PEP Search in MyCompoundID: Detection and Identification of Dipeptides and Tripeptides Using Dimethyl Labeling and Hydrophilic Interaction Liquid Chromatography Tandem Mass Spectro...

ARTICLE *in* ANALYTICAL CHEMISTRY · MARCH 2014

Impact Factor: 5.64 · DOI: 10.1021/ac500109y · Source: PubMed

4 AUTHORS, INCLUDING:

Yanan Tang
University of Alberta
17 PUBLICATIONS **27** CITATIONS
SEE PROFILE

Ronghong Li
University of Alberta
3 PUBLICATIONS **34** CITATIONS
SEE PROFILE

analytical chemistry

Article

pubs.acs.org/ac

PEP Search in MyCompoundID: Detection and Identification of Dipeptides and Tripeptides Using Dimethyl Labeling and Hydrophilic Interaction Liquid Chromatography Tandem Mass Spectrometry

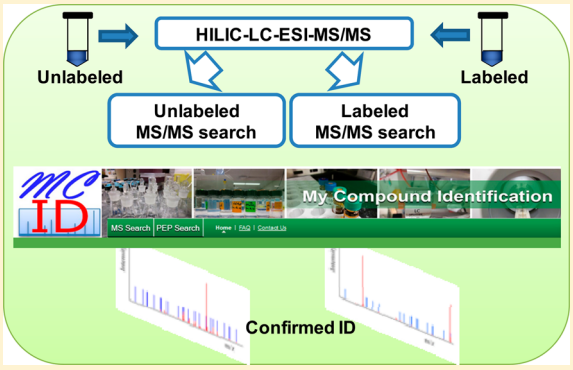
Yanan Tang,[†] Ronghong Li,[‡] Guohui Lin,[‡] and Liang Li*[,†]

[†]Department of Chemistry, University of Alberta, 11227 Saskatchewan Drive, Edmonton, Alberta, T6G 2G2 Canada

[‡]Department of Computing Science, University of Alberta, 2-21 Athabasca Hall, Edmonton, Alberta, T6G 2E8 Canada

Supporting Information

ABSTRACT: Small peptides, such as dipeptides and tripeptides, are naturally present in many biological samples (e.g., human biofluids and cell extracts). They have attracted great attention in many research fields because of their important biological functions as well as potential roles as disease biomarkers. Tandem mass spectrometry (MS/MS) can be used to profile these small peptides. However, the type and number of fragment ions generated in MS/MS are often limited for unambiguous identification. Herein we report a novel database-search strategy based on the use of MS/MS spectra of both unlabeled and dimethyl labeled peptides to identify and confirm amino acid sequences of di/tripeptides that are separated using hydrophilic interaction (HILIC) liquid chromatography (LC). To facilitate the di/tripeptide identification, a database consisting of all the predicted MS/MS spectra from 400 dipeptides and 8000 tripeptides was created, and a search tool, PEP Search, was developed and housed at the MyCompoundID website (www.mycompoundid.org/PEP). To evaluate the identification specificity of this method, we used acid hydrolysis to degrade a standard protein, cytochrome c, to produce many di/tripeptides with known sequences for LC/MS/MS. The resultant MS/MS spectra were searched against the database to generate a list of matches which were compared to the known sequences. We correctly identified the di/tripeptides in the protein hydrolysate. We then applied this method to detect and identify di/tripeptides naturally present in human urine samples with high confidence. We envisage the use of this method as a complementary tool to various LC/MS techniques currently available for small molecule or metabolome profiling with an added benefit of covering all di/tripeptide chemical space.



Small peptides have recently attracted great attention in various research fields including pharmacology, molecular biology, physiology, and other areas, as a series of biologically active dipeptides and tripeptides have been found to play important roles in neuron signal transmission,[1,2] intercellular signal transmission,[3] blood coagulation,[4] and metabolic processes.[5] Some di/tripeptides are involved in disease developing process, and their presence could indicate specific diseases, i.e., serve as disease biomarkers.[6−9]

Many efforts have been put into the identification and quantification of di/tripeptides or other small peptides in biological or food samples.[7,10−18] Mass spectrometry (MS) coupled to liquid chromatography (LC) is one of the most commonly applied technologies in di/tripeptides analysis. Identification of di/tripeptides from biological samples using MS is mainly done by *de novo* sequencing[15,19] or matching the accurate mass and retention time to those of standards.[7,20] However, because of the sequence diversity of di/tripeptides and chemical complexity of biological samples, *de novo* sequencing or using di/tripeptide standards for peptide identification can be difficult and time-consuming, particularly for untargeted analysis, such as in metabolome profiling work. While database search strategy using MS/MS spectra of peptide ions is commonly used for peptide identification,[21] current search algorithms are often limited to peptides containing 4 or more amino acids. Di/tripeptides usually do not produce a sufficient number of unique fragment ions for unambiguous sequence identification. Moreover, they do not retain on reversed phase (RP) LC column well and thus elute at the void volume without much separation in a conventional LC/MS/MS workflow used for peptide identification. Another way of identifying di/tripeptides is to use an experimental MS/MS spectral library of standard peptides. However, due to limited availability of peptide standards, current spectral libraries, such as NIST MS/MS database (http://peptide.nist.gov/),[22] METLIN (http://metlin.scripps.edu),[23] and MassBank (http://www.massbank.jp/?lang=en),[24] do not contain many

Received: January 9, 2014
Accepted: February 26, 2014
Published: February 26, 2014

di/tripeptide MS/MS spectra. In addition, as it will be shown in this paper, the MS/MS spectra of closely related di/tripeptides are very similar, and thus, in these cases, it is difficult to use spectral match for unambiguous identification.

