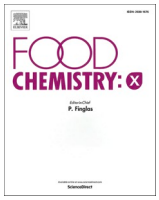
Food Chemistry: X 22 (2024) 101249

Contents lists available at ScienceDirect

Food Chemistry: X

journal homepage: www.sciencedirect.com/journal/food-chemistry-x



PeposX-Exhaust: A lightweight and efficient tool for identification of short peptides

Wanshun Liu [a], Mouming Zhao [a,b], Lishe Gan [a,c], Baoguo Sun [d], Shiqi He [a], Yang Liu [a,e], Lei Liu [a], Wu Li [a], Jing Chen [a], Yang Liu [a], Jianan Zhang [b,*], Jucai Xu [a,*]

[a] Guangdong Provincial Key Laboratory of Large Animal Models for Biomedicine, School of Pharmacy and Food Engineering, Wuyi University, Jiangmen 529020, China
[b] School of Food Science and Engineering, South China University of Technology, Guangzhou 510640, China
[c] College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China
[d] Beijing Advanced Innovation Center for Food Nutrition and Human Health, Beijing Technology & Business University, Beijing 100048, China
[e] College of Food Science and Technology, Hunan Agricultural University, Changsha 410128, China

ARTICLE INFO

Keywords:
Short peptides
Identification
Food-derived peptides
Peptidomics
Protein hydrolysates
Sequencing method

ABSTRACT

Short peptides have become the focus of recent research due to their variable bioactivities, good digestibility and wide existences in food-derived protein hydrolysates. However, due to the high complexity of the samples, identifying short peptides still remains a challenge. In this work, a tool, named PeposX-Exhaust, was developed for short peptide identification. Through validation with known peptides, PeposX-Exhaust identified all the submitted spectra and the accuracy rate reached 75.36%, and the adjusted accuracy rate further reached 98.55% when with top 5 candidates considered. Compared with other tools, the accuracy rate by PeposX-Exhaust was at least 70% higher than two database-search tools and 15% higher than the other two de novo-sequencing tools, respectively. For further application, the numbers of short peptides identified from soybean, walnut, collagen and bonito protein hydrolysates reached 1145, 628, 746 and 681, respectively. This fully demonstrated the superiority of the tool in short peptide identification.

1. Introduction

As known, short peptide sequences are important substances with good digestibility and various bioactivities such as antioxidation, anti-hypertension, et al. (Agyei et al., 2018; Nasri, 2017; Priya, 2019). They have been widely used in biological medicines(Kaur et al., 2021) and nutraceuticals(Suleria et al., 2015) due to their potential health profits. For food processing, short peptides are generally released from the enzymatic hydrolysis or fermentation of food-derived proteins. However, due to the general cutting sites and low purity of the widely used food-grade proteases, an extremely complex peptide composition is usually presented in the hydrolysates(Nasri, 2017). In order to reveal the release mechanisms and explore the bioactivities of peptides, it is crucial to characterize the peptide composition of the food-derived protein hydrolysates.

Following with the rapid development of proteomics, peptidomics has also advanced significantly during the past decades (Dallas et al., 2015; Zhang et al., 2014). The present proteomics tools, such as MASCOT(Perkins et al., 1999), SEQUEST(Eng et al., 1994), MaxQuant (Cox & Mann, 2008), ProteoWizard(Kessner et al., 2008), PepNovo (Frank & Pevzner, 2005), pNovo3(Yang et al., 2019), Peaks(Ma et al., 2003), et al., provide efficient and mature techniques for reading, converting, and visualizing mass data, as well as the identification of peptides. This greatly facilitates the analysis of food-derived protein hydrolysates and ferments in peptidomic analysis. Nevertheless, despite the applicability in peptidomic analysis, limitations in identification of short peptides are evidently present for database-search proteomics tools like MASCOT, SEQUEST, etc. (Sayd et al., 2018) In contrast, de novo-sequencing methods (such as PepNovo, pNovo3, etc.) perform better and are usually suggested for use to bypass the limitations, but the performance and identification rate in short peptide identification still remains to be improved (Martini et al., 2021). To solve the problem, several stopgap semi-automated methods were established, through the construction of home built-in database of the preset peptides with a length range of 2–4 by Matlab and subsequent filtration of mass spectra by Compound Discoverer 3.0 (Cerrato et al., 2020; Piovesana et al., 2019). However, manual processes were still needed to differentiate peptide isomers with a low efficiency for the method. Besides, the

* Corresponding authors at: Room 323, Building 13 of SCUT, No.381 Wushan Rd., Tianhe District, Guangzhou 510640, China(Jianan Zhang).
E-mail addresses: fez.jianan@foxmail.com (J. Zhang), xujucai.happy@163.com (J. Xu).

https://doi.org/10.1016/j.fochx.2024.101249
Received 2 November 2023; Received in revised form 16 February 2024; Accepted 18 February 2024
Available online 20 February 2024

identification of short peptides with a length over 4 is not supported due to the intolerance of the extremely enlarged database. The high complexity of the methods and dependence on commercial software also bring much trouble to users. To fundamentally address the problem, an open-source, fully-automatic and efficient tool for peptidomic analysis of short peptides is necessitated to facilitate the exploration of food-derived peptides.

In this work, a tool named PeposX-Exhaust was developed and introduced to solve the challenge of identifying short peptides from food-derived protein hydrolysates. The accuracy of the tool was evaluated through the test on LC-MS/MS datasets acquired from known peptide mixtures, and the related comparison was also performed among the present tools. Besides, the tool was further applied in analyses of several common food-derived protein hydrolysates to illustrate the feasibility and superiority of the method in identifying short peptides from highly complex samples. We hope this work will help to facilitate the exploration of short peptides from no matter food-derived or biological samples.

2. Materials

2.1. Materials and chemicals

Glutathione (purity $\geq$ 95 %) was purchased from Macklin Biochemical Co., Ltd (Shanghai, China) and prepared at a concentration of 1 mg/ml for use to preliminarily validate the method. The peptide mixtures were prepared as reported in our previous work(Xu et al., 2019). soybean, walnut, collagen, and bonito protein hydrolysates were purchased from Guangdong Huapeptides Biotechnology Co., Ltd and dissolved in ultrapure water at a concentration of 2 mg/ml. LC-MS grade methanol, acetonitrile and formic acid were purchased from Macklin Biochemical Co., Ltd (Shanghai, China).

