Fast Profiling of Protease Specificity reveals similar substrate specificities for cathepsins K, L and S

Matej Vizovišek[1,2,3], Robert Vidmar[1,3], Emmy Van Quickelberghe[4,5], Francis Impens[4,5,6], Uroš Andjelković[7], Barbara Sobotič[1,3], Veronika Stoka[1], Kris Gevaert[4,5], Boris Turk[1,2,8] and Marko Fonović[1,2*]

[1] Department of Biochemistry and Molecular and Structural Biology, Jozef Stefan Institute, Jamova cesta 39, SI-1000 Ljubljana, Slovenia
[2] Centre of Excellence for Integrated Approaches in Chemistry and Biology of Proteins, Jamova cesta 39, SI-1000 Ljubljana, Slovenia
[3] International Postgraduate School Jozef Stefan, Jamova cesta 39, SI-1000 Ljubljana, Slovenia
[4] Department of Biochemistry, Ghent University, B-9000 Ghent, Belgium
[5] Department of Medical Protein Research, VIB, B-9000 Ghent, Belgium
[6] Unité des Interactions Bactéries-Cellules, Institut Pasteur, 25 Rue du Docteur Roux, 75015 Paris, France
[7] Department of Chemistry, Institute of Chemistry, Technology and Metallurgy, University of Belgrade, Studentski trg 12-16, 11000 Belgrade, Serbia
[8] Faculty of Chemistry and Chemical Technology, University of Ljubljana, Aškerčeva cesta 5, SI-1000 Ljubljana, Slovenia

* Corresponding author: M. Fonović
Jozef Stefan Institute
Jamova 39, 1000 Ljubljana
Phone: +386 1 477 3474
Fax: +386 1 477 3984
e-mail: marko.fonovic@ijs.si

Keywords: proteomics, N-terminomics, intact protein-based cleavage site discovery, cathepsin protease specificity.

Received: 29-Sep-2014; Revised: 02-Dec-2014; Accepted: 22-Jan-2015.

Abbreviations:

AcD_3-NHS, N-hydroxysuccinimide ester of trideutero-acetate;

COFRADIC, combined fractional diagonal chromatography;

FPPS, fast profiling of protease specificity;

PEP, posterior error probability;

PICS, proteomic identification of protease cleavage sites;

SAX, strong anion exchanger;

TAILS, terminal amine isotopic labeling of substrates;

TNBS, 2,4,6-trinitrobenzenesulfonic acid

ABSTRACT: Proteases are important effectors of numerous physiological and pathological processes. Reliable determination of a protease's specificity is crucial to understand protease function and to develop activity-based probes and inhibitors. During the last decade, various proteomic approaches for profiling protease substrate specificities were reported. Although most of these approaches can identify up to thousands of substrate cleavage events in a single experiment, they are often time consuming and methodologically challenging as some of these approaches require rather complex sample preparation procedures. For such reasons their application is often limited to those labs that initially introduced them. Here we report on a fast and simple approach for proteomic profiling of protease specificities (Fast Profiling of Protease Specificity - FPPS), which can be applied to complex protein mixtures. FPPS is based on trideutero-acetylation of novel N-termini generated by the action of proteases and subsequent peptide fractionation on StageTips containing ion-exchange and reverse phase chromatographic resins. FPPS can be performed in two days and does not require extensive fractionation steps. Using this approach, we have determined the specificity profiles of the cysteine cathepsins K, L and S. We further validated our method by comparing the results with the specificity profiles obtained by the N-terminal COFRADIC method. This comparison pointed to almost identical substrate specificities for all three cathepsins and confirmed the reliability of the FPPS approach.

1. Introduction

Proteolytic degradation is one of the most common irreversible protein modifications. In its selective form, also known as proteolytic processing, it can alter protein activity, structural integrity and/or cellular localization. Through proteolytic processing proteases influence a large variety of cellular processes such as cell migration, proliferation, differentiation, tissue remodelling, immune response and apoptosis [1]. Therefore, identification of protease substrates and determination of their specificities can provide important insights into their function. Although recent developments in the field of degradomics have enabled large scale identification of protease substrates, substrate data for many proteases still remain scarce [1]. In the past decade, several proteomic approaches for identification of protease cleavage sites were developed, each relying on various strategies for peptide labelling and enrichment. The new peptides generated by proteolytic cleavage can be isolated by positive enrichment strategies such as N-terminal biotinylation coupled to avidin-based affinity chromatography or by negative enrichment approaches which remove internal (tryptic) peptides from the sample [2]. The two most commonly used negative enrichment strategies are TAILS and COFRADIC. TAILS (Terminal Amine Isotopic Labeling of Substrates) is based on reductive methylation of N-termini with isotopicaly labelled formaldehyde and subsequent removal of unlabelled tryptic peptides by chemical binding to a polyglycerol aldehyde polymer [3, 4]. In COFRADIC (COmbined FRActional DIagonal Chromatography), N-termini are trideutero-acetylated and internal tryptic peptides are removed by reversed phase chromatography after being tagged by a hydrophobic trinitrophenyl group [5-7]. N-terminal peptide enrichment, either by positive or negative approaches, enables in-depth identification of proteolytic events in complex samples. Although the aforementioned methods have proven to be valuable tools for identification of protease cleavage events, they often rely on several experimental steps, amongst others involving sample labelling and/or several cycles of chromatographic separation. They are therefore often difficult to implement in laboratories that are not dedicated to such degradomics methodologies and that only have limited access to the required instrumentation. Since the protease field is constantly expanding with new proteases being discovered and characterized, many laboratories would greatly benefit from the development of a method that enables a quick, simple and reliable determination of protease substrate specificity.

Here we present a simple and straightforward approach for proteomic identification of protease specificities in a complex proteome background. By combining *in solution* isotopic labelling [6] with Stage Tip fractionation [8-10] we developed a simple method to profile protease specificities. We tested our approach by profiling the substrate specificities of the cathepsins K, L and S, three lysosomal proteases known to play important roles in numerous molecular processes such as tissue remodelling, antigen processing and presentation, cell cycle regulation and prohormone processing [11, 12]. Moreover, their upregulated or misslocalized activity has been related to numerous pathological processes, including cancer progression, inflammatory diseases and neurodegeneration [13-17]. The cathepsin substrate specificity was historically studied using oxidized insulin beta chain [18-20] and later by combinatorial chemistry and small peptidic inhibitors [11, 21]. Only recently, a proteomic approach based on proteome-derived peptide libraries (Proteomic Identification of protease Cleavage Sites - PICS) was applied to determine the specificity of cathepsins B, K, L and S [22, 23]. In this approach, a cell lysate was first digested by trypsin or endoproteinase GluC and such generated peptide library was subsequently treated with a given cathepsin. The downside of this peptide-based approach is that some of the putative cleavage sites are destroyed during the proteolytic digestion needed for peptide library preparation. Our approach, which we named FPPS (Fast Profiling of Protease Specificity), enabled us to identify over 1800 cleavage events by cathepsins K, L and S in more than 700 proteins. Overall, a strong preference for aliphatic but also aromatic residues in the P2 position was observed for all three cathepsins, while other positions did not appear to have much influence on cathepsin cleavages, which is in good agreement with our current knowledge on the specificites of cysteine cathepsins [11]. The reliability of our FPPS approach was verified by similar analyses done with the N-terminal COFRADIC approach, which is one of the most used approaches for the in-depth identification of proteolytic cleavage events in complex samples. Both methods showed almost identical specificity profiles for all tested cathepsins.

