Published in final edited form as:

Anal Chem. 2010 February 15; 82(4): 1234–1244. doi:10.1021/ac9021083.

Size-Sorting Combined with Improved Nanocapillary-LC-MS for Identification of Intact Proteins up to 80 kDa

Adaikkalam Vellaichamy^{1,2}, John C. Tran¹, Adam D. Catherman¹, Ji Eun Lee¹, John F. Kellie¹, Steve M.M. Sweet¹, Leonid Zamdborg^{1,2}, Paul M. Thomas^{1,2}, Dorothy R. Ahlf^{1,2}, Kenneth R. Durbin¹, Gary A. Valaskovic³, and Neil L. Kelleher^{*,1,2}

¹Department of Chemistry, 600 S. Mathews Ave.

²Institute for Genomic Biology, 1206 W. Gregory Dr., University of Illinois at Urbana-Champaign, Urbana, IL 61801.

³New Objective Inc., 2 Constitution Way, Woburn, MA 02139.

Abstract

Despite the availability of ultra-high resolution mass spectrometers, methods for separation and detection of intact proteins for proteome-scale analyses are still in a developmental phase. Here we report robust protocols for on-line LC-MS to drive high-throughput top-down proteomics in a fashion similar to bottom-up. Comparative work on protein standards showed that a polymeric stationary phase led to superior sensitivity over a silica-based medium in reversed-phase nanocapillary-LC, with detection of proteins >50 kDa routinely accomplished in the linear ion trap of a hybrid Fourier-Transform mass spectrometer. Protein identification was enabled by nozzle-skimmer dissociation (NSD) and detection of fragment ions with <5 ppm mass accuracy for highly-specific database searching using custom software. This overall approach led to identification of proteins up to 80 kDa, with 10-60 proteins identified in single LC-MS runs of samples from yeast and human cell lines pre-fractionated by their molecular weight using a gel-based sieving system.

Keywords

Nozzle-skimmer Dissociation; 'Low/High' data acquisition; GELFrEE; Polymeric reversed phase; nanocapillary LC-MS

INTRODUCTION

Despite the availability of high-performance mass spectrometers, methods for mass spectrometry-compatible separation and high-throughput identification of intact proteins are as yet underdeveloped. Therefore, top-down analysis is often implemented as a technique for precise characterization of one or a few protein targets. Here, we report a robust option for nanocapillary chromatography to achieve identification of intact proteins on a high-throughput basis.

While substantial improvements have been realized for mass spectrometers^{1, 2} and software³ for top-down proteomics, fractionation options for robust handling of complex cell extracts

*Corresponding author. kelleher@scs.uiuc.edu; Fax: 217 244-8068.

SUPPORTING INFORMATION AVAILABLE:

Additional information as noted in text. This material is available free of charge via the Internet at <http://pubs.acs.org>.

prior to LC-MS analysis have been less forthcoming. Both one-dimensional and two-dimensional protein separation strategies have been reported for intact protein analysis⁴⁻¹⁴. However, such methods for online nano-LC-MS for high-throughput top-down proteomics have been relatively few and none have shown the MS/MS robustness of bottom-up methods.

With the aim of making top-down analysis a viable option for large-scale, comparative proteomics, we integrated a combination of strategies for the online identification of both low- (10-20 kDa) and high-molecular weight (60-80 kDa) intact proteins. These strategies include efficient solution-phase protein pre-fractionation, online separation, robust protein detection/fragmentation, and high-throughput database searching. The gel-eluted liquid fraction entrapment electrophoresis (GELFrEE) protocol published by Tran and Doucette¹⁵ was further improved for increased protein load and recovery¹⁶, and was used as our default protein prefractionation system. Solution phase isoelectric focusing (sIEF)^{17, 18} was optionally used to reduce sample complexity prior to GELFrEE.

Though identification of very high molecular weight (>80 kDa) proteins is possible with top-down proteomics^{19, 20}, routine, high-throughput identification in the moderately-high mass (40-80 kDa) regime is particularly important to increase proteome coverage. The approach shown here achieves on-line identification of high mass proteins using a PLRP stationary phase separation and data acquisition via a “Low/High” strategy involving detection of intact proteins with a unit-resolution ion trap scan (*i.e.*, “low” resolution) and fragmentation products at Fourier-Transform resolution (*i.e.*, “high”) after nozzle skimmer dissociation^{21, 22}. With the refined protocols and the “Low/High” approach, we are able to readily identify yeast and human proteins in the 70-80 kDa regime, and 10-60 proteins from each nano-LC run. Utilization of the refined nano-LC-MS approach closes the performance gap between top-down and bottom-up and will allow for proteome-scale profiling of intact proteins pre-fractionated by one or two-dimensional separations^{15, 18}.

EXPERIMENTAL SECTION

Protein isolation from Human cell lines

HeLa-S3 cells, obtained from American Type Culture Collection (ATCC), were grown as suspension cultures in minimal essential medium (MEM) supplemented with 10% calf serum. $1-3 \times 10^8$ cells were collected by centrifugation, resuspended in nuclei isolation buffer (15 mM Tris-HCl, pH 7.5, 60 mM KCl, 15 mM NaCl, 5 mM MgCl₂, 1 mM CaCl₂, 250 mM sucrose, 1 mM dithiothreitol, 10 mM sodium butyrate, 0.3% NP-40) at a 10:1 (v/v) ratio), and incubated on ice for 5 min. After centrifugation at $600 \times g$ for 5 min., the cytosolic proteins (supernatant) were collected^{15, 16}. Human lung cancer cells H1299 (ATCC) were grown in RPMI medium supplemented with 10% fetal bovine serum and cell lysate was obtained in RIPA buffer (25 mM Tris HCl pH 7.5, 150 mM NaCl, 0.5% NP-40, 0.05% SDS). Samples were reduced in SDS loading buffer containing beta-mercaptoethanol, and alkylated with iodoacetamide at a molar ratio of about 80:1 (iodoacetamide: protein) in the dark for 60 min.