2.2. UHPLC-ESI-QTOF-MS/MS analysis of samples

The UHPLC-ESI-QTOF-MS/MS analysis of samples were carried out on a Nexera UHPLC system (Shimadzu, Kyoto, Japan) coupled to an electrospray ionization quadrupole time-of-flight mass spectrometry (AB Sciex X500R, Framingham, MA USA). The mobile phase consisted of A (acetonitrile) and B (water) with 0.1 % (v/v) formic acid, and the gradient elution procedure was set as follows: 0–4.00 min (5.0 % A isocratic), 4.00–6.00 min (5.0–10.0 % A), 6.00–30.00 min (10.0–40.0 % A), 30.00–34.00 min (40.0–90.0 % A), 34.00–40.00 min (90 % A isocratic), 40.00–42.00 min (90.0–5.0 % A), 42.00–52.00 min (5.0 % A isocratic). A HSS T3 column (1.8 μm 100 Å, Waters, USA) with the inner diameter narrowed at 1.0 × 100 mm was used for the separation. The injection volume was 1 μL. The flow rate was 0.05 mL/min, and the column temperature was set at 45 °C. The mass spectrometer was operated in positive mode with top four precursor ions automatically fragmented for MS/MS analysis. The mass detection range was set to 50–1200 m/z. For other parameters, the default values recommended by the workstation software SCIEX OS 2.0 (AB Sciex, Framingham, MA USA) were used.

2.3. Data format conversion

The raw mass spectrum data were converted to files in the format (*.mgf) using the freely accessible freeware ProteoWizard (version 3.0.22) (Kessner et al., 2008). For this process, the "peakPicking" was enabled with vendor msLevel = 1–2. The absolute signal threshold was set to 30. The option "Use zlib compression" was disabled.

2.4. Identification of short peptides

For peptide identification by PeposX-Exhaust, the peptide mass tolerance was 0.005 Da and the fragment ion tolerance was 0.02 Da. The length range of peptides was set 2–4 for glutathione and 2–6 for other complex samples, respectively. For comparison with other tools including MASCOT (Perkins et al., 1999), SEQUEST (Eng et al., 1994) packaged in Proteome Discoverer (Orsburn, 2021), PepNovo+ (Frank & Pevzner, 2005; Frank, 2009) packaged in DenovoGUI (Muth et al., 2014) and pNovo3 (Yang et al., 2019), the peptide mass and fragment ion tolerance were all set 0.005 Da and 0.02 Da, respectively. The protein database used for MASCOT and SEQUEST were downloaded from UniProt (The UniProt Consortium, 2023) in May 2023. More detailed parameters used for analysis of complex samples are shown in supplemental Table 1S.

3. Results and discussion

3.1. Overview

The schematic diagram of the tool "PeposX-Exhaust" for short peptide identification from protein hydrolysate is shown in Fig. 1A. From the figure, the identification procedure can be divided into three parts: (1) preparing: pre-setting relevant variables and processing LC-MS/MS data. (2) peptide identification and screening: enumerating peptides at the given length range with generation of the theoretical MS and MS/MS information, and comparing the theoretical information with experimental MS and MS/MS spectra, followed by screening peptide candidates. (3) scoring and ranking: scoring the peptide candidates and screening the identified peptides. The three parts contribute to the lightweight, effective, and non-missing search performance of the tool together.

3.2. Preparation

3.2.1. Pre-setting relevant variables

Reasonable setting of variables is important for the efficient identification of peptides. Fig. 1S (see Supporting Information) presents a detailed instruction for variable settings before using the tool. Notably, the tool supports specifying the length range of target peptides (usually 2–6 amino acid residues), product ion type, the signal intensity threshold of precursor ions and mass-to-charge ratio (m/z) tolerances, etc. In general, the signal intensity threshold of precursor ions is recommended to be set as more than 3 times larger of the instrument noise intensity.

3.2.2. Reading and pre-processing LC-MS/MS data

The LC-MS/MS data in format (*.mgf) are firstly loaded and read using a built-in module packaged in PeposX-Exhaust, and then the preprocessing of the data begins. This plays a key role in providing high-quality MS and MS/MS information for the further peptide identification. The corresponding procedure can be divided into four steps as follows (Fig. 1B):

(1) Noise abatement. For output data obtained from mass spectrometer, usually only about less than 10 % of the peaks are useful, while the others are usually physical noise generated by the instrument, or the isotopic peaks (also called isotopic noise) which may confuse the accurate results (Michalski et al., 2011). Thus, it is essential to abate the noise in mass data. In this work, noise abatement is performed through the filtration of signal intensities of the fragment ions, and those with an intensity lower than the given threshold are discarded (Cannataro et al., 2005).
(2) Multi-to-single charge conversion. To standardize the mass spectra information and simplify the process, all multi-charged ions including precursor ions and fragment ions are converted to single-charged ions by m/z calculation.
(3) Charge correction of precursor ions. There is a common phenomenon that molecules may aggregate together and appear as double-molecule ions ($2\ M + H^+$) when analytes are injected in