2. Materials and methods

2.1. Fast Profiling of Protease Specificity (FPPS)

Cell culture and cell lysate preparation

Breast carcinoma cells (MDA-MB-231) were routinely maintained in DMEM medium (Lonza) supplemented with fetal bovine serum (10%), glutamine (1%) and penicillin streptomycin (1%) at 37°C and 5% CO_2. For whole cell lysate preparation, cells were grown to confluency, washed twice with PBS (Lonza, Basel, Switzerland) and detached with Hank's based enzyme-free cell dissociation solution (Millipore, Billerica, MA, USA). Detached cells were centrifuged and lysed on ice for 15 min with lysis buffer (20 mM sodium phosphate buffer pH 8.0, 150 mM NaCl, 0.05% SDS, 0.05% NP-40, 1 mM EDTA, 1 mM PMSF) followed by a short sonication (10 pulses of 5 s each). Insoluble material was removed after centrifugation at 14,000 g for 5 min. Next, the protein concentration was determined with the Bio-Rad assay, and aliquots containing 0.5 mg protein were frozen at -80°C until further use.

In vitro digestion of cell lysates with recombinant cathepsins

Recombinant human cathepsins K, L and S were expressed in the methylotropic yeast expression system *Pichia pastoris* (Invitrogen, Carlsbad, CA, USA) according to standard protocols [24, 25]. Pure mature proteins were titrated with the broad spectrum cysteine cathepsin inhibitor E-64 yielding active concentrations of 14 µM, 42 µM and 19 µM for catK, catL and catS, respectively. Total cell lysates were dialysed (or buffer exchanged) in a microfilter device with a cut-off of 3,000 Da (Millipore, Billerica, MA, USA) against 50 mM phosphate buffer, containing 150 mM NaCl and 2.5 mM DTT, pH 6.0. Next, the pH was checked and the sample was transferred to a low-binding Eppendorf tube (500 µg of total protein). Recombinant cathepsin K, L or S were then added to an approximate 1:100 enzyme/substrate molar ratio based on the assumption that the average Mw of a protein in the sample was 50,000 Da. Samples were then incubated at 37°C for 1 h before adding E-64 at a final concentration of 25 µM to block cathepsin activity.

In solution N-terminal labelling and sample preparation using microfilter devices

The cathepsin-treated samples were transferred to a 500 µl microfilter device with a cut-off of 3,000 Da (Millipore) and the buffer was exchanged to 100 mM phosphate buffer, pH 8.5. The volume of the samples was adjusted to 400 µl with 100 mM phosphate buffer, pH 8.5, before 2 mg of an N-hydroxysuccinimide ester of trideutero-acetate (AcD_3-NHS, made in-house [7]) was dissolved in the sample. The samples were incubated for 1 h at 30°C, after which the labelling step was repeated. In order to reverse partial labelling of serine, threonine and tyrosine side-chains (O-acylation), 10 µl of 50% NH_2OH was added per sample and left to incubate at room temperature for 20 min. Urea was then added to the samples to a final concentration of 8 M and proteins were reduced with 10 mM dithiothreitol (DTT) for 1 h at room temperature. Cysteine residues were alkylated by addition of iodoacetamide to a final concentration of 50 mM and incubation for 1 h at room temperature in the dark. To quench unreacted iodoacetamide, the samples were incubated with 50 mM DTT for 30 min at room temperature. After exchanging the buffer with 25 mM ammonium bicarbonate pH 7.8 by spinning the microcolumns in the centrifuge at 6000 x g for 15 min, the sample volume was adjusted to 250 µl with 25 mM ammonium bicarbonate pH 7.8. Sequencing-grade porcine trypsin (Promega, Madison, WI, USA) was added at a 1:100 (w/w, enzyme/substrate) ratio and incubated overnight at 37°C. The next day, the peptide-containing flowthrough was collected by spinning the microcolumns in the centrifuge at 6000 x g for 10 min and concentrated down to 50 µl.

Sample fractionation using SAX-C18 StageTips

The generated peptide mixture was mixed with Britton & Robinson buffer (20 mM acetic acid, 20 mM phosphoric acid and 20 mM boric acid), pH 11 (the pH was adjusted with 1 M NaOH if necessary). Anion exchanger tips were prepared by stacking 6 discs of Empore/Disk Anion Exchange (1214-5012) (Varian, Palo Alto, CA, USA) in a 200 µl pipet tip, while C18 tips were prepared by stacking 4 discs of Empore/C18 (Varian, Palo Alto, CA, USA). Peptides were fractionated on SAX-C18 StageTips with buffers at pH 11, 10, 9, 8, 7, 6, 5, 4 and 3 as described elsewhere [8, 9] and subsequently desalted on C18 tips with elution in 50 µl 60% (v/v) acetonitrile, 0.1% (v/v) formic acid in water. Eluted samples were concentrated to 15 µl by vacuum drying prior to further LC-MS/MS analysis.

LC-MS/MS analysis

LC-MS/MS analysis was performed using an Orbitrap LTQ Velos mass spectrometer coupled to a Proxeon nano-LC HPLC unit, which were automatically operated under the Xcalibur software (Thermo Fisher Scientific, Waltham, MA, USA). The peptide samples were loaded on a C18 trapping column (Proxeon EASY-Column^{TM}) (Thermo Fisher Scientific, Waltham, MA, USA) using solvent A (0.1% (v/v) formic acid in water) and separated on a C18 PicoFrit^{TM} AQUASIL analytical column (New Objective, Woburn, MA, USA). Elution was performed using a 120 min 5-50% linear gradient of solvent B (100% acetonitrile, 0.1% formic acid) at a flow rate of 300 nl/min. During the gradient, full MS spectra were recorded and MS/MS spectra were obtained by CID fragmentation of the nine most intense precursor ions from the full MS scan. Dynamic exclusion was enabled with repeat count of 2 and 60 s exclusion time. Full MS spectra were recorded in Orbitrap at resolution of 30,000, MS/MS spectra were recorded in profile mode in the linear ion trap at resolution of 7,500. Precursors with nonasigned charge state were not chosen for MS/MS.

Analysis of MS/MS data

Data analysis, including database searches, was performed using the MaxQuant software package version 1.4.0.5 and the Andromeda search engine [26, 27]. The raw spectral data were searched against the human protein sequences in the UniProt/Swiss-Prot database (UniProtKB, *Homo sapiens*, canonical database containing 20,249 sequences, release of May 1, 2013) with trideutero-acetylation of peptide N-termini (+45.034 Da) and methionine oxidation (+15.99 Da) as variable modifications and trideutero-acetylation of lysines (+45.034 Da) and carbamidomethylation of cysteines (+57.02 Da) as fixed modifications. Database searches were performed with semi-ArgC/P as enzyme setting allowing for 1 missed cleavage. Precursor ion and fragment ion mass tolerances were set to 20 ppm and 0.5 Da respectively. A reversed database search was performed and the false discovery rate (FDR) was set at 1% for peptide and protein identifications. Peptides with posterior error probability (PEP) score below 0.05 were retained as positive hits (Supplementary table S1). Peptides with trideuteroacetylated N-termini, which were present in the cathepsin treated sample but not in the negative control, were considered to be the product of cathepsin cleavage. The P1'-P4' positions were determined from the peptide N-terminus, whilst the P4-P1 positions were determined bioinformatically. The iceLogos were generated using frequencies of positional amino acid occurrence normalized to natural amino acid abundances in the human Swiss-Prot database [28].

