Peptidomics plays an important role in clinical proteomics and disease-associated biomarker discovery. It has exhibited mounting potential in early noninvasive diagnosis, prognosis, and treatment evaluation of diseases. This article presents an introduction of peptidomics, the entire peptidomic workflows for serum and urine samples, and a brief overview of recent works in this area. The review is designed to enable researchers to find the most suited strategy for their peptidome studies.

**Introduction**

The term ‘peptidomics’, initially mentioned by four groups, was introduced in the beginning of this century. This concept was derived from genomics and proteomics. Peptidomics refers to the qualitative, quantitative, and functional description of the whole biologically active peptides in a cell, tissue, or organism. The peptidome is the low-molecular-weight (LMW) subset of the proteomes, including peptides and small proteins with molecular weights ranging from 0.5 to 15 kDa. From the view of molecular weights, the peptidome studies can fill the gap between proteomics and metabonomics.

Peptides in the peptidome can be divided into two categories (as shown in Fig. 1). One category is the bioactive peptides shed from all cell types in the microenvironment. They may serve as reporters for cell-to-cell communications, such as hormones and cytokines. The other is the peptide fragments cleaved by enzymes resulting from in vivo resident proteins which could reflect biological enzymatic states of individuals. Body fluids are rich in disease-specific peptide and protein candidates, which are the primary resource for peptidome studies. Peptidome biomarkers in human body fluids are expected to forecast disease, diagnose various disorders, guide clinical therapy, and monitor medicine response.

The analysis of body fluids for clinical diagnosis can reduce the need of invasive tests. Serum and urine samples, which could be obtained easily and repeatedly, are sufficient to provide valuable information about dynamic variations within human body. The blood circulating around nearly every organ including tumors keeps direct and close contact with all cells and tissues and contains large quantities of proteolytical peptides, proteins, and protein fragments. Recent studies have shown that the serum peptidome has been a tempting field. Urine, except for the noninvasive feature, has other appealing properties such as the low concentration of proteins and easiness for enrichment. Among body fluids, urine is especially attractive for biomarker discovery in urological diseases.

**Experimental methods**

As so-called little sister or daughter of proteomics, analytical process of peptidomics is quite similar to the common strategy of proteomics. The peptidome workflow of biofluid samples usually contains sample storage, preparation, separation, detection, data mining, and biomarker identification and quantification. Several aspects have to be taken into account during the development of a method for peptidome studies. Hereby, to follow a standardized protocol is necessary since it could provide a guideline for performing highly efficient peptidome studies and achieving accuracy results. The main technical procedures are introduced here.

1 **Sample storage and preparation**

Since the peptidomics focuses on endogenous peptides, sample storage becomes an important work which influences final results. Referring to the regular steps, fasting blood samples are
collected to clot at room temperature for 1 hour and then centrifuged at the speed ranging from 1000 to 4000g for about 10 minutes. The clotting time and temperature show great influences on serum qualities, whereas freeze–thaw cycles are not so important. Serum (supernatant) samples are supplemented with a cocktail of protease inhibitors, and then kept in aliquots at −80 °C until utilized. Before being processed for extraction, thawed serum samples are centrifuged at more than 10 000g for 15 min to exclude most of the lipids and insoluble components.

First-void morning urine samples should be kept at −80 °C after immediate collection for further peptide extraction. Quintana et al. added protease inhibitors into each sample in order to avoid proteolysis. Before utilization, samples are thawed at room temperature for 15 min and centrifuged at approximately 1000g to 1500g, 4 °C for 10 minutes to remove cell debris.

Ahmad et al. suggested not to store samples over −20 °C in order to prevent protein degradation. For long-term storage, keeping in liquid nitrogen is preferred. All samples should be collected consecutively, but not at distinct times to reduce the storage-induced noise in high-throughput experiments.

