# Chapter 6
# Identification of Protease Substrates by Mass Spectrometry Approaches-2

**Anna Prudova, Ulrich auf dem Keller, and Christopher M. Overall**

**Abstract** Proteolysis is a major posttranslational modification of proteins with critical functional consequences to the protein, cell, and organism. The most effective way to monitor proteolytic events is to analyze the proteins directly. This chapter summarizes advantages and limitations of different mass spectrometry-based approaches for detection of proteolysis products. In general, liquid chromatography separation-based proteomics approaches are superior to 2D gel-based techniques and, in turn, quantitative proteomics have a significant advantage over label-free methods. Isotopic labeling of samples helps to identify substrates but fails to detect the exact cleavage site. Techniques that enrich for peptides containing the N-terminus of each protein provide a more relevant context for protease substrate discovery – they focus on the analysis of the neo-N-termini resulting from proteolysis. These techniques identify not only the substrates but also the prime side of the cleavage sites with a potential to extract further information of the protease sequence site specificity, thus setting the gold standard for the future of the degradomics field.

## Introduction

Having identified the components of the protease degradome gives the researcher a good picture of the proteolytic potential of the system. The ultimate information about the function of the identified degradome components and the resulting effects on the biological system in question, however, can be evaluated only by defining the direct action of the proteases, that is, the proteolytic modification of their substrates. This requires first, the identification of potential substrates in a given active

C.M. Overall
Centre for Blood Research, University of British Columbia, 4.401 Life Sciences Centre, 2350 Health Sciences Mall, Vancouver, BC V6T 1Z3, Canada, e-mail: chris.overall@ubc.ca

D. Edwards et al. (eds.), *The Cancer Degradome*.

degradome and second, the specific cleavage sites (auf dem Keller et al. 2007). In view of the overwhelming complexity of the protease web (Overall and Kleifeld 2006), exhaustive identification of the protease substrate repertoire with the corresponding cleavage sites can be a very daunting task.

Since proteolysis directly acts on proteins themselves as a posttranslational modification, the most obvious way to monitor this event is to analyze the proteins directly. Until recently, this was done by serial *in vitro* incubations of mostly recombinant substrate candidates with the protease under study and subsequent analysis by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE). Thereby, only one substrate candidate could be analyzed at a time under conditions far from *in vivo* physiology. For identification, the cleavage site fragments had to be isolated and subjected to time consuming and expensive chemical amino acid sequencing (Niall 1973). Even if the substrate under study was efficiently cleaved under *in vitro* conditions, it did not mean that the same would have happened in a complex biological system *in vivo*. Simply the fact that it *can* cleave does not mean that it *does* cleave under physiological conditions (Tam et al. 2004). Furthermore, substrate candidates had first to be identified in a separate experiment.

The development of mass spectrometry (MS) technology for the analysis of large biomolecules, particularly proteins, fundamentally changed the way proteins and their modifications are analyzed (Fenn et al. 1989). Now, the protein band of interest could be digested in the gel with a specific protease, such as trypsin, followed by protein identification based on its peptide mass fingerprint and database data analysis (Fenyo 2000). The invention of tandem mass spectrometry (MS/MS) for peptide analysis enabled further fragmentation of the protein-derived peptides, leading to the possibility to unambiguously identify proteins in more and more complex mixtures from their sequences (Wilm et al. 1996). In this chapter, we will summarize recent developments in the application of MS-based techniques for protease substrate discovery.

## *Two-Dimensional Gel Electrophoresis*

The combination of MS with two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) facilitated the simultaneous identification of hundreds of proteins in a complex biological mixture (Shevchenko et al. 1996). The high resolving power of 2D-PAGE and the development of various staining procedures to visualize these protein "spots" made it a popular method of choice for identifying protein abundance changes between two proteome samples.

Hwang et al. (2004) were the first to employ 2D-PAGE in combination with MS for protease substrate discovery. The authors incubated human plasma proteins with matrix metalloproteinase (MMP)-14 and compared the treated and untreated protein mixtures by 2D-PAGE. Subsequently, protein spots which disappeared or, in contrast, appeared in the MMP-14-treated sample due to the proteolytic cleavage were analyzed by peptide mass fingerprinting. This allowed the simultaneous identification of six known and nine new MMP-14 substrates in a complex

biological mixture. In a similar study, new substrates for caspase-3 were identified in human breast cancer cell lines by incubating the lysates of caspase-3-deficient MCF-7 cells with the recombinant protease (Lee et al. 2004b). Another study implemented the 2D-PAGE approach to characterize substrates for the intracellular serine protease-1 from *Bacillus subtilis* (Lee et al. 2004a). More recently, Major et al. (2006) used 2D-PAGE to assess the substrate range of the yeast mitochondrial matrix protease Pim1 *in vivo* using wild-type and Pim1delta strains. In addition to peptide mass fingerprinting, the authors extended the coverage of identified proteins by using MS/MS (Major et al. 2006).

One major limitation of conventional 2D-PAGE analyses is the reliability of protein identifications and the relatively high threshold for their quantification. Indeed, the two samples (with and without protease activity) have to be first electrophoresed in a very reproducible manner on two separate gels and then altered spots can be quantified by a densitometric image analysis. Thereby, the detection of only slight changes (which, however, can result in a strong biological phenotype) is still a major challenge. This drawback was recently offset by the introduction of two-dimensional difference gel electrophoresis (2D-DIGE), which involves labeling of samples with different fluorescent dyes (Cy3, Cy5) with subsequent analysis of two conditions on a single gel. This technique in combination with MS/MS was successfully used to identify new granzyme A and B substrates by incubating murine cell lysates with the corresponding recombinant proteases (Bredemeyer et al. 2004). A more recent study employed the 2D-DIGE technique to identify ADAMTS-1 substrates in a cell-based screen (Canals et al. 2006).