Protein isolation from yeast

Wild type *S. cerevisiae* sample was prepared as described previously²³. Briefly, the cells were grown to log-phase ($OD_{600} = 0.7$) in yeast extract peptone dextrose (YPD) liquid medium, harvested by centrifugation ($4,000 \times g$, 5 min.), followed by two water rinses and centrifugation. Cell membranes were disrupted by boiling in SDS solution (50 mM Tris-HCl, pH 7.5, 5% SDS, 5% glycerol, 50 mM dithiothreitol) with the complete protease inhibitors cocktail (Invitrogen, Carlsbad, CA). Additionally, cells were lysed with two passes through a French pressure cell (American Instrument Company, Silver Spring, MD) at 8,000 psi. The lysate was clarified at $13,000 \times g$, and the supernatant was stored at -20°C .

sIEF and GELFrEE separations

S. cerevisiae proteins (3 mg) were precipitated with cold acetone, resuspended in sIEF buffer (4 M urea, 2 M thiourea, 50 mM DTT, 1% w/v Biolyte 3/10 carrier ampholytes (Bio-Rad Laboratories, Hercules, CA)), and focused using an in-house eight-channel sIEF system¹⁷. After separation at 2W, the liquid fractions were transferred to separate vials. The chambers were further washed with 100 μ L of 1% SDS solution and these washes were combined with the respective sample fractions. Proteins in sIEF fractions were precipitated using cold acetone, and subsequently separated with multiplexed GELFrEE as previously described^{15, 18}. A commercial version of the GELFrEE separation platform is also available from Protein Discovery, Inc., (Knoxville, TN). Briefly, the GELFrEE buffer system used was either tris-glycine (0.192 M glycine, 0.025 M tris, 0.1% SDS) or tris-tricine (0.1 M tricine, 0.1 M tris, 0.1% SDS)¹⁶. Tube gels were cast to 15% T (1 cm length) for the resolving and 4% T for the stacking gels (300 μ L volume). About 200 μ g of proteins in approximately 100 μ L of sample buffer was loaded onto a GELFrEE column. GELFrEE fractions (150 μ L) for yeast (14) and HeLa (28) samples were collected for 1.5 h starting after the elution of the dye front.

Analytical SDS-PAGE slab gels

SDS-PAGE slab gel visualization of the GELFrEE fractions was employed to assess the resolution of separated proteins. One-fifteenth of the GELFrEE sample was loaded onto a 15% T (tris-glycine or tris-tricine) resolving slab gel. Gels were silver-stained following a previously published protocol²⁴.

Liquid Chromatography - Tandem Mass Spectrometry

GELFrEE liquid fractions containing proteins from 5 to 100 kDa were subjected to clean-up based on a method described previously²⁵. Briefly, methanol, chloroform, and water were added sequentially at 4:1:3 volumes of the sample volume with a brief vortexing between each solvent addition. Proteins became precipitated at the interphase between the upper methanol-water and lower chloroform layers subsequent to centrifugation at 13000 rpm for 5 minutes. The methanol-water layer was carefully removed without disturbing the interphase, and three volumes of methanol were added to the remaining solution. The solution was mixed gently by inverting the centrifuge tube and centrifuged at 13000 rpm for 10 minutes to pellet the proteins. The pellet was washed with three volumes of methanol after decanting the supernatant and dried in room temperature. Protein pellets were resuspended in 40 μ L of buffer A (95% H₂O: 5% acetonitrile, both containing 0.2% formic acid) and 10 μ L of the resuspended protein sample was injected using an autosampler (Eksigent, Dublin, CA). Nanobore analytical columns (75 μ m $\times$ 10 cm) with an integral fritted nanospray emitter (PicoFrit, New Objective, Inc., Woburn, MA) containing 5 μ m polymeric reversed-phase (PLRP) media (300, 1000, or 4000 Å pore size) or 5 μ m C4 derivatized porous silica (300 Å pore size) were prepared. Trap columns (150 μ m i.d. $\times$ 2 cm) contained identical chromatographic media. The Eksigent 1D Plus nano-HPLC system was operated at a flow rate of 300 nL/min. A 75 min. gradient with buffer A (as above) and B (5% H₂O: 95% acetonitrile, both containing 0.2% formic acid) was used for separation of complex protein samples consisted of the following concentration of B: 5% for 3 min; 30% at 10 min; 55% at 50 min. Within the next 5 min., B was ramped to 98% and remained at 98% for 3 min. before declining to 5% in another 5 min. The column was equilibrated in 5% B further up to 75 min. For the separation of standard proteins from either C4 or PLRP columns, a 60 min. gradient (B = 3% up to 3 min., 30% at 10 min., 55% at 35 min., 98% from 40-43 min. and 2% from 48 to 60 min.) was used. Standard protein mixtures were made as 2 mg/mL stocks in mass-spectrometry grade water and diluted in HPLC solvent A just before loading onto analytical column without the use of a trap column.

