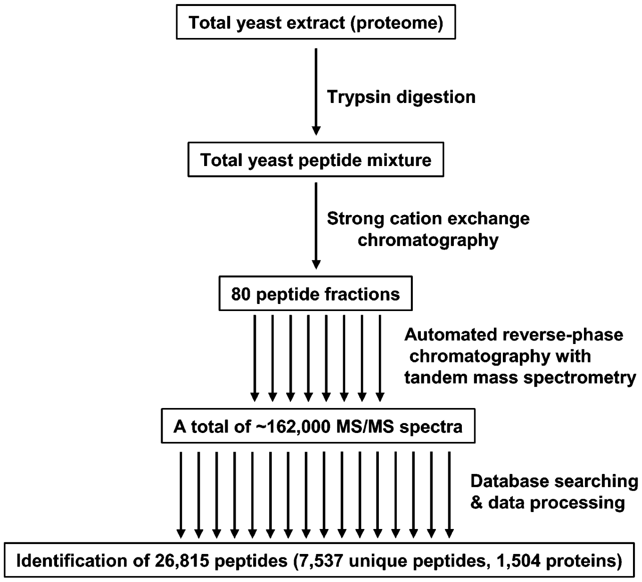

Total yeast extract (proteome)
Trypsin digestion
Total yeast peptide mixture
Strong cation exchange chromatography
80 peptide fractions
Automated reverse-phase chromatography with tandem mass spectrometry
A total of ~162,000 MS/MS spectra
Database searching & data processing
Identification of 26,815 peptides (7,537 unique peptides, 1,504 proteins)

Figure 1. Yeast proteome analysis using a combination of two-dimensional chromatography and tandem mass spectrometry. Whole yeast protein (1 mg) was first digested with trypsin (see the Experimental Section). The resulting highly complex peptide mixture was analyzed by SCX chromatography during an 80-min gradient with fraction collection. Each fraction ($n = 80$) was subjected to automated reversed-phase (RP) nanoscale capillary liquid chromatography (LC) coupled with tandem mass spectrometry (MS/MS). All resulting spectra were searched against protein sequence databases with the Sequest algorithm for peptide identification.

Data Processing. All MS/MS spectra were searched using the Sequest algorithm.[21,30] The database utilized was a composite of the known yeast ORFs (6139 proteins) and a second database containing 6139 bogus proteins created by precisely reversing the order of the amino acid sequence for each protein so that the C terminus became the N terminus. This is similar to a database created by Moore et al.[26] A total of 162 000 MS/MS spectra were searched against the database using the following criteria: no requirement for trypsin digestion, variable modification of methionine (+16 Da) and cysteine (+57 Da). The reason for the addition of the reverse-database is to provide an estimate of the false-positive rate of peptide identification by Sequest under various parameters.

Results

A flowchart of the experiment is presented in Figure 1. Whole yeast extract (1 mg protein) was trypsinized into peptides in the presence of SDS and urea. The resulting peptides were