Two-dimensional gel electrophoresis techniques, however, remain limited in their sensitivity, making it very difficult to identify biologically relevant low-abundant proteins in complex proteomes. Furthermore, very large and small molecular weight proteins evade detection by this method due to their electrophoretic migration behavior on PAGE gels. Difficulties also exist with highly hydrophobic proteins, such as all membrane proteins, and those with extreme pI values. Finally, 2D-PAGE resolution is often insufficient, as shown by the MS-based detection of up to six proteins in one gel spot when analyzing yeast cell extract proteins (Gygi et al. 2000). For many proteases, substrates are cleaved by less than 10 residues Overall and Blobel 2007), often resulting in significant alteration of biological activity. For such substrates, 2D-PAGE lacks the resolution to detect these subtle but important changes to a protein substrate.

## *Shotgun Proteomics*

While 2D-PAGE is still widely used as a reliable, robust, and inexpensive method, researchers tried to enhance the number of identified proteins in a complex proteome by exploring alternative methods for the separation of proteins and their tryptic peptides before MS analysis. This led to the development of so-called "shotgun" proteomics, a solution-based approach which involves fractionation of

trypsin-generated peptides in two dimensions of liquid chromatography (LC) before MS/MS analysis, and thereby termed 2D-LC-MS/MS (Washburn et al. 2001). The 2D-LC most commonly involves a combination of strong cation exchange (SCX) and reverse-phase $C_{18}$ chromatography. The first dimension LC can be performed either off-line on a conventional high-performance liquid chromatography (HPLC) system or in-line with the nano LC-MS/MS system—a method also known as multidimensional protein identification technology (MudPIT) (Wolters et al. 2001). Combining this approach with prior protein fractionation further enhances the proteome coverage (Chen et al. 2006). This technique was recently used to obtain comprehensive proteome maps of different eukaryotic samples, including mammalian tissues (Kislinger et al. 2006; Brunner et al. 2007).

## *ICATs Quantitative Proteomics*

The above-described powerful solution-based techniques are aimed for a maximal number of proteins to be unambiguously identified in a complex biological sample. However, initially they lacked an easy possibility to also quantify the identified proteins, a prerequisite for the detection of proteolytic events. This problem was solved by the introduction of stable isotopic tags, with the most widely used being isotope-coded affinity tags (ICATs) (Gygi et al. 1999). ICATs comprise a trifunctional structure: (1) a cysteine-reactive group allowing for covalent binding to reduced cysteine residues of peptides; (2) a linker region with nine carbon atoms which can be synthetized either with "light" ($^{13}C_0$) or with "heavy" ($^{13}C_9$) isotopes; and (3) a cleavable biotin moiety as a handle to isolate ICAT-labeled peptides from the mixture. In this approach, protein samples to be compared are trypsin digested and the peptide mixtures are subsequently reacted with either the light or the heavy ICAT label. Hereby, the labels are incorporated into all cysteine-containing peptides. Afterward, both samples are combined and the labeled peptides positively selected via an avidin affinity column. Upon reductive elution, the peptides are subjected to 2D-LC-MS/MS analysis. Thereby, each peptide is represented in MS1 mode by a pair of peaks with a mass difference of 9 Da corresponding to the heavy and light ICAT labels. The areas of these peaks are integrated and used to determine the relative abundance of the peptide in both samples to be compared.

The first to employ ICAT labeling as a proteomic method for substrate discovery were Tam et al. (2004). Here the authors analyzed the substrate degradome of MT1-MMP in a cell-based system, where the protease and its substrates were present in the relevant context of a complete proteolytic pathway, including cofactors, binding proteins, inhibitors, and other modifying agents. Thereby, numerous novel bioactive substrates including connective tissue growth factor (CTGF), secreted leukocyte protease inhibitor (SLPI), and tumor necrosis factor $\alpha$ (TNF$\alpha$) as well as the death receptor-6 were identified. These findings underlined the important functions of MMPs as signaling proteases and pioneered the use of quantitative proteomics for protease substrate discovery in a complex biological system under physiological

conditions. Consistently, ICAT-based quantitative proteomics also revealed that the inhibition of MT1-MMP overexpressed in MDA-MB-231 cells using small molecule inhibitors, resulted in decreased shedding of cell-surface proteins with concomitant increase in the uncleaved protein levels on the plasma membrane (Butler and Overall 2007). The myriad of substrates so identified using MMP inhibitors underscore the complex problem of using MMPs as targets for disease intervention. Inhibition of proteolysis of many protease substrates may lead to a loss of significant biological functions which cannot be "buffered" by robust compensatory pathways and thus result in drug side effects.

In a more recent and also cell-based study, ICAT labeling was used to identify novel bioactive proteins as MMP-2 substrates and mechanistically dissect the angiogenic function of this MMP (Dean et al. 2007). Here, vascular endothelial growth factor (VEGF)-binding proteins (connective tissue growth factor, CTGF and heparin affin regulatory peptide, HARP) were found to be substrates of MMP-2, the cleavage of which mobilized VEGF and its angiogenic function.

## *iTRAQ Quantitative Proteomics*

While ICAT labeling was successfully used to identify novel protease substrates, the shortcoming of this method is that only cysteine-containing peptides can be analyzed, thus limiting proteome coverage to ~93% of proteins. To overcome this limitation, a new generation of labels for quantitative proteomics that is named iTRAQ (isobaric tag for relative and absolute quantification) (Choe et al. 2005) can be used. The iTRAQ consists of a group that is reactive toward primary amino groups, the linker region, and a reporter group that gives rise to a highly diagnostic low-mass reporter ion upon fragmentation of a tagged peptide in MS/MS mode. The currently available set of iTRAQ reagents contains four different isobaric variants which have the same total mass but are different in the corresponding masses of their linker and reporter regions. Thus, when four different samples containing equal amounts of the same peptide are labeled with the four variants and then combined at a one-to-one ratio, the MS mode will show a single peak representing this peptide. However, upon fragmentation, four different reporter ion peaks will appear in the 114–117 m/z spectra region, with the areas of these peaks corresponding to the peptide amounts in each of the original four samples. Therefore, labeling of a peptide N-terminus and/or lysine residues allows peptide sequence information to be obtained together with its relative quantification in MS/MS mode without doubling the spectra complexity in the MS mode (as observed with any other nonisobaric labeling techniques, such as ICAT, SILAC, acetylation, or reductive dimethylation).

