THE CHALLENGES OF PEPTIDOMICS IN COMPLEMENTING PROTEOMICS IN A CLINICAL CONTEXT

Evelyne Maes,[1,2,3] Eline Oeyen,[1,2] Kurt Boonen,[1,2] Karin Schildermans,[1,2] Inge Mertens,[1,2] Patrick Pauwels,[4] Dirk Valkenborg,[1,2,5] and Geert Baggerman[1,2]*

[1]*Flemish Institute for Technological Research (VITO), Mol, Belgium*
[2]*Centre for Proteomics, University of Antwerp, Antwerp, Belgium*
[3]*Food and Bio-Based Products, AgResearch Ltd., Lincoln, New Zealand*
[4]*Molecular Pathology Unit, Department of Pathology, Antwerp University Hospital, Edegem, Belgium*
[5]*Center for Statistics, Hasselt University, Diepenbeek, Belgium*

Received 14 September 2016; accepted 1 October 2018

Published online in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/mas.21581

Naturally occurring peptides, including growth factors, hormones, and neurotransmitters, represent an important class of biomolecules and have crucial roles in human physiology. The study of these peptides in clinical samples is therefore as relevant as ever. Compared to more routine proteomics applications in clinical research, peptidomics research questions are more challenging and have special requirements with regard to sample handling, experimental design, and bioinformatics. In this review, we describe the issues that confront peptidomics in a clinical context. After these hurdles are (partially) overcome, peptidomics will be ready for a successful translation into medical practice.

Keywords: *peptidomics; biofluids; mass spectrometry; extraction; identification; neuropeptide*

I. INTRODUCTION

In nature, gene-coded translation products not only represent full-length proteins but also small peptides. Although the definition of "a peptide" is a bit arbitrary, they are mostly defined as a chain of 2–100 amino acids that represent molecules within a mass range of 200 Da to 10 kDa. Peptides are widely present in nature and play a significant regulatory role in several biological processes, including signal transduction and anti-hypertensive, anti-viral, and anti-microbial tasks and detecting self from non-self (immunopeptidomics) (Faridi et al., 2016; Martelli et al., 2014). These endogenous peptides include peptides translated from short open reading frames but also peptides released from larger precursor proteins as well as protein-processing and-degradation products. The peptides considered in this review should be clearly discerned from their *in vitro* generated proteolytic counterparts, typically used in bottom up proteomics.

In line with other "omics" fields, detection, identification, and quantification of all endogenous peptides present in a cell, tissue, or organism at a certain time point is defined as "peptidomics" (Schrader & Schulz-Knappe, 2001; Schulz-Knappe et al., 2001). The study of a peptidome has two major goals: i) identify and characterize new (bioactive) peptides or ii) quantify (relative or absolute) peptide levels in a variety of samples. Either way, peptidomics helps to elucidate the role of these endogenous peptides in their biological environment, and to obtain insights in the pathways that they are involved in. Moreover, peptidomics can be classified in different subfields; one might be interested in bioactive peptides, precursor proteins, protein processing and degradation products, antigen presentation, or a combination of all the above.

Since the advent of peptidomics more than 20 years ago, a broad range of biological samples was analyzed with peptidomics technologies. The majority of published studies have as subject hormone and neuropeptide research in invertebrates (Caers et al., 2015; De Haes et al., 2015; Secher et al., 2016). Clinical peptidomics applications are more limited. This limited number of studies is not due to the lack of interest of peptides in a clinical setting. On the contrary, peptides are key regulatory molecules and their discovery and functional characterization is of high importance in medicine and the lack of studies reflects the challenging nature of peptidomic analysis and data interpretation. As such, current applications of clinical peptidomics include identifying candidate peptide disease biomarkers, identifying peptides involved in cell-cell communication, which peptides elicit immune responses and signal transduction, and map dietary protein digestion (Dallas et al., 2015). Furthermore, peptides are becoming more important as therapeutic agents, and today several peptide-based vaccines are already tested in clinical trials (Skwarczynski & Toth, 2016).

Although, at first sight, it seems that proteomics and peptidomics are not so different because both are just a chain of amino acids (i.e., peptides are smaller, no enzymatic digestion is necessary, and several methodological links can be found with traditional proteomic approaches), some technological differences are present (Schrader et al., 2014). These differences cause issues that are the result of three major peptide characteristics: i) bioactive peptides of interest can be present in very low concentrations within complex biological samples (e.g., peptide hormones in plasma); ii) endogenous peptides range considerably in size (from 2 to over 100 amino acids) and overlap with metabolite and lipid mass range; and iii) heterogeneity in physiochemical properties of endogenous peptides varies widely, and as a consequence, for example, charge states of endogenous peptides as detected in a mass spectrometer differ compared to *in vitro* controlled digested peptides in proteomics approaches, such as tryptic peptides. Endogenous peptides

Dirk Valkenborg and Geert Baggerman share senior authorship.

**Correspondence to*: Professor Geert Baggerman, Flemish Institute for Technological Research (VITO), VITO Centre for Proteomics, Boeretang 200, 2400 Mol, Belgium. E-mail: geert.baggerman@vito.be

sometimes have no positive charge because often basic amino acids are removed in the conversion from precursor to active peptide and commonly post-translational modifications are present at the N-terminus. It is therefore sometimes easier to detect intermediates of the processing than to see the mature peptide. Other, larger peptides might have multiple charge states of 5+ to 10+, which make MS/MS sequence determination not trivial. These characteristics complicate both mass spectrometry measurements/methods as well as data analysis. Moreover, heterogeneity in chemical properties of these endogenous peptides make optimization of separation by liquid chromatography, a trial and error exercise. Furthermore, modifications will impact column retention properties of peptides largely and contribute to the variability in the chromatographic profile. In this review, we will, therefore, emphasize challenges of clinical peptidomics research (Fig. 1).

II. CHALLENGE 1: PEPTIDE EXTRACTION

In complex biological samples, peptides of interest must be preferentially enriched because these matrices typically contain lipids, salts, proteins, and carbohydrates that can decrease the ionization efficiency of peptides due to suppression effects. Sample preparation methods in peptidomics are highly diverse because peptides themselves vary in many characteristics such as size, charge, and hydrophobicity. In addition, many peptides can have a multitude of post-translational modifications with oxidation, acetylation, c-terminal amidation, pyroglutamic acid formation, and glycosylation as the most important ones. Also, different sample types or peptides of interest need different sample preparation methods; the concept of one-size-fits-all often used in proteomics does not apply on peptidomic sample treatment. The most common peptide extraction procedures include: 1) ultrafiltration with different types of molecular weight cut-off membranes (e.g., 10 kDa and 30 kDa) to separate low-molecular weight and high molecular weight fractions. Although this filtration step is fast and easy to apply, it does not permit complete separation of a specific mass range without partial loss and/or partial contamination from undesired fractions (Dallas et al., 2015). 2) Selective precipitation of the (larger) protein fraction with organic solvents (e.g., acetonitrile, acetone, methanol), acids (e.g., trifluoroacetic acid), or the addition of chaotropic agents (e.g., ammonium sulphate) (Vitorino et al., 2012). However, no complete protein removal can be achieved with precipitation and some peptides might aggregate and be lost in the precipitate (Dallas et al., 2015). 3) Solid phase extraction (SPE) columns retain analytes from complex mixtures and can be used to remove interfering compounds and concentrate the sample. The most popular SPE columns for peptidomics applications are Hydrophilic-Lipophilic Balance (HLB) or C18 columns. Alternatively, size exclusion chromatography (SEC) can be applied, and although it is time consuming, it allows isolation of specific mass ranges. 4) Magnetic beads for extraction of target peptides in complex samples. These magnetic beads can be functionalized for affinity purification (Safarik & Safarikova, 2004). In the case of immunopeptidomics, peptides can be extracted after immunopurifying the major histocompatibility (MHC) complexes. 5) Peptides can be loaded on a gel and separated from contaminants and proteins. The peptides can be re-extracted from the gel after separation. A general overview of a peptidomics protocol is shown in Figure 2.

