Published in final edited form as:

Proteomics. 2005 April ; 5(5): 1263–1273.

Improved Proteome Coverage by Using High Efficiency Cysteinyl Peptide Enrichment: The Human Mammary Epithelial Cell Proteome

Tao Liu, Wei-Jun Qian, Wan-Nan U. Chen, Jon M. Jacobs, Ronald J. Moore, David J. Anderson, Marina A. Gritsenko, Matthew E. Monroe, Brian D. Thrall, David G. Camp II, and Richard D. Smith

Biological Sciences Division and Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, WA, USA

Abstract

Automated multidimensional capillary liquid chromatography-tandem mass spectrometry (LC-MS/MS) has been increasingly applied in various large scale proteome profiling efforts. However, comprehensive global proteome analysis remains technically challenging due to issues associated with sample complexity and dynamic range of protein abundances, which is particularly apparent in mammalian biological systems. We report here the application of a high efficiency cysteinyl peptide enrichment (CPE) approach to the global proteome analysis of human mammary epithelial cells (HMECs) which significantly improved both sequence coverage of protein identifications and the overall proteome coverage. The cysteinyl peptides were specifically enriched by using a thiol-specific covalent resin, fractionated by strong cation exchange chromatography, and subsequently analyzed by reversed-phase capillary LC-MS/MS. An HMEC tryptic digest without CPE was also fractionated and analyzed under the same conditions for comparison. The combined analyses of HMEC tryptic digests with and without CPE resulted in a total of 14,416 confidently identified peptides covering 4,294 different proteins with an estimated 10% gene coverage of the human genome. By using the high efficiency CPE, an additional 1,096 relatively low abundance proteins were identified, resulting in 34.3% increase in proteome coverage; 1,390 proteins were observed with increased sequence coverage. Comparative protein distribution analyses revealed that the CPE method is not biased by protein molecular weight, pI, cellular location, or biological functions. These results demonstrate that the use of the CPE approach provides improved efficiency in comprehensive proteome-wide analyses of highly complex mammalian biological systems.

Keywords

Cysteinyl peptide enrichment; Human mammary epithelial cells; Mass spectrometry; Low abundance; Proteome coverage

Abbreviations

CPE, cysteinyl peptide enrichment; HMEC, human mammary epithelial cell; SCX, strong cation exchange; QCET, quantitative cysteinyl-peptide enrichment technology

*Correspondence: Dr. Richard D. Smith, Environmental Molecular Sciences Laboratory, MSIN: K8-98, Pacific Northwest National Laboratory, P.O. Box 999, Richland, WA 99352, USA. Email: rds@pnl.gov; Fax: (509) 376-7722..

1 Introduction

The increasing availability of partial or completed genome sequences for a diversity of organisms [1–4] (including human [5]), along with the technological advances in proteomics, microarray analysis, and metabolomics, provides biologists the first opportunity to pursue “systems biology,” the systematic and quantitative analysis of all of the components that constitute the system. Proteomic characterization is of significant interest because proteins, the primary bioeffectors, are involved in essentially all biological activities and the diversity of roles they play in biological events can be readily altered through transient modifications, enabling the delicate regulation of cellular processes on a systems-wide basis. Comprehensive global proteome analysis provides a particularly rich source of biological information and will contribute to detailed descriptions of the structure, function and control of biological systems in healthy and diseased states [6].

Automated multidimensional capillary liquid chromatography-tandem mass spectrometry (LC-MS/MS) is routinely applied in large scale proteome profiling [7–10]. However, the confined instrumental dynamic range limits the extension of its ability to explore low abundance proteins which are of particular interest in mammalian proteomic samples. To tackle this problem, either subcellular fractionation or enrichment strategies of “examine less to see more” or “divide and conquer” have been more widely adopted for proteomic studies besides orthogonal fractionation prior to MS analysis. Functional protein complexes or organelles (e.g., ribosome [11], spliceosome [12], phagosome [13], lysosome [14], exosome [15], vesicle coat [16], microsome [17], mitochondria [18], Golgi apparatus [19], nucleolus [20], etc.) have been selectively isolated and comprehensively analyzed. The collective data obtained in these studies provide a more informative cell map, leading to an increased understanding of complex cellular processes in healthy and diseased states [21]. Experimental approaches capable of concentrating and enriching subsets of peptides/proteins displaying similar physical or chemical features can also be employed to further reduce sample complexity and to facilitate increasingly comprehensive studies on specific subsets of proteins present at low cellular abundance. Specific examples include the selective isolation of glycopeptides [22] and phosphopeptides [23] as well as their chemically modified forms [24–29], and chemically reactive cysteinyl peptides [30–32].

