Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Chromatography B, 810 (2004) 69–76

JOURNAL OF CHROMATOGRAPHY B

www.elsevier.com/locate/chromb

Characterizing complex peptide mixtures using a multi-dimensional liquid chromatography–mass spectrometry system: *Saccharomyces cerevisiae* as a model system

Dawn M. Maynard[a,b,*], Junichi Masuda[a], Xiaoyu Yang[a], Jeffrey A. Kowalak[a], Sanford P. Markey[a]

[a] *Laboratory of Neurotoxicology, National Institute of Mental Health, NIH, Bethesda, MD 20892-1262, USA*
[b] *Medical Genetics Branch, National Human Genome Research Institute, NIH, Bethesda, MD 20892-1851, USA*

Received 5 April 2004; accepted 15 July 2004
Available online 7 August 2004

Abstract

A rugged, reproducible, multi-dimensional LC–MS system was developed to identify and characterize proteins involved in protein–protein interactions and/or protein complexes. Our objective was to optimize chromatographic parameters for complex protein mixture analyses using automated peptide sequence recognition as an analytical end-point. The chromatographic system uses orthogonal separation mechanisms by employing strong cation exchange (SCX) in the first dimension and reversed phase (RP) in the second dimension. The system is fully automated and sufficiently robust to handle direct injections of protein digests. This system incorporates a streamlined post analysis results comparison, called DBParser, which permitted comprehensive evaluation of sample loading and chromatographic conditions to optimize the performance and reproducibility. Peptides obtained from trypsin digestion of a yeast soluble extract provided an open-ended model system containing a wide variety and dynamic range of components. Conditions are described that resulted in an average ($n = 4$) of 1489 unique peptide identifications, corresponding to 459 non-redundant protein sequence database records (SDRs) in the 20 μg soluble fraction digest.

Keywords: Multi-dimensional LC–MS; 2D-LC–MS/MS; Peptide mixtures; Sequence database records (SDRs); *S. cerevisiae*; DBParser

1. Introduction

A current trend in proteomic-based protein analysis is to identify different subsets of gene products using mass spectrometry. Smaller subsets of proteins, such as those involved in protein complexes and/or protein–protein interactions, can be enriched and effectively characterized. A variety of affinity-based strategies [1] can be utilized to isolate and purify protein complexes while sub-cellular fractionation can be used to monitor dynamic changes in the sub-cellular distribution of proteins. Sub-cellular fractionation produces proteins that range in complexity, but typically contain 100–400 proteins [2]. Tandem Affinity Purification employs serial affinity separations that yield mixtures containing 1–200 proteins, always accompanied by contaminating proteins [3]. To study these more highly defined cellular proteome sub-sets, we developed an automated two-dimensional LC–MS/MS system to detect and identify peptides from proteolytic digests of the protein mixtures [4].

The optimization of chromatographic conditions in a 2D-LC–MS/MS system for analyses of mixtures of unknown peptides requires testing verified by dataset analyses. Visual inspection of UV or reconstructed ion chromatograms (RICs) is a useful guide for the selection of parameters leading to an even distribution of analytes. However, LC–MS/MS

* Corresponding author. Building 10, Room 10C-107, MSC 1851, SHBG, MGB, NHGRI, NIH, Bethesda, MD 20892-1851, USA.
Tel.: +1 301 402 6622; fax: +1 301 402 7290.
E-mail address: maynardd@mail.nih.gov (D.M. Maynard).

1570-0232/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.jchromb.2004.07.015

experiments produce hundreds or thousands of MS/MS data files, and the comparison of quality of data requires objective endpoints. We used stringent data analysis criteria with the Mascot search engine [5] to identify unmodified tryptic peptides (no post-translational modifications), followed by automated comparisons using a program called DBParser, which sorts, analyzes and compares the peptides and protein sequence database records (SDRs) [6].

Initially, standard protein digests were used to evaluate system performance parameters such as retention time reproducibility and sensitivity [4]. We chose tryptic digests of protein extracts from *Saccharomyces cerevisiae* as a test model because we sought an abundant, consistent sample that contains a wide variety and dynamic range of proteins. Even though the soluble yeast proteome is more complex than one we would expect to encounter from TAP purification or isolation of sub-cellular components, the yeast genome has been sequenced and its proteome has been studied using other multi-dimensional separation systems [7,8], so that it provides a good benchmark for optimization and comparison of technologies.

2. Experimental

2.1. Materials and reagents

Yeast strain BY4741 was purchased from Research Genetics (Invitrogen Corporation, Carlsbad, CA, USA). YPD broth was purchased from Qbiogene Inc. (Carlsbad, CA, USA). Bovine serum albumin, equine apomyoglobin, bovine β-casein, ammonium formate (99.995%), ammonium sulfate and other standard chemicals used in this work were obtained from Sigma–Aldrich (St. Louis, MO, USA). Formic acid (98%) was purchased from Fluka (Buchs SG, Switzerland). HPLC-grade acetonitrile was purchased from Burdick & Jackson (Muskegon, MI, USA). Endoproteinase Lys-C was purchased from Roche Applied Science (Indianapolis, IN, USA) while modified porcine trypsin was purchased from Promega (Madison, WI, USA). The strong cation exchange column used was a PolySulfoethyl A (50 mm × 1 mm i.d., 5 μm, 200 Å) purchased from PolyLC (Columbia, MD, USA). The reversed phase column used was BetaBasic C18 (100 mm × 0.3 mm i.d., 5 μm, 150 Å) purchased from Thermo Hypersil-Keystone Scientific Operations (Bellefonte, PA, USA).

