The Implications of Proteolytic Background for Shotgun Proteomics

Originally published in:
Molecular & Cellular Proteomics 6(9), https://doi.org/10.1074/mcp.M700029-MCP200

The Implications of Proteolytic Background for Shotgun Proteomics*[S]

Paola Picotti‡, Ruedi Aebersold‡§¶, and Bruno Domon‡‖

The analysis by liquid chromatography coupled to tandem mass spectrometry of complex peptide mixtures, generated by proteolysis of protein samples, is the main proteomics method used today. The approach is based on the assumption that each protein present in a sample reproducibly and predictably generates a relatively small number of peptides that can be identified by mass spectrometry. In this study this assumption was examined by a targeted peptide sequencing strategy using inclusion lists to trigger peptide fragmentation attempts. It was found that the number of peptides observed from a single protein is at least one order of magnitude greater than previously assumed. This unexpected complexity of proteomics samples implies substantial technical challenges, explains some perplexing results in the proteomics literature, and prompts the need for developing alternative experimental strategies for the rapid and comprehensive analysis of proteomes. ***Molecular & Cellular Proteomics 6:1589–1598, 2007.***

Proteomics aims at analyzing in one single experiment the complete proteome of a cell type, a tissue, or a species. The studies in which the proteins constituting a proteome are identified and quantified are particularly informative and relevant from a biological and biomedical point of view. At present this is most frequently attempted by a shotgun, tandem mass spectrometry-based strategy (1). Although various implementations of this strategy differ in the specifics, they all share the same operating steps (2): the proteins present in a sample are initially digested in solution with a highly specific protease, typically trypsin. The resulting peptide mixtures are subjected to one-, two-, or three-dimensional fractionation, and the peptides eluting from the last separation step, typically reverse-phase chromatography, are analyzed by MS/MS. Lastly the MS/MS spectra collected are assigned to peptide sequences using a suite of software tools (3, 4).

Shotgun proteomics is founded on the assumption that each protein present in a sample reproducibly generates a relatively small number of peptides, the boundaries of which conform to the cleavage specificity of the protease used. Trypsin cleaves at the C termini of arginine and lysine residues, and based on the occurrence of these two amino acids in proteins, an average of 10 peptides is expected for a stretch of a hundred residues. The validity of this assumption is critical to the success of proteomics experiments for several reasons. First, the number of peptides present in a protein digest determines the number of MS/MS cycles minimally required to fully analyze the sample. Because the MS-MS/MS duty cycle is given for a particular mass spectrometer (better than 1 Hz for modern instruments), in theory the number of peptides to be analyzed in the sample relates to the minimal duration of a proteomics experiment. Second, the sample complexity determines the practical dynamic range of proteome analyses. The nominal dynamic range of a mass spectrometer (at best 3–4 orders of magnitude) can in practice be significantly reduced by the fact that automated precursor ion selection (data-dependent acquisition (DDA)[1] primarily focuses on the most intense MS signals. In the case of very complex samples, only the highest intensity ions are fragmented, whereas ions of lower intensity pass through the system unselected even though their signal is well within the nominal dynamic range of the instrument. Third, database searches are often constrained to full tryptic peptides to restrict the searching time. Unspecific cleavages will thus produce peptides not anticipated by the search parameters, leading to misassignments or missed identifications.

Despite the development of peptide separation systems with higher peak capacity (5–7) and of tandem mass spectrometers with faster acquisition rate, the proteome of any species has yet to be fully mapped. The complete analysis of even moderately complex samples such as isolated organelles (8) or macromolecular complexes (9) has required enormous efforts. All these considerations suggest that the comprehensive analysis of a complex sample is more difficult than initially anticipated, and one of the reasons for this could be a degree of complexity resulting from proteolysis of the protein sample that is higher than expected.

To test this hypothesis and to assess the number of peptides actually generated by proteolysis of a protein, an in-depth characterization of the products of tryptic digestion of well defined proteins was carried out. Five pure bovine standard proteins, β-lactoglobulin, carbonic anhydrase, serum albumin, transferrin, and β-casein, were subjected individually

From the ‡Institute of Molecular Systems Biology, ETH Zürich, CH-8093 Zürich, Switzerland, §Faculty of Sciences, University of Zürich, CH-8006 Zürich, Switzerland, and ¶Institute for Molecular Systems Biology, Seattle, Washington 98103

Received, January 24, 2007, and in revised form, May 4, 2007

Published, MCP Papers in Press, May 28, 2007, DOI 10.1074/mcp.M700029-MCP200

[1] The abbreviation used is: DDA, data-dependent acquisition.

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

to tryptic digestion, and the resulting peptide mixtures were extensively characterized by LC/MS/MS. To maximize the number of peptides identified in the proteolytic digests, a targeted MS/MS sequencing strategy (10–12) was applied and was compared with the intensity-driven data-dependent acquisition method. The targeted approach is based on inclusion lists of precursor ions to trigger collision-induced dissociation. In this approach, the samples were initially analyzed in a high accuracy mass spectrometer in full scan mode. Data were then extensively processed off-line to extract and inventory monoisotopic ions of all the peptide ions observed. The samples were then subjected to MS/MS sequencing multiple times using inclusion lists to trigger fragmentation of the ions of interest, present in full MS scan regardless of their intensity, with retention time and charge state as the only constraints. The approach was shown to be a robust and effective means to sequence low abundance ions and revealed a number of peptides produced from proteolysis of a protein that is at least 10 times higher then previously assumed.

MATERIALS AND METHODS

Chemicals—Porcine trypsin (modified, sequencing grade) was purchased from Promega (Madison, WI). The standard bovine proteins β-lactoglobulin, β-casein, serum albumin, transferrin, and carbonic anhydrase and DTT were obtained from Sigma. Recombinant His-tagged human Pax-8 protein was overexpressed in *Escherichia coli* (BL21–3DE, Stratagene, Heidelberg, Germany) and purified by means of a HiTrap chelating nickel column (GE Healthcare). Tris(2-carboxyethyl)phosphine and iodoacetamide were purchased from Fluka (Buchs, Switzerland). HPLC-grade water and acetonitrile were purchased from Riedel-de Haën (Seelze, Germany).