Through an *in silico* digestion of the human Swiss-Prot protein database, only 22.5% of tryptic peptides were determined to contain cysteine amino acid residues. Thus, the specific selection and enrichment of the cysteinyl peptides provides a significant technical advantage for detecting low abundance proteins by effectively reducing sample complexity and dynamic range, as previously demonstrated in different biological systems [30–32]. Several methods have been developed to enrich cysteinyl peptides from complex proteome samples, including the use of isotope coded affinity tag (ICAT) reagents [30], reversible biotinylation of cysteinyl peptides [31], and reversible labeling of cysteinyl peptides using Ellman’s reagents and subsequent isolation of these modified cysteinyl peptides by combined fractional diagonal chromatography (COFRADIC™) [32]. A common feature of these methods includes the initial modification of cysteine residues with different reagents (reversibly or irreversibly) changing the affinity or physical properties of the cysteinyl peptides, that can then be exploited for subsequent isolation and enrichment of those modified cysteinyl peptides from complex mixtures using various chromatographic separations, such as avidin affinity chromatography and COFRADIC™. Nevertheless, several limitations associated with avidin affinity chromatography have been observed, including difficulty in removing nonspecifically bound peptides as well as reduced sample recovery due to irreversible binding of a subpopulation of the biotinylated peptides. The solid-phase approach for capturing cysteinyl peptides eliminates the need for avidin affinity chromatography, but contributes the new challenge of fragment ion production from the cysteine-bound tags during collision induced dissociation (CID)

complicating the analysis [33]. Reversible labeling of cysteinyl peptides with 5, 5'-dithiobis (2-nitrobenzoic acid) changed the peptide hydrophobicity, which allows their specific isolation utilizing COFRADIC™; however, additional chromatographic steps are required in this method.

We have recently developed a simplistic and high efficiency cysteinyl peptide enrichment (CPE) method using a thiol-specific covalent resin. This CPE method has been incorporated into the newly developed quantitative cysteinyl-peptide enrichment technology (QCET) for high throughput comparative proteome analysis [34]. The CPE method is highly specific, lends itself to automation for high throughput studies, and can readily be adapted for quantitative measurements. Furthermore, the reversible capture and release reaction of the CPE method for cysteinyl peptides generates no side reactions and the enriched cysteinyl peptides do not have a cumbersome, fragmentable tag. Here we report the application of this new CPE method to the global proteome analysis of human mammary epithelial cells (HMECs) using two-dimensional offline strong cation exchange (SCX) chromatography and reversed-phase (RP) capillary LC-MS/MS. By effectively reducing sample complexity, numerous additional proteins presumed to be of low abundance were identified by comparing the results from an identical analysis of a global digest with the CPE results. These additional identified proteins include protein kinases, phosphatases, small GTPases, transcription factors, and membrane receptors. Significant increases in sequence coverage for specific proteins were also observed due to the enhanced identification of cysteinyl peptides.

2 Materials and methods

2.1 Preparation of HMEC protein digests and standard mixture digest

The nontumorigenic human mammary epithelial cell line HMEC 184 A1L5 [35] was used in this study. The use of the HMEC cells was reviewed by the Pacific Northwest National Laboratory IRB for human subjects research in accordance with federal regulations. Cells were routinely cultured in DFC-1 medium (Gibco BRL, Gaithersburg, MD, USA) as previously described [36] until achieving 90% confluence. Cells were harvested, rinsed 3 times with ice-cold phosphate-buffered saline, and lysed with 0.4 mL/dish of lysis buffer (10 mM Tris, 150 mM NaCl, 1% NP-40, 1 mM NaVO₃, 10 mM NaF, and protease inhibitor cocktail, pH 7.4) using intermittent sonication on ice. The lysate was centrifuged for 20 min at 4 °C, 14,000×g to pellet any cell debris. The supernatant was collected and divided into aliquots and stored at -80 °C until further use. The protein concentration was determined using the bicinchoninic acid (BCA) protein assay (Pierce, Rockford, IL, USA). A total of 3 mg of protein from the cell lysate was subjected to tryptic digestion after disulfide bonds in the denatured proteins were reduced (50 mM Tris buffer, pH 8.2, 8 M urea, 10 mM tributyl phosphine) for 1 h at 37 °C. The reduced protein mixture was diluted 10 times using 20 mM Tris buffer, pH 8.2 and sequencing grade, modified porcine trypsin (Promega, Madison, WI, USA) was added at a trypsin:protein ratio of 1:50 and incubated at 37 °C overnight. The tryptic digest was loaded on a 1-mL SPE C18 column (Supelco, Bellefonte, PA, USA) and washed with 4 mL of 0.1% trifluoroacetic acid (TFA)/5% acetonitrile. Peptides were eluted from the SPE column with 1 mL of 0.1% TFA/80% acetonitrile and dried under reduced pressure. Peptide samples were stored at -80 °C until CPE treatment. All samples without CPE treatment were prepared in the same manner except the samples were alkylated immediately after the reducing step.

A five-protein mixture containing the following 5 proteins at the same concentration (15 pmol/mL) were obtained from Sigma (St. Louis, MO, USA): bovine serum albumin, bovine ribonuclease A, chicken lysozyme, chicken ovalbumin, and rabbit glyceraldehyde-3-phosphate dehydrogenase. A total of 100 µg of mixture was reduced, digested, desalted, and dried as described above.