2.2. Growth and lysis of *S. cerevisiae*

Strain BY4741 was grown to mid-log phase ($OD_{600} = 1.1$) in YPD (YEPD) broth at 30 °C, according to a TAP purification protocol [9]. The pellet was washed with deionized water three times and stored at −80 °C.

Five grams of cells were solubilized in 20 mL lysis buffer containing 100 mM NH_4HCO_3 supplemented with protease inhibitors [10], vortexed, and disrupted using a Mini-Bead Beater [11] (Biospec Products, Bartlesville, OK, USA). The mixture was centrifuged (Sorvall RT7 Plus, rotor RTH-750) at 3500 rpm for 5 min at 4 °C to pellet the glass beads and cell debris. After disruption [12], lysate from the Bead Beater was centrifuged (Sorvall RC-5B Refrigerated Superspeed Centrifuge with SS-34 rotor, Kendro Laboratory Products, Asheville, NC, USA) at 5 °C and 12,000 g (10,000 rpm) for 10 min. Supernatants were collected, combined and designated S1. The sample was adjusted to pH 8.0 with 100 mM ammonium bicarbonate (NH_4HCO_3) and the protein concentration was determined using the BioRad Protein Assay using BSA as standard [13]. Aliquots were stored at −80 °C.

2.3. Digestion and preparation of yeast fractions for LC–MS

A sample of S1 extract was denatured with 8 M urea in 0.4 M NH_4HCO_3. Disulfide bonds were reduced with 3 mM DTT, and then carbamidomethylated with 5 mM iodoacetamide [14]. Then the sample was diluted 4-fold and endoproteinase Lys-C was added to a final substrate-to-enzyme ratio of 100:1 (w/w). The sample was incubated for 15 h at 37 °C, after which, modified porcine trypsin was added at a final substrate-to-enzyme ratio of 50:1 (w/w) and incubated for 8 h at 37 °C. Aliquots of digested S1 were stored at −20 °C.

S1 digest was subsequently thawed to room temperature and the calculated volumes for 5-35 μg were added to individual 1.5 mL Eppendorf tubes. For sample preparation designated +AmBicarb, each aliquot was dried for 30 min on the Speed Vac (Automatic Environmental Speed Vac(R) System AES2010, ThermoSavant, Holbrook, NY, USA) at room temperature (low). For sample preparation designated −AmBicarb, 50 μL deionized H_2O was added to each previously dried sample, after which it was vortexed and dried for 60 min on the Speed Vac. Then, 25 μL deionized H_2O was added to each sample, after which it was vortexed and dried for 60 min on the Speed Vac. Each sample was re-constituted in 40 μL SCX-A buffer for 2D-LC and transferred to a Shimadzu auto-sampler vial.

2.4. 2D-LC–MS/MS technology

The fully-automated 2D HPLC system [4] was built using LC-VP Series components, consisting of two SCL-10AVP controllers, five LC-10ADVP pumps with micro flow control kit, a SIL-10ADVP automatic injector, a CTO-10ACVP column oven, maintained at 30 °C, and a SPD-10AVP UV detector (all from Shimadzu Corporation, Kyoto, Japan). Six peptide CapTraps (0.5 mm × 2 mm i.d., 0.5 μL) (Michrom BioResource Inc., Auburn, CA, USA) mounted on two 6-position rotary valves (FCV-14AH) were used as trapping columns. Two additional 2-position switching valves (FCV-12AH) were used as solvent selectors for trapping, desalting and loading samples onto the reverse phase column. The fused silica spray capillary was a non-coated New Objectives TaperTip capillary (50 cm, 360 μm o.d., 50 μm i.d., 50 μm i.d. tip) which led directly into a ThermoFinnigan LCQ Classic ESI-ion trap mass spectrometer (San Jose, CA, USA).

The auto-sampler was used to inject samples onto the SCX column, after which they were eluted onto six peptide cap traps using a stepwise gradient of 0, 1, 5, 10, 30, and 100% SCX-B (each 5 min, 10 bed volumes) at 80 μL/min. Peptides on the 6 cap traps were desalted using RP-C at 80 μL/min and then eluted sequentially onto the RP column and into the mass spectrometer using the following program: 10% B (3 min), a linear gradient of 10–60% B (80 min), 60–80% B (10 min), 80% B (2 min) at 10 μL/min. Mobile phase buffers were, for SCX-A, 10 mM ammonium formate buffer (HCO_2NH_4/HCO_2H), pH 3.8; for SCX-B, A + 100 mM ammonium sulfate ($(NH_4)_2SO_4$); for RP-A, water/acetonitrile/formic acid = 94.9/5/0.1 (v/v); for RP-B, water/acetonitrile/formic acid = 19.9/80/0.1 (v/v); and for RP-C, water/formic acid = 99.9/0.1 (v/v). The SCX mixer volume was 10 μL and the RP mixer volume was 2 μL.

