Most known organisms encode proteases that are crucial for constitutive proteolytic events. In the present paper, we describe a method to define these events in proteomes from *Escherichia coli* to humans. The method takes advantage of specific N-terminal biotinylation of protein samples, followed by affinity enrichment and conventional LC (liquid chromatography)–MS/MS (tandem mass spectrometry) analysis. The method is simple, uses conventional and easily obtainable reagents, and is applicable to most proteomics facilities. As proof of principle, we demonstrate profiles of proteolytic events that reveal exquisite in vivo specificity of methionine aminopeptidase in *E. coli* and unexpected processing of mitochondrial transit peptides in yeast, mouse and human samples. Taken together, our results demonstrate how to rapidly distinguish real proteolysis that occurs in vivo from the predictions based on in vitro experiments.

**Key words:** mass spectrometry, methionine aminopeptidase (MetAP), mitochondrial peptidase, N-terminal labelling, peptidase, protease, signal peptidase.

**INTRODUCTION**

Approx. 1–5% of a genome encodes proteolytic enzymes, depending on the species [1,2]. Proteases were originally identified over a century ago as protein-degrading enzymes in digestive juices and tissue homogenates, leading to the concept that they are protein destructors. The current view, however, sees proteases much more as conducting specific limited proteolysis to play pivotal roles in most biological processes [3]. The principle of proteolysis in vivo is to instigate irreversible changes to a set of protein substrates that alters their function, thereby generating the required biological event. The sum total of the proteases and their target substrates operating in a physiological pathway therefore defines the global proteolytic signature of that pathway [4]. Dysregulation of proteolysis in organisms is deleterious, and many abnormal developmental problems and diseases are attributed to aberrant proteolytic activity [2]. Since the job of proteases is to cleave protein substrates, then a vital part of understanding the role of proteolysis in health and disease is to determine the products of this proteolysis.

Efforts to understand protease and substrate interactions have primarily focused on defining a protease’s cleavage site specificity. Protease specificity has been defined by synthetic peptide libraries or phage display technology [5–12]. Although this information is important, it does not directly lead to natural substrate identification, and (paradoxically) has sometimes produced substrate predictions that do not occur naturally [5]. Identifying the natural substrates cleaved in vivo is technically challenging, but more insightful, because it delivers information not on what a protease can do, but on what it does. To achieve the goal of in vivo substrate determination, one needs to identify the N-terminal sequence of proteolytic products by proteomics, for example Edman degradation or MS.

Proteolytic events can be considered to be either constitutive or regulated. In contrast with regulated proteolytic events that depend on a specific trigger, constitutive proteolytic events remove unwanted portions of proteins following translation. Examples of constitutive proteolytic events are methionine removal by MetAPs (methionine aminopeptidases) [13], signal peptide removal by signal peptidases [14] and mitochondrial transit peptide removal by mitochondrial peptidases [15]. Protein databases contain most of the predicted proteins expressed in a given species and frequently contain sites predicted to be targeted by these proteases. However, only a fraction have been determined experimentally. The challenge is to develop a technology that will determine the magnitude and specificity of these constitutive proteolytic events in vivo.

In search of protease specificity in vivo, we have developed a proteomics approach that combines specific N-terminal tagging of proteins with affinity enrichment and LC (liquid chromatography)–MS/MS (tandem MS) detection, and which uses readily available reagents and can be adopted easily by any laboratory. We present extensive validation of the approach, and demonstrate a series of conventional and unusual proteolytic events as a result of the action of proteases on natural substrates.

**EXPERIMENTAL**

**Analysis of aprotinin guanidination**

Aprotinin (10.3 mg) was denatured (in 100 mM Hepes, pH 7.5, 50 mM NaCl and 6 M guanidine) and reduced [in 10 mM DTT]
(dithiothreitol) at 50 °C for 60 min. Following alkylation (in 50 mM iodoacetamide), sample pH was raised to approx. 10.3 by addition of NaOH. α-Methylisourea was added to a final concentration of 0.5 M, the pH was readjusted to 10.3, and the sample was incubated at 4 °C for 1 or 16 h. Samples were prepared for MS using C18 ZipTips (Millipore) according to the manufacturer’s protocol and spotted on to the MALDI (matrix-assisted laser-desorption ionization) target in 2 μl of matrix solution [10 mg/ml α-cyano-4-hydroxycinnamic acid in 50% acetonitrile/0.1% TFA (trifluoroacetic acid)]. MALDI–MS was performed on an Applied Biosystems Voyager-DE PRO Biospectrometry Workstation. Total amino acid analysis was performed with 250 μg (38.5 nmol) of guanidinated aprotinin. Protein digestion was performed in a vacuum hydrolysis tube (Pierce) in 6 M HCl at 110 °C overnight. Samples were evaporated, dissolved in coupling buffer (acetonitrile//pyridine/trietylamine/water, 10:5:2:3, by vol.) and derivatized with phenylisothiocyanate (Pierce) for 5 min at room temperature (23 °C). Samples were evaporated and dissolved in 50 mM ammonium acetate (pH 6.8). Aliquots (20 μl) were injected on to a Varian Microsorb C18 reverse-phase HPLC column, equilibrated to 52 °C and eluted with a gradient of 100 mM ammonium acetate (pH 6.8) and 50% acetonitrile using a Beckman System Gold HPLC system.

