Precise and Parallel Characterization of Coding Polymorphisms, Alternative Splicing, and Modifications in Human Proteins by Mass Spectrometry* [S]

Michael J. Roth, Andrew J. Forbes, Michael T. Boyne II, Yong-Bin Kim, Dana E. Robinson, and Neil L. Kelleher‡

The human proteome is a highly complex extension of the genome wherein a single gene often produces distinct protein forms due to alternative splicing, RNA editing, polymorphisms, and postranslational modifications. Such biological variation compounded by the high sequence identity within gene families currently overwhelms the complete and routine characterization of mammalian proteins by MS. A new data base of human proteins (and their possible variants) was created and searched using tandem mass spectrometric data from intact proteins. This first application of top down MS/MS to wild-type human proteins demonstrates both gene-specific identification and the unambiguous characterization of multifaceted mass shifts ($\Delta m$ values). Such $\Delta m$ values found from the precise identification of 45 protein forms from HeLa cells reveal 34 coding single nucleotide polymorphisms, two protein forms from alternative splicing, and 12 diverse modifications (not including simple N-terminal processing), including a previously unknown phosphorylation at 10% occupancy. Automated protein identification was achieved with a median expectation value of $10^{-13}$ and often occurred simultaneously with dissection of diverse sources of protein variability as they occur in combination. Top down MS therefore has a bright future for enabling precise annotation of gene products expressed from the human genome by non-mass spectrometers. Molecular & Cellular Proteomics 4:1002–1008, 2005.

Due to the presence of polymorphisms, alternative splicing, and postranslational modifications (PTMs) the human proteome is highly complex, often encoding multiple protein forms for a given gene (1). This biological complexity poses a significant analytical and bioinformatic challenge to the detailed analysis of mammalian proteomes by MS and is exacerbated by the presence of gene families sharing high sequence identity (2, 3). Protein modifications are often indicative of changes in cellular or tissue dynamics and therefore play central roles in regulation of the cell cycle or development of disease. Whether for new diagnostics or understanding molecular mechanisms in cell biology, protein identification using tryptic peptides has revolutionized the analysis of complex mixtures by mass spectrometry (1, 4).

High throughput platforms based on MALDI (5) and ESI use MS/MS engines capable of spectral acquisition at a rate of $>10^3$/week (6, 7). Recent studies indicate significant inefficiencies associated with such large scale “bottom up” analyses in mammalian systems including imperfect enzymatic cleavage (8, 9) and some MS/MS spectra requiring manual interpretation/validation for identification. Despite the lingering difficulties with peptide analysis, it provides the best and most general method for large scale protein identification today with information on nonsynonymous coding single nucleotide polymorphisms (cSNPs), alternative splicing (10), and PTMs challenging to obtain (2).

Recent developments by MacCoss et al. (11), Wu et al. (12), and Zhu et al. (13) use three proteases and multidimensional protein identification technology (“MudPIT”) or isoelectric focusing, reversed-phase chromatography, and three mass spectrometers (13), respectively, to obtain mass information on ~70–99% of the primary protein structure. Combining intact protein measurement with near exhaustive peptide analysis of five proteins from human cells allowed detection of N-terminal modifications and one alternatively spliced transcript (13). Although cSNP analysis of abundant blood proteins is possible (14), a general informatic strategy has yet to systematically integrate DNA and RNA level data with the MS-based interrogation of the human proteome. This is accomplished here using a data base of human proteins tailored

From the Department of Chemistry, University of Illinois Urbana-Champaign, Urbana, Illinois 61801

Received, March 6, 2005, and in revised form, April 27, 2005

Published, MCP Papers in Press, April 28, 2005, DOI 10.1074/mcp.M500064-MCP200

© 2005 by The American Society for Biochemistry and Molecular Biology, Inc.

This paper is available on line at http://www.mcponline.org

© 2005 by The American Society for Biochemistry and Molecular Biology, Inc.

This paper is available on line at http://www.mcponline.org
for the “top down” MS approach by combinatorial consideration of protein variability during a search (i.e. “shotgun annotation”) (15). Although nucleic acid-based approaches represent the highest throughput and best overall methods for capturing information about SNPs, proteomics-based approaches allow CSNP genotyping concurrent to modification and splice variant identification.