The LCQ was operated in positive ion mode with dynamic exclusion set to repeat count = 2, repeat duration = 0.35 min, exclusion duration = 1 min, exclusion mass width = 3 amu. Spectra were acquired in a data dependent manner with the top five most intense ions in the MS scan selected for MS/MS.

2.5. Database searching

Raw MS/MS files were submitted to the NIH Mascot [5] Cluster using Mascot Daemon. Data were searched against the SwissProt_Trembl database using a restricted taxonomy of *S. cerevisiae* (baker's yeast), enzymatic cleavage = trypsin, fixed modification = carbamidomethyl (C), variable modification = oxidation (M), monoisotopic mass, peptide tolerance = 1.5 Da, MS/MS tolerance = 0.8 Da, two missed cleavages, charge state = 1+, 2+, and 3+, instrument = ESI − TRAP.

2.6. Mascot output analysis using DBParser

DBParser [6] version 2.0 is a perl program that takes the output from Mascot flat files, stores data in a MySQL database, and generates user-friendly html output reports, which can be used for subsequent analysis and comparison. Peptide identifications included in DBParser output files are only those whose individual ions scores meet or exceed their Mascot identity score thresholds. When the Mascot ions score exceeds the identity score, there is less than 5% probability of the match being a random event, and a logarithmic relationship decreasing that probability as the difference between ion score and identity score grows. This stringent criterion was chosen because spectra interpreted manually consistently verified the automated Mascot selection when the ion scores exceeded the identity scores by several log values. Peptide identifications whose ions scores fell below the Mascot identity score but above the homology threshold were prone to mis-identification such that their inclusion in an automated parsing and evaluation strategy could not be accepted consistently. Peptide identifications were automatically rejected if their ions scores fell below the Mascot homology threshold.

DBParser was used to generate two reports with lists of peptides or proteins unique to or common between particular samples. The DBParser *Single Report* takes a single file or a group of files and produces a list of the non-redundant protein SDRs and the unique peptides in each protein SDR. The *Multiple Comparison Report* compares two to six files or groups of files, but does not compare the rejected protein SDRs. When tallying the number of identifications, only the unique (non-redundant) peptides were counted. Protein SDR identifications were made based on the unique peptide identifications. Only non-redundant protein SDRs were counted.

3. Results

3.1. Evaluating the microspray 2D-LC–MS/MS system

We examined some basic parameters of the system; namely optimal salt percentage used to elute the peptides, optimal loading amount, detection reproducibility and peptide carry-over between fractions. The optimum percentage of salt used to elute the peptides in each SCX fraction was determined by visual inspection of the RP reconstructed ion chromatograms (RICs) for the even distribution of peptides as well as by the maximum number of unique peptide identifications, as determined by DBParser. Fig. 1 shows the RICs from the RP analysis under optimized SCX conditions, in which the difference in percent salt used to elute the peptides is small between fractions initially (0, 1, 5, 10, 30, and 100% SCX-B). The first fraction, 0% SCX-B, contains peptides not bound to the SCX resin, or flow-through. Under these conditions, a 20 μg S1 digest yielded 24% more unique peptide identifications than it did when separated under conditions in which the difference in percent salt used was large (0, 10, 20, 30, 50, and 100% SCX-B) (data not shown).

Second, we determined the optimal loading amount of the system given six salt fractions in the first dimension and a 90 min RP gradient in the second dimension. Because

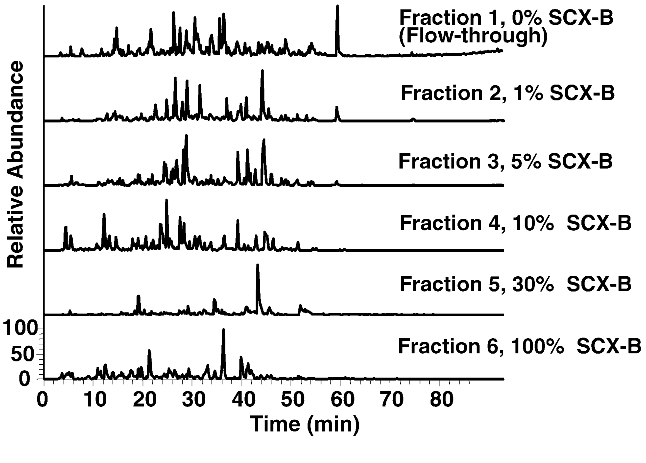

Fig. 1. Chromatographic fractionation of peptides from a 20 μg S1 digest.