In this work, we report a strategy to detect and identify di/tripeptides in complex biological samples with high confidence and speed. In this method, di/tripeptides are separated using hydrophilic interaction (HILIC) LC and detected by MS and MS/MS. The acquired MS/MS spectra are searched against a database consisting of predicted MS/MS spectra of all the di/tripeptides (8400 in total). However, the search results often contain many false matches. To increase the specificity of the search, the same sample is treated with dimethyl reaction via reductive amidation[18,25,26] to label the N-terminal of the peptides and ϵ-amine of lysine, followed by running the labeled sample using the same HILIC-LC/MS/MS setup. The MS/MS spectra of the labeled peptides display a prominent a_1 ion peak along with a few other fragment ion peaks that can be used to narrow down the list of matches from the unlabeled peptides to arrive at a unique match. In this work, we used a number of di/tripeptide standards for method development and performance evaluation. We then applied this method for analyzing di/tripeptides in a protein hydrolysate generated using microwave-assisted acid hydrolysis to assess its performance and general utility for detecting a diverse array of di/tripeptides. Finally, we applied this method to identify di/tripeptides in human urine to illustrate its applicability to handle real world samples.

EXPERIMENTAL SECTION

Chemicals and Reagents. All chemicals and reagents, except those specifically noted, were purchased from Sigma-Aldrich Canada (Markham, ON, Canada). LC/MS grade water and acetonitrile (ACN) were purchased from Thermo Fisher Scientific (Edmonton, AB, Canada).

Microwave-Assisted Acid Hydrolysis of Proteins. A standard protein, cytochrome c [Bovine], was hydrolyzed by microwave-assisted acid hydrolysis (MAAH) to produce di/tripeptides. A 100 μg portion of protein was dissolved in deionized water to 1 mg/mL and mixed with an equal volume of 6 M HCl.[27,28] The MAAH time was optimized to generate di/tripeptides that would cover a large portion of the protein sequence. After 30 min of MAAH, excess HCl was evaporated by vacuum concentrator. The protein hydrolysate was redissolved in 70% ACN and 30% MeOH to 0.5 μg/μL for analysis.

Dimethyl Labeling of Peptides. There were 38 di/tripeptide standards purchased from Sigma-Aldrich. These standards and the peptides generated from the hydrolysis of cytochrome c were dimethyl labeled on their N-terminal and ϵ-amino group of lysine residues as reported[26] with some modification. In this work, 50 μL of the cytochrome c hydrolysate (0.5 μg/μL) or peptide standard solution in 70% ACN and 30% MeOH were mixed with 5 μL of 1 M sodium acetate buffer (pH 5.0). After vortexing, 4 μL of freshly prepared 1 M sodium cyanoborohydride solution was added to each peptide solution, followed by the addition of 1 μL of 4% formaldehyde (in water). The solution was vortexed and incubated at 37 °C for 30 min.

Preparation of Human Urine. Human urine was collected from an healthy individual with ethics approval from the University of Alberta. A 100 μL portion of urine was mixed with 300 μL of precooled MeOH (0 °C) and incubated on ice for 15 min to precipitate the proteins. After incubation, supernatant was collected by centrifuging at 12 000 × g at 4 °C for 15 min. After evaporating the solvents in the supernatant, the urine sample was reconstituted in 100 μL of 70% ACN and 30% MeOH. A 50 μL portion of reconstituted urine sample was dimethyl labeled with formaldehyde as described above.

LC/MS/MS. Before LC/MS/MS analysis, all samples were acidified with 1% formic acid to pH 2. Samples were analyzed using a Bruker MAXIS IMPACT ESI quadrupole time-of-flight (QTOF) mass spectrometer (Bruker, Billerica, MA) linked to an Agilent 1100 series binary HPLC system (Agilent, Palo Alto, CA). They were separated on a hydrophilic interaction (HILIC) TSKgel Amide-80 column (1 mm × 250 mm, 5 μm particle size, 100 Å pore size; Tosoh Bioscience LLC, King of Prussia, PA). Solvent A was 10 mM ammonium formate in 95% (v/v) acetonitrile (pH 4.5), and solvent B was 10 mM ammonium formate in water (pH 4.5). Peptides were separated using a 70-min solvent gradient: 0−10 min, 0−10% solvent B; 10−45 min, 10−30%; 45−55 min, 30−90%; 55−60 min, 90%; 60−60.01 min, 90−0%; 60.01−70 min, 100% solvent A. The flow from HILIC-LC was directed to the ESI source at a flow rate of 100 μL/min. A survey MS scan was acquired at m/z 100−700 for 0.5 s, followed by 4 data-dependent MS/MS scan at m/z 20−700 for 0.5 s each. All mass spectra were collected in positive ion mode.

For the initial method development work, RPLC was examined as a possible tool for separating di/tripeptides. The RPLC column tested was ACQUITY BEH C18 (2.1 mm × 50 mm, 1.7 μm, 130 Å pore size; Waters). Solvent A was 0.1% formic acid in water, and solvent B was 0.1% formic acid in acetonitrile. The gradient used to separate di/tripeptides was the following: 0−5 min, flushing column with 2.5% solvent B; 5−35 min, 2.5−50%; 35−45 min, 50−85%; 45−50 min, 85%; 50−50.01 min, 85−2.5%; 50.01−60 min, 2.5%.

Web-Based Search Tool. A theoretical MS/MS spectrum of a peptide is a collection of all possible fragment ion masses (i.e., b, y, and other types of ions). The di/tripeptide database consists of the theoretical MS/MS spectra of all possible sequences of 400 dipeptides and 8000 tripeptides. Theoretical MS/MS spectra of dimethyl labeled di/tripeptides were also generated for the 8400 peptides. With dimethyl labeling, di/tripeptides are very efficient to produce a_1 ions. In the process of matching the di/tripeptides from the unlabeled sample with the ones in the dimethyl labeled sample, the a_1 ion information was taken into special consideration when calculating the match scores of the input mass spectra. Specifically, the score will be multiplied by 1 plus the number of a_1 hit if there are a_1 ions matched with the identification (see below).