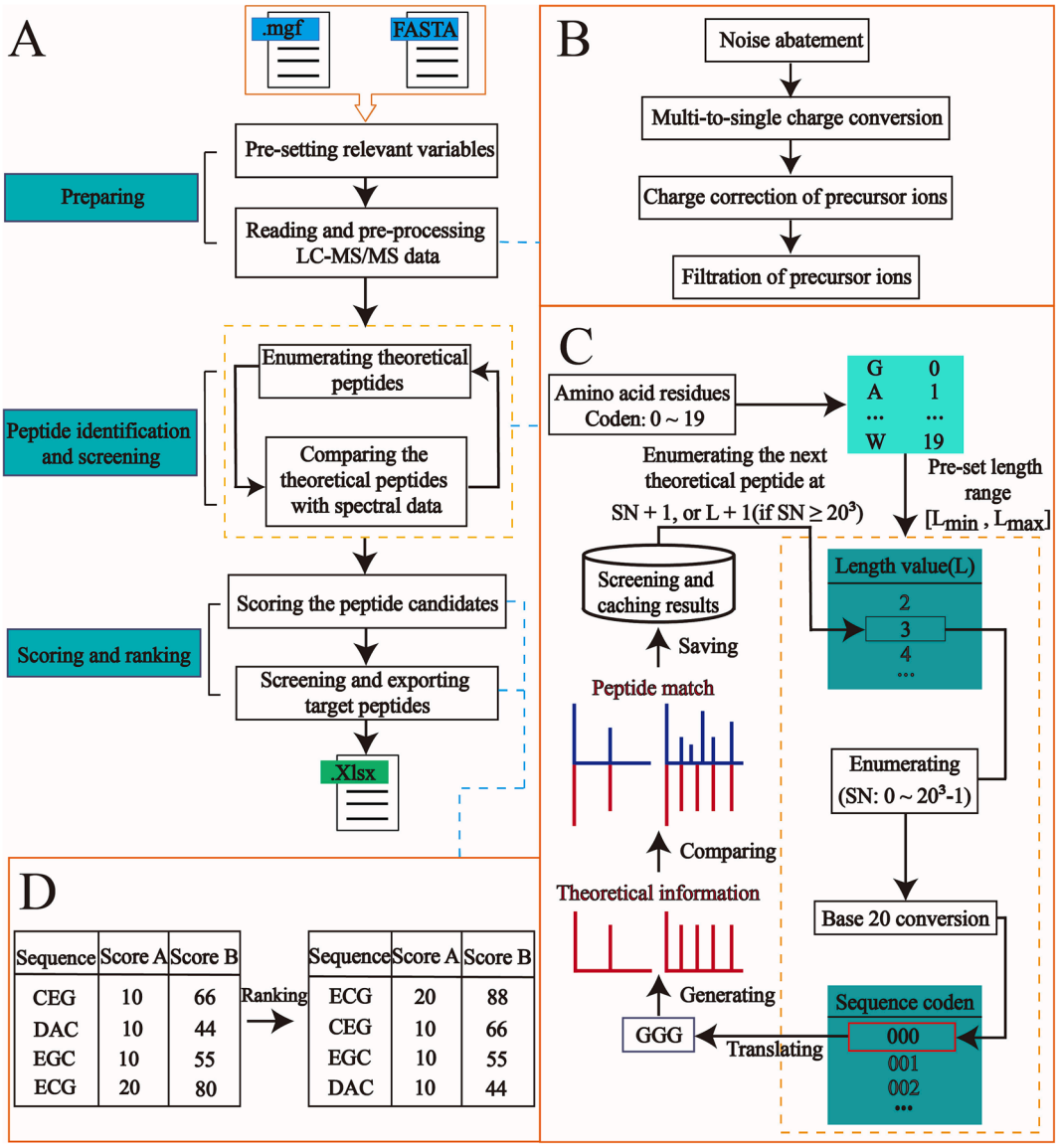


Fig. 1. (A) The schematic workflow of the tool for short peptide identification. (B) The schematic presentation of processing LC-MS/MS data. (C) The schematic diagram of peptide identification and screening process. (D) The schematic table of peptide scoring and ranking.

high concentration to mass spectrometry equipped with electrospray ionization source. In that case, considering the simultaneous existence of single-molecule ions ($M + H^+$), the single charged precursor ions may be mistakenly recognized as double charged ions in charge calculation by related-ion deconvolution. An example is presented in the supporting information Fig. 2S(A) for the phenomenon. From the figure, the single charged ion of glutathione was mistakenly marked as a double charged ion by the software DataAnalysis 4.0 (Bruker Daltonics GmbH, Germany). To address the above issue, a charge correction step has been added in the newly designed workflow of PeposX-Exhaust. Specifically, all precursor ions with multiple charges are screened to ensure that charge values displayed in their MS and MS/MS spectra match. If they do not, charge values from MS/MS spectrum are used. As shown in the Fig. 2S(B), the charge value (+1) was used for the precursor ion of glutathione. This approach can help to improve the accuracy of subsequent spectrum identification, which is one of the advantages of this tool over other software.

(4) Filtration of precursor ions. After charge check and correction, filtration is performed in two steps: (a) screening the compounds with precursor ion signal intensities higher than the pre-set threshold (marked as set S1); (b) retaining the compounds from S1 with precursor ion *m/z* values within the pre-set range (marked as set S2). In this way, valid MS and MS/MS data of the compounds are obtained for the subsequent identification.

3.3. Peptide identification and screening

In order to identify short peptides with efficiency and accuracy, an enumeration model based on the given peptide length range (the lower value, L_{min}; the upper value, L_{max}) and amino acid residues (see supporting information Table 2S) is proposed by the tool. As illustrated in the Fig. 1C, the specific steps for the procedure are as follows: (1) Coding the pre-defined amino acid residues with a total number of N (usually N = 20) as 0 ~ N-1; (2) Selecting a length value (initially, $L = L_{min}$) within the pre-set peptide length range; (3) Calculating the total number of possible theoretical peptides (N^L) and coding each peptide as sequence number from 0 to N^L-1; (4) Selecting an SN value (initially, SN = 0) and

Transferring the value to sequence codon through the base N conversion; (5) Translating the sequence codon to a peptide sequence; (6) Generating the theoretical MS and MS/MS information; (7) Comparing the theoretical information with experimental spectra contained in S2 and performing match analysis; (8) Screening and caching the matched results in memory; (9) Enumerating the next peptide at SN = SN + 1 (repeating step 4–8 until SN > N^L-1) or L = L + 1 (when SN > N^L-1, repeating step 2–8 until L > L_{max}).

During the procedure, the use of sequence number and its transformation to sequence codon is the most interesting part of the method, which is beneficial to theoretically enumerate all possible peptides at the given length range. For step 6, the theoretical MS information (mass-to-charge ratio, *m/z*) is calculated according to the formula 1, while the MS/MS information (the *m/z* lists of the product ions including a, b and y ions) is generated according to the formulas (2–4), respectively.

$$Calc.mz = M(H^+) + \sum_{j=1}^{L} M(A_j) + M(H_2O) \tag{1}$$

$$mz(a_k) = \frac{M(H^+) + \sum_{j=1}^{k} M(A_j) - M(CO)}{1} \tag{2}$$

$$mz(b_k) = \frac{M(H^+) + \sum_{j=1}^{k} M(A_j)}{1} \tag{3}$$

$$mz(y_k) = \frac{M(H^+) + \sum_{j=L+1-k}^{L} M(A_j) + M(H_2O)}{1} \tag{4}$$

where Calc.mz is the *m/z* value of the theoretical precursor ion; M(Ma et al.) is the molar mass of the j'th amino acid residue contained in the peptide; M(H^+) is the molar mass of the hydrogen ion; M(H_2O) is the molar mass of the water molecule; L is the number of composed amino acid residues of the peptide (also the length of the peptide), and mz(a_k), mz(b_k) and mz(y_k) represent the mz values of k'th a, b, and y fragment ion, respectively. M(CO) is the molar mass of CO (carbonyl).

For step 7, the *m/z* differences of precursor ions between the theoretical peptide and experimental compounds from S2 are firstly calculated, and the compounds with absolute difference values lower than the pre-defined MS tolerance threshold (E_1) are retained and marked as set S3. Then the MS/MS spectra of the compounds contained in S3 are further compared with the theoretical MS/MS information one by one, and the fragment ions of each compound with the absolute *m/z* difference values lower than the pre-defined MS/MS tolerance threshold (E_2) are marked and counted as F. Subsequently in step 8, the compounds with cover rate (C_r, calculated according to the formula 5) of matched fragments higher than the pre-defined threshold are screened and the theoretical peptide is cached as a potential candidate of these compounds. It should be noted that the maximal value of C_r is selected for the screening process for each compound while match analysis is performed simultaneously in a, b, and y ions.

$$C_r = \frac{F}{L} \tag{5}$$

The design of the above procedure is the core for this tool to identify short peptides as more as possible within the given length range. Particularly, this allows the short peptide identification without a source protein database, which can greatly facilitate the peptidomic analysis. Besides, the above designed procedure can commendably avoid excessive memory usage of loading any databases at once during the program execution. This also contributes a lot to developing the parallel processing capability of the tool through the distribution of different theoretical peptides to different processing cores, which can greatly accelerate the processing speed.