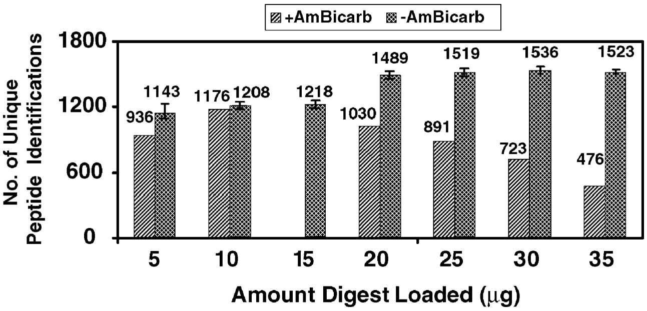

Fig. 2. The number of unique peptides identified as a function of increasing digest amount and different sample preparations. The number of unique peptides was determined using DBParser's Single Report. Sample preparation affects binding of peptides to the SCX column and therefore the number of detected peptides. In +AmBicarb, samples were dried on the Speed Vac, re-constituted in SCX-A buffer, and analyzed by 2D-LC–MS/MS ($n = 1$). In −AmBicarb, dried samples were subjected to successive water additions and evaporations to remove residual/remaining ammonium bicarbonate from the digestion buffer ($n = 4$; 15 μg sample was not run using +AmBicarb).

peptides are analyzed in the mass spectrometer, the maximum number of unique peptide identifications in each salt fraction was used to indicate the upper limit of the peak capacity (the ability to separate complex samples into as many peaks as possible) and therefore the optimal loading amount of the system. Digest amounts corresponding to 5–35 μg peptides from the S1 digest were used for this purpose. The calculated volume of digest was dried in the Speed Vac, re-dissolved in SCX-A buffer and loaded onto the 2D-LC–MS/MS system for detection and analysis. We observed that the number of unique peptide identifications increased from 5 to 10 μg, but with greater quantities the number of identifications decreased sharply. We had expected to see a plateau in the number of SDR identifications as the amount of digest increased, indicating a separation capacity limit and reasoned that remaining buffer from the LC sample preparation caused the unexpected phenomenon [15]. Either a high amount of residual ammonium ions (NH_4^+) competed with binding of positively charged peptides to the SCX column (ionic strength too high) or the loading buffer increased the pH such that peptides were not protonated adequately for binding. Consequently, digests were dried in the Speed Vac and subjected to successive cycles of deionized water addition and evaporation to facilitate removal of residual/remaining ammonium bicarbonate. Samples were again injected in 5–35 μg amounts ($n = 4$) and results shown in Fig. 2 (−AmBicarb) compared with the previously described method, (+AmBicarb). The average number of peptide identifications increased for each amount loaded when ammonium bicarbonate in the digest sample was decreased. A plateau was reached around 20 μg, after which no appreciable gain in the number of identifications was observed. On average, 1489 peptide identifications corresponding to 459 protein SDRs were detected in the 20 μg S1 digest.

To determine chromatographic and detection reproducibility of the 2D-LC–MS/MS system, the 20 μg S1 digest was analyzed four times and the resulting MS/MS files were submitted for database searching. Peptides from each salt fraction (fractions 1–6) in runs 1–4 were compared simultaneously using DBParser's multiple comparison report to determine the number of peptides unique to and common between the four runs. Results in Table 1 indicate the total number of peptides in each fraction is similar. A unique set of peptides was detected in the *same fraction* from *different runs*, demonstrating that each salt fraction is still highly complex and not all peptides are resolved or detected consistently in the second RP dimension. However, the number of common peptides between all four runs of the same fraction is high, indicating good chromatographic reproducibility and adequate MS sampling of *most* of the peptides in each fraction under the current conditions. Fig. 3 shows the retention time reproducibility of a peptide from translationally controlled tumor protein homolog (TCTP) in the same salt fraction from four different runs. The average retention time RSD for this peptide is 0.5% and the range is 0.5–2% for the majority of peptides analyzed.

Because the peptide mixture complexity is high in each salt fraction and it is known that the same components can be partitioned into more than one fraction in a multi-dimensional LC–MS experiment, we investigated the peptide carry-over in *adjacent salt fractions* from a 20 μg S1 digest on our 2D-LC–MS/MS system. This was accomplished by performing pair-wise comparisons of the detected peptides between adjacent fractions, using DBParser's multiple comparison report, to determine those unique to and common between the fractions. When the difference in percent salt used to elute the peptides from the SCX column is smaller between fractions, as it is between fractions 1–2, 2–3, 3–4, more peptides are found in common between the fractions (average of 58% with

Table 1
Run-to-run comparison of the number of unique peptide identifications in each salt fraction

Salt fraction	#Peptides found only in run 1	#Peptides found only in run 2	#Peptides found only in run 3	#Peptides found only in run 4	# Peptides common in all four runs
1	46	45	37	47	174
2	65	54	64	56	259
3	59	64	60	47	282
4	89	57	63	60	264
5	92	36	40	47	160
6	53	27	36	30	132

DBParser's multiple comparison report was used to generate lists of peptides unique to a particular run. Those peptides in common between all runs are listed in the common column. There are more common peptides between all four runs than there are unique peptides in each run, which indicates good reproducibility.