2 Extraction of peptides

Polypeptides are required to be separated from high molecular weight (HMW) proteins, other organic compounds, and inorganic components before analysis. With the wide dynamic range of protein abundance and the interference of highly abundant HMW proteins, the extraction and enrichment of polypeptides in complex biological mixtures remain a challenging task. For peptidomic analysis, the basic separation mechanism is the molecular weight difference between proteins and peptides. Serum, which contains proteins in a large dynamic range, is more complex than the urine sample. Depletion of HMW proteins in serum seems especially critical. Nevertheless, the inherent interferences in urine samples such as pH and salt concentration also yield great background and noise. Theodorescu et al. believed that urinary peptides are stable and will not undergo significant proteolysis within several hours of collection. In contrast, activation of proteases in blood samples is often associated with collection.

In order to achieve high extraction yield of LMW proteins and peptides, various works have aimed at depletion of high-abundance proteins. So far, many methods and technologies have been applied to extract peptides and small proteins, such as organic solvent precipitation, magnetic beads, solid-phase extraction (SPE), affinity removal column, strong cation exchange (SCX), etc. Snap-freezing extraction has also been used, which could lower the protein degradation and reduce the sample complexity. Furthermore, restricted access materials (RAM) have been proved to be effective and convenient to exclude HMW proteins and retain LMW proteins and peptides in the pores. These materials are based on the mechanisms of size exclusion chromatography (SEC) and reversed phase (RP) adsorption. A novel mixed mode of strong cation exchange/size exclusion capillary trap column was developed to selectively extract peptides for the injection into nano liquid chromatography. C-8-functionalized magnetic nanoparticles, C-18 absorbent, nanoporous silica particles, derivatized mesoporous silica beads, ZnO–poly(methyl methacrylate) nanobeads, mesoporous silica chips, and self-assembled TiO₂ nanocrystal clusters have been reported with various characteristics. It is also reported that 2,5-dihydroxybenzoic acid allows the effective extraction of endogenous peptides from tissues and the extracts can be stored without frozen for a long time. Appropriate sample extraction methods lead to the simplification and acceleration of the whole process for peptidome analysis. Collectively, all these measures are dedicated to identify LMW proteins and peptides. Each method has its own strength and weakness. A most appropriate method can be chosen from all these approaches according to their efficiency of HMW protein depletion and LMW protein (peptide) extraction. Here we focus on three convenient and widely used methods, which are precipitation, ultrafiltration, and SPE.

Organic solvent precipitation has been employed to remove HMW proteins. The addition of organic precipitants into protein solution results in the precipitation of denatured proteins, while peptides and LMW proteins could dissolve in the high concentration organic solvents. Polson et al. interpreted the process in two aspects. The dielectric constant of the protein mixtures decreases when organic solvent precipitants are added in, which makes the close attraction of charged molecules. On the other hand, organic solvent displaces the ordered water molecules around proteins. Hydrophobic interactions lead to protein aggregation and precipitation. Especially, acetonitrile (ACN) depletion was considered as the most effective and reproducible approach. Kawashima et al. developed a method called differential solubilization to extract peptides and LMW proteins in the subnanomolar range; four biomarker candidates of colon cancer were successfully discovered. However, precipitation presents some limitations, such as non-specific depletion and dilution of the sample. On the whole, the organic-precipitation method is rapid, simple, and cheap.

Ultrafiltration is realized by filtering the sample using a membrane of a defined molecular weight cut-off (MWCO) based on size exclusion mechanism. Greenwood and Simpson compared four commercial filter membranes whose nominal MWCOs are similar, and chose the one with the best overall performance for optimization. Zougman et al. explored centrifugal ultrafiltration to analyse the peptidome of the cerebrospinal fluid (CSF) samples; the result shows the confident identification of 563 peptide forms. Although some drawbacks have been encountered in mass transfer efficiency, flux limitation, and simultaneous concentration of LMW contaminants, ultrafiltration is a simple and routine method.

