PRESnovo: Prescreening Prior to *de novo* Sequencing to Improve Accuracy and Sensitivity of Neuropeptide Identification

Kellen DeLaney[1,#], Weifeng Cao[1,#], Yadi Ma[2], Mingming Ma[3], Yuzhuo Zhang[1], and Lingjun Li[1,3*]

[1]Department of Chemistry, [2]Department of Computer Sciences and [3]School of Pharmacy, University of Wisconsin-Madison, 777 Highland Ave., Madison, WI 53705

[#] These authors contributed equally to this work.

*To whom correspondence should be addressed.

Running Title: PRESnovo

Keywords: PRESnovo, *de novo* sequencing, tandem mass spectrometry, neuropeptide, peptide identification, motif

Address reprint requests to Dr. Lingjun Li. Mailing Address: 5125 Rennebohm Hall, 777 Highland Avenue, Madison, WI 53706-2222. Phone: (608)265-8491, Fax: (608)262-5345. E-mail: **lingjun.li@wisc.edu**.

Abstract

Identification of peptides in species lacking fully-sequenced genomes is challenging due to the lack of prior knowledge. *De novo* sequencing is the method of choice, but its performance is less than satisfactory due to algorithmic bias and interference in complex MS/MS spectra. The task becomes even more challenging for endogenous peptides that do not involve an enzymatic digestion step, such as neuropeptides. However, many neuropeptides possess common sequence motifs that are conserved across members of the same family. Taking advantage of this feature to improve *de novo* sequencing of neuropeptides, we have developed a method named PRESnovo (prescreening precursors prior to *de novo* sequencing) to predict the motif from a MS/MS spectrum. A neuropeptide sequence is broken into a motif with conserved amino acid residues and the remaining partial sequence. By searching against a predefined motif database constructed from known homologous sequences, PRESnovo assigns the most probable motif to each precursor via a sophisticated scoring function. Performance analysis was conducted with 15 neuropeptide standards, and 11 neuropeptides were correctly identified with PRESnovo compared to 1 identification by PEAKS only. We applied PRESnovo to assign motifs to peptide sequences in conjunction with PEAKS for assigning the rest of the peptide sequence in order to discover neuropeptides in tissue samples of green crab, *C. maenas*, and Jonah crab, *C. borealis*. Collectively, a large number of neuropeptides were identified, including 13 putative neuropeptides identified in green crab brain, 77 in Jonah crab brain, and 47 in Jonah crab sinus glands for the first time. This PRESnovo strategy greatly simplifies *de novo* sequencing and enhances the accuracy and sensitivity of neuropeptide identification when common motifs are present.

Introduction

With the advancement of high-throughput mass spectrometry (MS), shotgun proteomics has been employed as the major tool for performing large-scale analysis of biological samples. With current MS instruments and workflows, thousands of MS and MS/MS spectra are produced from a single analysis [1-3]. With the large numbers of spectra being generated, traditional manual analysis is no longer feasible, and numerous data analysis strategies have been developed to identify proteins from the high volume of MS/MS spectra. The methods of choice for annotating these spectra mainly include spectral matching, database searching, and *de novo* sequencing. Spectral matching can be performed with a variety of software platforms, including SpectraST [4], X1Hunter [5], and Bibliospec [6]. However, this method of analysis requires the generation of spectral libraries, which can be time consuming and costly. Database searching employs search engines such as Mascot, SEQUEST, OMSSA, MSFragger and X!Tandem [7-11] to interpret tandem MS spectra by scoring the similarities between the experimental and theoretical spectra generated through *in silico* prediction. The matches are then ranked such that the match with the highest score is the best predicted peptide spectrum match. While this method has proven to be highly effective for large-scale identifications, it requires prior knowledge of the sequences of peptides in the sample, typically through the use of genomic data. This limits its applicability to organisms with a fully-sequenced genome.

For those species without a fully-sequenced genome, no protein database is available, and so *de novo* sequencing is the main approach employed for peptide identification. With *de novo* sequencing, peptide sequences are derived from the masses of their fragment ions as shown in a MS/MS spectrum. When performing *de novo* sequencing, no protein sequence database is used

for reference, and so no prior knowledge is required of peptides in the sample. A number of algorithms and software packages have been developed recently for *de novo* peptide sequencing [12-15]. Most software use “spectral graph” or "probabilistic model" to interpret MS/MS spectra. A spectral graph is constructed from a MS/MS spectrum in which nodes represent fragment peaks and two nodes are connected by an edge if the mass difference between these two nodes matches a known amino acid residue. The algorithm tries to find t he longest possible path of connected edges from the N-terminus to the C-terminus in order to determine the combination of amino acids that best represents the peptide sequence. However, any incompleteness of fragmentation can cause gaps which disconnect the longest path from N- to C-termini. In some cases, the longest path fails to represent the correct sequence or partial sequences due to interfering fragments. Instead of constructing spectral graphs, some algorithms such as PEAKS [15] compute peptide sequences among all possible amino acid combinations and then map these sequences directly onto each MS/MS spectrum to find the best sequence match, which improves the likelihood of a correct sequence assignment despite interfering ions. While this method is more computationally expensive, dynamic programming is always employed to increase the computing efficiency. As a result, this method has been employed successfully for numerous *de novo* sequencing methods. However, in comparison to the database search strategy, the accuracy and sensitivity of *de novo* sequencing are far less satisfactory [16]. The combination of database searching and *de novo* sequencing to some extent improves the accuracy of prediction [16, 17]. Although these hybrid methods enhance peptide prediction, they require protein databases available for the species of interest. For species without available genomes, improving peptide identification remains an unmet challenge.

One example of a group of species without a complete genome is crustaceans, which are important model organisms for studying dynamic neural networks and neuromodulation. Neuropeptides, or short endogenous peptides involved in neuronal signaling, are an important class of molecules related to this system. Much effort has been placed on characterizing the neuropeptides involved in these signaling pathways with MS [18-24] and *in silico* prediction [25]. However, this class of molecules is particularly challenging to study because their workflow does not involve a digestion step [26]. With tryptic digests from proteins, the search space is able to be reduced based on how trypsin digests the proteins with well-defined amino acid residue at the carboxyl-end of resulting tryptic peptides. With neuropeptides, no enzymatic digestion parameters can be used to reduce the search space. However, highly-conserved peptide sequences like neuropeptides share conserved motifs, which can aid *de novo* sequencing. Herein, we focus on interpreting a part of a tandem MS spectrum to extract the conserved sequence motif instead of sequencing the entire MS/MS fragmentation spectrum. The determination of a motif will simplify *de novo* sequencing of the rest of the sequence and increase the accuracy of peptide identification. In order to determine the neuropeptide sequence motif, a database search method is employed similar to commonly-used methods previously described. First, motifs are derived from known peptide sequences from homologous species. Second, theoretical *in silico* fragmentation is performed for these motifs. The resulting b- and y- ions as well as their neutral loss fragments together with the corresponding motif are used to construct a motif database. This database is then searched by inputting experimental fragments to find the best matched motif for each precursor. A sophisticated scoring function, based on the sum of ratios of lengths of fragments to total motif lengths, ensures the correct assignment of motif to a given precursor. This strategy, named PRESnovo, simplifies the subsequent *de novo* sequencing step and

increases the accuracy and sensitivity of peptide identifications by performing a preprocessing function prior to *de novo* sequencing.

Methods

Construction of Motif Database.

