A new opening for the tricky untargeted investigation of natural and modified short peptides

Andrea Cerrato[1], Sara Elsa Aita[1], Anna Laura Capriotti[1*], Chiara Cavaliere[1], Carmela Maria Montone[1], Aldo Laganà,[1,2] Susy Piovesana[1]

1. Department of Chemistry, Università degli Studi di Roma La Sapienza, Piazzale Aldo Moro 5, 00185 Rome, Italy

2. CNR NANOTEC, Campus Ecotekne, University of Salento, Via Monteroni, 73100 Lecce, Italy

*__Corresponding author__:

Prof. Anna Laura Capriotti

Department of Chemistry, Sapienza Università di Roma, Piazzale Aldo Moro 5, 00185 Rome, Italy

E-mail: annalaura.capriotti@uniroma1.it

Telephone: (+39) 06 4991 3062

Fax (+39) 06 4906 31

Abstract

Short peptides are of extreme interest in clinical and food research fields, nevertheless they still represent a crucial analytical issue. The main aim of this paper was the development of an analytical platform for a considerable advancement in short peptides identification. For the first time, short sequences presenting both natural and post-translationally modified amino acids were comprehensively studied thanks to the generation of specific databases. Short peptide databases had a dual purpose. First, they were employed as inclusion lists for a suspect screening mass-spectrometric analysis, overcoming the limits of data dependent acquisition mode and allowing the fragmentation of such low-abundance substances. Moreover, the databases were implemented in Compound Discoverer 3.0, a software dedicated to the analysis of short molecules, for the creation of a data processing workflow specifically dedicated to short peptide tentative identification. For this purpose, a detailed study of short peptide fragmentation pathways was carried out for the first time. The proposed method was applied to the study of short peptide sequences in enriched urine samples and led to the tentative identification more than 200 short natural and modified short peptides, the highest number ever reported.

Keywords

Short peptides; modified peptides; post-translational modifications; Compound Discoverer; suspect screening analysis; high-resolution mass spectrometry

1. Introduction

Peptides can be classified in terms of their chain length as medium sized, namely those possessing five to twenty amino acids in their sequence, and small sized peptides, whose number of amino acids is lower than five[1]. Along with their molecular weight, peptides are also classified according to their nature, since they could be encrypted in proteins and therefore obtained by proteolysis with trypsin or other commercially available enzymes or they can be naturally produced by endogenous peptidases[2]. While medium size peptides, both tryptic and endogenous, have been extensively studied over the past two decades both in food and in clinical fields[2–6], the analysis of endogenous small peptides, which represent a crucial target in several fields, has long been an analytical challenge in terms of purification, separation and identification[2,7]. At present, in fact, a dedicated short-peptidomics approach is lacking, even though metabolomics studies have highlighted the importance of this class of compounds as possible biomarkers. In this context, for instance, metabolomics studies indicated a tetra-peptides and a tri-peptides as possible biomarkers in early bladder cancer urine samples[8] or some glycated peptides in dried blood spot from patients with cystic fibrosys[9], or some amino acid urinary markers potentially associated with autism spectrum disorder[10].

Recently, in our previous works the main attention was given on exploring several chromatographic technologies for dealing with the extremely uneven physico-chemical properties of this class of compounds by employing reversed phase (RP) C18 column, porous graphitic carbon (PGC) and a zwitterionic hydrophilic interaction (zic-HILIC) columns[11,12]. Subsequently, our attention was moved to sample preparation systems for their enrichment and purification considering the low endogenous abundance of these substances in biological samples. Therefore, an analytical platform based on graphitized carbon black (GCB) enrichment led to successfully identify more than 150 short peptides in urine samples[11]. Alongside, short peptides were also identified in plasma sample by employing a purification step based on Phree™ Phospholipid removal cartridge in combination with solid phase extraction (SPE) on a GCB sorbent[13]. Small endogenous peptides were also identified in milk sample by cotton-HILIC based purification[14]. Despite these analytical improvements, the challenge related to those species identification, in terms of both their MS/MS fragmentation study and the setup of a data processing workflow for routine application in clinical studies still remains. The main issue in the identification of short amino acidic sequence is due to the need for an extensive manual mass-spectrometric (MS) investigation. The most common software programs for proteomics and medium-sized peptidomics studies cannot, in fact, be used for the automatized identification of such short sequences[15].

Moreover, all these papers did not deal with the occurrence of post-translational modification (PTMs) on amino acid residues, deriving from the widespread amount of modification in protein in cells and tissues. PTMs play a key role in functional proteomics, since they regulate activity, localization and interaction with different molecules such as proteins, lipids, cofactors and nucleic acids[16,17]. Furthermore, PTMs containing peptides are also of great interest in nutraceutical field[18–20]. In nature, more than 200 different PTMs are known but only a small portion has been extensively investigated in peptidomic studies, leading to the tentative identification of modified medium-size peptides.[16] Since the study of amino acid sequences is a surrogate for proteins activity alteration and considering the biological significance as biomarkers of such class, a comprehensive study of those sequences, bearing in mind also the occurrence of the most common PTMs, is essential to understand how short peptides can be recognized, and how they crosstalk with one another to control fundamental biological processes.

Those technical bottlenecks were overcome by creating a novel customized workflow for small peptide analysis, which was implemented on Compound Discoverer 3.0, a software dedicated to the identification of small molecules based on HRMS data. This tool was set up with the purpose of leading to a comprehensive identification of all combinations of di-, tri and tetra peptides, deriving both from natural and modified amino acids, and it is the first of its kind, avoiding highly time-consuming data analysis of a large set of features. First, a database with all short peptide molecular formulas and masses deriving from the combination of the 20 natural amino acids and 14 residues presenting the most common PTMs was compiled by MatLab and employed both from data acquisition and processing on Compound Discoverer 3.0, which was implemented of a customized workflow specific for short amino acid sequences identification. The developed data processing workflow was applied to the first urine short peptide profiling, comprehensive of PTMs on amino acid residues. For this purpose, a detailed study of short peptide peculiar fragmentation pathways was achieved for the first time. In order to test the potentiality, advantages and benefits of our developed methodology, a comparison with traditional *de novo* sequencing approach was also carried out. Our study highlights the many benefits of a tool for rapid, facile and partially automatized data analysis for the tentative identification of short peptides reducing the number of false positives and enlarging the knowledge on such important class of biological compounds for further applications in several research fields.