In order to get high enriching capacity, SPE has been widely used in order to desalt and retain peptides. Various materials in different mechanisms are used, which contain nonpolar (e.g. C-8, C-18), polar (e.g. –CN, –NH₂), and other affinity materials (e.g. PS-DVB adsorbent). In recent applications, SPE is often combined with other pre-separation techniques, such as novel RP-SCX SPE precolumn and SEC/SPE-RAM online extraction, for the maximum coverage of peptidome. A study has been carried out to evaluate the performance of SPE, ultrafiltration, ACN precipitation, and SEC; and the result indicates that SPE exhibits the best peptide recovery. Collectively, SPE presents a good capability of HMW protein depletion and LMW protein (peptide) extraction, but a relatively lower reproducibility than ACN-depletion and ultrafiltration.
3 Fractionation (separation)

Though peptides have been extracted from serum and urine samples, they cannot be injected immediately into the mass spectrometer due to the high complexity of the extracted sample. For example, the peptides with the same amino acid composition but different sequences generate the same mass spectrometry (MS) spectra of molecular ions, when MS is used as the analysing technique. Therefore further separation steps are essential and even indispensable for the biomarker identification and discovery. The one-dimensional (1D) and two-dimensional (2D) gel electrophoresis (GE), which have sprung up over the last decades, could provide the staining intensity of a particular protein spot for quantification purposes. However, with the popularity of MS, the signal intensity from the mass spectrometer for quantification is preferred rather than the protein staining intensity used in the gels. Many ongoing quantification methods are contributed to the gel-free separation approaches. Thus, qualification and quantification by GE will not be covered here.

Capillary electrophoresis (CE) is gaining increasing attention. It offers unique advantages of high resolving power, high efficiency, short analysis time, low sample consumption, and low running cost. However, CE requires an adequate ionic strength buffer, which is not compatible with ESI-MS. Therefore three types of interfaces have been developed to solve this problem: sheathless, liquid junction, and sheathflow interfaces. Now CE has become an established technique in biofluid analyses, particularly for the extraction from biologic samples. Wang et al. have explored the offline coupling of CE to MS analysis and optimized the separation, deposition, and subsequent MS detection processes. In their experiment, the use of 2,5-dihydroxybenzoic acid as a multi-functional agent offers a simplified and improved protocol for the enhanced extraction and improved spectral quality of neuropeptide analysis.

As a hyphenated technique for MS, liquid chromatography (LC) remains the most popular method for peptide separation. It overcomes the disadvantage of the limited loading capacity in CE technique. In general, up to 100 microlitre quantities can be loaded into an LC column, whereas a CE can only be filled with a maximum of 1 μL, commonly around 50 nL. Bakun et al. have successfully used LC-MS system to study the global peptides for two sets of serum samples from healthy individuals and patients.

Owing to the high complexity of peptidome, 1D separation is insufficient in resolving power. Therefore, multidimensional separation technology with different mechanisms has been developed for the in-depth analysis of biosamples. Tenorio-Laranga et al. have combined 2D-RPLC to MS to identify the peptides in rat brain. Theoretically, more than two dimensional separations can be applied in order to achieve high resolution and large peak capacity. However, the relatively long period of analysis time and complicated data processing could hamper its use for high-throughput profiling of clinical samples.

4 Detection and identification

Mass spectrometer is an accurate and sensitive device which enables simultaneous analysis of a great number of peptides and proteins. However, characterization of endogenous bioactive peptides is challenging due to the high complexity of biofluid samples. Detection of numerous fragments necessitates the development of high-throughput, accurate, and sensitive mass spectrometer. Meanwhile the development of MS has been accelerated by the genome project and the proteomic research. Currently MS is regarded as a gold standard for protein profiling. Matrix-assisted laser desorption/ionization (MALDI) and electrospray ionization (ESI) are two commonly used soft ionization techniques which can effectively change solid or liquid samples into gaseous ions without excessive fragments.

MALDI is mostly used as an offline interface from LC or CE to MS. It is considered as a useful tool to detect biological samples owing to the ability of analysing complex samples and its high tolerance of salt concentration. Integrated LC-ESI-MS platforms are preferred for the analyses of complex biofluid samples. High performance liquid chromatography has been widely used with ESI-MS for bioactive peptide profiling, mainly due to the possibility and availability of online coupling. In addition, surface-enhanced laser desorption/ionization time of flight mass spectrometry (SELDI-TOFMS), which combines the pre-selection of peptides and proteins on a specific surface with a time-of-flight mass spectrometer, can detect multiple proteins and peptides in biological samples. This technique has successfully been applied for disease biomarker discovery in tissue and blood samples for various cancers.