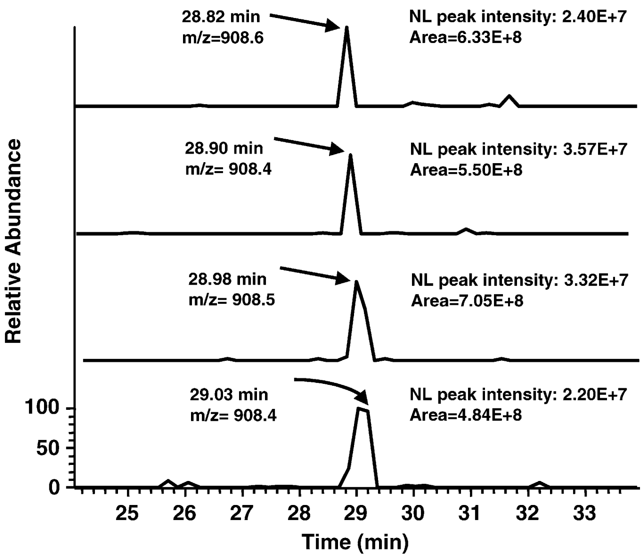

Fig. 3. Reconstructed ion chromatograms plotting the mass range 908.3–908.6 in fraction 1 from runs 1–4 of the 20 μg S1 digest. The peptide corresponding to the base peak at 908.5 (±1.5 Da) is DIFSNDELLSDAYDAK, from TCTP_Yeast, as determined by Mascot from the MS/MS spectrum (data not shown). The retention time RSD = 0.5% for this peptide. On average, the retention time RSD range was 0.5–2% for those peptides examined. The normalized (NL) peak intensity and area under the curve are shown. The area RSD = 16%.

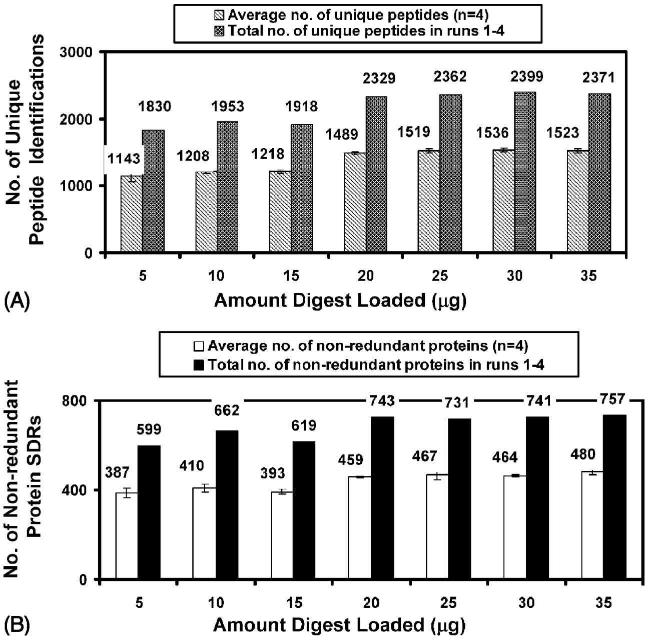

Fig. 4. (A) The average number of unique peptide identifications ($n = 4$) and the total number of unique peptide identifications, by combining results from runs 1–4, as a function of digest amount loaded. (B) The average number of non-redundant protein SDR identifications ($n = 4$) and the total number of non-redundant protein SDR identifications, by combining results from runs 1–4, as a function of digest amount loaded. All values obtained using DBParser's Single Report.

respect to the first fraction being compared). However, when the difference in percent salt is greater between fractions, as it is between fractions 4–5 and 5–6, fewer peptides are found in common and more are unique to each adjacent fraction (approximately 13% with respect to the first fraction being compared). This trend is independent of the amount of digest loaded on the system. Larger salt cuts (0, 10, 20, 30, 50, and 100% SCX-B) produce fewer common peptides between early fractions but identify fewer peptides overall than our optimized conditions.

We also determined how many peptides were carried over in *non-adjacent fractions*; that is, how many peptides identified in fraction 1 were observed in subsequent fractions. A pair-wise comparison indicated those peptides unique to and common between fraction 1 and fractions 2–6 from the same 20 μg yeast sample. The percentage of peptides in common between fraction 1 and fractions 2–6 (with respect to fraction 1) was 71, 40, 20, 6, and 2%, respectively. The percent of common peptides decreases as the percent difference between salt cuts increases to 10, 30 or 100% SCX-B as in fractions 1–4, 1–5, 1–6. This indicates that the carry-over from fraction 1 through subsequent fractions is high initially and decreases as the percentage of salt increases dramatically. This is in agreement with the trend observed in adjacent fractions.