3.4. Scoring, ranking and selection of peptide candidates

After the primary identification and screening of peptides described above, a set of peptide candidates can be obtained, followed by the scoring of these candidates to associate the target peptides with the highest confidence. In this paper, the scoring method involves two indicators, and the combined utilization of the two indicators is also proved of high confidence to achieve satisfying results in the subsequent experiments.

The first indicator (S_A) is related to the kinds of fragment ion series generated from the peptide candidate that can be perfectly matched with the experimental spectra, like a, b and y ion series *etc*. In detail, when a kind of fragment ion series generated from the peptide candidate can be completely matched with the experimental MS/MS fragments, 10 points are scored to the candidate as S_A. For example, if there are 2 kinds of fragment ion series or clusters completely matched with the MS/MS spectra, 20 points are then awarded as $S_A = 20$.

The second indicator mainly describes the proportion of the matched product ions to the whole theoretical product ions. For calculation of the indicator, the number of matched product ion species is counted as N_M, and then the indicator can be obtained as follows:

$$S_B = \frac{N_M}{C \times L} \times 100 \tag{6}$$

where C is the number of types of the fragment ion series pre-defined by the user. Through the scoring of peptide candidates, only those with score S_A and S_B higher than the pre-defined threshold are reserved and saved to the results. Preferably before the exhibition and exportation of the results, peptide candidates of each spectrum are sorted according to their values of S_A, S_B, E_1 and E_2 at the priority order of $S_A > S_B > E_1 > E_2$, and generally, the first candidate is selected for each spectrum as the final results. With consideration of such a scoring and screening method, wherein the higher scores of S_A and S_B, the higher credibility of the identification results, and the smaller errors of E_1 and E_2, the higher confidence of the candidates. It's worth noting that all the peptide candidates satisfied with the pre-defined requirements are exported and presented in the final results, so that users can manually modify or select the candidates according to the sample features and users' experiences. This is very important for identification of short peptides, since there is usually more than one peptide candidate that can be perfectly matched with the experimental MS/MS spectra, which is also very helpful for researchers to know more about their samples. In addition, researchers are also allowed to load the protein database downloaded from UniProt (The UniProt Consortium, 2023) to locate the protein sequences associated with the peptide candidates (Fig. 3S), which can aid in further studies of targeted peptides.

3.5. Application and discussion

3.5.1. Analysis of the standard sample (glutathione) using PeposX-Exhaust

In this section, the standard sample (glutathione, GSH, γ-Glu-Cys-Gly) was first used for the test of PeposX-Exhaust. By referring to the methods previously described in section 2.2 and 2.3, the MS and MS/MS data of GSH were obtained and exported as file in *.mgf format. For further data process and peptide identification, the length range of target peptides was set as 2–4, and product ion series including a, b and y ions were all enabled. The MS and MS/MS tolerances were set as mentioned in section 2.4, and the results were exported as a *.csv file and presented in Table 1. From the table, there were 4 candidates given by the tool, and all of them were found with at least one kind of product ion series that could be matched with the experimental spectra. However, benefiting from the excellent scoring method and ranking mechanism, the correct peptide sequence (ECG) was listed in the first row of the table. Besides, the identification process tool only 32 s to achieve analysis and the size of the packaged algorithm was only 25.6 MB. This

Table 1
The identification results of the compound spectra of standard sample (glutathione) by PeposX-Exhaust.

Spectrum title	Rank	Peptide	Measured *m/z*	Theoretical *m/z*	Score A (S_A)	Score B (S_B)	E_1 (Da)	E_2 (Da)
guguanggantai2.401.401.1 File:"", NativeID:"sample = 1 period = 1 cycle = 316 experiment = 2"	0	ECG	308.0906	308.0911	20	88	0.0005	0.0012
	1	CEG	308.0906	308.0911	10	66	0.0005	0.0029
	2	EGC	308.0906	308.0911	10	55	0.0005	0.0028
	3	DAC	308.0906	308.0911	10	44	0.0005	0.0105

Note: *m/z*, the mass-to-charge ratio of the ions; S_A and S_B, score indicators calculated as described in section 3.4, respectively; E_1 and E_2, the MS and MS/MS tolerances, respectively.

fully exhibited the high efficiency, outstanding accuracy and lightweight size of PeposX-Exhaust. It should also be noted that MS/MS spectra of short peptides might be usually presented with many candidates with complete product ion series matched possibly due to the high abundance in food-derived protein hydrolysates. To solve the problem, a strict scoring and screening method as described in this work was one of the effective means. The other feasible way was to evaluate the existence probability of the candidate peptide in samples, which would be included in our future work. In conclusion, the performance of PeposX-Exhaust was remarkable due to its success in identifying the standards, and more interesting experiments were deserved in the following sections.

3.5.2. Performance of the tool in identification of short peptides from the known peptide mixtures and its comparison with other exiting tools

For further validation of the tool (PeposX -Exhaust), it was subsequently tested in the identification of short peptides from the known peptide mixtures, and a comparison with other tools, including MASCOT, SEQUEST, PepNovo + and pNovo3, was also performed in this section. One should note that not all standard peptides could be successfully monitored in MS and MS/MS detection due to the limitations in separating and detection methods, and interactions occurring between each other of the peptides might also affect the mass detection. This might interfere the objective evaluation of the accuracy rate by different tools. Hence, the experimental spectra were filtrated through the manual annotations (Biotools 3.2, Bruker Daltonics, Germany) and only those with at least one kind of fragment ion series fully matched were kept. Besides, only one spectrum with the best manual annotation effect was retained for each known peptide. As shown in the supporting information Table 3S, a total of 69 experimental spectra were finally retained and listed in a *.mgf file. It was then subsequently submitted to the peptide identification tools.

(1) Performance of the tool in identification of short peptides from the known peptide mixtures.