The iTRAQ-based labeling in application to protease substrate discovery was first employed by Dean et al. (2007) (Dean and Overall 2007) using MMP-2 as a model secreted protease. In this study, the authors examined the conditioned medium from *Mmp-2–/–* murine fibroblasts transfected with active MMP-2 or its

inactive mutant form (so as to present a "naive" proteome that had not been exposed to the protease). The secreted and shed cell-surface proteins in the serum-free medium were collected and then denatured, alkylated, and digested with trypsin. Following peptide labeling with two different iTRAQ reagents, the samples were mixed in one-to-one ratio and analyzed by 2D-LC-MS/MS. Comparison of the relative abundances of each peptide identified in the two samples (derived from the corresponding iTRAQ ratios for each peptide) identified proteins that were degraded (and therefore represented by low iTRAQ ratios) and the proteins that were shed from the cell surface into the medium (and therefore presented by high iTRAQ ratios). The peptides that belonged to the proteins (or their regions) that were unaffected by MMP-2 cleavages exhibited iTRAQ ratios of ~1. In further experiments with four different iTRAQ labels, different time points were analyzed in a multiplex approach. On the basis of iTRAQ ratios for known previously reported substrates of MMP-2 observed in their analyte mixture, the authors established an iTRAQ ratio cutoff of fourfold. Thus, proteins with ratios less than 0.25 or greater than 4 were considered potential substrates, with some of them tested and confirmed in *in vitro* cleavage assays using purified proteins and recombinant MMP-2. In total, the analysis yielded known substrates (thus validating this approach) and also identified many previously undescribed substrates. In addition, mapping of the peptides with altered iTRAQ ratios compared to those peptides showing no change to the corresponding protein sequences, together with the results of *in vitro* cleavage assays, allowed identification of the region and even the exact site of the cleavage event in some proteins.

To summarize, the above-described solution-based quantitative proteomics approaches have a significant advantage over label-free methods in as much they identify substrates and may highlight a specific region of the molecule which is cleaved based on the differences in label ratios between specific peptides. However, the success of this strategy largely depends on the completeness of sequence coverage of the protein in question. In reality, the exact peptide(s) showing altered label ratios often remain undetected due to the complexity of a proteome mixture in general and its wide dynamic range, resulting in undersampling of lower-abundance proteins. Therefore, identifying potential substrates in a complex proteome still heavily resembles looking for a proverbial needle in a haystack.

ICAT-based substrate discovery methods somewhat address this issue since the proteomic mixture is simplified as it is being enriched for cysteine-containing peptides via label-dependent affinity pullout. As a result, the mixture contains fewer peptides, thus improving statistical chances for the identification of lower-abundance proteins. However, the same chemical bias excludes from analysis all the proteins without any cysteine (7% of all proteins) and limits the coverage of proteins containing only one cysteine residue in their sequence (35% of proteins), thus limiting the utility of this strategy for substrate discovery (Dean and Overall 2007). Consistent with this notion, comparison of iTRAQ and ICAT-based strategies within the same cellular context resulted in identification of higher numbers of total identified proteins (9-fold), known substrates (8-fold), protease inhibitors (4-fold), and proteases (31-fold) in iTRAQ-labeled samples (Dean and Overall 2007). Therefore, while an enrichment of proteomic

samples for cysteine-containing peptides by ICAT may offer an advantage in analysis of cysteine-rich potential substrates (e.g., many extracellular matrix proteins and cytokines), it is not beneficial for a system-wide unbiased substrate discovery. On the contrary, enrichment for N-terminal portion(s) of each protein provides a far superior context for protease substrate discovery, as every act of proteolytic processing results in a neo N-terminus representing the prime side of the cleavage site. Thus, by analyzing the neo N-terminal peptide resulting from the proteolytic cleavage one can identify not only the candidate substrate but also the exact cleavage site with a potential to extract further information of the protease sequence site specificity (Overall and Dean 2006). The methods for selective N-termini recovery, their MS and data analysis, and the application to protease substrate identification will be discussed in the next section in the chronological order in which each method was first reported.

## *N-terminal Enrichment Methods for Protease Substrate Discovery*

There have been a number of strategies reported that aim to selectively isolate protein N-terminal peptides for gel-free proteomic characterization of proteolysis. These include (1) differential N-terminal labeling to modify peptide hydrophobicity before diagonal chromatography (Gevaert et al. 2003); (2) N-terminal acetylation, then trypsin digestion, biotinylation, and affinity pullout of the internal peptides (McDonald et al. 2005); and (3) N-terminal-specific protein biotinylation with consequent affinity enrichment (Timmer et al. 2007).

The very first study using N-terminal enrichment to examine proteolytic processing of proteins, utilized an elegantly designed combined fractional diagonal chromatography (COFRADIC) approach (Gevaert et al. 2003). In this strategy, proteins are first acetylated with acetic anhydride at their N-terminal and lysine amino groups, digested with trypsin, and separated by reversed-phase HPLC. Internal peptides in each of the resulting 12 fractions are then chemically modified by 2,4,6-trinitrobenzenesulfonic acid (TNBS) at their N-termini to form very hydrophobic trinitrophenyl (TNP) derivatives, followed by a second reversed-phase fractionation. Because of their higher hydrophobicity, TNP-peptides are bound stronger by the column, elute in later fractions, and can be discarded. The acetylated peptides representing the original N-terminal portion of each protein are eluted in earlier fractions, collected, and then subjected to MS analysis. This secondary reversed-phase fractionation is performed for each of the 12 primary fractions, resulting in a total of 96 fractions and hence in 96 LC-MS/MS analyses per average experiment. The COFRADIC technique was first tested using human thrombocytes-derived proteomes, where 264 proteins were identified from 305 different peptides, with one peptide per protein, on average. About 10% of identified peptides were contaminating internal peptides that start with proline and pyroglutamate residues, and therefore have low or no reactivity toward 2,4, 6-trinitrobenzenesulfonic acid. In addition, another 74 internal tryptic peptides