A motif database is required to run PRESnovo. Here, motif refers to a representation of the similarity of different sequences in the same peptide family. For accurate results, it is recommended that all possible motifs for the peptides of interest are included. In this study, previously known neuropeptides from invertebrate species, mainly crustaceans, were collected from public databases including Uniprot knowledgebase and publications [25, 27-29]. In order to estimate the false positive rate, some vertebrate motifs were also included as decoy motifs because these motifs were not observed in crustacean neuropeptides . As the targeted neuropeptide sequences in most cases contain fewer than 20 amino acid residues, longer sequences such as proteins or receptors were excluded from the list. **Figure 1** shows how a motif database is constructed. The known neuropeptides collected from the various sources were clustered according to their family names followed by a multiple alignment procedure with ClustALW2 (http://www.ebi.ac.uk/Tools/msa/clustalw2/) for each family. Afterwards, the most aligned region was truncated to extract motifs with WebLogo3 (http://weblogo.threeplusone.com/). Longer peptides whose motifs were not easily determined were removed from the list, such as Bursicon, crustacean hyperglycemic hormone (CHH), molt inhibiting hormone (MIH), etc. When using WebLogo3, multiple motifs were adopted to represent each family. For example, all 5 C-terminal motifs, YAFGL, YDFGL, YNFGL, YEFGL, and YSFGL, were used to represent allatostatin A-type (AST-A) peptides.

The PRESnovo method works by mapping tandem MS spectra onto a home-built motif database in order to find the best suitable motif for each precursor. The practical implementation is to compare the experimental fragments associated with a precursor to those theoretical fragments associated with a motif. Therefore, a list of theoretical fragments must be created for each motif. The extracted motifs were *in silico* fragmented by MS-Product (http://prospector.ucsf.edu/prospector/cgi-bin/msform.cgi?form=msproduct), and the resulting b ion set (including b ions and neutral losses, b- H_2O and b-NH_3) were adopted as the theoretical fragments for N-terminal motifs while y ion set (including y ions and neutral losses, y- H_2O and y-NH_3) were adopted as the theoretical fragments for C-terminal motif. Post-translational modifications (PTM) were also included when fragmenting these motifs. For example, for motif YXFGLamide, C-terminal amidation was considered for the production of theoretical fragments. These were selected based on common PTMs observed for neuropeptide families [27, 30].

We used these extracted motifs and the associated theoretical fragments to construct our motif database (*see supplemental file*) in which 87 motifs covering 25 families were compiled for crustacean neuropeptides (2 families of 7 mammalian neuropeptides (Angiotensinogen (5) and Arg_vasopressin (2) were also included for test only, to be used as decoys to assess the false positive rate). The distribution of the motifs is shown in **Figure 2**.

Scoring Function and Motif Assignment.

Accurately assigning a motif to a given precursor is crucial. A well-defined scoring function can ensure the unbiased assignment of a motif in most cases. As such, a score is needed to evaluate the comparison between experimental and theoretical fragments within PRESnovo. Several considerations must be taken to generate a scoring function. These factors include significance of different-length theoretical fragments, the number of matches between experimental and

theoretical fragments, and the percentage of the total number of experimental fragments with respect to the total number of theoretical fragments. We first define the score of a theoretical fragment via dividing the length of fragment by the total length of the motif. This definition is suitable because the significance of longer fragments is higher than that of shorter fragments [31]. This score definition is also beneficial to the interpretation of non-informative tandem MS spectra because any high-scoring longer fragments can ensure a reliable motif assignment without consideration of gaps. After computing scores for fragments, these fragment scores were used to evaluate the motif assignment. Given that the more theoretical fragments are matched by experimental fragments the more reliable the corresponding motif assignment is, the score for a motif assignment is defined by the following formula:

$$S(P(\vec{F}_E) \sim M(\vec{F}_T)) = \frac{\sum S_E(\vec{F}_E \mid M(\vec{F}_T))}{\sum S_T(\vec{F}_T \mid M(\vec{F}_T))}$$

$S(P(\vec{F}_E) \sim M(\vec{F}_T))$ is the score for motif M（$M(\vec{F}_T)$）assigned to a precursor（$P(\vec{F}_E)$）, $\sum S_E(\vec{F}_E \mid M(\vec{F}_T))$ is the sum of the scores of all matched experimental fragments while $\sum S_T(\vec{F}_T \mid M(\vec{F}_T))$ is the sum of the scores of all theoretical fragments. In our motif database, the scores of fragments were calculated in advance and stored along with motifs. The neutral loss fragments share the same score as the associated b or y ions.

The workflow for using PRESnovo to assign the best matched motif to a given precursor is shown in **Figure 3**. The user-defined mass error tolerance was set to 0.5 Da for QTOF data and 0.02 Da for Orbitrap data for both precursor and fragments. The experimental fragments were compared to the theoretical fragments associated with each motif and the matched pairs were stored. If a neutral loss ion pair and the corresponding b or y ion pair were both found, only

one pair was kept for the following calculation. In the list of matched fragment pairs, all the matched theoretical fragment scores were used to calculate the score of each motif assignment. As a result, for a given precursor, all motifs were tentatively assigned to it with calculated scores. These motifs were then ranked in descending order based on the score for each motif assignment and the first five ranked motifs were reported.

NanoLC-ESI-QTOF Analysis for Peptide Standards and Tissue Sample.

A peptide standard mixture containing 15 neuropeptides (Supplemental Table S1) was subjected to nano-LC-ESI-QTOF (Waters Corp., Milford, MA) and the resulting MS data were used to test the performance of PRESnovo. Furthermore, PRESnovo was used for real tissue sample to identify endogenous neuropeptides. Tissue samples were extracted from the brain of European green crabs, *C. maenas* and brain and sinus glands of Jonah crabs, *C. borealis*. Green crab tissue was offline HPLC fractionated, and the collected fractions were analyzed on a nano-LC-ESI-QTOF (Waters Corp., Milford, MA). The Jonah crab tissue was analyzed on a nano-LC-ESI Q Exactive Orbitrap mass spectrometer (Thermo Scientific, Bremen, Germany). The details about experimental protocols including animal dissection, tissue extraction, offline HPLC fractionation and MS analysis can be found in the **Supporting Information** and followed those previously described [32]. After MS analysis, the QTOF raw data were converted into pkl formatted data with ProteinLynx (Waters Corp., Milford, MA). Parameters for ProteinLynx were set with default settings except noise reduction threshold at 10%. The Orbitrap raw data was converted into mzXML format using MSConvert with default settings [33].

De novo *Sequencing with the Aid of PRESnovo*.

The resulting pkl and mzXML files were used as input to PRESnovo. The current version of PRESnovo supports pkl and mzXML formats. Parsing of mzXML data was performed with

jmzReader [34]. Other data formats need to be converted into one of these two formats. Mass tolerances for both precursor and fragments were set to 0.5 Da for QTOF data and 0.02 Da for Orbitrap data. PRESnovo outputs a .csv file in which each precursor is assigned five motifs as the default setting according to a descending order of score. The number of motif assignments can be changed by the user. The scored fragments associated with each motif are also included in the output. By setting a motif score threshold, the user can filter out low-confidence motif assignments. Once a motif is obtained, the rest of the sequence can be determined by either manual sequencing or PEAKS. Given that PEAKS can report high-confidence sequence tags with individual positional confidence scores [15], it is very useful to combine PEAKS and PRESnovo together to sequence a peptide.

As a comparison, raw MS/MS data was also processed with PEAKS. The raw data were directly loaded into PEAKS (PEAKS 7, Bioinformatics Solutions Inc., Waterloo, ON) for *de novo* sequencing. The setting for PEAKS were as follows: mass error tolerances for both precursor and fragments 0.5 Da for QTOF data and 0.02 Da for Orbitrap data, no enzyme digestion, and pyro-glutamine (pQ), pyro-glutamic acid (pE), oxidation of methionine, and C-terminal amidation as variable PTMs.

Results and Discussion

Performance of PRESnovo.