2.2 COFRADIC Method

Cell culture and SILAC labelling

HL60 cells (European Collection of Cell Cultures, Salisbury, UK) were cultured in RPMI medium without arginine, lysine and glutamine (Silantes, München, Germany) supplemented with 10% dialyzed fetal bovine serum (Invitrogen, Carlsbad, CA, USA), 2 mM GlutaMAX^{TM} (Invitrogen), 25 units/ml penicillin (Invitrogen), 25 µg/ml streptomycin (Invitrogen), 146 mg/L L-Lysine.HCl (Sigma-Aldrich, Saint Louis, MO, USA) and either natural L-Arginine.HCl (light, Sigma-Aldrich) or $^{13}C_6$ L-Arginine.HCl (heavy, Silantes). The concentration of L-Arginine.HCl was reduced to 16.8 mg/L (8.4% of the normal concentration in RPMI) to prevent metabolic conversion of arginine to proline. Cells were cultured at 37°C and in 5% CO_2 for at least six population doublings to ensure complete incorporation of the labelled arginine. Upon completion, aliquots of 10^7 SILAC-labelled cells were washed three times with PBS and cell pellets were frozen at -80˚C until further use.

Cell lysis and cathepsin treatment

SILAC-labelled cells were lysed in a buffer containing 50 mM 2-(N-morpholino)ethanesulfonic acid (MES), 50 mM sodium phosphate pH 8, 150 mM NaCl, 1 mM EDTA and 1 mM DTT by three rounds of freeze-thawing. Lysates were cleared by centrifugation for 15 min at 16,000 x g, followed by acidification to pH 5.5 using 1 M HCl. For cathepsin treatment, we used a differential sample mixing strategy to allow software-based quantification and annotation of protein processing events, as reported previously [29]. Here, the control sample, made from three parts of light-labelled cell lysate, was incubated with 10 µM E-64, whereas the treated sample, made from one part light-labelled and one part heavy-labelled cell lysate, was incubated with 200 nM recombinant cathepsin L, K or S for 15 min at 37°C. After incubation, the control and cathepsin-treated samples were mixed, 4 M guanidinium hydrochloride was added and the pH was raised to 7.4 to proceed with COFRADIC isolation of N-terminal peptides.

COFRADIC isolation of N-terminal peptides

The N-terminal COFRADIC protocol was performed as described before [7]. Briefly, proteins were reduced and alkylated using 15 mM tris(2-carboxyethyl)phosphine (TCEP) and 30 mM iodoacetamide. After desalting the protein mixtures on NAP-10 columns (GE Healthcare, Little Chalfont, UK) in 2 M guanidinium hydrochloride and 50 mM sodium phosphate buffer at pH 8, free amino groups were acetylated by incubation with 20 mM of an N-hydroxysuccinimide (NHS) ester of trideutero-acetate (made in-house) for 1 h at 30°C followed by addition of 10 mM glycine to quench residual NHS esters and 120 mM hydroxylamine to reverse potential O-acetylation and incubation for 20 min at room temperature. The protein samples were again desalted on NAP-10 columns in 20 mM ammonium bicarbonate, boiled for 5 min, put on ice for 5 min and digested overnight with sequencing-grade trypsin (Promega, Madison, WI, USA) (enzyme/substrate of 1/100, w/w) at 37°C. The resulting peptide mixtures were incubated with 1,250 mU Q-cyclase (Qiagen, Germatown, MD, USA) and 625 mU activated pGAPase (Qiagen) to drive the formation of N-terminal pyroglutamate to completion and proteolytically remove this residue, respectively. N-terminal peptides were then pre-enriched by strong cation exchange (SCX) chromatography on disposable SCX cartridges (Agilent Technologies, Santa Clara, CA, USA) in a buffer containing 50% acetonitrile in 10 mM sodium phosphate, pH 3.0. N-terminal peptides, which were not retained on the SCX resin under these conditions, were collected in the run-through fraction, concentrated by vacuum drying and re-suspended in 10 mM ammonium acetate pH 5.5 in 2% acetonitrile. Methionine residues were oxidized by incubating the peptides with 0.5% H_2O_2 for 30 min at 30°C, followed by immediate injection of the samples on a capillary RP-HPLC column (Zorbax 300SB-C18, 2.1 mm internal diameter, 150 mm length, Agilent Technologies) for the first COFRADIC separation. Peptides were separated by a linear gradient of acetonitrile (from 2% to 70% in 100 min) and peptides that eluted between 20 and 80 min were collected in 15 primary COFRADIC fractions of 4 min each. Each primary fraction was incubated four times with 15 nmol of 2,4,6-trinitrobenzenesulfonic acid (TNBS) and each fraction was re-separated by RP-HPLC under the same conditions as during the primary separation. N-terminal peptides, which cannot get modified by TNBS, eluted from the column during the same time interval and were collected in 16 fractions of 0.5 min. Internal and C-terminal peptides, which were modified by TNBS, underwent a hydrophobic shift and were not collected. Secondary fractions with 12 min difference in retention time were pooled to a total of 48 samples for LC-MS/MS analysis.

LC-MS/MS and data analysis

Samples were analyzed by LC-MS/MS on a linear trap quadrupole (LTQ) Orbitrap Velos or a LTQ Orbitrap XL mass spectrometer (Thermo Fisher Scientific, Waltham, MA, USA) and peak list files were generated from the peptide fragmentation spectra as described before [29, 30]. Spectra were searched with MASCOT (version 2.4.0, www.matrixscience.com) in the Swiss-Prot database (releases of November 11, 2011, January 25, 2012 and April 18, 2012) with restriction to human proteins and quantified by the MASCOT Distiller software as described elsewhere [29]. Briefly, semiArgC/P was used as enzyme setting with one missed cleavage allowed. Methionine oxidation, trideutero-acetylation of lysine and alkylation of cysteine were set as fixed modifications, while variable modifications included acetylation and trideutero-acetylation of N-termini and pyroglutamate formation of N-terminal glutamine residues. Mass tolerance of precursor and fragment ions was set to 10 ppm and 0.5 Da, respectively. To allow identification of peptides from both light and heavy labelled samples, a quantitation method with two different components was made, defining $^{12}C_6$ Arg and $^{13}C_6$ Arg as exclusive modifications at any position. Only peptides that were ranked first and scored above the threshold score set at 99% confidence were withheld. The FDR was calculated for every search as described previously [31] and was always found to be lower than 0.78%. Potential false positive peptide identifications were selected and automatically removed by the Peptizer software application exactly as described before [32]. Only peptides that fulfilled the following three conditions were considered as neo-N-terminal peptides reporting true cathepsin cleavage sites: (1) the start position of the peptide in the protein sequence is >2, (2) the N-terminal alpha amino group of the peptide is modified by trideutero-acetylation and (3) the light/heavy (L/H) ratio is < 2 (L/H ratios of neo-N-terminal peptides are distributed around 1, determination of the ratio cut-off value is based on the boundary of a one-sided 99% quantile of the ratio distribution) as described before [29]. Neo-N-terminal peptides were then loaded into the TOPPR database [33] and re-mapped onto a later version of the Swiss-Prot database (release of January 22, 2014) to generate the final lists of total cleavage sites (Supplementary table S2).