**Table 6.1** Summary of current techniques for protease substrate discovery by enrichment of N-terminal peptides

| Technique | COFRADIC | McDonald et al. | Timmer et al. | Enoksson et al. |
|---|---|---|---|---|
| Major steps | 1. Chemical acetylation of protein N-termini and lysines<br>2. First round RP LC separation<br>3. Trypsin digest with concomitant C-terminal isotopic labeling of generated peptides<br>4. TNBS labeling of neo-N-termini of tryptic peptides<br>5. Second round of RP LC separation; internal peptides are discarded<br>6. LC-MS/MS | 1. Chemical acetylation of protein N-termini and lysines<br>2. Trypsin digest<br>3. Biotin coupling and pullout of neo-peptides resulting from trypsin digest or their pullout using NHS-activated Sepharose<br>4. LC-MS/MS | 1. Lysine gyanidinylation<br>2. Biotin coupling of protein N-termini<br>3. Trypsin digest<br>4. Avidin affinity enrichment of biotinylated peptides<br>5. LC-MS/MS | 1. Lysine gyanidinylation<br>2. iTRAQ labeling of protein N-termini<br>3. Trypsin digest<br>4. MALDI MS/MS |
| Method of N-terminal selection/ enrichment | Negative selection of all N-terminal peptides (including *in vivo* modified, e.g. acetylated, protein N-termini) | Negative selection of all N-terminal peptides (including *in vivo* modified, e.g. acetylated, protein N-termini) | Positive selection of peptides with unblocked amino-termini; *in vivo* modified peptides (e.g. acetylated, myristoylated etc.) are excluded from the analyte mix | Virtual enrichment for protein N-termini during the first "survey MS/MS", followed by their sequencing and quantification in the second MS/MS |

| | | | | |
|---|---|---|---|---|
| Quantification | Quantification in MS mode | Not reported; quantification in MS mode is possible, if stable-isotope labeled acetic anhydrate is used for the acetylation | Not reported; quantification in MS mode is possible, if stable-isotope labeled biotin is used | Quantification in MS/MS mode |
| Comments for consideration | – Powerful, but labor- and equipment-intense technique<br>– Does not distinguish between chemically and *in vivo* acetylated protein N-termini in a single experiment | – Fast and robust<br>– Has a potential to distinguish between chemically and *in vivo* acetylated protein N-termini in a single experiment if isotope-labeled acetic anhydrate is used | – Allows analysis of only unblocked protein N-termini and not *in vivo* modified proteins | – Limited proteome coverage due to ion suppression<br>– Has a potential for multiplex analysis using 4-plex iTRAQ reagents |

were identified, suggesting incomplete TNP-modification, regardless of the first N-terminal amino acid. Based on the theoretical *in silico* digestion of the human proteome yielding 17.5 internal peptides per N-terminal peptide, the authors' results demonstrate a significant N-terminal enrichment. To evaluate the true efficacy of this enrichment technique, it should also be experimentally shown how many N-terminal peptides are omitted from the analysis due to incomplete acetylation, overlapping elution with internal TNP-modified peptides, or sample loss during multiple handling steps, that is, chemical modification and cleanup steps as well as the multiple LC analyses.

Since the COFRADIC technique is based on the MS analysis of the initially chemically acetylated and therefore retained peptides, it does not allow for distinguishing these from the protein N-termini which are retained due to their acetylation *in vivo*. While *in vivo* N-terminal acetylation is absent in prokaryotes, it is estimated to occur in up to 80% of eukaryotic proteins (Polevoda and Sherman 2003). To differentiate between *in vivo* and *in vitro* acetylation, the authors performed two sets of COFRADIC experiments for the same sample, with and without the first acetylation step in the workflow. Omitting the acetylation reaction, results in the retention and analysis of the N-terminal peptides of *in vivo* N-terminally blocked proteins. However, if these N-terminal peptides contain lysine residues, their side-chain amino groups will be TNP-modified and the peptides excluded from the analysis during the secondary reversed-phase separation. According to the authors, this chemical bias results in the loss of approximately half of all *in vivo* acetylated proteins.

Among the identified protein N-termini, the authors observed the following posttranslational proteolytic modifications: (1) removal of the initiator methionine; (2) propeptide or signal peptide removal; and (3) internal cleavages. For example, the study uncovered a previously undescribed truncated form of actin starting at amino acid 29. In some instances, a proteolytic processing predicted by homology with other proteins has been verified and corrected. For example, dihydroorotate dehydrogenase was found to start at residue 28 rather than predicted position 11. While the study demonstrates the utility of the N-terminal enrichment approach to describe proteolytic processes in a biological sample, it does not, however, allow for strict quantification of these events.

To address this issue, the original COFRADIC technique was slightly modified to introduce the quantitative differential aspect (Van Damme et al. 2005). To incorporate the label, the samples are digested with trypsin in the presence of water with $^{18}O$ isotope. Thus, trypsin-catalyzed incorporation of two $^{18}O$ atoms at the C-terminus of the newly cleaved-off peptide results in a 4 Da mass difference compared to the same peptide created in the presence of light $^{16}O$ water (Staes et al. 2004). In this approach, two samples representing protease(s) treated and untreated proteomes are differentially labeled during the digest, then mixed in a one-to-one ratio (total peptide amount), COFRADIC-sorted, and MS analyzed. The N-terminal peptides equally present in both samples will be represented by a 4 Da-different doublet in the first dimension of MS analysis (MS1). Given their equal representation, the area under the corresponding peaks should be the same with the ratio of ~1.