2. Experimental Section

2.1. Chemicals and Materials

Optima LC–MS grade water, acetonitrile (ACN), and methanol (MeOH) were purchased from Thermo Fisher Scientific (Waltham, Massachusetts, USA). Trifluoroacetic acid (TFA) was supplied by Romil Ltd. (Cambridge). Formic acid and ammonium formate were purchased from Sigma-Aldrich (Germany). Dichloromethane (DCM) was provided by VWR International (Milan, Italy). Cartridges packed with 500 mg Carbograph 4 were supplied from Lara S.R.L (Lara S.r.l., Formello, RM, Italy).

2.2. Preparation of Urine Samples and Short Peptide Enrichment

The first urine of the day was collected from 10 healthy volunteers. The collected samples were pooled, centrifuged at 1000 × g, acidified with HCl to pH 2, aliquoted, and stored at −20 °C until further processing. Urine aliquots were thawed at room temperature and centrifuged at 8000 × g to remove any insoluble debris. SPE and cleanup of short peptides was carried out by cartridges packed with 500 mg Carbograph 4. The applied procedure was optimized in our previous work. [11] Briefly, after the manually packing of GCB stationary phase into 6 mL polypropylene tubes (Sigma-Aldrich) with 500 mg of Carbograph 4 bulk material (130 m^2/g surface area, 20/400–120/200 mesh size), the cartridge was washed with 5 mL of DCM/MeOH, 80:20 (*v/v*) with 20 mmol L^{-1} TFA and 5 mL of MeOH with 20 mmol L^{-1} TFA. The activation was carried out by flushing with 10 mL of 0.1 mol L^{-1} HCl and finally conditioned with 10 mL of 20 mmol L^{-1} TFA. Then, 2 mL urine was diluted in 8 mL H_2O with 20 mmol L^{-1} TFA and loaded onto the cartridge, which was sequentially washed with 2 mL of 20 mmol L^{-1} TFA and 0.5 mL MeOH. Finally, analytes were eluted by back flushing elution with 10 mL of DCM/MeOH and 80:20 (v/v) with 20 mmol L^{-1} TFA. The eluate was evaporated at room temperature in a Speed-Vac SC250 Express (Thermo Savant, Holbrook, NY, USA) and the residue reconstituted in 200 μL water and 200 μL of ACN/H_2O, 75:25 (v/v) for RP and HILIC separation, respectively.

2.3. Ultra-High Performance Liquid Chromatography-HRMS Analysis

A Vanquish binary pump H (Thermo Fisher Scientific, Bremen, Germany), equipped with thermostated autosampler and column compartment, was used for short peptides chromatographic separation on two orthogonal columns: a Kinetex XB-C18 (100 × 2.1 mm, 2.6 μm particle size, Phenomenex, Torrance, USA) for RP separation and a iHILIC-Fusion UHPLC Column, SS (100 ×

2.1 mm, 1.8 μm particle size, Hilicon, Umea, Sweden) for HILIC separation. Chosen flow, column temperature and gradient parameters are reported in our previous work without any modification [11]. The chromatographic system was coupled to a hybrid quadrupole-Orbitrap mass spectrometer Q Exactive (Thermo Fisher Scientific) using a heated ESI source. The ESI source was operated in positive mode and set up as previously reported [11]. HRMS top 5 data dependent acquisition (DDA) mode was performed in the range *m*/*z* 150-750 with a resolution (full width at half maximum, FWHM, *m*/*z* 200) of 70,000. Higher-energy collisional dissociation (HCD) fragmentation was performed at 40% normalized collision energy at resolution of 35,000 (FWHM @*m*/*z* 200). Two different inclusion lists, obtained by filtering the databases described in section “Short Peptide Databases Compilation”, presenting unique masses of precursor ions for natural or post-translational modified short peptides were implemented on the MS method for performing a suspect screening analysis. Raw data files were acquired by Xcalibur software (version 3.1, Thermo Fisher Scientific).

2.4. Short Peptide Database Compilation

Short peptide database was generated using MatLab R2018a by combining masses and molecular formulas of natural amino acids and several residues with modified side chains, in order to take into consideration the most common PTMs occurring on proteins. The combination of the 20 natural amino acids in di-, tri- and tetrapeptides resulted in 168,400 different sequences, which were filtered to remove duplicate masses. The filtered mass list (4980 unique masses) was intended to be employed as inclusion list, while the complete list was implemented into Compound Discoverer (v. 3.0, Thermo, Waltham, USA). Thirteen amino acids presenting modified side chains (citrulline, hydroxylysine, hydroxyproline, methionine sulfoxide, pyroglutamic acid, methylarginine, acetyllysine, methyllysine, dimethyllysine, trimethyllysine, succinyllysine, phosphoserine and sulfotyrosine) and lactic acid were chosen based on data in the literature for investigating PTMs on short peptide sequences[21–30]. However, the number of unique masses obtained by combining the aforementioned fourteen residues to the twenty natural amino acids was considerably higher than 5,000, which is the upper limit for inclusion lists on Q Exactive instruments. Since the tetra-peptides represented both the largest number in the list and the least abundant sequences that were identified in urine in our previous study[11], tetrapeptides presenting modified side chains were therefore not included in the definitive list of 3179 unique masses.