Protein Digestion—Each protein was solubilized in 0.1 M ammonium bicarbonate buffer containing 8 M urea at a final concentration of 3 mg/ml. After tris(2-carboxyethyl)phosphine reduction and iodoacetamide alkylation of cysteine residues, the solution was diluted to final 1 M urea, the pH was adjusted to 8.0, and the proteins were digested with trypsin at 37 °C. The digestion was repeated on the same set of proteins, after HPLC purification, to eliminate protein degradation products and other contaminants potentially present in the commercial protein preparations. Briefly proteins were loaded onto a macroporous reverse-phase C_{18} column (mRP-C18, 4.6 × 50 mm, Agilent Technologies, Waldbronn, Germany). Elution was carried out with a linear gradient of water/acetonitrile containing 0.1 and 0.08% (v/v) TFA, respectively, from 5 to 65% in 60 min at a flow rate of 0.75 ml/min. The eluate was monitored by absorption measurements at 226 and 280 nm. Fractions containing the protein species were collected at the peak apex, lyophilized, and then subjected to trypsin digestion with the protocol described previously. A range of different enzyme to substrate ratios (1:10 to 1:500) and different incubation times (from 2 to 24 h) were used for each protein digestion. In the case of β-lactoglobulin, digestion was also conducted with trypsin immobilized on agarose beads (Pierce) with recombinant trypsin from *Pichia pastoris* and with the enzyme Lys-C from *Lysobacter enzymogenes* (Roche Diagnostics). The digestion was stopped by acidification with formic acid to a final pH of 4.0. The peptide mixtures were cleaned by OASIS HLB or Sep-Pak tC18 cartridges (Waters, Milford, MA) eluted with 80% acetonitrile. Alternatively a protocol based on ammonium bicarbonate buffer exchange by gel filtration desalting columns (cutoff, 7000 Da; Pierce) prior to trypsin addition was used. Cleaned peptide samples were evaporated on a vacuum centrifuge to dryness, resolubilized in 0.1% formic acid, and immediately analyzed. Control samples were also prepared, resulting from the same protocol steps, without addition of the protein substrate.

Mass Spectrometry Analysis—Samples were analyzed on a hybrid LTQ-FT mass spectrometer (Thermo Electron, San Jose, CA) equipped with a nanoelectrospray ion source. Chromatographic separations of peptides were performed on an Agilent 1100 micro-HPLC system (Waldbronn, Germany) equipped with a 10-cm fused silica emitter (150- μm diameter) packed with a Magic C_{18} AQ 5-μm resin (Michrom BioResources, Auburn, CA). Peptides were loaded on the column from a cooled (4 °C) Agilent autosampler and separated with a linear gradient of acetonitrile/water containing 0.1% formic acid at a flow rate of 1.2 μl/min. A gradient from 5 to 50% acetonitrile in 50 min was used. For each peptide sample a standard DDA on the three most intense ions per MS scan was first performed. Three MS/MS spectra were acquired in the linear ion trap per each FT-MS scan, the latter acquired at 100,000 full-width half-maximum nominal resolution, resulting in an overall cycle time of ~1 s. Charge state screening was used, allowing fragmentation of singly and multiply charged ions and rejecting ions of unknown charge state. A threshold of 200 ion counts was set to trigger an MS/MS attempt. A manual extraction of all the *m/z* features (monoisotopic peaks, ^{12}C) was performed on the first FT-MS file resulting from the analysis of single protein tryptic digests down to a level of intensity of signal-to-noise ratio of 5.

The extracted ions were incorporated into an inclusion list for subsequent analysis. The features were sorted by intensity and divided into multiple lists containing up to 150 ions each. Each digest was reanalyzed under targeted, inclusion list-driven selection of precursor ions for MS/MS analysis, a number of times corresponding to the number of sublists generated. Previously derived attributes of peptide ions, such as charge state and retention time, were used as constraints for triggering MS/MS attempts. For low intensity signals (ion counts <1000) the acquisition time was increased to improve the quality of MS/MS spectra (up to five averaged MS/MS microscans; trapping time, up to 500 ms).

Data Analysis—Raw MS/MS data were searched against the bovine National Center for Biotechnology Information (NCBI) non-redundant database using BioWorks 3.2 (Sequest, Thermo Electron). Human keratins and porcine trypsin were added to the bovine protein database. Precursor ion tolerance was set to 0.55, and fragment tolerance was set to 0.5 Da. MS/MS data acquired on the tryptic digest of human Pax-8 were searched against the human NCBI non-redundant database. Data were searched allowing oxidation of methionine as a variable modification with carboxyamidomethylation of cysteine residues as a static modification. In the standard searches, two missed cleavages and one non-tryptic terminus were allowed. Data were also searched with no enzyme constraints against the sequence of the protein of interest. Each peptide assignment (BioWorks Peptide Probability score ≤1) was subjected to manual validation, and only high quality matches were accepted. Additionally data were searched for urea-induced carbamylation as a variable modification, at either the N terminus or Lys or Arg residues, with carboxyamidomethylation of cysteine residues as a static modification, allowing in this case only full tryptic cleavage and two missed cleavages. For each peptide assignment, the difference in ppm between the experimental and the theoretical monoisotopic values was calculated. The high accuracy of FT measurements was used as an additional and independent data filtering criterion.

RESULTS

To assess the number of peptides generated by proteolysis of a protein, we undertook an in-depth characterization of the actual products of tryptic digestion of well characterized proteins. Five pure bovine standard proteins, β-lactoglobulin,

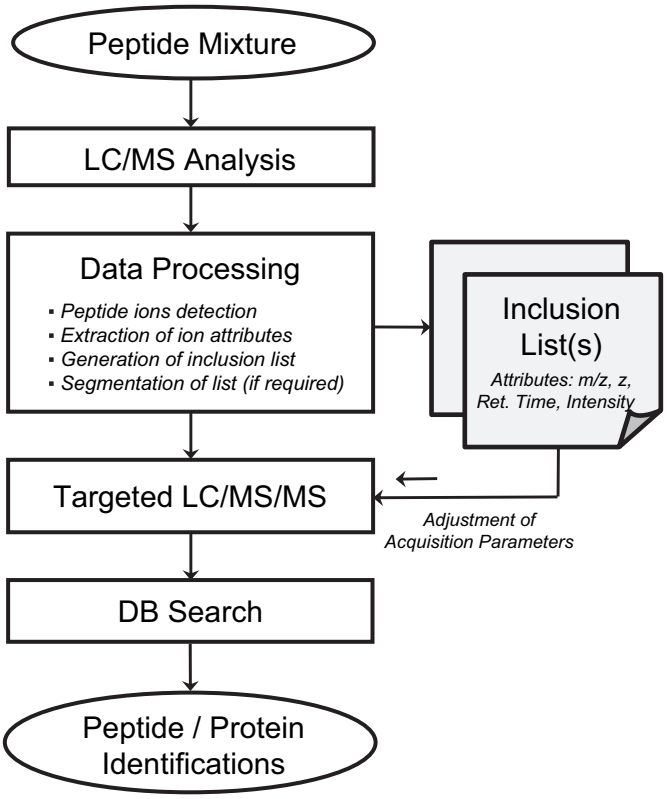

FIG. 1. **Work flow used for targeted MS/MS peptide sequencing.** *DB*, database.

carbonic anhydrase, serum albumin, transferrin, and β-casein (purity confirmed by SDS-PAGE) were subjected individually to tryptic digestion, and the resulting peptide mixtures were extensively characterized by LC/MS/MS. To exclude a significant bias of the composition of the proteolysis product induced by specific experimental conditions, the analyses were repeated using protocols in which critical parameters such as the enzyme to substrate ratio, incubation times, the type of commercial enzyme, digestion, and peptide cleanup protocols were varied. Digestion was also repeated on the same set of proteins subjected to an additional purification step, by either reverse-phase HPLC or gel filtration (see the supplemental materials), to eliminate possible low abundance degradation products potentially present in the commercial protein preparations. Tryptic digestion was also conducted on the human recombinant Pax-8 protein expressed in *E. coli* to exclude the presence of multiple isoforms in the starting material.