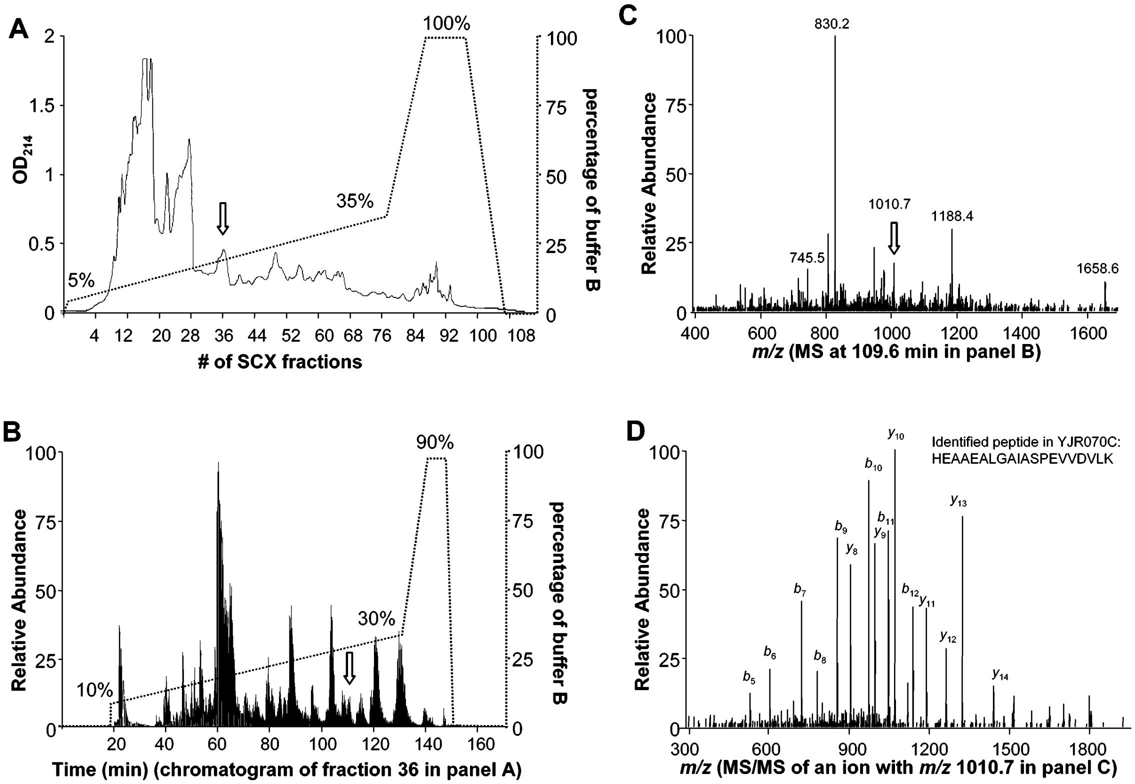


Figure 2. Large-scale analysis of yeast proteome by LC/LC−MS/MS. (A) SCX elution profile of typsinized total yeast extract. Peptides were eluted in a gradient from 5% to 35% buffer B and monitored by UV absorption at 214 nm. Fractions were collected every minute. (B) Base-peak chromatogram of SCX fraction 36 during RP LC−MS/MS analysis. Peptides were loaded and desalted in the first 20 min and eluted in a 10−30% gradient of buffer B over 120 min. Selected peptides were subjected to MS/MS analysis. (C) An MS survey scan during LC−MS/MS analysis at time 109.6 min. Many peptides can be seen coeluting. The mass spectrometer sequentially selected five peptide ions for further sequence analysis by collision-induced dissociation. (D) MS/MS scan of precursor ion *m/z* 1010.7. This was one of the five peptide ions chosen for analysis which was later matched to the peptide amino acid sequence shown. Some *b*- and *y*-ions derived from the peptide ion are also indicated.

separated by strong cation exchange (SCX) chromatography with fraction collection. Each fraction ($n = 80$) was collected and analyzed by automated reversed-phase (RP) chromatography coupled with tandem mass spectrometry using a vented column approach.[29] In total, more than 162 000 MS/MS spectra were automatically collected and searched against the composite database (see the Experimental Section). More than 26 800 peptides were identified (7537 unique peptides, see the Supporting Information).

As an example, the analysis of fraction 36 is shown in Figure 2. The off-line SCX analysis (Figure 2A) resulted in the collection of 80 discrete fractions. The LC−MS/MS analysis of fraction 36 is shown in Figure 2B. Figure 2C shows a single MS survey scan at the point of 109.6 min during the analysis. Figure 2D shows an MS/MS scan of the precursor ion, *m*/*z* 1010.7 triggered from the survey scan.

To avoid loading each SCX fraction manually, we utilized a vented-column approach for the RP-LC−MS/MS analysis[29] (Figure 3). This allowed for fast loading of SCX fractions and desalting (Figure 3A). To test the reproducibility and sensitivity of the system, a standard peptide (bradykinin) was loaded from 10 fmol to 1 pmol (Figure 3B). The system could readily detect the peptide at 10 fmol with a signal-to-noise ratio of about 4 and had a consistent performance indicated by the similar retention time of the peptide in each analysis.

After acquisition of MS/MS spectra, the software algorithm Sequest[21,30] was employed for peptide/protein identification. Sequest matches an MS/MS spectrum with a peptide sequence using the following steps. (i) All theoretical peptides with molecular masses isobaric to that of the precursor ion for an MS/MS spectrum are extracted from a database. (ii) Each peptide is given a preliminary score by examining the number of predicted fragment ions from the database peptide that match the acquired fragment ions in the MS/MS spectrum. (iii) The 500 best-matching peptides undergo a more rigorous ion-matching algorithm that generates a Sequest cross-correlation score (X_{corr}). A list of the best-matching peptides is returned to the user with the top-scoring peptide being considered the best candidate. The difference of X_{corr} scores between the top peptide and the runner-up is represented by the function termed delta-correlation score (ΔC_n).

In most large-scale peptide sequence analysis studies where Sequest is utilized, the results of the search for each MS/MS spectrum are filtered according to some established criteria to achieve a list of peptide matches. The criteria utilized include three parameters: state of tryptic ends (fully- or partially tryptic), X_{corr}, and ΔC_n. Although this data processing procedure is sophisticated, the standard for accepting matched peptides is rather empirical, and a large number of spectra require manual interpretation.[13] This dataset presented a unique opportunity to evaluate the rate of false-positives achieved by varying these three parameters because of the size of the dataset. More than 162 000 MS/MS spectra and more than 20 000 spectra easily matched to correct peptides. Hence, the