2.5. Data Analysis and Short Peptide Validation

For each sample, raw data files from three experimental replicates and a blank sample were processed by Compound Discoverer using a workflow designed as follows (Figure S1). For short amino acid sequence raw data processing, the customized databases generated in section "Short peptide database compilation", complete of IDs, masses and molecular formulas, were implemented in the *mass lists* feature for the automatic matching of extracted *m/z* ratios. Moreover, parameters for the *predict composition* tool were adapted to short peptides analysis. *Compound class scoring* tool was implemented with a large set of fragments deriving from amino acids at N-terminus, C-terminus and in the middle of the sequence. Thus, experimental MS/MS spectra were automatically matched to the 34 compound classes, one for each natural or modified residue. The complete set of fragments composing the compound classes and Compound Discoverer parameters are reported in Tables S1, S2 and S3. Extracted masses from the chromatograms were aligned and filtered to remove background compounds present in the blank sample, features whose masses were not present in the databases and those which were not fragmented. Filtered features were manually validated matching experimental spectra to those generated in silico by mMass 5.5[31]. For residues carrying PTMs, peptides were tentatively identified according to the characteristic fragmentation spectra. Raw files were also processed by pNovo 3.1.3[32].

3. Results and discussion

3.1. Customized Workflow on Compound Discoverer

Untargeted analysis based on high-resolution mass spectrometry (HRMS) gives the opportunity of tentatively identifying unknown or unexpected compounds. However, extremely difficult and time-consuming manual validation of the features is needed for the correct identification of uncharted compounds. Hence, when untargeted analyses are performed, the use of MS-based databases is almost essential for effectively associating retention times, mass to charge ratios and fragmentation spectra to known or unknown molecular structures. Nevertheless, even the most complete available databases do not possess exhaustive data for structure-related classes of compounds, resulting in often incomplete and fragmentary identifications. With the purpose of comprehensively identifying short amino acid sequences while assuring a great abridging of the manual validation, a different approach was chosen. As the number of combinations of amino acids in short peptide sequences is undoubtedly elevated but still finite, a customized database can be created combining the masses and formulas of natural and modified amino acids. The use of a customized database has a dual purpose. Firstly, it can be used as inclusion list for the mass-spectrometric method, allowing the selective MS/MS fragmentation of masses present in the database, which is particularly useful for the analysis of low-abundance compounds in complex matrices. Secondly, databases can be implemented in data analysis workflows for matching experimental features to listed compounds of a specific class. This approach was previously employed with the same logic for the comprehensive identification of phenolic compound conjugates[33].

When employing Thermo Q Exactive instrumentations, untargeted approaches are commonly performed with DDA methods, since more valuable data independent acquisition (DIA) approaches are highly time-consuming and show weak performances in slow orbitrap-based instruments [34,35]. However, DIA approaches, like all ion fragmentation (AIF), would grant the MS/MS fragmentation of all eluting precursor ions in a predefined isolation window, including low-abundance species like short peptides. On the other hand, DDA methods, in which top n ranked precursor ions are sequentially isolated and fragmented, would repeatedly cause high-abundance species to suppress less concentrated coeluting compounds, which would not be fragmented and, eventually, identified. When precursor ion databases are available, suspect screening MS approaches constitute a valuable alternative to DIA, granting the selective fragmentation of precursor ions present in the inclusion list and overcoming the limits of DDA mode. Thus, many low-abundance species that would have normally been neglected, were fragmented and validated. Raw data files were processed by Compound Discoverer 3.0 with a peculiar workflow which was specifically projected for short peptide identification. Even though manual validation is essential for appropriately assigning the correct order of the residues in the sequence, the customized workflow assured not merely a comprehensive identification of short peptide sequences, but also a decisive streamlining of the validation step. As showed in Figure 1, the sequentially applied filters (*Mass Lists* filter, MS2 filter and *Compound Class Scoring* filter) allowed a critical decrease of the original features, with the result of much fewer compounds to be manually validated. Furthermore, with the automatic matching of features to hypothetical sequences, the MS investigation is greatly simplified. Thanks to *Compound Class Scoring* tool, which matches experimental fragments to those of implemented compound classes and assigns a percentage score, MS/MS spectra study was even more streamlined.

3.2. Short Peptide Identification

Fragmentation patterns of peptide sequences with CID or HCD techniques are very well-known, studied and predictable [36,37], allowing the automation of their identification thanks to the use of spectral libraries or *de novo* approaches. Even though amino acids present considerably uneven physico-chemical properties, such as molecular weight, polarity or acid-base properties, medium to long-sized peptides generally present noticeably less variable characteristics, since single amino acid peculiarities are mutually mitigated. As a result, the common medium-sized peptides generated by in vitro digestion with enzymes are usually identified by *b* and *y* product ions, due to on-chain fragmentation in correspondence of amide bonds. When peptide sequences are very short, however, the peculiarities of the specific amino acids are crucial, and general rules cannot be applied to all peptides. Fragments deriving from amino acids with basic side chains, such as histidine, lysine and arginine, are typically the most abundant, regardless of the position on the sequence. Moreover, the number of *b* and *y* ions decreases significantly with the shortening of the sequences, thus not allowing a full attribution of the product ions in the spectra. Iminium ions, which correspond to *a* fragments for N-terminal amino acids, assume great importance, since they represent a very stable charged form due to the absence of acidic groups. Iminium ions are usually more abundant than the corresponding *b* fragments, except for glycine and alanine, whose small side chain size avoids their detection, and for lysine, whose $b\text{-}NH_3$ ion at *m/z* 129.1022 has generally high abundance (e.g. *Asp-Glu*, Figure S2). Asparagine, glutamine, lysine, arginine and their modified derivatives usually produce iminium ions with losses of ammonia (e.g. *m/z* 84.0808 rather than *m/z* 101.1073 for lysine), while aspartic and glutamic acid generate iminium ions with losses of water (e.g. *m/z* 84.0444 rather than *m/z* 102.0550 for glutamic acid). Even though the relative abundance of the iminium ions could indicate which amino acid is the N-terminus, some iminium ions (e.g. those deriving from histidine, proline, phenylalanine, tyrosine and tryptophan) are so much stable that are often base peaks. Luckily, for assigning the correct order of the residues, *y* ions are always present in the spectra, even when alkaline amino acids are present at N-terminus. As for iminium ions, also some *y* ions undergo losses of ammonia or water, which must be taken into consideration when attributing the signals. In particular, *y* ions undergoing losses of ammonia (which also correspond to *z* ions) are common for amino acids possessing amine or amide groups on their side chains as well as tyrosine (*m/z* 165.0546) and tryptophan (*m/z* 188.0706). Those neutral losses cause some tricky attributions when it comes to distinguishing asparagine from aspartic acid and glutamine from glutamic acid, as they produce iminium and *y* ions with the exact same *m/z* (e.g. 130.0499 corresponds both to *y*-NH_3 glutamine ion and *y*-H_2O glutamic acid ion). In most cases, however, *y* ions prior to neutral losses are present in MS/MS spectra even in low abundance. Even though multistage MS analysis has been proven to correctly discriminate between leucine and isoleucine by the relative abundance of iminium-NH_3 ion at *m/z* 69.0699 [38], the Q Exactive instrumentation does not allow performing MS^3 experiments. For this reason, when leucine or isoleucine are present in the amino acid sequence, they have been listed as *Xle*, a common abbreviation for indicating one of the two isomers. As regards amino acids with modified side chains, they usually behave as regular amino acid in terms of fragmentation pathways. Hydroxyproline, for example, produces high-abundance iminium and *y* ion at *m/z* 86.0600 and *m/z* 132.0655, respectively (e.g. *Hyp-Glu-Gly*, Figure S2). Discriminating between citrulline and arginine is sometimes highly insidious, since fragments produced by citrulline, which undergo both water and ammonia losses, assume the exact same masses as those produced by arginine (e.g. *m/z* 140.0818 corresponds both to *b*-NH_3 arginine ion and *b*-H_2O citrulline ion). The two residues have been discriminated based on the presence (or absence) of low-abundance *m/z* 113.0709, which corresponds to iminium-NH_3 ion of citrulline and cannot derive from arginine. Fragments containing methionine sulfoxide rapidly undergo losses of CH_4SO moieties, while tyrosine *O*-sulfate is widely subject to SO_3 losses. Those peculiar fragmentation patterns have been carefully studied to correctly attribute product ions (e.g. *Pro-sTyr*, Figure S2). Finally, lactoyl-bound sequences present very intense M-HCOOH losses, while *m/z* 73.0824, its *pseudo-b* ion, has very low abundance as a result of acidic compound scarce ionization efficiency.