Characterizing chemical derivatization

The labelling reagent NHS-SS-biotin [sulfosuccinimidyl-2-(biotinamido)ethyl-1,3-dithiopropionate] can react with side chains of serine, threonine, histidine and unreacted lysine residues. Since the database search only allows for modification of the N-terminal, MS/MS spectra corresponding to a peptide with a labelled side chain will be unassigned and not detected as a false-positive event. However, major side reaction could possibly cause saturation of the streptavidin and limit detection by MS/MS. To characterize N-terminal peptide recovery and quantify side reactions of the biotin tag, we prepared and analysed a defined test sample. We expressed and purified nine recombinant proteins with anticipated N-terminal peptides of favourable size after tryptic digestion (see Figs 2C and 2D). The proteins used were human caspase 3 C285A, human caspase 7 C285A, human caspase 9 C285A, human caspase 3/7 D9A/C285A, human caspase 7/9 D9A/C285A, human caspase 3/7/9 D9A/C285A, human caspase 3/7/9 D9A/C285A/ C285A, human caspase 3/7/9 D9A/C285A/ C285A, and three N-terminal mutants (SGPI, MVPI and ANPR) of Smac (second mitochondrial processing intein-terminus) will be unassigned and not detected as a false-positive event. MS/MS spectra corresponding to a peptide with a labelled side chain will be unassigned and not detected as a false-positive event. Since the database search only allows for modification of the N-terminal, MS/MS spectra corresponding to a peptide with a fixed N-terminal adduct, as well as potential side reactions.

Sample preparation

Escherichia coli strain MG1655 cultures were grown in 2 × YT medium [1.6 % (w/v) tryptone, 1 % (w/v) yeast extract and 0.5 % (w/v) NaCl] in baffled flasks with shaking at 200 rev./min at 37 °C to a D600 of 0.8 for exponential-phase cultures, or left overnight for stationary-phase cultures. FVB/N wild-type mice were housed and bred in compliance with the NIH (National Institutes of Health) guidelines and the Burnham Institute Animal Research Committee. Female mouse liver, kidney, heart and skeletal muscle were surgically removed following killing, briefly washed in PBS solution and snap-frozen in liquid nitrogen. C57/BL6 mouse peritoneal macrophages were elicited by thioglycolate injection and collected 3 days later, washed in PBS and kept on ice.

Yeast strain BY4741 map1Δ (map1::KanMX) was obtained from A.T.C.C. (Manassas, VA, U.S.A.). The map1Δ slow-growth phenotype was rescued by transformation with a single-copy plasmid containing the MAP1 gene under control of 1 kb of the endogenous UAS (upstream activating sequence) (pRS415MAP1) [16]. Yeast transformation was performed with lithium acetate. For analysis, yeast cultures (100 ml) were grown to a D600 of 1.0 in YPD [1% (w/v) yeast extract, 2% (w/v) peptone and 2% (w/v) glucose]. Cells were pelleted by centrifugation at 2000 g for 5 min, the pellets were washed once with water and then frozen at −80 °C until analysis.

HEK-293A (human embryonic kidney) cells were grown in DMEM (Dulbecco’s modified Eagle’s medium) supplemented with 10% fetal bovine serum, 100 units/ml penicillin, 100 μg/ml streptomycin and 2 mM glutamine at 37 °C in a humidified atmosphere containing 5% CO2. The cells were harvested by scraping in cold PBS and washed twice in cold PBS. Cytosolic HEK-293A cell extracts were prepared using hypotonic buffer (20 mM Pipes, pH 7.4, 10 mM KCl, 5 mM EDTA, 2 mM MgCl2 and 4 mM DTT), essentially as described in [17]. Briefly, the cells were harvested by scraping in cold PBS, washed twice in cold PBS, and incubated in hypotonic buffer for 30 min on ice to induce cell swelling. Extracts were prepared by cell membrane shearing using 20- and subsequently, 27-gauge needles followed by centrifugation at 1000 g for 20 min. The supernatants were centrifuged a second time and the resulting supernatants were collected and stored at −80 °C.

Human serum samples were prepared from whole blood drawn from the antecubital vein into vacutainer tubes containing either no additive or acid citrate dextrose solution (Becton Dickinson). Whole blood containing no additives was allowed to clot for 15 min at room temperature. Samples were then centrifuged at 1500 g for 15 min at 4 °C to pellet cellular blood components. Serum, obtained from the tube containing no additives, and plasma, obtained from the tube containing acid citrate dextrose solution, were divided into aliquots and stored at −80 °C. The samples were filtered (0.45 μm Whatman filter). Albumin and IgG were removed in some of the samples using ProteoExtract™ albumin/IgG removal kit (Calbiochem), according to the manufacturer’s instructions. Guanidine (6 M) and DTT (10 mM) were added immediately after depletion of IgG and albumin (depleted samples, 60 μl) or after filtering (non-depleted samples, 350 μl).