In order to evaluate the performance of PRESnovo, a standard mixture comprised of 15 peptides was analyzed with LC-MS/MS on a Waters QTOF instrument. The data were processed with both PRESnovo and PEAKS *de novo* sequencing to compare performance. As five predictions were output from both software packages, it was important to differentiate which was the most

suitable. The predicted sequences from each software were compared to the actual sequences of the standards. Interestingly, in most cases, the first prediction was the best one in the PRESnovo results. In the PEAKS results, this was not always the case, as other sequence predictions rather than the first one were the best prediction for some peptides. **Table 1** summarizes the results obtained from PRESnovo and PEAKS. Five output sequences are provided from both software, and the sequence matching the standard the closest is displayed. As shown, 11 out of 15 peptides were predicted correctly by PRESnovo while only 1 was computed correctly by PEAKS with at least six consecutively correct residues. The 11 peptides identified by PRESnovo produced doubly-charged precursors leading to an almost even number of b and y ions, which facilitated the detection of fragments associated with their motifs by PRESnovo. Two examples are shown in **Figure 4**, from which it can be seen that almost all motif-related fragments were produced, although some of them were at low intensities. Consequently, two motifs were confidently assigned to the precursors, respectively. However, PEAKS lacks specificity for identification of these endogenous peptides because PEAKS is originally designed for general use in bottom-up proteomics where tryptic peptides are the main targets. With tryptic peptides, certain patterns exist that inform the resulting *de novo* sequences obtained. As a result, PEAKS by itself is not optimal for non-tryptic peptide prediction [15]. Because PRESnovo is designed to recognize patterns specific to endogenous neuropeptides, it provides improved accuracy in sequencing these signaling peptides when used in conjunction with PEAKS.

The 4 peptides not identified correctly by PRESnovo fall into two categories: AST-A peptides (GDGRLYAFGLa and APSGAQRLYGFGLa) and disulfide bond bridging peptides (PFCNAFTGCa and CYFQNCPRGa). AST-A peptides are generally ionized as singly charged precursors which tend to produce predominantly N-terminal (primarily b-series) fragments.

These fragments (a/b/c ions) dominate the MS/MS spectrum, as shown in **Figure 5A**. However, the characteristic motif of AST-A peptides, YXFGLamide (X is a variable residue), is located at the C-terminus, which causes the motif database to deduce y-series fragments from this motif. As a result, the mismatch between the experimental b-series fragments and *in silico* y-series fragments leads to misidentification of these peptides. Although this demonstrates a limitation of PRESnovo's predicting power, it is not difficult to manually sequence AST-A peptides with the aid of PEAKS, given that most peptides belonging to this family have a simple pattern of b-ions dominating fragmentation. Disulfide bond bridging peptides are prone to producing complicated internal fragments that suppress the production of b/y ions (**Figure 5B**). As a result, not enough motif-associated b/y ions are available for PRESnovo to determine the motif for the precursor. While the inability to predict disulfide bond bridging peptide sequences is a limitation of PRESnovo, the same is true across many *de novo* sequencing software. Disulfide bond bridging peptides are relatively uncommon amongst neuropeptides, and only 4 possible sequences with a characteristic motif are present in the crustacean neuropeptide database. Therefore, this limitation is not expected to substantially impede the identification of crustacean neuropeptides.

The Factors Impacting on the Performance of PRESnovo.

The performance of PRESnovo is impacted mainly by the motif database being searched and the quality of tandem MS spectra being queried. The construction of a motif database is crucial to PRESnovo, as accurate, detailed compilation of motifs improves the likelihood of confident identifications being made. A motif database consists of two parts: a string of amino acids comprising each motif and its associated fragments. Motifs are generally collected from known peptide sequences present in homologous species. There is a tradeoff between the number of

motifs incorporated in a motif database and the prediction power of PRESnovo. If too few motifs are included such that all forms of a neuropeptide motif are not sufficiently represented, the search may return results with low specificity or inaccurate identifications. Conversely, having an excess number of motifs included in the database inflates the false positive rate and decreases the sensitivity of the assignment [35]. In order to maximize the detection of neuropeptides in tissue samples, a motif library containing at least 80 motifs is satisfactory, as indicated in **Figure S1a**. To avoid these issues, only biologically-related species should be considered when extracting motifs for a species of interest. For example, in this study, motifs were only extracted from known peptide sequences in other crustacean species. If the interest is in peptide identifications from a vertebrate species such as human or mouse, a vertebrate peptide motif database can be created specifically for the species of interest that only includes those from similar species. In this way, effort can be made to exclude interference from other motifs that are not likely to be present in the species of interest. Another possible solution is to use long motifs, which would reduce the likelihood of false positive prediction and would also simplify the subsequent *de novo* sequencing. *In silico* fragments associated with motifs also impact on the performance of PRESnovo. In our current motif database, only b and y series ions (i.e., b, y, and neutral loss ions) are included for each motif while no a/c or x/z ions are considered. The reason for this specification is that b and y ions are sufficient for determining a motif even though some of them may have low abundance in some cases.

The quality of tandem MS spectra also has significant impact on the performance of PRESnovo. Different from many automated *de novo* algorithms, PRESnovo can predict correct motifs for relatively low-quality MS/MS spectra provided that enough motif-associated fragments are present in the spectra. However, in the case of processing less informative spectra,

if the experimental fragments associated with motif are insufficient or high-score fragments (i.e. large fragments) are missed, PRESnovo will report low-confidence motifs, resulting in high false negative rate for peptide prediction. On the other hand, if a tandem MS spectrum contains too much noise and these interferences are not effectively removed, it will lead to the wrong motif assignment and thus increase the false positive rate [36-38]. Therefore, an efficient and effective preprocessing algorithm is needed to clean up tandem MS spectra. In this study we used ProteinLynx (Waters Co.) to remove noise or background signals from QTOF data prior to processing. ProteinLynx adopts Savitzky-Golay method to smooth the data and thus ensure fewer interfering signals are contained in the final pkl data files. While the quality of MS/MS spectra is important, the mass accuracy of MS/MS spectra does not appear to substantially impact the performance of PRESnovo, as shown in **Figure S1b**, where the search was run with alternating mass error tolerances and, except for very high mass errors (e.g. greater than 1,000 ppm), the results remained consistent.

***Application of PRESnovo in conjunction with PEAKS to identify neuropeptides in* C. maenas *and* C. borealis.**

While PRESnovo accurately predicts the sequence motif of a neuropeptide, its use in conjunction with PEAKS *de novo* software enables improved detection of neuropeptides present in real biological samples. An important characteristic of PEAKS is its ability to report the positional confidence for each amino acid in the predicted sequence [11]. This feature can be used to identify the rest of the sequence after PRESnovo predicts the motif. For instance, peptide TNFAFSPRLa shown in **Figure 2** was predicted with high confidence to possess the motif of FSPRLa (score: 0.93) by PRESnovo. Meanwhile, PEAKS reports a confident N-terminal

sequence tag TNFAFSP (as in **Table 1**) for the same peptide. By combining these two predictions, one can easily determine the sequence with manual verification.

PRESnovo and PEAKS were employed collectively to identify neuropeptides in tissue samples of green crab, *C. maenas*, and Jonah crab, *C. borealis*. These animals were chosen because they are well-characterized model organisms whose nervous systems have been previously studied using electrophysiology and immunohistochemistry [39-42]. However, while their dynamic neural networks are well-understood, there is currently no fully-sequenced genome available for either species, and so their neuropeptidomes are not yet fully characterized. Furthermore, with the inherent difficulties associated with MS characterization of neuropeptides, even detecting known neuropeptides remains challenging. To address these challenges and uncover more of the neuropeptides underlying neural modulation in these organisms, PRESnovo was employed in conjunction with PEAKS. First, PRESnovo was used to identify a motif for each precursor. The motif score threshold was set to 0.60 to filter out low-scoring motif assignments and the corresponding precursors. For high-score precursors, the PRESnovo results were compared to their corresponding PEAKS predictions. Manual sequencing was then used to combine this information and determine the final sequence for each precursor.

