SSPaQ: A Subtractive Segmentation Approach for the Exhaustive Parallel Quantification of the Extent of Protein Modification at Every Possible Site

Guillaume Gabant, Alain Boyer, Martine Cadene

SSPaQ: A subtractive segmentation approach for the exhaustive parallel quantification of the extent of protein modification at every possible site

*Guillaume Gabant, Alain Boyer, Martine Cadene**

Centre de Biophysique Moléculaire, Centre National de la Recherche Scientifique, UPR 4301, affiliated with University of Orléans, rue Charles Sadron, 45071 Orléans cedex 2

Running title: Subtractive parallel quantification of modifications

Keywords: Protein modification, covalent ligand, PTM, quantification, mass spectrometry

* Address reprint requests to:

Martine Cadene, CBM, CNRS, Rue Charles Sadron, CS 80054, 45071 Orléans, France. Tel: +33-238-25-5624, Fax: +33-238-63-1517. E-mail: martine.cadene@cnrs-orleans.fr.

Abstract

Protein modifications, whether chemically induced or post-translational (PTMs), play an essential role for the biological activity of proteins. Understanding biological processes and alterations thereof will rely on the quantification of these modifications on individual residues. Here we present SSPaQ, a subtractive method for the parallel quantification of the extent of modification at each possible site of a protein. The method combines uniform isotopic labeling and proteolysis with MS, followed by a segmentation approach, a powerful tool to refine the quantification of the degree of modification of a peptide to a segment containing a single modifiable amino acid. The strength of this strategy resides in: 1) quantification of all modifiable sites in a protein without prior knowledge of the type(s) of modified residues; 2) insensitivity to changes in the solubility and ionization efficiency of peptides upon modification; and 3) detection of missed cleavages caused by the modification for mitigation. The SSPaQ method was applied to quantify modifications resulting from the interaction of human Phosphatidyl Ethanolamine Binding Protein 1 (hPEBP1), a metastasis suppressor gene product, with locostatin, a covalent ligand and antimigratory compound with demonstrated activity towards hPEBP1. Locostatin is shown to react with several residues of the protein. SSPaQ can more generally be applied to induced modification in the context of drugs that covalently bind their target protein. With an alternate front-end protocol, it could also be applied to the quantification of protein PTMs, provided a removal tool is available for that PTM.

Introduction

Protein modifications play a crucial role in living organisms. By changing the physicochemical properties of proteins through chemical and biochemical reactions, modifications modulate protein function. Protein modifications can be the result of a cell process in the form of post-translational modifications (PTMs). Broad cellular functions including signaling and metabolism rely on PTMs that are produced from natural endogenous molecules and are finely regulated. Alternatively, modifications can be induced, intentionally or not, by exposure to any number of exogenous environmental or man-made agents. Induced modifications are usually the result of exposure to stress, infectious agents, food, drugs, lifestyle-derived xenobiotics such as tobacco and alcohol, pollutants, and radiations [1]. Developmental dysfunctions and human disease may arise as a consequence of PTM disorders and/or induced chemical modifications [2].

The quantification of protein modifications is of great interest for the understanding of biological processes and for clinical research investigations. Covalent drugs represent an ancient class of therapeutic agents whose appeal has been renewed in recent years [3]. The quantification of the extent of a residue modification by these drugs gives access to the addition kinetics. The quantification of *in vitro* induced modification is of great interest for recombinant therapeutic proteins such as antibodies, as oxidation could occur during the manufacturing process and storage, leading to the inactivation of the protein [4]. For PTMs, the determination of the extent of modification gives a measure of the effect of enzyme inhibitors (e.g. anti-kinase drugs) on modification (in this example, phosphorylation) at a given site. Today, a concerted effort is made to evaluate the impact of modifications, whether induced by a covalent drug or undergone by a biodrug, on treatment efficacy and human safety [4, 5].

Mass spectrometry (MS) is a powerful technique for the detection, localization and quantification of modifications [6]. Although it is hoped that top-down methods will someday solve the proteoform riddle, the complete separation of isoforms they require to prevent the loss of site filiation is still elusive at this point. Thus, a majority of efforts to this day focus on bottom-up methods, i.e. methods based on proteolysis of the target protein prior to MS analysis. Several approaches have been described to determine the degree of modification of proteins in terms of site occupancy. A number of studies have used the relative signal intensities of the modified peptide and its corresponding nonmodified form to infer a degree of modification. However, since a modification can alter the ionization efficiency of a peptide to an unpredictable degree [7], the extent of modification of a peptide cannot reliably be determined by a simple comparison of these signal intensities.

In a label-free approach, the ion count of modified and nonmodified peptides was titrated over the course of an experiment, thereby allowing for the calculation of normalized signals according to the saturation point [8]. A direct ratio of normalized signals at each time point was then used to quantify the extent of phosphorylation. This approach cannot compensate for any sample-to-sample variation during sample handling and depends on the analysis of a series of samples spanning the gamut to saturation.

Other approaches have used isotopically labeled internal standards to carry out absolute quantification of modified and nonmodified protein in the form of AQUA peptides [9] or PSAQ proteins [10]. The degree of modification was then deduced by calculation from the ratio of absolute quantities. These approaches require the costly synthesis and purification of numerous stable isotope-labeled internal standard peptides or proteins in their modified and nonmodified forms. Furthermore, they require prior knowledge of the nature of the modifications and the type of modified residue(s).

Elucidation of the fraction of modification has also been reported in the case of phosphorylation using a minimum of three different ratios representing protein, phosphopeptide and nonmodified peptide changes based on stable-isotope labeling with amino acids in cell culture (SILAC) [11]. However, the method necessitates that only one residue per peptide is modified. Moreover, SILAC approaches are generally based on Lys and Arg labeling, limiting the choice of proteolytic enzyme to be used.

Finally, another elegant strategy used a combination of deuteroformaldehyde/formaldehyde stable isotope chemical labeling and alkaline phosphatase treatment [12]. The sample of interest was divided into two aliquots for treatment with phosphatase and phosphatase-free control. Following differential chemical labeling of free amines with stable isotopes, both aliquots were recombined. Mass spectrometric analysis of the recombined mixture revealed the degree of phosphorylation by measuring the signal increase from the dephosphorylated peptide of the corresponding phosphopeptide [12]. However, this is a phosphorylation-centric method which cannot be applied to covalent chemical modifications.

In an attempt to circumvent the limitations of existing approaches, we set out to develop a method for a robust, reliable and comprehensive quantification of modification at every modification site. The ideal method should: 1) be capable of quantifying modifications in parallel and without prior knowledge of the type of target residues; 2) be insensitive to changes in the solubility and the ionization efficiency of peptides containing the modified residues; and 3) be able to detect and deal with missed cleavages caused by the modification. It should be universal in that it can be applied to all types of protein modification, and exhaustive, i.e. capable of quantifying all modifiable sites. Two main issues can impede the correct tally of modified residues. First, in a bottom-up approach, several peptide purification or detection factors can affect sequence coverage, so that some modifiable sites may be missed by MS. Second, if modification impairs the solubility and/or the ionization efficiency of peptides, or if

the modification undergoes degradation downstream of the addition step, quantification methods based on the detection of modified peptides may lead to an incomplete map of modified sites. This is a crucial point for quantitative approaches as the failure to identify a modified site is insufficient to claim its non-existence. Finally, the developed tools should allow for the straightforward visualization of the result.

Here we present SSPaQ, a subtractive method for the parallel quantification of the extent of modification at each possible site of a purified protein. In this strategy, we use isotopically labeled nonmodified protein as an internal standard. Uniform ^{15}N or ^{13}C labeling is agnostic with respect to the choice of proteinase for downstream proteolysis and is fairly cheap to implement. This approach is a powerful tool to list all the detectable modified sites on the protein, and to demonstrate the absence of modification on other residues. As a visual tool, it clearly shows, if they occur, the areas were no information could be drawn.

As a concrete example, this method was applied to the interaction of human Phosphatidyl Ethanolamine Binding Protein 1 (hPEBP1) with a covalent ligand, locostatin. The hPEBP1 protein, named according to the current classification of the PEBP family of proteins [13], and also named Raf Kinase Inhibitory Protein (RKIP) in mammals, is a metastasis suppressor gene product in different types of cancers [14, 15]. Locostatin, (S)-4-benzyl-3-crotonyl-oxazolidin-2-one, is the only known antimigratory compound with demonstrated activity towards PEBP1. To date, there is no known X-ray or NMR structure of the hPEBP1-locostatin complex, and the determination of the covalent site of addition of locostatin on hPEBP1 has proved a vexing analytical challenge to several teams. Thus, we have undertaken the quantification of the degree of modification of all modifiable sites of hPEBP1 by locostatin. The method we developed can also be applied to PTMs on the condition that a removal of the modification is feasible, as well as to covalent protein-small molecules complexes such as drug-target complexes or therapeutic protein conjugates.

Methods

Materials and Reagents. Unless otherwise stated, all chemicals used in this study were obtained from Sigma (St. Louis, MO). Locostatin was purchased from Acros Organics. Ammonium acetate and calcium chloride ($CaCl_2$) were procured from Merck (Darmstadt, Germany). Glycine was purchased from Eurobio (Courtaboeuf, France). Sequencing grade endoproteinases Asp-N, chymotrypsin, and trypsin were from Roche. Alpha-cyano-4-hydroxycinnamic acid was obtained from Bruker Daltonics (Bremen, Germany). Acetonitrile and isopropanol were procured from Carlo Erba (Milan, Italy). Formic acid 90% (FA) and trifluoroacetic acid (TFA) were purchased from Fisher Scientific (Illkirch, France). All solvents and buffers were prepared using 18 MΩ ultrapure water (MilliQ reagent grade system, Millipore).

Purification of recombinant hPEBP1. Recombinant human Phosphatidyl Ethanolamine Binding Protein 1 (hPEBP1) overexpressed in BL21DE3 *E.coli* was purified without tag using QAE Sephadex A-50 chromatography, isoelectrofocusing and Blue-Sepharose chromatography, as previously described [16]. The purified protein was stored at -80°C in ultrapure water with 2 equivalents of TCEP.

Interaction of hPEBP1 with locostatin. A solution of hPEBP1 protein at 7.3 µM was incubated with or without 1 mM locostatin for 4 h 30 at 37°C in incubation buffer consisting of 100 mM HEPES, pH 7.7, and 2% acetonitrile. The control and experiment samples were then spiked with a 5 µM solution of ^{15}N labeled hPEBP1 and the excess locostatin immediately removed by micro gel filtration on preconditioned spin columns with a 75 µL bed volume. Preconditioning of the gel filtration phase required 5 cycles of equilibration with 50 µL incubation buffer for efficient elimination of the manufacturer's preserving buffer. Each sample

was loaded onto the spin column and the collection performed immediately with a 30 s centrifugation at 1000 × g.