The database and the search tool, PEP Search, can be accessed in a public website, www.mycompoundid.org/PEP.

Database Search. Figure S1 in the Supporting Information shows the overview of the workflow used for database search, and Supporting Information Figure S2 shows the screenshot of the search results. The search website for MyCompoundID (MCID) also includes a tutorial and an example of search (see Supporting Information, notes 1 and 2). In the batch mode search, data generated from HILIC-LC/MS/MS of unlabeled and dimethyl labeled samples in CSV file format are separately uploaded onto PEP Search, and searched against the corresponding unlabeled and dimethyl labeled di/tripeptide databases. The two search results are automatically matched by a built-in program on the basis of their m/z of precursor ions (MH) and retention times (RT) within a user-defined tolerance (Tol) window. The search parameters applied in

this work from the data generated using the Bruker QTOF instrument were the following: MH Tol, 0.05 Da; fragment ion Tol, 0.01 Da; RT shift from −600 s; RT shift to 60 s; a_1 ion intensity threshold, 3 times of base intensity. While we have not tested the performance of the program for other instruments, the search program allows the user to adjust the parameters according to the performance of a given condition. The current program also allows manual search of an MS/MS spectrum of an unlabeled peptide. The match results can be manually inspected against the unlabeled (and labeled) spectrum to confirm di/tripeptide identification.

Match Scoring. To develop a scoring scheme for matching an experimental spectrum against a theoretical spectrum, we adopted a well-known "term frequency−inverse document frequency" (tf−idf) concept from information retrieval, which is a statistic way to extract unstructured information from a collection of information source, in text, audio, and/or video.[29] This concept is used for web-based data management and large-scale data analysis. In our study, when experimental MS/MS spectra are matched with the theoretical database, every peak from the MS/MS spectrum (the mass of a fragment ion) is regarded as a "term" (or keyword), and a theoretical spectrum is taken as a "document". Term frequency, tf, refers to the number of times the term occurs in a document, while the inverse document frequency, idf, is a measure of whether the term is common or rare across all the documents. tf−idf is a numerical statistic score which reflects how important a term is to a document in the overall MS/MS spectral database in our work. The tf−idf value of a term increases proportionally to the number of times the term appears in a document, but is offset by the frequency of the term in the database. This reflects the fact that some terms are generally more common than the others.

To score a match, if there is a peak in an experimental MS/MS spectrum of an unlabeled peptide matched to the unlabeled theoretical spectrum in the database, its tf−idf value would be calculated. The tf−idf values of all matched peaks in the MS/MS spectrum would be summed up to give the total match score. When searching the dimethyl labeled data set, for each precursor mass of a match, its corresponding dimethyl labeled mass would be searched within a retention time window of the unlabeled match, which was set to be +60 to −600 s in this work (this window is user-adjustable). Within the retention time window, if there are dimethyl labeled a_1 ions matched and the matched a_1 ion is from the N-terminal amino acid of the match, its score would be multiplied by (1+ number of a_1 ion matches). This multiplication results in a very high score for the match that is confirmed with the a_1 ion (i.e., the score would be greater than 100 while the scores of other matches without a_1 ion confirmation are usually much less than 100; see Supporting Information Table S1 as an example). In database search, the top ranked match with corresponding dimethyl labeled a_1 ion confirmation is normally considered to be the correct match (see Results and Discussion section on the confidence level of this approach).

RESULTS AND DISCUSSION

Figure 1 shows the overall workflow of di/tripeptide separation, HILIC-LC/MS detection, initial identification based on MS/MS database search, and confirmation with dimethyl labeling.

HILIC Separation. Reversed-phase liquid chromatography (RPLC)[15,30] has been reported for separating di/tripeptides. To improve retention and separation on RPLC column, di/tripeptides are usually chemically derivatized. However, many labeling reactions introduced a tag, such as N-dansyl[31] and N-AQC,[32] that resulted in the reduction of di/tripeptide backbone fragmentation in MS/MS, thereby losing structure or sequence information.

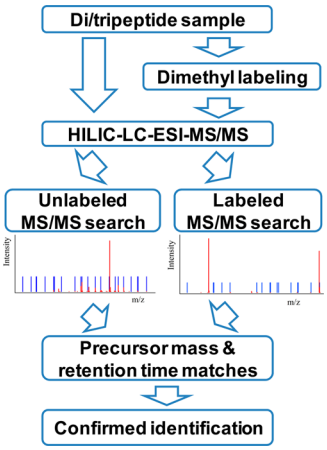


Figure 1. Workflow for di/tripeptide separation, identification, and validation with dimethyl labeling.

In our study, in order to generate a sufficient number of sequence-informative fragment ions of di/tripeptides for identification, di/tripeptides were first analyzed on LC-ESI-QTOF without chemical derivitization. In our initial development work, RPLC was used to separate a mixture of 38 di/tripeptide standards as well as di/tripeptides of the cytochrome c hydrolysate, followed by QTOF MS detection. Unfortunately, di/tripeptides did not retain well on RPLC column. Figure 2A shows the base-peak ion chromatogram obtained from RPLC-MS analysis of the mixture of 38 peptide standards. They eluted before 10 min.

In order to improve the retention and separation of di/tripeptides, HILIC chromatographic separation was selected. HILIC has been used to improve retention of hydrophilic analytes.[33−35] In addition, HILIC uses ESI-friendly mobile phase, such as acetonitrile, and thus can be readily coupled to ESI-MS.[35] In our experiment, the separation of 38 di/tripeptide standards on HILIC column was optimized. Figure 2B shows the base-peak ion chromatogram of the HILIC separation of the standard mixture. Most of the 38 di/tripeptides eluted after 10 min and distributed along the entire separation window.