Samples were analyzed on a 12 Tesla LTQ FT Ultra (Thermo Fisher Scientific, San Jose, CA) fitted with a digitally controlled nanospray ionization source (PicoView DPV-550, New

Objective, Inc.). Protein precursor ion intact masses and fragment masses were acquired in the LTQ (MS1) and FTICR (pseudo MS²) respectively, with different nozzle-skimmer dissociation (NSD) voltage settings at the 'Xcalibur' software (NSD is defined as 'SID' in Xcalibur). Based on preliminary analyses, NSD of 15 V was optimal for ion trap scans for the dissociation of weakly bound non-covalent adducts, while the NSD voltage for fragmentation was standardized as described in the 'Results and Discussion'.

Database search and protein identification

. Data from LC-MS/MS files were analyzed using ProSightPC 2.0³ (Thermo Fisher Scientific, San Jose, CA). For data acquired with the "Low/High" strategy, intact precursor and fragment masses from .raw files were determined using in-house software (called 'cRAWler') to generate files for ProSightPC 2.0. This software uses an embedded version of the deconvolution algorithm²⁶ for determining average, neutral intact masses and the 'THRASH'²⁷ algorithm for extracting monoisotopic, neutral fragment masses. These data in .puf (ProSight upload format) files were searched against a shotgun-annotated human (754,012 protein forms) or yeast (52,616 protein forms) proteome databases containing known post-translational modifications and alternative splice forms. Fragment masses from raw data for protein standards were obtained using 'Xtract' algorithm in QualBrowser (Thermo Fisher Scientific) and searched against a standard ProSight Warehouse. This database was built using ten protein standards and consisted of 7,361 protein forms.

In order to reduce the noise arising from low-abundance non-specific peaks, an in-house algorithm was used to trim the fragment mass list prior to database searching. For each .puf file, fragments were sorted into 50 or 100 Da mass bins and only the three or five most intense fragment ions within each bin were retained. This approach anticipates the regular spacing of "true" fragment ions and their variable intensities; retention is based on local intensity rather than overall intensity. The intensity-based reduction in the number of fragments per database search event improves the significance of search results by removing noise introduced by the THRASH algorithm.

To analyze data from yeast and human samples, our cRAWler software was modified to generate two files; one that contained all NSD spectral data with their deconvoluted intact masses and the other containing all other NSD spectra where our deconvolution algorithm was unable to assign an intact mass. In both type of searches, a fragment mass tolerance of 10 ppm was used. False discovery rates (FDR) were calculated based on searches against database of concatenated forward and reverse sequences.

RESULTS AND DISCUSSION

Comparison of stationary phases: sensitivity and resolution

Silica-based solid supports for reversed phase LC (e.g., C4, C8, and C18) have been used for separation of peptides and proteins, with the less hydrophobic media (shorter alkyl chains) typically being employed for intact protein separations^{11, 28-30}. However, given the touted benefits of polymeric media such, as uniform hydrophobicity and increased mechanical strength³¹⁻³⁵, we began a study of its performance for the chromatographic separation of intact proteins. Chromatographic peak widths and sensitivity during LC-MS were studied using detection in both the ion trap and the FTICR cell.

Ion trap base-peak chromatograms obtained from C4- and PLRP-nano-LC of three different equimolar amounts (0.3, 1, 3 pmol) of a seven-protein mix are shown in Figure 1A. Precursor ions for four of these proteins were detected from one LC-MS injection of 0.3 pmol total protein on C4 column. Examples of ion trap and FTICR spectra for carbonic anhydrase are shown in Figure 1B. When the protein amount was increased to 1 pmol, distinct chromatographic elution

of 6 of the proteins was observed along with a poorly resolved peak for ovalbumin (Figure 1A). As sample amount increased from 3 to 30 pmol, an increase in signal to noise (S/N) was observed for all proteins (except ovalbumin) on the C4 column.

LC-MS with PLRP media with the same 300 Å pore size gave an increased S/N. Ion trap base-peak chromatograms obtained from PLRP chromatography with three different concentrations of protein mix are shown in red on Figure 1A. The PLRP stationary phase enabled all seven proteins to be detected even at the lowest loading amount (0.3 pmol). Furthermore, with the same protein amount, higher S/N was observed with PLRP media than with the C4 media for all proteins tested. An example of the S/N difference (approximately 3 fold) is shown with carbonic anhydrase in Figure 1B. As previously observed with C4 media, further increases in sample load onto a PLRP column afforded higher S/N with the exception of ovalbumin whose peak appeared as a “hump” even at 30 pmol injected (data not shown). Overall, PLRP media exhibited higher protein recoveries which gave rise to an increased S/N of protein spectra by factors of 2 to 3.

In addition to the above benefits, PLRP media showed reduced chromatographic peak widths for some of the proteins. For example, 3-minute peak widths were obtained with 1.0 pmol of α -lactalbumin and myoglobin during C4 chromatography, while PLRP showed <1 minute peak widths (numbered 3 and 4, respectively in Figure 1A). Further, analysis of peak capacity using the formula $P = 1 + (t_g/w)$ (where P = Peak capacity, w = peak width at 13% height, and t_g = gradient time to 50% B)³⁶ demonstrated the differences between the C4 and PLRP media. For this calculation, four peaks representing a single protein each (numbered 3, 4, 6, and 7 in Figure 1) were used. At 1 pmol loading, a peak capacity of 21 and 30 was obtained for the C4 and PLRP media, respectively. When the loaded protein amount was increased to 3 pmol, the peak capacity of both media was affected only marginally (21 vs. 28). Both the observed chromatographic peak splitting with C4 (Figure 1), and differences observed in peak width could be attributed to the differences in the physicochemical properties of the stationary phases (*i.e.*, absence of silanol groups in PLRP media)^{31, 32, 35}.