3. Results

3.1 Identification of potential protease substrates and cleavage sites with FPPS

Figure 1 shows the workflow of the Fast Profiling of Protease Specificity (FPPS) method with its most important experimental steps. For evaluation of cathepsin cleavages, we treated proteins from soluble MDA-MB-231 cell lysates with recombinant cathepsins whereas lysates to which the general cathepsin inhibitor E-64 was added, were used as a negative control. The protein fragments generated by cathepsin cleavage were trideutero-acetylated on their N-termini in order to distinguish them from internal peptides that will be generated by trypsin in a next step. Since the label efficiently reacts with primary amino groups, it also labels the side-chains of lysines. As trypsin can no longer cleave at acetylated lysines, the majority of tryptic peptides end with an arginine at the C-terminus. Sample preparation for mass spectrometry analysis was performed *in solution* and the obtained peptide sample was subsequently fractionated using the Stage Tip protocol, which is based on the SAX capture of peptides followed by stage elution over C18 resin tips [8, 9]. Among all of the peptides identified by LC-MS/MS, cathepsin cleavage sites were recognized in the form of trideuteroacetylated neo-N-terminal peptides which were absent from the control sample. Using FPPS we identified 731 cleavages in 421 proteins for cathepsin K, 576 cleavages in 340 proteins for cathepsin L and 1062 cleavages in 539 proteins for cathepsin S, respectively (Figure 2A-B).The high number of common cleavage sites indicates that the three tested cathepsins have similar cleavage site preferences, which is in agreement with their high sequence homology, similar structural fold and reported physiological redundancy (reviewed in [11, 34]). In about 60% of the identified proteins only a single cleavage site was identified, suggesting that cathepsins processed, but not degraded proteins (Fig. 2C).

3.2 Sequence specificity analysis of identified cleavage sites

In order to compare the substrate specificity of the three cathepsins, the P1'-P4' residues were determined from the identified N-terminally labelled peptides, whereas the corresponding P1-P4 residues were determined by bioinformatic analysis. Using this information, substrate sequence logos composed of total unique cleavage sites for each of the tested cathepsins were generated [28] (Figure 3A-B). It is generally accepted that the substrate specificity of cysteine cathepsins is primarily governed by their S2 substrate binding pocket, which is preferentially occupied by hydrophobic residues [35]. This was confirmed by FPPS since hydrophobic

amino acids, such as leucine, valine and isoleucine were preferred in the P2 position of all three cathepsins. However, some differences among the cathepsins were also observed. Most important seems to be the cathepsin K preference for a proline in the P2 position, which was not observed for the other two cathepsins. In addition, cathepsin L exhibited a higher preference for aromatic residues (tyrosine and phenylalanine) in the P2 position than cathepsins S and K, while cathepsin S exhibited a higher preference for lysine in this position than the other two cathepsins.

Apart from the S2 binding site, which is a deep pocket, the S1 and S1' sites provide a substrate binding surface which is less well defined [35]. In the P1 position, glycine, arginine and lysine were the most common residues in all three cathepsins. In addition, cathepsin S, but not cathepsins L and K, also accepted threonine in this position. In the P1' position, cathepsin S preference for isoleucine was the only notable difference observed among the tested cathepsins (Figure 3A-B). In the P1 and P1' positions proline residues were strongly disfavoured for all three cathepsins. Specificity in the other positions was less pronounced, although a general preference for acidic residues (aspartates and glutamates) in the prime positions (P1'-P4') was observed for all three cathepsins. This was not observed in other reported datasets [21-23] and the reason for this could be a stronger retention of negatively charged peptides on anionic exchange resin (SAX). It should also be noted that charged regions on protein surface are more readily accessible to protease cleavage than buried hydrophobic regions. The sum of those two effects could therefore account for the observed enrichment of negatively charged residues at prime sites.

A direct pairwise comparison of substrate specificities of the three cathepsins showed additional minor differences in the P3 and the prime positions (Figure 4). Most notable is the cathepsin K preference for aspartate residues in the P2' and P4' positions and the cathepsin L preference for proline in the P4' position.

3.3 Comparison of FPPS with N-terminal COFRADIC

In order to evaluate the FPPS method, we compared it with the N-terminal COFRADIC (COmbined FRActional DIagonal Chromatography) approach [5]. The main feature of COFRADIC is the enrichment of newly formed N-termini that are tagged and blocked by *in vitro* trideutero-acetylation, by depletion of internal peptides, being the products of trypsin degradation. The bulk of internal peptides are removed in a first stage by binding to an SCX resin, and in a second stage by labelling with a highly hydrophobic trinitrophenyl group and subsequent removal by RP-HPLC. Identification of cathepsin K, L and S cleavage sites by COFRADIC was performed independently using a different cell line (HL-60). Due to the higher enrichment of neo-N-terminal peptides by elimination of internal tryptic peptides, COFRADIC was able to detect more cleavage events than FPPS. In general, N-terminal COFRADIC between 3,000 to 8,000 neo N-terminal peptides, while FPPS led to the identification of 500 to 1,000 cleavage events (Tables 1 and 2). However, a direct comparison of the specificity profiles obtained by both methods revealed almost identical profiles in the P1-P4 positions for all three tested cathepsins (Figure 5A-B). This demonstrates that the number of cathepsin cleavages identified with FPPS was sufficient for a reliable determination of their cleavage specificity. However, some differences between the two methods were observed in the P1'-P4' positions (Figure 5B). Most notable are the enrichment for histidine residues at prime positions in the FPPS dataset and the enrichment for lysine residues upon COFRADIC analysis, most probably due to differences in the sample preparation protocols of both methodologies. In the N-terminal COFRADIC protocol, during the initial SCX enrichment of N-terminal peptides histidine-containing peptides are removed since these residues are positively charged at the experimental pH (pH 3). On the other hand, lysine residues in both experimental approaches lose their charge due to deuteroacetylation of their side chain. In COFRADIC, they are therefore not retained by SCX resin, while in FPPS they are not efficiently separeted by SAX, which could lead to their observed underrepresentation. Interestingly, glycine residues in the P1' position were only found enriched in the FPPS analysis and not in the COFRADIC experiment, although a preference for glycine in this position has been reported previously [22]. In general, despite some subtle differences inherent to the experimental approaches, both approaches provided rich cleavage datasets and a reliable specificity profile for all tested cathepsins. Note that the fact that two different cell lines, MDA-MB-231 and HL-60 cells, were used when comparing both methods gave no influence on the determined specificity profiles. That was expected, as cathepsin cleavage preferences are independent from the actual protein composition of a sample.

4. Discussion

Although various approaches for proteomic determination of protease substrate specificities were reported in the last decade, this remains a challenging task, typically involving many steps of peptide labelling and separation. Here we report on the development and validation of a simple and straightforward degradomics procedure, which enables quick and reliable determination of protease substrate specificities that can be easily implemented in any laboratory setting. Moreover, the labelling reagents are very easily synthesized from commercialy available chemicals [7]. The procedure can be performed in two days and requires a single trideutero-acetylation labelling step, followed by peptide separation using home-made Stage Tips containing SAX and C18 resin which are known for efficient sample fractionation and good sample recovery [8-10]. Using this sample fractionation approach, we were able to identify 500-1,000 cleavage events per single protease treatment experiment, which proved to be sufficient for a detailed determination of protease specificity. For the purpose of simplicity, no relative quantitation was used in the approach described, although relative quantification methods such as SILAC could be included, if necessary. In our case, only N-terminally trideutero-acetylated peptides, which were present in the cathepsin-treated proteome and absent in the negative control, were used for profiling cleavage specificities. The reliability of our approach was verified by using N-terminal COFRADIC data on the same cathepsins, which gave almost identical specificity profiles. In a way, this confirmed our expectations in that the actual absence of in-depth quantitative data in FPPS did not introduce significant numbers of false positive assignments of neo N-terminal peptides. Furthermore, comparison of FPPS and COFRADIC data with cathepsin specificities obtained by peptide-based methods such as PICS and peptide library screens [21-23] shows numerous similarities in the obtained profiles, but also revealed some differences related to the methods used. Among the peptide-based approaches, some cleavage sites are lost in PICS during peptide library preparation [22] and peptide library screens cannot profile prime sites [1]. On the other hand, in protein-based approaches, cleavage sites containing hydrophobic residues can be underrepresented due to their lower availability or lover detectability in LC-MS/MS, while hydrophilic residues can be overrepresented for the same reason. Such observations suggest that various experimental approaches should be combined in order to obtain specificity profiles which are bias-free and most relevant for a given experimental setting.