However, if the parent N-terminus is cleaved by a protease, then its area will decrease, resulting in a ratio between protease treated/untreated samples less than 1. In addition, a neo N-terminus resulting from proteolytic cleavage will be generated and represented by a singlet in the MS1 lacking its 4 Da different counterpart due to its presence only in the protease-treated sample. The ratio of such peptides will be greater than 1.

The quantitative COFRADIC approach was applied to describe apoptosis-induced proteolytic events in anti-Fas antibody-treated versus untreated human Jurkat T lymphocytes (Van Damme et al. 2005). In addition to characterizing apoptosis-independent N-terminal processing (i.e., the baseline proteolytic activity of initiator methionine removal, signal peptide trimming, etc.), the analysis identified 93 apoptosis-induced cleavage sites in 71 proteins among 1,834 proteins detected in total. Consistent with the previously well-studied experimental model, most observed cleavages were found to be at the caspase consensus sites. The few cleavages showing other than aspartate P1 specificities represent either noncanonical caspase cleavages, additional protease classes activated during apoptosis, or false positives. To validate the findings, a few caspase-specific cleavage sites were investigated in *in vitro* cleavage assays, where recombinant caspases were used to cleave synthetic peptides harboring the identified candidate cleavage sites. Also, processing of four canonical caspase substrates was successfully detected in the activated Jurkat cell lysates by immunoblotting with the corresponding specific antibodies. However, processing of proteins with caspase unspecific cleavages was not tested by Western blotting, and therefore the possibility that such peptides are due to false-positive identifications was not addressed.

More recently, the same group used the quantitative COFRADIC approach to identify the substrates of an apoptosis-activated mitochondrial serine protease high-temperature requirement protein A2 (HtrA2/Omi) (Vande Walle et al. 2007). Recombinant wild-type HtrA2/Omi or its catalytically inactive S306A mutant was incubated with Jurkat T cell lysates which were then differentially labeled and N-terminally sorted. The analysis yielded 1,162 total protein identifications (from 1,964 peptides) and determined 50 cleavage sites in 15 proteins, represented mostly by cytoskeletal proteins. Several cleavage events were validated by specific immunoblotting in treated Jurkat T cells and by *in vitro* cleavage assays with recombinant HtrA2/Omi and *in vitro* translated susbstrates. Analysis of the 50 detected cleavage sites indicated a HtrA2/Omi preference to cleave after an aliphatic residue at P4 with the four positions C-terminal to the cleavage site being most commonly occupied by small or hydrophobic residues.

To summarize, COFRADIC is a powerful approach that allows for a significant sample simplification and N-terminal peptide enrichment, and therefore enables effective identification and quantification of proteolytic processing in a biological sample. However, the experimental design does not allow the analysis of *in vivo* modified (i.e., acetylated) and unblocked protein N-termini in a single experiment, thus somewhat limiting its use for characterization of N-terminal posttranslational modifications in eukaryotes. In addition, the above-discussed sequence bias of the labeling efficacy and the absence of a clear cutoff between N-terminal and internal

TNP-modified peptides during chromatographic separation limit the number of peptides being analyzed and reduce the proteome coverage. With 2 chemical labeling steps, 2 rounds of HPLC separation, and 96 LC-MS/MS runs per average experiment, this technique is rather time-, equipment-, and labor-intense and so far has not been adopted by a broad scientific community.

A different N-terminal enrichment approach designed by the Beynon laboratory also utilizes protein acetylation as the first step, followed by tryptic digestion (McDonald et al. 2005). Newly formed unblocked internal tryptic peptides are then coupled with *N*-hydroxysuccinimide (NHS), ester-derivative of biotin, retained by immobilized streptavidin and discarded. The remaining mixture consisting of protein N-terminal peptide(s) (naturally blocked or chemically acetylated) is then analyzed to yield information on the proteolytic processes in the sample. To distinguish between and quantify *in vivo* and chemically acetylated proteins, the authors suggest using stable isotope-labeled ($^{3}$H) acetic anhydrate that would result in a 3 Da mass shift in MS1. By analogy with COFRADIC, labeling with $^{18}$O at the C-terminus during trypsin digest (Van Damme et al. 2005) or with acetic anhydrate at the N-terminus has the potential to be utilized for comparative quantification of protease activity between two samples. This technique was tested on the soluble protein fraction of mouse skeletal muscle and a more complex mixture of soluble proteins from mouse liver (McDonald et al. 2005). As a proof of concept, qualitative comparison of matrix assisted laser desorption/ionization–time of flight (MALDI–TOF) spectra of an unfractionated peptide mixture with or without the N-terminal enrichment step indicated a significant spectra simplification, and enabled assignment of the highest intensity signals to true N-terminal peptides in the enriched samples. In contrast, without N-terminal selection the complexity of the sample prevented identification of any N-terminal peptides. LC-MS/MS analysis of N-terminally enriched mouse liver peptides yielded information on the N-terminal processing, such as removal of initiator methionine, loss of a signal peptide or propeptide, either known previously or inferred.

In order to yield more suitable sets of analytes and to increase the N-terminal sequence coverage of any given proteome, the authors suggest performing two parallel digests with proteases of different specificities. To test this hypothesis, *in silico* digest of 8,000 mouse liver proteins with trypsin and/or endopeptidase GluC was filtered to remove the peptides smaller than 500 Da and larger than 5,000 Da which are not suitable for MS analysis. The analysis of the remaining peptides indicated that tryptic or GluC digests alone would result in a respective 50% and 60% unambiguous proteome coverage (using the mass spectrometer with 20 ppm accuracy). The coverage reaches 80% when the sample is digested by the two proteases in parallel. The value can be improved even further when using higher accuracy MS (almost 90% coverage with two digests and 1 ppm instrument accuracy), further underlining the feasibility of proteome characterization via protein identification by a single peptide.