Effect of PLRP pore size on the retention time of proteins

Having determined the advantages of PLRP over C4 for online nano-LC separation of proteins, we questioned whether the differences in pore size of PLRP could effect retention time and peak width. Accordingly a protein mix consisting of six standard proteins including bovine serum albumin (BSA, 66 kDa) was used to compare three different pore sizes of PLRP stationary phase. As shown in Figure 2, the elution order of these proteins is similar to that reported for 4.6 mm i.d. PLRP analytical HPLC columns³¹, with the exception that BSA and myoglobin co-elute under conditions used here. This minor difference in chromatographic resolution could be attributed to the differences in column dimensions (nanobore vs. analytical), particle size (3 μ m vs. 10-15 μ m), and ion pairing agent used (formic acid vs. trifluoroacetic acid).

Co-elution of proteins (as seen with myoglobin and BSA) will ultimately affect the number of proteins identified in addition to requiring multiplexed top-down analysis. Thus, we interrogated the effect of PLRP pore size on chromatographic resolution. With a 1 pmol injection onto 300 Å PLRP media, distinct chromatographic peaks were obtained for BSA and myoglobin with only a small overlap (Figure 2). Though increasing the pore size provides narrower peak widths, coelution of BSA and myoglobin increased (Figure 2). As a result, fewer spectra unique to myoglobin and BSA were obtained with 1000 Å pore size, and no spectrum unique to BSA was obtained with 4000 Å PLRP.

Base peak widths observed with 300 Å pore size for cytochrome c, α -lactalbumin, myoglobin, carbonic anhydrase, and ovalbumin were 0.5, 0.5, 0.8, 1, and 4.5 min., respectively (Figure 2).

While chromatography with 1000 Å pore size media showed little change in the peak width for smaller proteins (cytochrome c and α -lactalbumin), noticeable reduction in peak width was observed for all other (larger) proteins (Figure 2). In addition, an approximately 5 minute delay in retention times was also observed (Figure 2). For the larger proteins, a further reduction in peak widths was observed when increasing the PLRP pore size to 4000 Å. This led us to conclude that increased PLRP pore size reduces the peak widths of proteins, and that this effect is more prominent with larger proteins. This effect was more pronounced in peak capacities calculated using four of the six proteins (peaks for cytochrome c, α -lactalbumin, carbonic anhydrase, and ovalbumin) separated using 300, 1000, and 4000 Å columns: values were 33, 36, and 40, respectively. This is in excellent agreement with the observation in 4.6 mm i.d. PLRP columns that increased peak width in smaller pore size (300 Å) results from restricted diffusion of higher molecular weight proteins³⁷. Based on the above results, we concluded 1000 Å PLRP as optimal to obtain better resolution without compromising on peak widths for proteins across a wide molecular weight range.

Robust detection of high mass proteins using ion trap

After determining the suitability of PLRP as a stationary phase for the online nano-LC of low to medium size proteins, we tried altering other mass spectrometry parameters to increase the detection ability of higher molecular weight proteins such as ovalbumin, BSA, and apotransferrin (78 kDa). Because of the advantages of the faster scan speed and the observed increase in S/N, we decided to use ion trap scans for intact mass detection rather than FTICR. Two main parameters previously reported for increasing the S/N of large protein spectra are 1) higher number of μ scans³⁸ and 2) increase of capillary temperature¹⁹. MS data were obtained for BSA, ovalbumin and apotransferrin with varying number of μ scans (Supplementary Figure S1). We observed that more μ scans translates into higher S/N for these proteins. Intact mass determination for ovalbumin required approximately 30 μ scans (Supplementary Figure S1). With the combination of online protein PLRP separation and increased μ scans in the ion trap, it was possible to routinely detect intact masses of proteins up to 78 kDa on a chromatographic time scale.

Robust fragmentation using nozzle-skimmer dissociation

Collisionally induced dissociation (CID) has been proven to be powerful method for fragmentation³⁹; however CID in an ion trap or FTICR cell usually involves an additional stage of mass selection prior to dissociation. Since it is difficult to isolate and fragment selected charge states of large proteins in an ion trap, we evaluated non-selective nozzle-skimmer dissociation (NSD) method for both low and high mass proteins. Dissociation of multiply-charged proteins using NSD was first reported by Loo *et al.* in 1988^{21, 22}, and was later reported in several other studies for the fragmentation of large biomolecules^{19, 27, 40}.

We applied NSD to the same standard mixture as above and evaluated the optimum dissociation settings. Examples of NSD-FTICR spectra obtained with increasing NSD voltage are given in Figure 3. The charge state distribution for BSA had a maximum charge state of 65+ (Figure 3A). Nozzle-skimmer voltage (Δ NS) of 15 V was found to reduce the chemical noise, but was insufficient to dissociate multiply charged protein ions. Therefore, Δ NS of 15 V was applied for all precursor ion scans (MS1). Increasing the Δ NS to 30 V caused the disappearance of charge states higher than ~65+ (Figure 3B). At Δ NS of 45 V and higher, more of the highly charged precursors dissociated yielding fragments with charges of +4 and +5 (Figure 3B).