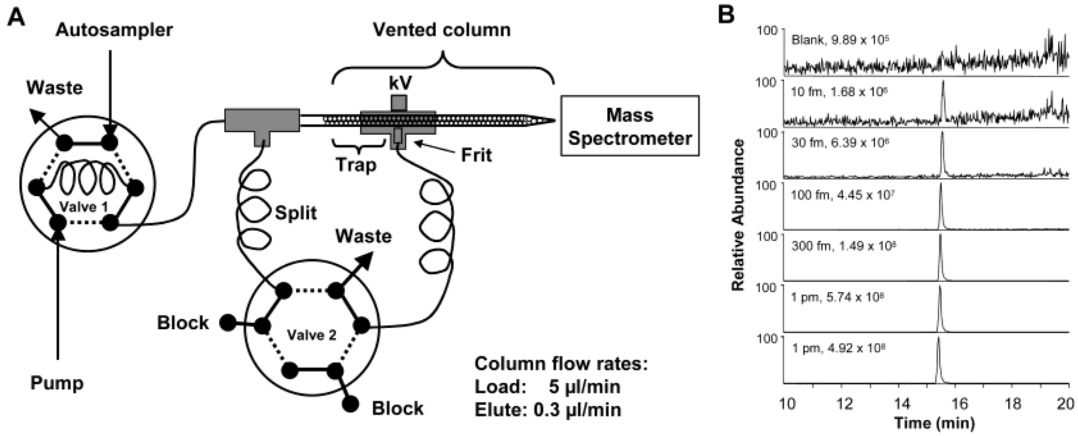


Figure 3. Automated nanoscale microcapillary LC−MS/MS using a vented column. (A) Integrated scheme including an autosampler, HPLC pump, two six-port valves, one micro-tee, one micro-cross, a vented RP column and a LCQ^{DECA} ion trap mass spectrometer. SCX fractions were first loaded onto a sample loop in valve 1 that was connected by dashed lines. Valves 1 and 2 were then switched (connected by solid lines), which closed the micro-tee (flow restrictor; split) and opened the micro-cross (vent). At this stage, the sample was transferred to the trap region of the vented-column and desalted by extensive wash at high flow rate (up to 5 μL/min). After sample loading at high flow rate, valve 2 was again switched (dashed-line connection), which opened the flow-restrictor (split) and closed the vent. Peptides bound to the trap were eluted during a RP gradient and ionized by electrospray with voltage (1.8 kV) applied via a gold wire to one arm of the cross. All 80 fractions were analyzed by this same procedure. (B) A standard peptide bradykinin (monoisotopic *m/z* of 530.8 for doubly charged ion) was sequentially titrated from 10 fmol to 1 pmol using a 100 μm i.d. nanoscale vented-column. Selected ion chromatograms for the doubly charged peptide ion (*m/z* 530.3−533.3) are shown. The amount of the peptide loaded is also indicated as well as the strength of signal recorded.

number of total peptides after filtering was large enough to easily evaluate false-positive rates of less than 1%.

To address the false-positive rate in database searching by Sequest, we exploited the use of a composite database containing the yeast ORFs in both the forward (correct) and reverse (incorrect) orientation. The reverse-database was positioned after the forward one. After searching the 162 000 MS/MS spectra, the results were filtered according to specific criteria. Any peptide passing the filtering parameters that was derived from the reverse-database is defined as a false-positive (followed by manual verification of all false-positives). Because the reverse-database is exactly the same size (number of amino acids), it can be assumed that the rate of false-positives is the same for both databases. Therefore the overall false-positive rate can be estimated by doubling the number of peptides found from the reverse-database and dividing the result by the total number of identified peptides from both databases according to following formula: $\%\ \mathrm{fal} = 2[n_{\mathrm{rev}}/(n_{\mathrm{rev}} + n_{\mathrm{real}})]$, where % fal is the estimated false-positive rate, n_{rev} is the number of peptides identified (after filtering) from the reverse-database, and n_{real} is the number of peptides identified (after filtering) from the real database.

We thus calculated the estimated false-positive rate of returned peptide identifications with regard to their charge state, tryptic state, X_{corr}, and ΔC_n. We first examined the effect of increasing the ΔC_n on the estimated false-positive rate for fully tryptic and doubly charged peptides. As expected, when the cutoff of ΔC_n increased for various *Xcorr* settings, false-positive rates decreased gradually (Figure 4) as did the number of peptides accepted (data not shown). To maximize the identified peptides and keep the false-positive rate below 1%, the cutoff ($\Delta C_n > 0.08$) showed a good compromise. After fixing the cutoff of ΔC_n, the false-positive rate was further derived as shown in Table 1. The estimated false-positive rate was highly dependent on peptide charge state and tryptic state. For example, for nontryptic peptides with a low X_{corr} score (<2), nearly all peptides matched in the database were false-positives, which was in good agreement with the fact that trypsin possesses high substrate specificity and produces very few completely nontryptic peptides during digestion. In addition, when the same threshold of X_{corr} was applied, fully tryptic peptides had a much lower false-positive rate than partially tryptic peptides. Based on the Table 1, we determined new criteria (highlighted in the Table 1) with the goal of an overall estimated false-positive rate of less than 1%: ΔC_n score is >0.08; in the case of fully tryptic peptides, *Xcorr* should be larger than 2.0, 1.5, or 3.3 for charge states of +1, +2, +3, respectively; partially tryptic peptides must have an X_{corr} score of >3.0 (+2 state) or >4.0 (+3 state).

We next compared these new criteria with those previously delineated by Yates and colleagues.[14] Our dataset was processed (filtered) according to both criteria (Figure 5A). Our criteria resulted in an approximately 3-fold lower estimated false-positive rate (10.6% vs 30.8%) for identified proteins before manual validation, which decreased the requirement of manual interpretation. To further analyze the false-positives in identified proteins, the proteins were categorized by the number of unique peptides contributing to their identification (Figure 5B). The false-positive rate among proteins identified by a single peptide was 26% (188 out of 716). This rate decreased dramatically when two peptides were utilized in the identification to 1.4% (4 out of 279). Finally within this very large dataset, there were no false-positives detected for proteins identified by three or more peptides. Therefore, we manually confirmed proteins identified by one or two peptides instead of all proteins matched by four or fewer peptides.[14] A total of 216 (12.6% of 1720) proteins were discarded, which was consistent with the predicted false-positive rate (10.6% in Figure 5A). In our dataset, proteins identified by a single peptide after manual interpretation represented 33.9% (510 out of 1504) of the entire dataset that was significantly less than that in the dataset published previously (42.1%, 617 out of 1465).[14] Furthermore, while 858