The pre-analytical phase is, besides the peptide extraction approach that is chosen, also challenged by several other variables, endogenous, and exogenous, which can affect the results. Although some of these variables are controllable (e.g., time and method of collection, freeze-thaw cycles), several others are not (e.g., ischemia time, diet, alcohol use), and both classes of variables can influence peptide profiling results (Vitorino et al., 2012). Influences of these factors are already reported in various peptidomics studies (de Jong et al., 2011; Leichtle et al., 2013). These deviations in sample collection and handling might significantly alter the clinical peptidome. Consistent sample handling during peptidomics experiments

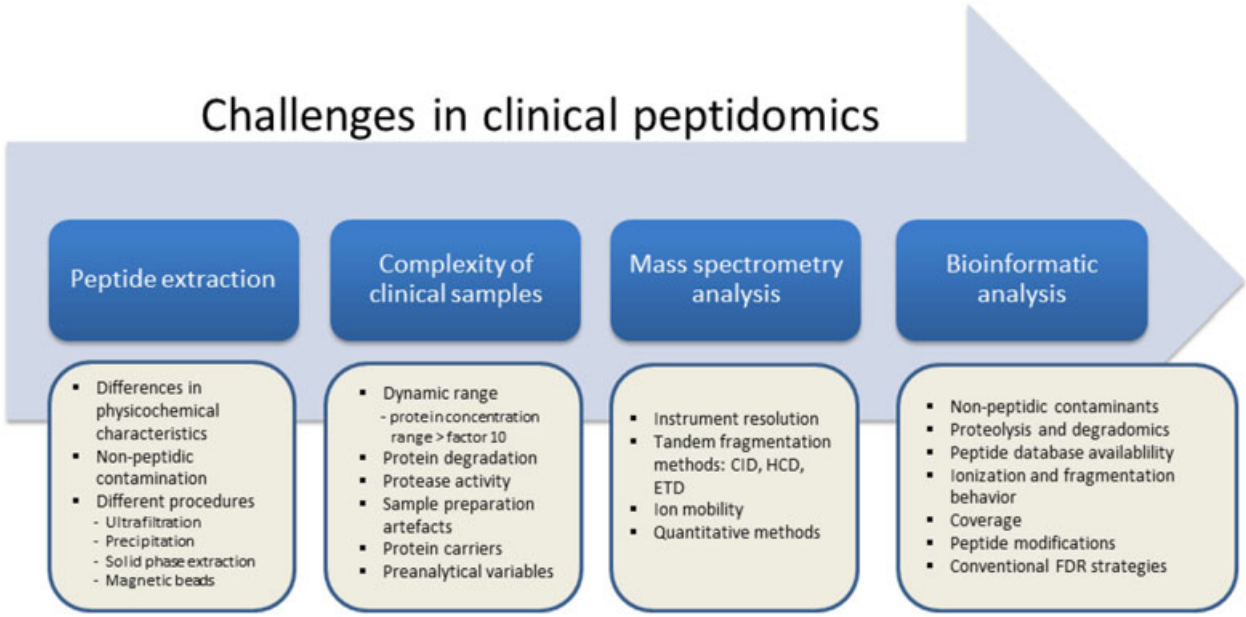

FIGURE 1. A schematic overview of the major challenges in clinical peptidomics.

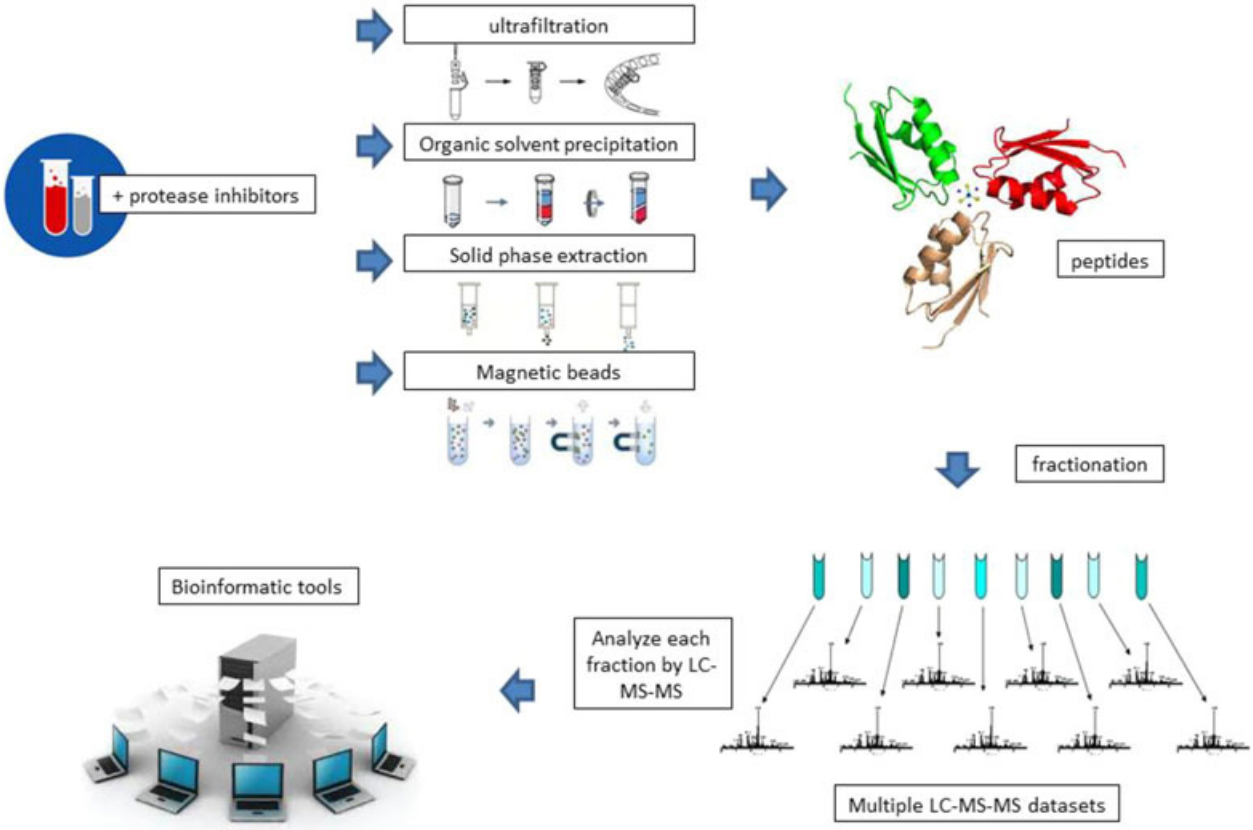

FIGURE 2. A Peptidomics workflow in a nutshell. Although a one-approach-fits all cannot be applied in clinical peptidomics, mostly, these different steps are applied.

and between different experimental groups, combined with a solid design of experiments is thus of critical importance. Additionally, minimization of peptidase activity with protease/peptidase inhibitors remains crucial in all peptidomics applications to circumvent proteome and peptidome degradation. These protease/peptidase inhibitors, however, are peptides (analogues) themselves and added to the sample in over-abundance, can suppress the original peptidome analysis.

III. CHALLENGE 2: THE COMPLEXITY OF CLINICAL SAMPLES

Analysis of endogenous peptides in biofluids or tissues can provide valuable insights into disease mechanisms (Sigdel et al., 2014). The study of peptide expression/protein degradation with peptidomics in order to find peptide signatures in disease versus healthy conditions is, therefore, very relevant. The detected peptides might also have utility as potential clinical biomarkers. Furthermore, because peptides are all low molecular weight molecules, they are more permeable between tissues compared to larger proteins, that might facilitate their detection in peripheral tissues and biofluids. A wide range of clinical tissue types have been used in peptidomics studies, mostly in the search for peptide-based biomarkers (Lai et al., 2015). Complexity of the clinical samples, however, complicates the biomarker search. In the next section, we provide an overview of this complexity in the most commonly used biofluids.

A. Blood

The protein and peptide content in blood reflects the secretion and release of several organs, tissues and cell types, and can be seen as a very dynamic and complex matrix with large potential in medical and pharmaceutical applications. Analysis of endogenous blood peptides, however, is hampered by the presence of vast amounts of peptides that result from blood peptidase activity and *in vivo* protein turn-over. Whereas peptidase activity during and after blood collection adds further complexity to the peptide content, strategies to preserve blood stability and to quench the functioning of *ex vivo* peptidases are necessary to study the endogenous blood peptidome. Part of the complexity introduced by degradation might be lowered with several extensive pre-fractionation steps; however, minimization of peptidase activity remains crucial because peptidase and protease activity will not only result in more protein degradation products in the sample but likely also degrade part of the *in vivo* peptidome.