Profiling of proteases with broad substrate specificities such as cysteine cathepsins is usually less straightforward than analysis of highly specific proteases, such as caspases. Cathepsins

are important players in numerous pathological processes and reliable profiling of their specificities is of high importance for a better understanding of their physiological function. With some of their inhibitors currently going through various phases of clinical testing, cathepsins are also considered as relevant therapeutical targets [36, 37]. In general, FPPS and COFRADIC results correlate well with peptide library screen results, but comparison of specificity profiles has also revealed some differences [21]. In the P1 position of all tested cathepsins, FPPS as well as COFRADIC showed enrichment of glycines, which was not observed in peptide library screens. Both FPPS and COFRADIC also identified lysines and arginines in this position, which is an important aspect of cathepsin substrate specificity, and was also observed in small peptide positional library screens [21], but was completely absent in the peptide-based PICS method, where trypsin was used for the preparation of peptide library [22, 23]. On the other hand, substrate library screens showed a strong preference for tryptophan in the P2 position, which was less prominent in PICS, but absent in COFRADIC and FPPS. The reason for this could be the low natural occurrence of tryptophan [38] and the overall lower detectability of hydrophobic peptides in MS-based approaches. A plausible explanation is that in native proteins, which comprise the starting proteome for cathepsin processing, hydrophilic regions that form the loops and other protease-exposed regions, are much more accessible to proteases than internal regions, which are mostly hydrophobic. Therefore, cleavages in hydrophobic regions are less likely to occur mainly due to more limited accessibility. Both FPPS and COFRADIC were able to detect a similar specificity in the P2 position for all tested cathepsins, which is in good agreement with previous studies [21-23]. Moreover, they also detected a proline specificity for cathepsin K, a distinguishing feature from cathepsins L and S, that was also observed previously and is likely to be of physiological relevance [21, 39]. A preference of cathepsin K for Pro in the P2 position was also demonstrated using mutants of the S2 pocket [40]. Indeed, cathepsin K is known to cleave proline-rich regions of collagen [41], an abundant structural protein composed of three type II polyproline helices [42]. In other positions, especially P1'-P4', there were only minor differences between FPPS and COFRADIC.

In summary, results obtained using FPPS are in good agreement with published data on substrate specificities of cysteine cathepsins. The observed differences between tested cathepsins indicated some level of individual substrate preferences; however, the fact that these differences are relatively small is in accordance with the reported overlapping substrate specificity between cathepsins K, L and S. The described proteomic approach could thus be a valuable tool in further studies of protein processing in complex proteomes. It should be noted however that neo-N-terminal peptide enrichment based approaches such as COFRADIC are expected to be less dependent on the scan speeds of mass spectrometers due to the overall lower sample complexity. However, the introduction of mass spectrometers with increasing scan speeds is expected to alleviate this issue for FPPS. As there was a good correlation between FPPS and COFRADIC data, we believe that due to its simplicity FPPS could become a method-of-choice for fast and reliable profiling of proteases in numerous laboratories. On the other hand, a possible limitation of the FPPS approach might be its application to profiling of highly specific proteases such as caspases that generate much lower numbers of cleavage sites. In such cases, the corresponding neo-N-terminal peptides might remain unidentified in the more complex FPPS peptide mixtures as compared to COFRADIC or TAILS peptide mixtures.

ACKNOWLEDGEMENTS

We wish to thank I. Klemenčič (Jožef Stefan Institute, Ljubljana) for excellent technical assistance and to S. Pečar and J. Mravljak (Faculty of Pharmacy, University of Ljubljana) for the preparation of AcD_3-NHS. This work was supported by grants from the Slovenian Research Agency (P1-0140, J1-3602 and N1-0022 to B.T.; J1-0185 to M. F.). Further, K.G. acknowledges support from the Research Foundation – Flanders (FWO-Vlaanderen, project numbers G.0048.08 and G.0C37.14N). The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium [43] via the PRIDE partner repository with the dataset identifier PXD001553 and PXD001536 (DOI: 10.6019/ PXD001536).

The authors have declared no conflict of interest.

REFERENCES

[1] Turk B., Turk D., Turk V., Protease signalling: the cutting edge. *EMBO J.* 2012, 31, 1630-43.
[2] Plasman K., Van Damme P., Gevaert K., Contemporary positional proteomics strategies to study protein processing. *Curr Opin Chem Biol.* 2013, 17, 66-72.
[3] Kleifeld O., Doucet A., auf dem Keller U., Prudova A. et al., Isotopic labeling of terminal amines in complex samples identifies protein N-termini and protease cleavage products. *Nat Biotechnol.* 2010, 28, 281-8.
[4] Prudova A., auf dem Keller U., Butler G.S., Overall C.M., Multiplex N-terminome analysis of MMP-2 and MMP-9 substrate degradomes by iTRAQ-TAILS quantitative proteomics. *Mol Cell Proteomics.* 2010, 9, 894-911.
[5] Gevaert K., Goethals M., Martens L., Van Damme J. et al., Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides. *Nat Biotechnol.* 2003, 21, 566-9.
[6] Van Damme P., Van Damme J., Demol H., Staes A. et al., A review of COFRADIC techniques targeting protein N-terminal acetylation. *BMC Proc.* 2009, 3 Suppl 6, S6.
[7] Staes A., Impens F., Van Damme P., Ruttens B. et al., Selecting protein N-terminal peptides by combined fractional diagonal chromatography. *Nat Protoc.* 2011, 6, 1130-41.
[8] Wisniewski J.R., Zougman A., Mann M., Combination of FASP and StageTip-Based Fractionation Allows In-Depth Analysis of the Hippocampal Membrane Proteome. J *Proteome Res.* 2009, 8, 5674-8.
[9] Wisniewski J.R., Zougman A., Nagaraj N., Mann M., Universal sample preparation method for proteome analysis. *Nat Methods.* 2009, 6, 359-62.
[10] Wisniewski J.R., Ostasiewicz P., Mann M., High Recovery FASP Applied to the Proteomic Analysis of Microdissected Formalin Fixed Paraffin Embedded Cancer Tissues Retrieves Known Colon Cancer Markers. *J Proteome Res.* 2011, 10, 3040-9.
[11] Turk V., Stoka V., Vasiljeva O., Renko M. et al., Cysteine cathepsins: from structure, function and regulation to new frontiers. *Biochim Biophys Acta.* 2012, 1824, 68-88.
[12] Brix K., Dunkhorst A., Mayer K., Jordans S., Cysteine cathepsins: cellular roadmap to different functions. *Biochimie.* 2008, 90, 194-207.
[13] Vasiljeva O., Reinheckel T., Peters C., Turk D. et al.,. Emerging roles of cysteine cathepsins in disease and their potential as drug targets. *Curr Pharm Des.* 2007, 13, 387-403.
[14] Fonovic M., Turk B., Cysteine cathepsins and their potential in clinical therapy and biomarker discovery. *Proteomics Clin Appl.* 2014, 8, 416-26.
[15] Pislar A., Kos J., Cysteine cathepsins in neurological disorders. *Mol Neurobiol.* 2014, 49, 1017-30.
[16] Reiser J., Adair B., Reinheckel T., Specialized roles for cysteine cathepsins in health and disease. *J Clin Invest.* 2010, 120, 3421-31.
[17] Mohamed M.M., Sloane B.F., Cysteine cathepsins: multifunctional enzymes in cancer. *Nat Rev Cancer.* 2006, 6, 764-75.
[18] Kargel H.J., Dettmer R., Etzold G., Kirschke H. et al., Action of cathepsin L on the oxidized B-chain of bovine insulin. *FEBS Lett.* 1980, 114, 257-60.
[19] Kirschke H., Wiederanders B., Bromme D., Rinne A., Cathepsin S from bovine spleen. Purification, distribution, intracellular localization and action on proteins. *Biochem J.* 1989, 264, 467-73.
[20] Bromme D., Steinert A., Friebe S., Fittkau S. et al.,. The specificity of bovine spleen cathepsin S. A comparison with rat liver cathepsins L and B. *Biochem J.* 1989, 264, 475-81.