The direct fragmentation of intact protein ions using FTMS now provides expectation values (Pscores) that are orders of magnitude better than searches based on tryptic peptides (16–18), a far more efficient and robust reconstruction process for the primary structure of the mature protein, and detection of more diverse mass discrepancies (Δm values) than targeted analysis approaches (e.g. for phosphopeptides). Major limitations for top down MS are difficulties in handling proteins >50 kDa routinely, low percent occupancy and multivalent PTMs (such as glycosylations) are difficult to detect, and only medium scale projects <<200 proteins from microorganisms have been achieved (19). The top down MS/MS approach using standard fragmentation methods or electron capture dissociation (ECD) has provided 100% coverage with localization of basic PTMs for proteins in Bacteria (17, 20), Archaea (16, 17), yeast (19, 21, 22), and a plant (23).

Here we demonstrate unparalleled characterization of human (nuclear) proteins revealing seven different types of modifications in regulation and maturation including a novel phosphoprotein. This was achieved by extending the data base concept of shotgun annotation from a single human histone (15) to a proteomic scale and required the integration of diverse DNA, RNA, and protein level information. This work establishes the basis for routine application of top down MS to capture coding haplotypes within a gene and allele-specific splicing and modification patterns on a far greater number of human proteins.

**EXPERIMENTAL PROCEDURES**

**Cell Culture and Lysate Fractionation**—Human HeLa-S3 cells were grown to a density of 0.6 × 10⁶ cells/ml using Joklik’s modified minimum essential medium and supplemented with 5% newborn calf serum. Cells were harvested using centrifugation at 2500 × g and two washes in cold PBS. The nuclei were precipitated and isolated using detergent washes, and the cytosol was extracted (24). The isolated nuclei were then resuspended, and for a portion of the extract, the chromatin (including DNA-binding proteins) was precipitated by adding 0.5 N NaCl and 5 mM MgCl₂. The proteins in solution were then loaded onto a prep cell (Bio-Rad) with a 12% T gel using an acid-labile surfactant (21). Proteins from the prep cell fractions were precipitated, treated at pH 2 for 1 h, and then separated using a Symmetry C₄ reversed-phase (RP) LC column (Waters, Milford, MA). For ~25% of identified proteins including barrier-to-autointegration factor (BAF) (see Fig. 4), the PF two-dimensional (2D) system (Beckman Coulter) was used for separation of proteins by pI, and then RPLC was carried out as outlined in the PF 2D manual.

**ESI/Q-FTMS**—Fractionated protein mixtures were suspended in ESI solution (49.5% MeOH, 49.5% H₂O, and 1% formic acid) and centrifuged at 14,000 rpm for 10 min. Sample solutions were then loaded into a 96-well plate and automatically introduced to the mass spectrometer using the NanoMate 100 (Advion BioSciences, Ithaca, NY). Approximately 10 µl of solution from each well were infused by automated nanospray into the heated metal capillary source. Typical samples enabled more than 40 min of stable nanospray providing sufficient time to acquire high quality broadband MS, threshold MS/MS, and ECD MS/MS scans for two to three intact proteins per sample. In cases of insufficient fragmentation for precise localization of PTMs, excess sample was used in a more targeted fashion, and in some cases a greater number of scans were summed for collisionally activated dissociation (CAD), infrared multiphoton dissociation (IRMPD), or ECD.

The instrument used in this study was a custom 8.5-tesla Q-FTMS of the Marshall design (25). In the case of CAD external to the magnet bore, ions were selected using the quadrupole and fragmented using electrostatic acceleration (10–45 V) into an octopole pressurized to ~10 millitorr with nitrogen gas. In the case of IRMPD or ECD, a SWIFT window 7 m/z wide was used. The isolated charge state was then dissociated using infrared laser radiation for 0.25–0.45 s (with a beam expander mounted in front of the laser, 40 watts, 75% power). After threshold dissociation, the quad-enhanced and SWIFT-isolated species was dissociated using ECD. Electrons were introduced to the cell for 100–200 ms using a dispenser cathode 35 inches from the center of the magnet. The kinetic energy of the electrons was controlled by placing a 1–2-V bias potential on the filament of the dispenser cathode.

**Automated Data Acquisition**—A custom TCL automation script first acquired 5–10 broadband scans followed by a quadrupole marching experiment, and upon completion a modified THRASH algorithm (26) automatically determined M values resulting in a peak list that was then used to select proteins for MS/MS analysis. The most abundant charge state of each protein was selectively accumulated using a notch-filtering quadrupole window 10 m/z wide automatically acquiring 5–10 scans. For targeted proteins, 25 or 50 scans of axial CAD or IRMPD were recorded to yield protein identifications. Automatically acquired ECD spectra were the sum of 100 scans.