For peptidomic analysis, the key challenges for MS are sensitivity, mass accuracy, and the ability to generate information-rich fragments with tandem MS. There are four commonly used types of mass analysers: quadrupole, time-of-flight, ion trap, and Fourier transform ion cyclotron resonance. These analysers are usually put together in tandem to strengthen their superiorities.

Quadrupole (Q) has a low resolving power but offers high stability when acting as a mass analyser; it is particularly useful as a transmission device or a collision cell when operated in rf-only mode. A time-of-flight (TOF) analyser offers good sensitivity and high resolving power. The MALDI source is normally coupled with the TOF analyser. The hybrid system (Q-TOF) is quite popular, in which a quadrupole system is coupled orthogonally to a TOF. It exhibits an increased mass accuracy and resolution, as well as the MS/MS function. Ion traps are capable of storing a number of ions and selecting specific ions for single- or multiple-stage fragmentation. Moreover, QTRAP (Q-Q/Q-LIT), which combines triple quadrupole with linear ion trap, offers the superior capability of specific modification detections. Fourier transform ion cyclotron resonance (FTICR) MS also traps the ions. It has the highest mass accuracy and resolving power among all types of mass analysers, and also provides MS/MS fragmentation function. Chen et al. made the evaluation of three commonly used types of MALDI instruments for neuropeptide expression analysis, including MALDI TOF/TOF, MALDI FTICR and MALDI LTQ Orbitrap mass spectrometers. The result shows that the MALDI LTQ Orbitrap provides the best overall performance in mass accuracy, dynamic range, and spectral resolution when it is used with high energy collisional dissociation for complete fragmentation.

In order to determine the amino acid sequence of a specific peptide, the tandem MS mode should be chosen in the operation.
The collision-induced dissociation fragmentation technique is incorporated in basically all types of tandem MS. It is based on electron transfer and cleavage of the precursor ions through collisions with inert gas molecules in the collision cell. New fragmentation techniques, such as electron capture dissociation (ECD) and electron transfer dissociation (ETD), have been developed. They are capable of offering explicit fragmental spectra for the determination of amino acid sequence, which are particularly useful in localization of modifications and complete profiling of amino acid sequence.\(^7\)

5 Quantitative evaluation

The advent of high-throughput MS-based proteomics has created a strong demand for compatible quantitative analysis methodologies. Clinical diagnosis further promotes the development of quantification methods. As a result, various techniques have emerged for gel-free quantitative analyses of proteomics and peptidomics samples. Each of these methods comes along with certain strengths and drawbacks. Researchers often excessively rely on the software such that the validity of results could not be correctly assessed. Quantification methods can be commonly divided into three main categories: stable isotope labeling, label-free methods, and multiple reaction monitoring (MRM).\(^44\)

5.1 Stable isotope labeling. For quantitative analysis with good accuracy, isotope-labeling technique is the most widely used approach. It has been applied to a variety of cell systems and biofluid samples. Stable isotope labeling is a rapidly evolving field. Many kinds of tag labeling methods have been developed. Isotope-labeling techniques can be used either after protein extraction with chemical labeling, or in cell culture with metabolic incorporation. Among these methods, the most useful techniques are iTRAQ (Isobaric tags for relative and absolute quantitative) and ICAT (Isotope coded affinity tags).\(^13\) The quantitation is introduced using differential isotopic tags for quantification. Peptidomics are commonly divided into three main categories: stable isotope labeling, label-free methods, and multiple reaction monitoring (MRM).\(^44\)

5.2 Label-free approach. As the name implies, label-free quantification could provide quantitative information without using any form of labeling reagents. Due to the high cost of isotopic labeling reagents, a label-free strategy has been a reasonable alternative when numerous samples need to be analysed. Peptidomic analysis prefers label-free methods, owing to the simplicity of experimental protocols, though it is a bit difficult to quantify small changes in the peptide level without isotopic labeling.\(^72\) In general, highly precise and accurate mass spectra are vital to the label-free quantification as well as the high-resolution mass spectrometer to complex mixture analyses. Rossbach et al.\(^75\) explored a label-free MS-based approach to identify and quantitate a variety of endogenous peptides from rat. They suggested some of the identified peptides as conceivable biomarker candidates for the drug-induced withdrawal behavior. Kultima et al.\(^74\) determined endogenous peptides by label-free relative quantification method and investigated 10 different normalization methods on peptide data generated with LC-MS. The result indicated that the best normalization method could decrease the median standard deviation by 43% on average compared with non-normalized method.