3.3. Short Peptidomic Analysis of Urine Samples

Natural and modified short peptides have not been extensively studied in peptidomics works. For this reason, in this paper an analytical workflow able to characterize the entire short peptidomic profile on urine sample was developed. An overview of the tested variables in the workflow is depicted in Figure 2.

Two orthogonal analytical columns (RP and HILIC) were used for urine sample separation. In our previous work on short endogenous peptides in urine samples[11], in fact, only 39% of the identified compounds was common to the two chromatographic columns. Moreover, the median Grand average of hydropathicity (GRAVY) values for C18 unique peptides was close to 0 value, while it was -1.78 for HILIC unique peptides, in agreement with the wide range of physico-chemical properties of short peptides. Thus, employing two orthogonal system could lead to broaden the hydropathicity range of short analyzed peptides and therefore enhancing the number of tentative identifications. We next choose to evaluate two mass spectrometric strategies: an untargeted approach and a suspect screening analysis based on inclusion list (Figure 2). Raw data were processed by Compound Discoverer software leading to the tentative identification of 216 and 42 short peptides following suspect and untargeted approach, respectively. Those results demonstrated that the suspect approach is essential for the comprehensive identification of low-abundance short peptides. Among the 216 tentative identifications, 154 sequences presented only natural amino acids, while 62 possessed modified residues (Table S4 and S5, respectively). The high percentage (roughly 30%) of modified sequences amongst the identified short peptides represent a pivotal outcome, considering how neglected those compounds usually are. However, it is not sure whether tentatively identified peptides containing modified amino acids were real PTMs or are artifacts generated after protein cleavage or during sample pre-treatment. Eight modifications among the 14 modifications inserted in the database were found; in particular, 17 pyroglutamic acid, 16 hydroxyproline, 11 lactic acid, 9 citrulline, 4 succinillysine, 3 sulfotyrosine, one methionine sulfoxide and one metylarginine containing peptides were identified. Regarding the four PTMs not found, factors including degradation effects, unfavorable ionization properties of non-tryptic peptides with PTMs, and the low *in vivo* abundance, can hinder their identification in a complex biological sample. In fact, despite the great potential of hyphenated high-resolution techniques, direct analysis is generally not possible without prior specific enrichment for some of these PTMs. Serine phosphorylations, for instance, are present in biological samples in sub-stoichiometric concentrations and possess poor ionization efficiency, resulting in the suppression of their signal. As a result, highly selective enrichment strategies, such as immobilized metal or metal oxide affinity chromatographies, for phospo-based PTMs prior to HPLC-MS/MS analysis have become mandatory for efficient detection. Specific enrichment analytical methodology have to be optimize and applied[39–41].

In order to increase the coverage and the confidence in peptide identification, a combination of two columns with a distinct selectivity mechanism was carried out (Figure 2). As shown in Table S4, among the total number of natural short identified peptides, 56 were exclusively identified by the C18 column, 45 unique peptides were identified by the zic-HILIC column, while 53 were in common between the two data sets. It was also demonstrated that the use of the two columns was the best choice for enlarging the number of identifications also for modified peptides, considering that 25 peptides were in common between the two separation strategies while 20 and 17 were univocally identified in RP and HILIC separation, respectively. Whereas HILIC separation avoids elution of peptides at the dead volume, RP chromatography possesses ability to better separate isomers, which have been found to co-elute instead with HILIC due to generally larger peak shapes. As shown in Table S4, for instance, peptides *Glu-Val* and *Asp-Xle* were successfully separated by C18 column (r_t 4.25 and 5.88, respectively), while co-eluted at r_t 6.50 using the zic-HILIC column. This peculiarity could be important in targeted analysis of specific short sequences with a relevant clinical value, especially because the fragmentation patterns generated by collisional dissociation do not include product ions which can discriminate between leucine and isoleucine, for instance.