Denaturation and Enzymatic Proteolysis. At this stage each sample was split into three tubes in order to perform parallel proteolyses. Proteins were thermally denatured for 10 min at 60°C in a solution of 1 M urea, 100 mM glycine, and 20% acetonitrile. Glycine was added during this step to prevent urea-induced carbamylation of lysines, arginines, cysteines, and the N-terminal amine. Denatured proteins were then proteolyzed in parallel for 2 h at a high enzyme:substrate (E:S) ratio, using conditions optimized for Asp-N, chymotrypsin, or trypsin cleavage. For Asp-N, the reaction was carried out at 37°C in 100 mM ammonium acetate, pH 7.0, at a 1:10 E:S ratio. Trypsin cleavage was carried out at 37°C in a solution of 100 mM ammonium acetate-ammonium bicarbonate, pH 7.5, at an E:S ratio of 1:2, while chymotrypsin cleavage was done at 25°C in the same buffer with 10 mM $CaCl_2$, at a 1:10 E:S ratio. Proteolyses were stopped by addition of TFA at a final concentration of 0.1% and immediate desalting/concentration on a C18 ZipTip from Millipore (Billerica, MA, USA). The volume of TFA solution was adjusted to dilute acetonitrile to a final concentration of 2.5% for efficient binding of the peptides on the ZipTip reverse phase.

Mass spectrometry. Unlabeled and ^{15}N labeled hPEBP1 protein samples were analyzed by MALDI-TOF MS. The matrix solution consisted of a saturated solution of 4-hydroxy-α-cyano-cinnamic acid (4-HCCA) in 3:1:2 formic acid:water:isopropanol. Proteins in the micromolar range were prepared by twenty-fold dilution into the matrix solution. The analyte-matrix samples were spotted onto a gold-plated sample probe using the ultra-thin layer method as described [17, 18]. Spots were washed with 0.1% TFA before acquisition. Analyses were performed using an Ultraflex I mass spectrometer (Bruker Daltonics, Germany) equipped with a 337 nm nitrogen laser and a gridless delayed extraction ion source. An accelerating voltage of 25 kV was used and the delay optimized at 250 ns to achieve a mass resolution greater than

1000 over the mass range of interest (10000–20000 Da). A deflection of matrix ions up to 800 Da was applied to prevent detector saturation. Spectra were acquired in the linear positive ion mode by accumulation of 1000-1200 laser shots. Calibration was performed externally using apomyoglobin and cytochrome c ion peaks acquired from a neighboring spot. The instrument was controlled using Bruker FlexControl 3.3 software and MALDI-TOF-MS spectra processed using FlexAnalysis 3.3 software from Bruker Daltonics.

Unspiked control and experiment samples were analyzed by Liquid Chromatography - High Resolution Mass Spectrometry (LC-HRMS). Acquisitions were performed on the UltiMate™ 3000 NanoRSLC System (Dionex, Sunnyvale, CA, USA) connected to a 4-GHz MaXis Ultra High Resolution quadrupole-TOF mass spectrometer (Bruker Daltonics) equipped with an electrospray ion source. The LC loading pump was used for these experiments. The LC-MS setup was controlled by the Bruker HyStar™ software version 3.2. Proteins were desalted online on a Waters MassPREP™, 2.1 x 10 mm, phenyl 1000 Å reverse-phase cartridge and eluted at a flow rate of 500 µL/min using a 5 to 90% gradient of acetonitrile in 0.1% formic acid. High resolution mass spectra were acquired in positive ion MS mode over a 700–4500 *m/z* range with a nebulizer gas pressure of 1.1 bars. The drying gas flow was 3 liter/min, and the temperature was 200 °C. The in-source collision induced dissociation (isCID) parameter was adjusted at 50eV to promote the observation of desolvated forms of protein. The acquisition rate was 1 Hz corresponding to spectra summations of 4504. External calibration was performed with the ESI-L Low Concentration Tuning Mix (Agilent Technologies). Mass spectra were processed and charge-deconvoluted using DataAnalysis 3.1 software (Bruker Daltonics) and the MaxEnt algorithm.

Proteolytic peptides were analyzed by nanoUltraHPLC-nanoESI UHR–QTOF MS. Experiments were performed using the above setup using an online nano-ESI ion source. Peptides were pre-concentrated online on a Dionex Acclaim PepMap100 C18 reverse-phase

pre-column (inner diameter 100 µm, length 2 cm, particle size 5 µm, pore size 100 Å), and separated on a nanoscale Acclaim Pepmap100 C18 column (inner diameter 75 µm, length 25 cm, particle size 2 µm, pore size 100 Å) at a flow rate of 450 nL/min using a 2–35% gradient of acetonitrile in 0.1% formic acid. Chromatographic peaks were about 3-4 s at Full Width at Half Maximum (FWHM), corresponding to a peak width at the base around 8-10 s. Mass spectrometer scans were set at a frequency of 1 Hz in MS mode only. These settings ensure that there is around 10 data points per extracted ion chromatogram (XIC), allowing for an accurate determination of the area under the curve (AUC). Mass spectra were acquired in positive ion mode from *m/z* 50–3000. Lock mass calibration was performed on *m/z* 622 (Hexakis(2,2-Difluoroethoxy)Phosphazene; CAS #: 186817-57-2) and 1222 (Hexakis(1H, 1H, 4H-hexafluorobutyloxy)Phosphazine; CAS #: 186406-47-3).

Skyline MS1 Quantification. MS data from nanoUltraHPLC-UHR-QTOF MS were processed using the open source software Skyline 2.1 [19] to calculate the area under the curve for each peak from XIC. XICs were constructed for all isotopic ion peaks of each peptide with a base peak threshold of 1% and an *m/z* tolerance of 0.005 Da. Each XIC was manually checked in terms of retention time alignment between light and heavy peptides as well as across acquisition replicates. The isotope dot product (idotp), which compares observed and theoretical isotopic distributions, was below 0.95 and 0.89 for light and heavy peptides, respectively. Idotp should be 1.0 for a perfect match. The area for each charge state of each peptide was exported to .csv format and used as an input for our in-house software dedicated to the quantification of modification at each site.

Results and Discussion

Strategy and general workflow for parallel quantification of the degree of site modification

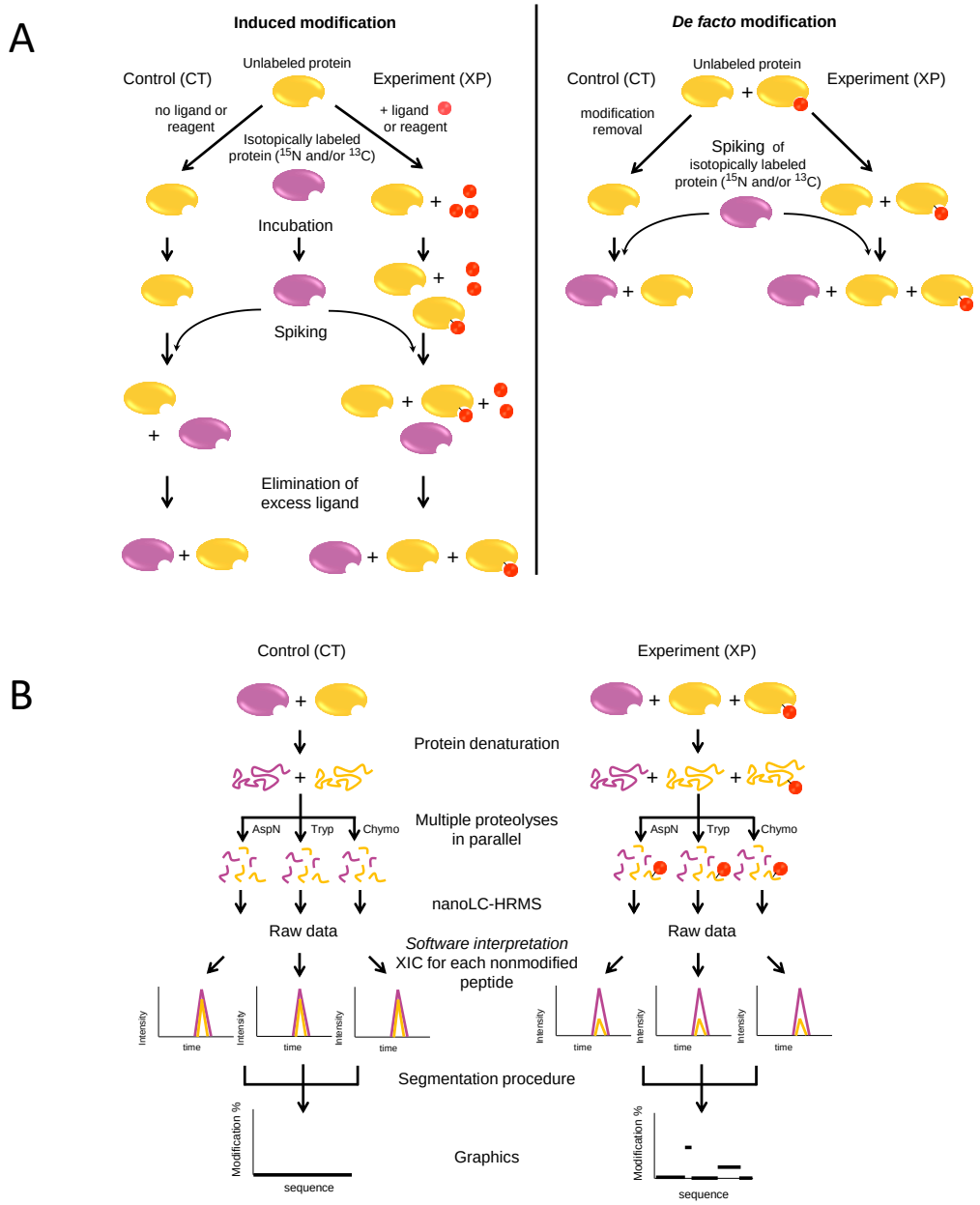
The quantification strategy presented here was developed in the context of an induced modification, i.e. a covalent protein-ligand complex. However, it can also be applied to *de facto* modifications such as PTMs, provided that the modification can be removed. In the case of phosphorylation for example, an enzymatic treatment using lambda-phosphatase can be performed for exhaustive removal of the PTM.

The strength and originality of the subtractive method is that it relies exclusively on signals derived from nonmodified peptides. For the sake of clarity, we describe the logic behind the strategy in the case of an induced modification of the protein, though for removable PTMs the same reasoning applies. The method is based upon the fact that the modification of a given site in a protein automatically leads to a decrease in the pool of protein that is nonmodified at this site. Consequently, the modification produces a proportional decrease of the signal associated with the peptide that includes this site and is left nonmodified in the sample. This phenomenon is even cumulative, as the subtractive effects of modification at two or more sites would add up to a total sum of decrease in the peptide's signal. Thus, by quantifying the decrease of the nonmodified peptide, we can indirectly quantify the increase in modified peptide, and thereby calculate the extent of modification. In the case of chemical modification, the modification is quantified globally, whether or not the modification moiety undergoes degradation downstream of the addition step.

Depending on whether an induced or *de facto* modification is measured, different front-ends are required to produce the samples and their controls for quantification (Figure 1-A). In the case of *de facto* modification such as a PTM, the ^{14}N protein control is generated by removal

of the modification by chemical or enzymatic means, and the experimental sample is left untouched. In the case of induced modifications, the ^{14}N protein in the absence of ligand or reagent is the control. The method then consists in performing three incubations in parallel: ^{14}N protein without ligand (or after modification removal) as control, ^{14}N protein with ligand (or with *de facto* modification) as experiment, and ^{15}N protein without ligand (or after modification removal).