Labelling procedure

Yeast, E. coli, mouse, HEK-293A cell and blood samples were immediately denatured and reduced in 6 M guanidine with 10 mM DTT, and boiled for 10 min to inactivate cellular proteases. Iodoacetamide (30 mM) was added to alkylate cysteine side chains. The pH of each sample was increased to 10.3 with NaOH before adding 0.5 M α-methylisourea. The pH was readjusted to 10.3, and the lysine guanidination reaction was carried out at 4 °C for 20 h. The proteins were desalted by buffer exchange (PD-10, Amersham Biosciences) into urea buffer (8 M urea, 50 mM Hepes, pH 7.8, and 50 mM NaCl). The urea stock solution was made fresh, deionized using AG 501-X8 resin (Bio-Rad) and filtered before use. The proteins were subsequently labelled by 5 mM EZ-Link sulfo-NHS-SS-biotin (Pierce Biotechnology) at room temperature for approx. 1 h. This NHS-reactive reagent is specific for the N-terminal of the proteins, i.e. native N-termini and proteolytic cleavage sites, owing to the previous blocking of cysteine and lysine side chains, and the biotin tag of the molecule allows for positive selection by immobilized streptavidin. Eventually, unreacted biotin reagent was quenched by the addition of 50 mM glycine for 30 min and subsequently excluded by buffer exchange into 4 M urea, 100 mM Hepes, pH 7.8, and...
100 mM NaCl. The samples were diluted 1:2 with distilled water before digestion overnight by sequencing-grade modified trypsin (Promega) or endoproteinase Glu-C (Roche). Further enzymatic activity was inhibited by boiling the samples for 5 min. The samples were centrifuged at 10 000 g for 5 min, and the resulting supernatant was added to immobilized streptavidin (Pierce Biotechnology) for 1 h at room temperature for selection and enrichment of the biotinylated N-terminal peptides. The streptavidin beads were washed extensively with AmmBic buffer (50 mM triethylammonium bicarbonate, pH 7.8), high-salt AmmBic buffer (AmmBic buffer containing 1 M NaCl) and finally by AmmBic buffer again. The flow-through and first AmmBic wash were collected and allowed to bind new streptavidin beads to minimize eventual loss of labelled peptides. The labelled peptides were eluted by the addition of 50 mM DTT, which cleaves the disulfide-linked biotin tag, leaving an 88 Da addition to the N-termini of the labelled peptides. The peptide elution was dried by vacuum to reduce sample volume, and desalted using C18 OMIX tips (Varian), according to the manufacturer’s instructions. The peptide solution was again dried by vacuum and redissolved in 0.1% TFA for analysis by LC–MS/MS.

Sample analysis by nano LC–MS

The automated NanoLC-LTQ system consists of an Eksigent Nano-2D LC autosampler, a switch valve, a C18 trap column (Agilent), a capillary separation column (100 µm internal diameter × 10 cm length, packed with Synergi 4 µm C18), and a LTQ ion-trap mass spectrometer (Thermo Electron). The separation column is mounted into the Finnigan Nanospray II ion source (Thermo Electron) and used as the electrospray tip as well. First, trypsin- or Glu-C-digested peptides (3–9 µl) were loaded by autosampler on to the trap column in 100 % solvent A [2 % acetonitrile and 0.1 % methanoic (formic) acid] using a flow rate of 10 µl/min for 4 min. After sample loading and washing, the valve was switched, and the gradient was delivered to the trap and separation column at 500 nL/min. Peptides were separated with a 100–120 min linear gradient of 10–60 % solvent B (80 % acetonitrile and 0.1 % methanoic acid), and was then eluted directly into the LTQ spectrometer. The fully automated NanoLC-LTQ was operated via an Instrument Method of Xcalibur. MS/MS spectra were collected automatically during the LC–MS runs. Each scan was set to acquire a full MS scan followed by four MS/MS scans of the four most intense ions from the preceding MS scan.

Database searching

After data acquisition, MS/MS spectra were then extracted and searched against the corresponding protein database (Swiss-Prot) using SEQUEST SorcererTM (SageN). For E. coli samples, a non-enzymatic peptide database was used. For human, mouse and yeast samples, semi-tryptic or Glu-C peptide databases were used. A molecular mass of 88 Da was added to the static search of all N-termini to account for carboxyamidomethylation. A differential search of amino acids includes methionine + 16 Da for oxidation and lysine +42 Da for guanidination. In some searches, we included +88 Da for possible side reactions of the biotinylation probe of serine, threonine, histidine or lysine residues. After SEQUEST searching, the results were filtered automatically, organized and displayed by PeptideProphet and ProteinProphet (ISB), which are installed in SorcererTM. A minimum probability score of 0.95 was set to assure low errors in peptide identification (see the Results and discussion section). All peptides must have cross-correlation values (Xcorr) of at least 2.0. The false-positive peptide identification rate was quantified using forward and reverse database searching, and was found to be only 0.53 %.