The importance of the enrichment step was also evaluated. The direct analysis of urine 1:4 diluted in mobile phase and analyzed by RP-UHPLC allowed identifying only 55 short peptides sequences. As a result, the enrichment step and the suspect screening method for MS analysis are both equally essential for reducing suppression due to other high-abundance metabolites. In particular, the MS method is even more crucial than a 20-fold enrichment and purification step as regards the number of identified peptides (42 sequences for the enriched sample analyzed in untargeted fashion *vs* 55 for the dilute and shoot sample analyzed with suspect screening method).

3.4. *De novo* Sequencing of Short Peptides

The common database-based proteomics software programs cannot be employed for short peptide analysis, since such short sequence identification would result in low level of confidence in associating the sequences to single proteins. Therefore, *de novo* sequencing programs represent the only viable option for comparing our developed methodology to already existing techniques[42,43]. *De novo* sequencing programs automatically predict amino acid sequences based on MS and MS/MS spectra and are therefore not depending on databases of known protein sequences. To date *de novo* sequencing was considered very promising and the only method for determining proteins from organisms with unknown genomes or for identifying blind PTMs[32,44]. pNovo is a freeware for *de novo* sequencing which has already shown to guarantee excellent results in terms of medium-sized peptide coverage[32]. However, as shown in Tables S4 and S5, it only allowed the tentative identification of 59 and 28 natural and modified short peptides, which is roughly 40% of the sequences identified by Compound Discoverer. The unsatisfactory results were largely due to the lower cut off at *m/z* 300, which excludes most dipeptides. Nonetheless, although MS/MS spectra were automatically associated to amino acid sequences, thousands of tentative identifications were listed, causing long manual validation for filtering the effective correct features. Moreover, some of the identified sequences were misinterpreted, since the program does not take peak abundances into account, and just matches experimental to in *silico* product ions. Considering the high cut-off and the several misinterpretations of short sequences, the employed de novo software is probably oriented to longer peptide identification. In a typical *de novo* peptide identification, there is no peak extraction and chromatogram building prior to MS/MS spectral interpretation and peptides are seldom manually checked. However, while standard tryptic digest analysis of medium-sized peptides rarely leads to false positive due to the protein rich nature of tryptic digests and the multicharged ion filter used in the mass spectrometric methods, which excludes most other compounds, short endogenous peptides are minor components of other singly charged metabolites rich extracts. Therefore, the issue of peak extraction and association to MS/MS spectra followed by careful manual validation is highly recommended. It must be noted, however, that the independence of pNovo from databases allowed the tentative identification of eight modified tetrapeptides, a class of short peptides that had to be left out as previously discussed (Table S5). In the end, as well as guaranteeing more thorough results, our developed approach is feasible for metabolomics applications both in clinical and agri-food fields for the evaluation of potential biomarker or bioactive sequences. Compound Discoverer, in fact, permits alignment, normalization and differential analysis for rigorous quality control-based studies, leading to still unexplored directions in short peptidomics studies.

4. Conclusions

Both in clinical and in food applications, the identification and characterization of endogenous short peptides is of great significance. However, a variety of issues makes short peptidomics less straightforward compared to medium-size peptide analysis. The paper describes the development of a data processing workflow by means of Compound Discoverer software for the identification of endogenous short and modified peptides, whose peculiar fragmentation pathways were for the first time discussed. The developed analytical workflow demonstrated the potentiality to obtain a larger number of short peptide sequence identifications, and above all it allowed the tentative identification of PTM-containing sequences, which is a hot topic in the fields of clinical and nutraceutical endogenous peptidomics. The obtained results revealed that short modified peptides, which are often neglected, constitute a large portion of this class of endogenous compounds. Moreover, the presented analytical workflow increases the computational speed while reducing the manual work of the analysis when compared to other current methods well described recently[2], which do not lead to the identification of short peptide sequences. The suspect screening method was considerably the best choice to overcome the issues of DDA analysis for those low-abundance compounds. Last but not least, this novel approach could result in the discovery of non-invasive biomarkers for diagnosing patients with different diseases, with the aim to ultimately improve clinical outcomes.

5. Acknowledgments

The work was supported by the PRIN project Prot. 2017Y2PAB8, entitled “Cutting Edge Analytical Chemistry Methodologies and Bio-Tools to Boost Precision Medicine in Hormone-Related Diseases”, provided by the Italian Ministry of Education, Universities and Research.

6. References

[1] A.C.-L. Lee, J.L. Harris, K.K. Khanna, J.-H. Hong, A comprehensive review on current advances in peptide drug development and design, Int. J. Mol. Sci. 20 (2019) 2383. https://doi.org/10.3390/ijms20102383.

[2] E. Maes, E. Oeyen, K. Boonen, K. Schildermans, I. Mertens, P. Pauwels, D. Valkenborg, G. Baggerman, The challenges of peptidomics in complementing proteomics in a clinical context, Mass Spectrom. Rev. 38 (2019) 253–264. https://doi.org/10.1002/mas.21581.

[3] R. Zenezini Chiozzi, A.L. Capriotti, C. Cavaliere, G. La Barbera, S. Piovesana, R. Samperi,

A. Laganà, Purification and identification of endogenous antioxidant and ACE-inhibitory peptides from donkey milk by multidimensional liquid chromatography and nanoHPLC-high resolution mass spectrometry, Anal. Bioanal. Chem. 408 (2016) 5657–5666. https://doi.org/10.1007/s00216-016-9672-z.

[4] S. Piovesana, A.L. Capriotti, C. Cavaliere, G. La Barbera, R. Samperi, R. Zenezini Chiozzi, A. Laganà, Peptidome characterization and bioactivity analysis of donkey milk, J. Proteomics. 119 (2015) 21–29. https://doi.org/10.1016/j.jprot.2015.01.020.

[5] R. Zenezini Chiozzi, A.L. Capriotti, C. Cavaliere, G. La Barbera, S. Piovesana, A. Laganà, Identification of three novel angiotensin-converting enzyme inhibitory peptides derived from cauliflower by-products by multidimensional liquid chromatography and bioinformatics, J. Funct. Foods. 27 (2016) 262–273. https://doi.org/10.1016/j.jff.2016.09.010.