[21] Choe Y., Leonetti F., Greenbaum D.C., Lecaille F., et al., Substrate profiling of cysteine proteases using a combinatorial peptide library identifies functionally unique specificities. *J Biol Chem.* 2006, 281, 12824-32.
[22] Biniossek M.L., Nagler D.K., Becker-Pauly C., Schilling O., Proteomic identification of protease cleavage sites characterizes prime and non-prime specificity of cysteine cathepsins B, L, and S. *J Proteome Res.* 2011, 10, 5363-73.
[23] Schilling O., Overall C.M., Proteome-derived, database-searchable peptide libraries for identifying protease cleavage sites. *Nat Biotechnol.* 2008, 26, 685-94.
[24] Mihelic M., Dobersek A., Guncar G., Turk D., Inhibitory fragment from the p41 form of invariant chain can regulate activity of cysteine cathepsins in antigen presentation. *J Biol Chem.* 2008, 283, 14453-60.
[25] Bromme D., Nallaseth F.S., Turk B., Production and activation of recombinant papain-like cysteine proteases. *Methods.* 2004, 32, 199-206.
[26] Cox J., Mann M., MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol.* 2008, 26, 1367-72.
[27] Cox J., Neuhauser N., Michalski A., Scheltema R.A. et al.,. Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. *J Proteome Res.* 2011, 10, 1794-805.
[28] Colaert N., Helsens K., Martens L., Vandekerckhove J., Gevaert K., Improved visualization of protein consensus sequences by iceLogo. *Nat Methods.* 2009, 6, 786-7.
[29] Impens F., Colaert N., Helsens K., Ghesquiere B. et al., A quantitative proteomics design for systematic identification of protease cleavage events. *Mol Cell Proteomics.* 2010, 9, 2327-33.
[30] Ghesquiere B., Colaert N., Helsens K., Dejager L., et al., In vitro and in vivo protein-bound tyrosine nitration characterized by diagonal chromatography. *Mol Cell Proteomics.* 2009, 8, 2642-52.
[31] Elias J.E., Gygi S.P., Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods.* 2007, 4, 207-14.
[32] Helsens K., Timmerman E., Vandekerckhove J., Gevaert K., Martens L., Peptizer, a tool for assessing false positive peptide identifications and manually validating selected results. *Mol Cell Proteomics.* 2008, 7, 2364-72.
[33] Colaert N., Maddelein D., Impens F., Van Damme P. et al., The Online Protein Processing Resource (TOPPR): a database and analysis platform for protein processing events. *Nucleic Acids Res.* 2013, 41, D333-7.
[34] Nagler D.K., Menard R., Family C1 cysteine proteases: biological diversity or redundancy? *Biol Chem.* 2003, 384, 837-43.
[35] Turk D., Guncar G., Podobnik M., Turk B., Revised definition of substrate binding sites of papain-like cysteine proteases. *Biol Chem.* 1998, 379, 137-47.
[36] Turk B., Targeting proteases: successes, failures and future prospects. *Nat Rev Drug Discov.* 2006, 5, 785-99.
[37] Palermo C., Joyce J.A., Cysteine cathepsin proteases as pharmacological targets in cancer. *Trends Pharmacol Sci.* 2008, 29, 22-8.
[38] Tourasse N.J., Li W.H., Selective constraints, amino acid composition, and the rate of protein evolution. *Mol Biol Evol.* 2000, 17, 656-64.
[39] Lecaille F., Weidauer E., Juliano M.A., Bromme D., Lalmanach G., Probing cathepsin K activity with a selective substrate spanning its active site. *Biochem J.* 2003, 375, 307-12.

[40] Lecaille F., Chowdhury S., Purisima E., Bromme D., Lalmanach G., The S2 subsites of cathepsins K and L and their contribution to collagen degradation. *Protein Sci.* 2007, 16, 662-70.
[41] Dejica V.M., Mort J.S., Laverty S., Antoniou J. et al. Increased type II collagen cleavage by cathepsin K and collagenase activities with aging and osteoarthritis in human articular cartilage. *Arthritis Res Ther.* 2012, 14, R113.
[42] Shoulders M.D., Raines R.T., Collagen structure and stability. *Annu Rev Biochem.* 2009, 78, 929-58.
[43] Vizcaíno J.A., Deutsch E.W., Wang R., Csordas A. et al. ProteomeXchange provides globally co-ordinated proteomics data submission and dissemination. *Nat Biotechnol.* 2014, 30, 223-226.

Figure 1. Outline of the FPPS method

The method combines in-solution sample labelling in microfilter devices with the Stage Tip sample separation protocol. In the first step, cells are lysed and a soluble lysate is treated with a recombinant cathepsin. Primary N-termini and neo-N-termini are trideutero-acetylated in order to be distinguished from internal peptides in later stages and proteins are then digested with trypsin according to standard procedures. Peptides are fractionated on SAX-C18 tips and analysed by LC-MS/MS, after which the lists of peptides identified in the negative control and in the sample treated with cathepsin are compared and only the neo-N-terminal peptides not present in the negative control are considered to result from cathepsin cleavage.