Figure 1. Quantification workflow. A. Preparation of the samples and their controls for quantification of *de facto* or induced modifications. B. Production of peptides and nanoLC-HRMS quantification of the nonmodified peptides.



The ^{15}N protein is subsequently used as an internal standard to compensate for losses during sample processing, for incomplete proteolysis, as well as for variations in the chromatographic injection volume and in ionization. After the initial incubation step, the same volume of internal standard solution is added to the control and experiment solutions. This step is the only source

of uncontrolled variability for quantification and must be done carefully. For induced modifications, there is an extra step where the excess ligand/reagent is removed using a size exclusion spin column to prevent non-specific modification of peptides (Figure 1-A, left panel). At this point both workflows merge (Figure 1-B). The next two steps aim at achieving the most complete proteolysis to prevent bias in the measurement of the extent of modification and to generate optimal peptide signal-to-noise ratios. They also ensure that the modified and nonmodified forms of the target protein undergo proteolysis with the same efficiencies despite possible differences in their tridimensional structure. This is achieved by the combined use of a denaturation step followed by a proteolysis in the presence of denaturing agent at a high enzyme:substrate ratio. Multiple proteolyses are carried out in parallel in order to maximize the coverage of sequence by combining data for different proteinases. Ideally, each and every residue of the protein sequence should be covered by at least one proteolytic peptide, so that no residue can escape the quantitative measurement of its modification. The proteolysis efficiency is assessed through detection of the intact protein by MALDI-TOF MS, and through counting of missed cleavages in product peptides. The mixtures of ^{14}N and ^{15}N peptide solutions are analyzed in triplicate by nanoUltraHPLC–UHR-QTOF MS. Data are then processed using quantification software such as Skyline to identify peptides, to extract related nonmodified ^{14}N and ^{15}N ion intensities for the construction of extracted ion chromatograms (XIC), and to calculate the area under the curve (AUC) of each ^{14}N and ^{15}N XIC peak. By dividing the ^{14}N area with the ^{15}N area, one obtains a standardized area ratio for ^{14}N peptides *versus* the ^{15}N standard. In the experiment, the area lost due to the modification finds itself subtracted from the ^{14}N area, so that upon modification, the ratio drops below the ratio measured in the control. In the case of chemical modification, the modification is quantified as a whole, whether or not the modification moiety undergoes degradation downstream of the addition step.

Data treatment for the calculation of bound peptide fractions

The fraction of modification for a given peptide $f_{bound\ p}$ is calculated using data from the nonmodified peptide in the control (CT) and experiment (XP) sample analyses as follows. Let ${}^{L}A_{p,z}^{ct}$ be the AUC for the ion signal of a given charge state *z* of a given nonmodified peptide *p* in the control, with ${}^{L}A_{p,z}^{xp}$ the corresponding AUC in the experiment. L refers to the light isotopic form of the protein, as opposed to the uniformly isotope-labeled H form. The internal standard added early in the protocol by spiking a solution of uniformly isotope-labeled protein (H) allows for the measurement of the ${}^{H}A_{p,z}^{ct}$ and ${}^{H}A_{p,z}^{xp}$ areas. Standardized area ratios are calculated by dividing the AUC areas from the light and heavy peptide signals (equations 1 and 2), which is done in quantification software, giving the input list of *m/z*, z, and areas.

$$A_{p,z}^{CT} = \frac{{}^{L}A_{p,z}^{ct}}{{}^{H}A_{p,z}^{ct}} \qquad (1)$$

$$A_{p,z}^{XP} = \frac{{}^{L}A_{p,z}^{xp}}{{}^{H}A_{p,z}^{xp}} \qquad (2)$$

Triplicate results are averaged to give the averaged standardized area $\bar{A}_{p,z}^{CT}$ and $\bar{A}_{p,z}^{XP}$. To assess the quality of the data based on measurements in the control, the dataset mean m($\bar{A}_{p,z}^{CT}$) and associated standard deviation $\sigma(\bar{A}_{p,z}^{CT})$ are calculated. Values of $\bar{A}_{p,z}^{CT}$ in the control that deviate by more than $2\sigma(\bar{A}_{p,z}^{CT})$ from the mean are considered outliers and the corresponding charge states eliminated from the CT and XP datasets. Then, the fraction of bound peptide in the experiment and in the control can be calculated by dividing these standardized areas by the standardized area in the control, which is equal to the total amount of peptide, and subtracting this ratio from the ratio obtained in the control, which is 1 by definition (equations 3 and 4):

$$f_{bound\ p,z}^{CT} = \frac{\bar{A}_{p,z}^{CT}}{\bar{A}_{p,z}^{CT}} - \frac{\bar{A}_{p,z}^{CT}}{\bar{A}_{p,z}^{CT}} = 1 - \frac{\bar{A}_{p,z}^{CT}}{\bar{A}_{p,z}^{CT}} \qquad (3)$$

$$f_{bound\ p,z}^{XP} = \frac{\bar{A}_{p,z}^{CT}}{\bar{A}_{p,z}^{CT}} - \frac{\bar{A}_{p,z}^{XP}}{\bar{A}_{p,z}^{CT}} = 1 - \frac{\bar{A}_{p,z}^{XP}}{\bar{A}_{p,z}^{CT}} \qquad (4)$$

The propagated error associated with these calculations is given by equations 5 and 6

$$\sigma_{p,z}^{CT} = f_{bound\ p,z}^{CT} \times \sqrt{2 \times \left(\frac{\sigma(\bar{A}_{p,z}^{CT})}{\mathrm{m}(\bar{A}_{p,z}^{CT})}\right)^2} \qquad (5)$$

$$\sigma_{p,z}^{XP} = f_{bound\ p,z}^{XP} \times \sqrt{2 \times \left(\frac{\sigma(\bar{A}_{p,z}^{CT})}{\mathrm{m}(\bar{A}_{p,z}^{CT})}\right)^2} \qquad (6)$$

In this experimental setup, the determination of $f_{bound\ p,z}^{XP}$ is based solely on applying ratios of standardized areas with the control data, so that it is insensitive to errors in light and heavy protein concentration measurements. In fact, the heavy protein concentration does not even have to be equal to the light protein concentration.

The bound fraction results from the *n* different charge states *z* of a given peptide *p* are averaged before further data treatment:

$$f_{bound\ p}^{CT} = \frac{1}{n} \sum_{i=1}^{n} (f_{bound\ p,z}^{CT})_i \qquad (7)$$

$$f_{bound\ p}^{XP} = \frac{1}{n} \sum_{i=1}^{n} (f_{bound\ p,z}^{XP})_i \qquad (8)$$

with the propagated errors calculated using equations 9 and 10:

$$\sigma_p^{CT} = \frac{1}{n} \sqrt{\sum_{i=1}^{n} \left(\sigma_{p,z}^{CT}\right)^2} \qquad (9)$$

$$\sigma_p^{XP} = \frac{1}{n} \sqrt{\sum_{i=1}^{n} \left(\sigma_{p,z}^{XP}\right)^2} \qquad (10)$$

In the control sample, as the protein is not modified, f_{bound} values are expected to be equal to zero for all peptides. Even if some peptide bonds suffer from suboptimal cleavage efficiency, the effect will be compensated by a correspondingly suboptimal efficiency for the internal standard. Finally, a graphical representation of the degree of modification of each peptide based on equation (11) provides a comprehensive view of the modified, nonmodified and nonobserved sequences of the protein.

$$f_{bound\ p}^{XP} = \text{f (sequence)} \qquad (11)$$

In the remainder of the paper, unless otherwise noted, experimental values are shown for the hPEBP1-locostatin complex.

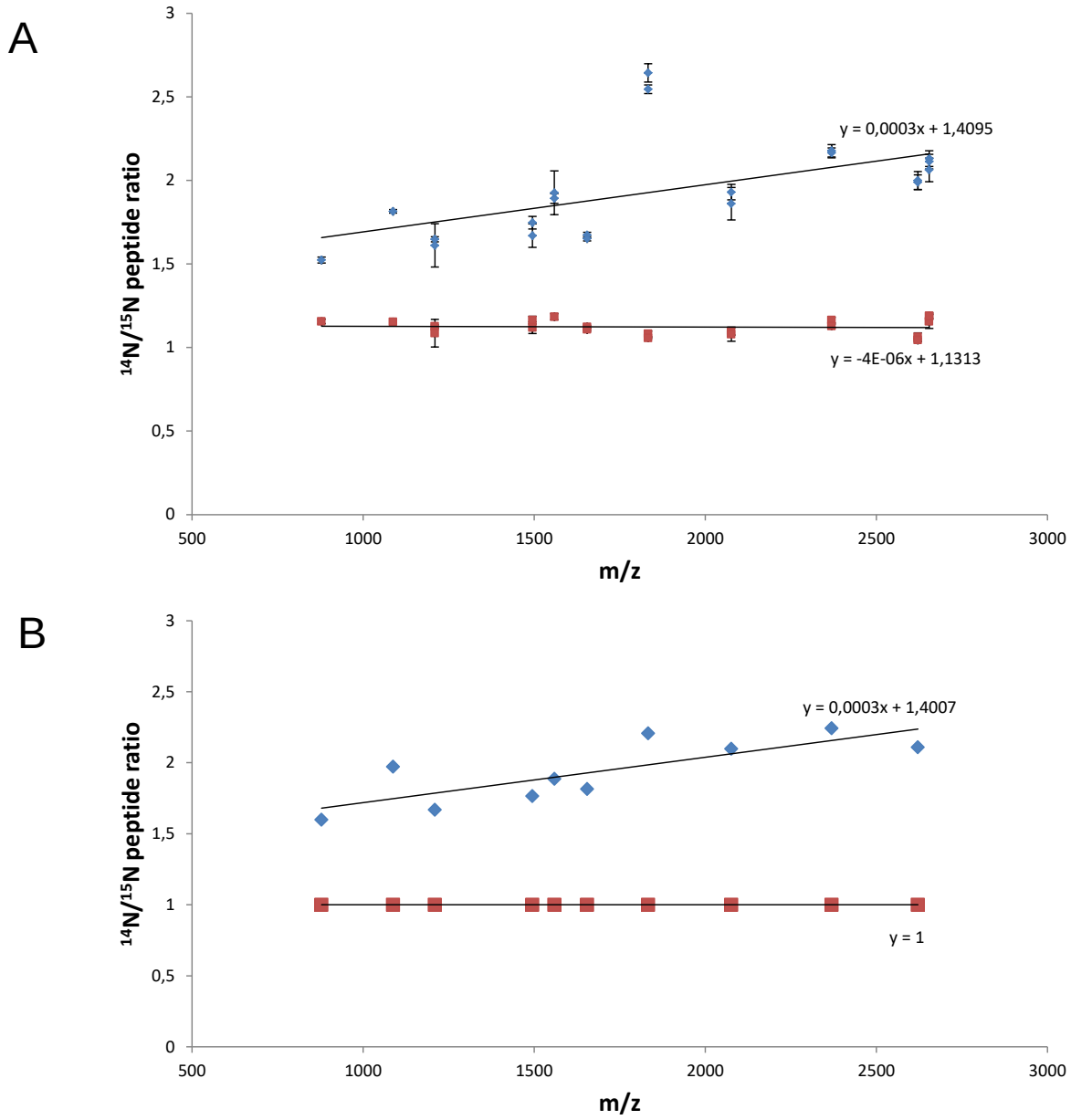
Evaluation of ^{15}N metabolic labeling and experimental considerations

To validate the use of the labeled protein as internal standard, isotopic metabolic labeling efficiency must be measured at the beginning of the workflow. Here, metabolic labeling with stable ^{15}N nitrogen isotope was performed. Metabolic incorporation was evaluated by MALDI-TOF MS analysis by measuring the delta mass between the labeled and nonlabeled protein peaks (Figure S1). The observed masses of ^{15}N hPEBP1 and ^{14}N hPEBP1 are 21,160 Da and 20,924 Da, respectively. The mass increase caused by ^{15}N labeling is 236 Da. Since there are 256 nitrogen atoms in the protein, the isotope incorporation efficiency is 92.2%. The incorporation efficiency of this sample, which was originally prepared for structural biology studies, is thus rather low. As a consequence, ^{15}N-labeled peptides will have a broad isotopic peak distribution.