Protein identification in top-down proteomics is performed by matching observed monoisotopic fragment masses with theoretical fragment masses³. Therefore, we next analyzed the effect of Δ NS on the identification metrics for both low and high mass proteins. Using Δ NS values up to 75 V, there is a direct increase in the number of fragment ions observed for

all three proteins with molecular weights from 12 to 66 kDa (Figure 3C), with the exception of carbonic anhydrase which produced fewer fragments at Δ NS 75 V than at Δ NS 60 V. It is also noteworthy that at Δ NS 30 V, carbonic anhydrase yielded a number of recognizable fragments with a ProSightPC E-value of 10^{-2} while no such identifications were obtained for cytochrome *c* and BSA (Figure 3C). Additionally, these three proteins differed slightly in the types of fragments produced by NSD. As noted by others³⁷, BSA produced only b ions independent of Δ NS; however cytochrome *c* and carbonic anhydrase gave both b and y ions in varying proportions. The overall performance of NSD is consistent with earlier reports for top-down analysis^{21, 37} and serves as the basis for an optimized platform for application to endogenous proteins of high molecular weight (*vide infra*).

Interestingly, sequence coverage (Figure 3C) for these proteins remained almost constant throughout the Δ NS range of 45 to 75 V (Figure 3D). An exception was observed for carbonic anhydrase, which showed a mild decrease the sequence coverage upon changing Δ NS from 45 to 75 V. The consistent sequence coverage obtained with increasing Δ NS is attributed to a shift in the average fragment ion charge states (e.g., from +4 to +3). At a Δ NS of 100 V, large numbers of smaller fragments were produced (mostly +1 ions) for all proteins. In the voltage ranges tested, NSD produced fragments with an average sequence coverage of 40-50% for cytochrome *c* (12.3 kDa) and carbonic anhydrase (29 kDa), but only 5-6% sequence coverage for BSA (66 kDa). With this performance on standards, this 'Low-High' (*i.e.*, ion trap MS1 scan followed by FT-ICR fragments scan) approach for top-down analysis was tested for use in high-throughput top-down proteomics.

A robust work flow for top-down proteomics

Standardization of the methods and conditions for efficient pre-fractionation, online HPLC, and robust fragmentation as described above helped us to adopt a work flow for the analysis of a complex proteome sample as presented in Figure 4. Proteins were isolated from either yeast or human cells in a buffer suitable for one dimensional fractionation by GELFrEE¹⁵ or two-dimensional separation by sIEF and GELFrEE¹⁸. After detergent removal of GELFrEE fractions²⁵, nano-LC-LTQ FT runs using the 'Low-High' mode described above were performed on samples from a variety of molecular weight ranges. Monoisotopic precursor and fragment neutral masses were extracted using in-house software, and ProSightPC 2.0 searching used a maximum E-value threshold of 10^{-1} with calculated false discovery rates (FDRs) of less than 1% for protein identification.

PLRP capillary nano-LC-LTQ FTICR analysis of complex mixtures

The integrated workflow outlined above was tested on HeLa S3 cytosol. Proteins were isolated from HeLa cells and 200 μ g were fractionated by GELFrEE¹⁵. After chloroform/methanol/water clean-up, 1/5th of the precipitated protein in the fractions was used to assess the quality of the sample preparation by analytical SDS-PAGE (Figure 5A). As described previously^{15, 18}, the size-based partitioning of the whole cytosolic proteome gave 23 liquid fractions with proteins from 10 kDa (fraction min. 5) to over 100 kDa (fraction min. 90) (Figure 5). Online PLRP-nano-LC-MS/MS was performed on selected fractions. A base peak chromatogram using ion trap scans obtained for the GELFrEE fraction collected at 8 minutes (*i.e.*, proteins eluted from 6-8 minutes) show more than 20 distinct peaks for intact proteins and ESI-MS spectra for three peaks are shown in Figure 5B. A ProSightPC database search of the high resolution fragment masses against a human top-down database in 'absolute mass' mode gave 15 unique protein identifications using a maximum E-value of 1×10^{-3} ; the best E-value was 1×10^{-20} (Figure 5A). With post-acquisition fragment filtering (outlined in the methods section), this result improved to 20 unique identifications with the best E-value being 3×10^{-30} (Table 1). Various noise filter windows used for the analysis of the 8-minute fraction are given in Supplementary Figure S2. The three most abundant fragments in any given 100

m/z window were used for identifying the above 20 proteins with the E-value cut-off set at 1×10^{-3} . This list consisted of proteins such as ribosomal protein S14 (16.3 kDa), eukaryotic translation initiation factor 5A-1-like (16.9 kDa), nucleoside diphosphate kinase B (17.3 kDa), and peroxiredoxins 1 (22.1 kDa) and 2 (15.1 kDa), with 95% of proteins in the expected molecular weight range of 7-22 kDa (Table 1).

Robust nano-LC- mass spectrometry analysis of other GELFrEE fractions with molecular weights up to 50-60 kDa gave an average of 15 unique protein identifications in each nano-LC run (data not shown). Although single-scan NSD spectra obtained for samples with >60 kDa proteins contained many fragment ions, ProSightPC search results showed poor E-value assignments due to low S/N of matching fragments and a large number of unassigned fragments. Therefore, precursor mass detection and protein identification from such spectra required some spectral averaging. With this implementation, identification of a 71 kDa protein from a GELFrEE fraction collected from 70-80 min. was possible (see Figures 5A and 5C). Summed MS spectra obtained from MS1 and NSD scans for one of the proteins from the above fraction is given on the left in Figure 5C. A ProSightPC search against the entire database using a precursor mass of 70,000 Da and mass tolerance of 100,000 Da gave several heat shock proteins, with the heat shock protein HSP7C (70,765 Da) as the top hit with an E-value of 3×10^{-4} . There were a total of 7 fragments that matched in this database search; each had a mass error of less than 1 ppm. Thus, it is apparent that the robust top-down proteomics analysis is capable of identifying proteins in the range of 70 kDa from a human cell line.