**Construction of the Custom Human Data Base**—A highly annotated data base of human protein forms was created within ProSight Warehouse (27) using conflict sequences, splicing data, PTMs from UniProt (28), SNP information from dbSNP, and a variety of manually entered data, such as new PTMs found in the primary literature. UniProt data bases were transformed from Swiss-Prot format by a custom data base loader created using Perl scripts and BioPerl libraries. To populate the data base with SNP information, dbSNP was queried for nonsynonymous, coding polymorphisms with an available corresponding protein accession number. The resultant information was populated to a local data base. Using a portion of dbSNP running locally, protein sequence information and function/description were obtained. Using custom Perl scripts, the results were converted to the necessary ProSight Warehouse format. A data base loader application then extracted the protein information and populated ProSight Warehouse with all possible protein forms based on combinations of known variations for each gene product (15). The current number of protein forms in the human data base is 2,823,267 yielding a structured query language data base of 3.5 gigabytes with 17,333 proteins containing 1–10 cSNPs for subsequent searching using ProSight Retriever (29).

**Data Analysis and Data Base Searching**—Intact protein MS and MS/MS data were analyzed by THRASH (26) resulting in a protein list and fragment ion list that were uploaded onto the ProSight PTM (27) web server for data base searching (prosightptm.scs.uiuc.edu). The criteria for data base searching were generally a ±2000 M₁ window and 5–20-ppm tolerance for fragment ions with default search options selected as follows; Met, on/off; acetyl, on/off; and SNPs, on. Pscores reported in this study were calculated as reported previously (16), and...
those <10⁻³ required no manual validation of the identification result. Unless noted otherwise, Mᵣ and fragment ion mass values reported are for neutral, monoisotopic peaks (using external calibration), and protein identification numbers are UniProt primary accession numbers.

RESULTS AND DISCUSSION

Genotyping by Top Down MS— With one SNP present every ~1 kb in the human genome and 50,973 cSNPs currently known in dbSNP alone, well over half of human genes contain cSNPs, and top down MS/MS should enable robust genotyping even in the presence of PTMs. Fractions generated from a previously reported 2D separation of intact proteins (21) typically contain multiple proteins of varying abundance as in the ESI/Q-FTMS spectrum of Fig. 1a. Of the seven components, proteins of 6657.71 Da and 11,644.8 Da were selectively accumulated and fragmented by CAD and separately using ECD (spectra not shown). The CAD fragmentation data of Fig. 1b identified the 6.7-kDa component as a mitochondrial proteolipid (Pscore, 4×10⁻⁷) containing a known cSNP encoding an I₉V residue change (Δm = 14.02 Da). Only the Ile-9 allele was observed with an intact mass error of 18 ppm. The 11.6-kDa component was identified from the Fig. 1c MS/MS data to be calgizzarin S₁₀₀C (Pscore, 1×10⁻¹²). The calgizzarin gene contains a cSNP translating to a 1-Da variability (E₃₆K), readily resolved for the Glu-36 allele observed in the background of N-terminal methionine loss/acetylation (overall 0.6-ppm error). This illustrates the efficiency of intact protein MS/MS for genotyping cSNPs, a feat not often possible using digestion-based approaches. Determination of minihaplotypes in coding regions (i.e. the co-occurrence of multiple alleles in a coding sequence) should also be possible using endogenous material itself instead of in vitro produced/artificial peptides from PCR products (30).

Gene-specific Identification and Genotyping of a Modified Protein— Two-dimensional fractionation of a nuclear protein

Fig. 1. Complete characterization of multiple cSNP-containing proteins from one fraction. a, partial ESI/Q-FT mass spectrum (10 scans) of an acid-labile surfactant PAGE/RPLC sample from human cells. b, tandem mass spectrum (50 scans) from collisional dissociation of a 6.7-kDa protein selectively accumulated and fragmented using the quadrupole enhancement to FTMS. c, tandem mass spectrum (50 scans, axial CAD) from dissociation of the 11.6-kDa species at 905 m/z. d and e, graphical fragment maps generated upon data base retrieval using the MS/MS spectra of proteins highlighted in a (insets). Tall and short markers represent fragment ions produced from CAD (b/y-type) and ECD (c/z'-type), respectively. The circle in e indicates N-terminal acetylation. The shaded residue occurs at known cSNP sites. theo, theoretical; exp, experimental.