Those peptides unique to each fraction demonstrate that the S1 digest is a highly complex mixture. However, we wanted to test whether those peptides that are unique to each of the four runs could be used advantageously. Therefore, the data from all four of the 20 μg runs were combined and a single report was generated using DBParser. The *average* number of peptides and the new *total* number of peptides as a function of digest amount are shown in Fig. 4A, along with the corresponding average number of proteins and total number of proteins in Fig. 4B. We observe a 43–47% increase in the number of unique identifications for both peptides and protein SDRs by combining the results from all four runs. This indicates that repeated analyses of samples with complexities exceeding column separation efficiency is valuable, in that additional data can be obtained. The number of identified peptides and therefore protein SDRs increases significantly by repetition because the mixture complexity exceeds the spectral sampling rate.

The total number of identified protein SDRs increases as a function of the amount of digest injected until 20 μg, after which the number of identifications reaches a plateau, concomitant with the number of peptide matches. Fig. 4B demonstrates that the biggest increase in the number of protein SDR identifications occurs from 15 to 20 μg. Using the total data from all four runs, there were over 450 protein SDRs in common between the two amounts loaded, while the number of protein SDRs unique to the 20 μg sample almost doubled from the 15 μg sample.

To understand this trend, we determined the average number of unique (non-redundant) peptides used to make protein SDR identifications for each amount loaded (see Fig. 5). Approximately 50% of protein SDRs were identified with

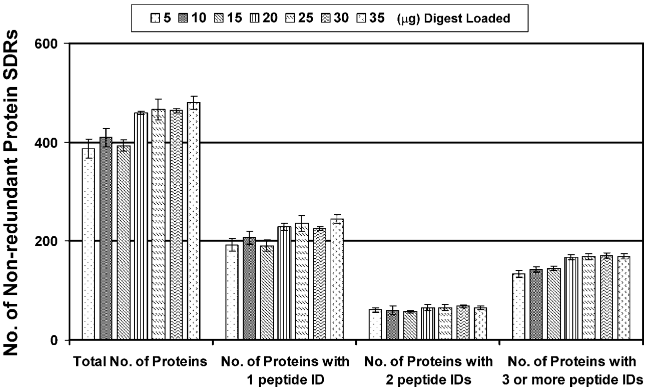

Fig. 5. The average number of non-redundant protein SDR identifications as a function of digest amount, and the number of protein SDRs identified with one, two, and three or more peptides. In general, more protein SDRs are identified with higher amounts of loaded digest. Approximately, 50% of protein SDR identifications are made from just one peptide identification.

a single peptide, while the remaining 50% of protein SDRs were identified by either two or three or more peptides. The general trend for the number of identifications per protein was one peptide > three or more peptides > two peptides, which was consistent across all amounts of digest loaded. The protein SDR with the greatest number of unique peptide matches in the 20 μg samples was pyruvate kinase 1 (KPY1) with 30 unique peptides, followed closely by elongation factor 2 (EF2) with 29 unique peptides and heat shock protein SSA2 (HS72) with 26 unique peptides. Compared with the 15 μg sample, almost 90% of the protein SDRs *unique* to the 20 μg samples were identified by just one peptide.

This trend did not change when determining the number of peptides per protein SDR for the *total* protein SDR count (sum of runs 1–4). That is, 50% of the total protein SDR identifications were made using one peptide, while 40% were made with three or more peptides and 10% were made with two peptides. Combining the results from multiple runs does not introduce bias in terms of the number of peptides used to make the identifications.

Because protein identifications based on a single peptide comprise almost 50% of the total protein SDRs, we determined how reproducible these peptide identifications were in multiple runs of the same sample. Table 2 shows protein SDR level comparisons of each salt fraction using DBParser. The number of protein SDRs common in all four of the 20 μg runs was high, indicating good chromatographic and detection reproducibility. Of those common protein SDRs, 50–60% were made using one discrete peptide [6] (a peptide that is assigned to one and only one protein SDR). High quality single peptides were observed consistently in multiple runs, giving higher confidence to the corresponding protein SDR assignments based on one peptide.

4. Discussion

Chromatographic conditions, such as the percentage of salt used to elute the peptides, were optimized to maximize the number of unique (non-redundant) peptide identifications in the yeast soluble protein lysate. Sample preparation was essential to improve and increase the number of identifications as the amount of loaded digest increased. Although solid phase extraction is not required prior to sample injection, residual ammonium bicarbonate must be removed prior to sample loading on the system to avoid inadvertent elution of peptides from the ion exchange column. The optimal loading amount was found to be 20 μg of lysate. Injection of greater quantities did not produce appreciably more peptide identifications or protein SDR matches.