RESULTS AND DISCUSSION

Method outline and validation

This method takes advantage of the fact that a single proteolytic event generates a new N-terminal amine (see Figure 1 for method outline). This new amine and the original N-terminal amine of the protein(s) can be specifically labelled by an amine-reactive tag, provided that other amines (e.g. lysine side chains) are blocked. Lysine guanidination using o-methylisourea has long been used to aid in the derivatization of amines for protein MS [18–21]. We optimized guanidination conditions using purified aprotinin (Figures 2A and 2B). Treatment of the test protein aprotinin with o-methylisourea resulted in a mass increase corresponding to guanidination of the four lysine residues (Figure 2A). Since the lysine signal in amino acid analysis was almost completely suppressed (Figure 2B), we conclude that only the lysine ε-amines were modified, and any guanidination of the aprotinin
Figure 2 Validation of chemical derivatization and peptide identification

(A) Guanidination of aprotinin. Aprotinin was denatured, alkylated and guanidinated (+42 Da/lysine residue) at 4°C for 1 h (ii) or overnight (iii). (i) MALDI spectra of aprotinin before addition of o-methylisourea hemisulfate. The main MALDI peak in (i) demonstrates the m/z for unmodified aprotinin, whereas in (iii), the main peak corresponds to successful modification of all four ε-NH₂ groups of the protein. Note the time-dependent shift in the main mass peak during the guanidination reaction. (B) Amino acid analysis of guanidinated aprotinin. Guanidinated aprotinin was exchanged into HCl for hydrolysis to amino acids, and analysed following pre-column derivatization and RPLC. The amino acid analysis revealed a composition characteristic of the protein, with individual peaks matching for untreated (black line) and guanidinated samples (grey line), with the exception of the lysine peak. Lower panels show magnifications of the homoarginine (hArg) peak region and the lysine peak region. Note the almost complete suppression of lysine and an accompanying formation of homoarginine. (C) Method validation using a purified protein mixture. Peptides were identified from spectra by searching with a fixed modification of all peptide N-terminal amino acids. Side reactions of the biotin tag with unintended amino acid side chains were identified using a differential modification on lysine (K), serine (S), threonine (T) or histidine (H) residues. Less than 7% of peptides identified had an undesired side chain modification with the biotin tag. (D) Over 80% of the peptides identified without side-chain adducts corresponded to the expected protein N-terminus. ORF, open reading frame. (E) Probability score. Distribution of probability score of all observed peptides in a typical E. coli sample. The arrow indicates the cut-off (0.95) for peptide assignment. A cross-correlation (Xcorr) cut-off at 2.0 was also included.

N-terminal amine was undetectable. We can not rule out that unwanted guanidination of proteins at the N-terminus may occur, preventing subsequent detection, but our validation experiments with aprotinin suggest that this would be a rare occurrence and would not have a substantial impact on data acquisition. Indeed, previous studies of protein and peptide guanidination revealed very little, if any, N-terminal modification under conditions similar to those of the present study [18–21]. The efficiency of lysine side-chain guanidination was >94%, tested on a mixture of nine purified proteins (results not shown), which can also be estimated from the peptide data in Supplementary Tables 1–5 (http://www.BiochemJ.org/bj/407/bj4070041add.htm).

Strategic N-terminal labelling using a disulfide-cleavable biotinylated tag enabled enrichment and selection of the tagged peptides following digestion of the protein sample. Potential side reactions of the biotin tag with the side chains of lysine, serine, threonine and histidine residues were found to constitute only 6.6% of all N-terminally tagged peptides (Figures 2C and 2D). We identified peptides by LC–MS/MS and parsed the data by filtering for peptides with the chemically modified N-terminus. Like in other proteomic analyses of post-translational modifications, we expected to identify a single peptide per protein, and so the inclusion criteria must be rigorous. To this end, we included peptides with a PeptideProphet probability score of >0.95 and cross-correlation (Xcorr) value of >2.0, which is extremely stringent (Figure 2E). Specifically, searching a representative E. coli dataset against a forward and reverse database, using these criteria, resulted in a false-positive peptide identification rate of only 0.53%. Thus we demonstrate that the accepted spectra were of high quality, and the resulting peptide identifications constituted sufficient evidence to confidently identify proteins of interest. We analysed each individual sample three times to increase our confidence in detected peptides (owing to higher spectral counts), as well as enhancing the possibility of detecting peptides of low abundance. We also analysed repeated preparations of each sample type and used both trypsin and Glu-C digestion to increase the total number of peptides and thus again increase spectral counts and our confidence in the MS data. Depending on the origin of the sample analysed, we obtained an overlap of 50–70%
the initiator methionine, in contrast with eukaryotes, where N-acetyltransferases block approx. 50% of yeast cytosolic proteins and upwards of 80% of mammalian ones [23]. Analysis of 365 N-terminal E. coli peptides provided a profile for the in vivo specificity of MetAP, which demonstrated a strict preference for small and uncharged amino acids in the P1′ position (Figure 4). Thus our ex vivo analysis of natural substrates cleaved in vivo supports the in vitro analysis of MetAP specificity using synthetic substrates [24,25].