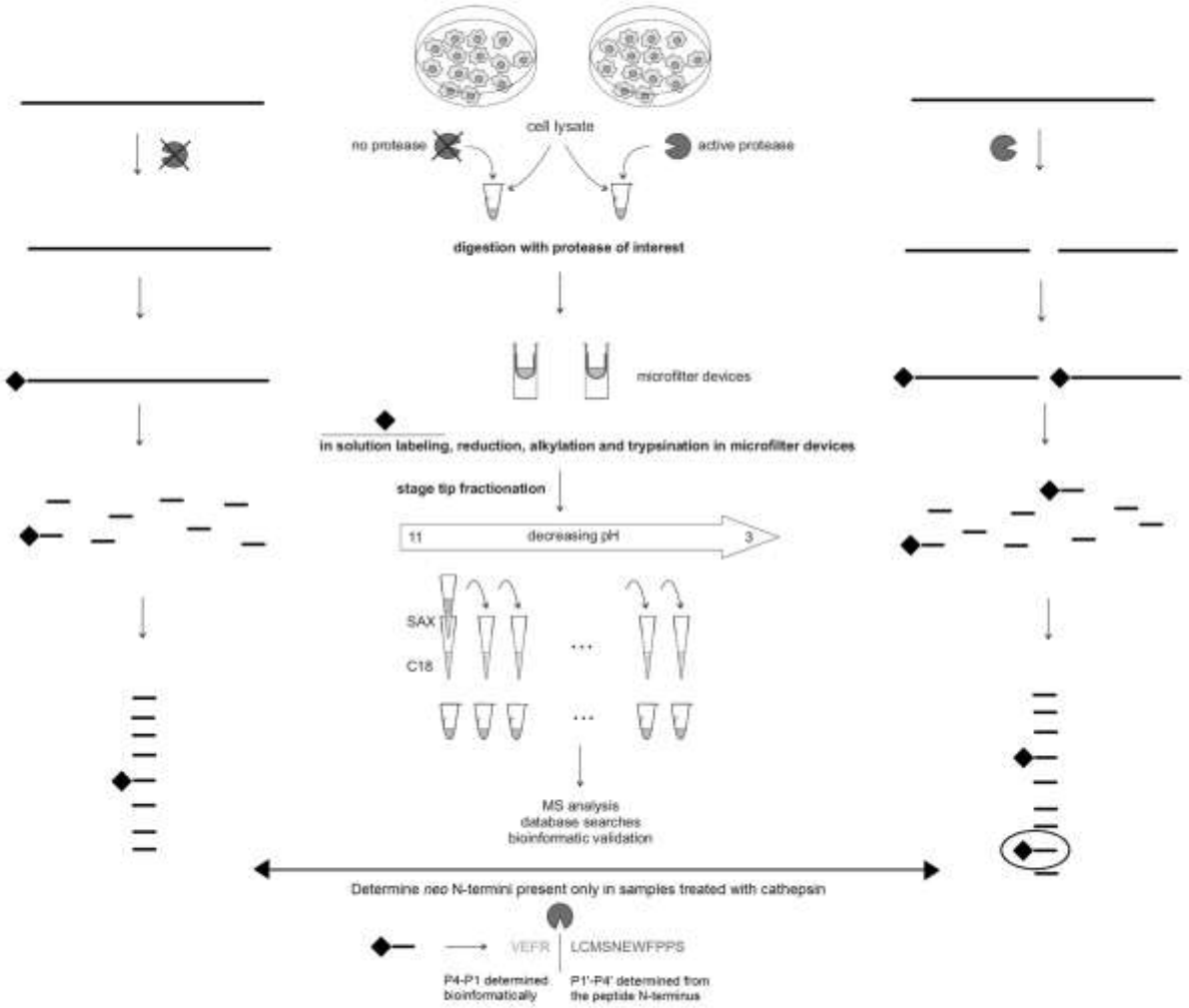

Figure 2. Identified cathepsin cleavage sites and cleaved substrate proteins

Over 1,800 cleavage sites in more than 700 proteins were identified for cathepsins K, L and S. The Venn diagrams show the overlap of identified substrate proteins (A) and cleavage sites (B) between the three tested cathepsins. The pie charts indicate the fraction of substrate proteins with 1, 2, 3, 4, 5 or > 5 detected cleavage sites (C).

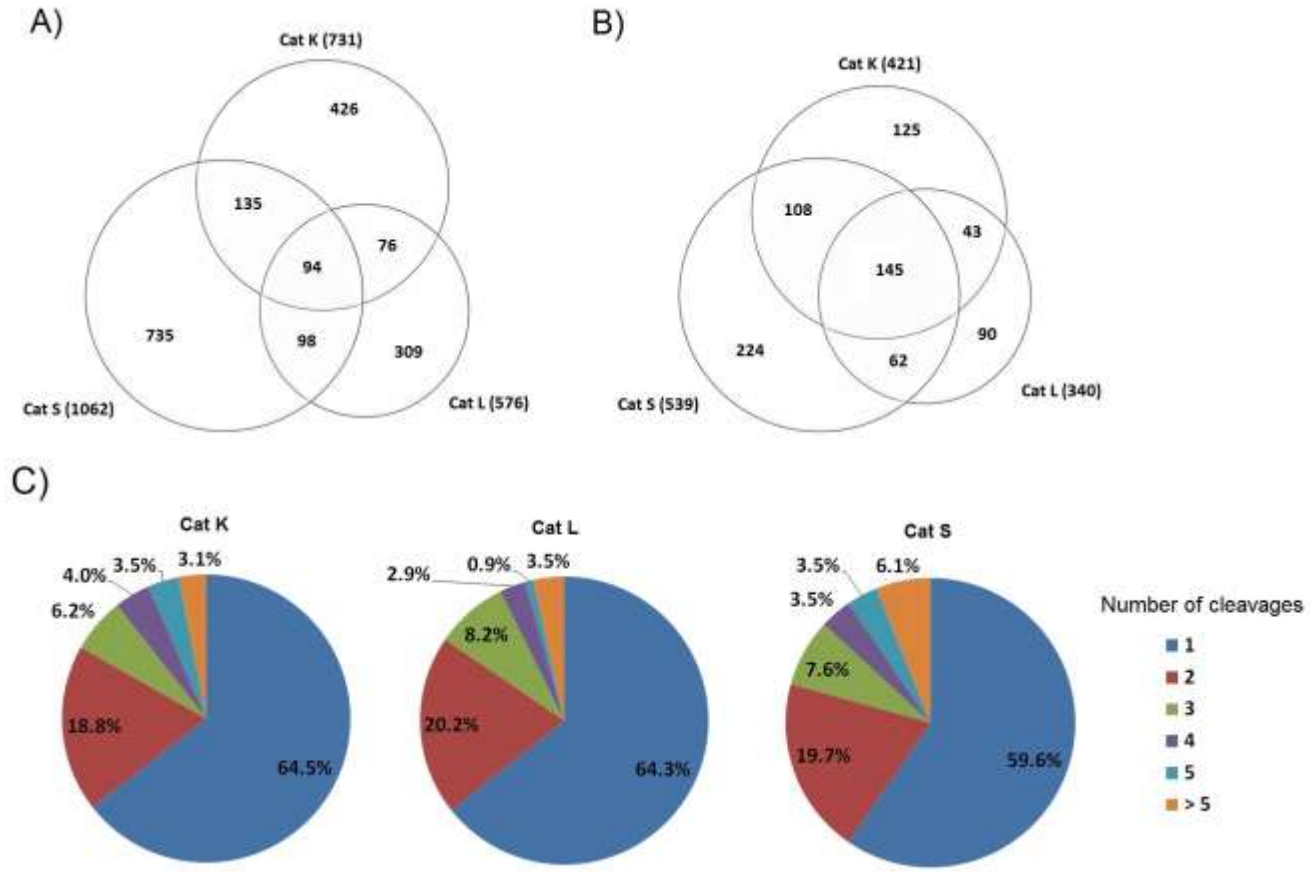

Figure 3: IceLogos indicating the cleavage site specificity for cathepsins K, L and S after FPPS analysis.

IceLogos based on all (A) or unique (B) cleavage sites detected with cathepsin K, L and S. At each position surrounding the cleavage site, significantly under- and over-represented amino acids (p-value was set to 0.01) are shown as letters whose size corresponds to the percentage of difference with their average proteome occurrency. Residues shown in pink were never observed at that position.