Fig. 2. Intact and MS/MS fragmentation spectra providing high retrieval specificity obtained for a modified, cSNP-containing member of a highly conserved gene family. a, broadband ESI/FTMS spectrum (10 scans) of an acid-labile surfactant PAGE/RPLC fraction from human HeLa cells. b, auto-SWIFT isolation spectrum (10 scans) of the 18+ charge state at 779 m/z. c, partial auto-ECD MS/MS spectrum (100 scans) of the species of b. d, the graphical fragment map generated upon data base retrieval from ECD and CAD fragmentation data illustrating the position of the cSNP within the histone H₂A gene. theo, theoretical; exp, experimental. Tall and short markers indicate fragment ions from CAD and ECD, respectively. The circle in d indicates N-terminal acetylation. The shaded residue occurs at a known cSNP site.
extract from asynchronous HeLa cells yielded various fractions containing core histones. Processing of one such sample by automated MS/MS provided ECD data for a 13,997.8-Da component of only 8% relative abundance (Fig. 2a). These MS/MS data (Fig. 2c) specifically identified histone H2A family member O from 29 distinct H2A forms (17 gene family members and their variants; Supplemental Fig. 1a) with a 10⁻¹⁸ Pscore. A sequence alignment was performed on H2A.O with the five most homologous protein forms in the H2A family (>80% identity; Supplemental Fig. 1b), revealing that four fragment ions (of the 19 automatically assigned) provided the specificity for precise and automatic identification of H2A.O versus the next best match. The H2A.O gene also contains a cSNP at residue 124 leading to a His→Tyr change (Δm = 26.00 Da). Only the His-124 form was observed indicating that these cells are homozygous at this locus. The observed intact mass contained a Δm of 42.01 ± 0.02 Da localized to the first five N-terminal residues (Fig. 2d). This Δm is most likely acetylation of the N terminus, although this same modification at Lys-5 is formally possible. Thus, an automated data flow can now differentiate between post-translationally modified and cSNP-containing isoforms even in highly conserved gene families.

Identification and Semiquantitative Analysis of Alternative Splice Variants—In a separate sample, Q-FTMS/MS analysis automatically identified a 11,977.9-Da protein as prothymosin α (ProTα; Fig. 3d). ProTα is encoded by six family members with high sequence homology (31, 32). The family member observed contains four introns and from EST data is known to be alternatively spliced due to a rare GAGGAG motif that creates adjacent AG acceptor sites at the intron 2/exon 3 boundary (Fig. 3e) (33). In most tissues, ~10% of this mRNA

![Fig. 3. Characterization and semiquantitation of alternative splice variants using top down MS/MS.](image)