2.2 Cysteinyll peptide enrichment using thiol-affinity resin

Unless otherwise noted, all solutions used in this step were degassed to prevent oxidation of the thiol content. The tryptic digests from 100 µg protein (small scale) or 3 mg protein (large scale) were reduced with 5 mM DTT in 20 µL of 50 mM Tris buffer, pH 7.5, 1 mM EDTA (coupling buffer) for 30 min at 37 °C, after which the samples were diluted to 100 µL by adding coupling buffer. 100 µL of Thiopropyl Sepharose 6B thiol-affinity resin (Amersham Biosciences, Uppsala, Sweden) were prepared from dried powder per the manufacturer's instruction. Briefly, the dried powder was rehydrated in water for 15 min and washed by 50 bed volumes of water, followed by 50 bed volumes of coupling buffer in a Handee Mini-Spin column (Pierce). The reduced peptide sample was then incubated with the resin for 1 h at room temperature (RT) with gentle mixing and the unbound portion was removed from the resin by spinning the column at low speed. The resin was washed in the spin column sequentially with 0.5 mL×5 of 50 mM Tris buffer, pH 8.0, 1 mM EDTA, (washing buffer), 2 M NaCl, 80% acetonitrile/0.1% TFA solution, and washing buffer. To release the captured cysteinyll peptides, 100 µL of 20 mM DTT freshly prepared in washing buffer was added to the resin and incubated for 30 min at RT. The resin was further washed with 100 µL of 80% acetonitrile. The pH of the combined eluates was adjusted to 8.0 and the sample was alkylated with 80 mM iodoacetamide for 30 min at RT in the dark. The samples were then desalted by SPE C18 column as described above and lyophilized. 50 µL of 25 mM NH₄HCO₃ was used to dissolve the small scale sample and 10 µL of each was used for LC-MS/MS analysis. A tryptic digest of 3 mg of HMEC proteins was used for the enrichment of cysteinyll peptides prior to SCX fractionation and subsequent LC-MS/MS analysis.

2.3 SCX chromatography

Enriched cysteinyll peptides from the large scale sample and the global tryptic digest were both reconstituted with approximately 1.0 mL of 10 mM ammonium formate, 25% acetonitrile, pH 3.0, and injected for strong cation exchange chromatography onto a Polysulfoethyl A 200 mm×4.6 mm column (PolyLC, Columbia, MD, USA) preceded by a 10 mm×10 mm guard column with a flow rate of 1 mL/min. The separations were performed with a Shimadzu LC-10A system (Columbia, MD, USA) utilizing a Unicam 4225 (Thermo Electron, Waltham, MA, USA) UV/vis detector with mobile phases consisting of solvent A: 10 mM ammonium formate, 25% acetonitrile, pH 3.0, and solvent B: 500 mM ammonium formate, 25% acetonitrile, pH 6.8. Once loaded, the run was isocratic for 10 min at 100% A. Peptides were then separated by a gradient from 0% to 50% B over 40 min, followed by a gradient of 50% to 100% B over 10 min. The gradient was then held at 100% solvent B for 10 min. A total of 35 fractions were collected using a Shimadzu FRC-10A fraction collector, after which each fraction was lyophilized to dryness and stored at -80 °C until analyzed.

2.4 Capillary LC-MS/MS analysis

The analyses of peptide samples were performed using a custom-built high pressure capillary LC system (described in detail elsewhere [37]) coupled online with a LCQ Deca XP ion trap mass spectrometer (ThermoFinnigan, San Jose, CA, USA) using an in-house-manufactured ESI interface. The reversed-phase capillary column was made by slurry packing 5-µm Jupiter C18 bonded particles (Phenomenex, Torrance, CA, USA) into a 65-cm-long, 150 µm-i.d.×360 µm-o.d. fused silica capillary (Polymicro Technologies, Phoenix, AZ, USA) that incorporated a 2-µm retaining stainless steel screen in an HPLC union (Valco Instruments Co., Houston, TX, USA). The mobile phase consisted of 0.2% acetic acid and 0.05% TFA in water (A) and 0.1% TFA in 90% acetonitrile/10% water (B). Mobile phases were degassed on-line using a vacuum degasser (Jones Chromatography Inc., Lakewood, CO, USA). After loading 10 µL of peptides onto the column, the mobile phase was held at 100% A for 20 min. Exponential gradient elution was performed by increasing the mobile-phase composition to 70% B over

150 min using a stainless steel mixing chamber. To identify the eluting peptides, the ion trap mass spectrometer was operated in a data-dependent MS/MS mode (m/z 400–2000), in which a full MS scan was followed by three MS/MS scans. The 3 most intensive precursor ions were dynamically selected and subjected to CID in the order of high intensity to low intensity, using a normalized collision energy setting of 45%. A dynamic exclusion duration of 3 min was used. The temperature of the heated capillary and the electrospray voltage were 200 °C and 2.2 kV, respectively.

2.5 MS/MS Data analysis

Peptides were identified by searching the MS/MS spectra against a nonredundant human International Protein Index (IPI) database (consisting of 47,306 protein entries at the time of analysis; available online at <http://www.ebi.ac.uk/IPI>) using SEQUEST (ThermoFinnigan) [38]. Static modification on cysteine residues was applied for the identification of iodoacetamide labeled cysteinyl peptides. The criteria used to filter raw SEQUEST identifications were similar to that reported by Washburn *et al.* [7]. Briefly, a delta correlation value (ΔC_n) greater than 0.1 and cross-correlation score (Xcorr) cutoff values of 1.9, 2.2, and 3.75 were used for fully and partially tryptic peptides with charge states of 1+, 2+, and 3+, respectively. Non-tryptic peptides were not included. In addition, an elution time constraint was placed upon peptide identifications based upon the predicted versus observed elution time in the LC separation prior to MS/MS analysis [39]. Such a technique has been previously shown to increase the overall confidence of the dataset [10]. Identified peptides which did not elute within $\pm 10\%$ of their predicted elution time were removed as possible false identifications. Protein names and descriptions were converted to the format of human Swiss-Prot, human NCBI, and TrEMBL databases for better reference.