Reduction in the complexity of the proteome is achievable by two-dimensional orthogonal separation of proteins in the liquid phase¹⁸. We believed that this reduction in complexity would help to increase the size limit of protein detection. For this experiment we extracted proteins from yeast cells and performed sIEF as described¹⁷. One of the eight IEF fractions (fraction 3) was subjected to GELFrEE separation, followed by the above chromatography and mass spectrometry protocols. Results of the protein identification process are shown in Supplementary Figure S3. The ion trap base peak chromatogram shows approximately 10 precursor ion peaks; NSD spectra with automated ProSightPC searching identified proteins (26.6 and 46.7 kDa) for two of the peaks shown in the chromatogram (Figure S3). Analysis of higher molecular weight proteins using our robust protein identification method identified proteins greater than 80 kDa. As an example, an 81 kDa molecular chaperone protein (HSC82) was identified through a ProSightPC search against the yeast database with a sequence tag region shown in Supplementary Figure 3D. The increased ability to identify proteins at high molecular weight is attributed to a variety of factors including the nano-LC media, the use of NSD, and the reduction of sample complexity after three dimensions of protein fractionation.

While multiple dimensions of pre-fractionation can help to resolve highly complex proteomes into approximately 128 fractions (8×16), the increase in proteome coverage also depends on the peak capacity of the online nano-LC. We therefore looked at how many proteins could be identified from a single nano-LC run coupled to the optimized robust top-down analysis. Accordingly, a lung cancer cell line (H1299) was lysed using RIPA buffer and subjected to GELFrEE separation with PLRP nano-LC-MS (Figure 6). A ProSightPC search of the data obtained using an extended (120 min.) gradient identified a total of 55 and 57 high confidence, non-redundant proteins from fractions 16 and 18 (15-40 kDa proteins), respectively. This is in agreement with a chromatographic peak capacity of 30 calculated for a 60-minute gradient using 1.0 pmol protein mix. However, analysis of the 7-minute fraction (8-20 kDa regime) data obtained using a 75-minute gradient resulted in 60 unique protein identifications (Figure 6, Supplementary Table 1), indicating a higher total capacity for identification of smaller proteins. Thus, it is conceivable that the current optimized PLRP nano-LC-MS coupled with pre-fractionation techniques such as GELFrEE could identify hundreds of proteins on a scale quite comparable to bottom-up proteomics.

In conclusion, it is apparent that the top-down protocols optimized here are capable of identifying proteins up to 80 kDa from human and yeast proteome samples. The fractionation approaches are reproducible and use hardware, columns, reagents and software that are available commercially. The “Low/High” strategy for data acquisition uses a fast, sensitive ion trap for MS1 and only requires high-resolution for MS2 data. This is sufficient for top-down experimentation across a wide range of masses and should extend the number of laboratories able to perform top-down proteomics in a routine fashion.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank members of Kelleher research group, particularly Haylee Thomas, Cong Wu, Mingxi Li, and Ioanna Ntai, for technical support and insightful discussions. We also acknowledge Carla Marshall-Waggett and Stanley Durand for facilitation and technical expertise. Funding was provided by the National Institutes of Health (GM 067193-07) to NK, the National Science Foundation (DMS 0800631) to LZ, the American Chemical Society (Division of Analytical Chemistry Fellowship) for PT, and the Institute of Genomic Biology (IGB Fellows Program) for AV.

REFERENCES

1. Breuker K, Jin M, Han X, Jiang H, McLafferty FW. *J Am Soc Mass Spectrom* 2008;19:1045–1053. [PubMed: 18571936]
2. Schaub TM, Hendrickson CL, Horning S, Quinn JP, Senko MW, Marshall AG. *Anal Chem* 2008;80:3985–3990. [PubMed: 18465882]
3. Zamdborg L, LeDuc RD, Glowacz KJ, Kim YB, Viswanathan V, Spaulding IT, Early BP, Bluhm EJ, Babai S, Kelleher NL. *Nucleic Acids Res* 2007;35:W701–706. [PubMed: 17586823]
4. Loo JA, Udseth HR, Smith RD. *Anal Biochem* 1989;179:404–412. [PubMed: 2774189]
5. Jensen PK, Pasa-Tolic L, Peden KK, Martinovic S, Lipton MS, Anderson GA, Tolic N, Wong KK, Smith RD. *Electrophoresis* 2000;21:1372–1380. [PubMed: 10826683]
6. Lubman DM, Kachman MT, Wang H, Gong S, Yan F, Hamler RL, O'Neil KA, Zhu K, Buchanan NS, Barder TJ. *J Chromatogr B Analyt Technol Biomed Life Sci* 2002;782:183–196.
7. Simpson DC, Ahn S, Pasa-Tolic L, Bogdanov B, Mottaz HM, Vilkov AN, Anderson GA, Lipton MS, Smith RD. *Electrophoresis* 2006;27:2722–2733. [PubMed: 16732621]
8. Sharma S, Simpson DC, Tolic N, Jaitly N, Mayampurath AM, Smith RD, Pasa-Tolic L. *J Proteome Res* 2007;6:602–610. [PubMed: 17269717]
9. Meng F, Cargile BJ, Patrie SM, Johnson JR, McLoughlin SM, Kelleher NL. *Anal Chem* 2002;74:2923–2929. [PubMed: 12141648]
10. Du Y, Meng F, Patrie SM, Miller LM, Kelleher NL. *J Proteome Res* 2004;3:801–806. [PubMed: 15359734]
11. Parks BA, Jiang L, Thomas PM, Wenger CD, Roth MJ, Boyne MT 2nd, Burke PV, Kwast KE, Kelleher NL. *Anal Chem* 2007;79:7984–7991. [PubMed: 17915963]
12. Pesavento JJ, Bullock CR, LeDuc RD, Mizzen CA, Kelleher NL. *J Biol Chem* 2008;283:14927–14937. [PubMed: 18381279]
13. Wenger CD, Boyne MT 2nd, Ferguson JT, Robinson DE, Kelleher NL. *Anal Chem* 2008;80:8055–8063. [PubMed: 18841935]
14. Zhou F, Hanson TE, Johnston MV. *Anal Chem* 2007;79:7145–7153. [PubMed: 17711353]
15. Tran JC, Doucette AA. *Anal Chem* 2008;80:1568–1573. [PubMed: 18229945]
16. Lee JE, Kellie JF, Tran JC, Tipton JD, Catherman AD, Thomas HM, Ahlf DR, Durbin KR, Vellaichamy A, Ntai I, Marshall AG, Kelleher NL. *J Am Soc Mass Spectrom* 2009;20:2183–2191. [PubMed: 19747844]
17. Tran JC, Doucette AA. *J Proteome Res* 2008;7:1761–1766. [PubMed: 18284188]
18. Tran JC, Doucette AA. *Anal Chem* 2009;81:6201–6209.