Di/tripeptide Identification by MS/MS. Although some commercial databases (e.g., NIST 12 MS/MS library and Wiley MS/MS library) and online spectral databases (e.g., METLIN and MassBank) containing a small number of di/tripeptide MS/MS spectra are available, to our knowledge, there are no public accessible database and search tool for automatic identification of di/tripeptides using a comprehensive sequence library. The most commonly used search tools, such as SEQUEST[21] and MASCOT,[36] are designed for proteomics applications, which identify longer peptides containing 4 or more amino acids. Di/tripeptides cannot be reliably identified. To address this problem, we constructed a database consisting of all 8400 sequence possibilities of di/tripeptides and their corresponding possible fragment ions. A search algorithm, PEP Search, was developed for matching the masses of the precursor ion and fragment ions of a query peptide with the theoretical or predicted values in the database. This tool also includes a sequence confirmation part to help determine the N-terminal

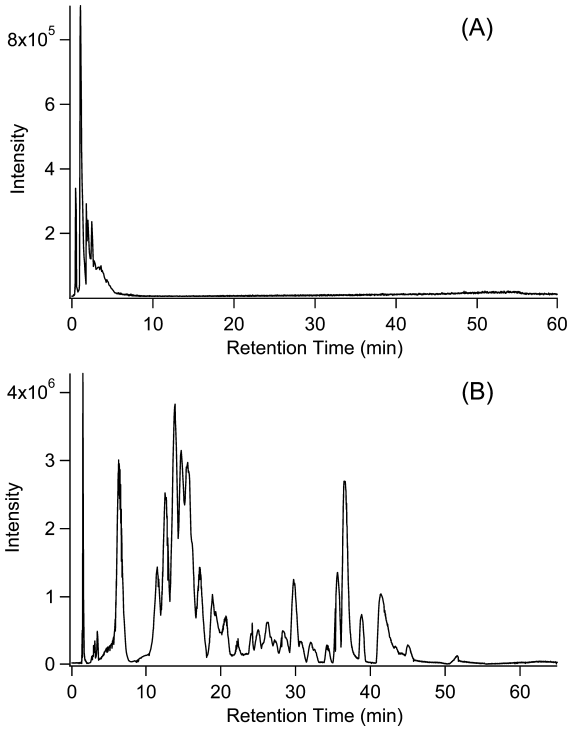


Figure 2. Base-peak ion chromatograms of 38 di/tripeptide standards separated using (A) RPLC and (B) HILIC-LC.

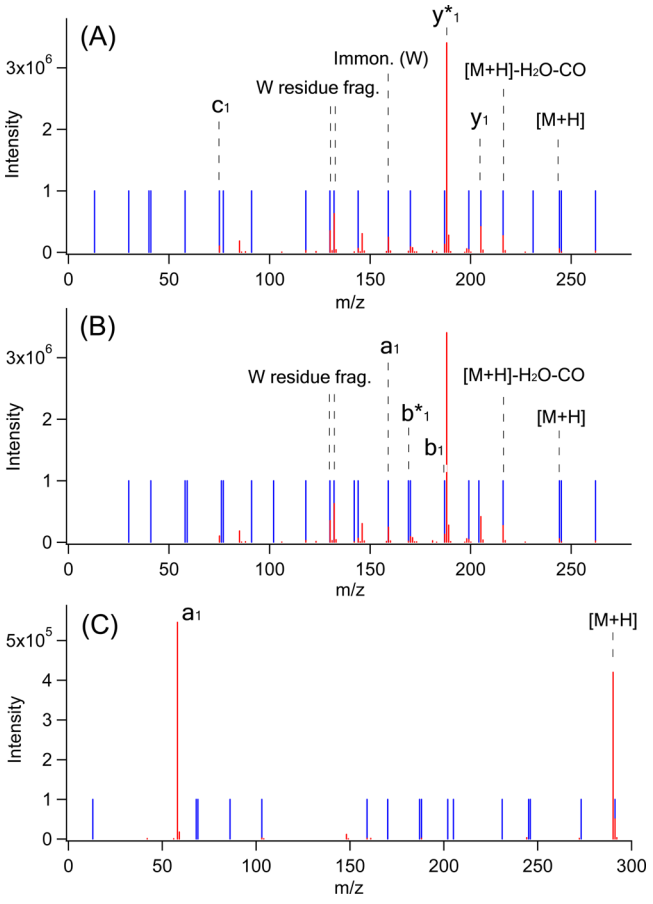


Figure 3. Search results of the experimental MS/MS spectrum obtained from the unlabeled peptide standard mixture matched with the theoretical fragment ions of (A) GW and (B) WG. (C) Experimental MS/MS spectrum of dimethyl labeled dipeptide matched with the theoretical fragment ions of dimethyl labeled GW.

amino acid of a di/tripeptide to eliminate false matches. To achieve this goal, dimethyl chemical derivatization was used to label the free amine group on the N-terminus and lysine residue of di/tripeptides. A corresponding fragment-ion database for dimethyl labeled di/tripeptides was constructed.

In our strategy (see Figure 1), a biological sample containing di/tripeptides is analyzed by HILIC-LC/MS/MS to obtain information on retention time, precursor ion m/z, and fragment ion m/z values. Then, the same sample is dimethyl labeled, followed by HILIC-LC/MS/MS analysis using the same LC gradient. Supporting Information Figure S1 illustrates the workflow used for database search. After uploading the two sets of MS/MS spectra, the program automatically searches the unlabeled and labeled databases for possible matches. The di/tripeptides in the sample with or without dimethyl labeling are matched with each other according to their precursor ion m/z and retention times (see later for more detailed discussion).