Two types of liquid blood fractions, plasma, or serum, can be obtained after blood collection. Obviously, the choice between plasma or serum for specific peptidomics studies depends on the study objective. Plasma is defined as the liquid part of blood after centrifugation, thus without cells. To collect plasma, blood is withdrawn in tubes with anti-coagulant additives, such as EDTA, citrate, or heparin, to minimize blood coagulation, whereas serum can only be obtained when blood clotting is finished, typically after 30–60 min. Serum is the

liquid fraction that remains after precipitation of the blood clot and the remaining cellular blood fraction. During clot formation, however, proteins and peptides can be non-specifically trapped. Additionally, clotting induces cell lysis and protease activity, because blood clotting is a multiprotease event that cleaves many (highly abundant) proteins and thus releases large quantities of mostly uninformative peptides. Although serum has often been used in clinical peptidomics studies, the peptide complexity is even higher compared to plasma, because serum contains a higher number of peptides that appear only after clotting events. To prevent protease and peptidase activity during blood withdrawal for plasma collection, sampling and processing several options are possible including the use of blood tubes that contain a special protease/peptidase inhibitor cocktail, the addition of acids or snapfreeze samples. All these methods obtain improved results compared to "crude" plasma, but none deliver optimal results. For example, adding protease/peptidase inhibitors will interfere with peptidomic analyses because they are peptide analogues themselves. Currently, the major obstacle for the use of plasma for clinical peptidomics studies is that peptides usually emanate from degradation processes to make it difficult to distinguish artifact peptides that originate during plasma collection and/or sample handling from those that are uniquely derived from the disease (Aristoteli et al., 2007). Additionally, binding of endogenous peptides to carrier proteins (e.g., the albuminome) is to be avoided to achieve complete peptide extraction in blood peptidomics studies (Gundry et al., 2007; Scumaci et al., 2011).

B. Urine

The easy and non-invasive nature of urine collection, combined with the availability of the large amount of sample that can be collected, make urine an attractive biofluid for clinical applications (Olivieri & Rai, 2010). Because it has been demonstrated that different disease states alter the profile of the urinary proteome and peptidome and/or increase urinary peptide excretion, "omics"-analysis of this clinical sample is growing (Martelli et al., 2014). However, the complexity of the "urineome" complicates the analysis. The urine peptidome, for example, is made up of small soluble peptides that derive naturally from plasma, and are excreted by filtration in the kidney or originate from the urogenital tract. In addition, urine is stored in the bladder for several hours before excretion resulting in degradation of proteins by the proteolytic activity of endogenous proteases which increases the peptide abundance substantially (Bauca et al., 2014). Furthermore, additional variation in the urine peptidome is created as a consequence of altered daily intake of fluids, diets, metabolic processes, and circadian rhythms (Martelli et al., 2014). Also, differences are noticed between first and second void and between first stream and mid-stream urine because bacteruria/hematuria is minimized in the later (Delanghe & Speeckaert, 2014). Despite the high biological variation, urine seems more stable compared to blood, because it is demonstrated that the urine peptide pattern is more stable at room temperature than peptide analysis of serum (Fiedler et al., 2007).

C. Other Clinical Biofluids

Besides the two most common biofluids applied in omics research, peptidomics analysis of other biofluids also has potential in clinical applications. Peptidomics studies have been reported in saliva, cerebrospinal fluid (CSF), tears, and miscellaneous fluids such as seminal fluid (Fung et al., 2004) and vitreous humor (Rollin et al., 2004).

Interest in the analysis of saliva originates from its non-invasive sample collection. Saliva is a body fluid secreted by glands of the oral cavity, where it lubricates the oral cavity and participates in food digestion and prevention of infections (Bauca et al., 2014). The salivary peptidome is complex and variable. Indeed, as protein digestion occurs as soon as proteins enter the oral cavity and continues after saliva sample collection, endogenous peptides, which, for example originate from acinar gland cells, will be overwhelmed with exogenous peptides (Martelli et al., 2014). Other pre-analytical variables such as gender, age, diet, and circadian rhythms also play roles in peptide composition of saliva (de Jong et al., 2011; Bauca et al., 2014).

CSF is a colorless liquid that surrounds the brain and spinal cord, and provides mechanical protection, waste product removal, metabolite circulation, and central nervous system homeostasis regulation. Peptides present in CSF fluid originate either from blood filtration or from the brain tissue itself (Martelli et al., 2014). The invasive nature of sample collection however, makes it inappropriate for general screening of presumably healthy individuals or all patients with neuropathologies (Bauca et al., 2014). Most CSF peptidomics analyses are, therefore, not aimed at biomarker research but are often devoted to the analysis of amyloid-beta peptides in relation to neurodegenerative diseases (Hansson et al., 2017; Martelli et al., 2014).

Peptidomic analysis of tears is stimulated by the non-invasive sample collection and can be useful in a wide range of clinical applications regarding ocular pathologies because its composition reflects the physiological condition of the underlying tissues (Azkargorta et al., 2017). Besides the major function of tears as humidifier of the eyes, tears also prevent infection and are a barrier to the outside environment (Martelli et al., 2014). Although initially not expected, the tear peptidome is quite complex, and the peptide concentration range spans several orders of magnitude (Hayakawa et al., 2013). Although inter-day variation seems low, a remarkable inter-individual variation adds even more complexity to the analysis (González et al., 2012). Analysis of the tear peptidome is mostly focused on the characterization of naturally cleaved tear peptides because they can be bioactive and fulfill specific functions not ascribed to their original protein parents such as antimicrobial activity or intercellular signaling (Azkargorta et al., 2017).

IV. CHALLENGE 3: MASS SPECTROMETRY (MS) ANALYSIS

Due to the fact that peptides have a large diversity of physicochemical properties, a wide variety of MS methods have been employed to tackle this challenge in peptidomics research. Electrospray ionization (ESI), matrix-assisted laser desorption/ionization (MALDI), and surface-enhanced laser desorption/ionization (SELDI) sources, as well as liquid chromatography (LC) and capillary electrophoresis (CE) have successfully been applied in peptidomics studies. In peptidomics, it is, furthermore, of crucial importance for peptide identification that the MS method has both an accurate mass measurement (instrument mass resolution is fundamental to reduce the number of peptide

possibilities) and provide tandem mass fragmentation to provide additional information for peptide sequencing (Dallas et al., 2015). The most commonly used peptide fragmentation method is collision induced dissociation (CID), highly modified peptides (e.g., glycopeptides) or larger peptides mostly require alternative fragmentation methods such as electron-transfer dissociation (ETD) and high-energy collision dissociation (HCD) or a combination of different fragmentation methods (Shen et al., 2011). A comprehensive overview of all these MS techniques applied in peptidomics can be found in Aristoteli et al. (2007), Cunningham et al. (2012), Dallas et al. (2015), and Romanova et al. (2013).

In clinical samples, peptide extractions are most often contaminated with metabolites and other non-peptidic features. This multitude of singly charged interferences can induce peptide ion suppression and thus represent an additional challenge. Large fractions of lipids and metabolites will also saturate the binding capacity (sometimes irreversibly) of the chromatographic columns. Although not frequently used, ion-mobility mass spectrometry (IM-MS) and high-field asymmetric waveform ion mobility spectrometry (FAIMS) could be implemented in the MS workflow. These electrophoretic gas-phase techniques allow one to distinguish and separate molecules based on their charge, mass, and mobility (Harvey et al., 2011). This way, it enables the separation of co-eluting species or low abundant features from chemical interferences, such as non-peptidic (e.g., metabolites) contaminants based on their structure (Xia et al., 2008). Additionally, it allows for separation of peptide isomers (Jia et al., 2014) or discrimination of different modification sites of the same modification on the same peptide backbone (Ibrahim et al., 2011).

In line with proteomics, quantitative analysis of peptide profiles under different physiological conditions is emerging and represents one of the most challenging tasks in the field. The most commonly used methods for peptide quantification include label-based (Boonen et al., 2018) (isobaric and isotopic) and label-free methods and methodological advances for data interpretation are evolving; however, this part remains challenging (Verdonck et al., 2016). Additionally, targeted methods can be applied to obtain absolute quantification results. An overview of all quantitative peptidomics methods is out of scope of this review but is provided in Romanova et al. (2013). A general rule in setting up LC-MS analysis of endogenous peptides is that for every research question, the experimental parameters (buffers, columns, gradients, spray, CID, etc.) must be optimized in a trial-and-error fashion, mainly due to the large peptide diversity in different tissues and physiological conditions.

V. CHALLENGE 4: BIOINFORMATIC ANALYSIS AND PEPTIDE IDENTIFICATION

Bioactive peptides and peptide products from proteolysis are both biopolymers composed out of a chain of amino acids. For peptide identification, strategies that are rooted in peptide-centric proteomic methods can be applied to endogenous peptides as well. However, a shift in focus on peptides rather than on proteins has repercussions on the data analysis in several aspects as will be discussed in this section (Fig. 3). First, we provide an overview of the classical paradigm used to interpret shotgun proteomics data.

First, the most popular search strategy in peptide-centric proteomics are database search methods that match a theoretical fragment spectrum from a candidate peptide sequence to an observed fragment spectrum. Typically, candidate peptides originate from an *in silico* digested protein database and are selected for comparison based on the mass of the precursor ion. Popular search engines are MASCOT (Perkins et al., 1999), SEQUEST (Eng et al., 1994), X!Tandem (Craig & Beavis, 2004), and OMSSA (Geer et al., 2004). A review of search algorithms is provided by Nesvizhskii et al. (2007) and Shteynberg et al. (2013).