2.6 Protein classification and categorization

The analyses of molecular weight (MW) and pI distributions of identified proteins were performed based upon information extracted from human IPI database and human NCBI database. Categorization of the biological function and cellular location of the identified proteins was performed with the assistance of Gene Ontology identification numbers (GO ID #s) downloaded from European Bioinformatics Institute (EBI) at <ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/HUMAN>. Cellular location (*cellular component*) and process (*biological process*) categorization of the detected proteins were based upon GO ID #s with relatively broad categories to better represent the data.

3 Results and discussion

3.1 Cysteinyl peptide enrichment

The process of specific capture/release of cysteinyl peptides is illustrated in Figure 1. Briefly, cysteinyl peptides are covalently captured by the resin through the formation of a mixed disulfide bond. The unbound, non-cysteinyl peptides and the released 2-thiopyridone are removed by washing. Stringent washing steps are used to remove the non-specifically bound peptides that could potentially be retained through either ionic or hydrophobic interactions with the resin matrix. The captured cysteinyl peptides are subsequently released by incubating the resin with a reducing reagent (e.g., DTT). The overall cysteinyl peptide capture/release reaction is simple and quantitative, as previously demonstrated [34].

The specificity and reproducibility of the CPE method were evaluated by preparing and analyzing three replicates of cysteinyl peptide-enriched samples from HMEC cell lysate. LC-MS/MS analysis of the three samples resulted in 427, 457, and 428 confident peptide identifications of which 97.4%, 96.1%, and 96.9% were determined to be cysteinyl peptides, respectively; the unbound, non-cysteinyl peptide samples were also analyzed and only less

than 1% of the identified peptides were determined to be cysteine-containing peptides, demonstrating the high specificity, efficiency, and reproducibility of this CPE method. The overlap of peptides identified (where the peptides have exactly the same sequences) between these 3 independent preparations was found to be 61.4%, 62.5%, and 59.3%, respectively. The ~60% overlap in peptide identifications was mainly due to the undersampling issue in one-dimensional data-dependent LC-MS/MS, since the overlap observed between replicate analysis of the same sample is also ~60%. However, the use of two-dimensional or multi-dimensional LC separation coupled to MS/MS could keep this problem at a minimum level.

The performance of the CPE method was further evaluated by comparing the peptide and protein identifications from HMEC global tryptic digests with and without CPE using one-dimensional LC-MS/MS analysis. The two samples were both analyzed in triplicate under identical conditions, and the average numbers of peptide and protein identifications were used for the comparison (Table 1, upper portion). With CPE the percentage of identified cysteinyl peptides increased from 23.7% to 96.8%, and proteins identified solely from cysteinyl peptides increased from 32.3% to 96.4%, demonstrating the specific isolation of cysteinyl peptides from the global complex tryptic digests. In addition, the total identified proteins increased by 48.7% (from 189 to 281) from LC-MS/MS analysis, presumably due to the reduced sample complexity/dynamic range from CPE. Another feature of this method, as previously demonstrated [34], is the ability to provide significantly greater coverage of cysteinyl peptides: a total of 42 out of 45 predicted tryptic cysteinyl peptides were confidently identified (93.3% coverage) from the five-protein mixture using this method, while only 15 of 44 predicted tryptic cysteinyl peptides (34.1% coverage) were detected from a six-protein mixture using the ICAT approach [40].

3.2 SCX-RPLC-MS/MS analyses

To increase overall proteome coverage, SCX and capillary RPLC were coupled offline for two-dimensional separation of peptide samples before MS/MS analyses. Briefly, following cell lysis, two identical aliquots of lysates were subjected to protein denaturation, reduction, and proteolytic digestion, one sample was alkylated by iodoacetamide while the other one was not alkylated. The tryptic digest without alkylation was treated by the thiol-specific covalent resin to enrich cysteinyl peptides. The resulting peptide mixture and the other alkylated tryptic digest without CPE were separately fractionated into 35 peptide fractions using SCX chromatography. All 70 SCX fractions were analyzed via high resolution capillary RPLC-MS/MS. A total of 202,538 MS/MS spectra were generated and 43,378 of them matched to peptides in the human protein database following stringent filtering of the SEQUEST results. Further screening of the data by applying the elution time constraint resulted in 14,416 different peptide identifications (86.1% fully tryptic peptides) of high confidence, from which a total of 4,294 different proteins were identified (a summary of these identified proteins is available as a supplementary table online, including protein database access number, descriptions, peptides identified, and charge state, Xcorr, and ΔC_n of each peptide identification). Compared to the results of a multidimensional proteome analysis of the same cell line without CPE [10], 1,224 of 1,574 previously identified proteins were observed in this study, representing a 77.8% overlap between these two independent experiments.