To evaluate quantification errors, Asp-N peptides from a 1:1 mixture of ^{14}N / ^{15}N hPEBP1 protein solutions were analyzed in decaplicate by nanoUltraHPLC-UHR-QTOF MS (MaXis,

Bruker), and the $^{14}N/^{15}N$ ratios for each peptide were plotted as a function of *m/z* (Figure 2-A, blue dots).

Figure 2. Influence of the number of isotopic distribution peaks used for area calculation on the quantification result. The $^{14}N/^{15}N$ ratios for each peptide were calculated using Skyline software and plotted as a function of *m/z*. A. Asp-N peptides from a 1:1 mixture of $^{14}N/^{15}N$ hPEBP1 protein solutions were analyzed in decaplicate by nanoUltraHPLC-UHR-QTOF MS. B. Theoretical relative isotopic abundances of the Asp-N peptides of hPEBP1. In blue: the first three peaks of the monoisotopic distribution are considered. In red: The whole monoisotopic distribution (isotope > 1 %) is considered.



As is the case for many quantification strategies, only the first three isotopic peaks of the monoisotopic distribution, a0, a1, and a2, were initially taken into account to calculate the AUC for $^{14}N/^{15}N$ peptide ratio determination. From Figure 2-A, it is immediately obvious that the $^{14}N/^{15}N$ ratio increases as a function of *m/z*. Furthermore, the ratio values range from 1.5 to 2.5, far from the expected value of 1. The problem with this approach is that it relies on the assumption that ^{15}N incorporation is 100% complete. When the isotope incorporation is

incomplete as is the case here, the contribution of the first three isotopic peaks to the overall distribution will depend on the molecular composition. To evaluate the effect of less-than-100% isotope incorporation efficiency, we compared the $^{14}N/^{15}N$ peptide ratios calculated using the first three isotopic peaks *versus* all isotopic peaks in the distribution (Figure 2-A, blue *versus* red dots). When the whole distribution is used, the regression curve is practically a horizontal line with a y-intercept at 1.13, i.e. much closer to the expected value of 1. The remaining deviation of 0.13 can be explained by errors in ^{14}N and ^{15}N protein concentrations combined with pipetting errors. The use of the first three peaks of the monoisotopic distribution in a context of incomplete incorporation of 15N (92.2% in our case) leads to a bias on the measured ratio which increases as a function of *m/z*, as shown in Figure 2-B. Equations 1 and 2 allow for error calculation expressed as accuracy and precision percentages:

$$Accuracy = \frac{measured\ ratio - expected\ ratio}{expected\ ratio} \times 100 \qquad (12)$$

$$Precision = \frac{\sigma\ (measured\ ratio)}{\mu\ (measured\ ratio)} \times 100 \qquad (13)$$

Equations 12 and 13 were applied to the measured ratios in Figure 2A, where the expected ratios are all equal to 1. This way we can evaluate how the selection of 3 peaks *versus* all peaks in the distribution affects the error on measured ratios when incorporation is incomplete. In this data set, a 3-peaks selection results in accuracies ranging 52.3 – 164.3 % and a 0.6 – 8.1 % precision. The use of the whole monoisotopic distribution decreases these ranges to 4.4 – 18.9 % accuracy and 0.4 – 7.6 % precision, respectively. As ^{15}N incorporation is rarely 100% complete, particularly for samples prepared for structural biology rather than specifically for mass spectrometry, we propose to systematically use the whole isotopic distribution to ensure a reliable and accurate quantification, not only for this particular method, but for all methods relying on isotopically labeled standards. The effects of the degree of modification and of the light-to-heavy protein concentration ratio, respectively, on the precision and accuracy of f_{bound} measurement are presented in Supplementary Material (Figures S2-S5).

Graphical representation of parallel peptides quantification

The hPEBP1 protein was incubated for 5 hours with or without locostatin and cleaved with Asp-N, chymotrypsin and trypsin. The proteolysis efficiency for each proteinase was assessed by checking for the absence of intact protein after proteolysis (data not shown). After nanoUltraHPLC-UHR-QTOF MS analysis, the extent of sequence coverage and the number of missed cleavages were checked. The quantification software Skyline was used to derive ^{14}N / ^{15}N peptide ratios, and the $f_{bound\ p}$ for each peptide was calculated as described above.

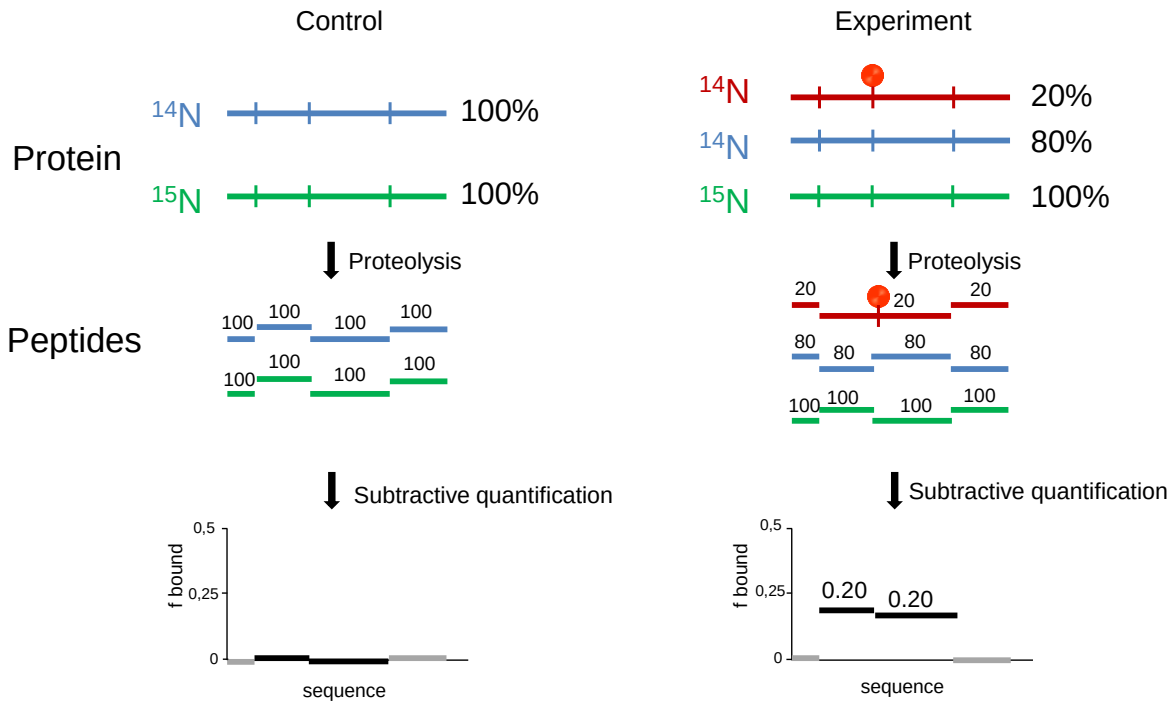
The bound fractions obtained in the presence of locostatin for proteinases Asp-N, chymotrypsin and trypsin are presented Figure S6. The standard deviations calculated from the control datasets were used to assess the likelihood that a f_{bound} significantly differs from zero. At twice the standard deviation, there is a risk of 5% to consider a f_{bound} as different from zero while it is not. Upper and lower threshold values corresponding to +/- $2\sigma_p^{CT}$ are represented in grey for each proteinase in Figure S6. Since the propagated deviation σ_p^{CT} depends on the number of charge states to average for each peptide, its value varies locally. The $f_{bound\ p}^{XP}$ values outside these threshold values are considered as significantly positive or negative.

At this stage, all f_{bound} $f_{bound\ p}^{XP}$ values different from zero were expected to correspond to modified peptides, and thus to have positive values. However, Figure S6 shows that some $f_{bound\ p}^{XP}$ values are significantly negative. This is the case for $f_{bound\ p}^{XP}$ values for the [56-69] and [175-187] Asp-N peptides, the [159-181] chymotrypsin peptide, and the [63-77], [83-93], [120-132], [133-141], [162-179] tryptic peptide. The significance of this observation will become apparent in the next section.

Detection of missed and shifted cleavages caused by the modification

In a typical proteolysis, a missed cleavage (MC) can be caused by the close proximity of cleavage sites, or by a sequence-specific hindrance of substrate binding to the proteinase. Trypsin proteolysis, for example, is hindered by the presence of a proline, lysine or arginine residue in the P1′ position, or of a negatively charged residue in the vicinity of the cleavage. In our protocol, the ^{14}N and ^{15}N proteins undergo cleavage at the same rates along the sequence. However, if a cleavable residue is modified in the experiment sample, it is should no longer be susceptible to cleavage, thus leading to a missed cleavage caused by the modification (MCm). Such MCm must be distinguished from the classical MC described above as contrary to an MC, a MCm leads to a quantitative decrease of both of the nonmodified peptides flanking the modified residue. In terms of quantification, the MCm gives the peptide adjacent to the modification an artificially high f_{bound} equal to the f_{bound} of the peptide which is really modified. The specific pattern of peptide f_{bounds} generated by this behavior is illustrated by the theoretical scheme in Figure 3.

Figure 3. **Bound fraction pattern caused by a missed cleavage caused by the modification.** This scheme is valid for both N- and C-endoproteinases. Protein and peptides are represented in color with associated abundances. The corresponding f_{bounds} are calculated based on changes in nonmodified peptides and graphed at the bottom.