Second, a spectral library search strategy can be adopted to identify tandem mass (MS^2) spectra. This approach attempts to match observed fragment spectra to a library of previously observed and high quality annotated spectra based on the precursor mass (Lam et al., 2007; Falkner et al., 2008). Software methods are MS InsPecT (Tanner et al., 2005), SpectraST (Lam et al., 2008) and BiblioSpec (Frewen et al., 2006), etc. Of course, you need a comprehensive spectral library of endogenous peptides for this approach. Neuropedia is a neuropeptide spectral library that can be queried with the M-Split spectral library search tool (Kim et al., 2011).

Third, a *de novo* algorithm can be employed to derive peptide identifications based on the MS^2 spectrum peak patterns (Nesvizhskii et al., 2006). The advantage of the latter method is that it does not require prior knowledge about a protein database or a reference library. Principles from computer science and mathematics are adopted to interpret the spectra as an optimization problem that tries to relate the masses of the fragment ions to a series of amino acids. Peaks software, PepNovo (Frank, 2009) and Novor (Ma, 2015) are examples of a *de novo* approach. An excellent review of *de novo* methods are provided by Nesvizhskii et al. (2007) and Menschaert et al. (2010).

Other search strategies exist that combine concepts of the previously mentioned strategies, such as sequence-tag assisted database searching (like Peaks DB), and its variations using "spectral dictionaries" and gapped peptides (Menschaert et al., 2010).

The length of some bioactive peptides can pose problems, and new strategies for peptidomics in the mass range between 3 and 8 kDa have been published recently (Budamgunta et al., 2018).

Because amino acid chains in digested peptides and bioactive peptides have similar structures, they could be analyzed and identified with the same bioinformatics machinery. However, some lacuna in the underlying assumptions could jeopardize a valid peptide identification. In this section we will use the database search strategies as a case example to point out the culprits that hamper peptidomic identification. The misconceptions that are illustrated in this section can be generalized to the other search strategies as well.

A. Contaminants

During peptide extraction, it is unavoidable that non-peptidic content, such as lipids and other types of metabolites, can also be enriched. Typically, LC-MS peptidome analysis is highly sensitive to the presence of these contaminants, and if these are not effectively removed, they will impair the performance of LC-MS analysis. (see Sigdel et al. (2014) in the case of urinary peptidomics). Strategies to remove these interfering agents

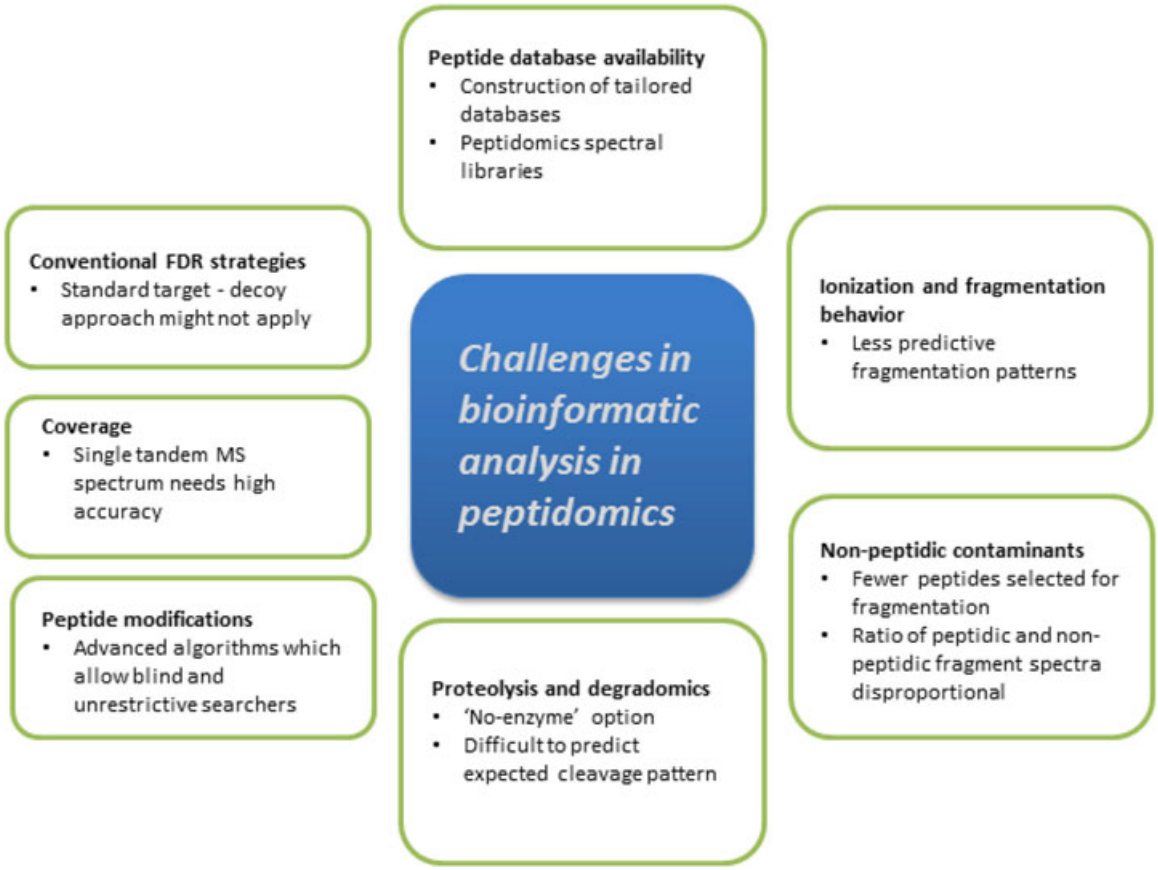

FIGURE 3. Overview of the different bioinformatic challenges in clinical peptidomics experiments.

include: ultrafiltration, lyophilization, and vacuum concentration, ion exchange or reverse phase strategies, solid phase extraction (SPE) or serial solid phase extraction (called modified SPE by Sigdel et al. (2018)), where peptides isolated from SPE are subjected to a second step of purification using a processed silicon carbide resin in a pH dependent manner (Sigdel et al., 2014).

Peptidomic samples can contain low amounts of peptides and it should be noted that the low total amount and the lower amount of different peptides makes the LC-MS analysis more prone to carry-over, especially if proteomic samples or QC samples (like BSA) are run on the same column. Low intense peaks from previous samples will be selected for fragmentation if the sample itself does not saturate the capacity for fragmentation of the instrument and method. A search against the whole proteome, including regularly occurring protein contaminants, is therefore always recommended.

In contrast to peptidomics, proteomics approaches are relatively robust against these contaminants. Generally, non-protein contaminants are removed from the sample with a combination of protein precipitation and a form of filtration. Protein digestion (with e.g., trypsin) will create and enrich the peptidic content in the sample with respect to the non-peptidic content. As a result, the influence of the non-peptidic content in the sample is diluted. This enrichment of peptides ensures that in a shotgun proteomics setting which applies with data-dependent acquisition, mostly peptides are selected for further fragmentation and downstream identification. When one applies the shotgun strategy and corresponding data dependent acquisition to a peptidomic sample with its non-peptidergic contaminants, the ratio of peptidic and non-peptidic fragment spectra is disproportional. Fewer peptides are selected for fragmentation and this depletion of information has consequences on the downstream analysis of the mass spectrometry data.

B. Proteolysis and Degradomics

Processing of endogenous peptides highlights a second challenge in peptidomics data analysis. Many related, yet slightly different, protein and peptide products are present in the sample that increase the complexity, we see for example for many neuropeptides multiples intermediates. A lack of knowledge about the enzyme activity requires employing database search strategies with the "no-enzyme" option because it is hard to predict an expected cleavage pattern. Another option is to use a database of mature peptides and choose a "no-cleave" option. This of course, needs to be complemented with additional searches afterward and can only be applied when the peptides originate from a relatively well known processing pathway. An especially difficult case is blood peptidomics, where a multitude of protease and peptidases are active, to generate a complex and diluted peptide content. This bioprocessing is less an issue in the bottom-up proteomics approach in which the protein content of a sample is digested by an enzyme with a known cleavage pattern. Obviously, semi-tryptic peptides caused by N-terminus and C-terminus processing will be present in the sample;

however the majority will originate from the controlled proteolysis with, for example, trypsin.