To maximize the number of peptides identified in the proteolytic digests, a targeted MS/MS sequencing strategy was applied and compared with the intensity-dependent DDA approach. The targeted approach is based on an inclusion list of precursor ions that are selected within a specified chromatographic elution window and trigger a collision-induced dissociation. The method is schematically illustrated in Fig. 1. Initially the sample of interest is analyzed in LC/MS mode (full scan) in a high mass accuracy mass spectrometer. Ideally the LC/MS analysis is replicated to determine the ion features reproducibly associated to the sample. The raw data are processed and analyzed in depth off line to extract all observed monoisotopic ions related to peptides with the pertaining attributes (*m/z*, charge state, retention time, and intensity). The sample is subsequently analyzed in MS/MS mode (several times if necessary), using inclusion lists containing the ions of interest to trigger collision-induced dissociation, with retention time and charge state as constraints. Sequencing parameters (MS/MS sampling time and triggering threshold) are adjusted according to the attributes of the features included. Number and length of inclusion lists are chosen based on the acquisition software inclusion capacity and on the instrument sequencing rate. Fragment ion spectra acquired by targeted sequencing analyses are then assigned to peptide sequences by database searching, and the assignments are pooled.

Because most of the scoring schemes of database search algorithms tend to favor longer peptide sequences with genuine tryptic ends, atypical tryptic digestion products might be underscored in the automatic database search process (see the supplemental material for a more detailed discussion). Consequently in this work, the conventional filtering of scores was not used to discriminate between good and lower quality assignments. Each peptide assignment (with BioWorks Peptide Probability $\leq$1) was subjected to manual inspection, and only high quality matches were accepted. In this way numerous good quality MS/MS spectra, with a valid sequence assignment, that would not have been retained using conventional filtering criteria were added to the dataset. Assignments were further corroborated by a difference in the precursor ion mass lower than 3.5 ppm using the accuracy of FT mass measurements as an independent filtering criterion. Examples of peptide assignments included in the dataset with a low score are illustrated in Fig. 2.

Fig. 3 shows the results of this type of analysis for a bovine β-lactoglobulin tryptic digest. Lactoglobulin is a small protein, composed of 162 amino acids, with a molecular weight of 18,281. *In silico* digestion yields 12 fully tryptic peptides containing more than five amino acids. Overall 32 peptides were identified using the conventional data-dependent analysis (Supplemental Table 1) with automated selection of highly intense precursor ions for fragmentation. Strikingly the targeted peptide sequencing approach yielded 117 distinct peptide identifications (Supplemental Table 1), 85 of which had not been observed by using the automated method. Among the peptides observed, 93 were derived from the dominant form of bovine β-lactoglobulin, whereas 19 sequences corresponded to allelic variants of β-lactoglobulin (G64D, A118V, and F105V). In addition five peptides were trypsin autolysis products or fragments of keratins. Most of the predicted fully tryptic peptides of β-lactoglobulin (11 of 12) were identified by both DDA and targeted methods. The latter approach showed