Excellent chromatographic and detection reproducibility was observed for those peptides found in common in the same fraction from multiple runs. The average retention time reproducibility is 0.5–2% RSD. This is notable considering the high complexity of the mixture, demonstrated by the number of unique peptides detected in the same fraction from multiple runs. Retention time variances occur mainly when ions elute near the boundaries of a particular MS time window [16]. Dynamic exclusion parameters specifying MS/MS selection may also affect which peptides are detected with high quality in a consistent manner, because once selected, a given *m*/*z* will not be repeated for MS/MS analysis during a specified time window (1 min). As a result, peptides may be detected at the beginning or end of a chromatographic peak (for example, see typical peak in Fig. 3). The survey scan parent ion intensity triggering an MS/MS event may not correspond to the maximum or optimum intensity for fragmentation and

Table 2
Reproducibility of protein SDR identifications made from just one peptide

Salt fraction	#Protein SDRs common in all four runs	Percentage of common protein SDRs identified by just one peptide (%)
1	106	60
2	139	56
3	147	58
4	138	52
5	85	53
6	81	57

DBParser's multiple comparison report was used to generate lists of protein SDRs identified in all four runs of the 20 μg S1 digest (common), indicating that the peptides identifying these protein SDRs are reproducibly detected. The percentage of these common protein SDRs identified by just one peptide ranges from 50 to 60%.

identification [17]. A faster MS sampling rate should profile the peaks better in MS mode.

From visual inspection of chromatograms, it appears that peptides are well distributed between six fractions. However, comparison of the peptides between adjacent and non-adjacent fractions using DBParser reveals there is also peptide carry-over. That is, certain peptides are localized in particular SCX fractions without carry-over, but other peptides elute in adjacent fractions (e.g., Fr 1–2, 2–3, etc.). Carry-over correlates directly with the percent salt difference between fractions. When this difference is small, the boundaries on the SCX column are not sharp and higher peptide carry-over results. Increasing the elution (trapping) time for the SCX mode did not increase the number of identifications. Even though experiments conducted using larger incremental salt steps identified fewer common peptides between fractions, fewer peptides were detected overall (data not shown). Therefore, step conditions were optimized to promote higher overall peptide recognition.

In order to understand peptide carry-over between non-adjacent fractions, peptides common between fractions 1 and 5 and fractions 1 and 6 were examined more closely to see if peptide physical properties could explain this behavior. While several peptides found in late eluting fractions are long (28–34 residues) and have a significant number of hydrophobic residues, their calculated p*I* values did not predict their fraction of elution in SCX mode. We observed three different types of peptides in the yeast soluble proteome. Some peptides are well behaved on SCX and RP (well fractionated on SCX, Gaussian distribution on RP), some peptides are well behaved on SCX but poorly resolved on RP (well fractionated on SCX, adsorptive tailing on RP), and other peptides are poorly resolved on both SCX and RP (poorly fractionated on SCX, adsorptive tailing or elution over broad range on RP). As expected, fractionation depends on the physical properties of each peptide, which determine the types of interactions with the SCX column. Factors that affect peptide fractionation include local basic regions, mixed modes of electrostatic and hydrophobic interactions with the column, or peptide aggregation. While the chosen set of conditions is successful for many of the peptides in the yeast soluble lysate, it cannot be optimized for every peptide present in such a complex mixture. Even peptides from the same protein behave very differently as evidenced by the different peak areas observed in digests of pure proteins. No single set of conditions is ideal for detecting and identifying all peptides therefore a general optimized strategy was selected.

A benchmark for multi-dimensional LC–MS/MS analysis of complex mixtures is Multidimensional Protein Identification Technology (MudPIT), a technique described by Washburn and co-workers [7]. Briefly, those researchers generated three fractions of *S. cerevisiae* using differential extraction and separately analyzed them using a 15-step MudPIT analysis. Results were combined to report 5540 peptide identifications corresponding to 1484 proteins. An offline multidimensional technique reported by Peng et al. [8] utilized 80 SCX fractions followed by RP-LC–MS/MS. They reported a total of 7537 unique peptides corresponding to 1504 proteins in a yeast soluble fraction. On average, we detected and identified 1489 unique tryptic peptides corresponding to 459 non-redundant protein SDRs in a six-step fractionation of a 20 μg yeast soluble fraction. This clearly meets the requirements for peptide/protein detection and identification in samples containing less than several hundred proteins, typical of tandem affinity purifications or separated subcellular organelles.

Total peptide identifications and protein SDR counts were increased substantially by combining results from all four runs of the 20 μg sample. By combining information from multiple runs [17], peptides that were identified in run 1 can be grouped with those from runs 2–4, which can increase the number of peptides used to make any one particular protein SDR assignment. Though no major change in the distribution of peptides per protein SDR was observed when comparing the average peptide and protein SDR results with the total peptide and protein SDR results, the total numbers increased significantly.

In terms of protein SDR identification, the largest increase in identifications occurs between 15 and 20 μg, after which no significant increase is observed. When comparing the peptide identifications and protein SDRs unique to and common between these two samples, almost 90% of the protein SDRs *unique* to the 20 μg samples are identified by one peptide. This would explain why the largest jump in peptide identifications corresponds to the largest jump in protein SDR identifications. The assignment of additional peptide identifications apparently corresponds to additional MS/MS spectra collected when parent ion intensity threshold requirements have been reached.