![Fig. 4. Characterization of a previously unknown phosphoprotein using top down MS/MS.](image)
| No. | Protein accession no. | Fragmentation method | Protein size | Δmass | Pscore | Function | Gene family | cSNP | Alternative splice | PTM | Notes |
|-----|----------------------|----------------------|--------------|-------|--------|----------|-------------|------|-------------------|------|--------|
| 1   | P39019               | CAD                  | 15.9 kDa     | 0.3   | 1.0E-03| 40 S ribosomal protein S19 | X             | E44G |
| 2   | P56378               | CAD                  | 6.7 kDa      | 0.12  | 4.0E-07| 6.8-kDa mitochondrial proteolipid PRO1574 | X             | I9V  |
| 3   | P18124               | CAD                  | 29.2 kDa     | 0.20  | 1.0E-02| 60 S ribosomal protein L7 | X             | E47K |
| 4   | P23411               | ECD                  | 8.1 kDa      | 0.03  | 2.0E-05| 60 S ribosomal protein L38 | X             | K52E,F34L |
| 5   | O75964               | CAD                  | 11.3 kDa     | 0.20  | 1.0E-10| ATP synthase, e.g. chain, mitochondrial EC subunit G | X             | K47R,A66V |
| 6   | P56381               | ECD                  | 5.6 kDa      | 0.12  | 8.0E-26| ATP synthase e chain, mitochondrial | X             | E30K |
| 7   | O75531               | ECD                  | 10.2 kDa     | 80.05 | 2.0E-10| Barrier-to-autointegration factor | X             | Phosphorylated near N terminus, N-term Ac |
| 8   | P61769               | CAD                  | 11.8 kDa     | 0.0   | 3.0E-10| β₂-microglobulin precursor HDCMA22P | X             | X |
| 9   | P06703               | ECD                  | 10.2 kDa     | 0.2   | 3.0E-14| Calcyclin | X             | N69S,H27R,G90D,B33T |
| 10  | P31949               | ECD,CAD              | 11.6 kDa     | 0.1   | 1.0E-12| Calgizzarin S100C protein MLN | X             | E36K |
| 11  | P02593               | ECD                  | 16.8 kDa     | 0.1   | 3.0E-24| Calmodulin | X             | Trimethylated on Lys-115 |
| 12  | P15954               | CAD                  | 5.4 kDa      | 0.02  | 1.0E-07| Cytochrome c oxidase polypeptide VIIc, mitochondrial precursor; EC 1.9.3.1 | X             | Loss of N-term 16 residues to mature form |
| 13  | P63241               | CAD                  | 17.0 kDa     | 0.1   | 6.0E-17| Eukaryotic translation initiation factor | X             | X |
| 14  | P49773               | ECD                  | 13.7 kDa     | 0.0   | 1.0E-30| Histidine triad nucleotide-binding protein 1 (adenosine 5'-monophosphoramidase) (protein kinase C inhibitor 1) | X             | G105R,E100A |
| 15  | P10412               | ECD                  | 21.7 kDa     | 0.2   | 2.0E-17| Histone H1.4 member H1b | X             | X |
| 16  | P20670               | ECD                  | 14.0 kDa     | 0.0   | 2.0E-18| Histone H2A member O | X             | X |
| 17  | P02278               | IRMPD                | 13.8 kDa     | *     | 7.0E-14| Histone H2B | X             | X |
| 18  | P16106               | CAD                  | 15.5 kDa     | 0.1   | 1.0E-02| Histone H3 | X             | X |
| 19  | P02304               | IRMPD                | 11.3 kDa     | 0.1   | 9.0E-18| Histone H4 | X             | X |
| 20  | P14174               | ECD,CAD              | 12.3 kDa     | 0.0   | 1.0E-04| Macrophage migration-inhibitory factor, phenylpyruvate tautomerase, glycocaylation-inhibitory factor | X             | Disulfide; N106S,I68T,P44L |
| 21  | O95167               | CAD                  | 9.2 kDa      | 0.14  | 7.0E-07| NADH-ubiquinone oxidoreductase B9 subunit | X             | Consistent w/N-term Ac,L81V |
| 22  | O00483               | CAD                  | 9.4 kDa      | 1.00  | 1.0E-22| NADH-ubiquinone oxidoreductase MLRQ subunit EC | X             | D80G |
| 23  | P26447               | ECD                  | 11.9 kDa     | 0.1   | 2.0E-28| Placental calcium-binding protein, calvasculin S100, calcium-binding protein A4 MTS1 | X             | E9V |
| 24  | P06454               | ECD                  | 12.0 kDa     | 0.0   | 9.0E-15| Prothymosin α | X             | X |
| 25  | P06454               | IRMPD                | 12.1 kDa     | 0.0   | 1.0E-09| Prothymosin α | X             | X |
| 26  | Q99584               | IRMPD                | 11.4 kDa     | 0.1   | 7.0E-06| S100 calcium-binding protein A13 | X             | Consistent with N-term Ac at Ala-2, E14D |
| 27  | P25815               | ECD                  | 10.5 kDa     | 0.1   | 2.0E-14| S-100P protein | X             | E32D |
| 28  | P08578               | ECD                  | 10.7 kDa     | 0.5   | 5.0E-07| Small nuclear ribonucleoprotein E, snRNP-E, Sm protein E, Sm-E, SmC | X             | G5C |
| 29  | Q9UEA3               | CAD                  | 6.5 kDa      | 0.1   | 2.0E-21| Ubiquinol-cytochrome c reductase complex, 6.4-kDa protein | X             | L24W |
| 30  | P02248               | ECD                  | 8.4 kDa      | 0.00  | 1.0E-04| Ubiquitin | X             | Loss of C-term GG to mature form |
| 31  | Q15834               | IRMPD                | 8.6 kDa      | 0.10  | 5.0E-15| Ubiquitin-like protein NEDD8 | X             | Loss of C-term GGLRQ to mature form |

* Mass errors are from comparison of intact mass values; fragment ion mass errors are typically 1–20 ppm.

*a* E-x represents $\times 10^{-x}$.

All cSNPs were verified using dbSNP; the major allele (in boldface type) was observed for cases where allelic frequencies were provided.

N-term, N-terminal; Ac, acetylation; C-term, C-terminal.

Multiple isobaric forms observed with identical Pscore values.