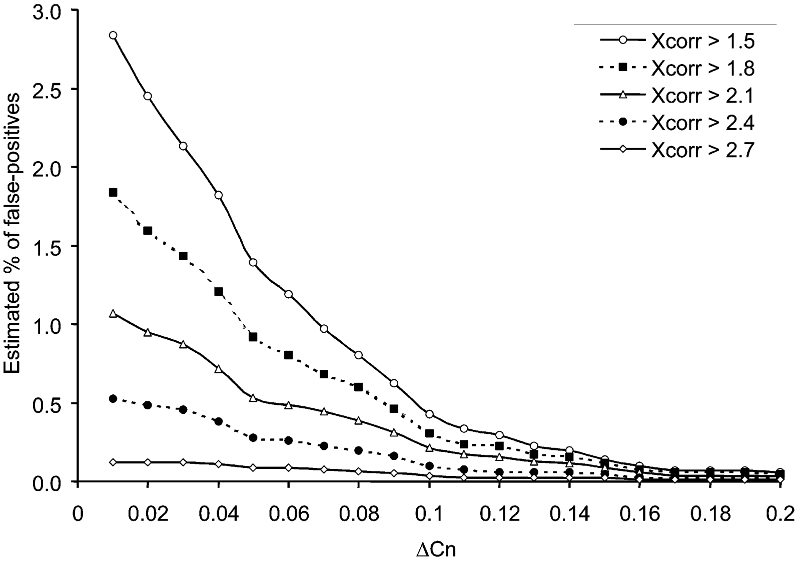


Figure 4. Evaluation of the effect of ΔC_n on the false-positive rate of peptide identifications. A total of 162 000 MS/MS spectra were searched against the chimeric forward- and reverse-yeast database. Sequest analysis results were filtered for fully tryptic peptide ends and a charge state of +2. The effect of increasing the ΔC_n cutoff at different X_{corr} values is shown. A false-positive was defined as the identification of a peptide from the reverse-database after filtering. The estimated false-positive rate was calculated as described in the text. The estimated percent of false-positives at an $X_{corr} > 1.5$ and a ΔC_n of >0.08 was less than 1% (170) with an *n* of more than 20 000.

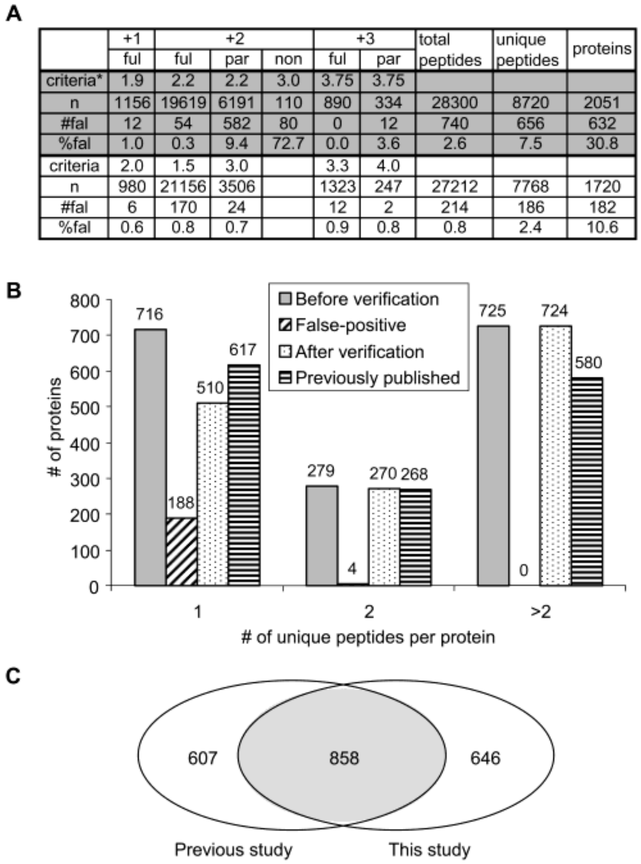
Table 1. Evaluation of the False-Positive Rate Using Different X_{corr} Cutoff Thresholds[a]