By applying this identification strategy, we can successfully differentiate and confirm the identification of dipeptides with reverse amino acid sequence in an isomeric pair. For example, as illustrated in Figure 3A,B, an MS/MS spectrum generated from an unlabeled dipeptide can be matched with both GW and WG. Due to a limited number of sequence-specific fragment ions generated in MS/MS, unambiguous identification to GW or WG is difficult, as they both have y-ions or b-ions matched and the numbers of matched fragment ions are similar: 8 peaks matched with GW and 7 peaks matched with WG. In this situation, when we searched the MS/MS spectrum of the dipeptide after dimethyl labeling, the dimethyl labeled a_1 ion is shown to be a dominant peak (see Figure 3C). This a_1 ion provides the information on the N-terminal amino acid of the dipeptide, allowing the confirmation of this dipeptide to be GW, not WG.

Because two separate data sets are generated and searched in our strategy, it is important to correlate the matches of the same peptide from the two LC/MS runs. In our work, the precursor ion mass and retention behavior of the matched di/tripeptides on HILIC column are used as the criteria for correlation. Theoretically, the mass of dimethylated peptide should be increased by 28.031 30 Da, compared to the unlabeled counterpart, due to the addition of $(CH_3)_2$ on amine after dimethyl labeling. If there is lysine (K) in a peptide, the ϵ-amine of lysine residue can also be dimethylated. Thus, the mass of dimethylated peptide containing lysine should be increased by $(n + 1) \times 28.031\ 30$ Da, where n = the number of lysine in the peptide sequence. In another situation, if the N-terminus of a di/tripeptide is proline (P), because of its cyclic structure, only one CH_3 can be added in dimethyl reaction. Thus, di/tripeptides with N-terminal proline would only have 14.015 65 Da increase in their precursor mass after dimethyl labeling.[37]

Besides the precursor mass, retention behavior changes of di/tripeptides before and after dimethyl labeling on HILIC column can also help correlate the same peptide in two runs. Generally, dimethyl labeling makes di/tripeptides more hydrophobic, resulting in a decrease in retention time on HILIC column. Figure 4A shows one example, YGG, where the retention time of dimethylated YGG was reduced, compared to that of the unlabeled YGG. Figure 4B shows the number distribution of

dx.doi.org/10.1021/ac500109y | Anal. Chem. 2014, 86, 3568−3574

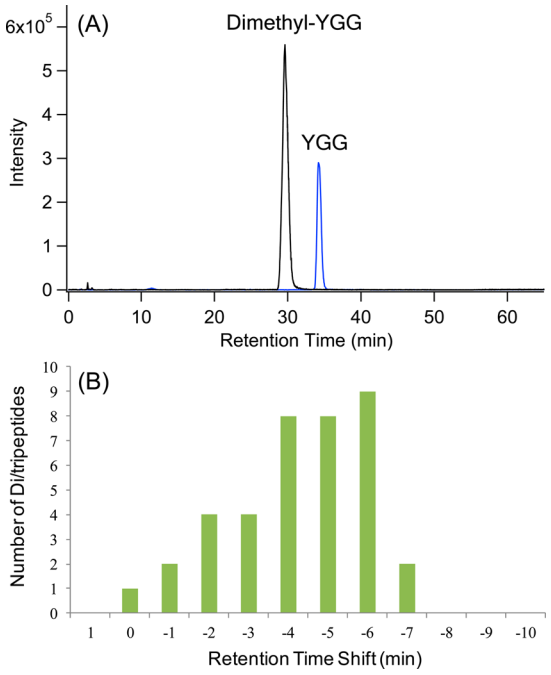


Figure 4. (A) Extracted ion chromatogram of unlabeled and dimethyl labeled YGG on HILIC-LC. (B) Distribution of retention time changes of 38 di/tripeptide standards after dimethyl labeling, compared to unlabeled di/tripeptide standards.

the 38 standard peptides as a function of the retention time shift. The time reduction ranges from 0 to 7 min. In the search program, this range can be adjusted by the user. In our work, in order to cover any possible larger shifts of other di/tripeptides that we have not tested, the retention time shift window before and after dimethyl labeling was conservatively set to be from +1 to −10 min.

Applying this method, we identified and confirmed all 38 di/tripeptide standards from the mixture. Without dimethyl labeling confirmation, a total of 46 matches of di/tripeptides were found from the unlabeled HILIC-LC/MS/MS analysis. Thus, the false identification rate was 17.4%. With dimethyl labeling confirmation, 8 false matches were eliminated, resulting in all correct 38 identifications of the di/tripeptide standards in the search result.

As a comparison, we also searched the MS/MS spectra of the 38 di/tripeptide standards against the publicly accessible experimental MS/MS spectral libraries. Searching the online NIST spectral library and the MassBank library did not result in any matches. Using the METLIN library, 21 putative peptide identifications were found and 3 of them were false, representing a false identification rate of 14.3%. Di/tripeptides with reversed amino acid sequences in isomeric pairs could not be distinguished in METLIN. These results indicate that the current experimental MS/MS spectral libraries have limited use for unambiguous identification of di/tripeptides.

Fragmentation Pattern of Dimethyl Labeled Di/tripeptides. The above results of analyzing the 38 standards illustrate the power of using dimethyl labeling to eliminate the false matches from the unlabeled peptides. This is made possible due to unique and complementary fragmentation patterns that can be generated from the unlabeled and labeled peptide ions. Dimethylation is a widely applied chemical labeling method for peptides and proteins because of its simplicity, and fast and almost complete reaction.[18,25,37,38] Dimethyl labeling can help enhance the peptide signals as it increases the basicity of the N-terminus of peptides.[39,40] Moreover, after dimethyl labeling, changes in the fragmentation patterns of peptides have been reported.[37] From the study of Hsu et al.,[37] dimethyl labeling was reported to have signal enhancement of a_1 ions derived from all 20 amino acids, because of the formation of stable quaternary immoniun ions. We also observed a_1 ion enhancement in MS/MS spectra of di/tripeptides. Without dimethyl labeling, a_1 ions were barely observed in the MS/MS spectra (see Supporting Information Figure S3A for AF as an example). The most commonly observed fragment ions from the unlabeled di/tripeptides are y and b ions, which are actually very useful to provide structure or sequence information on di/tripeptides. However, after dimethyl labeling, the signal of a_1 ions is greatly enhanced and usually becomes the base peak in the MS/MS spectrum, while other types of fragment ions are hardly observed (see Supporting Information Figure S3B for dimethylated AF).