5.3 MRM. Multiple reaction monitoring, also called selected reaction monitoring, is an emerging technology that partly overcomes the shortcomings of shotgun workflows. MRM usually exploits in a triple quadrupole mass spectrometer; the first and the third quadrupoles are to specifically select a predefined precursor and one fragment of it, whereas the second quadrupole acts as a collision cell for fragmentation. It has the potential to reliably quantify analytes of low abundance in complex mixtures with high sensitivity and specificity.\(^79\) The serum extracts after ACN-depletion were fractionated by a nanoflow LC/tandem MS and quantified by the MRM method by Kay et al.\(^21\) This method shows a viable alternative to the immunochemistry-based protein-depletion techniques commonly used for removing high abundance proteins from serum prior to MS-based proteomic analyses. Cirulli et al.\(^65\) exploited the IMAC (immobilized metal affinity chromatography) modified protocol combined with MRM-based triple quadrupole scan modes to achieve a sensitive and rapid analysis for phosphopeptide identification and characterization.

Data processing and bioinformatics

1 Data mining and statistical analysis

The increasing popularity of gel-free proteomic technologies has promoted the development of appropriate statistical methods. Nowadays chemometrics has become an essential method and contributed greatly to the development and flourish of the ‘-omics’ fields such as metabolomics\(^76\) and proteomics.\(^77\) In reality, biological MS data are extremely complex and hard to analyse. The vast amount of data is submitted to data mining
algorithms in order to obtain statistical interpretation. The
algorithms reduce the raw data to a new set of variables which
are statistically correlated to the original data. The few major
components could account for most of the variations in the
data. Multivariate clustering analysis, allowing to obtain
a highest discrimination ratio, is usually performed to dis-
tinguish among the categories of objectives on their peptide
profiles.

There are many data mining strategies for processing and
profiling of multivariate data. All these strategies fall into two
categories. The first is unsupervised methods, among which the
most classical approach is the principal component analysis
(PCA). PCA as an unsupervised method explains the maximum
variation between samples; it utilizes a bilinear decomposition
approach without information of any groupings. The
principal components are calculated from numerous original data, and
they maximize the variability of the original predictors across
samples. Supervised methods, the second category, are suitable
when the inner-group variance dominates over intra-group
variance, including mainly support vector machines (SVM) and
partial least-squares discriminant analysis (PLS-DA). The
SVM method classifies data based upon their maximal separa-
tion using a three-dimensional hyperplane. This method could
minimize the over-training of the data, but is unable to provide
levels of confidence to any classification. PLS-DA clusters two
defined groups in the maximum separation. It is done by a PLS
regression based on PCA analysis. Thus, mapping of data using
PLS-DA model that sharpens the separation is applied to
investigate the differences between two sets of samples and
further to predict classification of unknown samples. Orthogonal
PLS (O-PLS) is an extension of PLS. It selects peaks with a
significant differing median between groups of samples and
works well in metabonomics and proteomics currently. Forshed et al. illustrated that O-PLS models could give the
most potential and robust biomarker candidates with the
removal of structure noise. Skyyt et al. set up an optimized
platform for SELDI-TOFMS and O-PLS data analysis to search
biomarkers of the clinical significance in prostate cancer, which
allows the verification of peptidomic markers for early diagnosis
of the cancer.

The process of internal and external cross-validation should be
performed to validate the model obtained from the classification
analysis and avoid false estimates and overfitting. The data in
peptidome may be divided into two sets, a training set and
a validation set. The training data are used to form multivariate
models for feature selection and training the classifier to derive
a model that is then applied on the validation set. After cross-
validation, the models become much reliable and could well
interpret the variations. The biomarker candidates are then
evaluated using appropriate methods, such as receiver operating
characteristic curve (ROC), that offer prediction accuracy of
individual markers.