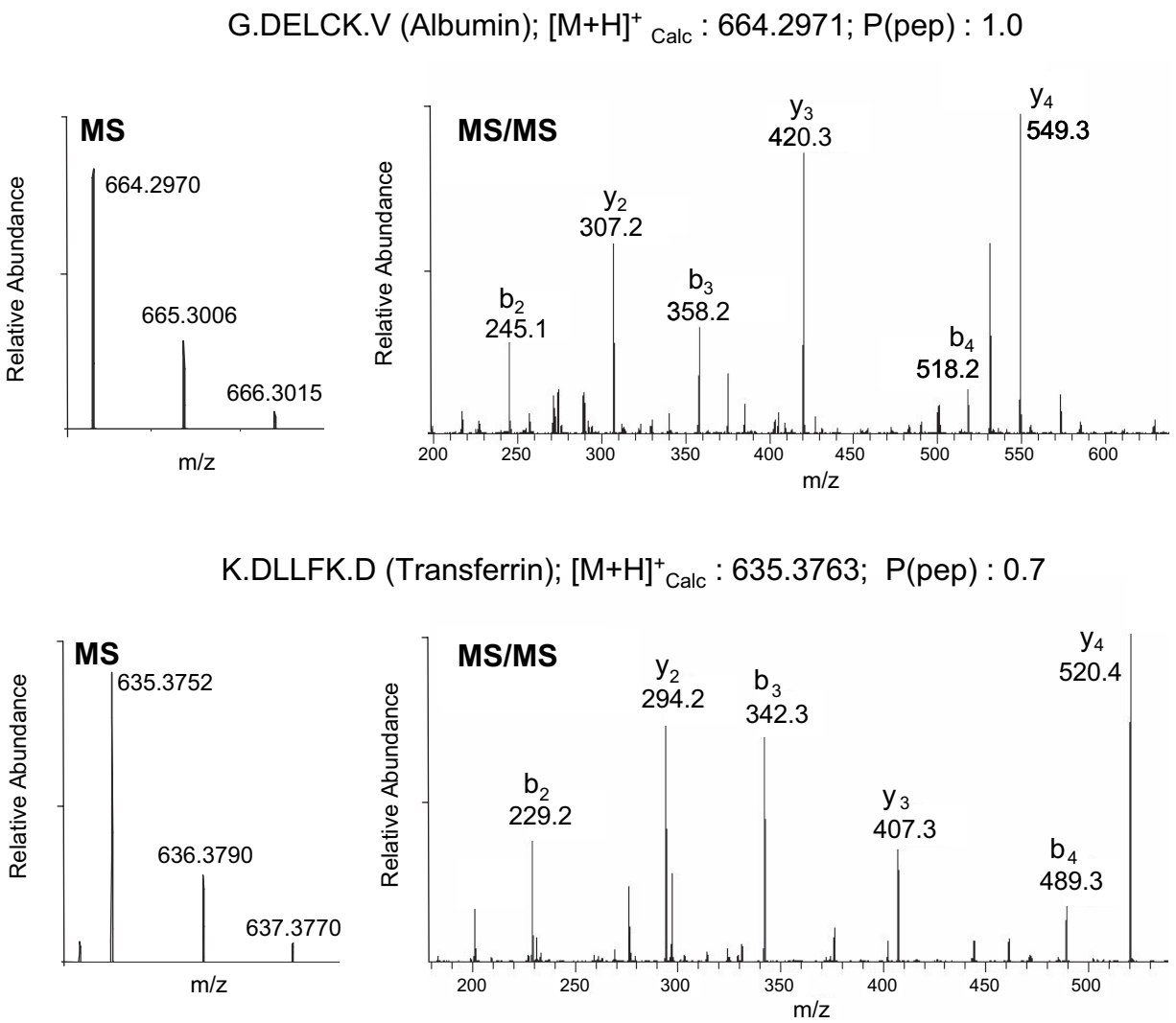

FIG. 2. **Selected examples of peptide identification.** Peptides are shown with good MS/MS sequence coverage poorly scored because of their short length. *$[M + H]^{+}_{Calc}$*, calculated $[M + H]^{+}$ mass of the peptide; *P(pep)*, BioWorks probability score.

an unprecedented capacity of identifying low abundance digestion products, among them peptides containing missed cleavages, modifications, amino acid substitutions, and a striking number of sequences (89 β-lactoglobulin-derived peptides) containing one non-tryptic terminus. Fig. 3 shows the clear prevalence of partly tryptic sequences with the non-tryptic cleavage at the N terminus. This is related to the fact that peptides retaining a highly basic Arg/Lys residue at the C terminus have high ionization efficiency and were therefore more easily detected. Of note is the presence of truncated peptides series with distinct elution times (Fig. 4). This observation provides strong evidence that the various truncated forms of a peptide were actually present in the sample prior to chromatographic separation and were not artifacts produced during ionization due to in-source fragmentation.

The color scale used in Fig. 3 indicates semiquantitatively abundances of the set of peptides from β-lactoglobulin observed in LC/MS/MS analyses. Although large in number (more than 75% of peptide identifications; Fig. 5), most of the partly tryptic peptides corresponded to low abundance signals. Collectively they contribute to less than 20% of the overall ion current (Fig. 5; calculation encompassing only the peptides formally identified). The proportion of partly and fully tryptic peptides identified remained approximately constant over the set of protein digests studied (see the supplemental material). Moreover this distribution was not significantly affected by the digestion conditions (see the supplemental materials); only a small increase in the occurrence of partly tryptic peptides was observed at higher enzyme to substrate ratios (1:10 compared with 1:500). Furthermore the overall number of peptides identified per protein increased with the molecular mass of the proteins. No significant variation in the proteolysis product composition was observed when digestion was performed on proteins purified by reverse-phase HPLC or gel

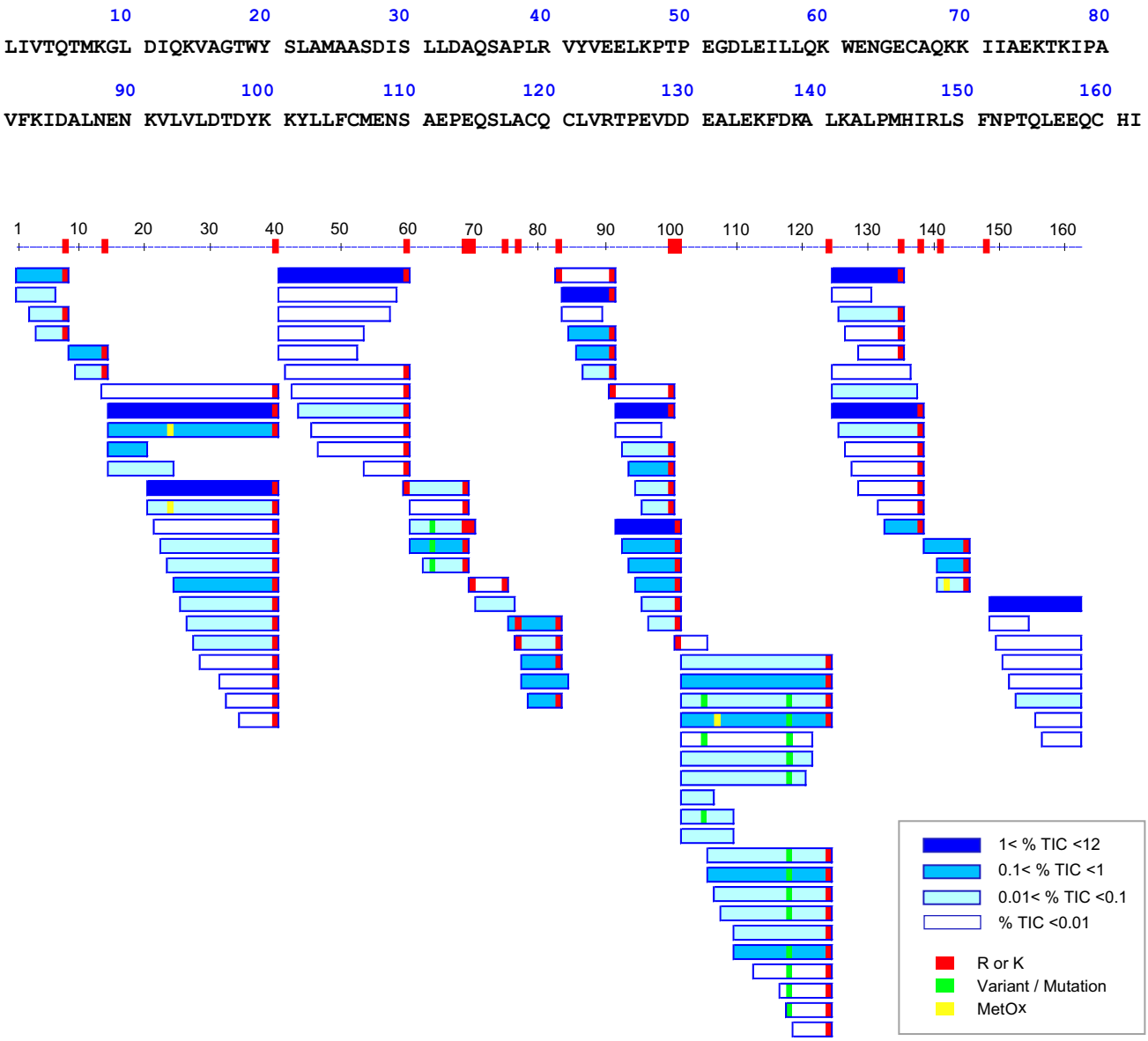

FIG. 3. **Map of peptides identified in the tryptic digest of β-lactoglobulin using the targeted MS/MS approach.** The *blue* color scale reflects the relative peptide abundance (measured as fraction of the total ion current (*TIC*) of the peptides identified). *MetOx*, oxidized Met.

filtration as well as on the recombinant protein Pax-8 (Supplemental Table 3). Similar results, in terms of occurrence of unspecific cleavages, were also obtained when Lys-C was used as an enzyme.