	ion state (+1)						ion state (+2)						ion state (+3)					
	ful		par		non		ful		par		non		ful		par		non	
x_{corr}	% fal	*n*	% fal	*n*	% fal	*n*	% fal	*n*	% fal	*n*	% fal	*n*	% fal	*n*	% fal	*n*	% fal	*n*
1.5	2.6	1713	19.7	1672	89.5	711	**0.80**	**21 156**	25.8	8371	98.8	6802	10.4	2433	78.3	4078	99.7	12 976
1.6	2.4	1605	16.9	1511	91.1	540	0.73	21 124	24.0	8159	98.6	6121	10.4	2430	77.8	4037	99.7	12 803
1.7	1.9	1456	14.3	1342	94.4	411	0.66	21 049	21.9	7895	99.1	5282	10.3	2426	77.9	3984	99.7	12 590
1.8	1.5	1295	12.9	1165	93.1	275	0.60	20 921	19.7	7598	99.2	4381	9.9	2413	77.7	3893	99.6	12 276
1.9	1.0	1156	9.8	977	93.9	179	0.55	20 744	17.0	7264	97.8	3566	9.6	2401	77.4	3771	99.7	11 832
2.0	**0.6**	**980**	8.1	815	94.2	121	0.46	20 475	14.3	6895	96.8	2770	9.2	2385	76.1	3615	99.8	11 282
2.1	0.7	848	8.0	650			0.39	20 099	11.9	6539	96.8	2109	8.6	2362	74.8	3421	99.7	10 553
2.2	0.6	682	7.0	513			0.28	19 619	9.4	6191	95.3	1581	8.2	2333	73.7	3173	100.1	9689
2.3	0.0	563	6.8	380			0.22	19 065	7.5	5803	93.4	1148	7.8	2290	70.7	2867	100.2	8650
2.4	0.0	454	5.8	275			0.20	18 417	5.8	5434	90.6	843	7.1	2244	67.4	2583	100.8	7580
2.5	0.0	364	6.7	208			0.09	17 722	4.5	5091	85.2	601	6.4	2167	64.4	2301	100.8	6359
2.6	0.0	270	6.6	152			0.09	16 964	3.4	4754	79.9	438	4.9	2079	60.6	1993	101.7	5136
2.7	0.0	199	8.9	112			0.06	16 170	2.4	4442	75.9	316	4.4	1978	55.7	1696	102.2	4031
2.8	0.0	150					0.03	15 374	1.9	4131	80.7	218	3.7	1873	48.7	1414	102.5	2985
2.9	0.0	111					0.01	14 563	1.3	3820	76.9	156	3.1	1757	41.9	1183	103.5	2167
3.0							0.00	13 734	**0.7**	**3506**	72.7	110	2.6	1638	35.2	999	104.7	1513
3.1							0.00	12 840	0.6	3238			2.0	1525	28.7	829	104.4	1027
3.2							0.00	11 903	0.3	2951			1.4	1419	24.4	713	102.9	657
3.3							0.00	10 954	0.2	2666			**0.9**	**1323**	19.6	613	102.1	433
3.4							0.00	10 059	0.1	2407			0.7	1218	14.3	531	105.9	270
3.5							0.00	9102	0.1	2156			0.5	1113	9.1	461	107.7	156
3.6							0.00	8243	0.1	1891			0.2	1016	6.5	403		
3.7							0.00	7431	0.1	1667			0.0	936	4.5	353		
3.8							0.00	6668	0.0	1471			0.0	855	3.8	319		
3.9							0.00	5868	0.0	1275			0.0	787	2.2	275		
4.0							0.00	5133	0.0	1110			0.0	717	**0.8**	**247**		
4.1							0.00	4483	0.0	954			0.0	641	0.0	229		
4.2							0.00	3913	0.0	832			0.0	582	0.0	210		
4.3							0.00	3411	0.0	700			0.0	523	0.0	193		
4.4							0.00	2915	0.0	596			0.0	491	0.0	175		
4.5							0.00	2451	0.0	493			0.0	448	0.0	156		

[a] ful, fully tryptic ends; par, partially tryptic ends; non, nontryptic; % fal, the percentage of false-positives; *n*, the sum of peptides based on the X_{corr} cutoff (shown in the first column). All peptides display ΔC_n of >0.08. The established thresholds are shown in boldface type.

proteins were found in both datasets, 607 proteins were unique to the published dataset and 646 proteins were found only in our dataset (Figure 5C).

We also examined the elution profile of the peptides with different solution charge states. At the pH utilized for the separation (pH 3.0), only lysine, arginine, histidine, and the amino-terminus contribute to the peptide solution charge state. Peptides with different solution charges were well separated mainly based on their ionic charge (Figure 6A) but also dependent on hydrophobicity (Figure 6B) during the elution.



A

	+1	+2			+3		total peptides	unique peptides	proteins
	ful	ful	par	non	ful	par			
criteria*	1.9	2.2	2.2	3.0	3.75	3.75			
n	1156	19619	6191	110	890	334	28300	8720	2051
#fal	12	54	582	80	0	12	740	656	632
%fal	1.0	0.3	9.4	72.7	0.0	3.6	2.6	7.5	30.8
criteria	2.0	1.5	3.0		3.3	4.0			
n	980	21156	3506		1323	247	27212	7768	1720
#fal	6	170	24		12	2	214	186	182
%fal	0.6	0.8	0.7		0.9	0.8	0.8	2.4	10.6

Figure 5. Refining Sequest search criteria with respect to reducing the number of false-positives. A total of 162 000 MS/MS spectra were searched against the chimeric database described in the text that contained yeast ORFs in both the forward and reverse direction. (A) Comparison of the criteria (marked by asterisk) previously described by Yates and colleagues (ref 14) and our new criteria from Table 1. While peptides with $\Delta C_n > 0.1$ were accepted in the previous criteria, we selected $\Delta C_n > 0.08$. Ion charge state (+1, +2, or +3) and tryptic state (ful, fully; par, partially; or non, nontryptic) are indicated. The sum of peptides ($n = n_{\mathrm{rev}} + n_{\mathrm{real}}$) in each category, the number of false-positives (#fal $= 2n_{\mathrm{rev}}$) and the percentage of false-positives (% fal) are also shown. After manual validation, we removed 216 proteins (12.6%) and accepted 1504 proteins. (B) Assignment of proteins according to how many unique peptides used in their identification for this study (1504 proteins) and one previously published (ref 14) (1465 out of 1486 proteins with extractable peptide information). (C) The overlap in the proteins detected in this study and those previously described (ref 14).

The number of peptides identified in each fraction was outlined in Figure 7A. The number of identified proteins increased steadily with the number of fractions analyzed and did not reach a plateau (Figure 7B). Although the slope appears to decreases gradually, if the curve is normalized according to the number of peptide identified in each SCX fraction (Figure 7A), it is clear that a plateau is not being reached (data not shown). To increase the throughput of the method, one strategy is to decrease the redundant MS/MS analysis of peptides from the same fraction. Analyzing every other fraction would reduce the analysis time by 50% and would still allow the identification of 84% of same proteins. Analyzing only every third fraction resulted in 74% of the same proteins being identified (Figure 7C).