[6] G. Arapidi, M. Osetrova, O. Ivanova, I. Butenko, T. Saveleva, P. Pavlovich, N. Anikanov, V. Ivanov, V. Govorun, Peptidomics dataset: blood plasma and serum samples of healthy donors fractionated on a set of chromatography sorbents, Data Br. 18 (2018) 1204–1211. https://doi.org/10.1016/j.dib.2018.04.018.

[7] S. Piovesana, A.L. Capriotti, C. Cavaliere, G. La Barbera, C.M. Montone, R. Zenezini Chiozzi, A. Laganà, Recent trends and analytical challenges in plant bioactive peptide separation, identification and validation, Anal. Bioanal. Chem. 410 (2018) 3425–3444. https://doi.org/10.1007/s00216-018-0852-x.

[8] C. Shen, Z. Sun, D. Chen, X. Su, J. Jiang, G. Li, B. Lin, J. Yan, Developing urinary metabolomic signatures as early bladder cancer diagnostic markers, Omi. A J. Integr. Biol. 19 (2015) 1–11. https://doi.org/10.1089/omi.2014.0116.

[9] A. DiBattista, N. McIntosh, M. Lamoureux, O.Y. Al-Dirbashi, P. Chakraborty, P. Britz-McKibbin, Metabolic signatures of cystic fibrosis identified in dried blood spots for newborn screening without carrier identification, J. Proteome Res. (2018) acs.jproteome.8b00351. https://doi.org/10.1021/acs.jproteome.8b00351.

[10] A. Liu, W. Zhou, L. Qu, F. He, H. Wang, Y. Wang, C. Cai, X. Li, W. Zhou, M. Wang, Altered urinary amino acids in children with autism spectrum disorders, Front. Cell. Neurosci. 13 (2019). https://doi.org/10.3389/fncel.2019.00007.

[11] S. Piovesana, A.L. Capriotti, A. Cerrato, C. Crescenzi, G. La Barbera, A. Laganà, C.M. Montone, C. Cavaliere, Graphitized carbon black enrichment and uhplc-ms/ms allow to meet the challenge of small chain peptidomics in urine, Anal. Chem. (2019). https://doi.org/10.1021/acs.analchem.9b03034.

[12] S. Piovesana, C.M. Montone, C. Cavaliere, C. Crescenzi, G. La Barbera, A. Laganà, A.L. Capriotti, Sensitive untargeted identification of short hydrophilic peptides by high performance liquid chromatography on porous graphitic carbon coupled to high resolution mass spectrometry, J. Chromatogr. A. 1590 (2019) 73–79. https://doi.org/10.1016/j.chroma.2018.12.066.

[13] S. Piovesana, A. Cerrato, M. Antonelli, B. Benedetti, A.L. Capriotti, C. Cavaliere, C.M. Montone, A. Laganà, A clean-up strategy for identification of circulating endogenous short peptides in human plasma by zwitterionic hydrophilic liquid chromatography and untargeted peptidomics identification, J. Chromatogr. A. (2019) 460699. https://doi.org/10.1016/j.chroma.2019.460699.

[14] C.M. Montone, A.L. Capriotti, A. Cerrato, M. Antonelli, G. La Barbera, S. Piovesana, A. Laganà, C. Cavaliere, Identification of bioactive short peptides in cow milk by high-performance liquid chromatography on C18 and porous graphitic carbon coupled to high-resolution mass spectrometry, Anal. Bioanal. Chem. 411 (2019) 3395–3404. https://doi.org/10.1007/s00216-019-01815-0.

[15] T. Välikangas, T. Suomi, L.L. Elo, A comprehensive evaluation of popular proteomics software workflows for label-free proteome quantification and imputation, Brief. Bioinform. (2017). https://doi.org/10.1093/bib/bbx054.

[16] G. Duan, D. Walther, The roles of post-translational modifications in the context of protein

interaction networks, PLOS Comput. Biol. 11 (2015) e1004049. https://doi.org/10.1371/journal.pcbi.1004049.

[17] Y.-C. Wang, S.E. Peterson, J.F. Loring, Protein post-translational modifications and regulation of pluripotency in human stem cells, Cell Res. 24 (2014) 143–160. https://doi.org/10.1038/cr.2013.151.

[18] L. Xing, T. Ding, C. Huang, Y. Xi, Q. Wu, W. Qian, F. Wan, B. Zhang, A comprehensive atlas of lysine acetylome in onion thrips (Thrips tabaci Lind.) revealed by proteomics analysis, J. Proteomics. 207 (2019) 103465. https://doi.org/10.1016/j.jprot.2019.103465.

[19] R.S. Jansen, R. Addie, R. Merkx, A. Fish, S. Mahakena, O.B. Bleijerveld, M. Altelaar, L. IJlst, R.J. Wanders, P. Borst, K. van de Wetering, N -lactoyl-amino acids are ubiquitous metabolites that originate from CNDP2-mediated reverse proteolysis of lactate and amino acids, Proc. Natl. Acad. Sci. 112 (2015) 6601–6606. https://doi.org/10.1073/pnas.1424638112.

[20] S. Yao, C.C. Udenigwe, Peptidomics of potato protein hydrolysates: implications of post-translational modifications in food peptide structure and behaviour, R. Soc. Open Sci. 5 (2018) 172425. https://doi.org/10.1098/rsos.172425.

[21] H. Yamada, T. Ozawa, H. Kishi, S. Okada, Y. Nakashima, A. Muraguchi, Y. Yoshikai, Cutting edge: b cells expressing cyclic citrullinated peptide–specific antigen receptor are tolerized in normal conditions, J. Immunol. 201 (2018) 3492–3496. https://doi.org/10.4049/jimmunol.1800826.

[22] M. Yamamoto, M.I. Pinto-Sanchez, P. Bercik, P. Britz-McKibbin, Metabolomics reveals elevated urinary excretion of collagen degradation and epithelial cell turnover products in irritable bowel syndrome patients, Metabolomics. 15 (2019) 82. https://doi.org/10.1007/s11306-019-1543-0.