The 2D LC separation and analysis of peptide samples with and without CPE is illustrated in Figure 2. Figure 2A and Figure 2B show the SCX chromatograms of the enriched cysteinyl peptides and the global tryptic digest, respectively. The RPLC-MS/MS analyses of two fractions eluted at the same time (~40 min) in these two SCX separations were shown in Figure 2C and Figure 2D (base-peak chromatograms). In the survey scans of the peptides eluted at ~75 min (Figure 2E and Figure 2F), precursor ions with m/z 1065.8 were selected for further sequence analysis by CID, and both of them were identified as the fully tryptic cysteinyl

peptide, IEAELQDICNDVLELLDK, from 14-3-3 protein β/α , as shown in Figure 2G and Figure 2H. The comparison of SCX fractionation, RPLC separation and MS and MS/MS detection of the cysteinyl peptide-enriched sample and the global tryptic digest demonstrated the effectiveness of the CPE method in reducing sample complexity (Figure 2, A versus B, C versus D, and E versus F, respectively) as well as the reliability and reproducibility of the analytical systems.

3.3 Detection of low abundance proteins and increased sequence coverage

Table 1 (lower portion) shows the comparison of the peptide and protein identifications from the SCX-RPLC-MS/MS analyses of the HMEC tryptic digest with and without CPE. With the application of CPE, 5,348 cysteinyl peptides were identified out of 5,924 total peptide identifications, representing a specificity of 90.3% for the CPE method even when extensive separations were used; 93.0% of the proteins were identified using only cysteinyl peptides. Figure 3 further illustrates the effectiveness of cysteinyl peptide enrichment. Note that in the analysis of the HMEC global tryptic digest without CPE (where the cysteine residues were alkylated), the percentage of detected cysteinyl peptides (23.1%) closely parallels the results obtained from the *in silico* digestion of human protein database (22.5%).

The numbers of total identified proteins from samples with and without CPE are comparable (3,198 without CPE; 2,904 with CPE); however, a total of 1,096 additional proteins were identified only from the cysteinyl peptide-enriched sample, representing an increase of 34.3% in proteome coverage. Most of these additionally identified proteins presumably are of relatively low abundance, including protein kinases, phosphatases, small GTPases, transcription factors, and membrane receptors. Several lines of evidence suggest CPE facilitates the detection of specific subsets of low abundance proteins, which are not accessible through the global analysis without CPE, due to the dynamic range limitation. For example, several different integral membrane metalloproteinases (ADAM 10, ADAM 15, and ADAM 17) were identified only from the cysteinyl peptide-enriched sample. They cleave the membrane-bound precursor of TNF- α to its mature soluble form and are responsible for the proteolytic release of several other cell-surface proteins. Ephrin type-A receptor 1, ephrin type-B receptor 3, ephrin-B2 (integral membrane proteins), and ephrin-A5 (attached to the membrane by a GPI-anchor) were also identified only from the CPE sample. Other examples include proteins involved in transcription regulation (activator 1 small subunit family members, glucocorticoid receptor, components of the origin recognition complex), growth regulation and cell cycle (band 4.1-like protein 3, cyclin-dependent kinase inhibitors), signal transduction (interleukin-6 and interleukin-9 receptors, toll-interacting protein, toll-like receptor 8, serine/threonine protein phosphatases 4, 5, and 6), and secreted proteins (kallikrein subfamily members, Kunitz-type protease inhibitor 1).

In addition to the increased detection of low abundance proteins, 1,390 proteins were observed as having greater sequence coverage among the 1,808 proteins observed in both analyses. Examples of increased sequence coverage for protein identifications with the use of CPE are shown in Table 2. The quality of the dataset is well demonstrated by the high sequence coverage per single protein in the combined results. In general, over 280 detected proteins have at least 10 or more unique peptide identifications, while more than 840 proteins have at least 5 or more unique peptide identifications. Totally, 54.2% of detected proteins (2,326) have at least 2 or more unique peptide identifications. The high efficiency CPE method not only significantly increases the total number of protein identifications, but also greatly improves the sequence coverage for specific proteins, thus providing broader proteome coverage and higher quality protein identifications.

3.4 Protein classification and categorization

All identified proteins were first classified based on their physiochemical properties. The distribution of pI values over a range of 3–13 for both the identified proteins and the entire human proteome (37,490 entries from the NCBI database) are shown in Figure 4A and Figure 4B, respectively. Similar to a previous computational analysis [41], our results showed that the distribution of the predicted pI values of human proteins (both the identified proteins and the human proteome) is typical of a eukaryotic cell, displaying an approximate tri-modal distribution. Comparing the pI value distributions of the identified proteins and the human proteome, it appears that our results have good coverage for both cytosolic and membrane proteins, but the coverage is slightly biased toward the identification of cytosolic proteins (cluster at pI 6), presumably due to the better solubility and hence detection of cytosolic proteins compared to membrane proteins (cluster at pI 9). Improved coverage of membrane proteins from membranes and subcellular organelles can be achieved by coupling CPE with further subcellular fractionation and better membrane-solubilizing strategies. The molecular weight (MW) distributions of the identified proteins and the entire human proteome are shown in Figure 4C and Figure 4D. Proteins in the 20–40 kDa range were the most prevalent in the identified proteins, while for the human proteome, proteins in the <20 kDa range were the most prevalent. This is probably due to the higher content of truncated forms of proteins, bioactive peptides with small MW in the human protein database. The overall trends of MW distribution of the identified proteins and the human proteome are very similar: the majority of proteins are in the 0–140 kDa range and comparatively few have a MW greater than 300 kDa.