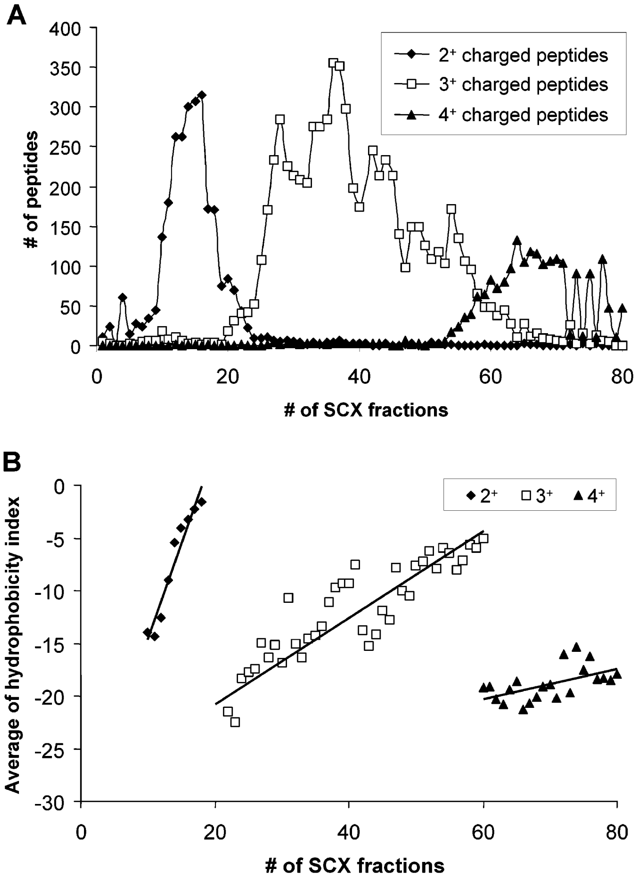


Figure 6. SCX separates peptides in a mixed mode. (A) Peptides identified in each SCX fraction with regard to their solution charge state. (B) The hydrophobic index average (ref 32) for all peptides in each fraction categorized by their charge state in solution. This shows that peptides were separated primarily by solution charge state and secondarily by hydrophobicity.

Discussion

Multidimensional chromatography coupled to tandem mass spectrometry (LC/LC−MS/MS) represents a promising alternative for proteome analysis. Multidimensional chromatography is most commonly accomplished by the combination of SCX and RP chromatography because of their orthogonal separation properties. In this report, we present the analysis of expressed yeast proteins by LC/LC−MS/MS techniques. More than 162 000 MS/MS scans were collected throughout the analysis with a final result of identification of 1504 proteins.

The sensitivity of peptide detection in LC−MS/MS is largely dependent on the concentration of the peptide eluted from the column, which is affected by the amount of sample loaded and the gradient length (or the slope of linear gradient). It is a challenge to optimize sample loading and the gradient length for each SCX fraction to maximize peptides detected by MS/MS. A common strategy is to maximize the sample loading and to run a longer gradient. However, when the sample is limited, a long gradient will decrease the detection sensitivity of many low abundant species and result in fewer peptides identified. On the other hand, overloading the column will cause peak broadening of abundant peptide species that will suppress the signal of coeluted minor species in electrospray. In the offline approach, the concentration of peptides in each fraction can be estimated based on the UV absorbance at a wavelength of

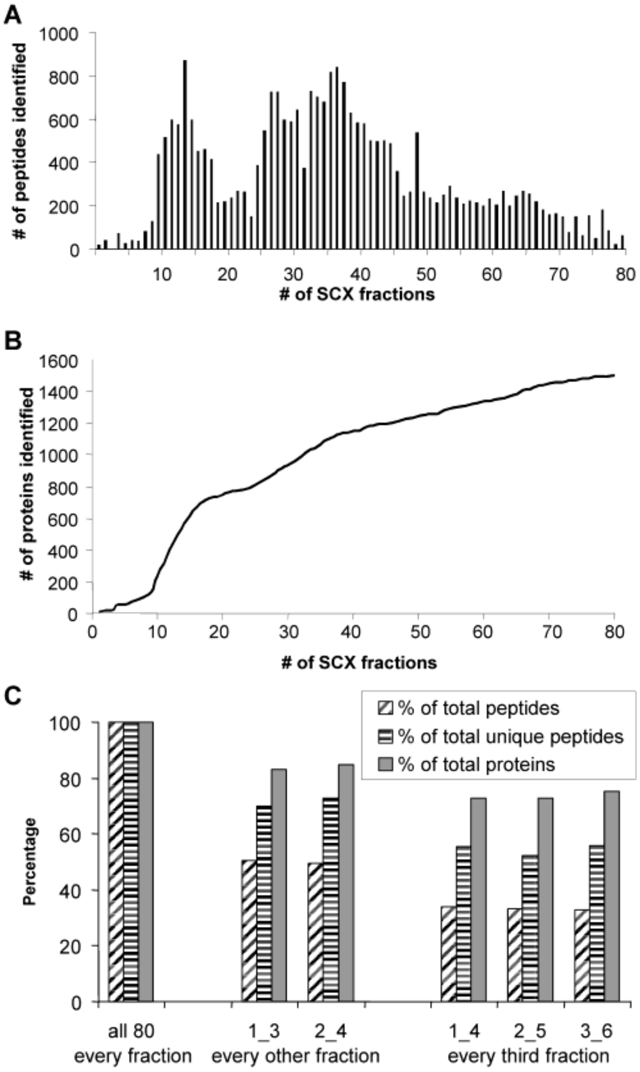


Figure 7. (A) Number of peptides identified in each SCX fraction. (B) Cumulative curve of proteins identified by LC−MS/MS. (C) Effect of analyzing only every other or every third fraction from the 80 SCX fractions on total number of peptides, unique peptide and proteins identified. The number of total peptides (26 815), unique peptides (7537) and proteins (1504) were normalized to 100%. Key: 1_3, all fractions with odd numbers; 2_4, all fractions with even numbers; 1_4, fractions 1, 4, 7, ...; 2_5, fractions 2, 5, 8, ...; 3_6, fractions 3, 6, 9, ...