Brain extract from green crab was analyzed on a Waters QTOF instrument and pre-processed with ProteinLynx prior to PRESnovo analysis. The resulting data included 77 neuropeptides, summarized in **Table 2**, of which 13 sequences are putative neuropeptides that have never before been identified in green crab [43]. Of the 13 novel sequences, 7 of them have previously been identified in other invertebrate species, such as *Callinectes sapidus* and *Homarus americanus*, and all were only identified with the assistance of PRESnovo. Both brain and sinus gland tissue extracts from Jonah crab were analyzed on a Thermo Q Exactive Orbitrap

instrument, and the raw files were converted to the open-source mzXML format. After analysis with PRESnovo and PEAKS, 100 neuropeptides were identified in the brain sample, 77 of these being putative novel neuropeptides never before identified in previous literature. The repeatability across technical replicates and overlap between biological replicates from distinct brain samples are shown in **Figure S2**. A total of 76 neuropeptides were identified in the sinus glad extract, with 47 of these sequences being putative novel neuropeptides not previously identified in any species. **Table 3** summarizes the neuropeptides detected in Jonah crab that match the crustacean database, while **Table 4** lists the putative novel neuropeptides identified in Jonah crab, 24 of which were only identified with the assistance of PRESnovo. **Figures 6a** and **6b** show MS/MS spectra of novel neuropeptides identified in brain and sinus gland tissue, respectively. **Figures S3 -S5** show the MS/MS spectra of all novel neuropeptides identified. As can be seen, PRESnovo predicted the characteristic sequence motif that led to the full neuropeptide sequence shown. **Figure 6c** shows a neuropeptide identified that is present in the crustacean database but was incorrectly assigned by PEAKS, demonstrating the improvement in identification afforded by PRESnovo pre-screening. The originally-predicted sequence in PEAKS scrambled the last three amino acids, but PRESnovo was able to assign them based on mapping the fragment ions to a common sequence motif. This example demonstrates the improved accuracy offered by combining PRESnovo with PEAKS for *de novo* sequencing, as identifying fragment ions characteristic of neuropeptide sequence motifs increases the likelihood of correct identifications.

Of the neuropeptides identified in the two species, the most common families were AST-A, FMRFamide-related peptide (FaRP), RYamide, orcokinin, tachykinin, and pyrokinin. AST-A and AST-B neuropeptides are distributed throughout the nervous and neuroendocrine system of

crustaceans and have been found to be inhibitors of neuromodulation [44, 45]. Several novel peptides belonging to these families were identified in both the brain and sinus glands, indicating that these peptides may also exhibit inhibitory effects. FaRP neuropeptides have been found to have a variety of functions within the nervous and neuroendocrine system, including as autocrines, paracrines, and circulating hormones [46-48]. Therefore, it is expected that these neuropeptides would be identified in both the sinus glands where they may be released as circulating hormones and in the brain, where they may serve a more local function within the neuropil . While there have not been many studies on the function of pyrokinins in crustaceans, they were found to have an effect on the gastric mill [49]. As the brain sends neuronal projection to innervate the stomach movement of the crab, pyrokinin neuropeptides were mostly identified in the brain tissue and the putatively identified novel pyrokinin peptides may also have a role in gastric activity. Tachykinin and orcokinin peptides are also more prominent in the stomatogastric nervous system, with a variety of functions including hindgut contractions [50-52]. In this study, putative novel tachykinin and orcokinin peptides were identified in the brain that may also have a role in modulating some stomach activity. Additionally, 3 tachykinin peptides were identified in the sinus gland, indicating that they may have a different modulatory role. Putative novel RYamide neuropeptides were identified in the brain and sinus glands as well. The biological activity of RYamides is not fully understood, but they have been previously identified in neuroendocrine tissue and central neuropil, suggesting functions both locally and as circulating hormones [53, 54].

Follow-up experiments will need to be performed in order to confirm the putative peptide identifications and determine their biological activities. However, these results demonstrate the great potential PRESnovo has for both facilitating the discovery of novel neuropeptides and

improving detection coverage of the crustacean neuropeptidome identified in analyses. This application provided here indicates that PRESnovo can greatly improve the discovery of neuropeptides across a variety of other species and indicates potential for improving identification of other endogenous peptides across a variety of other sample types, provided that commonly-shared sequence motifs exist.

Conclusions

In this work, a prescreening strategy, namely PRESnovo, was developed to improve the accuracy, specificity and sensitivity of peptide identification. In conjunction with *de novo* sequencing algorithms such as PEAKS, this method is powerful for identification of highly conserved peptides such as neuropeptides. The strategy we proposed in this manuscript can be easily extended to other species of interest, provided that a well-constructed motif database is obtained. Future directions may include incorporation of more sophisticated algorithms for sequencing disulfide bond bridging peptides and peptides with motifs that are difficult to detect, such as AST-A. The software and motif database used in this work can be freely downloaded via the following link: https://www.lilabs.org/resources.

Abbreviations:

AKH/RPCH: Adipokinetic hormone/red pigment concentrating hormone; AST-A: Allatostatin A; AST-B: Allatostatin B; AVP: Arginine vasopressin; CCAP: Crustacean cardioactive peptide; FaLP: FMRFamide-related peptide

Acknowledgements

This work was supported by a National Science Foundation grant (CHE-1710140) and the National Institutes of Health (NIH) through grant 1R01DK071801. The Orbitrap instruments were purchased through the support of an NIH shared instrument grant (NIH-NCRR S10RR029531). K.D. acknowledges a predoctoral fellowship supported by the National Institutes of Health-General Medical Sciences F31 National Research Service Award (1F31GM126870-01A1) for funding. LL acknowledges a Vilas Distinguished Achievement Professorship and Charles Melbourne Johnson Professorship with funding provided by the Wisconsin Alumni Research Foundation and University of Wisconsin-Madison School of Pharmacy.

Supporting Material Available

Experimental details for crustacean tissue collection, preparation, and MS analysis, table of peptide standards used for performance analysis, and MS/MS spectra of all putative novel neuropeptides are included in the supporting material.

References

1. Hunt, D.F., Michel, H., Dickinson, T.A., Shabanowitz, J., Cox, A.L., Sakaguchi, K., Appella, E., Grey, H.M., Sette, A.: Peptides presented to the immune system by the murine class II major histocompatibility complex molecule I-Ad. Science. **256**, 1817-1820 (1992)
2. Wolters, D.A., Washburn, M.P., Yates, J.R., 3rd: An automated multidimensional protein identification technology for shotgun proteomics. Anal Chem. **73**, 5683-5690 (2001)
3. Foster, L.J., de Hoog, C.L., Zhang, Y., Xie, X., Mootha, V.K., Mann, M.: A mammalian organelle map by protein correlation profiling. Cell. **125**, 187-199 (2006)
4. Lam, H., Deutsch, E.W., Eddes, J.S., Eng, J.K., King, N., Stein, S.E., Aebersold, R.: Development and validation of a spectral library searching method for peptide identification from MS/MS. Proteomics. **7**, 655-667 (2007)
5. Craig, R., Cortens, J.C., Fenyo, D., Beavis, R.C.: Using annotated peptide mass spectrum libraries for protein identification. J Proteome Res. **5**, 1843-1849 (2006)