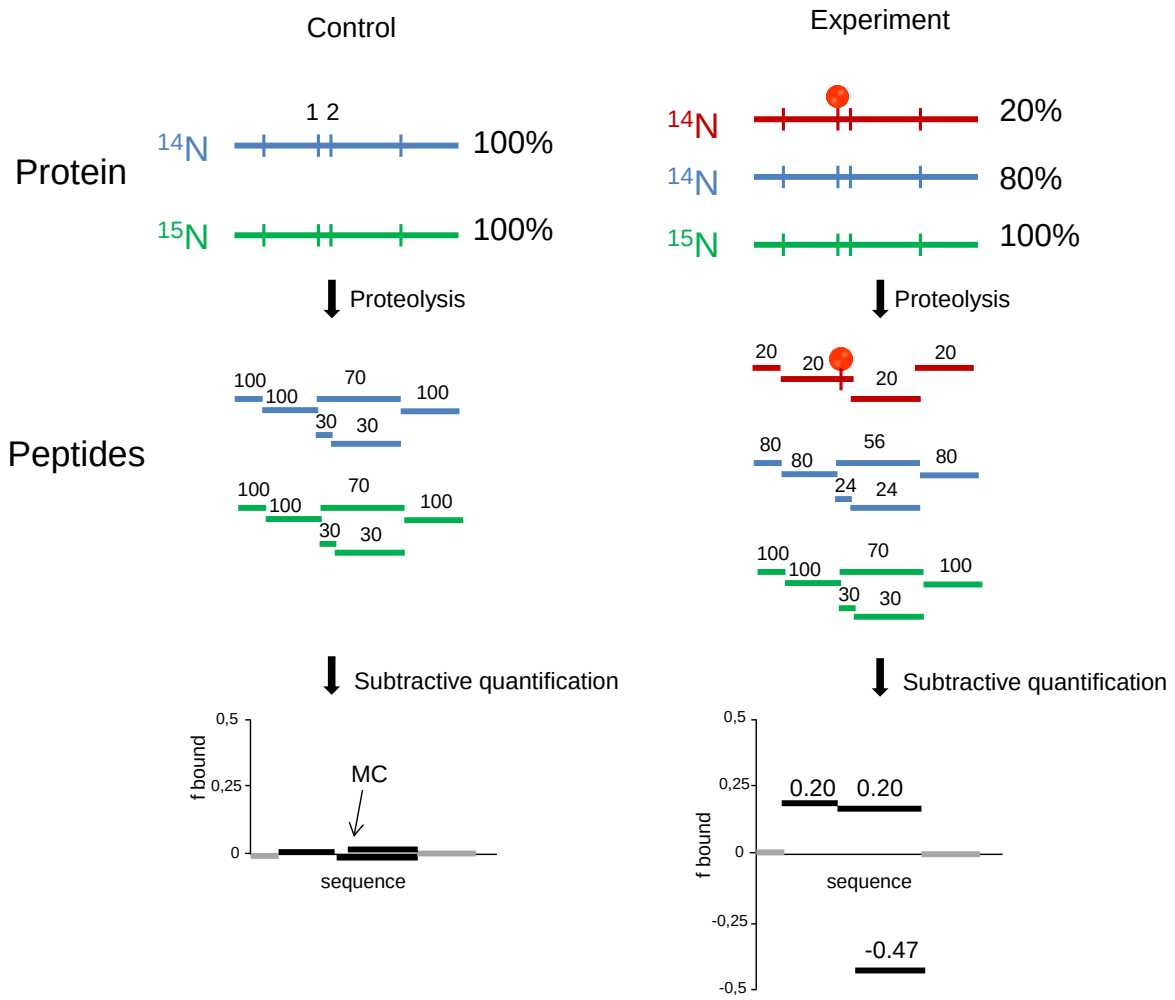


This pattern was observed experimentally in the trypsin graph, specifically with peptides [40-47] and [48-62] (Figure S6) which are detected as two contiguous peptides with near identical fractions of modification (0.042 and 0.046). This interpretation is corroborated by the fact that the residue in position 47 is a lysine, a residue which is both modifiable and cleavable. Of the two adjacent peptides, the data of the peptide that does not contain the cleavable residue, i.e. [48-62] in this example, should not be retained. From a qualitative point of view, since the proteinase cannot cleave a modified residue, a MCm constitutes a clue that the residue where it occurs is indeed modified. Lys47 can thus be proposed to be modified by locostatin.

Although the presence of MCm could be anticipated, negative f_{bound} values such as those described in the preceding section, were not. We thus set out to find the causal factor for this unexpected behavior. The presence of an interference in the isotopic distributions was first checked as described [20] and ruled out. If the observed negative f_{bound} is due to a change in the ^{14}N / ^{15}N area ratio, it can only mean that the ^{14}N / ^{15}N area ratio in the experiment increases relative to the ratio in the control, which is constant by design.

To explain a relative increase of the experiment ^{14}N / ^{15}N area ratio, we hypothesized the existence of a shifted cleavage (SC). The definition of a SC between experiment and control is a modification-induced difference in the proteolysis efficiency at a site adjacent to the modification. In other words, the SC involves the presence a cleavable residue in the immediate vicinity of a modified cleavable residue. In the control, the close proximity of the two cleavable residues leads to classical MC, so that cleavage at site #1 is much more efficient than at site #2 (see theoretical data Figure 4). In the experiment, based on the rational hypothesis that a cleavable residue which is modified can no longer be cleaved, the modification of the cleavable residue prevents cleavage, producing a MCm at site #1, and more importantly, promotes cleavage at site #2, thus increasing the ^{14}N / ^{15}N area ratio of the corresponding peptide. A recognizable pattern of three peptides is produced, two of which are adjacent with positive f_{bound} values due to the MCm, while the shorter third peptide has a negative f_{bound} (Figure 4). There is symmetry between the positive and negative f_{bound} values if and only if the cleavage is spread 50-50% between adjacent sites.

Figure 4. **Bound fraction pattern caused by a shifted cleavage.** This scheme is valid for both N- and C-endoproteinases. Protein and peptides are represented in color with associated abundances. The corresponding f_{bounds} are calculated using non modified peptides and graphed at the bottom. Two cleavable and modifiable sites separated by two residues are shown: site 1 and site 2. In the control, site 1 is completely cleaved while a missed cleavage of 70% is considered at site 2. In the experiment, 20% of 14N protein is modified at site 1, so that there is a MCm at site 1 and there is no more MC at site 2 in this modified population of protein. The di-peptide generated after cleavages at sites 1 and 2 is not represented in the bottom graphs because of its short length which impairs its detection by MS.



This SC pattern is observed for the [63-80], [81-93] and [83-93] peptides. A MC at Lys82 may be caused by the neighboring cleavable Lys80. If Lys80 is modified, it can no longer be cleaved, and a MCm is generated. The cleavage of Lys82 will thus be promoted, thereby increasing the area of the ^{14}N [83-93] peptide in the experiment and leading to the observation of a negative f_{bound}. The existence of a SC at Lys82 is corroborated by the relative increase of the ^{14}N [83-93] area compared to its ^{15}N counterpart in the experiment, while both ^{14}N and ^{15}N [83-93] areas are predictably small in the control. The negative value for peptide [63-77] may also be due to this SC. Because some areas lack coverage near peptides with negative f_{bound} values, the SC pattern may be missed. Thus, this interpretation cannot be applied unambiguously to peptides [120-132], [133-141], [162-179], and [159-181]. However, these negative f_{bound} cannot be taken at face value and were not used for subsequent data processing. At this stage, a lot of overlapping peptides both within and between proteolyses remain (Figure S6), preventing a straight interpretation of the data.

Narrowing down the modified positions

Usually, the quantification of modifications is performed at the peptide level, whereas the quantification of interest for the biologist is at the amino acid site level. The quantitative measurement from a peptide can be narrowed down to a residue only if it is the only modifiable residue in the peptide. If there are several modifiable residues in a peptide, qualitative MS/MS analysis is classically carried out to operate this narrowing down of candidate residues. However, the localization of modification sites is a more difficult task than the detection and identification of modified peptides. This is because MS/MS localization requires a good quality spectrum composed of relatively complete series of fragment ions to unambiguously identify the specific modified amino acid. The low ionization efficiency and/or low abundance of some

modified peptides lead to low signal to noise ratios which affect their selection for fragmentation and yield low fragment ion intensities. In some cases, current interpretation algorithms may actually lead to mislocalization and thus misquantification of the modified site, especially in the case of modification at different positions of co-eluting peptides isomers, i.e. isobaric modified peptides. There are thus great advantages in exploiting MS data for this purpose. The specifics of the experiment presented here led us to propose an original angle to address the problem.

The segmentation approach

Figure S6 shows that, in spite of our efforts to get complete cleavage with maximum sequence coverage, MC still occurred at a significant level, especially in the case of chymotrypsin due to the close proximity of sites. This leads to local ladders which differ by a few residues, sometimes as little as 1, 2, or 3. Modified residues could be contained in these short stretches of sequence difference. This means that there is precious, unexploited information contained in the dataset. To extract f_{bound} data from groups of ladder peptides, we developed a powerful tool, which we call the segmentation approach, to refine at the MS level the quantification of the extent of modification, with the ultimate goal of narrowing down the measurement to a single modifiable amino acid.

Peptides arrangements can be classified into four categories as shown in Scheme I.

Scheme I

Only class A pairs, corresponding to fully overlapping peptides with a common end, can be used for segmentation. The principle of segmentation is that the difference in f_{bound} of two

peptides with a common end is attributable to modification of the sequence that differs between them. It is based on the identification and use of class A peptides, within a given proteolysis and/or from different proteolyses, to deduce a f_{bound} value of the complementary sequence as shown in Scheme 2.

Scheme II

If segmentation generates more than one f_{bound} per segment of sequence, these f_{bound} values are averaged and assigned to this segment for the next cycle. The segmentation process can be iterated by reusing the complementary sequence for segmentation. Non-segmentable sequences are propagated to the next cycle until the end of the process. The segmentation process stops when only class B, C and D peptides remain. A script programmed to compute f_{bound} values is used for this step as well as subsequent steps in data interpretation. To prevent the generation of artificially high f_{bound} values, all artefacts caused by MCm and SC are removed manually from the dataset before launching the segmentation routine.

Propagation of error is expected to occur as a result of the simple math used in the segmentation process. The propagated error will depend on the number of applied segmentation cycles that generate subtractions and sometimes averages, and thus vary along the sequence. Given σ_α and σ_β, the standard deviation for the measurement of $f_{bound\ p}^{XP}$ for two given peptides α and β calculated as described above (σ_p^{XP}), applying a segmentation cycle *seg* generates a propagated error σ_{seg} due to the subtraction of f_{bound} values, which is given by equation 12:

$$\sigma_{seg} = \sqrt{\sigma_{\alpha}^2 + \sigma_{\beta}^2} \qquad (12)$$

Similarly, given σ_{γ} and σ_{δ}, the standard deviation associated with $f_{bound\ p}^{XP}$ for two segments of identical sequence, averaging the two f_{bound} values generates a propagated error σ_{ave} due to addition and to multiplication by a constant of $\frac{1}{2}$, according to equation 13:

$$\sigma_{ave} = \frac{1}{2} \cdot \sqrt{\sigma_{\gamma}^2 + \sigma_{\delta}^2} \qquad (13)$$

The propagation of error given in equations 12 and 13 was applied at every computation step as needed.

The data from all three proteolyses was combined in Figure S7 to provide a synthetic view of the results. Figure S7 shows a number of class B and C overlapping modified pairs which cannot be easily interpreted. Before applying a final processing step to attempt to improve resolution in the sequence dimension however, the f_{bound} obtained after segmentation can be used to test for coherence in the whole dataset.