214 nm. For most fractions, we loaded only one-half of the sample volume to maximize sensitivity. For several fractions with much higher peptide concentrations, some peak broadening effect was observed (data not shown) and we reduced the loading amount to 25%. To provide the ability for adjustment to the amounts of sample loaded and gradient length, we first analyzed every 8th fraction (fractions 8, 16, 24, 32, 40, 48, 56, 64, 72, and 80).

A two-dimensional separation and analysis of proteolyzed yeast protein has already been presented by Yates and colleagues.[14] They reported the analysis of 1486 yeast proteins by a combination of SCX and RP chromatography with peptide sequence analysis by tandem mass spectrometry. It should be noted that their report actually represents the pooled results of three separate experiments with each experiment being performed after differential protein extraction. Our results represent the results of a single experiment from a single yeast sample analyzed in an automated fashion. In addition, they utilized a biphasic column approach where the SCX and RP materials were contained in a single piece of fused silica. This did not allow for the use of detergents such as SDS or cholate in the lysis buffer. Our lysis buffer contained detergents that were removed by the SCX chromatography step and allowed for the detection of numerous membrane proteins (see the Supporting Information). A total of 858 proteins (about 60%) were identified by both approaches overall, while approximately 40% of proteins in the two datasets were different (Figure 5C). Furthermore, if we consider proteins identified by 1, 2, and greater than 2 peptides from our dataset with the previously published set, the proteins identified in common were 34%, 52%, and 75%, respectively. The reason for the relatively poor correlation between the two studies may come from several sources. (i) The previous dataset was based on the aforementioned pooling of samples. (ii) Because the sequencing technique is truly "shotgun", different peptide ions may have been selected. (iii) Manual interpretation was performed in each case that is also subject to some error. (iv) Finally, the strains of yeast were different and may have had different patterns of gene expression. In any case, more large-scale studies are needed for further comparisons.

We made use of a vented-column[29] for the automated analysis of the 80 SCX fractions. This technique allowed for nano-column (75−100 μm i.d.) loading at high flow rates (~5 μL/min) with subsequent peptide elution after closing the vent at low flow rates (~200 nL/min). Chromatography is greatly improved because the vented-column acts as a single column. The vented-column was especially useful in the analysis of SCX fractions because of the need for high flow washes to remove salts prior to analysis.

Manual interpretation of MS/MS spectra is often important to verify protein identifications derived from 2 peptides or less. With very large datasets such as the one in this study, manual interpretation of many thousands of peptide matches may be necessary. We therefore sought to evaluate the rate of false-positives after filtering database search results. We created a chimeric or composite database that contained protein sequences for all yeast ORFs in the forward direction followed by the same ORFs in a reverse-orientation from C terminus to N terminus. After performing a Sequest search of the 162 000 MS/MS spectra from this experiment against the composite database, we estimated the false-positive rate by determining number of peptides from the reverse-database (false sequence) that were identified. Because the number of peptides identified from the reverse-database was often small, it required a very large number of right-answer peptides to provide an accurate estimate. For example, 21 156 peptides were identified (85 from the reverse-database and 21 071 from the true database) when Sequest results were required to be fully tryptic, have an X_{corr} of >1.5, and display a ΔC_n of >0.08. The false-positive estimate was 0.8% [2 × 85/(21071 + 85)]. The previously utilized Xcorr cutoff for partially tryptic, doubly charged peptides with a ΔC_{n} > 0.1 was 2.2. This cutoff resulted in an estimated false-positive rate of nearly 10% (Figure 5A). In fact, this category let in nearly 600 peptides that would need to be removed by manual interpretation. By using an *Xcorr* cutoff of 3.0, this number was reduced to only 24. We generated new filtering criteria with the goal of an overall false-positive rate of <1%. The ΔC_n should be 0.08 or greater. In the case of peptides with fully tryptic ends, X_{corr} cutoff values of 2.0, 1.5, and 3.3 should be utilized for peptides with a charge state of +1, +2, and +3, respectively.

In the case of partially tryptic peptides, X_{corr} cutoff values of 3.0 and 4.0 for 2+ and 3+ peptides, respectively, are used. Partially tryptic peptides with a charge state of +1 and nontryptic peptides of any charge state are not accepted. If this dataset were filtered with the previously established criteria we would need to remove ~630 peptides compared to only ~180 by using the new criteria (Figure 5A). Furthermore, the new criteria resulted in identification of more proteins (1538 vs 1419) that were calculated by subtracting the estimated false-positives from the total proteins shown (Figure 5A).

We have evaluated an offline multidimensional chromatography peptide separation approach for the large-scale analysis of expressed yeast proteins. We identified 7537 peptides (1504 proteins) during a completely automated analysis. Our data suggest that additional protein or peptide separations would still be required to maximize the separation capacity to meet the challenge of the complexity of a higher eukaryotic proteome. For instance, enriching for cysteine-containing peptides before two-dimensional LC−MS/MS would decrease the peptide complexity, reduce the number of peptides identified for the same proteins and therefore improve throughput for protein identification.[31] Alternatively, removing the time factor could be accomplished by the deposition of RP-LC elutions directly onto plates for analysis by MALDI ionization.