19. Han X, Jin M, Breuker K, McLafferty FW. *Science* 2006;314:109–112. [PubMed: 17023655]
20. Karabacak NM, Li L, Tiwari A, Hayward LJ, Hong P, Easterling ML, Agar JN. *Mol Cell Proteomics* 2009;8:846–856. [PubMed: 19074999]
21. Loo JA, Udseth HR, Smith RD. *Rapid Communications in Mass Spectrometry* 1988;2:207–210.
22. Loo JA, Edmonds CG, Smith RD. *Science* 1990;248:201–204. [PubMed: 2326633]
23. de Godoy LM, Olsen JV, de Souza GA, Li G, Mortensen P, Mann M. *Genome Biol* 2006;7:R50. [PubMed: 16784548]
24. Shevchenko A, Wilm M, Vorm O, Mann M. *Anal Chem* 1996;68:850–858. [PubMed: 8779443]
25. Wessel D, Flugge UI. *Anal Biochem* 1984;138:141–143. [PubMed: 6731838]
26. Zheng H, Ojha PC, McClean S, Black ND, Hughes JG, Shaw C. *Rapid Commun Mass Spectrom* 2003;17:429–436. [PubMed: 12590391]
27. Horn DM, Zubarev RA, McLafferty FW. *J Am Soc Mass Spectrom* 2000;11:320–332. [PubMed: 10757168]
28. Badock V, Steinhilber U, Bommert K, Otto A. *Electrophoresis* 2001;22:2856–2864. [PubMed: 11565780]
29. Van den Bergh G, Arckens L. *Methods Mol Biol* 2008;424:147–156. [PubMed: 18369860]
30. Millea KM, Krull IS, Cohen SA, Gebler JC, Berger SJ. *J Proteome Res* 2006;5:135–146. [PubMed: 16396504]
31. Twesten, K. A. T. a. T. N. *Journal of Chromatography A* 1986;359:111–119.
32. Lloyd LL. *Journal of Chromatography* 1991;544:201–217.
33. Zhelev NZ, Barratt MJ, Mahadevan LC. *J Chromatogr A* 1997;763:65–70. [PubMed: 9129316]
34. Elgar DF, Norris CS, Ayers JS, Pritchard M, Otter DE, Palmano KP. *J Chromatogr A* 2000;878:183–196. [PubMed: 10866065]
35. Lloyd LL, Millichip MI, Watkins JM. *J Chromatogr A* 2002;944:169–177. [PubMed: 11831752]
36. Gilar M, Daly AE, Kele M, Neue UD, Gebler JC. *J Chromatogr A* 2004;1061:183–192. [PubMed: 15641361]
37. Loo JA, Edmonds CG, Smith RD. *Anal Chem* 1991;63:2488–2499. [PubMed: 1763807]
38. Courchesne PL, Jones MD, Robinson JH, Spahr CS, McCracken S, Bentley DL, Luethy R, Patterson SD. *Electrophoresis* 1998;19:956–967. [PubMed: 9638942]
39. Senko MW, Speir JP, McLafferty FW. *Anal Chem* 1994;66:2801–2808. [PubMed: 7978294]
40. Eyles SJ, Speir JP, Kruppa GH, Gierasch LM, Kaltashov IA. *J Am Chem Soc* 2000;122:495–500.

Figure 1. Protein sensitivity, resolution in C4 and PLRP nano-capillary columns

A) Chromatograms for RPLC separation of a mixture of seven protein standards, using C4 (blue) and PLRP (red) (75 μ m I. D. $\times$ 100 mm, 300 $\text{\AA}$, 5 μ m stationary phase) analytical columns. Standards used were: 1. Ubiquitin; 2. Cytochrome c; 3. α -lactalbumin; 4. Myoglobin; 5. α -casein; 6. Carbonic anhydrase; 7. Ovalbumin. B) Mass spectra for carbonic anhydrase obtained from online LC-MS using C4 (blue) and PLRP (red) media described in A.