C. Peptide Database Availability

The general phenomenon of incomplete protein databases of known and predicted proteins also challenges peptidomics studies. The success of a database search strategy for peptide identification depends on the presence of the correct or homologous amino acid sequence of the scrutinized peptide in the database (Costa et al., 2013). A first approach is to search the protein database with the "no-enzyme" option that results in all possible peptide sequences that takes into account every possible cleavage site of a protein. Such an unconstrained search will lead to an enormous search space and require more time and memory. This problem can be (partially) addressed by new search engines that are much faster than the ones traditionally used, like Morpheus (Wenger & Coon, 2013) or especially MSFragger (Kong et al., 2017). In this case a particular spectrum is compared against a set of candidate peptides that are not likely to be present in the analyzed sample which in turn inflates the false-positive findings and increases the stringency threshold for true positive. Opening up the search is generally not recommended as some search engines lack sensitivity in their scoring function (Nesvizhskii, 2010). A more conservative approach would be to generate a database that contains known and predicted peptide precursor proteins, to increase identification rates (Falth et al., 2006; Menschaert et al., 2009a; Menschaert et al., 2010). This is of course only justified if the sample contains only endogenous peptides, highlighting again the importance of sample handling and peptide enrichment in peptidomics. However, unknown biotransformation and biological peptide processing can hamper the construction of such tailored databases to result in an incomplete peptide database. A very stringent approach, on the other hand, is to allow only bioactive peptides in a database that correspond to the study objective by only selecting those proteins that contain cleavage sites for known proteases and peptidases in the studied sample. Many fragment spectra will remain unidentified with this approach because the knowledge of processing patterns is very incomplete Spectra that originate from other types of peptides in the databases (e.g., by being modified or that have N or C-terminal extensions) can be identified afterward by clustering methods (such as Bonanza, Menschaert et al., 2009b). In addition, the presence of a number of endogenous peptides encoded in unconventional coding regions such as short open-reading frames are reported (Kondo et al., 2007; Ingolia et al., 2011; Hayakawa et al., 2013), making a comprehensive peptide database more difficult. A solution to include these "unconventional peptides" is to consider the six-frame translation of the entire genome (the study of Hayakawa et al. (2013) combines ETD and CID to improve sensitivity when querying against all possible reading frames) (Costa et al., 2013; Hayakawa et al., 2013), potentially the use of RNA-seq reads as in proteogenomics studies (for example (Renuse et al., 2011; Helmy et al., 2012; Woo et al., 2014)) and the use of peptidomic spectral libraries (Wang & Bandeira, 2013).

D. Ionization and Fragmentation Behavior

Ionization and fragmentation of peptides that are the product of a controlled proteolysis are well understood and very abundant. Cases in point are tryptic peptides because cleavage at lysine (K) and arginine (R) ensures basic amino acids at the carboxyl terminus. The basic C-terminus together with the basic N-terminus (amino terminus) yield multiple charges during electrospray ionization, to hence facilitate ionization and fragmentation. Multiply charged ions will produce a series of b- and y-ions that are expected with typical shotgun proteomics search methods, like SEQUEST and MASCOT. Some neuropeptide or peptide hormone intermediates have basic C-terminal amino acid extensions, meaning that the processing intermediates are sometimes easier to detect. However, mature bioactive peptides and non-tryptic endogenous peptides can have unfavorable ionization properties (also negatively influenced by PTMs), to result in lower ion intensities, and in addition, will produce less predictive and often less informative fragmentation patterns when compared to tryptic peptides. This unpredicted behavior might complicate peptide identification strategies with bioinformatics tools that are tailored toward proteomics (Menschaert et al., 2010).

E. Coverage

In peptidomics studies, sometimes only one specific bioactive peptide and/or its breakdown products is/are observed for each precursor. Endogenous peptide identifications thus often rely on a limited number of tandem MS spectra to make high accurate and reliable peptide-to-spectrum matching very important for endogenous peptide identification (Hayakawa et al., 2013). Of course the presence of breakdown products and intermediates increases the confidence of endogenous peptides but are not always easy to identify taking into account the roadblocks that hamper confident peptide-to-spectrum matches mentioned previously. Peptide to spectrum matches (PSMs) from similar peptides (e.g., intermediates) can increase the confidence of PSMs that are based on low quality fragmentation spectra (since real hits cluster around endogenous peptide sequences whereas false hits have equal probability throughout the [decoy] database). However, it should be noted that spectra from intermediates are correlated and cannot be considered fully independent confirmations. Identifications on the precursor level that are based on more than one peptide can be identified with high confidence and make detection of specific forms of a peptide more reliable, but in general only a few separate peptides originate from a precursor and in some cases only a single bioactive peptide is deduced from a single peptide precursor is.

This is in contrast to proteomics where a protein digestion typically results in 20–50 peptides per protein. In this case, some flexibility is allowed because there are multiple opportunities to target and identify a protein based on the corresponding peptides. In proteomics studies the concept of protein coverage is used to indicate the extent of amino acids in a particular protein sequence that is covered by the observed peptide fragments. Obviously, the more peptides that are linked to a protein the higher the confidence that this protein is truly present in the sample. This paradigm is reflected in how protein probabilities are calculated whereas highly abundant proteins are identified confidently with various peptides. Generally, a minimum of two peptides per protein ("two-peptide-rule") is required to boost confidence in the resulting protein identification. However, in some cases, protein identification might depend

on the evidence of only one peptide. These "one-hit-wonders" identification are more demanding on the confidence of the peptide level and often are manually curated in order to reduce the chance on false positive findings. The latter situation describes clearly the conditions that peptidomics studies are operating in, rendering the probabilities of single peptide identifications lower by requiring higher quality fragmentation spectra and individual peptide scores.

Identifications of new peptides therefore need to be rigorously validated, particularly when it is based on a single peptide. Additional confirmation can be acquired by taking into account fragment ion intensities, fragmentation patterns, and retention times. The most cost-efficient way to confirm an identification is to confirm the fragmentation spectrum with a predicted spectrum (Budamgunta et al., 2018). Software that predicts fragmentation patterns (e.g., MS2PIP) including their relative intensities are very accurate these days (Degroeve et al., 2015). Although this software is trained on proteomics data, they do perform well for the prediction of non-tryptic peptides (although this is not checked exhaustively, in part due to the lack of well validated, sufficiently large peptidomics databases to train the algorithms). More confidence can be acquired if the peptide of interest is synthesized and measured on the same LC-MS setup. The fragmentation pattern and relative intensities will be more accurate since they do depend on the ionization and fragmentation settings. Additionally, accurate retention times will provide valuable extra information concerning the identification. The expenses and effort are certainly warranted if the peptide of interest will be used for further targeted measurements.

F. Peptide Modifications

Endogenous peptides are exposed to biotransformation and regulation. As such, these peptides carry post-translational modifications (PTMs) to become biologically active or to improve stability. The most frequently observed PTMs of bioactive peptides are C-terminal amidation, acetylation, pyroglutamate formation at the N-terminus, and sulfatation (Boonen et al., 2008; Menschaert et al., 2009b). Some of these PTMs hamper enzymatic degradation of peptides with peptidases and/or are required for biological activity. Chemical modifications during sample preparation, whether intentionally or not, can further complicate matters. Modifications have an influence on ionization and fragmentation patterns that can complicate their analysis. Database search engines can account for variable PTMs, but the addition of variable PTMs results in a combinatorial growth of the search space, on top of the "no-enzyme" issue or uncertainty about amino acid substitution in the peptide sequence. As a consequence there is a large increase in search time and a decrease in sensitivity that introduces more false positive findings. In the case of proteomics, a constrained view on PTMs can be adopted because missing some exotic modifications would not jeopardize the identification of a protein, because it is very unlikely that all tryptic peptides would carry a PTM. In peptidomic studies, an open view on the possible PTMs has to be considered. For this purpose, conventional strategies must be replaced with more advanced algorithms that allow for blind and unrestrictive searches, as is reported for MSFragger Kong et al. (2017) and Chick et al. (2015). Another option is a multi-stage identification processes that searches a dataset repeatedly, by adding a few variable modifications at each new search step would circumvent the combinatorial increase in the search space. Database search engines can also be combined in a pipeline with software that is more suited for open PTM searches (Menschaert et al., 2009b). These strategies can make use of the concept of sequence tags. The restriction on the search space imposed by a sequence tag is more limiting than the restriction imposed by the mass of a candidate modified peptide (Liu et al., 2006; Na et al., 2008; Dasari et al., 2010). Other algorithms that allow for an unrestricted PTM search are spectrum-to-sequence alignment (Tsur et al., 2005; Tanner et al., 2008; Chen et al., 2009; Ahrne et al., 2010), spectral clustering (Falkner et al., 2008; Menschaert et al., 2009b), peptide motif analysis (Liebler et al., 2002; Liu et al., 2008), or other methods (Havilio & Wool, 2007; Baumgartner et al., 2008).