The identification results by PeposX-Exhaust for the peptide mixtures are presented in Fig. 2. The sub-figure A shows the *m/z* differences between the experimental precursor ions and the theoretical precursor ions of the first peptide candidate for each spectrum. Most identified peptides were found with a *m/z* difference between −0.002 and 0.002, suggesting a relatively high accuracy by the tool. Fig. 2B shows the charge distribution of different precursor ions, and indicates a tendency that short peptides may tend to be single-charged during the electrospray ionization. This was somewhat different from those for long peptides, which were usually multi-charged in electrospray ionization(Tu et al., 2018; Wang et al., 2018). It should be noted that no charge number error was detected in the experiment, so no charge correction was performed as

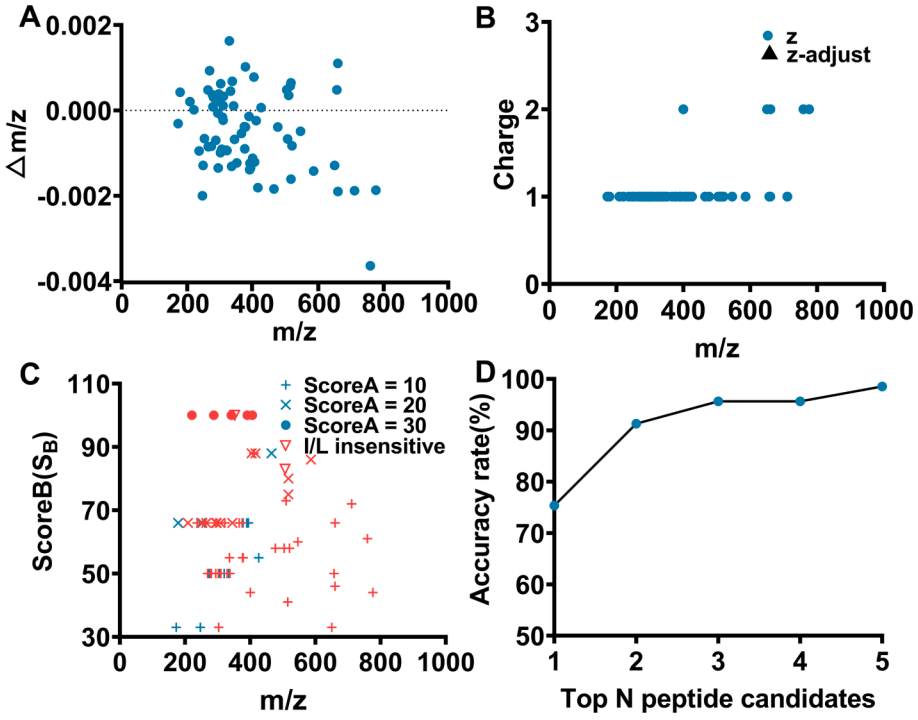


Fig. 2. The identification results by PeposX-Exhaust for the peptide mixtures: (A) The mass-to-charge ratio (*m/z*) distribution of the precursor ions of the identified peptides; (B) The charge distribution of the precursor ions (no charge correction performed in the analysis); (C) The score distribution of the identified peptides with correct identifications marked in red; (D) The change in correct identifications by PeposX-Exhaust when considering the top 1–5 candidates. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

displayed in the sub-figure B.

Fig. 2C shows the score distributions of the identified peptides at different precursor ion masses with correct ones marked in red. From the figure, most of the compounds could be identified to peptides with high accuracy through the results screening at $S_A \geq 10$. When attention was paid to the identified peptides with $S_A \geq 20$, the erroneous identifications could be evidently deceased though accompanied with an evident decrease in number of identified peptides. Hence, the value of 10 was usually recommended for use of S_A in most cases, while the value of 20 was recommended in cases when higher accuracy was demanded. The value of S_B was an associated indicator mainly used for the ranking of peptide candidates for each compound spectrum and screening of the results. From the figure, the results screening at $S_B > 50$ could also help to decrease mistakes. This fully demonstrated the importance and feasibility of the scoring method. It was interesting that errors were rarely observed for compounds with precursor ion mass high than 400, possibly due to the increasing difficulty for MS/MS spectra to be perfectly matched with a kind of product ion series. One should also note that the algorithm developed in this work could not yet differentiate residues (leucine and isoleucine) contained in peptides as shown in the Fig. 2C, and their impact on identification results was ignored as many previous reports (Cox et al., 2011).

The change in correct identifications is presented in Fig. 2D when top 1–5 candidates were considered. From the figure, 52 of the 69 spectra (accuracy rate = 75.36 %) were exactly identified when with only top 1 candidate (ranked 0) considered. While with top 2 candidates considered, the number of correctly identified spectra was evidently increased and reached as high as 63 (approximately, the accuracy rate = 91.3 %). What's more, more correctly identified spectra and higher accuracy rate were obtained when with more candidates considered as shown in the figure. This fully demonstrated the feasibility and validity of the newly developed tool.

(2) The performance comparison of the PeposX-Exhaust with other tools in identifying short peptides from the known peptide mixtures.

The short peptide identification results by different tools are displayed in Fig. 3. From the Fig. 3A and B, the newly developed PeposX-Exhaust successfully identified all the submitted spectra (identification rate 100 %), and the lengths of the identified peptides were basically consistent with the truth of the known peptides, exhibiting the excellent ability in discovering short peptide. In contrast, the number of identified spectra was only 19 (identification rate 27.54 %) and 2 (identification rate 2.90 %) for SEQUEST and MASCOT, respectively. Besides, the peptides identified by MASCOT and SEQUEST were presented with a length not less than 6 and 5, respectively, possibly due to the algorithm design. The main purpose of SEQUEST and MASCOT was to analyze proteins contained in samples, and hence they were designed more dedicated in identifying long or specific peptides. This made it unsuitable for these tools to identify short peptides of low or poor specificity (Martini et al., 2021). The performance by de novo-sequencing tools (PepNovo + and pNovo3) was relatively better than MASCOT and SEQUEST. From the figure, the numbers of spectra identified by PepNovo + and pNovo3 were 61 (identification rate 88.41 %) and 49 (identification rate 71.14 %), respectively. The performance difference between the two tools mainly lies in the identification of dipeptides, with 15 more dipeptides identified by PepNovo +. This well explained the applicability of de novo-sequencing tools in short peptide identification, although the performance could be further improved.

For further validation and comparison of these tools, the accuracy rate with top 1 candidate considered and the adjusted accuracy rate with top 5 candidates considered were surveyed and displayed in the Fig. 3C. From the figure, the highest accuracy rate (75.36 %) and adjusted accuracy rate (98.55 %) were both obtained by PeposX-Exhaust, respectively. Notably, the accuracy rate was at least 70 % higher than the two database-search tools (MASCOT and SEQUEST), and 15 % higher than the other two de novo-sequencing tools (PepNovo + and pNovo3). The accuracy rate cumulative curves of the five tools were also figured out in