6. Frewen, B.E., Merrihew, G.E., Wu, C.C., Noble, W.S., MacCoss, M.J.: Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. Anal Chem. **78**, 5678-5684 (2006)
7. Eng, J.K., Mccormack, A.L., Yates, J.R.: An Approach to Correlate Tandem Mass-Spectral Data of Peptides with Amino-Acid-Sequences in a Protein Database. J Am Soc Mass Spectr. **5**, 976-989 (1994)
8. Perkins, D.N., Pappin, D.J., Creasy, D.M., Cottrell, J.S.: Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis. **20**, 3551-3567 (1999)
9. Craig, R., Beavis, R.C.: TANDEM: matching proteins with tandem mass spectra. Bioinformatics. **20**, 1466-1467 (2004)
10. Geer, L.Y., Markey, S.P., Kowalak, J.A., Wagner, L., Xu, M., Maynard, D.M., Yang, X., Shi, W., Bryant, S.H.: Open mass spectrometry search algorithm. J Proteome Res. **3**, 958-964 (2004)
11. Kong, A.T., Leprevost, F.V., Avtonomov, D.M., Mellacheruvu, D., Nesvizhskii, A.I.: MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. Nat Methods. **14**, 513-520 (2017)
12. Taylor, J.A., Johnson, R.S.: Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. Rapid Commun Mass Spectrom. **11**, 1067-1075 (1997)
13. Fischer, B., Roth, V., Roos, F., Grossmann, J., Baginsky, S., Widmayer, P., Gruissem, W., Buhmann, J.M.: NovoHMM: a hidden Markov model for de novo peptide sequencing. Anal Chem. **77**, 7265-7273 (2005)
14. Frank, A., Pevzner, P.: PepNovo: de novo peptide sequencing via probabilistic network modeling. Anal Chem. **77**, 964-973 (2005)
15. Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A., Lajoie, G.: PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. Rapid Commun Mass Spectrom. **17**, 2337-2342 (2003)
16. Frank, A.M., Savitski, M.M., Nielsen, M.L., Zubarev, R.A., Pevzner, P.A.: De novo peptide sequencing and identification with precision mass spectrometry. J Proteome Res. **6**, 114-123 (2007)
17. Kim, S., Gupta, N., Bandeira, N., Pevzner, P.A.: Spectral dictionaries: Integrating de novo peptide sequencing with database search of tandem mass spectra. Mol Cell Proteomics. **8**, 53-69 (2009)
18. Zhang, Y., DeLaney, K., Hui, L., Wang, J., Sturm, R.M., Li, L.: A Multifaceted Mass Spectrometric Method to Probe Feeding Related Neuropeptide Changes in Callinectes sapidus and Carcinus maenas. J Am Soc Mass Spectrom. **29**, 948-960 (2018)
19. DeLaney, K., Li, L.: Data Independent Acquisition Mass Spectrometry Method for Improved Neuropeptidomic Coverage in Crustacean Neural Tissue Extracts. Anal Chem. **91**, 5150-5158 (2019)
20. DeLaney, K., Li, L.: Capillary electrophoresis coupled to MALDI mass spectrometry imaging with large volume sample stacking injection for improved coverage of C. borealis neuropeptidome. Analyst. **145**, 61-69 (2019)
21. Liu, Y., Buchberger, A.R., DeLaney, K., Li, Z., Li, L.: Multifaceted Mass Spectrometric Investigation of Neuropeptide Changes in Atlantic Blue Crab, Callinectes sapidus, in Response to Low pH Stress. J Proteome Res. **18**, 2759-2770 (2019)

22. Ma, M.M., Bors, E.K., Dickinson, E.S., Kwiatkowski, M.A., Sousa, G.L., Henry, R.P., Smith, C.M., Towle, D.W., Christie, A.E., Li, L.J.: Characterization of the Carcinus maenas neuropeptidome by mass spectrometry and functional genomics. General and Comparative Endocrinology. **161**, 320-334 (2009)
23. Hui, L.M., D'Andrea, B.T., Jia, C.X., Liang, Z.D., Christie, A.E., Li, L.J.: Mass spectrometric characterization of the neuropeptidome of the ghost crab Ocypode ceratophthalma (Brachyura, Ocypodidae). General and Comparative Endocrinology. **184**, 22-34 (2013)
24. Dickinson, P.S., Armstrong, M.K., Dickinson, E.S., Fernandez, R., Miller, A., Pong, S., Powers, B.W., Pupo-Wiss, A., Stanhope, M.E., Walsh, P.J., Wiwatpanit, T., Christie, A.E.: Three members of a peptide family are differentially distributed and elicit differential state-dependent responses in a pattern generator-effector system. J Neurophysiol. **119**, 1767-1781 (2018)
25. Christie, A.E., Pascual, M.G.: Peptidergic signaling in the crab Cancer borealis: Tapping the power of transcriptomics for neuropeptidome expansion. Gen Comp Endocrinol. **237**, 53-67 (2016)
26. DeLaney, K., Buchberger, A.R., Atkinson, L., Grunder, S., Mousley, A., Li, L.: New techniques, applications and perspectives in neuropeptide research. J Exp Biol. **221**, (2018)
27. Christie, A.E., Stemmler, E.A., Dickinson, P.S.: Crustacean neuropeptides. Cell Mol Life Sci. **67**, 4135-4169 (2010)
28. Wang, Y., Wang, M., Yin, S., Jang, R., Wang, J., Xue, Z., Xu, T.: NeuroPep: a comprehensive resource of neuropeptides. Database (Oxford). **2015**, bav038 (2015)
29. Consortium, U.: UniProt: a worldwide hub of protein knowledge. Nucleic Acid Res. **47**, D506-D515 (2019)
30. Hook, V., Funkelstein, L., Lu, D., Bark, S., Wegrzyn, J., Hwang, S.R.: Proteases for processing proneuropeptides into peptide neurotransmitters and hormones. Annu Rev Pharmacol Toxicol. **48**, 393-423 (2008)
31. Allet, N., Barrillat, N., Baussant, T., Boiteau, C., Botti, P., Bougueleret, L., Budin, N., Canet, D., Carraud, S., Chiappe, D., Christmann, N., Colinge, J., Cusin, I., Dafflon, N., Depresle, B., Fasso, I., Frauchiger, P., Gaertner, H., Gleizes, A., Gonzalez-Couto, E., Jeandenans, C., Karmime, A., Kowall, T., Lagache, S., Mahe, E., Masselot, A., Mattou, H., Moniatte, M., Niknejad, A., Paolini, M., Perret, F., Pinaud, N., Ranno, F., Raimondi, S., Reffas, S., Regamey, P.O., Rey, P.A., Rodriguez-Tome, P., Rose, K., Rossellat, G., Saudrais, C., Schmidt, C., Villain, M., Zwahlen, C.: In vitro and in silico processes to identify differentially expressed proteins. Proteomics. **4**, 2333-2351 (2004)
32. DeLaney, K., Buchberger, A., Li, L.: Identification, Quantitation, and Imaging of the Crustacean Peptidome. Methods Mol Biol. **1719**, 247-269 (2018)
33. Adusumilli, R., Mallick, P.: Data Conversion with ProteoWizard msConvert. Methods Mol Biol. **1550**, 339-368 (2017)
34. Griss, J., Reisinger, F., Hermjakob, H., Vizcaino, J.A.: jmzReader: A Java parser library to process and visualize multiple text and XML-based mass spectrometry data formats. Proteomics. **12**, 795-798 (2012)
35. Wang, G., Wu, W.W., Zhang, Z., Masilamani, S., Shen, R.: Decoy Methods for Assessing False Positives and False Discovery Rates in Shotgun Proteomics. Anal Chem. **81**, 146-159 (2009)
36. Xu, H., Freitas, M.A.: A Dynamic Noise Level Algorithm for Spectral

Screening of Peptide MS/MS Spectra. BMC Bioinformatics. **11**, 1-8 (2010)
37. Sadygov, R.G., Cociorva, D., Yates, J.R.r.: Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. Nat Methods. **1**, 195-202 (2004)
38. Zhang, J., Gonzalez, E., Hestilow, T., Haskins, W., Huang, Y.: Review of Peak Detection Algorithms in Liquid-Chromatography-Mass