Careful examination of the MS/MS fragmentation patterns of the 38 labeled di/tripeptide indicates that all of them could produce an intense a_1 ion and 34 of them (90%) actually generated the a_1 ions only in their MS/MS spectra. Over 94% (32 out of 34) of the labeled dipeptides did not generate any y/b ions. However, two out of four tripeptides produced some fragment ions, such as y_2 or b_2 ions (see Supporting Information Figure S3C for FFF). This is understandable considering that the longer sequence of tripeptides should increase the probability of causing backbone breakage. Some dimethyl labeled dipeptides with N-terminus tryptophan (W) could generate fragment ions at the tryptophan side chain,[41] but these fragment ions could not provide information on the sequence of the peptide. Overall, dimethyl labeling greatly enhances the a_1 ion signal of di/tripeptides at the cost of losing the y/b ions. Thus, using unlabeled di/tripeptides to obtain sequence information from the y/b ions is required to generate the list of possible matches. The enhanced a_1 ion from the dimethyl labeled di/tripeptides, which provides information on the identity of the N-terminal amino acid of a peptide, can then be used to confirm the identification of di/tripeptides from the initial matches.

Analysis of Cytochrome c Hydrolysate. To evaluate the performance of this method for identifying di/tripeptides in a complex sample, we analyzed an acid hydrolysate of a standard protein, cytochrome c. In this case, the sequences of di/tripeptides can be predicted from the protein sequence and thus can be used to gauge the accuracy of the matches and identifications. The protein hydrolysate generated from microwave-assisted acid hydrolysis using 6 M HCl was analyzed using HILIC-LC/MS/MS. Supporting Information Figure S5A shows the base-peak ion chromatogram of the unlabeled sample, and Supporting Information Figure S5B shows the chromatogram of the labeled sample. The MS/MS data from the unlabeled and labeled hydrolysates were searched using PEP Search to generate a list of di/tripeptide matches.

Table 1 summarizes the results obtained from cytochrome c hydrolysates prepared using different hydrolysis time. As Table 1 shows, using 30 min hydrolysis time, 65 di/tripeptides were matched from the unlabeled sample, and 48 of them were confirmed from the labeled sample. Manual check of the protein sequence, and hence the expected di/tripeptide sequences, as well as the MS/MS spectra, confirmed these 48 matches to be correct. Thus, the false discovery rate (FDR) was zero from the combined results of unlabeled and labeled

Table 1. Summary of the Peptide Identification Results Obtained from the Acid Hydrolysates of Cytochrome c Prepared Using Different Microwave Irradiation Times

MAAH time (min)	20	30	40	50	60
no. of total matches from the unlabeled sample	66	65	42	35	32
no. of peptides confirmed	48	48	36	28	28
sequence coverage (%) from the confirmed peptides	72.1	74.0	60.6	50.0	47.1
no. of false matches	15	14	4	5	3
FDR without confirmation	22.7%	21.5%	9.5%	14.3%	9.4%
FDR with confirmation	0	0	0	0	0
no. of peptides in protein sequence, but not confirmed	6	7	3	3	2
no. of manually confirmed peptides	3	3	2	2	1
false negative rate (%)	5.9	5.9	5.3	6.7	3.4

samples. These 48 di/tripeptides cover 74% of the cytochrome c sequence. Similarly, we examined the hydrolysates generated using other hydrolysis times and found that, in all cases, correct identifications were made from the combined results (see Table 1).

For the unconfirmed matches in the 30-min hydrolysates, 10 out of 17 unconfirmed matches could not be found in the protein sequence; these matches were obviously incorrect. The remaining 7 matches were not confirmed in the dimethyl labeling experiment, but they are in the sequence of cytochrome c. We manually checked their MS/MS spectra, especially looking for the y ion or b ion series. We found 3 tripeptides whose matches could be considered to be correct, as they had y ions or b ions present in the spectra. This example illustrates that our method may miss some true di/tripeptides in a biological sample; however, the false negative rate in this case was 3 out of 51 (48 + 3) peptides. The results of other MAAH experiments are also presented in Table 1, and the false negative rate ranges from 3.4% to 6.7%. In contrast, the false discovery (or positive) rate from the unlabeled samples was 21.5% [i.e., (17 − 3)/65] for the 30-min hydrolysate. As Table 1 shows, the FDR ranges from 9.4% to 22.7% in the search results of the unlabeled samples.

We note that PEP Search was developed for identification of di/tripeptides naturally present in a biological sample (see an example shown below), not for protein identification. The sequences of di/tripeptides are too short for unambiguous protein identification based on sequence search alone. The above analysis of the acid hydrolysate of cytochrome c with known sequences of di/tripeptides was intended to examine the performance of our method in terms of sensitivity and specificity for di/tripeptide identification.

Analysis of Human Urine. Many di/tripeptides are present naturally in biofluids such as urine. These small peptides may potentially be biomarkers like other types of small molecules for disease diagnosis. To identify di/tripeptides in human urine, the unlabeled and dimethyl labeled human urine samples were separately analyzed by HILIC-LC/MS/MS. The resultant ion chromatograms are shown in Supporting Information Figure S5C,D. In a search of the data from the unlabeled urine, 59 matches were returned in the result (Supporting Information Table S1). When the unlabeled and labeled data were searched together, 13 di/tripeptide matches were confirmed (Table 2). Among the 13 confirmed di/tripeptides, we validated the identification of AL with a standard peptide by comparing their retention time and MS/MS spectrum (see Supporting Information Figure S4).