DISCUSSION

This work provides the first in-depth characterization of the actual components of tryptic digests in terms of expected products and digestion by-products. The application of a targeted peptide sequencing strategy allowed a drastic increase of the number of peptide identifications compared with conventional DDA analysis. In particular it promoted sequencing of low abundance peptides, which had a low probability of being selected for collision-induced dissociation in an automated mode. This revealed an unexpected complexity of tryptic digests mostly given by a substantial number of peptides with a non-tryptic terminus.

The stringency of trypsin specificity was recently emphasized by Mann and co-workers (13) in an analysis of a mouse liver proteome fraction. The study provides a good representation of products normally identified in standard proteomics experiments on complex digests. However, the sample used in that study was complex, and it was analyzed by automated selection of precursor ions, thus selecting and fragmenting predominantly the high intensity precursors. Therefore, the claim of strict specificity of trypsin for Lys and Arg residues is consistent with the data in this study in as much as the high intensity precursor ions are concerned. The present work

demonstrates that the low intensity precursor ions that very likely were not analyzed in the prior study (13) are enriched for peptides with partially tryptic or non-tryptic ends. Furthermore a lack of trypsin specificity has been discussed previously and widely documented in the LYSIS (14, 15) database created by collecting literature data on proteolytic cleavages (see the supplemental materials for a discussion on the underlying reasons for trypsin unspecificity).

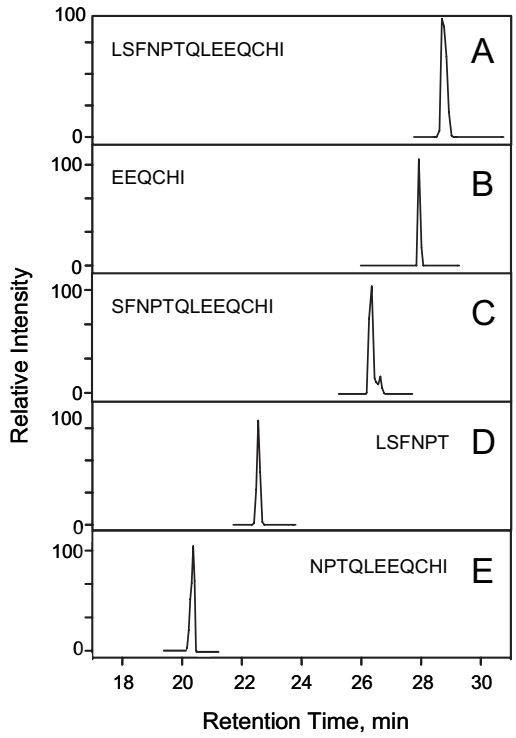

FIG. 4. **Extracted ion chromatograms of a subset of observed peptides.** Panels show differently truncated forms (*B–E*) of the same tryptic (*A*) peptide.

Partly tryptic peptides have occasionally been reported in the proteomics literature to occur as low abundance by-products of tryptic digestion. In addition, recent comprehensive studies have highlighted the presence of numerous unexpected species in tryptic digests, including partly tryptic, modified, and short peptides (16). In many cases, full trypsin specificity was chosen as a database search criterion, precluding the identification of partly tryptic peptides. However, the possibility of identifying partly tryptic peptides, even when specified as search parameter, might be affected by database search scoring or filtering algorithms, many of which tend to favor full tryptic peptides, attributing them a higher probability (see the supplemental materials for a detailed discussion on data analysis issues). For these reasons peptides with partially tryptic or non-tryptic ends are substantially underrepresented in the current proteomics literature.

In this study, the number of peptides identified in the β-lactoglobulin digest (117 peptides) using a targeted sequencing method was almost one order of magnitude greater than predicted (12 peptides). The actual complexity of such digests might be even greater if one considers that the peptides identified were only those present in sufficient amount to be detected by the mass spectrometer (*i.e.* ionizable and within the dynamic range of the instrument). Furthermore additional peptides could potentially be found in the digests by searching MS/MS data against single nucleotide polymorphism databases or by using *de novo* sequencing tools (17). Taking

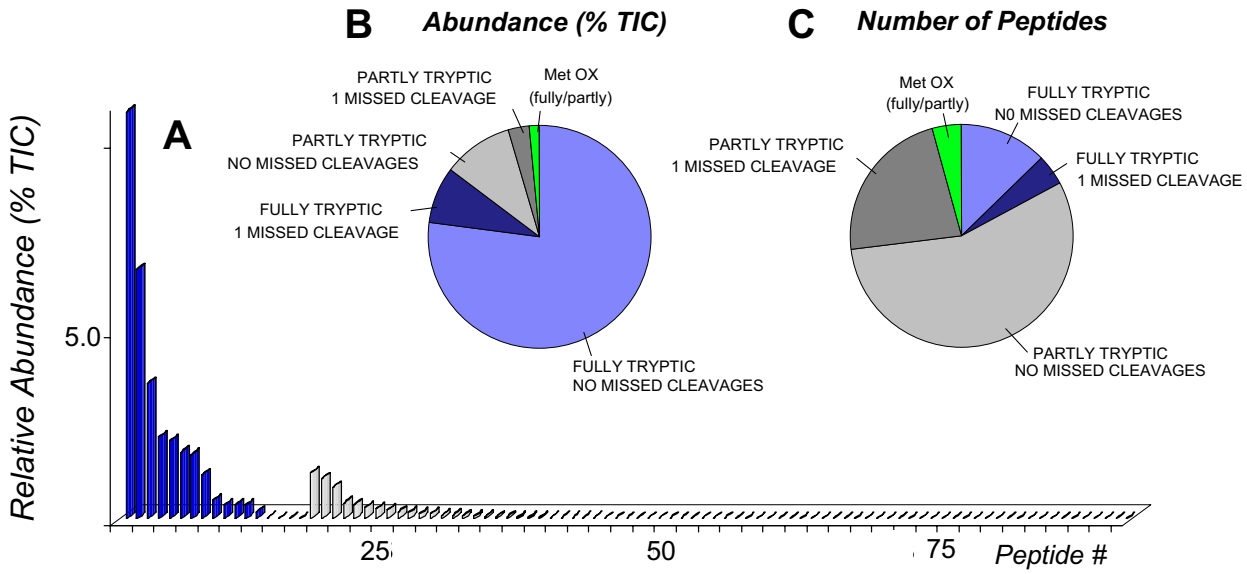

FIG. 5. **Semiquantitative analysis of β-lactoglobulin peptides.** *A*, histogram of peptide abundances expressed as fraction of the total ion current (*TIC*) of peptides identified. Fully tryptic peptides are in *blue*, and partly tryptic peptides are in *gray*. Pie charts show the cumulative abundances of partly and fully tryptic peptides containing one or no missed cleavages or containing an oxidized methionine (*Met OX*) in terms of fraction of the total ion current (*B*) and of number of peptides identified (*C*).

into account also a wide range of potential polypeptide modifications occurring naturally or as a consequence of sample handling (18), the actual complexity of tryptic digests is likely to exponentially increase. As an example, a database search was performed on the raw MS/MS data of the bovine albumin digest in which peptide carbamylation (at the N terminus or Arg or Lys residues) was allowed on fully tryptic peptides (Supplemental Table 4). Forty different carbamylated peptides were identified as a result of the reaction with isocyanic acid derived from urea decomposition. This number is likely to substantially increase if half-tryptic cleavage is allowed.