Figure 2. Protein retention is a function of PLRP pore size

HPLC chromatograms obtained after separation of six protein standards ranging from 12 - 66 kDa on PLRP solid support with varying pore size. Protein standards used were: 1. Cytochrome c; 2. α-lactalbumin; 3. Myoglobin; 4. Carbonic anhydrase; 5. Ovalbumin; 6. Bovine serum albumin.

Figure 3. Nozzle-skimmer dissociation (NSD), and identification of intact proteins

A) Precursor ion spectra for BSA acquired using low-resolution ion trap with 20 μ scans and NSD voltage (Δ NS) of 15 V. B) NSD spectra for BSA obtained with increase in Δ NS. C) Total number of ProSightPC assigned fragment ions for cytochrome c, carbonic anhydrase and BSA with varying NSD voltages. D) Percentage protein coverage for cytochrome c, carbonic anhydrase and BSA with varying Δ NS.

Figure 4. Overview of robust top-down analysis work flow

Outline of the work flow for robust intact protein identification from either yeast or mammalian samples after standardizing some of the parameters and conditions with commercially available protein standards. Proteins were isolated using a lysis buffer containing SDS followed by ultrasonication (see methods). 2 or 0.2 mg of protein was used, respectively, for separation by sIEF or GELFrEE. Fractions from GELFrEE underwent detergent removal by chloroform/methanol precipitation, and the resultant proteins were subjected to PLRP nano-LC-MS. A low-resolution full-scan event at the ion trap was followed by high-resolution FTICR scan after nozzle-skimmer dissociation (NSD). Absolute masses were extracted from NSD scan using in-house software that uses THRASH algorithm, and the resultant files were searched against Uniprot human protein database by ProSightPC 2.0 search engine. Proteins were identified with E-value threshold and its associated FDR in each database search.

Figure 5. Pre-fractionation and PLRPS nano-LC-MS/MS of HeLa sample

Silver-stained gel image of fractions collected from GELFrEE separation of HeLa cell lysates (A). A selected fraction (highlighted in dotted arrow) was injected to nano-LC-MS/MS analysis. A Δ NS of 75V was applied for protein fragmentation and high-resolution spectra were collected in the FTICR (B). Proteins were identified through a ProSightPC 2.0 database search in absolute search mode against a protein database built using the Database Manager portion of ProSight PC 2.0. Precursor ion spectrum, NSD spectrum, and ProSightPC search result showing the identification of a 70.8 kDa protein, HSP7C from 80 min. fraction (C).

^a = Non-redundant Uniprot protein accessions identified with <1% FDR

Figure 6. Pre-fractionation and PLRPS nano-LC-MS/MS of lung cancer cell proteins

Polyacrylamide gel image showing GELFrEE fractions from a human lung cancer cell line (H1299) collected from 0 to 20 minutes (A). Ion trap base peak chromatogram obtained from PLRP nano-LC-MS of the 20-minute fraction is shown in B. Nano-LC separation, mass spectrometry data acquisition, and database searching were performed using the protein identification strategy described in the manuscript and in Figure 4. The total number of proteins identified using different E-value threshold levels and the corresponding false discovery rates (FDR) is shown in C. With <1% FDR, a total of 60 and 57 non-redundant Uniprot accession numbers were assigned to spectra obtained from the 7-minute and 20-minute fractions highlighted in panel A.

Table 1
Proteins identified through robust top-down analysis of a single (8-minute) GELFREE fraction of human HeLa S3 sample.

No.	Matching Fragments	Theoretical Mass	N-terminal Modification	P Score	E-Value	Uniprot Accession
1	18	15093.60	N-acetyl-L-alanine	4.3E-36	3.2E-30	B7Z5A2
2	18	11662.80	N-acetyl-L-serine	3.4E-28	2.6E-22	B8ZZQ6
3	13	16901.40	N-acetyl-L-alanine	4.7E-26	3.6E-20	Q6IS14
4	16	11950.80	N-acetyl-L-serine	6.5E-24	4.9E-18	Q15204
5	13	10521.60		4.3E-23	3.2E-17	P05114
6	13	7046.58		2E-20	1.5E-14	P02795
7	12	19502.00		1.4E-19	1.1E-13	B4DN70
8	11	16060.50		4.3E-18	3.3E-12	B7ZB63
9	11	22235.30	N-acetyl-L-serine	8.1E-18	6.1E-12	B5BU26
10	13	9256.02		1.3E-17	9.9E-12	P05204
11	10	17172.20	N-acetyl-L-alanine	2.5E-17	1.9E-11	B7Z6S5
12	10	17871.30		3.6E-17	2.7E-11	B4DJC8
13	10	6717.97		1.1E-15	8E-10	B8ZZI3
14	6	18991.60		4.3E-12	3.3E-06	Q5T9W8
15	7	38861.30	N-acetyl-L-aspartic acid	1.4E-11	1.1E-05	B4DW52
16	7	15153.80	N-acetyl-L-alanine	3.5E-11	2.6E-05	A6NIW5
17	7	12173.70	N-acetyl-L-serine	1.4E-10	0.0001	B9ZVP7
18	7	17097.00	N-acetyl-L-threonine	1.4E-10	0.00011	Q9H7Z5
19	6	21245.70		2E-10	0.00015	B4DNW1
20	7	20611.70		1.1E-09	0.0008	P18085