Protein trafficking across membranes is governed by N-terminal sequences that are proteolytically removed upon translocation. Signal peptidases remove the signal peptides required to drive translocation of proteins through the cell membrane in prokaryotes, or the secretion apparatus in eukaryotes. Hallmark features of signal peptides are (i) a basic N-terminus, (ii) a hydrophobic membrane-spanning stretch, and (iii) a C-terminal polar region terminating in an Ala-Xaa-Ala motif [26]. We analysed peptides that corresponded to a cleavage site between residues 15 and 50 as a search criterion for signal peptides. In E. coli, we observed 28 signal peptide events, of which 21 had been previously determined experimentally and seven predicted (Supplementary Table 1). Our sampling of E. coli signal peptidase cleavage sites confirmed the accuracy of predictions for this prokaryote. However, we found evidence for miss-annotations of signal peptidase cleavage sites in mammalian samples (Supplementary Tables 3–5). Thus predictions of signal peptidase are more accurate in E. coli than in mammals. Mitochondrial proteins are imported into the matrix following transit peptide removal by MPP (mitochondrial processing peptidase). Transit peptides are heterogeneous in length, generally contain arginine in the P3 or P2 position, and the P1′ residue is often aromatic or hydrophobic [27]. The matrix-localized MIP (mitochondrial intermediate peptidase) can subsequently process imported peptides by removing the N-terminal eight amino acids [27]. Additionally, matrix proteins can be translocated to the intermembrane space by IMP (inner membrane protease) 1 and 2, which recognize and process N-terminal transit peptides [28]. Many mitochondrial proteins are processed by these proteases, however, the exact cleavage sites follow very loose consensuses, making prediction problematic [27], and are often unknown or inferred from a few well-studied examples. In the yeast, mouse and human datasets, we find N-termini confirming annotated and predicted cleavage sites of mitochondrial transit peptides. Strikingly, we find substantial discrepancies between the predicted and observed limits of mitochondrial transit peptides (Table 1 and Supplementary Tables 2–5). Representative high-quality spectra identify new transit peptide-cleavage sites of mitochondrial proteins (see Supplementary Figure 1 at http://www.BiochemJ.org/bj/407/bj4070041add.htm). Alternatively, heterogeneity in processing can occur through aminopeptidase activity. In sampling these proteomes, we present a strategy that can be pivotal in refining the authentic cleavage sites used in vivo by the several MPPs.

In human serum, we observed hallmarks of specific limited proteolysis of the blood clotting cascade, and considerable trimming of N-termini (Figure 5). N-terminal trimming of serum proteins has been observed previously [29], and we detected a series of nested N-terminal peptides consistent with the action of ectopic cell-surface proteases, including enzymes that remove one residue at a time (aminopeptidase N), or two residues at a time (APN, aminopeptidase N) and DPPIV (dipeptidyl peptidase IV)]. These cell-surface proteases have been reported to modify tumour cell behaviour [30,31]. DPPIV releases a dipeptide from circulating glucagon-like peptides, gastric inhibitory polypeptide and members of the enteroglucagon/GRF (growth-hormone-releasing factor) superfamily, resulting in their

when comparing the peptides found in the same sample run three times on two different occasions, which is comparable with that reported previously (~70% reproducibility in LTQ-MS/MS analysis of yeast samples [22]).

Identification of N-termini from proteolytically processed and unprocessed proteins validates our methodology. Indeed, our analysis of E. coli revealed peptides corresponding to 8.7% of all predicted protein N-termini (365 native protein N-termini and 28 with the signal peptide removed, from 4506 predicted genes). The method confirms predicted proteolytic events in vivo and elucidates previously uncharacterized cleavage events. Examples of constitutive proteolytic events are methionine removal by MetAPs [13], signal peptide removal by signal peptidases [14] and mitochondrial transit peptide removal by mitochondrial peptidases [15]. We set out to profile these constitutive proteolytic events in several biological samples.

**Figure 3** Summary of annotated events (A) and unique N-termini (B)

(A) Summary of annotated events. Distribution of original N-termini and proteolytic events from different species. Blue, initiator methionine; red, initiator methionine removed; yellow, signal peptide removed; green, mitochondrial transit peptide removed; dark blue, pro-peptide removed. (B) Summary of unique N-termini. Distribution of unique N-termini observed from different species. Blue, initiator methionine; red, initiator methionine removed; yellow, signal peptide removed; green, mitochondrial transit peptide removed; dark blue, pro-peptide removed; light blue, unascribed peptides.