Spectrometry. Current Genomics. **10**, 388-401 (2009)
39. DeLaney, K., Buchberger, A.R., Atkinson, L., Grunder, S., Mousley, A., Li, L.J.: New techniques, applications and perspectives in neuropeptide research. Journal of Experimental Biology. **221**, (2018)
40. Marder, E., Gutierrez, G.J., Nusbaum, M.P.: Complicating connectomes: Electrical coupling creates parallel pathways and degenerate circuit mechanisms. Dev Neurobiol. **77**, 597-609 (2017)
41. Nusbaum, M.P., Blitz, D.M., Marder, E.: Functional consequences of neuropeptide and small-molecule co-transmission. Nat Rev Neurosci. **18**, 389-403 (2017)
42. White, R.S., Spencer, R.M., Nusbaum, M.P., Blitz, D.M.: State-dependent sensorimotor gating in a rhythmic motor system. J Neurophysiol. **118**, 2806-2818 (2017)
43. Ma, M., Bors, E.K., Dickinson, E.S., Kwiatkowski, M.A., Sousa, G.L., Henry, R.P., Smith, C.M., Towle, D.W., Christie, A.E., Li, L.: Characterization of the Carcinus maenas neuropeptidome by mass spectrometry and functional genomics. Gen Comp Endocrinol. **161**, 320-334 (2009)
44. Skiebe, P., Schneider, H.: Allatostatin peptides in the crab stomatogastric nervous system: inhibition of the pyloric motor pattern and distribution of allatostatin-like immunoreactivity. Journal of Experimental Biology. **194**, 195-208 (1994)
45. Jorge-Rivera, J., Marder, E.: Allatostatin decreases stomatogastric neuromuscular transmission in the crab Cancer borealis. J Exp Biol. **200**, 2937-2946 (1997)
46. Dockray, G.J.: The expanding family of -RFamide peptides and their effects on feeding behaviour. Exp Physiol. **89**, 229-235 (2004)
47. Bechtold, D.A., Luckman, S.M.: The role of RFamide peptides in feeding. J Endocrinol. **192**, 3-15 (2007)
48. Findeisen, M., Rathmann, D., Beck-Sickinger, A.G.: RFamide Peptides: Structure, Function, Mechanisms and Pharmaceutical Potential. (2011)
49. Saideman, S.R., Ma, M., Kutz-Naber, K.K., Cook, A., Torfs, P., Schoofs, L., Li, L., Nusbaum, M.P.: Modulation of rhythmic motor activity by pyrokinin peptides. J Neurophysiol. **97**, 579-595 (2007)
50. Li, L., Pulver, S.R., Kelley, W.P., Thirumalai, V., Sweedler, J.V., Marder, E.: Orcokinin peptides in developing and adult crustacean stomatogastric nervous systems and pericardial organs. J Comp Neurol. **444**, 227-244 (2002)
51. Christie, A.E., Cashman, C.R., Stevens, J.S., Smith, C.M., Beale, K.M., Stemmler, E.A., Greenwood, S.J., Towle, D.W., Dickinson, P.S.: Identification and cardiotropic actions of brain/gut-derived tachykinin-related peptides (TRPs) from the American lobster Homarus americanus. Peptides. **29**, 1909-1918 (2008)
52. Sousa, G.L., Lenz, P.H., Hartline, D.K., Christie, A.E.: Distribution of pigment dispersing hormone- and tachykinin-related peptides in the central nervous system of the

copepod crustacean Calanus finmarchicus. General and Comparative Endocrinology. **156**, 454-459 (2008)
53. Li, L.J., Kelley, W.P., Billimoria, C.P., Christie, A.E., Pulver, S.R., Sweedler, J.V., Marder, E.: Mass spectrometric investigation of the neuropeptide complement and release in the pericardial organs of the crab, Cancer borealis. Journal of Neurochemistry. **87**, 642-656 (2003)
54. Ma, M., Wang, J., Chen, R., Li, L.: Expanding the Crustacean neuropeptidome using a multifaceted mass spectrometric approach. J Proteome Res. **8**, 2426-2437 (2009)

Figure Legends

Figure 1. The diagram illustrates the construction of a motif database. Three resources including public databases (e.g. NCBI), previous publications and the neuropeptide discoveries by our lab are used to generate a collection of crustacean neuropeptides which are clustered into families according to their conserved motifs. Then each family of neuropeptides are aligned with WebLogo (version 3.0, http://weblogo.threeplusone.com/) to extract detailed motifs followed by *in silico* fragmentation of these extracted motifs with MS-product (http://prospector.ucsf.edu/prospector/cgi-bin/msform.cgi?form=msproduct). Finally, the motif and the corresponding b- or y- series fragments are compiled into a motif database.

Figure 2. The distribution of motifs according to their families. The number represents the total number of motifs associated with each family.

Figure 3. The schematic representation of motif assignment. The experimental fragments associated with the precursor (in .pkl file) are searched against a predefined motif database in which each motif contains y-series ions (y, y-NH_3, y-H_2O) or b-series ions (b, b-NH_3, -H_2O). The scores from the matched experimental and theoretical fragments (S1, S2, S3, S4) are used to calculate the overall score for motif assignment as shown in Output file. The motif with the highest score is assigned to the precursor.

Figure 4. The representative tandem MS spectra and PRESnovo output. A) MS/MS spectrum of TNFAFSPRLa. PRESnovo found four motif-associated fragments (y2, y3, y4 and y5) and then assigned C-terminal motif of FSPRLa to this precursor; B) MS/MS spectrum of NFDEIDRSGFGFN. PRESnovo found six motif-associated fragments (b2, b3, b4, b5, b6 and b7) and then assigned N-terminal motif of NFDEIDR to this precursor.

Figure 5. PRESnovo fails to interpret tandem MS spectra of AST-A peptide and disulfide bond bridging peptide. A) a/b/c ions dominate in MS/MS spectrum of GDGRLYAFGLa; B) incomplete motif-associated fragments in MS/MS spectrum of CYFQNCPRGa.

Figure 6. MS/MS spectra of putative novel neuropeptides detected in (A) brain and (B) sinus gland of Jonah crab, *C. borealis*, as well as a neuropeptide matching the crustacean neuropeptide database that PEAKS failed to assign correctly. In all 3 cases, PRESnovo was able to identify fragment ions indicative of common motifs.

Table 1. The identified peptide standards by PRESnovo and PEAKS 7, with the displayed sequence being the output from each software that most closely matches the correct sequence.

Family	Mass	Real Sequence	PRESnovo (Motif)	PEAKS 7
CCAP	955.37	PFCNAFTGCa	FIRFa	MCSAAGACAAT
RYamide	975.44	SGFYANRYa	**FYANRYa**	GRNTAAGGGDT
Pyrokinin	1036.53	SGGFAFSPRLa	**FSPRLa**	**SGGFA**FM(O)AFFFGT
Pyrokinin	1050.55	TNFAFSPRLa	**FSPRLa**	GS**FAFSP**VGIa
AST-A	1066.54	GDGRLYAFGLa	FGPRLa	TGGAQPEQLPAa
AVP	1083.44	CYFQNCPRGa	NSELINSILG	M(O)SQEPAAHM(O)**Ga**
FLP	1104.57	GAHKNYLRF	**NYLRFa**	HQGAGGVPMRPa
AST-B	1106.50	QWSSMRGAWa	**WSSMRGAWa**	TGRSSGAAAGADS
AST-B	1259.64	SGKWSNLRGAWa	**WSNLRGAWa**	QCARSVAGGSASAPa
Angiotension	1281.65	DRVYVHPFHL	**DRVYVHPF**	SAGPVEGGDLMLH
AST-A	1334.70	APSGAQRLYGFGLa	pQVNFSPNWa	**APSGA**TCSSMGVGVLa
AST-B	1469.68	VPNDWAHFRGSWa	**WAHFRGSWa**	VMDLGSAGSGNGMQM(O)
Orcokinin	1472.66	NFDEIDRSGFGFAa	**NFDEIDR**	**NFDELD**AG**SG**S**FG**GT
Orcokinin	1473.64	NFDEIDRSGFGFA	**NFDEIDR**	CTG**DELD**AGPGGAPGGT
Orcokinin	1516.65	NFDEIDRSGFGFN	**NFDEIDR**	GCT**DE**NNNM(O)EGGCGT

*the bold font type refers to the correct identifications. a: amide. The underlined residues are the correctly identified by PEAKS 7. 11 were identified by PRESnovo while 1 by PEAKS 7 (NFDEIDRSGFGFAa, based on criteria of six consecutively correct residues)