For comparison, the proteins identified in the cysteinyl peptide-enriched sample and the global sample were categorized separately according to their biological process and cellular location using Gene Ontology (GO) categorization (see Materials and Methods for details). A total of 91.8% of identified proteins were categorized, and the cellular location and biological processes of the identified proteins are illustrated in Figure 5. The majority of identified proteins originated from either the nucleus, membranes, or cytoplasm. Particularly, 292 proteins were categorized as mitochondrial proteins, even without prior organelle purification using sucrose gradients [18]. In general, the distributions of cellular location for proteins identified from the cysteinyl peptide-enriched sample and the global sample were nearly identical. (Figure 5A and Figure 5B). A wide range of proteins involved in a diversity of biological processes were identified in this study and were classified into 17 general categories. Cellular/physiological process is a diverse category including proteins responsible for cell cycle, cytoskeleton organization, cell mobility, cell adhesion, cell proliferation, etc.; it was found to be the largest category for all identified proteins (763 proteins). Particularly, 16 proteins related to epidermal differentiation were identified in the Development category (126 proteins), including long chain collagen, cornifin A, cornifin B, envoplakin, desmoplakin, S100 calcium-binding protein A7, and various forms of keratins. The signal transduction category (397 proteins) includes 65 protein kinases, 28 phosphatases, 36 Ras-related proteins, and 7 different 14-3-3 proteins. Many additional proteins were classified in categories related to the metabolisms of different biomolecules (1,240 proteins), such as amino acids, carbohydrates, lipids, nucleotides, proteins, and RNA/DNAs, and proteins involved in energy pathways (57 proteins). Again, no significant difference was observed between the distributions of the functionally classified proteins that were identified in the samples with and without CPE (Figure 5C and Figure 5D). Moreover, the classifications of the proteins identified in these two samples showed similar MW, pI, and gene location distributions (data not shown), indicating that the CPE method is not biased by these characteristics, supported by the fact that 96.4% protein entries in the human database contain cysteine residues.

4 Concluding remarks

The work described here demonstrates the effectiveness of the incorporation of the high efficiency cysteinyl peptide enrichment toward more comprehensive proteome analysis of the complex human cellular system. The incorporation of cysteinyl peptide enrichment significantly increases the overall proteome coverage (34.3% increase over global analysis without cysteinyl peptide enrichment) and also greatly improves the sequence coverage for specific proteins, providing for the identification of more low abundance proteins as well as higher confidence protein identifications. In addition, cysteinyl peptide enrichment is not biased towards protein pI, MW, cellular location, and biological function. The effectiveness of this method in reducing sample complexity/dynamic range has enabled high throughput qualitative and quantitative proteome analyses of human cell lines through the application of accurate mass and time tags [34]. Combining the cysteinyl peptide enrichment with other separation strategies prior to MS analysis, such as organelle purification/enrichment and multidimensional/nano-scale separations, will further reduce the sample complexity/dynamic range thus enabling global identification of low abundance regulatory proteins that are involved in a diversity of important cellular events.

Acknowledgements

We thank the NIH National Center for Research Resources (RR018522) for support of this research and the Environmental Molecular Sciences Laboratory (EMSL) at Pacific Northwest National Laboratory (PNNL) for use of the instrumentation applied in this research. The EMSL is a national scientific user facility sponsored by the U.S. Department of Energy's (DOE) Office of Biological and Environmental Research. PNNL is operated by Battelle for the DOE under Contract DE-AC06-76RLO 1830.

References

1. Ren SX, Fu G, Jiang XG, Zeng R, et al. *Nature* 2003;422:888–893. [PubMed: 12712204]
2. Venter JC, Remington K, Heidelberg JF, Halpern AL, et al. *Science* 2004;304:66–74. [PubMed: 15001713]
3. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, et al. *Nature* 2002;420:520–562. [PubMed: 12466850]
4. Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, et al. *Nature* 2004;428:493–521. [PubMed: 15057822]
5. Venter JC, Adams MD, Myers EW, Li PW, et al. *Science* 2001;291:1304–1325. [PubMed: 11181995]
6. Patterson SD, Aebersold R. *Nat Genet* 2003;33:311–323. [PubMed: 12610541]
7. Washburn MP, Wolters D, Yates JR. *Nat Biotechnol* 2001;19:242–247. [PubMed: 11231557]
8. Florens L, Washburn MP, Raine JD, Anthony RM, et al. *Nature* 2002;419:520–526. [PubMed: 12368866]
9. Peng J, Elias JE, Thoreen CC, Licklider LJ, Gygi SP. *J Proteome Res* 2003;2:43–50. [PubMed: 12643542]
10. Jacobs JM, Mottaz HM, Yu LR, Anderson DJ, et al. *J Proteome Res* 2004;3:68–75. [PubMed: 14998165]
11. Link AJ, Eng J, Schieltz DM, Carmack E, et al. *Nat Biotechnol* 1999;17:676–682. [PubMed: 10404161]
12. Rappsilber J, Ryder U, Lamond AI, Mann M. *Genome Res* 2002;12:1231–1245. [PubMed: 12176931]
13. Garin J, Diez R, Kieffer S, Dermine JF, et al. *J Cell Biol* 2001;152:165–180. [PubMed: 11149929]
14. Journet A, Chapel A, Kieffer S, Roux F, Garin J. *Proteomics* 2002;2:1026–1040. [PubMed: 12203898]
15. Wubbolts R, Leckie RS, Veenhuizen PT, Schwarzmann G, et al. *J Biol Chem* 2003;278:10963–10972. [PubMed: 12519789]
16. Wasiak S, Legendre-Guillemain V, Puertollano R, Blondeau F, et al. *J Cell Biol* 2002;158:855–862. [PubMed: 12213833]
17. Han DK, Eng J, Zhou H, Aebersold R. *Nat Biotechnol* 2001;19:946–951. [PubMed: 11581660]