G. Conventional False Discovery Rate (FDR) Strategies

A commonly accepted strategy in large-scale proteomics is to control the confidence of peptide-to-spectrum-matches by false discovery rate (FDR). This FDR can be calculated "locally" by assigning posterior error probabilities to each peptide-spectrum match (Keller et al., 2005; Choi & Nesvizhskii, 2008; Kall et al., 2008) or "globally" by calculating the proportion of decoy count-based identifications from a target-decoy approach (TDA) (Elias et al., 2005; Elias & Gygi, 2007) in which a database search is conducted on, for example, a concatenated databases, that is, the target database of interest and a decoy database that represents the null. The decoy database is mostly generated by reversing the protein sequences of the target database, but other strategies are possible as well, for example, such as a randomization approach that preserves the distributional properties of amino acids, cleavage sites, and peptide lengths in the target database while minimizing the number of peptide sequences in common between the target and decoy database. A target-decoy approach permits the estimation of the likelihood that a PSM is correct given that it came from a large collection of PSMs for which the false positive finding is controlled for Elias and Gygi (2007). Key in the FDR strategy is that both alternative and null distribution are properly sampled in order to obtain a good estimate of the FDR and that these distributions are well discernable among each other. If count statistics are flawed, several sources of nuisances will bias the FDR estimate, making it impractical to use. In general, for proteomics studies the count statistics are sufficient to confidently estimate the FDR, that is, many high scoring PSM from the target database. On the other hand, the FDR strategy might fail in peptidomics studies (or even be invalid when using restricted search databases) and suffer similar limitations as the case for proteomics in non-model organisms, the search for hypothetical proteins or rare splice-variant in proteogenomics, or in MS-based clinical microbiological typing, where evidence for single PSM findings ("one-hit-wonder") is required. The reason is that the number of chance findings from the decoy database is often close to the number of PSMs from the target database due to factors that are previously described in this section. These factors compromise the identification rate and influence the FDR estimate. According to Gupta et al. (2011) TDA is not designed for evaluating the reliability of individual peptide identification. Furthermore, in a TDA search strategy, peptide identifications are not necessarily

reflecting the actual progenitor sequence. Colaert et al. (2011) state that whenever a peptide closely enough resembles the originator peptide in terms of their fragmentation spectra, the distinction between the real hit and its proxy is nearly impossible and result in "close-but-not-perfect" matches. Note that these "close-but-not-perfect" matches are also found in the decoy database and not necessarily indicate a random hit but a match that is homologous to the actual peptide. Especially, in peptidomics where there is an uncertainty about the actual *in silico* progenitor sequences, these non-random finding in the decoy database might become problematic and cause conservative thresholds is the FDR scheme because target and null distribution severely overlap.

H. Way Out?

Especially for peptidomics data, a combined use of search engines and search strategies can increase the confidence in peptide identifications when a PSM is confirmed with multiple search engines, even when different searches rate a PSM below the threshold confidence. These multistage strategies are already reported for proteomics approaches (Keller et al., 2005; Alves et al., 2008; Kwon et al., 2011) and peptidomics (Menschaert et al., 2010). However, when combining information from multiple search methods caution should be applied. Shen et al. (2011) expressed their concerns for using an FDR decoy search strategy for peptide identification with multiple search strategies, because different sets of peptides were identified at the same low FDR level from the same set of spectra. This observation raised questions on the accuracy of the FDR evaluation in the case of degradomic-peptidomic analysis. Though some tools might be able to tack this ambiguity (Kwon et al., 2011), special attention and additional work are still required on the development of statistical error rate estimation methods that are applicable to multistage peptidomics approaches.

Also, methods and pipelines that can handle the unknown PTMs are important. For this purpose, several spectral matching/clustering strategies have been developed to facilitate identification of new and unexpected modifications, to provide the unmodified and the modified sequence that are present in the sample. The Bonanza clustering method of Menschaert et al. (2009b) explains modified bioactive peptides by their known peptide precursors. Additionally, implementation of bioinformatics tools that try to predict the active proteases based on the observed degradome to sort out protein remnants from bioactive peptides might be extremely helpful (Song et al., 2010, 2012).

In order to deal with non-peptidic contaminants and interferences, Jeong et al. (2012) suggest to remove unidentifiable spectra prior to a search strategy because it would reduce the computation time and positively influence the false discovery rate because unidentifiable spectra can only generate false PSMs. An effort in this direction is provided by "Lipid Centrifuge" of Dittwald et al. (2014) that provides a framework to recognize and triage non-peptidic contaminants prior to data interpretation, in this case lipids.

With mass spectrometry improvements in mass accuracy and advances in peptide fragmentation techniques, *de novo* interpretation methods will become increasingly important. The combination of *de novo* algorithms with classical database search algorithms, like for example, Peaks and Peaks DB (Zhang et al., 2012) are currently available as user friendly software. A very promising development in this context is that *de novo*-assisted method can enable blind PTM search that allow for hundreds of modifications (Han et al., 2011). The restriction that short sequence tags imposed on a database compensates the increase of the search space caused by allowing many PTMs. For further reading on bioinformatics solution for peptidomics, we refer to Menschaert et al. (2010) who reported on tools, strategies, and methodologies within the peptidomics field and the application thereof. Furthermore universal search tools, like MS-GF+ (Kim & Pevzner, 2014; Wu et al., 2015), a mass-tolerant database search approach (Chick et al., 2015) and MS Fragger (Kong et al., 2017), that allow for open and unrestricted database searches can overcome the difficulties in peptidomic spectral identifications laid out in this section.

VI. CONCLUSION

In clinical applications, peptides have large potential, both in use as biomarkers and as well as potential peptidic treatments. Their pharmacological characterization is, therefore, of crucial importance. The study of the endogenous peptide content in clinical samples is more relevant than ever. However, a variety of issues (peptide extraction, complexity of clinical samples, MS, and bioinformatic analysis) make peptidomics research less straightforward compared to proteomics/genomics and other—omics applications. In recent years, several of these hurdles are partially circumvented with technological enhancements that include an improved sample preparation, more sensitive mass spectrometry instruments, and improved bioinformatic solutions. However, continued progress in the peptidomics field is still required to put this "omics" technology to the next level, where peptidomics results might be translated into the clinic.

REFERENCES

Ahrne E, Muller M, Lisacek F. 2010. Unrestricted identification of modified proteins using MS/MS. Proteomics 10:671–686.

Alves G, Wu WW, Wang G, Shen RF, Yu YK. 2008. Enhancing peptide identification confidence by combining search methods. J Proteome Res 7:3102–3113.

Aristoteli LP, Molloy MP, Baker MS. 2007. Evaluation of endogenous plasma peptide extraction methods for mass spectrometric biomarker discovery. J Proteome Res 6:571–581.

Azkargorta M, Soria J, Acera A, Iloro I, Elortza F. 2017. Human tear proteomics and peptidomics in ophthalmology: Toward the translation of proteomic biomarkers into clinical practice. J Proteomics 150:359–367.

Bauca JM, Martinez-Morillo E, Diamandis EP. 2014. Peptidomics of urine and other biofluids for cancer diagnostics. Clin Chem 60:1052–1061.

Baumgartner C, Rejtar T, Kullolli M, Akella LM, Karger BL. 2008. SeMoP: a new computational strategy for the unrestricted search for modified peptides using LC-MS/MS data. J Proteome Res 7:4199–4208.

Boonen K, Landuyt B, Baggerman G, Husson SJ, Huybrechts J, Schoofs L. 2008. Peptidomics: the integrated approach of MS, hyphenated techniques and bioinformatics for neuropeptide analysis. J Sep Sci 31:427–445.

Boonen K, De Haes W, Van Houtven J, Verdonck R, Baggerman G, Valkenborg D, Schoofs L. 2018. Quantitative peptidomics with isotopic and isobaric tags. Methods Mol Biol 1719:141–159.

Budamgunta H, Olexiouk V, Luyten W, Schildermans K, Maes E, Boonen K, Menschaert G, Baggerman G. 2018. Comprehensive peptide analysis of mouse brain striatum identifies novel sORF-encoded polypeptides. Proteomics 18:e1700218.

Caers J, Boonen K, Van Den Abbeele J, Van Rompay L, Schoofs L, Van Hiel MB. 2015. Peptidomics of neuropeptidergic tissues of the tsetse fly

glossina morsitans morsitans. J Am Soc Mass Spectrom 26: 2024–2038.

Chen Y, Chen W, Cobb MH, Zhao Y. 2009. PTMap–a sequence alignment software for unrestricted, accurate, and full-spectrum identification of post-translational modification sites. Proc Natl Acad Sci USA 106:761–766.