This time-efficient protocol was further improved to decrease the number of steps and therefore to increase sample recovery. The biotinylation step is now excluded and replaced by a direct coupling and removal of internal peptides via a

commercially available amino-reactive immobilized reagent, NHS-activated Sepharose (McDonald and Beynon 2006). Thus, in the final protocol the internal peptides can be removed directly after the digest, with the flow-through being analyzed without further treatments. The protocol was tested using LC-MS/MS with soluble proteins from *Escherichia coli* and identified ~300 proteins by their N-termini with relatively few internal peptides. As the authors suggested, the proteome coverage might be further increased by employing additional fractionation steps before LC-MS/MS. In principle, this approach could be used to identify substrates of a specific protease, but this has yet to be reported.

Yet another interesting approach for N-terminal enrichment has been demonstrated by Timmer et al. (2007). Here, the proteins are first denatured, reduced, and alkylated and then the amino groups of lysine side chains are protected by lysine-specific guanidinylation. In the next step, the N-termini of proteins are selectively labeled with NHS-biotin. Following tryptic digest, biotin-labeled N-terminal peptides are positively selected by immobilized streptavidin. The captured peptides are then reductively cleaved off from the column using dithiothreitol (DTT) and analyzed by LC-MS/MS. This approach was tested on *E. coli*, yeast, mouse and human cell lines, and serum proteomes to profile constitutive proteolytic events in these samples. To increase the confidence and coverage, the samples were digested with both GluC and trypsin and run three times using dynamic exclusion criteria. Consistent with previous reports, multiple runs yielded a 50–70% overlap between the same sample analyzed by MS several times. The coverage of the proteome ranges from ~350 peptides in serum to ~500 peptides in *E. coli*, yeast, and mouse tissues and ~1,000 peptides in 293A human embryonic kidney cell line. These values are comparable to the number of peptides identified by McDonald and colleagues in *E. coli* (McDonald and Beynon 2006), but are lower than the ones reported using the COFRADIC technique on a similar cell-line model (Van Damme et al. 2005). This can be due, at least partially, to a higher degree of sample fractionation before LC-MS/MS analysis in COFRADIC or might be inherent to the technique. However, in contrast to COFRADIC where the majority of the identified peptides represented N-terminal peptides (Van Damme et al. 2005) (Gevaert et al. 2003), Timmer et al. (2007) reported that many of the identified peptides belong to internal sequences and can not be ascribed to any known proteolytic modifications (e.g., initiator methionine removal, propeptide removal). With the exception of *E. coli*, where such unascribed peptides constitute less then 50% of total peptides identified, the rest of the samples exhibit a broader range—from 70% of unascribed peptides in yeast and mouse tissues to 80–90% in human cell lines and serum. A possible explanation for such a high percentage of proteolysis in the samples could be a high general protease activity induced by cell disruption that was not completely inhibited before sample denaturation. Using this technique, the authors observed and characterized methionine aminopeptidase activity and removal of signal peptides as well as N-terminal trimming of proteins in serum samples.

In contrast to COFRADIC and McDonald et al., N-terminal enrichment strategies where naturally acetylated N-termini are automatically included in the analyte mix, the present protocol results in retention of only the N-termini of unblocked proteins.

Thus, such selection excludes from the analysis up to 80% of total natural N-termini in eukaryotic samples. In contrast, retaining naturally modified (i.e., acetylated) N-termini helps to curb sample loss and has an additional advantage of higher confidence protein identifications, as it is then based on a positionally anchored original N-terminal peptide (McDonald et al. 2005).

Following from the design of the present method, the degree of N-terminal enrichment (in terms of contamination with internal peptides) of the final analyte largely depends on the efficacy/completeness of protein lysine residue guanidinylation in the beginning of the protocol and on the absence of side reactions during biotin coupling. As noted by the authors, incomplete and/or side reactions will lead to biotin coupling to lysine and/or serine, threonine, or histidine residues and will result in contamination of the final analyte with internal peptides, and therefore must be strictly controlled for. It should be noted that such spectra pollution will decrease true N-termini coverage and further complicate data analysis leading to a higher rate of false-positive identifications.

While the current protocol by Timmer et al. (2007) does not readily allow for quantification of proteolytic events in the analyzed sample, the authors propose that it can be modified to incorporate stable isotope-labeled biotin for N-terminal labeling, or to include C-terminal trypsin-dependent $^{18}O$ exchange as seen in COFRADIC. Both such potential modifications would result in a mass shift in MS1 and enable relative quantification of the same peptide in two samples.

A different strategy for coping with overwhelming sample complexity has been offered by Enoksson et al. (2007). In this N-terminal pseudo-enrichment approach, the sample is not treated to physically remove all internal peptides, but is left complex, and then filtering for N-terminal peptides is applied at the sample analysis stage on the MALDI–TOF/TOF. Briefly, the proteins are first lysine-specific guanidinylated to block their reactive side-chain amino groups and then labeled with the iTRAQ reagent at the protein N-terminus. Following mixing of two samples (that have been treated or not with a protease) at one-to-one ratio, the sample is digested by trypsin, chromatographically separated and analyzed on a MALDI–TOF/TOF instrument.

This strategy takes advantage of the fact that in contrast to LC-coupled mass spectrometers with electrospray ionization, the MALDI instruments allow for multiple scans/analysis of the same peptide(s) in the sample. Thus, during the first low-energy scan the sample is surveyed for the presence of peptides with the iTRAQ reporter ion in the spectra to form a data-dependent inclusion list for the second scan. The first low-energy scan results in low sample consumption, and its limited fragmentation is not sufficient for peptide sequencing but is suitable for indicating diagnostic iTRAQ reporter ions. Therefore, in the second higher-energy scan only the previously selected peptides with iTRAQ tag will be fragmented for high confidence identification and quantification. In this workflow, the iTRAQ-bearing peptides represent original N-termini of the proteins as well as protease-generated neo-N-termini, with iTRAQ ratios allowing discrimination between the two. Thus, the original N-termini equally present in both samples will have iTRAQ ratio of ~1, while protease-dependent neo N-termini will have a singleton or greater than 1 iTRAQ signal ratio.