---

**Characterization of SNPs, PTMs, and Splice Variants**

---

**TABLE I**

Partial list of human proteins identified and characterized using top down MS

- **No.**
- **Protein accession no.**
- **Fragmentation method**
- **Protein size**
- **Δmass**
- **Pscore**
- **Function**
- **Gene family**
- **cSNP**
- **Alternative splice**
- **PTM**
- **Notes**
contains an extra GAG codon (encodes for an extra Glu) versus 90% of ProTα transcripts where the more 5’ acceptor site is used, producing a form with one less residue (33). Upon examination of the broadband spectrum, both species were observed in a ~10:1 ratio of light versus heavy protein (Fig. 3a). The minor species was subsequently fragmented (Fig. 3c), and the extra Glu residue was precisely localized (Fig. 3d, right).

The presence of the GAGAG motif was recognized as a possible acceptor site by only NetGene2 (www.cbs.dtu.dk/services/NetGene2), one of five intron/exon prediction programs tested. Using BLAST to search human EST libraries (www.ncbi.nlm.nih.gov/dbEST), more than 1300 dbEST entries were attributed to ProTα with only ~150 matching the longer form, consistent with an earlier finding that the ~9:1 ratio of short:long is not tissue-specific (33). Also using BLAST, the GAGAG motif at this locus was found only in primates. Neither rat nor mouse have the extra splice acceptor site and have evolved only the long form of the protein, which is actually the less favorable form in humans.

Identification of a Novel Phosphoprotein—As a last illustration of new advantages provided by the top down MS approach, the 10,191.1-Da BAF protein was identified in a nuclear extract and exhibited a +79.95 ± 0.05-Da satellite peak at ~10% occupancy consistent with phosphorylation (Fig. 4a). The data from automated MS/MS localized the phosphorylation to the 11 N-terminal residues. Manual MS/MS using electrons further confirmed a Met off/acetylated N terminus and narrowed the region of phosphorylation to Thr-2 or Ser-3 (Fig. 4b). This well studied protein directly binds to chromatin, is thought to be involved in attachment of chromatin to the inner nuclear membrane (34), and is known to be modified. No other forms of this protein have been observed in adjacent fractions, and the pI change caused by phosphorylation is small enough to allow coelution of both forms in identical fractions during chromatofocusing and RPLC. With the two-dimensional fractionation behavior of this modified protein now known, detection of this protein from nuclear extracts was reproduced twice more. This now allows targeted studies on this protein from synchronized HeLa cells in a straightforward manner. Such a platform for biochemical interrogation of targeted proteins after RNA interference, chemical perturbation, or cell synchronization will be highly valuable for capturing a more detailed picture of functional regulation mechanisms involving PTM dynamics.

Summary of Findings and Outlook—Using a dual ion fragmentation approach to automatically analyze two to three small human proteins per fraction by top down MS/MS, 45 proteins were identified with a median probability score of 10^-13 (Table I). A main advantage of the top down strategy is that information on the entire primary structure of the mature protein is obtained, allowing reliable dissection and abundance measurements of highly related gene products from genetic or transcriptional variation and enzymatic modification. Due to the complimentary nature of ion fragmentation using electrons and collisions with gas, precise localization of PTMs, polymorphisms, and amino acids at splice junctions is indeed possible. For the identified proteins, 45% were found in forms not present in UniProt’s Human Proteome Initiative, and ~40% contained SNPs for which only single alleles were observed. Over 85% of the identifications required no manual validation of the data base retrieval result: Δm localization sometimes improved upon inspection of the raw data. Characterization of closely related protein forms (e.g. different PTM isomers or SNP forms) sometimes required manual scrutiny of the output from ProSight PTM with the correct form yielding the highest score in the retrieval list in ~90% of cases.

The ability to automatically genotype cSNPs and characterize PTMs with gene-specific identifications is enabled by the new informatic strategy of shotgun annotation (15), the combinatorial consideration of diverse sources of Δm values. This strategy represents a major shift in curation philosophy for protein data bases (35), is well suited for a top down approach using FTMS, and recognizes that detailed information on SNPs, mutations (36), splice variants (37), and PTMs (38) will be increasingly known and even somewhat predictable (36). By embedding such variability tightly within a MS retrieval engine, the current study drastically improves identification metrics, enables known biological events to be characterized as they occur in combination, and allows unknown biology to be uncovered more efficiently. Shotgun annotation actually increases the quality of most retrievals by allowing more absolute mass values of fragment ions observed in a top down MS/MS experiment to match those values generated from protein forms housed in a data base. The examples highlighted here illustrate an overall process that can simply be called “proteotyping.” The term proteotyping is akin to genotyping at the DNA level but captures all the variability of proteins as they occur in populations and change over time. Fragmentation of intact proteins represents an emergent method for “reverse annotation” of the human genome, and top down MS can now be embraced by organizations such as the Human Proteome Organization.