18. Taylor SW, Fahy E, Zhang B, Glenn GM, et al. Nat Biotechnol 2003;21:281–286. [PubMed: 12592411]
19. Bell AW, Ward MA, Blackstock WP, Freeman HN, et al. J Biol Chem 2001;276:5152–5165. [PubMed: 11042173]
20. Dundr M, Misteli T. Mol Cell 2002;9:5–7. [PubMed: 11804579]
21. Brunet S, Thibault P, Gagnon E, Kearney P, et al. Trends Cell Biol 2003;13:629–638. [PubMed: 14624841]
22. Kaji H, Saito H, Yamauchi Y, Shinkawa T, et al. Nat Biotechnol 2003;21:667–672. [PubMed: 12754521]
23. Ficarro SB, McClelland ML, Stukenberg PT, Burke DJ, et al. Nat Biotechnol 2002;20:301–305. [PubMed: 11875433]
24. Wells L, Vosseller K, Cole RN, Cronshaw JM, et al. Mol Cell Proteomics 2002;1:791–804. [PubMed: 12438562]
25. Zhang H, Li XJ, Martin DB, Aebersold R. Nat Biotechnol 2003;21:660–666. [PubMed: 12754519]
26. Zhou H, Watts JD, Aebersold R. Nat Biotechnol 2001;19:375–378. [PubMed: 11283598]
27. Oda Y, Nagasu T, Chait B. Nat Biotechnol 2001;19:379–382. [PubMed: 11283599]
28. McLachlin DT, Chait BT. Anal Chem 2003;75:6826–6836. [PubMed: 14670042]
29. Qian WJ, Goshe MB, Camp DG 2nd, Yu LR, et al. Anal Chem 2003;75:5441–5450. [PubMed: 14714534]
30. Gygi SP, Rist B, Griffin TJ, Eng J, Aebersold R. J Proteome Res 2002;1:47–54. [PubMed: 12643526]
31. Spahr CS, Susin SA, Bures EJ, Robinson JH, et al. Electrophoresis 2000;21:1635–1650. [PubMed: 10870950]
32. Gevaert K, Ghesquiere B, Staes A, Martens L, et al. Proteomics 2004;4:897–908. [PubMed: 15048972]
33. Borisov OV, Goshe MB, Conrads TP, Rakov VS, et al. Anal Chem 2002;74:2284–2292. [PubMed: 12038753]
34. Liu T, Qian WJ, Strittmatter EF, Camp DG, et al. Anal Chem. 2004In press.
35. Stampfer MR, Yaswen P. Cancer Surv 1993;18:7–34. [PubMed: 8013001]
36. Chen WNU, Woodbury RL, Kathmann LE, Opresko LK, et al. J Biol Chem 2004;279:18488–18496. [PubMed: 14978035]
37. Shen Y, Zhao R, Belov ME, Conrads TP, et al. Anal Chem 2001;73:1766–1775. [PubMed: 11338590]
38. Eng JK, McCormack AL, Yates JR. J Am Soc Mass Spectrom 1994;5:976–989.
39. Petritis K, Kangas LJ, Ferguson PL, Anderson GA, et al. Anal Chem 2003;75:1039–1048. [PubMed: 12641221]
40. Gygi SP, Rist B, Gerber SA, Turecek F, et al. Nat Biotechnol 1999;17:994–999. [PubMed: 10504701]
41. Schwartz R, Ting CS, King J. Genome Res 2001;11:703–709. [PubMed: 11337469]

Figure 1.

Enrichment of cysteinyl peptides using thiol-specific covalent resin. Cysteinyl peptides were captured by the resin through the formation of a mixed disulfide bond. A stringent wash was applied to remove the non-specifically, noncovalently bound peptides, after which a reducing reagent (e.g. DTT) was added to release the bound cysteinyl peptides.

Figure 2.

Comparison of SCX-RPLC-MS/MS analyses of HMEC global tryptic digests with (A, C, E, G) and without (B, D, F, H) cysteinyl peptide enrichment. Examples are provided for a peptide identified in both analyses. Samples were fractionated by SCX chromatography (A, B), and the fractions that eluted at ~40 min in each analysis were analyzed by RPLC-MS/MS under the same conditions (C, D). Corresponding peptides eluted at ~75 min during the LC-MS/MS analyses were shown in E and F, respectively. Peptides with m/z 1065.8 in each analysis were selected for data-dependent MS/MS and both were subsequently identified as the fully tryptic cysteinyl peptide IEAEIQDICNDVLELLDK from 14-3-3 protein β/α (G, H).