Acknowledgment. We thank D. Moazed for providing the yeast strains and cell culture equipment. We also thank S. Gerber and C. Gartner for encouraging discussions. This work was supported in part by NIH Grant No. HG00041 (S.P.), the Giovani−Armenise Harvard Foundation (S.P.), and the Jane Coffin Childs Memorial Fund for Medical Research (J.P.).

Supporting Information Available: Table of unique peptide sequences. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) O'Farrell, P. H. *J. Biol. Chem.* **1975**, *250*, 4007−21.

(2) Gorg, A.; Obermaier, C.; Boguth, G.; Harder, A.; Scheibe, B.; Wildgruber, R.; Weiss, W. *Electrophoresis* **2000**, *21*, 1037−53.

(3) Pandey, A.; Mann, M. *Nature* **2000**, *405*, 837−46.

(4) Opiteck, G. J.; Lewis, K. C.; Jorgenson, J. W.; Anderegg, R. J. *Anal. Chem.* **1997**, *69*, 1518−24.

(5) Opiteck, G. J.; Jorgenson, J. W. *Anal. Chem.* **1997**, *69*, 2283−91.

(6) Tong, W.; Link, A.; Eng, J. K.; Yates, J. R., III. *Anal. Chem.* **1999**, *71*, 2270−8.

(7) Wall, D. B.; Kachman, M. T.; Gong, S. S.; Parus, S. J.; Long, M. W.; Lubman, D. M. *Rapid Commun. Mass Spectrom.* **2001**, *15*, 1649−61.

(8) Wall, D. B.; Kachman, M. T.; Gong, S.; Hinderer, R.; Parus, S.; Misek, D. E.; Hanash, S. M.; Lubman, D. M. *Anal. Chem.* **2000**, *72*, 1099−111.

(9) Yates, J. R., 3rd; Link, A. J.; Schieltz, D. *Methods Mol. Biol.* **2000**, *146*, 17−26.

(10) Aebersold, R.; Goodlett, D. R. *Chem. Rev.* **2001**, *101*, 269−95.

(11) Mann, M.; Hendrickson, R. C.; Pandey, A. *Annu. Rev. Biochem.* **2001**, *70*, 437−73.

(12) Peng, J.; Gygi, S. P. *J. Mass Spectrom.* **2001**, *36*, 1083−91.

(13) Link, A. J.; Eng, J.; Schieltz, D. M.; Carmack, E.; Mize, G. J.; Morris, D. R.; Garvik, B. M.; Yates, J. R., III. *Nat. Biotechnol.* **1999**, *17*, 676−82.

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

(15) Wolters, D. A.; Washburn, M. P.; Yates, J. R., III. *Anal. Chem.* **2001**, *73*, 5683−90.

(16) Washburn, M. P.; Ulaszek, R.; Deciu, C.; Schieltz, D. M.; Yates, J. R., III. *Anal. Chem.* **2002**, *74*, 1650−7.

(17) MacCoss, M. J.; McDonald, W. H.; Saraf, A.; Sadygov, R.; Clark, J. M.; Tasto, J. J.; Gould, K. L.; Wolters, D.; Washburn, M.; Weiss, A.; Clark, J. I.; Yates, J. R., III. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 7900−5.

(18) Liu, H.; Lin, D.; Yates, J. R., III. *Biotechniques* **2002**, *32*, 898−902.

(19) Sanders, S. L.; Jennings, J.; Canutescu, A.; Link, A. J.; Weil, P. A. *Mol. Cell. Biol.* **2002**, *22*, 4723−38.

(20) Ohi, M. D.; Link, A. J.; Ren, L.; Jennings, J. L.; McDonald, W. H.; Gould, K. L. *Mol. Cell. Biol.* **2002**, *22*, 2011−24.

(21) Eng, J.; McCormack, A. L.; Yates, J. R., 3rd *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976−89.

(22) Mann, M.; Wilm, M. *Anal. Chem.* **1994**, *66*, 4390−9.

(23) Clauser, K. R.; Baker, P.; Burlingame, A. L. *Anal. Chem.* **1999**, *71*, 2871−82.

(24) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20*, 3551−67.

(25) Zhang, W.; Chait, B. T. *Anal. Chem.* **2000**, *72*, 2482−9.

(26) Moore, R. E.; Young, M. K.; Lee, T. D. *J. Am. Soc. Mass Spectrom.* **2002**, *13*, 378−86.

(27) Gygi, P. M.; Licklider, L. J.; Peng, J.; Gygi, S. P. In *Protein Analysis: A Laboratory Manual*; Simpson, R., Ed.; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, 2002.

(28) Burke, T. W.; Mant, C. T.; Black, J. A.; Hodges, R. S. *J. Chromatogr.* **1989**, *476*, 377−89.

(29) Licklider, L. J.; Thoreen, C. C.; Peng, J.; Gygi, S. P. *Anal. Chem.* **2002**, *74*, 3076−83.

(30) Yates, J. R., III. *Electrophoresis* **1998**, *19*, 893−900.

(31) Gygi, S. P.; Rist, B.; Griffin, T. J.; Eng, J.; Aebersold, R. *J. Proteome Res.* **2002**, *1*, 47−54.

(32) Deber, C. M.; Wang, C.; Liu, L. P.; Prior, A. S.; Agrawal, S.; Muskat, B. L.; Cuticchia, A. J. *Protein Sci.* **2001**, *10*, 212−9.

PR025556V