**Analysis of** E. coli, yeast, mouse and human proteomes

Peptides were characterized using Swiss-Prot features corresponding to annotated proteolytic events and grouped into functional categories (Figure 3). Spectra were collected by the mass spectrometer using dynamic exclusion criteria; however, abundant proteins were often identified by multiple spectra corresponding to the same N-terminal peptide. The majority of N-terminal peptides in E. coli were derived from open reading frames with or without
Figure 4  *E. coli* MetAP specificity in vivo

Extent of MetAP processing depending on the residue that follows the initiator methionine, shown as percentage of the total for each residue. Black bars show removal of methionine, and grey bars show retained methionine at protein N-termini. The total number of unique N-termini observed per residue are shown below. *E. coli* MetAP displays a strict and efficient processing preference for small and uncharged amino acids in the P1′ position.

Table 1 Mitochondrial transit peptides

Disagreements between our data (P1 observed) and the corresponding Swiss-Prot annotations (P1 annotated) were most severe when we analysed proteins annotated in Swiss-Prot as mitochondrial proteins. In each case, the observed cleavage site was consistent with prior removal of the transit peptides that help sort the proteins to mitochondria. In addition to our newly discovered transit peptide cleavage sites (NOVEL), we observed a substantial portion of different cleavage sites (NEW TRANSIT), sometimes consistent with the action of MIPs that generally remove additional octapeptides from proteins localized in the mitochondrial matrix. Occasionally, we observed trimming (TRIMMING) of single residues, presumably by aminopeptidases. Proteins with two distinct peptides are highlighted in grey. Spectral counts indicate the number of times the unique peptide was identified, and guanidinated lysine, which gives a homoarginine derivative is abbreviated as K#.

<table>
<thead>
<tr>
<th>YEAST</th>
<th>SwissProt ID</th>
<th>Peptide sequence</th>
<th>Spectral counts</th>
<th>P1 observed</th>
<th>P1 annotated</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BYHE YEAST</td>
<td>F.LATYAQPQGKX Y</td>
<td>5 14 12</td>
<td>NEW TRANSIT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CHYE YEAST</td>
<td>L.ATVHQPGQGR Y</td>
<td>26 15 15</td>
<td>ANNOTATED</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CIIH YEAST</td>
<td>G.MATFATAAATPR L</td>
<td>2 18 23</td>
<td>NEW TRANSIT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CIIIH YEAST</td>
<td>M.ATTA4KATHPR L</td>
<td>4 19 32</td>
<td>NEW TRANSIT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GLYM YEAST</td>
<td>G.LLTSQASLTVQKIPVSQDDPHEPFLQLQRE H</td>
<td>3 19 20</td>
<td>NEW TRANSIT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INFY YEAST</td>
<td>L.HKKPTITLNE N</td>
<td>3 21 unknown NOVEL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EFPA YEAST</td>
<td>Y.ASLTFRHSSQKQgeoisv S</td>
<td>12 22 unknown NOVEL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DGCH YEAST</td>
<td>N.DQSHURLNKBPIYRVSQPOQAVR Y</td>
<td>23 32 47</td>
<td>NEW TRANSIT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UCR1 YEAST</td>
<td>L.QLSSLTSLKETVR T</td>
<td>2 22 30</td>
<td>NEW TRANSIT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CD02 YEAST</td>
<td>F.KTSTIEIPMAELEISQILKRETK N</td>
<td>1 71 unknown NOVEL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MOUSE</th>
<th>SwissProt ID</th>
<th>Peptide sequence</th>
<th>Spectral counts</th>
<th>P1 observed</th>
<th>P1 annotated</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BYHL MOUSE</td>
<td>A.PBALAQEG R C</td>
<td>4 17 unknown NOVEL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALDP MOUSE</td>
<td>L.SAATHLQAPKCHDPE V</td>
<td>13 19 19</td>
<td>ANNOTATED</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALDP MOUSE</td>
<td>A.AATLSAVAPKPHQPE V</td>
<td>2 20 19</td>
<td>ANNOTATED + TRIMMING</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PMML MOUSE</td>
<td>A.VISTSSMLPVQK # I</td>
<td>5 20 27</td>
<td>NEW TRANSIT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OTC MOUSE</td>
<td>H.PLQSSQKIQPSDVSQOLKJ R</td>
<td>2 24 32</td>
<td>NEW TRANSIT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CYS MOUSE</td>
<td>H.SASSHTNKHVSQSLNPQEGAR I</td>
<td>3 26 27</td>
<td>NEW TRANSIT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SARC MOUSE</td>
<td>L.ATEQPPRTESISKPVQG Y T</td>
<td>4 26 unknown NOVEL TRANSIT @ 30 + MIP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NDL86 MOUSE</td>
<td>A.HTAMQEVLPGVPSR Y T</td>
<td>1 29 29</td>
<td>ANNOTATED + TRIMMING</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP9A MOUSE</td>
<td>A.MMTAALQQGQESTPDGTGQ2QRRPR L</td>
<td>7 36 32</td>
<td>NEW TRANSIT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DSIA MOUSE</td>
<td>K.LASAKQSWPQTSVPPVDEHE F</td>
<td>5 42 43</td>
<td>NEW TRANSIT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DSHA MOUSE</td>
<td>A.SASHQSDATQYTPVVEHE F</td>
<td>9 42 43</td>
<td>ANNOTATED</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DSHA MOUSE</td>
<td>B.AVHDSTQNPVTQVDE F</td>
<td>1 44 43</td>
<td>ANNOTATED + TRIMMING</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATIA MOUSE</td>
<td>L.OHVTATKMSLEIBER I</td>
<td>17 43 43</td>
<td>ANNOTATED</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATPA MOUSE</td>
<td>Q.MTQTASMSLDEER I</td>
<td>11 44 43</td>
<td>ANNOTATED + TRIMMING</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PRD3 MOUSE</td>
<td>H.YPVQFQAPMTHKG</td>
<td>1 52 63</td>
<td>NEW TRANSIT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DQNS3 MOUSE</td>
<td>A.KEDYQVGLSVPR N</td>
<td>5 90 unknown NOVEL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HUMAN</th>
<th>SwissProt ID</th>
<th>Peptide sequence</th>
<th>Spectral counts</th>
<th>P1 observed</th>
<th>P1 annotated</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AKSS HUMAN</td>
<td>G.LHHKLAQVLAV R R</td>
<td>1 19 32</td>
<td>NEW TRANSIT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATFC HUMAN</td>
<td>P.FAKMLFPVRPVQVQEGDR Y</td>
<td>1 23 23</td>
<td>ANNOTATED</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATFC HUMAN</td>
<td>P.FAKMLFPVRPVQVQEGDR Y</td>
<td>2 24 23</td>
<td>TRANSIT + TRIMMING</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CSY HUMAN</td>
<td>H.ASASSHTNKHVSQSLNPQEGAR I</td>
<td>24 26 27</td>
<td>NEW TRANSIT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CHSA HUMAN</td>
<td>G.FQTVQVQVRK R A</td>
<td>1 32 42</td>
<td>NEW TRANSIT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ECHI HUMAN</td>
<td>L.TOSQOQRESOAVLAGEAPHDYSQVSRE I</td>
<td>1 33 unknown NOVEL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EFTH HUMAN</td>
<td>B.ASASSHTNKHVSQSLNPQEGAR I</td>
<td>2 41 45</td>
<td>NEW TRANSIT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GNNS3 HUMAN</td>
<td>A.KEDYQVGLSVPR N</td>
<td>13 90 unknown NOVEL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Figure 5  N-terminal trimming of human blood serum proteins