Table 2. List of Di/tripeptides Identified from a Human Urine Sample

di/tripeptide	m/z	retention time (min)	dimethyl labeled a_1 ion
LGG	246.1093	13.9	L(I)
GE	205.0982	14.7	G
GSG	220.0617	19.2	G/V
KD	262.1655	28.5	K
AV	189.1236	31.7	A
GA	147.0766	34.5	G
RD	290.1605	37.0	R/K
AL	203.1506	41.4	A/S
SP	203.1506	41.4	A/S
QD	262.1288	41.7	Q
VG	175.0229	44.9	V/L(I)
VA	189.1601	47.0	V/L(I)
LG	189.1601	47.0	V/L(I)

There were 46 unconfirmed di/tripeptide matches from urine. Upon close examination of these MS/MS spectra, we found most of these matches had poor matches with the theoretical spectra, such as only one or two peaks matched with the theoretical fragments, and generally, no y/b ions matched. These were most likely the random matches because of the chemical complexity of the urine sample where there were many other nonpeptide chemicals that could have similar precursor ion masses as the peptides. Furthermore, 33 of the 46 matches were eluted out before 10 min in the HILIC-LC gradient run, indicating that they were hydrophobic and not consistent with the expected retention behavior of the di- and tripeptides. This example illustrates that, by applying our search strategy, we can confirm di/tripeptide matches and significantly decrease the number of false identifications in a complex biological sample. Although we used the retention time information in this example to judge the peptide identification, it would be highly useful if more accurate prediction of the retention behavior of di/tripeptides were available. Thus, we expect that future work on accurate retention behavior prediction of these peptides should result in further improvement of the method.

Future Perspective. There are several research areas worth pursuing to improve the performance and applicability of the PEP Search method. In our current search algorithm, peak abundance information is not used. Recent work has suggested that it is possible to predict relative intensities of fragment ions in MS/MS spectra of small molecules and peptides.[42,43] Incorporating the predicted fragment ion abundance information in the search algorithm will likely improve the overall matching performance in terms of sensitivity and specificity. In addition, expanding the current theoretical di/tripeptide spectral library by adding possible modifications to these peptides, such as formylation and acetylation, will be very useful for identifying modified di/tripeptides. Modified di/tripeptides have shown to play important biological roles.[44] We also note that in our current library only the 20 common amino acids were used to build the di/tripeptide sequences. However, over 70 uncommon amino acids have been found, and thousands of other amino acid structures have been recently predicted.[45] Adding these amino acids to the di/tripeptide sequence library should be useful for identifying and discovering small peptides

containing one or more uncommon amino acids. Finally, in principle, our method should be expandable to tetra- and pentapeptides. The size of the spectral library would be significantly increased, which would increase the search time. Validation of the performance would require the synthesis of many tetra- and pentapeptides.

CONCLUSIONS

We have developed a strategy for the identification of di/tripeptides in complex biological samples. In our method, one aliquot of a sample is analyzed by HILIC-LC/MS/MS directly, and another aliquot of the same sample is labeled using dimethyl reaction, followed by HILIC-LC/MS/MS. The two sets of the MS/MS data are entered into a search program, PEP Search, in the MycompoundID website, which is developed specifically for di/tripeptide identification. The peptide matches from the unlabeled sample which contain many false positives are sorted and confirmed by the labeled sample. Future work will be directed toward the development of enabling sample preparation and separation methods to enrich the low abundance di/tripeptides present in a metabolomic sample in order to improve their detectability.

ASSOCIATED CONTENT

Supporting Information

Additional information as noted in text. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: Liang.Li@ualberta.ca.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the Natural Sciences and Engineering Research Council of Canada, the Canadian Institutes of Health Research, the Canada Research Chairs program, Genome Canada, and Genome Alberta.