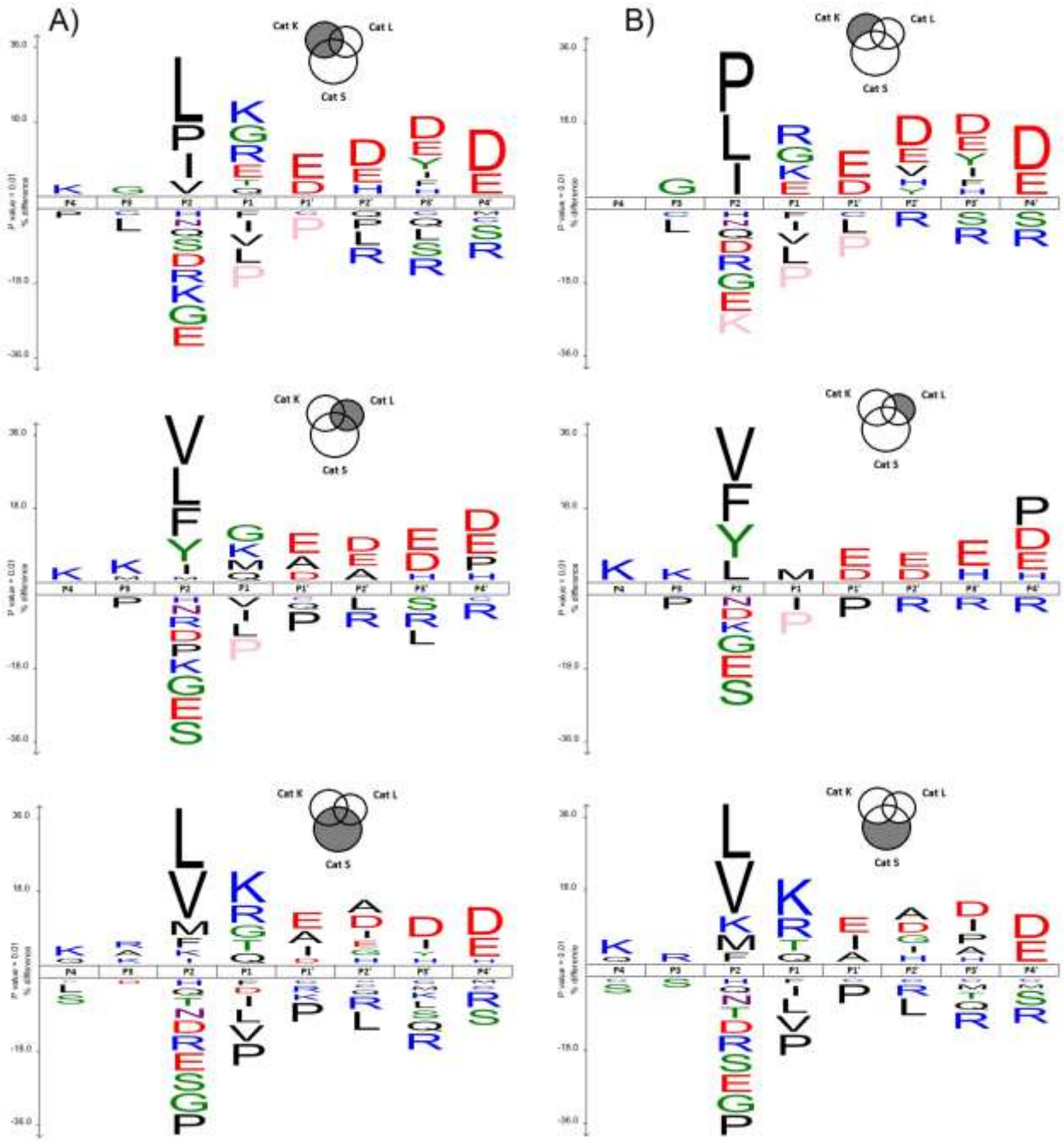

Figure 4: Pairwise comparison of substrate preferences for cathepsin K, L and S detected with FPPS.

Cleavage sites specific for cathepsin K, L or S were used to generate iceLogos for direct comparaison of the cleavage site specificity. At each position surrounding the cleavage site, significantly under- and over-represented amino acids (p-value was set to 0.01) are shown as letters whose size corresponds to the percentage of difference with their average proteome occurrency. Residues shown in pink were never observed at that position.

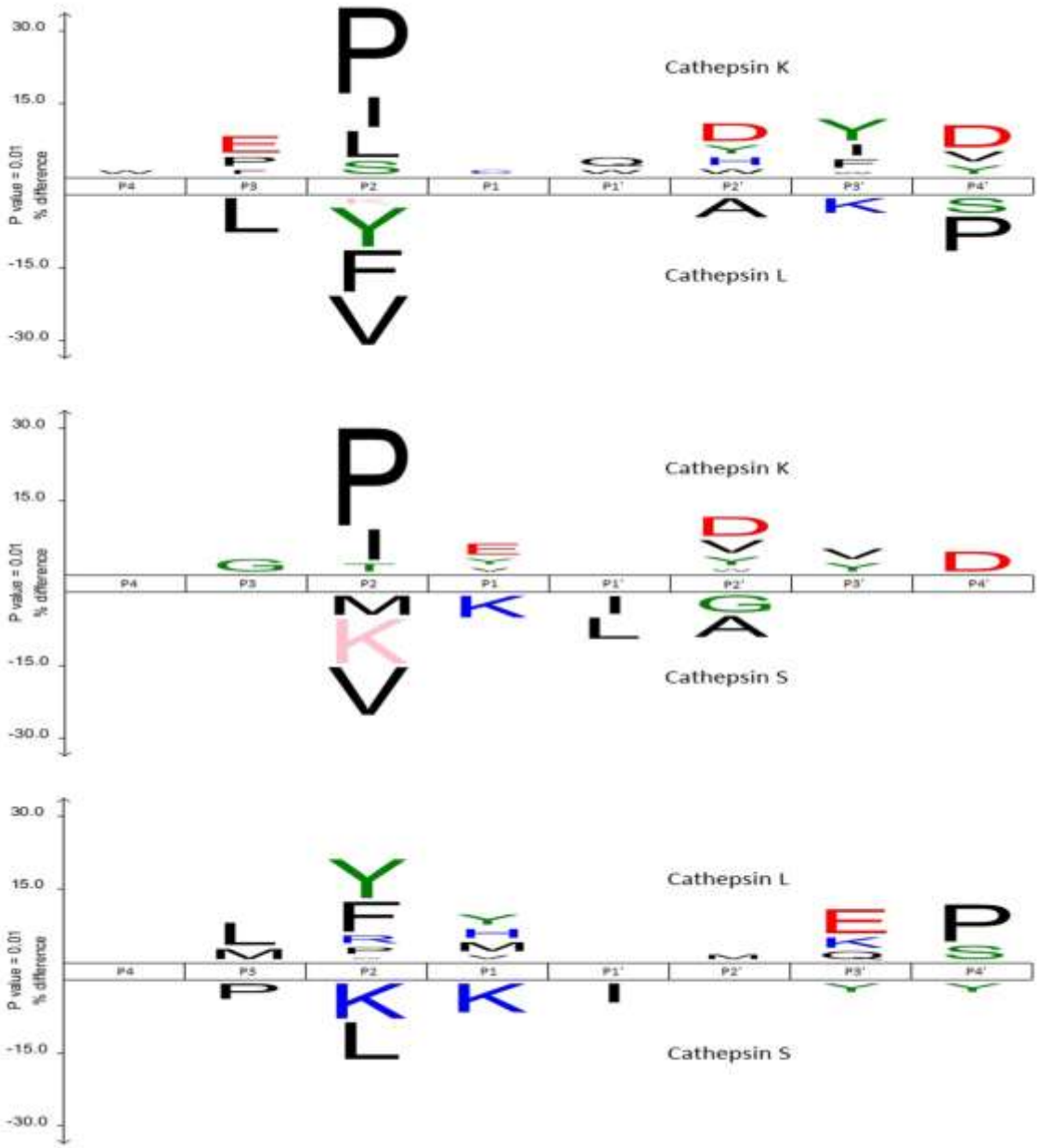

Figure 5: Comparison of the cleavage site specificity detected with FPPS and COFRADIC.

IceLogos based on all cleavage sites detected with cathepsin K, L and S after N-terminal COFRADIC analysis. (B) Comparison between FPPS and COFRADIC data revealed almost identical substrate specificities in the P4 – P4' positions. At each position surrounding the cleavage site, significantly under- and over-represented amino acids (p-value was set to 0.01) are shown as letters whose size corresponds to the percentage of difference with their average proteome occurrency. Residues shown in pink were never observed at that position.