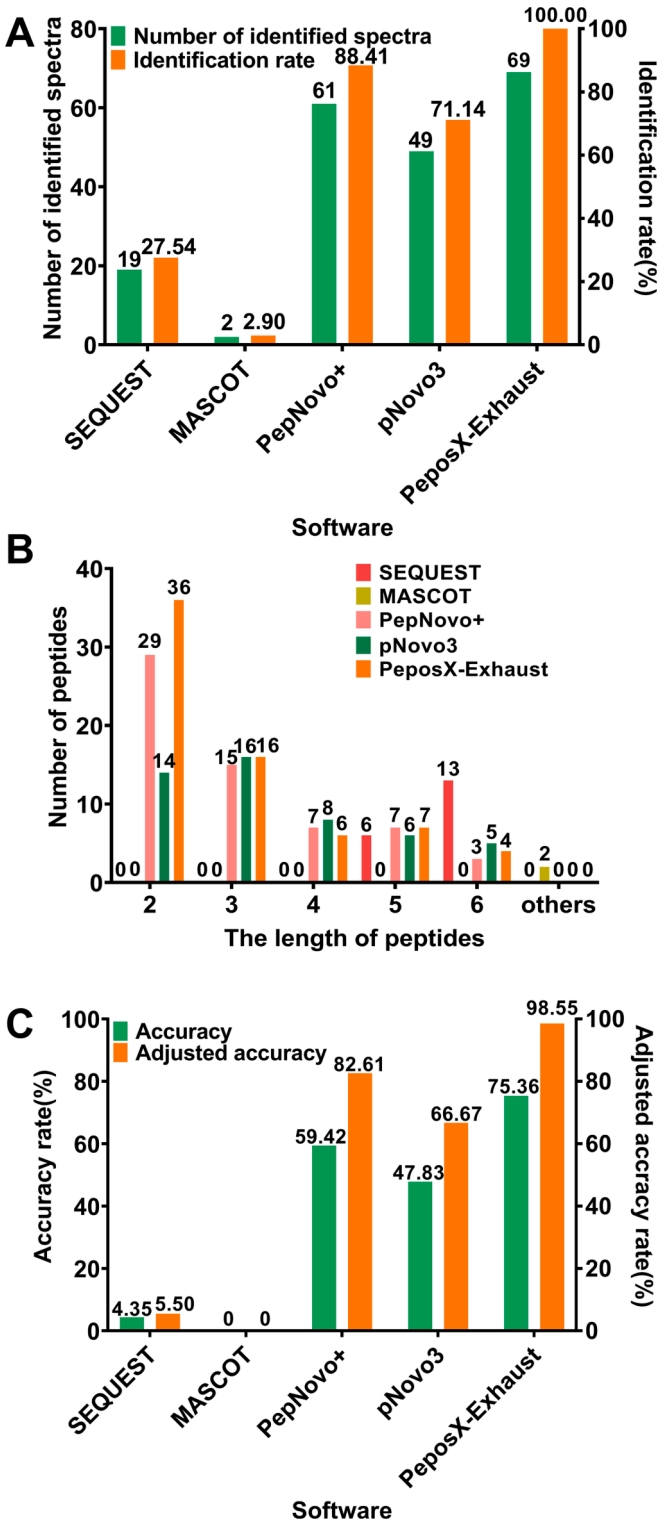


Fig. 3. The performance and comparison of different tools in identifying short peptides from the known peptide mixture: (A) the number of spectra identified and the corresponding identification rate, which represents the percentage of the number of identified peptides to the total number of submitted spectra; (B) the length distribution of the identified peptides; (C) the accuracy rate and adjusted accuracy rate. The accuracy rate represents the percentage of MS/MS spectra that were correctly identified with the use of the top 1 peptide candidate, while the adjusted accuracy rate represents the percentage of MS/MS spectra that could be correctly identified with the use of top 5 peptide candidates.

Fig. 5S, and evident improvements in identification performance were also observed by PeposX-Exhaust. This fully demonstrated the superiority, validity, and feasibility of the newly developed algorithm, which could greatly facilitate the identification of short peptides from complex samples.

3.5.3. *Application and comparison of the algorithms in identification of short peptides from common food-derived protein hydrolysates*

Four data sets acquired from the LC-MS/MS analyses of complex samples including the common soybean, walnut, collagen and bonito protein hydrolysates, consisting of 8619, 7777, 7237 and 7335 experimental MS/MS spectra, were used for the further validation and comparison of the PeposX-Exhaust with other existing tools in this section.

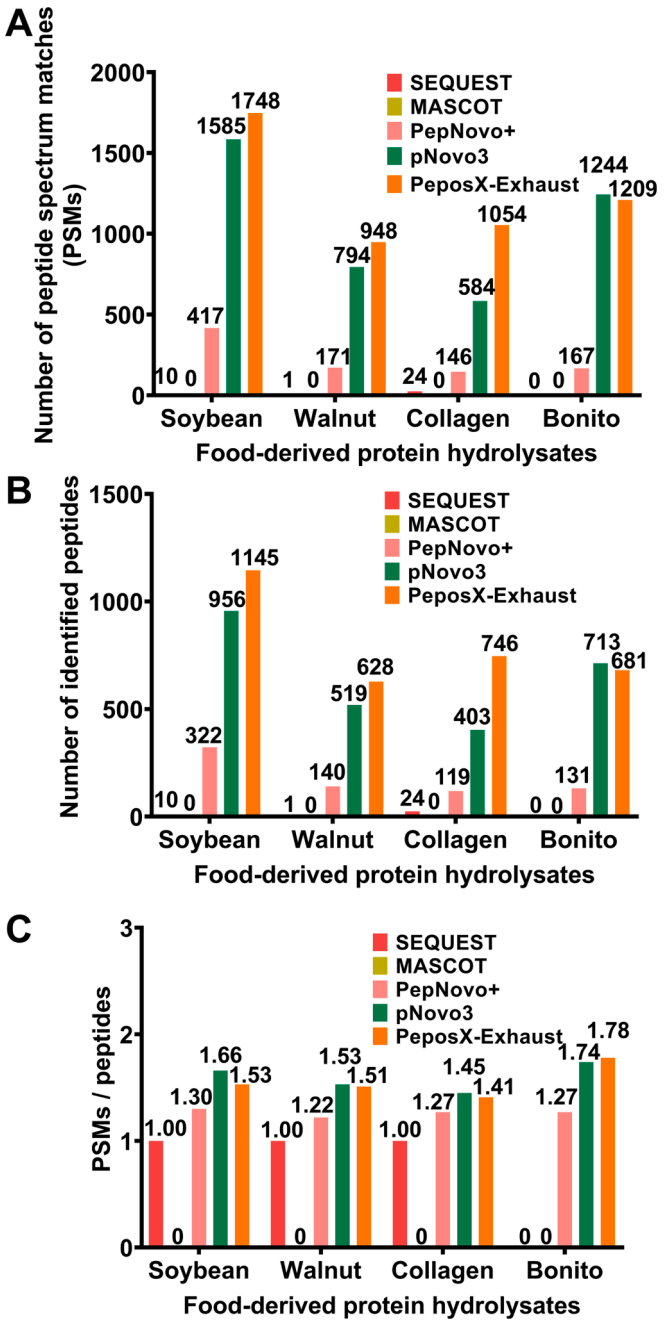


Fig. 4. The performance of different tools in identifying short peptides from the common food-derived protein hydrolysates: (A) the number of peptide spectra matches (PSMs);(B) the number of identified short peptides; (C) the ratio of PSMs to peptides.