This approach was first tested on a mixture of seven purified *E. coli* proteins containing putative caspase cleavage sites, which were treated with wild-type caspase-3 or its catalytically inactive mutant C285A. A total of 12 cleavage sites in 6 proteins were identified in the MS analysis compared to 5 cleavage fragments identified by SDS-PAGE and 8–10 indicated by Western blotting analysis. Further, the method was tested on a cell-free apoptosis model using HEK293 hypotonic extracts, where 20 different cleavage sites were identified, mostly previously undescribed but with canonical caspase cleavage site specificity. Cleavage of one identified substrate, actin, was further confirmed by Western blotting.

As discussed in previous sections of this chapter, iTRAQ labeling does not result in a mass shift in MS1 and therefore does not lead to doubling of the sample complexity. A very serious limitation of this technique is in the fact that identification of iTRAQ-bearing peptides in the first "survey MS/MS" will be limited by ion suppression due to the physical presence of many internal peptides, that is, many N-terminal peptides simply will not be ionized and detected under such conditions. This notion is supported by observations of McDonald et al. made under very similar conditions using a MALDI–TOF instrument to analyze fractionated mouse liver proteins (McDonald et al. 2005). Thus, McDonald et al. reported that high sample complexity prevented N-termini detection when internal peptides are physically present in the sample. Consistent with ion suppression being a limiting factor, Enoksson et al. (2007) detected only 20 cleavage sites compared to 93 cleavages identified by the COFRADIC method in a similar cell-based apoptosis model (Van Damme et al. 2005). However, it should be noted that the use of a different cell line might be a contributing factor as well. Also, in contrast to all the other above-described techniques, the latter method can only be used with MALDI mass spectrometers. Thus, the virtual N-terminal enrichment technique of Enoksson et al. is a suitable method for detection and quantification of protease cleavage sites in defined protein sets of test substrates or in less complex proteomes.

As a future direction, another technique for substrate discovery is in development by the authors' laboratory termed "terminal amine isotope labeling of substrates", TAILS (Kleifeld et al., manuscript in preparation). In this approach, the sample is enriched for N-terminal peptides of each protein, thus allowing for neo-N-termini resulting from proteolysis to be identified with higher probability (sample complexity reduction) and confidence (positional information). Specifically, the proteomes of two samples containing active and inactive protease (control) are first reduced, alkylated, and labeled with amino-reactive isotope-containing reagents (such as formaldehyde or iTRAQ). Such labeling selectively modifies lysine residues and protein N-termini. Following trypsin digestion, the newly created and therefore unblocked internal peptides are selectively removed by an amine scavenging polymer or beads. The remaining N-terminome fraction is then analyzed by MS/MS. For example, when the iTRAQ reagents 114 and 115 are used to differentially label samples containing active and inactive protease, respectively, this results in protease-cleaved neo-N-termini identified as spectra with singletons (i.e., containing only one isotopic signature, 114). These are distinct from noncleaved peptides, which exhibit both isotopic signatures, 114 and 115, at 1:1 ratio. Therefore, MS/MS sequencing identifies protease substrates and defines

the sequence of the cleavage site, with iTRAQ labeling allowing for an estimate of how much a particular substrate is processed.

To summarize, this approach utilizes the power of multiplex isotope labeling and negative selection of internal peptides by use of a highly soluble, highly derivatized polymer. Thus, it enriches for protease cleavage neo-peptides and natural N-termini (acetylated and nonacetylated), allowing determination of both protease substrates and their cleavage sites, as well as annotation of N-terminal posttranslational proteome processing in a single experiment.

## Conclusions

A number of reported techniques that are different in their labeling, enrichment and quantification strategies all aim at proteome simplification and N-terminal enrichment in order to enable more efficient protease substrate identifications. When selecting a suitable MS technique for determining the substrate(s) of a particular protease, one may choose to consult the following checklist: (1) proteome coverage achieved by the method; in general, the higher it is the better is the chance for finding the substrate(s); (2) quantification aspect; quantification always strengthens qualitative findings and allows for subtraction of basal proteolysis. MS/MS-based quantification methods, such as those based on iTRAQ, offer some advantages, including a possibility for the simultaneous analysis of up to eight samples in one experiment; (3) reagent availability and level of expertise required to perform the protocol and to analyze the data; (4) instrumentation, time, labor, and cost efficiency.

**Acknowledgments** Christopher M. Overall was supported by a Canada Research Chair in Metalloproteinase Proteomics and Systems Biology, by research grants from the National Cancer Institute of Canada (with funds raised by the Canadian Cancer Association) and from the Canadian Breast Cancer Research Alliance Special Program Grant on Metastasis, and by a Centre Grant from the Michael Smith Research Foundation. Ulrich auf dem Keller was supported by the Deutsche Forschungsgemeinschaft Germany. Anna Prudova was supported by a postdoctoral fellowship from the Centre for Blood Research, UBC.

## References

auf dem Keller, U., Doucet, A., et al. 2007. “Protease research in the era of systems biology.” Biol Chem 388: 1159–62.

Bredemeyer, A. J., Lewis, R. M., et al. 2004. “A proteomic approach for the discovery of protease substrates.” Proc Natl Acad Sci USA 101: 11785–90.

Brunner, E., Ahrens, C. H., et al. 2007. “A high-quality catalog of the Drosophila melanogaster proteome.” Nat Biotechnol 25: 576–83.

Butler, G. S. and Overall, C. M. 2007. “Proteomic validation of protease drug targets: Pharmacoproteomics of matrix metalloproteinase inhibitor drugs using isotope-coded affinity tag labelling and tandem mass spectrometry.” Curr Pharm Des 13: 263–70.

Canals, F., Colome, N., et al. 2006. “Identification of substrates of the extracellular protease ADAMTS1 by DIGE proteomic analysis.” Proteomics 6(Suppl. 1): S28–35.