Determination of the total modified stoichiometric fraction

Determination of the total modified stoichiometric fraction is performed at the protein level. It corresponds to the sum of the extent of modification at each modified sites. Mass spectrometry is one of the few methods that can simultaneously measure the amplitude of a proteoform signal and calculate the modification stoichiometry in terms of number of sites per protein chain for a given modification type. Contrary to the commonly used total modified fraction, which corresponds to the proportion of modified chains in the total pool of protein chains, the total modified stoichiometric fraction (TMSF) takes into account the stoichiometry of modification (equation 14):

$$\text{Total modified stoichiometric fraction} = \frac{\sum (number\ of\ modification\ of\ a\ proteoform\ n \times signal\ of\ a\ proteoform\ n)}{\text{signal of total protein}} \quad (14)$$

If each site is 100% modified, i.e. if saturation is attained at all sites, the TMSF value tends to the total integer stoichiometry of the modification fixed on the protein. For example, if a protein has 3 modification sites capable of reaching saturation, TMSF tends to 3, whereas the total modified fraction tends to 1.

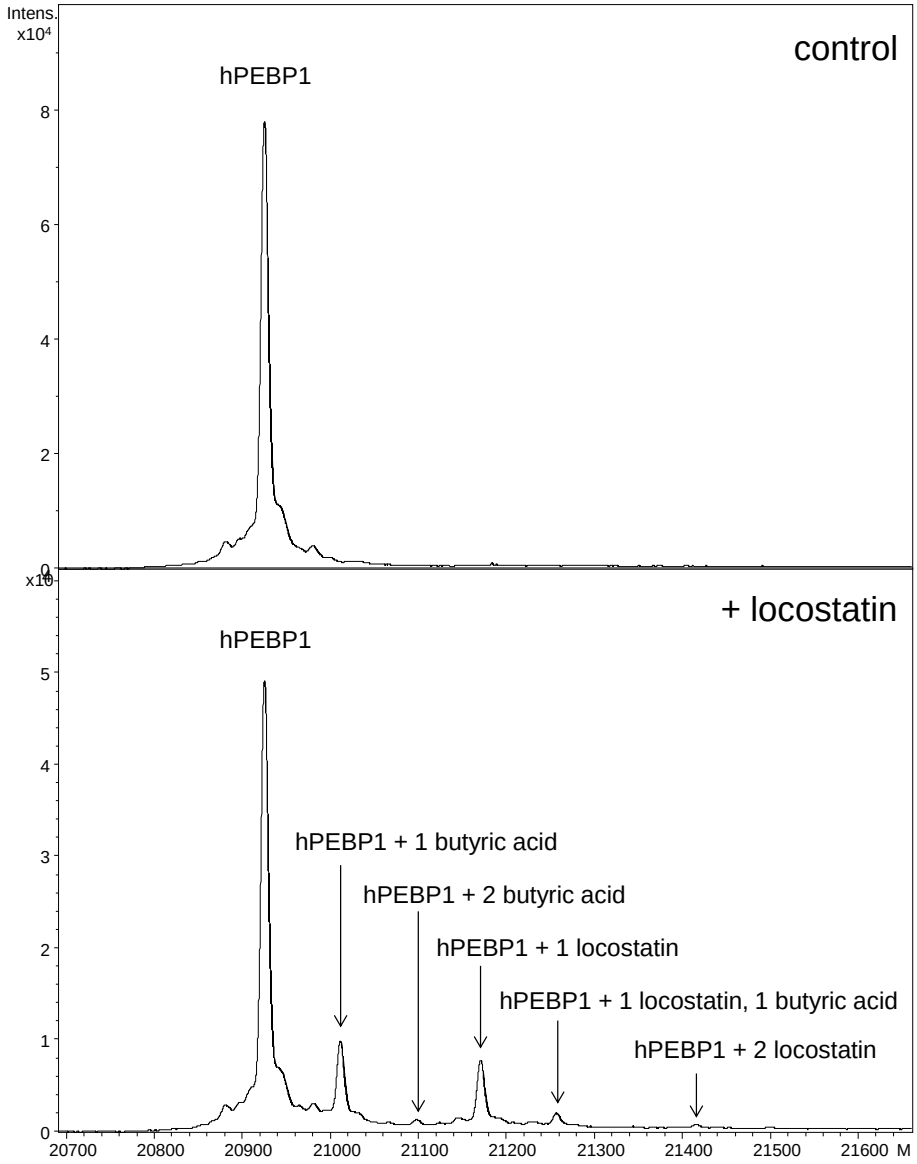
In most cases, the detrimental or beneficial effect of a modification on the ionization efficiency of a whole protein is greatly diluted by the large number of protonation sites on the protein. When using TMSF as a direct quantitative measurement of protein modification, the approximation that different proteoforms have the same ionization efficiency can thus be made. The TMSF measurement may become unreliable if the modifiers are so large and/or their polarity so different from that of the protein that they alter the solubility and/or ionization of the protein. If the instrument itself displays nonlinearity with respect to ion production and/or intensity detection across a wide mass range, large modifiers may also prove problematic.

The point of measuring TMSF is that it can then be compared to the sum of individual f_{bound} measured by parallel quantification in a bottom-up approach such as the subtractive method. The comparison gives a measure of the coherence of the extent of modification found on the whole protein with the data obtained at the peptide level.

The hPEBP1 and locostatin partners were incubated for 5 h at 37°C before removing the excess of locostatin by micro gel filtration. The covalent complex was then analyzed by HRMS (Figure 5). Figure 5 shows a peak at +245.40 Da corresponding to the hPEBP1-locostatine complex which bears a theoretical mass increment of 245.27 Da compared to locostatin. An additional peak at +86.13 Da is detected and corresponds to a hydrolysis product of bound locostatin we previously identified as butyric acid (theoretical mass increment of 86.09 Da) and which is still

covalently attached to hPEBP1 [21]. Taking into account the intensity of the hPEBP1-locostatin and hPEBP1-butyric acid complexes, a TMSF of 0.37 is found. In this experiment, the ^{15}N labeled internal standard was not added because the corresponding peak partially overlaps with the peak of ^{14}N hPEBP1-locostatin complex, thus preventing its detection and quantification.

Figure 5. **Measurement of TMSF for the hPEBP1-locostatin complex.** hPEBP1 and locostatin were incubated for 5 h at 37°C and excess locostatin removed by micro gel filtration. The covalent complex was then analyzed by HRMS with online desalting.



We compared the extent of modification on the whole protein measured as TMSF with the sum of individual f_{bound} values for all segments. Thus for each proteinase, a percentage of recovery of modification at the peptide level is obtained (Table 1).

Table 1. Comparison of the extent of modification on the whole protein measured as TMSF with the sum of individual f_{bound} values after segmentation.

Proteinase	TMSF	Σf_{bounds}	recovery (%)
Asp-N proteolysis	0.37	0.27	73
Chymotrypsin proteolysis	0.37	0.36	97
Trypsin proteolysis	0.37	0.16	43

Since SSPaQ is a subtractive method, the measured f_{bound} values represent the total locostatin binding fractions, whether the molecule stays intact or is partially degraded into butyric acid *in situ* following binding. Thus Σf_{bounds} and TMSF values in Table 1 are directly comparable and should be equal. The closest this was achieved is with chymotrypsin proteolysis. For AspN and trypsin and proteolysis, only 73 and 43% of TMSF is recovered at the peptide level after segmentation. The remainder may correspond to modification that is widely distributed over slightly modified and short segments and/or nonobserved sequences. The origin of the difference in recovery between proteinases is not clearly identified yet, but it may be due, at least in part, to the difference in average length of proteolytic peptides with a resulting difference in f_{bound} values per segment on average.

Local Minimum

As mentioned above, Figure S7 still shows a number of class B and C pairs of overlapping segments between proteolyses, which lead to residual ambiguity. The reason for this is that, contrary to class A pairs, any f_{bound} difference between peptides cannot be unambiguously assigned to a modification located on either side of the sequence in common. While f_{bound} difference calculations are meaningless for class B and C pairs, one piece of information can be derived from the comparison of f_{bound} values in the common sequence, and used to further refine the data as shown hereafter.

The principle of the local minimum is that in the sequence that is common to two or more partially overlapping class B or C pairs or groups, the lowest f_{bound}, i.e. the local minimum (LocMin), gives the highest possible value of the extent of modification in that region. If one of the segments has a higher value, logic dictates that it can only be due to a modification in a nonoverlapping region. As stated above however, the exact f_{bound} of this modification cannot be calculated from the difference since these are not a class A pair. The Local Minimum concept is thus simply based on the selection, for each residue, of the minimum value of f_{bound} found for all segments containing this residue. Results with LocMin treatment applied on the 3 combined proteolyses data are shown in Figure S8. A f_{bound} histogram as a function of sequence is shown in Figure 6. The propagated σ calculated in software was used for each proteinase to assign an error bar to the LocMin of each segment in the figure. The histogram shows that, in spite of careful filtering of the data at several stages of the process, segments with negative values remain. Table 2 summarizes the top 5 modified segments identified, along with modifiable amino acids they contain and the observation of a MCm or SC in the segment.

Figure 6. **Histogram representation of parallel sites quantification for the hPEBP1-locostatin complex.** Error bars correspond to twice the standard deviation.