Table 2. Identified neuropeptides in brain tissue of green crab, *C. maenas*

Family	Mass	Sequence
AST-A	768.38	EAYAGFLa
AST-A	779.39	NPYAFGLa
AST-A	779.39	GGPYAFGLa
AST-A	780.38	DPYAFGLa
AST-A	793.41	AGPYAFGLa
AST-A	794.39	EPYAFGLa
AST-A	807.43	AAPYAFGLa
AST-A	809.40	AGPYSFGLa
AST-A	823.42	ASPYAFGLa
AST-A	850.47	GKPYAFGLa
AST-A	852.40	EPYEFGLa
AST-A	878.47	RGPYAFGLa
AST-A	896.44	FSGASPYGLa
AST-A	908.48	ARPYSFGLa
AST-A	924.50	LKAYDFGLa
AST-A	925.46	ATGQYAFGLa
AST-A	938.49	TRPYSFGLa
AST-A	922.52	KLPYSFGLa
AST-B	1106.57	QWSSMRGAWa
AST-B	1259.70	SGKWSNLRGAWa
AST-B	1292.62	STNWSSLRSAWa
AST-B	1469.69	VPNDWAHFRGSWa
CCAP	956.38	PFCNAFTGCa
FaRPs	734.40	GPFLRFa
FaRPs	850.49	RNFLRFa
FaRPs	886.55	PSLRLRFa
FaRPs	904.50	PSMRLRFa
FaRPs	920.50	PSM(O)RLRFa
FaRPs	937.52	NRSFLRFa
FaRPs	953.52	SRNYLRFa
FaRPs	964.53	NRNFLRFa
FaRPs	965.52	DRNFLRFa
FaRPs	976.51	PQGNFLRFa
FaRPs	1021.55	GNRNFLRFa
FaRPs	1022.53	GDRNFLRFa
FaRPs	1047.56	APQGNFLRFa
FaRPs	1103.60	GAHKNFLRFa
FaRPs	1104.62	SMPSLRLRFa

FaRPs	1123.62	GLSRNYLRFa
FaRPs	1136.58	DGNRNFLRFa
FaRPs	1157.61	YGNRSFLRFa
FaRPs	1207.62	DQNRNFLRFa
FaRPs	1270.64	pQDLDHVFLRFa
FaRPs	1270.67	PELDHVFLRFa
FaRPs	1287.67	QDLDHVFLRFa
FaRPs	1288.62	QDNDHVFLRFa
FaRPs	1313.77	DARTAPLRLRFa
Orcokinin	936.42	DEIDRSGFa
Orcokinin	1197.54	NFDEIDRSGFa
Orcokinin	1227.55	NFDEIDRSSFa
Orcokinin	1255.54	NFDEIDRSGFG
Orcokinin	1269.55	NFDEIDRSGFA
Orcokinin	1285.55	NFDEIDRSSFG
Orcokinin	1299.57	NFDEIDRSSFA
Orcokinin	1473.65	NFDEIDRSGFGFA
Orcokinin	1546.67	NFDEIDRSSFGFN
Orcomyotropin	1185.51	FDAFTTGFGHS
Others	843.47	HIGSLYRa
Others	915.53	KIFEPLVA
Others	1371.78	KIFEPLRDKNL
PDH	1926.01	NSELINSLLGIPKVMNDAa
Pyrokinin	877.51	LYFAPRLa
Pyrokinin	1023.55	TSFAFSPRLa
Pyrokinin	1108.56	TDGFAFSPRLa
RPCH	929.43	pQLNFSPGWa
SIFamide	1160.64	RKPPFNGSIFa
SIFamide	1380.73	GYRKPPFNGSIFa
Tachykinin	765.39	SGFLGMRa
Tachykinin	862.45	PSGFLGMRa
Tachykinin	878.44	PSGFLGM(O)Ra
Tachykinin	933.48	APSGFLGMRa
Tachykinin	934.46	APSGFLGMR
Tachykinin	949.48	APSGFLGM(O)Ra
Tachykinin	963.49	TPSGFLGMRa
Tachykinin	979.49	TPSGFLGM(O)Ra
Tachykinin	991.49	APSGFLGMRG
Tachykinin	1007.48	APSGFLGM(O)RG

*Red color represents novel neuropeptides identified in green crab. “a” indicates C-terminal amidation. “p” indicates pyroglutamate Gln. “O” stands for oxidation of Met.

Table 3. Identified neuropeptides matching to the database in brain and sinus gland tissue of Jonah crab, *C. borealis*

Family	Mass	Sequence	Tissue
AST-B	1106.5081	AGWSSMRGAWa	Brain, SG
AST-B	1106.5081	QWSSMRGAWa	Brain, SG
AST-B	1292.6262	STNWSSLRSAWa	Brain, SG
Corazonin	1368.6211	pQTFQYSRGWTNa	Brain
FaRP	1146.6411	APQRNFLRFa	SG
FaRP	965.5195	DRNFLRFa	Brain, SG
FaRP	961.5246	ERNFLRFa	Brain
FaRP	1103.5989	GAHKNYLRFa	SG
FaRP	1021.5569	GNRNFLRFa	Brain, SG
FaRP	1145.5981	GYSKNYLRFa	Brain, SG
FaRP	816.4758	HVFLRFa	Brain
FaRP	1103.6353	KHKNYLRFa	Brain
FaRP	694.3915	NFLRFa	Brain, SG
FaRP	964.5355	NRNFLRFa	Brain, SG
FaRP	1270.6458	pQDLDHVFLRFa	Brain, SG
FaRP	1287.6724	QDLDHVFLRFa	Brain, SG
FaRP	850.4926	RNFLRFa	Brain
FaRP	1180.6101	SENRNFLRFa	Brain, SG
Proctolin	648.3595	RYLPT	SG
Pyrokinin	835.4704	FAFSPRLa	Brain
Pyrokinin	877.5174	LYFAPRLa	Brain, SG
Pyrokinin	1036.5454	SGGFAFSPRLa	SG
Pyrokinin	1050.561	TNFAFSPRLa	Brain
Ryamide	1029.4668	EGFYSQRYa	SG
RYamide	783.4028	FVGGSRYa	SG
RYamide	861.4133	FYSQRYa	SG
RYamide	1113.5679	RSSFVGGSRYa	SG
RYamide	975.4562	SGFYANRYa	Brain, SG
Ryamide	1113.5679	SSRFVGGSRYa	SG
Tachykinin	933.4854	APSGFLGMRa	Brain, SG
Tachykinin	621.3421	FLGMRa	SG
Tachykinin	765.3956	SGFLGMRa	Brain, SG
Tachykinin	963.496	TPSGFLGMRa	Brain, SG

"a" represents C-terminal amidation. "p" means pyroglutamate Gln or Glu. "SG" indicates sinus gland.

Table 4. Identified putative novel neuropeptides in the brain and sinus gland tissue of Jonah crab, *C. borealis*

Family	Mass	Sequence	Tissue
AST-A	1065.55	APTDLYAFGLa	Brain
AST-A	1065.55	PATDLYAFGLa	Brain
AST-A	983.4824	QRDYSFGLa	Brain
AST-A	939.4926	RQAYSFGLa	SG
AST-B	1259.652	KGSWSNLRGAWa	Brain
AST-B	1292.608	MGNWSSLRSAWa	Brain
FaRP	964.5242	AAQNFLRFa	SG
FaRP	1103.599	AGHKNYLRFa	SG
FaRP	1146.641	APAGRNFLRFa	Brain
FaRP	1146.641	APGARNFLRFa	Brain, SG
FaRP	1146.641	APGRANFLRFa	Brain, SG
FaRP	1006.535	APGSNFLRFa	Brain
FaRP	976.5242	APNNFLRFa	Brain
FaRP	1267.624	CAENRNFLRFa	Brain
FaRP	1267.607	CCPGGRNFLRFa	Brain
FaRP	1267.607	CCPNRNFLRFa	Brain
FaRP	1324.646	DGMGNRNFLRFa	Brain
FaRP	1270.639	DHVCHVFLRFa	SG
FaRP	1394.567	DSGPDDYGHMRFa	SG
FaRP	1394.567	DSPGDDYGHMRFa	Brain
FaRP	1180.61	DTNRNFLRFa	SG
FaRP	1394.567	pEGTSDDYGHMRFa	SG
FaRP	1267.624	EACNRNFLRFa	Brain
FaRP	1222.621	EERNNFLRFa	Brain
FaRP	1532.762	EESAEVPPNFLRFa	Brain
FaRP	1022.53	EGAANFLRFa	Brain
FaRP	1233.589	EQANDNFLRFa	Brain
FaRP	1769.896	EQQPHAGLSAGNFLRFa	Brain
FaRP	1180.61	ESNRNFLRFa	SG
FaRP	1180.61	ESRNNFLRFa	Brain, SG
FaRP	1324.664	ESSNGRNFLRFa	Brain
FaRP	1032.598	FALAGRPRFa	Brain
FaRP	1180.614	FGAPNNFLRFa	Brain
FaRP	1103.599	HAGKNYLRFa	Brain, SG
FaRP	1270.639	HDVCHVFLRFa	Brain
FaRP	1146.594	HEVSNFLRFa	Brain