REFERENCES

(1) Snyder, S. H. *Science* **1980**, *209*, 976−983.
(2) Bonfanti, L.; Peretto, P.; De Marchis, S.; Fasolo, A. *Prog. Neurobiol.* **1999**, *59*, 333−353.
(3) Weichart, D.; Gobom, J.; Klopfleisch, S.; Hasler, R.; Gustavsson, N.; Billmann, S.; Lehrach, H.; Seegert, D.; Schreiber, S.; Rosenstiel, P. *J. Biol. Chem.* **2006**, *281*, 2380−2389.
(4) Scheffler, J. E.; Berliner, L. J. *Biophys. Chem.* **2004**, *112*, 285−291.
(5) Vanneste, Y.; Michel, A.; Dimaline, R.; Najdovski, T.; Deschodt-Lanckman, M. *Biochem. J.* **1988**, *254*, 531−537.
(6) Fonteh, A. N.; Harrington, R. J.; Tsai, A.; Liao, P.; Harrington, M. G. *Amino Acids* **2007**, *32*, 213−224.
(7) Wu, M. H.; Xu, Y.; Fitch, W. L.; Zheng, M.; Merritt, R. E.; Shrager, J. B.; Zhang, W. R.; Dill, D. L.; Peltz, G.; Hoang, C. D. *Rapid Commun. Mass Spectrom.* **2013**, *27*, 2091−2098.
(8) Osborn, M. P.; Park, Y.; Parks, M. B.; Burgess, L. G.; Uppal, K.; Lee, K. C.; Jones, D. P.; Brantley, M. A. *PLoS One* **2013**, *8*, 10.
(9) Husek, P.; Svagera, Z.; Vsiansky, F.; Franekova, J.; Simek, P. *Clin. Chem. Lab. Med.* **2008**, *46*, 1391−1397.
(10) Jandke, J.; Spiteller, G. *J. Chromatogr.* **1986**, *382*, 39−45.
(11) Nakashima, E. M. N.; Kudo, A.; Iwaihara, Y.; Tanaka, M.; Matsumoto, K.; Matsui, T. *Anal. Biochem.* **2011**, *414*, 109−116.
(12) Cimlova, J.; Kruzberska, P.; Svagera, Z.; Husek, P.; Simek, P. *J. Mass Spectrom.* **2012**, *47*, 294−302.
(13) Bobba, S.; Resch, G. E.; Gutheil, W. G. *Anal. Biochem.* **2012**, *425*, 145−150.
(14) Stressler, T.; Eisele, T.; Fischer, L. *Int. Dairy J.* **2013**, *30*, 96−102.
(15) Takahashi, K.; Tokuoka, M.; Kohno, H.; Sawamura, N.; Myoken, Y.; Mizuno, A. *J. Chromatogr. A* **2012**, *1242*, 17−25.
(16) Fan, Y.; Rubakhin, S. S.; Sweedler, J. V. *Anal. Chem.* **2011**, *83*, 9557−9563.
(17) Li, L. J.; Sweedler, J. V. In *Annual Review of Analytical Chemistry*; Annual Reviews: Palo Alto, CA, 2008; Vol. *1*, pp 451−483.
(18) Fu, Q.; Li, L. J. *Anal. Chem.* **2005**, *77*, 7783−7795.
(19) Ubhi, B. K.; Davenport, P. W.; Welch, M.; Riley, J.; Griffin, J. L.; Connor, S. C. *J. Chromatogr. B: Anal. Technol. Biomed. Life Sci.* **2013**, *934*, 79−88.
(20) Zheng, J.; Dixon, R. A.; Li, L. *Anal. Chem.* **2012**, *84*, 10802−10811.
(21) Link, A. J.; Eng, J.; Schieltz, D. M.; Carmack, E.; Mize, G. J.; Morris, D. R.; Garvik, B. M.; Yates, J. R. *Nat. Biotechnol.* **1999**, *17*, 676−682.
(22) Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; King, N.; Stein, S. E.; Aebersold, R. *Proteomics* **2007**, *7*, 655−667.
(23) Smith, C. A.; O'Maille, G.; Want, E. J.; Qin, C.; Trauger, S. A.; Brandon, T. R.; Custodio, D. E.; Abagyan, R.; Siuzdak, G. *Ther. Drug Monit.* **2005**, *27*, 747−751.
(24) Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K.; Oda, Y.; Kakazu, Y.; Kusano, M.; Tohge, T.; Matsuda, F.; Sawada, Y.; Hirai, M. Y.; Nakanishi, H.; Ikeda, K.; Akimoto, N.; Maoka, T.; Takahashi, H.; Ara, T.; Sakurai, N.; Suzuki, H.; Shibata, D.; Neumann, S.; Iida, T.; Funatsu, K.; Matsuura, F.; Soga, T.; Taguchi, R.; Saito, K.; Nishioka, T. *J. Mass Spectrom.* **2010**, *45*, 703−714.
(25) Hsu, J. L.; Huang, S. Y.; Chow, N. H.; Chen, S. H. *Anal. Chem.* **2003**, *75*, 6843−6852.
(26) Ji, C.; Li, L. *J. Proteome Res.* **2005**, *4*, 734−742.
(27) Wang, N.; Li, L. *J. Am. Soc. Mass Spectrom.* **2010**, *21*, 1573−1587.
(28) Tang, Y.; Li, L. *Anal. Chim. Acta* **2013**, *792*, 79−85.
(29) Robertson, S. *J. Doc.* **2004**, *60*, 503−520.
(30) Gil-Agusti, M.; Esteve-Romero, J.; Carda-Broch, S. *J. Chromatogr. A* **2008**, *1189*, 444−450.
(31) Zheng, J. M.; Li, L. *Int. J. Mass Spectrom.* **2012**, *316*, 292−299.
(32) Ullmer, R.; Plematl, A.; Rizzi, A. *Rapid Commun. Mass Spectrom.* **2006**, *20*, 1469−1479.
(33) Guo, K.; Ji, C. J.; Li, L. *Anal. Chem.* **2007**, *79*, 8631−8638.
(34) Schlichtherle-Cerny, H.; Affolter, M.; Cerny, C. *Anal. Chem.* **2003**, *75*, 2349−2354.
(35) Guo, Y.; Gaiki, S. *J. Chromatogr. A* **2011**, *1218*, 5920−5938.
(36) Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20*, 3551−3567.
(37) Hsu, J. L.; Huang, S. Y.; Shiea, J. T.; Huang, W. Y.; Chen, S. H. *J. Proteome Res.* **2005**, *4*, 101−108.
(38) Lo, A.; Tang, Y. A.; Chen, L.; Li, L. *Anal. Chim. Acta* **2013**, *788*, 81−88.
(39) Ji, C. J.; Lo, A.; Marcus, S.; Li, L. *J. Proteome Res.* **2006**, *5*, 2567−2576.
(40) Wu, C. J.; Hsu, J. L.; Huang, S. Y.; Chen, S. H. *J. Am. Soc. Mass Spectrom.* **2010**, *21*, 460−471.
(41) Fu, Q.; Li, L. J. *J. Am. Soc. Mass Spectrom.* **2006**, *17*, 859−866.
(42) Galezowska, A.; Harrison, M. W.; Herniman, J. M.; Skylaris, C. K.; Langley, G. J. *Rapid Commun. Mass Spectrom.* **2013**, *27*, 964−970.
(43) Pechan, T.; Gwaltney, S. R. *BMC Bioinf.* **2012**, *13 (Suppl 15)*, S13.
(44) Merlin, D.; Si-Tahar, M.; Sitaraman, S. V.; Eastburn, K.; Williams, I.; Liu, X.; Hediger, M. A.; Madara, J. L. *Gastroenterology* **2001**, *120*, 1666−1679.
(45) Meringer, M.; Cleaves, H. J.; Freeland, S. J. *J. Chem. Inf. Model.* **2013**, *53*, 2851−2862.