These findings have deep implications for the widely used shotgun strategy and for proteomics strategies based on reproducible LC/MS feature maps such as the accurate mass and time tag method (19). The latter approach relies on accurate mass measurements and normalized chromatographic retention times to identify peptides as a surrogate for MS/MS analysis. Preliminary bioinformatics analysis indicated that the number of peptides of a given mass within a window of a 1–2 ppm is at least 10–20 times higher if half-tryptic peptides are included compared with fully tryptic peptides only. The results of this study, therefore, do not only indicate that predictions based on strict trypsin specificity seriously underestimate the actual sample complexity but also that the assumptions on which the shotgun and accurate mass tag strategies are founded are greatly oversimplified.

Unspecifically cleaved peptides produced during trypsinization of single proteins were demonstrated to be generally of relatively low abundance (orders of magnitude lower) in the digestion mixture compared with the expected, specific peptides. Although this is not a significant problem for the analysis of single proteins or simple mixtures, for a complex biological sample, an extremely large number of low abundance peptides is expected to create a dense proteolytic background in mass spectra. Considering the wide range of protein concentrations in biological samples (20), the proteolytic background resulting from abundant components will tend to hide the signal of true tryptic peptides originating from lower abundance proteins. This hypothesis was confirmed by analyzing a fraction of a human serum digest enriched for glycopeptides (21) in which the β-lactoglobulin digest previously characterized was spiked in at the concentration of 1% by weight (see the supplemental materials). Several β-lactoglobulin tryptic peptides known to be present in the sample from the previous analysis were not detected in full-scan mass spectra, whereas several half-tryptic peptides from highly abundant serum proteins were clearly identified.

In addition, to evaluate the abundance of partly tryptic peptides in already existing proteomics datasets, a large publicly available proteomics data repository, PeptideAtlas, was searched with respect to the yeast and plasma builds (www.peptideatlas.org (22–24)). A major fraction of the overall number of peptides identified, 23.3 and 40.1% for yeast and plasma databases, respectively (Fig. 6), corresponds to partly tryptic sequences. The much higher value obtained for serum might reflect the presence of a high number of serum proteases with different specificities and the number of spectra represented in the respective databases (4 million *versus* 14 million). However, both figures are extremely high considering that database search constraints and filtering systems tend to penalize and underrepresent unexpected digestion products. This is explained by the fact that PeptideAtlas is a collection of many different proteomics experiments (conducted in different laboratories with different fractionation/enrichment protocols). As a consequence, highly abundant tryptic peptides will tend to be observed consistently in different samples and replicated in the database, whereas each low abundance half-tryptic peptide is associated to a much lower number of observations as illustrated in the plot shown in Fig. 6. For the same considerations, one might evoke a potentially high false positive rate of half-tryptic peptide identifications in PeptideAtlas. However, in compiling the database extreme care has been taken to address this issue by using statistical models and reverse database searches (22–24). This was further evaluated by manually inspecting a subset of MS/MS spectra stored in PeptideAtlas corresponding to partly tryptic peptides, most of which were reliable assignments. Therefore the presence in the database of such a high number of partly tryptic peptides is a clear indication of their existence in tryptic digests. Collectively these data show that the observations made with purified single proteins also apply to the analysis of more complex protein mixtures. This is a serious limitation that might partly explain the failure of shotgun proteomics approaches to detect low abundance proteins in very complex biological samples. For instance, despite the considerable efforts (millions of MS/MS spectra collected in conjunction with extensive fractionation) (25), the comprehensive mapping of complex proteomes such as the human serum proteome by shotgun approach remains a major challenge. Even relatively small proteomes such as the yeast proteome have not yet been fully covered (26).

Strategies based on depletion of highly abundant proteins (27) or extensive serial fractionation at the protein or peptide levels (28–31) may increase the identification rate of low abundance proteins but do not eliminate the intrinsic limitation of the shotgun technique. A broader use of targeted sequencing approaches, in which the detection of peptide ions and their actual sequencing are decoupled and performed in two consecutive experiments as described in this work, is a valuable approach that yields more in-depth characterization of peptide mixtures but remains insufficient to fully map a complex proteome such as the human serum proteome. Therefore, alternative strategies will have to be explored for rapid and comprehensive proteome coverage. One option is to shift to a hypothesis-driven paradigm (32) leveraging the knowledge from already existing proteomics (*e.g.* peptide/protein libraries (23, 33)) or genomics data to specifically screen for known or suspected proteins in a sam-