[23] M. Huang, H. Zhao, S. Gao, Y. Liu, Y. Liu, T. Zhang, X. Cai, Z. Li, L. Li, Y. Li, C. Yu, Identification of coronary heart disease biomarkers with different severities of coronary

stenosis in human urine using non-targeted metabolomics based on UPLC-Q-TOF/MS, Clin. Chim. Acta. 497 (2019) 95–103. https://doi.org/10.1016/j.cca.2019.07.017.

[24] J. Moskovitz, Detection and localization of methionine sulfoxide residues of specific proteins in brain tissue, Protein Pept. Lett. 21 (2013) 52–55. https://doi.org/10.2174/09298665113209990068.

[25] K. Pagano, D. Galante, C. D’Arrigo, A. Corsaro, M. Nizzari, T. Florio, H. Molinari, S. Tomaselli, L. Ragona, Effects of prion protein on aβ42 and pyroglutamate-modified aβpε3-42 oligomerization and toxicity, Mol. Neurobiol. 56 (2019) 1957–1971. https://doi.org/10.1007/s12035-018-1202-x.

[26] J. Murn, Y. Shi, The winding path of protein methylation research: milestones and new frontiers, Nat. Rev. Mol. Cell Biol. 18 (2017) 517–527. https://doi.org/10.1038/nrm.2017.35.

[27] S. Paolella, B. Prandi, C. Falavigna, S. Buhler, A. Dossena, S. Sforza, G. Galaverna, Occurrence of non-proteolytic amino acyl derivatives in dry-cured ham, Food Res. Int. 114 (2018) 38–46. https://doi.org/10.1016/j.foodres.2018.07.057.

[28] D. He, Q. Wang, M. Li, R.N. Damaris, X. Yi, Z. Cheng, P. Yang, Global proteome analyses of lysine acetylation and succinylation reveal the widespread involvement of both modification in metabolism in the embryo of germinating rice seed, J. Proteome Res. 15 (2016) 879–890. https://doi.org/10.1021/acs.jproteome.5b00805.

[29] W. Li, R.A. Boykins, P.S. Backlund, G. Wang, H.-C. Chen, Identification of phosphoserine and phosphothreonine as cysteic acid and β-methylcysteic acid residues in peptides by tandem mass spectrometric sequencing, Anal. Chem. 74 (2002) 5701–5710. https://doi.org/10.1021/ac020259v.

[30] D. Asakawa, H. Takahashi, S. Sekiya, S. Iwamoto, K. Tanaka, Sequencing of sulfopeptides using negative-ion tandem mass spectrometry with hydrogen attachment/abstraction dissociation, Anal. Chem. 91 (2019) 10549–10556. https://doi.org/10.1021/acs.analchem.9b01568.

[31] M. Strohalm, D. Kavan, P. Novák, M. Volný, V. Havlíček, mMass 3: a cross-platform software environment for precise analysis of mass spectrometric data, Anal. Chem. 82 (2010) 4648–4651. https://doi.org/10.1021/ac100818g.

[32] H. Yang, H. Chi, W.-F. Zeng, W.-J. Zhou, S.-M. He, pNovo 3: precise de novo peptide sequencing using a learning-to-rank framework, Bioinformatics. 35 (2019) i183–i190. https://doi.org/10.1093/bioinformatics/btz366.

[33] A. Cerrato, G. Cannazza, A.L. Capriotti, C. Citti, G. La Barbera, A. Laganà, C.M. Montone, S. Piovesana, C. Cavaliere, A new software-assisted analytical workflow based on high-resolution mass spectrometry for the systematic study of phenolic compounds in complex matrices, Talanta. 209 (2020) 120573. https://doi.org/10.1016/j.talanta.2019.120573.

[34] J.F. Xiao, B. Zhou, H.W. Ressom, Metabolite identification and quantitation in LC-MS/MS-based metabolomics, TrAC Trends Anal. Chem. 32 (2012) 1–14. https://doi.org/10.1016/j.trac.2011.08.009.

[35] R. Wang, Y. Yin, Z.-J. Zhu, Advancing untargeted metabolomics using data-independent acquisition mass spectrometry technology, Anal. Bioanal. Chem. 411 (2019) 4349–4357. https://doi.org/10.1007/s00216-019-01709-1.

[36] S. Li, R.J. Arnold, H. Tang, P. Radivojac, On the accuracy and limits of peptide fragmentation spectrum prediction, Anal. Chem. 83 (2011) 790–796. https://doi.org/10.1021/ac102272r.

[37] M.S. Montaudo, Mass spectra of copolymers, Mass Spectrom. Rev. 21 (2002) 108–144. https://doi.org/10.1002/mas.10021.

[38] Y. Xiao, M.M. Vecchi, D. Wen, Distinguishing between leucine and isoleucine by integrated lc–ms analysis using an orbitrap fusion mass spectrometer, Anal. Chem. 88 (2016) 10757–10766. https://doi.org/10.1021/acs.analchem.6b03409.

[39] G. La Barbera, A.L. Capriotti, C. Cavaliere, F. Ferraris, M. Laus, S. Piovesana, K. Sparnacci, A. Laganà, Development of an enrichment method for endogenous phosphopeptide

characterization in human serum, Anal. Bioanal. Chem. 410 (2018) 1177–1185. https://doi.org/10.1007/s00216-017-0822-8.