As shown in the Fig. 4, the performance by PeposX-Exhaust in identifying short peptides was evidently better than that by the other four tools. From the Fig. 4A and B, hundreds or even thousands of the spectra and short peptides were identified by PeposX-Exhaust, respectively. Besides, the ratio of peptide spectrum matches (PSMs) to short peptides for different samples was between 1.40 and 1.80 (Fig. 4C), suggesting a relatively high credibility of the results by PepOS-Exhaust and a strong capability of the algorithm to explore short peptides from complex samples. In comparison, the number of spectra and short peptides identified by pNovo3 were slightly lower than PeposX-Exhaust, while PepNvo + performed less well. One should note that only a few or even no spectra and short peptides were identified by SEQUEST and MASCOT, respectively, showing a powerlessness and unsuitability of the two tools to identify short peptide from the complex food-derived protein hydrolysates. Frankly, SEQUEST and MASCOT were much better in identification of specific long peptides rather than non-specific short peptides. Their outstanding superiorities and advantages in proteomic analysis have been widely reported and known by many researchers, offering the guidance to developments of peptidomics.

As mentioned above, there were usually a lot of short peptides contained in food-derived protein hydrolysates resulting from the use of crude proteases of general cutting sites. Fig. 5 shows the length distribution of short peptides identified from the four data sets by PepNovo+, pNovo3 and PeposX-Exhaust. From the figure, the lengths of peptides identified by different tools were mainly concentrated in the range of 2–5 as expected, indicating a necessity and worth to efficiently identify short peptides in order to further explore bioactive peptides. Compared with the results by PepNovo + in the Fig. 5A, evidently higher number of identified peptides across different length values were obtained by pNovo3 and PeposX-Exhaust. Interestingly, similar variation trend in the number of identified peptides was roughly observed between the results by pNovo3 (Fig. 5B) and PeposX-Exhaust (Fig. 5C). The main difference between the two tools still lied in the performance in identifying dipeptides as mentioned before in the section 3.5.2. Besides, the numbers of peptides with a length range of 5–6 identified by PeposX-Exhaust were also a slightly higher than that by pNovo3, possibly due to the differences in the algorithm method design. It should be noted that, there were quite a few peptides identified with a length between 5 and 6 from the collagen hydrolysates, illustrating a possible need of further hydrolysis of the product for better digestibility. This fully illustrated the utility and feasibility of PeposX-Exhaust in identifying short peptides from complex food-derived samples.

One should also note that the algorithm was mainly tested and validated in short peptides with a length range of 2–6, but the identification of peptides with a length longer than 6 should be also achievable. Nevertheless, it would be very time consuming. To speed the processing procedure, multiprocessing technology was also equipped for the algorithm. Users could freely control the number of processor cores at the beginning of the processing. Generally, the time would be 20 times longer than before when the target peptide length was additionally extended with one residue. Therefore, in most cases, the algorithm was mainly recommended for identifying peptides with lengths no more than 6, and target peptides with lengths over 6 were still advised to be processed using the existing tools including MASCOT, SEQUEST, PepNovo+, etc.

4. Conclusion

In this study, a short peptide identification tool named as PeposX-Exhaust was developed. Through the test and validation on known peptide mixtures, the algorithm was proved of high accuracy, feasibility and efficiency for identification of short peptides. The accuracy rate and adjusted accuracy rate for the peptide mixtures reached as high as 75.36 % and 98.55 %, respectively, which was much better than those by MASCOT, SEQUEST, PepNovo + andpNovo3, respectively. For further application in analysis of complex food-derived protein hydrolysates,

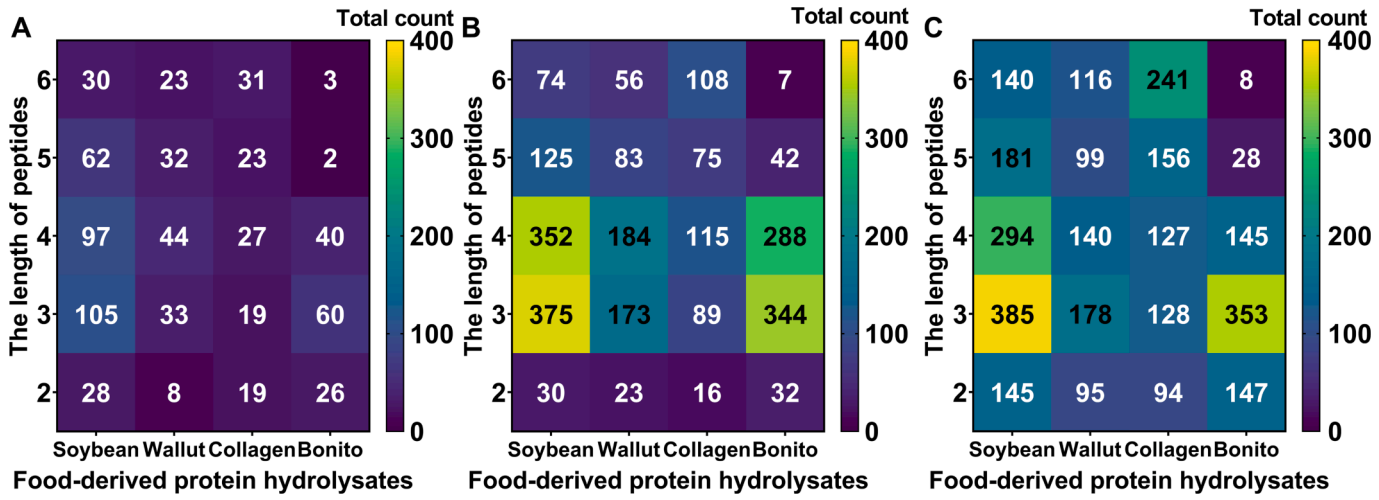


Fig. 5. The length distributions of the identified short peptides from common food-derived protein hydrolysates by PepNovo+, pNovo3 and PeposX-Exhaust.

relatively good performance was also achieved by the algorithm with hundreds of short peptides identified. In general, PeposX-Exhaust was highly recommended for use to identify short peptides with a length no longer than 6. It provides a lightweight, feasible and efficient way for such analyses, and will greatly facilitate the exploration of food-derived peptides. The future work will be directed towards the enhancement of such an algorithm to improve the accuracy at top 1 candidate and overcome the drawbacks in long peptide identification, so that the peptide composition of the food-derived protein hydrolysates can be easily and efficiently revealed. One should note that, although the algorithm was only tested in food-derived protein hydrolysates, its extension to biological samples should at most need some minor revision. The packaged algorithm and parameter settings are available at www.peposx.com.