Chick JM, Kolippakkam D, Nusinow DP, Zhai B, Rad R, Huttlin EL, Gygi SP. 2015. A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. Nat Biotechnol 33:743–749.

Choi H, Nesvizhskii AI. 2008. Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics. J Proteome Res 7:254–265.

Colaert N, Degroeve S, Helsens K, Martens L. 2011. Analysis of the resolution limitations of peptide identification algorithms. J Proteome Res 10:5555–5561.

Costa EP, Menschaert G, Luyten W, De Grave K, Ramon J. 2013. PIUS: peptide identification by unbiased search. Bioinformatics 29:1913–1914.

Craig R, Beavis RC. 2004. TANDEM: matching proteins with tandem mass spectra. Bioinformatics 20:1466–1467.

Cunningham R, Ma D, Li L. 2012. Mass spectrometry-based proteomics and peptidomics for systems biology and biomarker discovery. Front Biol 7:313–335.

Dallas DC, Guerrero A, Parker EA, Robinson RC, Gan J, German JB, Barile D, Lebrilla CB. 2015. Current peptidomics: applications, purification, identification, quantification, and functional analysis. Proteomics 15:1026–1038.

Dasari S, Chambers MC, Slebos RJ, Zimmerman LJ, Ham AJ, Tabb DL. 2010. TagRecon: high-throughput mutation identification through sequence tagging. J Proteome Res 9:1716–1726.

De Haes W, Van Sinay E, Detienne G, Temmerman L, Schoofs L, Boonen K. 2015. Functional neuropeptidomics in invertebrates. Biochim Biophys Acta 1854:812–826.

Degroeve S, Maddelein D, Martens L. 2015. MS(2)PIP prediction server: compute and visualize MS(2) peak intensity predictions for CID and HCD fragmentation. Nucl Acids Res 43:W326–W330.

de Jong EP, van Riper SK, Koopmeiners JS, Carlis JV, Griffin TJ. 2011. Sample collection and handling considerations for peptidomic studies in whole saliva; implications for biomarker discovery. Clin Chim Acta 412:2284–2288.

Delanghe J, Speeckaert M. 2014. Preanalytical requirements of urinalysis. Biochem Med (Zagreb) 24:89–104.

Dittwald P, Nghia VT, Harris GA, Caprioli RM, Van de Plas R, Laukens K, Gambin A, Valkenborg D. 2014. Towards automated discrimination of lipids versus peptides from full scan mass spectra. EuPA Open Proteom 4:87–100.

Elias JE, Gygi SP. 2007. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods 4:207–214.

Elias JE, Haas W, Faherty BK, Gygi SP. 2005. Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. Nat Methods 2:667–675.

Eng JK, McCormack AL, Yates JR. 1994. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J Am Soc Mass Spectrom 5:976–989.

Falkner JA, Falkner JW, Yocum AK, Andrews PC. 2008. A spectral clustering approach to MS/MS identification of post-translational modifications. J Proteome Res 7:4614–4622.

Falth M, Skold K, Norrman M, Svensson M, Fenyo D, Andren PE. 2006. SwePep, a database designed for endogenous peptides and mass spectrometry. Mol Cell Proteomics 5:998–1005.

Faridi P, Aebersold R, Caron E. 2016. A first dataset toward a standardized community-driven global mapping of the human immunopeptidome. Data Brief 7:201–205.

Fiedler GM, Baumann S, Leichtle A, Oltmann A, Kase J, Thiery J, Ceglarek U. 2007. Standardized peptidome profiling of human urine by magnetic bead separation and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. Clin Chem 53:421–428.

Frank AM. 2009. A ranking-based scoring function for Peptide−Spectrum matches. J Proteome Res 8:2241–2252.

Frewen BE, Merrihew GE, Wu CC, Noble WS, MacCoss MJ. 2006. Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. Anal Chem 78:5678–5684.

Fung KY, Glode LM, Green S, Duncan MW. 2004. A comprehensive characterization of the peptide and protein constituents of human seminal fluid. Prostate 61:171–181.

Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH. 2004. Open mass spectrometry search algorithm. J Proteome Res 3:958–964.

González N, Iloro I, Durán JA, Elortza F, Suárez T. 2012. Evaluation of inter-day and inter-individual variability of tear peptide/protein profiles by MALDI-TOF MS analyses. Mol Vis 18:1572–1582.

Gundry RL, Fu Q, Jelinek CA, Van Eyk JE, Cotter RJ. 2007. Investigation of an albumin-enriched fraction of human serum and its albuminome. Proteom Clin Appl 1:73–88.

Gupta N, Bandeira N, Keich U, Pevzner PA. 2011. Target-decoy approach and false discovery rate: when things may go wrong. J Am Soc Mass Spectrom 22:1111–1120.

Han X, He L, Xin L, Shan B, Ma B. 2011. PeaksPTM: Mass spectrometry-based identification of peptides with unspecified modifications. J Proteome Res 10:2930–2936.

Hansson KT, Skillback T, Pernevik E, Kern S, Portelius E, Hoglund K, Brinkmalm G, Holmen-Larsson J, Blennow K, Zetterberg H, Gobom J. 2017. Expanding the cerebrospinal fluid endopeptidome. Proteomics 17:1600384(1-6).

Harvey SR, MacPhee CE, Barran PE. 2011. Ion mobility mass spectrometry for peptide analysis. Methods 54:454–461.

Havilio M, Wool A. 2007. Large-scale unrestricted identification of post-translation modifications using tandem mass spectrometry. Anal Chem 79:1362–1368.

Hayakawa E, Menschaert G, De Bock PJ, Luyten W, Gevaert K, Baggerman G, Schoofs L. 2013. Improving the identification rate of endogenous peptides using electron transfer dissociation and collision-induced dissociation. J Proteome Res 12:5410–5421.

Hayakawa E, Landuyt B, Baggerman G, Cuyvers R, Lavigne R, Luyten W, Schoofs L. 2013. Peptidomic analysis of human reflex tear fluid. Peptides 42:63–69.

Helmy M, Sugiyama N, Tomita M, Ishihama Y. 2012. Mass spectrum sequential subtraction speeds up searching large peptide MS/MS spectra datasets against large nucleotide databases for proteogenomics. Genes Cells 17:633–644.

Ibrahim YM, Shvartsburg AA, Smith RD, Belov ME. 2011. Ultrasensitive identification of localization variants of modified peptides using ion mobility spectrometry. Anal Chem 83: 5617–5623.

Ingolia NT, Lareau LF, Weissman JS. 2011. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell 147:789–802.

Jeong K, Kim S, Bandeira N. 2012. False discovery rates in spectral identification. BMC Bioinformatics 13:S2.

Jia C, Lietz CB, Yu Q, Li L. 2014. Site-specific characterization of d-amino acid containing peptide epimers by ion mobility spectrometry. Anal Chem 86:2972–2981.

Kall L, Storey JD, MacCoss MJ, Noble WS. 2008. Posterior error probabilities and false discovery rates: two sides of the same coin. J Proteome Res 7:40–44.

Keller A, Eng J, Zhang N, Li XJ, Aebersold R. 2005. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. Mol Syst Biol 1:2005.0017.

Kim S, Pevzner PA. 2014. MS-GF+ makes progress towards a universal database search tool for proteomics. Nat Commun 5:5277.

Kim Y, Bark S, Hook V, Bandeira N. 2011. NeuroPedia: neuropeptide database and spectral library. Bioinformatics 27:2772–2773.

Kondo T, Hashimoto Y, Kato K, Inagaki S, Hayashi S, Kageyama Y. 2007. Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. Nat Cell Biol 9:660–665.

Kong AT, Leprevost FV, Avtonomov DM, Mellacheruvu D, Nesvizhskii AI. 2017. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. Nat Methods 14:513.

Kwon T, Choi H, Vogel C, Nesvizhskii AI, Marcotte EM. 2011. MSblender: a probabilistic approach for integrating peptide identifications from multiple database search engines. J Proteome Res 10:2949–2958.

Lai ZW, Petrera A, Schilling O. 2015. The emerging role of the peptidome in biomarker discovery and degradome profiling. Biol Chem 396: 185–192.

Lam H, Deutsch EW, Eddes JS, Eng JK, King N, Stein SE, Aebersold R. 2007. Development and validation of a spectral library searching method for peptide identification from MS/MS. Proteomics 7:655–667.

Lam H, Deutsch EW, Eddes JS, Eng JK, Stein SE, Aebersold R. 2008. Building *consensus* spectral libraries for peptide identification in proteomics. Nat Methods 5:873.

Leichtle AB, Dufour JF, Fiedler GM. 2013. Potentials and pitfalls of clinical peptidomics and metabolomics. Swiss Med Wkly 143:w13801.