Chen, E. I., Hewel, J., et al. 2006. "Large scale protein profiling by combination of protein fractionation and multidimensional protein identification technology (MudPIT)." Mol Cell Proteomics 5: 53–6.

Choe, L. H., Aggarwal, K., et al. 2005. "A comparison of the consistency of proteome quantitation using two-dimensional electrophoresis and shotgun isobaric tagging in Escherichia coli cells." Electrophoresis 26: 2437–49.

Dean, R. A. and Overall, C. M. 2007. "Proteomics discovery of metalloproteinase substrates in the cellular context by iTRAQ labeling reveals a diverse MMP-2 substrate degradome." Mol Cell Proteomics 6: 611–23.

Dean, R. A., Butler, G. S., et al. 2007. "Identification of candidate angiogenic inhibitors processed by matrix metalloproteinase 2 (MMP-2) in cell-based proteomic screens: Disruption of vascular endothelial growth factor (VEGF)/heparin affin regulatory peptide (pleiotrophin) and VEGF/Connective tissue growth factor angiogenic inhibitory complexes by MMP-2 proteolysis." Mol Cell Biol 27: 8454–65.

Enoksson, M., Li, J., et al. 2007. "Identification of proteolytic cleavage sites by quantitative proteomics." J Proteome Res 6: 2850–8.

Fenn, J. B., Mann, M., et al. 1989. "Electrospray ionization for mass spectrometry of large biomolecules." Science 246: 64–71.

Fenyo, D. 2000. "Identifying the proteome: Software tools." Curr Opin Biotechnol 11: 391–5.

Gevaert, K., Goethals, M., et al. 2003. "Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides." Nat Biotechnol 21: 566–9.

Gygi, S. P., Rist, B., et al. 1999. "Quantitative analysis of complex protein mixtures using isotope-coded affinity tags." Nat Biotechnol 17: 994–9.

Gygi, S. P., Corthals, G. L., et al. 2000. "Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology." Proc Natl Acad Sci USA 97: 9390–5.

Hwang, I. K., Park, S. M., et al. 2004. "A proteomic approach to identify substrates of matrix metalloproteinase-14 in human plasma." Biochim Biophys Acta 1702: 79–87.

Kislinger, T., Cox, B., et al. 2006. "Global survey of organ and organelle protein expression in mouse: Combined proteomic and transcriptomic profiling." Cell 125: 173–86.

Lee, A. Y., Goo Park, S., et al. 2004a. "Identification of the degradome of Isp-1, a major intracellular serine protease of Bacillus subtilis, by two-dimensional gel electrophoresis and matrix-assisted laser desorption/ionization-time of flight analysis." Proteomics 4: 3437–45.

Lee, A. Y., Park, B. C., et al. 2004b. "Identification of caspase-3 degradome by two-dimensional gel electrophoresis and matrix-assisted laser desorption/ionization-time of flight analysis." Proteomics 4: 3429–36.

Major, T., von Janowsky, B., et al. 2006. "Proteomic analysis of mitochondrial protein turnover: Identification of novel substrate proteins of the matrix protease pim1." Mol Cell Biol 26: 762–76.

McDonald, L. and Beynon, R. J. 2006. "Positional proteomics: Preparation of amino-terminal peptides as a strategy for proteome simplification and characterization." Nat Protoc 1: 1790–8.

McDonald, L., Robertson, D. H., et al. 2005. "Positional proteomics: Selective recovery and analysis of N-terminal proteolytic peptides." Nat Methods 2: 955–7.

Niall, H. D. 1973. "Automated Edman degradation: The protein sequenator." Methods Enzymol 27: 942–1010.

Overall, C. M. and Blobel, C. P. 2007. "In search of partners: linking extracellular proteases to substrates." Nat Rev Mol Cell Biol 8: 245–57.

Overall, C. M. and Dean, R. A. 2006. "Degradomics: Systems biology of the protease web. Pleiotropic roles of MMPs in cancer." Cancer Metastasis Rev 25: 69–75.

Overall, C. M. and Kleifeld, O. 2006. Validating matrix metalloproteinases as drug targets and anti-targets for cancer therapy." Nat Rev Cancer 6: 227–39.

Polevoda, B. and Sherman, F. 2003. "N-terminal acetyltransferases and sequence requirements for N-terminal acetylation of eukaryotic proteins." J Mol Biol 325: 595–622.

Shevchenko, A., Jensen, O. N., et al. 1996. "Linking genome and proteome by mass spectrometry: Large-scale identification of yeast proteins from two dimensional gels." Proc Natl Acad Sci USA 93: 14440–5.

Staes, A., Demol, H., et al. 2004. "Global differential non-gel proteomics by quantitative and stable labeling of tryptic peptides with oxygen-18." J Proteome Res 3: 786–91.

Tam, E. M., Morrison, C. J., et al. 2004. "Membrane protease proteomics: Isotope-coded affinity tag MS identification of undescribed MT1-matrix metalloproteinase substrates." Proc Natl Acad Sci USA 101: 6917–22.

Timmer, J. C., Enoksson, M., et al. 2007. "Profiling constitutive proteolytic events in vivo." Biochem J 407: 41–8.

Van Damme, P., Martens, L., et al. 2005. "Caspase-specific and nonspecific in vivo protein processing during Fas-induced apoptosis." Nat Methods 2: 771–7.

Vande Walle, L., Van Damme, P., et al. 2007. "Proteome-wide identification of HtrA2/Omi substrates." J Proteome Res 6: 1006–15.

Washburn, M. P., Wolters, D., et al. 2001. "Large-scale analysis of the yeast proteome by multidimensional protein identification technology." Nat Biotechnol 19: 242–7.

Wilm, M., Shevchenko, A., et al. 1996. "Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry." Nature 379: 466–9.

Wolters, D. A., Washburn, M. P., et al. 2001. "An automated multidimensional protein identification technology for shotgun proteomics." Anal Chem 73: 5683–90.