CRediT authorship contribution statement

Wanshun Liu: Conceptualization, Writing – original draft. **Mouming Zhao:** Writing – review & editing. **Lishe Gan:** Data curation, Software. **Baoguo Sun:** Formal analysis, Visualization. **Shiqi He:** Investigation, Resources. **Yang Liu:** Supervision, Validation. **Lei Liu:** Validation. **Wu Li:** Funding acquisition. **Jing Chen:** Methodology. **Jianan Zhang:** Funding acquisition. **Jucai Xu:** Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

The authors gratefully acknowledge the GuangDong Basic and Applied Basic Research Foundation (No. 2022A1515110711 and 2022A1515011543), the Science and Technology Planning Project of Guangdong Province (2021B1212040016), the fellowship of China Postdoctoral Science Foundation (No. 2021 M691068) for their financial supports and the Jiangmen Science and Technology Plan Project (no. 2220002000277 and 2020JC01030).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.fochx.2024.101249.

References

Agyei, D., Tsopmo, A., & Udenigwe, C. C. (2018). Bioinformatics and peptidomics approaches to the discovery and analysis of food-derived bioactive peptides. *Analytical and Bioanalytical Chemistry, 410*(15), 3463–3472.

Cannataro, M., Guzzi, P., Mazza, T., & Veltri, P. (2005). Preprocessing, management, and analysis of mass spectrometry proteomics data. *In Workflows Management: New Abilities for the Biological Information Overflow, the Network Tools and Applications in Biology (NETTAB) workshop*.

Cerrato, A., Aita, S. E., Capriotti, A. L., Cavaliere, C., Montone, C. M., Laganà, A., & Piovesana, S. (2020). A new opening for the tricky untargeted investigation of natural and modified short peptides. *Talanta, 219*, Article 121262.

Cox, J., & Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology, 26*(12), 1367–1372.

Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., & Mann, M. (2011). Andromeda: A peptide search engine integrated into the MaxQuant environment. *Journal of Proteome Research, 10*(4), 1794–1805.

Dallas, D. C., Guerrero, A., Parker, E. A., Robinson, R. C., Gan, J., German, J. B., Barile, D., & Lebrilla, C. B. (2015). Current peptidomics: Applications, purification, identification, quantification, and functional analysis. *Proteomics, 15*(5–6), 1026–1038.

Eng, J. K., McCormack, A. L., & Yates, J. R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry, 5*(11), 976–989.

Frank, A., & Pevzner, P. (2005). PepNovo: De novo peptide sequencing via probabilistic network modeling. *Analytical Chemistry, 77*(4), 964–973.

Frank, A. M. (2009). A ranking-based scoring function for peptide-spectrum matches. *Journal of Proteome Research, 8*(5), 2241–2252.

Kaur, A., Kehinde, B. A., Sharma, P., Sharma, D., & Kaur, S. (2021). Recently isolated food-derived antihypertensive hydrolysates and peptides: A review. *Food Chemistry, 346*, Article 128719.

Kessner, D., Chambers, M., Burke, R., Agus, D., & Mallick, P. (2008). ProteoWizard: Open source software for rapid proteomics tools development. *Bioinformatics, 24*(21), 2534–2536.

Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A., & Lajoie, G. (2003). PEAKS: Powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry, 17*(20), 2337–2342.

Martini, S., Solieri, L., & Tagliazucchi, D. (2021). Peptidomics: New trends in food science. *Current Opinion in Food Science, 39*, 51–59.

Michalski, A., Cox, J., & Mann, M. (2011). More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent lc-ms/ms. *Journal of Proteome Research, 10*, 1785–1793.

Muth, T., Weilnböck, L., Rapp, E., Huber, C. G., Martens, L., Vaudel, M., & Barsnes, H. (2014). DeNovoGUI: An open source graphical user interface for de novo sequencing of tandem mass spectra. *Journal of Proteome Research, 13*(2), 1143–1146.

Nasri, M. (2017). Chapter four - protein hydrolysates and biopeptides: Production, biological activities, and applications in foods and health benefits. a review. In F. Toldrá (Ed.), *Advances in Food and Nutrition Research* (Vol. 81, pp. 109–159). Academic Press.

Orsburn, B. C. (2021). Proteome discoverer-a community enhanced data processing suite for protein informatics. *Proteomes, 9*(1), 15.

Perkins, D. N., Pappin, D. J., Creasy, D. M., & Cottrell, J. S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis, 20*(18), 3551–3567.

Piovesana, S., Capriotti, A. L., Cerrato, A., Crescenzi, C., La Barbera, G., Lagana, A., Montone, C. M., & Cavaliere, C. (2019). Graphitized carbon black enrichment and uhplc-ms/ms allow to meet the challenge of small chain peptidomics in urine. *Analytical Chemistry, 91*(17), 11474–11481.

Priya, S. (2019). Therapeutic perspectives of food bioactive peptides: A mini review. *Protein and Peptide Letters, 26*(9), 664–675.

Sayd, T., Dufour, C., Chambon, C., Buffière, C., Remond, D., & Santé-Lhoutellier, V. (2018). Combined in vivo and in silico approaches for predicting the release of bioactive peptides from meat digestion. *Food Chemistry, 249*, 111–118.

Suleria, H. A., Osborne, S., Masci, P., & Gobe, G. (2015). Marine-Based Nutraceuticals: An Innovative Trend in the Food and Supplement Industries. *Marine Drugs, 13*(10), 6336–6351.

The UniProt Consortium. (2023). UniProt: The universal protein knowledgebase in 2023. *Nucleic Acids Research, 51*(1), 523–531.

Tu, M., Wang, C., Chen, C., Zhang, R., Liu, H., Lu, W., Jiang, L., & Du, M. (2018). Identification of a novel ace-inhibitory peptide from casein and evaluation of the inhibitory mechanisms. *Food Chemistry, 256*, 98–104.

Wang, C., Tu, M., Wu, D., Chen, H., Chen, C., Wang, Z., & Jiang, L. (2018). Identification of an ace-inhibitory peptide from walnut protein and its evaluation of the inhibitory mechanism. *International Journal of Molecular Sciences, 19*(4), 1156.

Xu, J., Zheng, L., Su, G., Sun, B., & Zhao, M. (2019). An improved peak clustering algorithm for comprehensive two-dimensional liquid chromatography data analysis. *Journal of Chromatography. A, 1602*, 273–283.

Yang, H., Chi, H., Zeng, W. F., Zhou, W. J., & He, S. M. (2019). pNovo 3: Precise de novo peptide sequencing using a learning-to-rank framework. *Bioinformatics, 35*(14), i183–i190.

Zhang, Z., Wu, S., Stenoien, D. L., & Paša-Tolić, L. (2014). High-throughput proteomics. *Annual Review of Analytical Chemistry, 7*, 427–454.