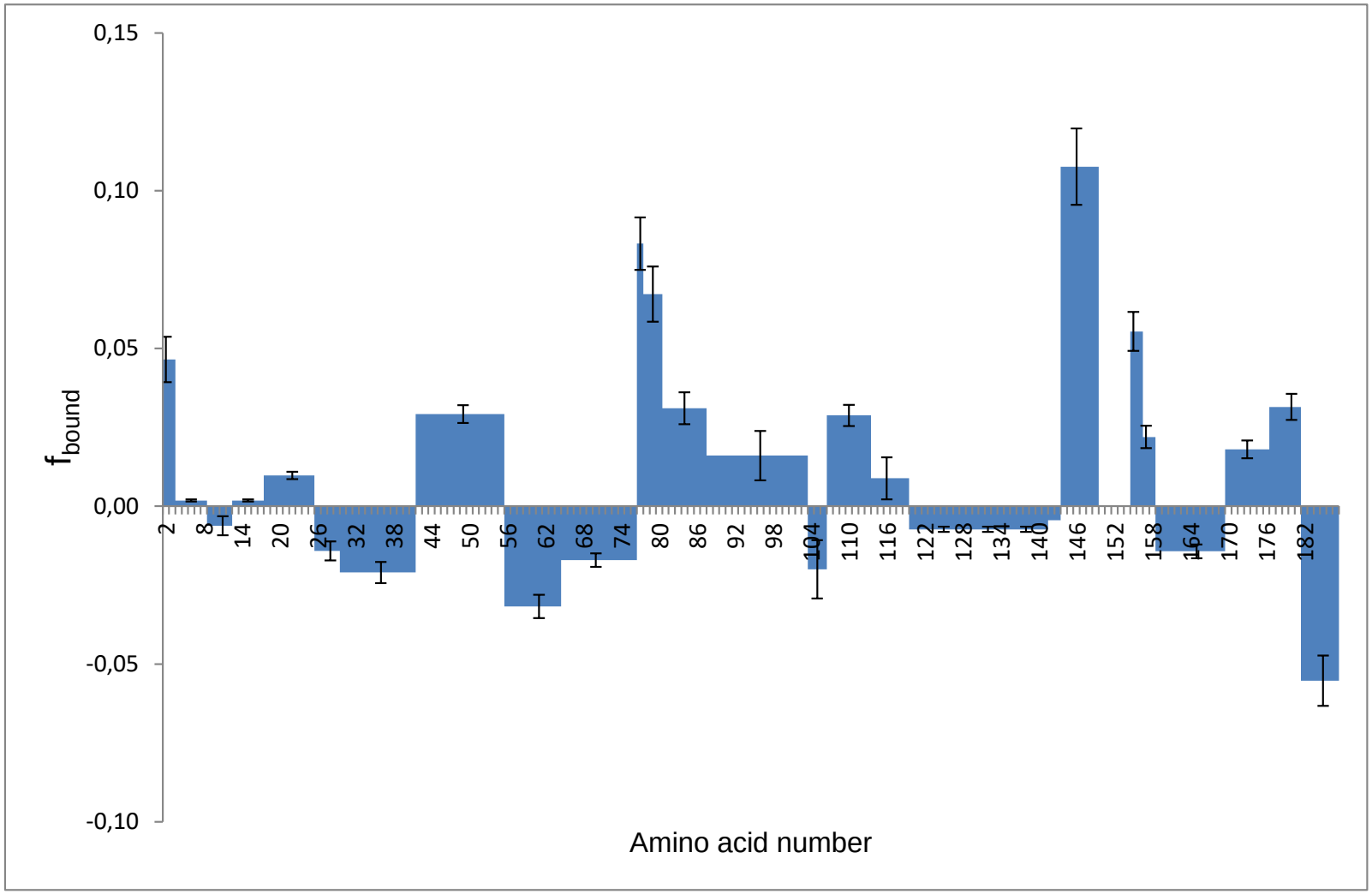


Table 2. Top 5 modified segments identified after LocMin narrowing of modified sequences. Standard deviations are local and depend on the proteolysis dataset which generated the corresponding LocMin f_{bound}.

Sequence segment	LocMin f_{bound}	Modifiable sites	Site leading to MCm or SC due to a modification	Unambiguously identified site
[2-3]	0.047	N-terminal		N-terminal
[77]	0.083	Lys77		Lys77
[78-80]	0.067	Lys80	Lys80	Lys80
[144-149]	0.108	His145, Lys148		
[155-156]	0.055	Lys156		Lys156

Care should be taken in the interpretation of LocMin f_{bound} values, as class B pairs generate a falsely high f_{bound} on at least one of the nonoverlapping sides. The fact that the LocMin f_{bound} shown for these flanking nonoverlapping sides cannot be unambiguously distributed between sides is represented by a small linking curve between the corresponding sequences in Figure S8. In this dataset, an example of such potential problem areas are [4-8] and [13-17] with f_{bound} around zero. LocMin f_{bound} values generated by class C groups can also be overestimated, because modifications leading to these f_{bound} values can be located outside the common sequence for all segments in the group. There may even be cases where the common sequence does not bear any modification. So, while LocMin minimizes errors by representing the upper boundary for the true f_{bound} in the common sequence, this minimization may still not be enough, preventing unambiguous assignment of the measured f_{bound} to a modifiable residue in that sequence. In this dataset, LocMin calculations from class C groups leads to a potentially overestimated f_{bound} for sequences [81-87] and [144-149].

Table 2 shows the 5 most modified segments to consider at the end of the narrowing process. Four segments contain only one modifiable residue: [2-3], [77], [78-80] and [155-156]. The

f_{bound} values observed in these areas can thus be attributed to these sites, i.e. the N-terminal site, Lys77, Lys80 and Lys156, with some level of confidence. For the remaining segment [144-149], two residues may be modified, thus preventing a quantitative measurement at the amino acid level.

Lys47 was identified as modified in the [42-55] segment based on the generation of a MCm with trypsin cleavage. After the segmentation and LocMin processes, the f_{bound} associated with this segment indicates modification at this site is minor.

Biological significance

At a cellular level, locostatin prevents cell migration by binding to PEBP1, as demonstrated by a chemical genetics approach [22, 23]. The unique observation of the *in vitro* anti-migration effect of locostatin, together with the anti-metastasis effect of its target protein, is the basis for the study of the hPEBP1-locostatin complex *in vitro* with the aim to obtain valuable information for the design of new molecules as potential anti-metastatic leads.

The most striking aspect of the parallel quantification of locostatin reaction on hPEBP1 is that no single residue appears with an f_{bound} value that is clearly superior to any other. In the present study, apart from the [150-154] segment, the whole sequence is covered. In the covered area, all the sites modified by locostatin were detected and quantified in a controlled reaction, i.e. in conditions of single-hit statistics with elimination of the excess of locostatin before complete proteolysis. Locostatin binding appears nonspecific as from the top 5 modified segments, at least 4 distinct residues bear a modification. This result explains why, in spite of years of our best efforts as well as several other teams' [24, 25], the locostatin binding site was never clearly identified.

The [150-154] segment, which contains Ser153, a residue which is pivotal to signalization since its phosphorylation by PKC switches rPEBP1 or hPEBP1 from the Raf1-MAP kinase pathway to GRK2 [26], is non-observed in this dataset. Based on the 97% modification recovery obtained with chymotrypsin, this segment could only bear a small fraction of modification at this site. Modification by locostatin followed by its hydrolysis into butyric acid would introduce a negative charge in this area. If located on Ser153, the negatively charged group could mimic phosphorylation at Ser153. Downstream signaling of this "always-on phosphorylation" switch may then favor anti-migratory effects, for example through hPEBP1's ability to modulate the nF-κB pathway [27]. On the other hand, if a butyrate group was located on another residue near Ser153, it could prevent GRK2 kinase binding to Ser153, preventing the switch away from Raf1, thus re-enforcing the inhibition of the Raf1 MAP kinase pathway. This could indirectly trigger Aurora with positive anti-tumoral and anti-metastasis effects. . Interestingly, the [81-87] segment contains Tyr81 and His86, two residues that are part of the evolutionarily conserved anionic-ligand binding pocket of PEBP1 proteins [28]. As mentioned above, the localization of the locostatin binding site has been the subject of several unsuccessful attempts, based on NMR or MS approaches. Shemon et al. (2009) attempted to study the complex between rPEBP1 and locostatine by NMR but the analysis could not be performed because of protein precipitation induced by adding locostatin to high concentrations of rPEBP1. To circumvent this difficulty, these authors used (S)-4-benzyloxazolidin-2-one, the hydrolysis product of locostatin, which does not contain the crotonyl group, and showed that this molecule binds to the anion pocket (Shemon et al., 2009). Although we previously obtained the same result with hPEBP [28], the biological relevance of the non-covalent complex between hPEBP1 and the hydrolysis product of locostatin is not established. Another study using mass spectrometry, attempted the localization of the locostatin binding site on hPEBP1 [25]. However, the authors used a high concentration of hPEBP1 in the presence of an excess of locostatin, leading again to protein

precipitation which is indicative of a profound disruption of the protein structure. Conclusions drawn from complexes obtained in conditions of precipitation are likely to be biologically irrelevant. Nevertheless, MS results were obtained after resolubilization of the insoluble fraction of the hPEBP1-locostatin incubation. From this, the authors pointed to His86, among several modified residues, as the primary target site of locostatin, based on the fact that it is a highly conserved residue. Despite the fact that locostatin, as an electrophile, likely binds to deprotonated nucleophilic residues, these results were used in support of a computational simulation for locostatin binding in the anion pocket [29]. Our results show that modification in the [81-87] segment is in fact minor, and thus residues therein cannot be considered as a specific binding site.

In light of the present quantification study of locostatin modifications on hPEBP1, both functional and structural studies could be undertaken, with the aim to better characterize the binding site responsible for the anti-metastatic activity. The rational design of activators of hPEBP1's natural antimetastasis effect would greatly benefit from these developments. It should be noted that locostatin binding was measured here independent of further degradation upon binding. If separate quantifications of locostatin binding and its subsequent degradation at each site are desired, then a different method should be applied.

The use of uniform isotope labeling makes this parallel quantification method accessible to numerous protein samples made for the purpose of structural biology studies, while the inclusion of all isotopic peaks ensures success even in the case of suboptimal enrichment. The present method can measure f_{bound} with accuracies up to 30%, and precision values increasing with f_{bound}. The reliability of the method thus depends on modifications generating sufficiently high f_{bound} values. In the present study, the individual f_{bound} values are relatively low due to the high number of parallel sites that were modified, while TMSF was kept at a low 0.37 in keeping with single-hit statistics conditions. For proteins with a low number of modified sites, the

relative error should be comparatively low. The method is also essentially modular: segmentation could be omitted if none of the proteolyses generate MCs. Proteolysis results can be combined before or after segmentation, and LocMin can be applied or not depending on the presence of class B and C groups. The script we developed will be made available as open-source code at the time of publication at the following address [url to be made available].

With the modified front-end described in Figure 1, this approach could serve to quantify PTMs, the main limitation being that only one type of PTM can be quantified at a time. Whether quantifying PTMs or induced modifications, the labeled protein should ideally be produced in the same expression system as the unlabeled protein, so as to work with the same background of PTMs as the unlabeled sample. PTMs that are not the focus of the quantification may be removed if they impact the sequence coverage of the protein of interest. Phosphorylation and N-glycosylation for example, can be enzymatically treated in the control and experiment samples before proteolysis.

SSPaQ can be considered as complementary to top-down methods still being developed, and/or original fragmentation approaches such as photodissociation (PD) [30] and electron transfer dissociation (ETD) [31] which may better preserve side-chain modifications than collision induced dissociation (CID). In principle, the method could be applied to protein mixtures instead of purified protein as was done here. One could grow cells with uniform isotopic labeling to create a pool of internal standard proteins. At the expense of exhaustivity of modification sites coverage, whole cell quantification of modification could then be applied using cell-wide parallel quantification of modifications with SSPaQ.

Conclusions

The subtractive method developed here is well adapted to address some of the common issues associated with the quantitation of protein modifications at all sites of protein. Because SSPaQ deals only with nonmodified peptides, it circumvents problems linked to ionization efficiencies and the potential loss of integrity of the modification after covalent addition, while decreased solubility of modified peptides becomes a non-issue. In the case of induced modifications, it can also work without prior knowledge of the types of modified residue, provided that a sufficient set of proteinases is used to cover the whole sequence and that cleavages are achieved at least once between each modified residue. With a slightly different front end and appropriate enzymatic removal tools, it could be applied more generally to PTM quantification.

Acknowledgements

The financial support of FEDER (grant #2699-33931, SyMBioMS), Région Centre Val de Loire, Université d'Orléans and CNRS for high resolution mass spectrometry is gratefully acknowledged. We thank Cancéropole Grand Ouest and Ligue contre le Cancer for funding the production of hPEBP1 and the ANR for a grant to characterize the hPEBP1-locostatin (METASUPP, ANR-08-BLAN-0033). We also thank Yannick Berteaux for helpful discussions while writing the segmentation script and Emmanuelle Mebold for preliminary proteolysis experiments.