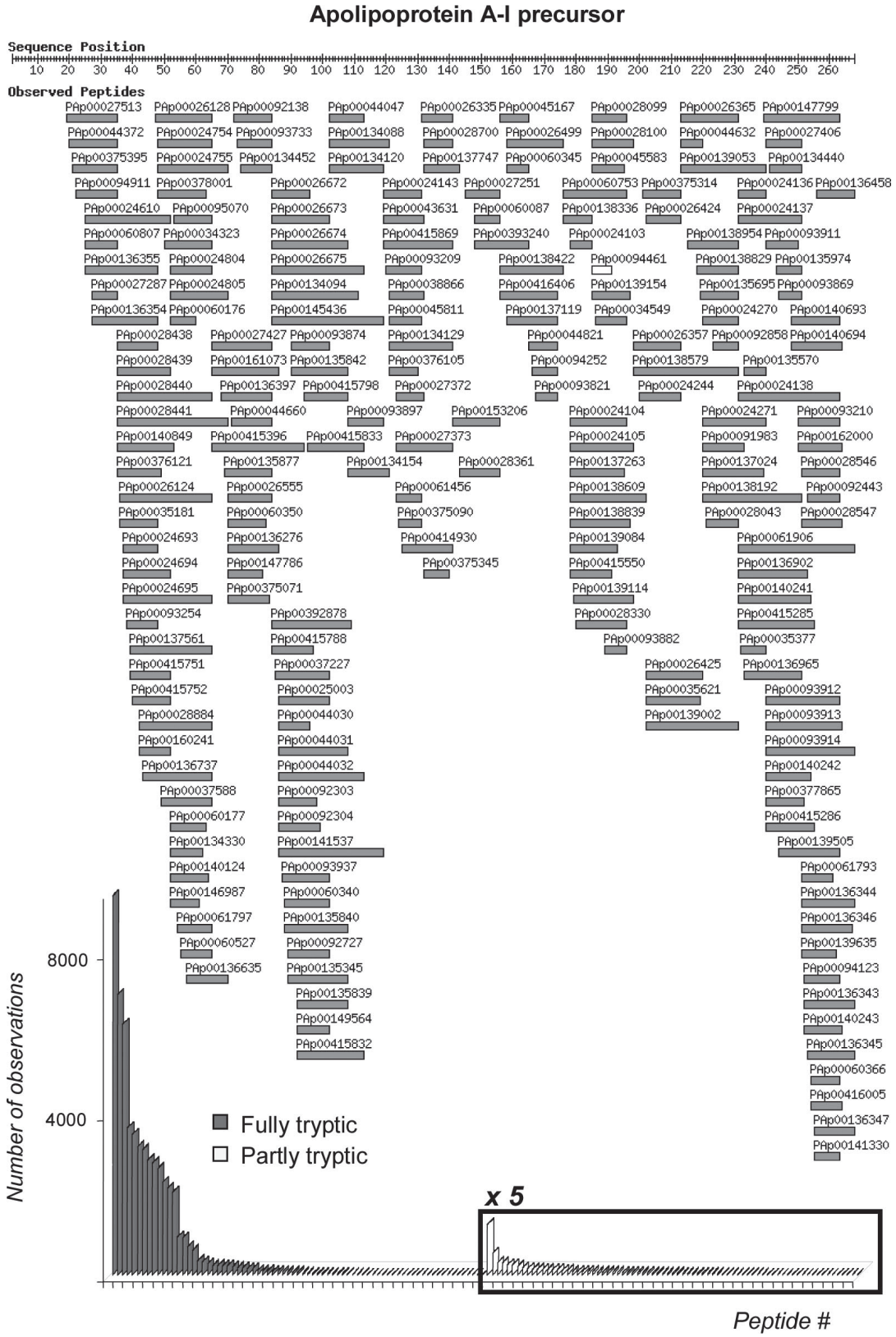

FIG. 6. **Peptide identifications reported in the proteomics data repository PeptideAtlas (screen shot from www.peptideatlas.org) as exemplified with the protein apolipoprotein A-I.** The histogram illustrates the occurrence of the 150 most represented peptides of apolipoprotein A-I expressed in terms of number of observations reported in PeptideAtlas (plasma build). The histogram in the *box* is magnified 5-fold.

ple. Such information will be used to determine proteotypic peptides (33, 34) and their fragmentation properties to screen for sets of proteins by multiple reaction monitoring (35, 36). In such experiments, performed on triple quadrupole-type instruments, pairs of precursor/fragment ions specific to a peptide are serially selected and monitored. This is a very powerful technique to detect analytes in complex samples at a sensitivity and specificity that is difficult to reach by shotgun proteomics approaches.

Acknowledgments—We thank Nichole King, Markus Mueller, Lukas Reiter, and Martin Hubalek for valuable assistance and discussion. We also thank Luca Codutti for providing an aliquot of the recombinant Pax-8 protein and Reto Ossola for providing a serum peptide sample. P. P. acknowledges Prof. Angelo Fontana for insightful discussions and support.

* This work was supported in part by federal funds from the NHLBI of the National Institutes of Health under Contract N01-HV-28179 and by the Swiss National Science Foundation under Contract 3100A0–107679 and the Competence Center for Systems Physiology and Metabolic Diseases, Zurich. 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.

S The on-line version of this article (available at http://www.mcponline.org) contains supplemental material.

‖ To whom correspondence should be addressed: Inst. of Molecular Systems Biology, ETH Zürich, Wolfgang-Pauli-Str. 16, HPT D 74, 8093 Zürich, Switzerland. Tel.: 41-44-633-20-88; E-mail: domon@imsb.biol.ethz.ch.