Acknowledgments—We thank Rich LeDuc for technical assistance and Hugh Robertson for valuable discussions. We also are grateful to John Hobbs and Jeff Chapman of Beckman Coulter and Tim Barder of Eprogen for assistance with the PF 2D system.

* This work was supported by National Science Foundation Career Award CH 0134953, National Institutes of Health Grant GM 067193, the Sloan Foundation, the University of Illinois Urbana-Champaign Center of Neuroproteomics (p30_DAO18310), the Research Corporation (Cottrell Scholars Program), and the Henry and Lucille Packard Foundation. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

‡ To whom correspondence should be addressed: Dept. of Chem-

Molecular & Cellular Proteomics 4.7 1007
Characterization of SNPs, PTMs, and Splice Variants

istry, University of Illinois Urbana-Champaign, 39 RAL, 600 S. Matthews, Urbana, IL 61801. E-mail: Kelleher@scs.uiuc.edu.

REFERENCES

1. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. Nature 422, 198–207
2. Yates, J. R. (2004) Mass spectral analysis in proteomics. Annu. Rev. Bio-
phys. Biomol. Struct. 33, 297–316
3. Sam-Yellowe, T. Y., Florens, L., Johnson, J. R., Wang, T., Drazba, J. A., Le Roch, K. G., Zhou, Y., Batalov, S., Carucci, D. J., Winzeler, E. A., and Yates, J. R., III (2004) A Plasmodium gene family encoding Maurer’s cleft membrane proteins: structural properties and expression profiling. Ge-
ome Res. 14, 1052–1059
4. Rappisilber, J., Ryder, U., Lamond, A. I., and Mann, M. (2002) Large-scale proteomic analysis of the human spliceosome. Genome Res. 12, 1231–1245
5. Plines, W. M., Parker, K., Peltert, J., Patterson, D. H., Vestal, M. L., and Martin, S. A. (1998) Protein identification and protein characterization by high-performance time-of-flight mass spectrometry. J. Protein Chem. 17, 525–526
6. Haynes, P. A., and Yates, J. R., III (2000) Proteome profiling-pitfalls and progress. Yeast 17, 81–87
7. Gygi, S. P., Rist, B., Griffin, T. J., Eng, J., and Aebersold, R. (2002) Proteome analysis of low-abundance proteins using multidimensional nano
ion-trap and isotope-coded affinity tags. J. Proteome Res. 1, 47–54
8. Thiede, B., Lamer, S., Mattow, J., Siejak, F., Dimmler, C., Rudel, T., and Jungblut, P. R. (2000) Analysis of missed cleavage sites, tryptic oxidation and N-terminal pyroglutamylation after in-gel tryptic digestion. Rapid Commun. Mass Spectrom. 14, 496–502
9. Konig, S., Zeller, M., Peter-Katalinic, J., Roth, J., Sorg, C., and Vogt, T. (2001) Use of nonspecific cleavage products for protein sequence anal-
yzis as shown on calcyclin isolated from human granulocytes. J. Am. Soc. Mass Spectrom. 12, 1180–1185
10. Field, H. I., Fenyo, D., and Beavis, R. C. (2002) RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in a relational database. Proteomics 2, 36–47
11. MacCoss, M. J., McDonald, W. H., Saraf, A., Sadygov, R., Clark, J. M., Tasto, J. J., Gould, K. L., Wolters, D., Washburn, M., Weiss, A., Clark, J. I., and Yates, J. R., III (2002) Shotgun identification of protein modifi-
cations from protein complexes and lens tissue. Proc. Natl. Acad. Sci. U. S. A. 99, 7900–7905
12. Wu, C. C., MacCoss, M. J., Mardones, G., Finnegan, C., Mogelsvang, S., Yates, J. R., III, and Howell, K. E. (2004) Organelle proteomics reveals Golgi arginine dimethylation. Mol. Biol. Cell 15, 2907–2917
13. Zhu, K., Kim, J., Yoo, C., Miller, F. R., and Lubman, D. M. (2003) High sequence coverage of proteins isolated from liquid separations of breast cancer cells using capillary electrophoresis-time-of-flight MS and MALDI-TOF MS mapping. Anal. Chem. 75, 6209–6217
14. Gatlin, C. L., Eng, J. K., Cross, S. T., Detter, J. C., and Yates, J. R., III (2002) Shotgun identification of protein modifications from intact proteins and lens tissue. Proc. Natl. Acad. Sci. U. S. A. 99, 7900–7905