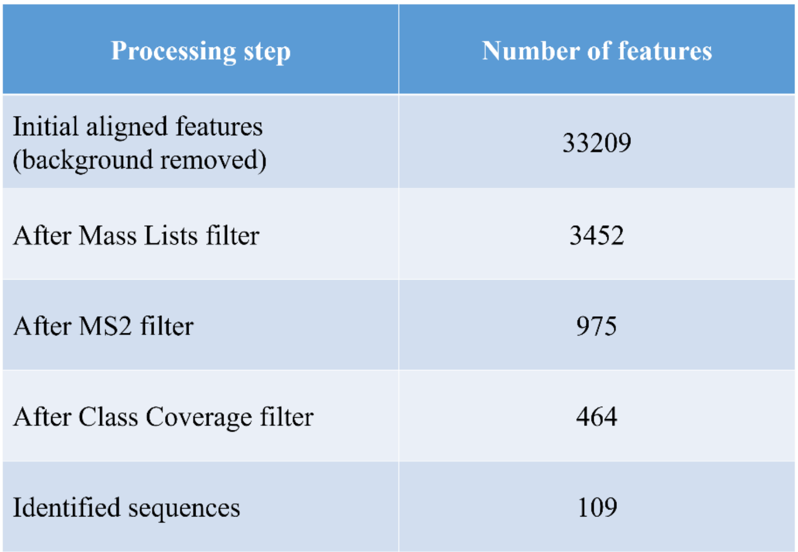
[40] G. La Barbera, A.L. Capriotti, C. Cavaliere, F. Ferraris, C.M. Montone, S. Piovesana, R. Zenezini Chiozzi, A. Laganà, Saliva as a source of new phosphopeptide biomarkers: Development of a comprehensive analytical method based on shotgun peptidomics, Talanta. 183 (2018) 245–249. https://doi.org/10.1016/j.talanta.2018.02.085.

[41] S. Piovesana, A.L. Capriotti, C. Cavaliere, F. Ferraris, D. Iglesias, S. Marchesan, A. Laganà, New magnetic graphitized carbon black TiO 2 composite for phosphopeptide selective enrichment in shotgun phosphoproteomics, Anal. Chem. 88 (2016) 12043–12050. https://doi.org/10.1021/acs.analchem.6b02345.

[42] M. Cermeño, J. Stack, P.R. Tobin, M.B. O'Keeffe, P.A. Harnedy, D.B. Stengel, R.J. FitzGerald, Peptide identification from a Porphyra dioica protein hydrolysate with antioxidant, angiotensin converting enzyme and dipeptidyl peptidase IV inhibitory activities, Food Funct. 10 (2019) 3421–3429. https://doi.org/10.1039/C9FO00680J.

[43] A.B. Nongonierma, S. Paolella, P. Mudgil, S. Maqsood, R.J. FitzGerald, Identification of novel dipeptidyl peptidase IV (DPP-IV) inhibitory peptides in camel milk protein hydrolysates, Food Chem. 244 (2018) 340–348. https://doi.org/10.1016/j.foodchem.2017.10.033.

[44] C. Hughes, B. Ma, G.A. Lajoie, De novo sequencing methods in proteomics, in: 2010: pp. 105–121. https://doi.org/10.1007/978-1-60761-444-9_8.



Processing step	Number of features
Initial aligned features (background removed)	33209
After Mass Lists filter	3452
After MS2 filter	975
After Class Coverage filter	464
Identified sequences	109

Figure 1. An exemplary application of the *Mass Lists*, MS2 and *Class Coverage* filter to natural short peptide analysis raw files obtained by RP separation and suspect screening MS method.

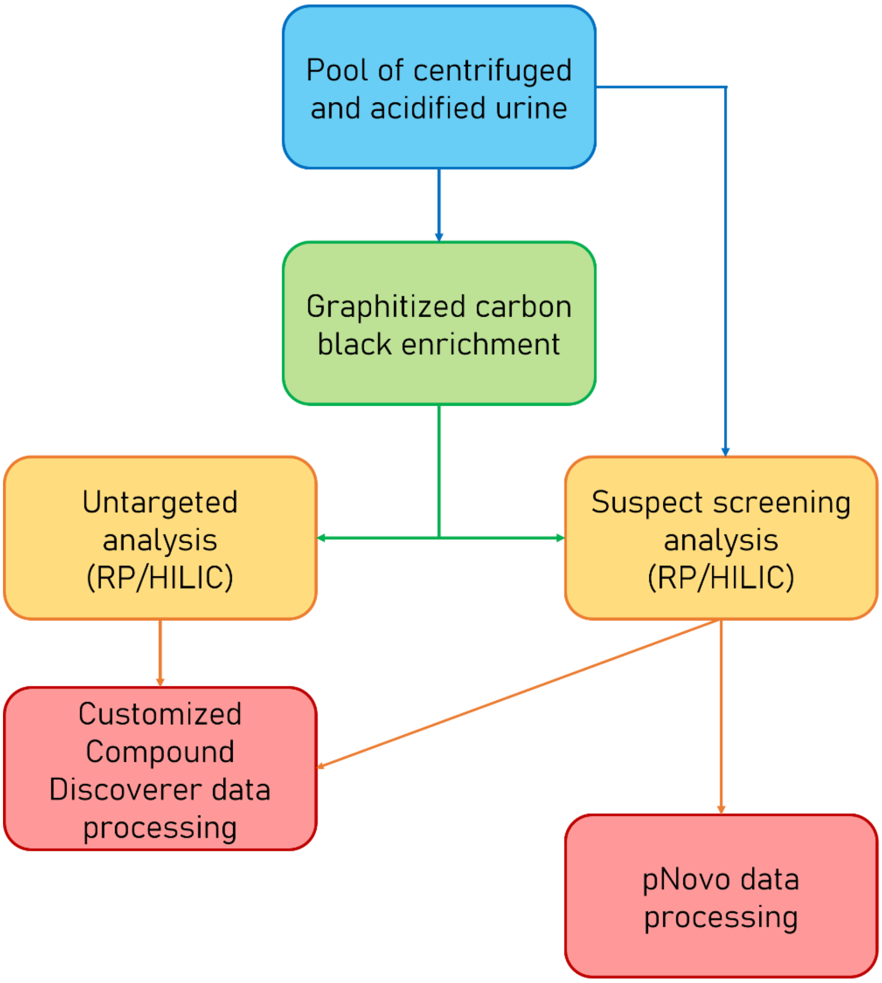

Pool of centrifuged and acidified urine
Graphitized carbon black enrichment
Untargeted analysis (RP/HILIC)
Suspect screening analysis (RP/HILIC)
Customized Compound Discoverer data processing
pNovo data processing

Figure 2. Flow chart representing the compared strategies for short peptide identification.