REFERENCES

1. Washburn, M. P., Wolters, D., and Yates, J. R., III (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19,** 242–247
2. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* **422,** 198–207
3. Sadygov, R. G., Cociorva, D., and Yates, J. R., III (2004) Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nat. Methods* **1,** 195–202
4. Nesvizhskii, A. I. (2006) Protein identification by tandem mass spectrometry and sequence database searching. *Methods Mol. Biol.* **367,** 87–120
5. Wolters, D. A., Washburn, M. P., and Yates J. R., III (2001) An automated multidimensional protein identification technology for shotgun proteomics. *Anal. Chem.* **73,** 5683–5690
6. Plumb, R. S., Rainville, P., Smith, B. W., Johnson, K. A., Castro-Perez, J., Wilson, I. D., and Nicholson, J. K. (2006) Generation of ultrahigh peak capacity LC separations via elevated temperatures and high linear mobile-phase velocities. *Anal. Chem.* **78,** 7278–7283
7. Luo, Q., Shen, Y., Hixson, K. K., Zhao, R., Yang, F., Moore, R. J., Mottaz, H. M., and Smith, R. D. (2005) Preparation of 20-μm-i.d. silica-based monolithic columns and their performance for proteomics analyses. *Anal. Chem.* **77,** 5028–5035
8. Andersen, J. S., Lam, Y. W., Leung, A. K., Ong, S. E., Lyon, C. E., Lamond, A. I., and Mann, M. (2005) Nucleolar proteome dynamics. *Nature* **433,** 77–83
9. Ranish, J. A., Yi, E. C., Leslie, D. M., Purvine, S. O., Goodlett, D. R., Eng, J., and Aebersold, R. (2003) The study of macromolecular complexes by quantitative proteomics. *Nat. Genet.* **33,** 349–355
10. Domon, B., and Broder, S. (2004) Implications of new proteomics strategies for biology and medicine. *J. Proteome Res.* **3,** 253–260
11. Domon, B., Picotti, P., Ossola, R., Stahl-Zeng, J., and Aebersold, R. (2005) Identification and quantification of biomarkers in serum samples, in *Proceedings of the 53rd ASMS Conference on Mass Spectrometry, San Antonio, Texas, June 5–9, 2005*, TH0Apm-3, American Society for Mass Spectrometry, Santa Fe, NM
12. Picotti, P., Lee, H., Domon, B., and Aebersold, R. (2006) Novel approach to identify low abundance biomarkers in serum, in *Proceedings of the 54th ASMS Conference on Mass Spectrometry, Seattle, Washington, May 28–June 1, 2006*, MOB-6, American Society for Mass Spectrometry, Santa Fe, NM
13. Olsen, J. V., Ong, S. E., and Mann, M. (2004) Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Mol. Cell. Proteomics* **3,** 608–614
14. Keil, B. (1992) *Specificity of Proteolysis*, pp. 66–69, Springer-Verlag, New York
15. Keil, B., and Tong, N. T. (1993) *Lysis: A Proteolysis Data Base*, pp. 66–73, Springer-Verlag, New York
16. Chalkley, R. J., Baker, P. R., Hansen, K. C., Medzihradszky, K. F., Allen, N. P., Rexach, M., and Burlingame, A. L. (2005) Comprehensive analysis of a multidimensional liquid chromatography mass spectrometry dataset acquired on a quadrupole selecting, quadrupole collision cell, time-of-flight mass spectrometer: I. How much of the data is theoretically interpretable by search engines? *Mol. Cell. Proteomics* **4,** 1189–1193
17. Savitski, M. M., Nielsen, M. L., Kjeldsen, F., and Zubarev, R. A. (2005) Proteomics-grade de novo sequencing approach. *J. Proteome Res.* **4,** 2348–2354
18. Tsur, D., Tanner, S., Zandi, E., Bafna, V., and Pevzner, P. A. (2005) Identification of post-translational modifications by blind search of mass spectra. *Nat. Biotechnol.* **23,** 1562–1567
19. Smith, R. D., Anderson, G. A., Lipton, M. S., Pasa-Tolic, L., Shen, Y., Conrads, T. P., Veenstra, T. D., and Udseth, H. R. (2002) An accurate mass tag strategy for quantitative and high-throughput proteome measurements. *Proteomics* **2,** 513–523
20. Anderson, N. L, and Anderson, N. G. (2002) The human plasma proteome: history, character, and diagnostic prospects. *Mol. Cell. Proteomics* **1,** 845–867
21. Zhang, H., Li, X. J., Martin, D. B., and Aebersold, R. (2003) Identification and quantification of N-linked glycoproteins using hydrazide chemistry, stable isotope labeling and mass spectrometry. *Nat. Biotechnol.* **21,** 660–666
22. Desiere, F., Deutsch, E. W., King, N. L., Nesvizhskii, A. I., Mallick, P., Eng, J., Chen, S., Eddes, J., Loevenich, S. N., and Aebersold, R. (2006) The PeptideAtlas project. *Nucleic Acids Res.* **34,** D655–D658
23. Deutsch, E. W., Eng, J. K., Zhang, H., King, N. L., Nesvizhskii, A. I., Lin, B., Lee, H., Yi, E. C., Ossola, R., and Aebersold, R. (2005) Human Plasma PeptideAtlas. *Proteomics* **5,** 3497–3500
24. King, N. L., Deutsch, E. W., Ranish, J. A., Nesvizhskii, A. I., Eddes, J. S., Mallick, P., Eng, J., Desiere, F., Flory, M., Martin, D. B., Kim, B., Lee, H., Raught, B., and Aebersold, R. (2006) Analysis of the Saccharomyces cerevisiae proteome with PeptideAtlas. *Genome Biol.* **7,** R106
25. Omenn, G. S., States, D. J., Adamski, M., Blackwell, T. W., Menon, R., Hermjakob, H., Apweiler, R., Haab, B. B., Simpson, R. J., Eddes, J. S., Kapp, E. A., Moritz, R. L., Chan, D. W., Rai, A. J., Admon, A., Aebersold, R., Eng, J., Hancock, W. S., Hefta, S. A., Meyer, H., Paik, Y. K., Yoo, J. S., Ping, P., Pounds, J., Adkins, J., Qian, X., Wang, R., Wasinger, V., Wu, C. Y., Zhao, X., Zeng, R., Archakov, A., Tsugita, A., Beer, I., Pandey, A., Pisano, M., Andrews, P., Tammen, H., Speicher, D. W., and Hanash, S. M. (2005) Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. *Proteomics* **5,** 3226–3245
26. Ghaemmaghami, S., Huh, W. K., Bower, K., Howson, R. W., Belle, A., Dephoure, N., O'Shea, E. K., and Weissman, J. S. (2003) Global analysis of protein expression in yeast. *Nature* **425,** 737–741
27. Ettorre, A., Rosli, C., Silacci, M., Brack, S., McCombie, G., Knochenmuss, R., Elia, G., and Neri, D. (2006) Recombinant antibodies for the depletion of abundant proteins from human serum. *Proteomics* **6,** 4496–4505
28. Malmstrom, J., Lee, H., Nesvizhskii, A. I., Shteynberg, D., Mohanty, S., Brunner, E., Ye, M., Weber, G., Eckerskorn, C., and Aebersold, R. (2006) Optimized peptide separation and identification for mass spectrometry based proteomics via free-flow electrophoresis. *J. Proteome Res.* **5,** 2241–2249

29. Heller, M., Ye, M., Michel, P. E., Morier, P., Stalder, D., Junger, M. A., Aebersold, R., Reymond, F., and Rossier, J. S. (2005) Added value for tandem mass spectrometry shotgun proteomics data validation through isoelectric focusing of peptides. *J. Proteome Res.* **4,** 2273–2282
30. Heller, M., Michel, P. E., Morier, P., Crettaz, D., Wenz, C., Tissot, J. D., Reymond, F., and Rossier, J. S. (2005) Two-stage Off-Gel isoelectric focusing: protein followed by peptide fractionation and application to proteome analysis of human plasma. *Electrophoresis* **26,** 1174–1188
31. Righetti, P. G., Castagna, A., Herbert, B., Reymond, F., and Rossier, J. S. (2003) Prefractionation techniques in proteome analysis. *Proteomics* **3,** 1397–1407
32. Domon, B., and Aebersold, R. (2006) Mass spectrometry and protein analysis. *Science* **312,** 212–217
33. Kuster, B., Schirle, M., Mallick, P., and Aebersold, R. (2005) Scoring proteomes with proteotypic peptide probes. *Nat. Rev. Mol. Cell Biol.* **6,** 577–583
34. Mallick, P., Schirle, M., Chen, S. S., Flory, M. R., Lee, H., Martin, D., Ranish, J., Raught, B., Schmitt, R., Werner, T., Kuster, B., and Aebersold R. (2007) Computational prediction of proteotypic peptides for quantitative proteomics. *Nat. Biotechnol.* **25,** 125–131
35. Anderson, L., and Hunter, C. L. (2006) Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins. *Mol. Cell. Proteomics* **5,** 573–588
36. Domon, B., Stahl-Zeng, J., and Aebersold, R. (2006) Novel strategy for rapid screening and quantification of biomarkers in serum, in *Proceedings of the 54th ASMS Conference on Mass Spectrometry, Seattle, Washington, May 28–June 1, 2006*, Toc-88, American Society for Mass Spectrometry, Santa Fe, NM