Another statistical method to verify biomarker candidates is to
calculate the \( p \)-value between each group using the homosce-
dastic Student’s \( t \)-test. \( P \)-Values interpret the level of significance
for discrimination between two defined groups. As criteria, the
significant biomarker will be confirmed only if it has a \( p \)-value of
less than 0.05 in the \( t \)-test. Candidates with \( p \)-value greater than
0.05 are usually eliminated.

Mass spectra are analysed using search engines. The most
commonly used search engines are Mascot, SeQuest algorithms,
X!Tandem, OMSSA and MS-Fit. There are several peptide sequencing
information databases such as non-redundant NCBI, Erop-Moscow,
SwePep, PeptideDB, Peptidome, PeptideBank, human protein
database and so on. For instance, SwePep is a gathering of information on a number of
endoenous peptides, and the validation is performed by
comparing the fragmentation between the two spectra. Pep-
Bank, created in 2007, is a free database of peptides compiled
from sequence text-mining and public peptide data sources.
This database includes biological and clinical applications, which
is especially useful for the prediction of the binding partners of
biologically interesting peptides.

Following the steps of biomarker discovery and characteristics,
validation should be taken into account according to the pro-
teomic patterns. High false-discovery rate hampers the veracity
of biomarker candidates. Perfect biomarkers could discrimi-
nate disease from healthy controls. It becomes obvious that
separate markers would not be sufficient, but a combination of
well-selected markers could be meaningful. Therefore, specific,
sensitive, and high-throughput immunoassays are primarily
employed to verify the biomarkers. Since suitable and high-
quality antibodies with sufficient specificity and sensitivity are
expensive and hard to optimize, a straightforward validation
method has emerged recently. A multiplexed MRM approach
coupled with stable isotope dilution provides an alternative assay
method. Meanwhile, problems of the reproducibility and
transferability of MRM assays still need to be overcome.

In recent years, protein separation and detection technologies
have reached a matured level such that proteomic processes of
biological samples can be analysed reproducibly and accu-
rately. Based on the successful experiences of proteomics,
researchers could avoid detours in peptidome studies. Many
methods and techniques applied to proteomics are readily
applicable to the peptidome. Although peptidome is a subfield of
proteome, it has some inherent characteristics. First of all, due
to its low molecular weight and low concentration, the peptidome
could be assessed without tryptic digestions. Secondly, the pep-
tidome aims at the identification of endogenous peptides; the
inactive peptides derived from proteolysis are not its research
objectives. Finally, peptidomics focuses on the sequence infor-
mation of the peptides, whereas proteomics emphasizes the
protein identifications.

There is an urgent need for discovering valid biomarkers to
improve diagnostic accuracies for clinical purposes. A wide

Conclusion and future perspectives

This journal is © The Royal Society of Chemistry 2011


This is an Accepted Manuscript, which has been through the Royal Society of Chemistry's internal editorial review.
It is fully citable and will be indexed by Science Citation Index Expanded and Scopus.

Published on 17 February 2011. doi:10.1039/C0AY00705F
attention has been paid to clinical peptidomics and biomarker discovery. The peptidome researches of tissues, CSF, serum, and urine samples are simultaneously ongoing. While tissues and CSF could offer relatively higher specificity and clinical relevance, serum and urine are still more attractive with higher availability and a lower factor of invasiveness. 

Discovery of biomarkers that are of prognostic and diagnostic relevance has been a major focus of proteomic research during the past decade. Nowadays peptidomics is an evolving tool to understand the pathogenesis of disease process and improve clinical outcomes. In particular, many attempts have been made to identify and verify peptide biomarkers in biofluids with proteomic patterns. We hope that this review will serve as a basis for peptidomic biomarker validation will be applied for routine healthcare and guideline to diseases in the future.

Acknowledgements

This work was supported by the Natural Science Foundation of China and Fujian Province Department of Science & Technology.

References

2 N. Minamino, Tampakushi Kakusan Koso, 2001, 46, 1510–1517.