Liebler DC, Hansen BT, Davey SW, Tiscareno L, Mason DE. 2002. Peptide sequence motif analysis of tandem MS data with the SALSA algorithm. Anal Chem 74:203–210.

Liu C, Yan B, Song Y, Xu Y, Cai L. 2006. Peptide sequence tag-based blind identification of post-translational modifications with point process model. Bioinformatics 22:e307–e313.

Liu J, Erassov A, Halina P, Canete M, Nguyen DV, Chung C, Cagney G, Ignatchenko A, Fong V, Emili A. 2008. Sequential interval motif search: unrestricted database surveys of global MS/MS data sets for detection of putative post-translational modifications. Anal Chem 80:7846–7854.

Ma B. 2015. Novor: real-time peptide *de novo* sequencing software. J Am Soc Mass Spectrom 26:1885–1894.

Martelli C, Iavarone F, Vincenzoni F, Cabras T, Manconi B, Desiderio C, Messana I, Castagnola M. 2014. Top-down peptidomics of bodily fluids. Peptidomics 1:47–64.

Menschaert G, Vandekerckhove TT, Baggerman G, Landuyt B, Schoofs L, Luyten W, Van Criekinge W. 2009. Bioinformatics solutions for MS/MS interpretation tailored to the peptidomics field. Commun Agric Appl Biol Sci 74:67–73.

Menschaert G, Vandekerckhove TT, Landuyt B, Hayakawa E, Schoofs L, Luyten W, Van Criekinge W. 2009. Spectral clustering in peptidomics studies helps to unravel modification profile of biologically active peptides and enhances peptide identification rate. Proteomics 9:4381–4388.

Menschaert G, Vandekerckhove TT, Baggerman G, Schoofs L, Luyten W, Van Criekinge W. 2010. Peptidomics coming of age: a review of contributions from a bioinformatics angle. J Proteome Res 9: 2051–2061.

Na S, Jeong J, Park H, Lee KJ, Paek E. 2008. Unrestrictive identification of multiple post-translational modifications from tandem mass spectrometry using an error-tolerant algorithm based on an extended sequence tag approach. Mol Cell Proteomics 7:2452–2463.

Nesvizhskii AI, Roos FF, Grossmann J, Vogelzang M, Eddes JS, Gruissem W, Baginsky S, Aebersold R. 2006. Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. Mol Cell Proteomics 5:652–670.

Nesvizhskii AI, Vitek O, Aebersold R. 2007. Analysis and validation of proteomic data generated by tandem mass spectrometry. Nat Methods 4:787–797.

Nesvizhskii AI. 2010. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. J Proteomics 73:2092–2123.

Olivieri B, Rai AJ. 2010. A primer on clinical applications and assays using urine: focus on analysis of plasma cell dyscrasias using automated electrophoresis and immunofixation. Methods Mol Biol 641:13–26.

Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20:3551–3567.

Renuse S, Chaerkady R, Pandey A. 2011. Proteogenomics. Proteomics 11:620–630.

Rollin R, Mediero A, Martinez-Montero JC, Suarez-Leoz M, Vidal-Fernandez P, Cortes-Valdes C, Fernandez-Cruz A, Fernandez-Durango R. 2004. Atrial natriuretic peptide in the vitreous humor and epiretinal membranes of patients with proliferative diabetic retinopathy. Mol Vis 10:450–457.

Romanova EV, Dowd SE, Sweedler JV. 2013. Quantitation of endogenous peptides using mass spectrometry based methods. Curr Opin Chem Biol 17:1–13. https://doi.org/10.1016/j.cbpa.2013.1005.1030.

Safarik I, Safarikova M. 2004. Magnetic techniques for the isolation and purification of proteins and peptides. Biomagn Res Technol 2:7.

Schrader M, Schulz-Knappe P. 2001. Peptidomics technologies for human body fluids. Trends Biotechnol 19:S55–S60.

Schrader M, Schulz-Knappe P, Fricker LD. 2014. Historical perspective of peptidomics. EuPA Open Proteom 3:171–182.

Schulz-Knappe P, Zucht HD, Heine G, Jurgens M, Hess R, Schrader M. 2001. Peptidomics: the comprehensive analysis of peptides in complex biological mixtures. Comb Chem High Throughput Screen 4:207–217.

Scumaci D, Gaspari M, Saccomanno M, Argiro G, Quaresima B, Faniello CM, Ricci P, Costanzo F, Cuda G. 2011. Assessment of an *ad hoc* procedure for isolation and characterization of human albuminome. Anal Biochem 418:161–163.

Secher A, Kelstrup CD, Conde-Frieboes KW, Pyke C, Raun K, Wulff BS, Olsen JV. 2016. Analytic framework for peptidomics applied to large-scale neuropeptide identification. Nat Commun 7:11436.

Shen Y, Tolic N, Xie F, Zhao R, Purvine SO, Schepmoes AA, Moore RJ, Anderson GA, Smith RD. 2011. Effectiveness of CID, HCD, and ETD with FT MS/MS for degradomic-peptidomic analysis: comparison of peptide identification methods. J Proteome Res 10:3929–3943.

Shteynberg D, Nesvizhskii AI, Moritz RL, Deutsch EW. 2013. Combining results of multiple search engines in proteomics. Mol Cell Proteomics 12:2383–2393.

Sigdel TK, Nicora CD, Hsieh SC, Dai H, Qian WJ, Camp DG, Sarwal MM. 2014. Optimization for peptide sample preparation for urine peptidomics. Clin Proteomics 11:7.

Sigdel TK, Nicora CD, Qian WJ, Sarwal MM. 2018. Optimization for peptide sample preparation for urine peptidomics. Methods Mol Biol 1788:63–72.

Skwarczynski M, Toth I. 2016. Peptide-based synthetic vaccines. Chem Sci 7:842–854.

Song J, Tan H, Shen H, Mahmood K, Boyd SE, Webb GI, Akutsu T, Whisstock JC. 2010. Cascleave: towards more accurate prediction of caspase substrate cleavage sites. Bioinformatics 26:752–760.

Song J, Tan H, Perry AJ, Akutsu T, Webb GI, Whisstock JC, Pike RN. 2012. PROSPER: an integrated feature-based tool for predicting protease substrate cleavage sites. PLoS ONE 7:e50300.

Tanner S, Shu H, Frank A, Shen Z, Wilmarth PA, David LL, Loomis WF, Briggs SP, Bafna V. 2005. InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. Anal Chem 77: 4626–4639.

Tanner S, Payne SH, Dasari S, Shen Z, Wilmarth PA, David LL, Loomis WF, Briggs SP, Bafna V. 2008. Accurate annotation of peptide modifications through unrestrictive database search. J Proteome Res 7:170–181.

Tsur D, Tanner S, Zandi E, Bafna V, Pevzner PA. 2005. Identification of post-translational modifications by blind search of mass spectra. Nat Biotechnol 23:1562–1567.

Verdonck R, De Haes W, Cardoen D, Menschaert G, Huhn T, Landuyt B, Baggerman G, Boonen K, Wenseleers T, Schoofs L. 2016. Fast and reliable quantitative peptidomics with labelpepmatch. J Proteome Res 15:1080–1089.

Vitorino R, Barros AS, Caseiro A, Ferreira R, Amado F. 2012. Evaluation of different extraction procedures for salivary peptide analysis. Talanta 94:209–215.

Wang M, Bandeira N. 2013. Spectral library generating function for assessing spectrum-spectrum match significance. J Proteome Res 12: 3944–3951.

Wenger CD, Coon JJ. 2013. A proteomics search algorithm specifically designed for high-resolution tandem mass spectra. J Proteome Res 12:1377–1386.

Woo S, Cha SW, Merrihew G, He Y, Castellana N, Guest C, MacCoss M, Bafna V. 2014. Proteogenomic database construction driven from large scale RNA-seq data. J Proteome Res 13:21–28.

Wu C, Monroe ME, Xu Z, Slysz GW, Payne SH, Rodland KD, Liu T, Smith RD, et al. 2015. An optimized informatics pipeline for mass spectrometry-based peptidomics. J Am Soc Mass Spectrom 26:2002–2008.

Xia Y-Q, Wu ST, Jemal M. 2008. LC-FAIMS-MS/MS for quantification of a peptide in plasma and evaluation of FAIMS global selectivity from plasma components. Anal Chem 80:7137–7143.

Zhang J, Xin L, Shan B, Chen W, Xie M, Yuen D, Zhang W, Zhang Z, Lajoie GA, Ma B. 2012. PEAKS DB: *de novo* sequencing assisted database search for sensitive and accurate peptide identification. Mol Cell Proteomics 11:1–8.