Blood serum contained several proteins with ragged N-termini. The longest derivative of each corresponds to the N-terminus following signal peptide removal, as annotated in Swiss-Prot, and the trimming is consistent with the activity of cell-associated aminopeptidases and DPPs. Guanidinated lysine, which gives a homoarginine derivative is abbreviated as K#. *The annotated signal peptide-cleavage site for CD5 antigen-like protein in Swiss-Prot.

---

**Table:**

<table>
<thead>
<tr>
<th>Protein</th>
<th>Event</th>
<th>Peptide</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpha-1-antichymotrypsin</td>
<td>signal removed</td>
<td>HPNSPLDEENLTQENQDR+G...</td>
</tr>
<tr>
<td></td>
<td>trimmed</td>
<td>NSPLDEENLTQENQDR+G...</td>
</tr>
<tr>
<td>Hemopexin</td>
<td>signal removed</td>
<td>TPPLPTSAHGNVAEGTK#PDPDVTER1C...</td>
</tr>
<tr>
<td></td>
<td>trimmed</td>
<td>PTPSAGHNVAEGTK#PDPDVTER1C...</td>
</tr>
<tr>
<td></td>
<td></td>
<td>HGNVAEGTK#PDPDVTER1C...</td>
</tr>
<tr>
<td>Alpha-2-HS-glycoprotein</td>
<td>signal removed</td>
<td>AHPGPGLYRQ...</td>
</tr>
<tr>
<td></td>
<td>trimmed</td>
<td>HPGPLYRQ...</td>
</tr>
<tr>
<td>Alpha-2-macroglobulin</td>
<td>signal removed</td>
<td>SVSGK#PQYMVLVPSLLHTE1T...</td>
</tr>
<tr>
<td></td>
<td>trimmed</td>
<td>K#PQYMVLVPSLLHTE1T...</td>
</tr>
<tr>
<td>Apolipoprotein-C-II</td>
<td>signal removed</td>
<td>GTQQPQQDEMPSTPTLTQV#1E...</td>
</tr>
<tr>
<td></td>
<td>trimmed</td>
<td>TQQPQQDEMPSTPTLTQV#1E...</td>
</tr>
<tr>
<td>CD5 antigen like</td>
<td>signal removed</td>
<td>ASPSVGRLVGLHRC1E...</td>
</tr>
<tr>
<td></td>
<td>trimmed</td>
<td>SPSVGRLVGLHRC1E...</td>
</tr>
<tr>
<td></td>
<td></td>
<td>*SGVRVLGGLHRC1E...</td>
</tr>
<tr>
<td>Fibrinogen alpha chain</td>
<td>signal removed</td>
<td>ADGSEQDFALAGGGVR#1G...</td>
</tr>
<tr>
<td></td>
<td>trimmed</td>
<td>DSGEQDFALAGGGVR#1G...</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SSEDFALAGGGVR#1G...</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EGDFALAGGGVR#1G...</td>
</tr>
<tr>
<td>Haptoglobin</td>
<td>signal removed</td>
<td>VDSGVNVTDAIADDGLP#PPEIAH#YVHS#1VY...</td>
</tr>
<tr>
<td></td>
<td>trimmed</td>
<td>GNDVTDAIADDGLP#PPEIAH#YVHS#1VY...</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DVTDAIADDGLP#PPEIAH#YVHS#1VY...</td>
</tr>
</tbody>
</table>