15. Pesavento, J. J., Kim, Y. B., Taylor, G. K., and Kelleher, N. L. (2004) Targeted analysis and discovery of posttranslational modifications in proteins of Arabidopsis thaliana using quadrupole FT hybrid mass spectrometry for top down proteomics. Anal. Chem. 76, 2852–2858
16. Cargile, B. J., McLuckey, S. A., and Stephenson, J. L., Jr. (2001) Identification of bacteriophage MS2 coat protein from E. coli lysates via ion trap cationic activation of intact protein ions. Anal. Chem. 73, 1277–1285
17. Meng, F., Cargile, B. J., Patric, S. M., Johnson, J. R., and Kelleher, N. L. (2002) Processing complex mixtures of intact proteins for direct analysis by mass spectrometry. Anal. Chem. 74, 2923–2929
18. Forbes, A. J., Mazur, M. T., Patel, H. M., Walsh, C. T., and Kelleher, N. L. (2001) Toward efficient analysis of >70 kDa proteins with 100% se-
quence coverage. Proteomics 1, 927–933
19. Zabrouskov, V., Giacomelli, F., Van Wijl, and McLafferty, F. W. (2003) A new approach for plant proteomics: characterization of chloroplast proteins of Arabidopsis thaliana by top-down mass spectrometry. Mol. Cell. Proteomics 2, 1253–1260
20. Allis, C. D., Glover, C. V., and Gorovsky, M. A. (1979) Micronuclei of Tetrahymena contain two types of histone H3. Proc. Natl. Acad. Sci. U. S. A. 76, 4857–4861
21. O’Donovan, C., Apweiler, R., and Bairoch, A. (2001) The human proteome project (HPP). Trends Biotechnol. 19, 178–181
22. Horvath, M. M., Fondon, C. L., Pasa-Tolic, L., Marto, J. A., White, F. M., Guan, S., and Marshall, A. G. (1996) Electrospray ionization Fourier transform ion cyclotron resonance at 9.4 T. Rapid Commun. Mass Spectrom. 10, 1824–1828
23. Taylor, G. K., Kim, Y. B., Forbes, A. J., Meng, F., McCarthy, R., and Kelleher, N. L. (2003) Web and database software for identification of intact proteins using “top down” mass spectrometry. Anal. Chem. 75, 4081–4086
24. Telmer, C. A., Retchless, A. C., Kinsey, A. D., Conley, Y., Rigatti, B., Gorin, M. B., and Jarvik, J. W. (2003) Detection and assignment of mutations and minihaplotypes in human DNA using peptide mass signature geno-
yping (PMSG): application to the human RDS/peripherin gene. Genome Res. 13, 1944–1951
25. Pineiro, A., Cordero, O. J., and Nogueira, M. (2000) Fifteen years of pro-
thyosmin α: contradictory past and new horizons. Peptides 21, 1433–1446
26. Eschenfeldt, W. H., Manrow, R. E., Krug, M. S., and Berger, S. L. (1989) Isolation and partial sequencing of the human prothymosin α gene family. Evidence against export of the gene products. J. Biol. Chem. 264, 7546–7555
27. Manrow, R. E., and Berger, S. L. (1993) GAG triplets as splice acceptors of last resort. An unusual form of alternative splicing in prothymosin α pre-mRNA. J. Biol. Chem. 268, 234–281
28. Segura-Totten, M., Kowalski, A., and Wilson, K. (2002) Barrier-
to-autointegration factor: major roles in chromatin decondensation and nuclear assembly. J. Cell Biol. 158, 475–485
29. Mann, M., and Jensen, O. N. (2003) Proteomic analysis of post-translational modifications. Nat. Biotechnol. 21, 255–261
30. Horvath, M. M., Fondon, J. W., III, and Garner, H. R. (2003) Low hanging fruit: a subset of human cSNPs is both highly non-uniform and predict-
able. Gene (Amst.) 312, 197–206
31. Davis, K. E., Atanelov, L., Modrek, B., and Xing, Y. (2003) ASAP: the alternative splicing annotation project. Nucleic Acids Res. 31, 101–105
32. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O’Donovan, C., Phan, L., Pit-
bout, S., and Schneider, M. (2003) The SWISS-PROT protein knowledge-
base and its supplement TrEMBL in 2003. Nucleic Acids Res. 31, 365–370