Figure 3.

Efficiency of cysteinyl peptide enrichment. HMEC tryptic digests with and without cysteinyl peptide enrichment were fractionated into 35 fractions by SCX chromatography, and each resulting fraction was analyzed by RPLC-MS/MS. The distributions of cysteinyl peptide and total peptide identifications in each fraction for untreated (A) and cysteinyl peptide-enriched sample (B) are shown. Without cysteinyl peptide enrichment, only 23.1% of identified peptides contain cysteine, while 90.3% of peptides were determined to be cysteine-containing peptides with cysteinyl peptide enrichment.

Figure 4.

Classification of total identified proteins (A, C) and the human proteome (B, D) based on physiochemical properties. pI distributions were plotted for the total identifications (A) and all human proteins (B) using predicted pI values extracted from NCBI human protein database, over increments of 0.1 pH units. MW distributions were plotted for the total identifications (C) and all human proteins (D) using 20 kDa increments.

Figure 5.

Comparison of the protein categorization using GO identification numbers (*cellular component* or *biological process*). Identified proteins from samples with (A) and without (B) cysteinyl peptide enrichment were categorized based upon their cellular location.

Categorization based upon the predicted biological roles was also performed for proteins identified from samples with (C) and without (D) cysteinyl peptide enrichment. Results shown here represent 91.8% of total identified proteins.

Table 1

Summary of HMEC peptide and protein identifications with and without cysteinyl peptide enrichment (CPE).

Sample	Total Cys-peptides ^{a)}	Total peptides	Cys-peptide %	Total Cys-proteins ^{b)}	Total proteins	Cys-protein %
<i>LC-MS/MS</i> ^{c)}						
Without CPE ^{d)}	88	372	23.7	61	189	32.3
With CPE	423	437	96.8	271	281	96.4
<i>SCX-RPLC-MS/MS</i>						
Without CPE	2,405	10,408	23.1	1,327	3,198	41.5
With CPE	5,348	5,924	90.3	2,701	2,904	93.0
Overlap	1,624	1,916	84.8	1,137	1,808	62.9
Total	6,129	14,416	42.5	2,891	4,294	67.3

^{a)}Cys-peptide is cysteinyl peptide^{b)}Cys-protein means protein identifications containing cysteinyl peptide hit(s)^{c)}Samples were analyzed in triplicate and listed are the average numbers^{d)}All cysteinyl residues were alkylated for peptide samples prepared without CPE

Table 2

Increased sequence coverage for specific proteins with the application of cysteinyl peptide enrichment.

Increase ^{a)}	Protein Name and ID ^{b)}	Sequence	Cys	Global
83.3%	Gemin 5 GEM5_HUMAN	R.VSGFTFSHHPGQYNLCATSSDDGTVK.I R.LLCVAWSPDPDCIYSGADDFCVHK.W R.SKEELHQDCLVLATAK.H R.TIFCLTCSPHHEDLVAIGYK.D K.LICTIQQHHK.L K.TVVTEHALHQHTISTLHWSPR.V	Yes ^{c)} Yes Yes Yes Yes	Yes ^{d)}
80.0%	Osmotic stress protein 94 OS94_HUMAN	K.SKELDNFCNPIYKPK.P R.NFDEALVDYFCDEFK.T K.ELDNFCNPIYKPK.P R.AQFEQLCASLLAR.V K.KPVADCVISIPSFFTDAER.R	Yes Yes Yes Yes Yes	Yes
75.0%	WD-repeat protein 11 WD11_HUMAN	K.ILITQEKPSAEFITLNDCLQLAYLPSK.R K.LAEGVQLLCLIDK.A K.LSNPLDICYDVLCEWAYFQK.F K.NFPTITALEWSPSHNLK.S	Yes Yes Yes Yes	Yes
62.5%	Diacylglycerol kinase, alpha KDGA_HUMAN	K.DVVCLNDVSCYFSLLEGGRPEDK.L R.HLSLALFQSFETGHCLNETNVTG.G K.ALPCEVSTYAK.S R.ILVCGGDGTVGWILETIDK.A K.DVVCLNDVSCYFSLLEGGRPEDKLEFTFK.L R.RPHGDIYGINQALGATAK.V R.VAEYLDWDVSELRPIQEMMK.E R.AGATTVPLLVLGLEMTLK.D	Yes Yes Yes Yes Yes Yes Yes Yes	Yes Yes Yes
50.0%	Cullin 4B CU4B_HUMAN	K.HECGAAFTSK.L K.YNLEELYQAVENLCSYK.I K.VDHIIDICFLK.N R.SLLSMLSDLQIYQDSFEQR.F K.TLSHNLLVSEVYNQLK.F K.RPNKPAELIAK.Y	Yes Yes Yes	Yes Yes Yes

^{a)} Percentile increased in sequence coverage by using cysteinyl peptide enrichment^{b)} Protein IDs are given according to Swiss-Prot nomenclature, proteins names are extracted from the same database^{c)} Peptides identified in HMEC tryptic digest with cysteinyl peptide enrichment (Cys)^{d)} Peptides identified in HMEC tryptic digest without cysteinyl peptide enrichment (Global)