---

**Legend:**

- *The annotated signal peptide-cleavage site for CD5 antigen-like protein in Swiss-Prot.

---

biological inactivation. Consequently, DPPIV inhibitors are under clinical trials for therapeutic potential to enhance insulin secretion and overcome Type 2 diabetes (reviewed in [32]). We do not know whether the trimming of the N-termini by FAPα or DPPIV-like activity has a biological consequence, but we are struck by the potential that our discovery may have as a diagnostic biomarker of the efficacy of therapeutic treatment. As expected, cellular proteins are very low in the blood samples, whereas these samples are enriched in signal peptide-cleavage products corresponding to secreted proteins. Complete datasets are presented in Supplementary Tables 1–5.

Peptides that cannot be readily ascribed to the constitutive proteolytic events described above comprise a substantial portion of all datasets (Figure 3B). Because they are not annotated in protein sequence databases, these events probably originate from previously undocumented cleavages. They may represent biologically relevant peptides, trimming by aminopeptidases or results of natural protein turnover. However, it is unlikely that many of them are artefacts arising from sample preparation procedures. In our samples of mixtures of purified proteins, which model a complex proteome, we observed 81% of peptides that represent native protein N-termini (Figure 2D), and therefore > 80% of the unascribed peptides in our biological samples are likely to represent the results of in vivo proteolysis. With the exception of well-characterized pro-peptides (mainly in serum proteins, see Supplementary Table 5), internal cleavages in proteins are rarely detected by techniques used previously. The frequency of these unascribed cleavage events is unknown, but from our data, one can predict that internal proteolysis may be far more common than appreciated previously. However, in the absence of more direct data, we do not yet wish to ascribe these events to specific biologies.

**Characterizing proteolytic profiles using positive selection**

The approach we have described to identify N-termini of proteins is closely related in outcome to N-terminal analysis by Edman degradation. Indeed, the N-terminal sequences of 223 *E. coli* proteins were deduced following two-dimensional PAGE followed by Edman degradation of the excised protein spots [33]. However, this procedure would take months to complete, and is extremely reagent-intensive. Our methodology is complementary to two other MS-based techniques based on negative selection of modified tryptic peptides [34,35]. Our more direct strategy, in contrast, enriches N-termini by positive selection. Positive selection has certain advantages over negative selection: (i) we use it as a filter to simplify datasets, because only N-terminally labelled peptides are counted in our analysis, (ii) datasets are simplified further because all N-acetylated proteins are discarded, and (iii) if the biotin N-terminal probe is replaced by fluoresceinated amine-reactive dyes, it is possible to utilize the protocol to assess differences in constitutive (or even regulated) proteolysis by differential gel electrophoresis [36]. Although we have focused on constitutive proteolytic events, we realize that simply replacing the N-terminal label with one that is isotope-coded would allow us to quantify regulated proteolytic events, much as described by the COFRADIC (combined fractional diagonal chromatography)
Post-translational modification in health and disease.

Our approach addresses the limitations hindering the current understanding of proteolysis as a diagnostic tool. Profiling protease activity in vivo can implicate distinct proteases as pathological targets for therapeutics. In addition, specific substrates can be used as biomarkers for diagnosis and early detection of pathology. Our approach addresses the limitations hindering the current understanding of proteolysis as a post-transcriptional modification in health and disease.

This work was supported by NIH (National Institutes of Health) grant RR19752 and by a NIH Roadmap Initiative National Biotechnology Resource Center grant RR02643 for the Center on Proteolytic Pathways. M. E. is supported by the Swedish Society for Medical Research and the Wenner-Gren Foundation. We thank Andrei Osterman, Pavel Pevzner, Stephen Tanner and Salvatore Secchi for advice.

REFERENCES