Most protein laboratories require some additional experimental verification before reporting protein SDRs identified by a single peptide. A single peptide identification could be based upon one irreproducible MS/MS spectrum and may be spurious. This is troubling considering that almost 50% of the protein identifications are based upon one peptide, regardless of the amount loaded. There are two ways to gain confidence in single peptide identifications. First, if this *discrete* peptide (a peptide that is assigned to one and only one protein SDR) is identified in the same fraction from more than one run of the same sample with a high quality MS/MS spectrum, confidence is increased that this peptide, and the resulting protein SDR, is present. This is precisely what we found when comparing the proteins in common between four runs of a 20 μg S1 digest. Between 50 and 60% of the common proteins are identified with one discrete peptide meaning that Mascot identified these peptides in four different runs. Based upon the Mascot peptide identifications, our confidence is high that these proteins are present in the mixture. Manual inspection of these MS/MS spectra supported the automated search results and comparison. Secondly, the recorded MS/MS data files may match additional peptides from post-translationally modified proteins that would confirm the protein SDR with a second peptide. All of the results reported in this study utilized Mascot searches allowing no post-translational modifications or non-tryptic cleavages. Further data mining would be required, but it may support protein SDRs made from single peptide identifications with confirming data.

5. Conclusions

An automated 2D-LC–MS/MS system can be optimized with respect to maximizing peptide identifications. Successive cycles of water addition and subsequent evaporation removed residual volatile buffer from the protein digests so there was no need for solid phase extraction of samples prior to their injection and sample losses could be minimized. Due to the implementation and copious desalting of the reverse phase traps between the two separation columns, a wide variety of salts are permitted in the first dimension. This automated system allows samples to run continuously and is limited only by the number of spaces in the auto-sampler tray.

Based on the results shown, we detected peptides reproducibly and reliably from complex mixtures exceeding 400 proteins. Compared with other published reports, our system used relatively low quantities and produced a high number of peptide identifications for the yeast soluble fraction. There is a significant advantage to combining mass spectral files from repeated analyses when protein mixture complexity exceeds the separation capacity of an analytical system. A 45% increase in the number of peptide and protein SDR identifications was observed when combining results from four runs of the same 20 μg sample. Combined with automated database searching and DBParser filtering, we report an efficient and streamlined way to analyze and identify peptides, and therefore proteins, in complex mixtures.

Acknowledgements

The authors would like to thank the following people for their helpful contributions to this manuscript: Susan Chacko, Cai Yun Chen, Vijay Dondeti, Jonathan Epstein, Lewis Geer, and Edward Unsworth.

References

[1] A. Bauer, B. Kuster, Eur. J. Biochem. 270 (2003) 570.
[2] M. Dreger, Eur. J. Biochem. 270 (2003) 589.
[3] J.M. Cronshaw, A.N. Krutchinsky, W. Zhang, B.T. Chait, M.J. Matunis, J. Cell Biol. 158 (2002) 915.
[4] J. Masuda, D.M. Maynard, M. Nishimura, T. Ueda, J.A. Kowalak, S.P. Markey, J. Chromatogr. A, submitted.
[5] D.N. Perkins, D.J. Pappin, D.M. Creasy, J.S. Cottrell, Electrophoresis 20 (1999) 3551.
[6] X. Yang, V. Dondeti, R. Dezube, D.M. Maynard, S.P. Markey, L.Y. Geer, J. Epstein, X. Chen, J.A. Kowalak, J. Proteome Res., in press.
[7] M.P. Washburn, D. Wolters, J.R. Yates 3rd, Nat. Biotechnol. 19 (2001) 242.
[8] J. Peng, J.E. Elias, C.C. Thoreen, L.J. Licklider, S.P. Gygi, J. Proteome Res. 2 (2003) 43.
[9] G. Rigaut, A. Shevchenko, B. Rutz, M. Wilm, M. Mann, B. Seraphin, Nat. Biotechnol. 17 (1999) 1030.
[10] D.G. Drubin, K.G. Miller, D. Botstein, J. Cell Biol. 107 (1988) 2551.
[11] http://www.mc.vanderbilt.edu/vumcdept/cellbio/gould/html/taptag.pdf.
[12] M.P. Molloy, B.R. Herbert, B.J. Walsh, M.I. Tyler, M. Traini, J.C. Sanchez, D.F. Hochstrasser, K.L. Williams, A.A. Gooley, Electrophoresis 19 (1998) 837.
[13] http://www.biorad.com/LifeScience/pdf/Bulletin_9004.pdf.
[14] K.L. Stone, K.R. Williams, in: P. Matsudaira (Ed.),Academic Press Inc., San Diego, 1993, p. 55.
[15] T. Le Bihan, H.S. Duewel, D. Figeys, J. Am. Soc. Mass Spectrom. 14 (2003) 719.
[16] M.T. Davis, J. Beierle, E.T. Bures, M.D. McGinley, J. Mort, J.H. Robinson, C.S. Spahr, W. Yu, R. Luethy, S.D. Patterson, J. Chromatogr. B, Biomed. Sci. Appl. 752 (2001) 281.
[17] Y. Shen, J.M. Jacobs, D.G. Camp 2nd, R. Fang, R.J. Moore, R.D. Smith, W. Xiao, R.W. Davis, R.G. Tompkins, Anal. Chem. 76 (2004) 1134.