References

1. Rappaport, S.M., Smith, M.T.: Environment and Disease Risks. Science. 330, 460–461 (2010).
2. Rubino, F.M., Pitton, M., Di Fabio, D., Colombi, A.: Toward an " *omic* " physiopathology of reactive chemicals: Thirty years of mass spectrometric study of the protein adducts with endogenous and xenobiotic compounds. Mass Spectrom. Rev. 28, 725–784 (2009).
3. Singh, J., Petter, R.C., Baillie, T.A., Whitty, A.: The resurgence of covalent drugs. Nat. Rev. Drug Discov. 10, 307–317 (2011).
4. Torosantucci, R., Brinks, V., Kijanka, G., Halim, L.A., Sauerborn, M., Schellekens, H., Jiskoot, W.: Development of a Transgenic Mouse Model to Study the Immunogenicity of Recombinant Human Insulin. J. Pharm. Sci. 103, 1367–1374 (2014).
5. Johnson, D.S., Weerapana, E., Cravatt, B.F.: Strategies for discovering and derisking covalent, irreversible enzyme inhibitors. Future Med. Chem. 2, 949–964 (2010).
6. Olsen, J.V., Mann, M.: Status of Large-scale Analysis of Post-translational Modifications by Mass Spectrometry. Mol. Cell. Proteomics MCP. 12, 3444–3452 (2013).
7. Gropengiesser, J., Varadarajan, B.T., Stephanowitz, H., Krause, E.: The relative influence of phosphorylation and methylation on responsiveness of peptides to MALDI and ESI mass spectrometry. J. Mass Spectrom. JMS. 44, 821–831 (2009).
8. Steen, H., Jebanathirajah, J.A., Springer, M., Kirschner, M.W.: Stable isotope-free relative and absolute quantitation of protein phosphorylation stoichiometry by MS. Proc. Natl. Acad. Sci. U. S. A. 102, 3948–3953 (2005).
9. Gerber, S.A., Rush, J., Stemman, O., Kirschner, M.W., Gygi, S.P.: Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. Proc. Natl. Acad. Sci. U. S. A. 100, 6940–6945 (2003).
10. Ciccimaro, E., Hanks, S.K., Yu, K.H., Blair, I.A.: Absolute quantification of phosphorylation on the kinase activation loop of cellular focal adhesion kinase by stable isotope dilution liquid chromatography/mass spectrometry. Anal. Chem. 81, 3304–3313 (2009).
11. Olsen, J.V., Vermeulen, M., Santamaria, A., Kumar, C., Miller, M.L., Jensen, L.J., Gnad, F., Cox, J., Jensen, T.S., Nigg, E.A., Brunak, S., Mann, M.: Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. Sci. Signal. 3, ra3 (2010).
12. Wu, R., Haas, W., Dephoure, N., Huttlin, E.L., Zhai, B., Sowa, M.E., Gygi, S.P.: A large-scale method to measure absolute protein phosphorylation stoichiometries. Nat. Methods. 8, 677–683 (2011).
13. Zeng, L., Imamoto, A., Rosner, M.R.: Raf kinase inhibitory protein (RKIP): a physiological regulator and future therapeutic target. Expert Opin. Ther. Targets. 12, 1275–1287 (2008).
14. Lee, H.C., Tian, B., Sedivy, J.M., Wands, J.R., Kim, M.: Loss of Raf kinase inhibitor protein promotes cell proliferation and migration of human hepatoma cells. Gastroenterology. 131, 1208–1217 (2006).
15. Klysik, J., Theroux, S.J., Sedivy, J.M., Moffit, J.S., Boekelheide, K.: Signaling crossroads: the function of Raf kinase inhibitory protein in cancer, the central nervous system and reproduction. Cell. Signal. 20, 1–9 (2008).
16. Jaquillard, L., Saab, F., Schoentgen, F., Cadene, M.: Improved accuracy of low affinity protein-ligand equilibrium dissociation constants directly determined by electrospray ionization mass spectrometry. J. Am. Soc. Mass Spectrom. 23, 908–922 (2012).

17. Cadene, M., Chait, B.T.: A Robust, Detergent-Friendly Method for Mass Spectrometric Analysis of Integral Membrane Proteins. Anal. Chem. 72, 5655–5658 (2000).
18. Gabant, G., Cadene, M.: Mass spectrometry of full-length integral membrane proteins to define functionally relevant structural features. Methods San Diego Calif. 46, 54–61 (2008).
19. Schilling, B., Rardin, M.J., MacLean, B.X., Zawadzka, A.M., Frewen, B.E., Cusack, M.P., Sorensen, D.J., Bereman, M.S., Jing, E., Wu, C.C., Verdin, E., Kahn, C.R., MacCoss, M.J., Gibson, B.W.: Platform-independent and Label-free Quantitation of Proteomic Data Using MS1 Extracted Ion Chromatograms in Skyline: APPLICATION TO PROTEIN ACETYLATION AND PHOSPHORYLATION. Mol. Cell. Proteomics. 11, 202–214 (2012).
20. Zhang, G., Ueberheide, B.M., Waldemarson, S., Myung, S., Molloy, K., Eriksson, J., Chait, B.T., Neubert, T.A., Fenyo, D.: Protein Quantitation Using Mass Spectrometry. Methods Mol. Biol. Clifton NJ. 673, 211–222 (2010).
21. Gabant, G., Beaufour, M., Schoentgen, F., Cadene, M.: A detailed characterization of the interaction between the PEBP/RKIP protein and locostatin, a potential antimetastatic lead.
22. Mc Henry, K.T., Ankala, S.V., Ghosh, A.K., Fenteany, G.: A non-antibacterial oxazolidinone derivative that inhibits epithelial cell sheet migration. Chembiochem Eur. J. Chem. Biol. 3, 1105–1111 (2002).
23. Zhu, S., Mc Henry, K.T., Lane, W.S., Fenteany, G.: A chemical inhibitor reveals the role of Raf kinase inhibitor protein in cell migration. Chem. Biol. 12, 981–991 (2005).
24. Shemon, A.N., Eves, E.M., Clark, M.C., Heil, G., Granovsky, A., Zeng, L., Imamoto, A., Koide, S., Rosner, M.R.: Raf Kinase Inhibitory Protein protects cells against locostatin-mediated inhibition of migration. PloS One. 4, e6028 (2009).
25. Beshir, A.B., Argueta, C.E., Menikarachchi, L.C., Gascón, J.A., Fenteany, G.: Locostatin Disrupts Association of Raf Kinase Inhibitor Protein With Binding Proteins by Modifying a Conserved Histidine Residue in the Ligand-Binding Pocket. Forum Immunopathol. Dis. Ther. 2, 47–58 (2011).
26. Deiss, K., Kisker, C., Lohse, M.J., Lorenz, K.: Raf kinase inhibitor protein (RKIP) dimer formation controls its target switch from Raf1 to G protein-coupled receptor kinase (GRK) 2. J. Biol. Chem. 287, 23407–23417 (2012).
27. Yeung, K.C., Rose, D.W., Dhillon, A.S., Yaros, D., Gustafsson, M., Chatterjee, D., McFerran, B., Wyche, J., Kolch, W., Sedivy, J.M.: Raf kinase inhibitor protein interacts with NF-kappaB-inducing kinase and TAK1 and inhibits NF-kappaB activation. Mol. Cell. Biol. 21, 7207–7217 (2001).
28. Tavel, L., Jaquillard, L., Karsisiotis, A.I., Saab, F., Jouvensal, L., Brans, A., Delmas, A.F., Schoentgen, F., Cadene, M., Damblon, C.: Ligand binding study of human PEBP1/RKIP: interaction with nucleotides and Raf-1 peptides evidenced by NMR and mass spectrometry. PloS One. 7, e36187 (2012).
29. Rudnitskaya, A.N., Eddy, N.A., Fenteany, G., Gascón, J.A.: Recognition and reactivity in the binding between Raf kinase inhibitor protein and its small-molecule inhibitor locostatin. J. Phys. Chem. B. 116, 10176–10181 (2012).
30. Shemon, A.N., Heil, G.L., Granovsky, A.E., Clark, M.M., McElheny, D., Chimon, A., Rosner, M.R., Koide, S.: Characterization of the Raf kinase inhibitory protein (RKIP) binding pocket: NMR-based screening identifies small-molecule ligands. PloS One. 5, e10479 (2010).
31. Shaw, J.B., Li, W., Holden, D.D., Zhang, Y., Griep-Raming, J., Fellers, R.T., Early, B.P., Thomas, P.M., Kelleher, N.L., Brodbelt, J.S.: Complete protein characterization using top-down mass spectrometry and ultraviolet photodissociation. J. Am. Chem. Soc. 135, 12646–12651 (2013).

32. Syka, J.E.P., Coon, J.J., Schroeder, M.J., Shabanowitz, J., Hunt, D.F.: Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. Proc. Natl. Acad. Sci. U. S. A. 101, 9528–9533 (2004).

Table legends

Table 1. Comparison of the extent of modification on the whole protein measured as TMSF with the sum of individual f_{bound} values after segmentation.

Table 2. Final segments identified as modified after LocMin narrowing of modified sequences. The italicized sites in the table correspond to f_{bound} that are higher than σ yet within 2 σ of zero. The non-italicized sites are significantly different from zero.

Figure legends

Figure 1. Quantification workflow. A. Preparation of the samples and their controls for quantification of *de facto* or induced modifications. B. Production of peptides and nanoLC-HRMS quantification of the nonmodified peptides.

Figure 2. Influence of the number of isotopic distribution peaks used for area calculation on the quantification result. The $^{14}N/^{15}N$ ratios for each peptide were calculated using Skyline software and plotted as a function of *m/z*. A. Asp-N peptides from a 1:1 mixture of $^{14}N/^{15}N$ hPEBP1 protein solutions were analyzed in decaplicate by nanoUltraHPLC-UHR-QTOF MS. B. Theoretical relative isotopic abundances of the Asp-N peptides of hPEBP1. In blue: the first three peaks of the monoisotopic distribution are considered. In red: The whole monoisotopic distribution (isotope > 1 %) is considered.

Figure 3. Bound fraction pattern caused by a missed cleavage caused by the modification. This scheme is valid for both N- and C-endoproteinases. Protein and peptides are represented in color with associated abundances. The corresponding f_{bounds} are calculated based on changes in nonmodified peptides and graphed at the bottom.

Figure 4. Bound fraction pattern caused by a shifted cleavage. This scheme is valid for both N- and C-endoproteinases. Protein and peptides are represented in color with associated abundances. The corresponding f bounds are calculated using non modified peptides and graphed at the bottom. Two cleavable and modifiable sites separated by two residues are shown: site 1 and site 2. In the control, site 1 is completely cleaved while a missed cleavage of 70% is considered at site 2. In the experiment, 20% of 14N protein is modified at site 1, so that there is a MCm at site 1 and there is no more MC at site 2 in this modified population of protein. The di-peptide generated after cleavages at sites 1 and 2 is not represented in the bottom graphs because of its short length which impairs its detection by MS.

Figure 5. Measurement of TMSF for the hPEBP1-locostatin complex. hPEBP1 and locostatin were incubated for 5 h at 37°C and excess locostatin removed by micro gel filtration. The covalent complex was then analyzed by HRMS with online desalting.

Figure 6. Histogram representation of parallel sites quantification for the hPEBP1-locostatin complex. Error bars correspond to twice the standard deviation. The average negative f_{bound} is represented by a dotted bar