FaRP	1249.647	HFDRNFLRFa	Brain
FaRP	1248.663	HFNRNFLRFa	Brain
FaRP	1146.594	HSDLNFLRFa	Brain
FaRP	1532.832	KAAPSNRNNFLRFa	Brain
FaRP	1146.666	KAPRNFLRFa	SG
FaRP	1498.746	KCSTDGRGNFLRFa	Brain
FaRP	1146.666	KGAPVNFLRFa	SG
FaRP	966.5399	KGSNFLRFa	Brain
FaRP	1103.599	KHAGNYLRFa	SG
FaRP	1103.599	KHQNYLRFa	SG
FaRP	1146.666	KLPNNFLRFa	Brain
FaRP	1146.678	KPARNFLRFa	SG
FaRP	1248.709	KQQLGNFLRFa	Brain
FaRP	1146.678	KRAPNFLRFa	SG
FaRP	1305.749	KRMVPNFLRFa	Brain
FaRP	1145.598	KSGYNYLRFa	Brain, SG
FaRP	1333.737	KSPNGRNFLRFa	Brain
FaRP	1145.598	KSYGNYLRFa	SG
FaRP	1146.666	KVPQNFLRFa	SG
FaRP	1333.737	LQAGNRNFLRFa	Brain
FaRP	1021.553	MVPNFLRFa	Brain
FaRP	976.5242	NPANFLRFa	Brain
FaRP	992.5192	NSPNFLRFa	Brain
FaRP	1544.748	NSYSERNNFLRFa	Brain
FaRP	1146.641	PAGARNFLRFa	SG
FaRP	976.5242	PAGGNFLRFa	Brain
FaRP	1006.535	PNTNFLRFa	Brain
FaRP	1146.605	pQGQRNFLRFa	SG
FaRP	1222.625	pQQAAFNFLRFa	Brain
FaRP	1551.78	PVMEMRNNFLRFa	Brain
FaRP	1103.599	QHKNYLRFa	Brain, SG
FaRP	1180.599	QKDDNFLRFa	Brain
FaRP	1333.737	QLAGNRNFLRFa	Brain
FaRP	1146.641	QPARNFLRFa	SG
FaRP	1248.695	QRNRNFLRFa	Brain
FaRP	1146.641	RAGAPNFLRFa	SG
FaRP	965.5195	RDNFLRFa	Brain, SG
FaRP	964.5355	RGGNFLRFa	Brain, SG
FaRP	1021.557	RGNNFLRFa	SG
FaRP	1180.61	RNDTNFLRFa	Brain
FaRP	1180.61	RNESNFLRFa	Brain, SG

FaRP	1021.557	RNGNFLRFa	Brain, SG
FaRP	964.5355	RNNFLRFa	Brain, SG
FaRP	1180.61	RNSENFLRFa	SG
FaRP	3692.954	RPGQLLLAEASSWLPTQQEGTKRGYSKNYLRFa	Brain
FaRP	1248.695	RQNRNFLRFa	Brain
FaRP	1276.679	SAGPNRNFLRFa	Brain
FaRP	1219.658	SAPNRNFLRFa	Brain
FaRP	1394.567	SDGPDDYGHMRFa	Brain
FaRP	1180.61	SERNNFLRFa	SG
FaRP	1483.71	STDEPYPNFLRFa	Brain
FaRP	1180.599	STPNSNFLRFa	SG
FaRP	1180.61	TDRNNFLRFa	SG
FaRP	1180.635	TSAQVNFLRFa	SG
FaRP	1544.729	TSDELTTCNFLRFa	Brain
FaRP	1180.599	TSPNSNFLRFa	SG
FaRP	1180.599	TSSNPNFLRFa	SG
FaRP	1532.836	VPGFAPNRNFLRFa	Brain
FaRP	1004.523	YFNFLRFa	Brain
FaRP	1145.598	YGSKNYLRFa	SG
Pyrokinin	2380.269	FNGPKPLAKYVDTNFAFSPRLa	Brain
Pyrokinin	2379.252	FNPGKLPKSQMTTNFAFSPRLa	Brain
Pyrokinin	1036.545	GSGFAFSPRLa	Brain
Pyrokinin	1050.561	QSFAFSPRLa	Brain, SG
RYamide	1029.467	pEGFYSQRYa	SG
RYamide	975.4562	GSFYANRYa	Brain, SG
RYamide	1029.467	pQGFYSQRYa	Brain
RYamide	2133.035	SADRTQLTERSGFYANRYa	Brain
RYamide	1113.568	SRSFVGGSRYa	SG
RYamide	2133.035	TGARDGTLTERSGFYANRYa	Brain
Tachykinin	1143.622	APPLSGFLGMRa	SG
Tachykinin	1175.623	KNAPSGFLGMRa	Brain
Tachykinin	1145.613	KSAHTFLGMRa	SG
Tachykinin	1145.576	KSDHGFLGMRa	SG
Tachykinin	1253.547	NCCAPSGFLGMRa	Brain

“a” represents C-terminal amidation. “p” means pyroglutamate Gln or Glu. “SG” indicates sinus gland.

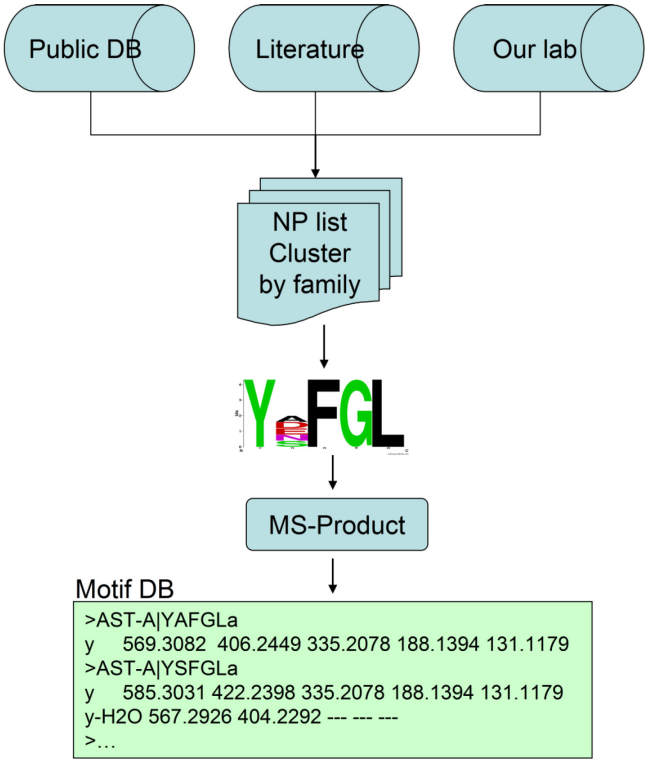

Figure 1. The diagram illustrates the construction of a motif database. Three resources including public databases (e.g. NCBI), previous publications and the neuropeptides discovered by our lab are used to generate a collection of crustacean neuropeptides which are clustered into families according to their conserved motifs. Then each family of neuropeptides are aligned with WebLogo (version 3.0, http://weblogo.threeplusone.com/) to extract detailed motifs followed by *in silico* fragmentation of these extracted motifs with MS-product (http://prospector.ucsf.edu/prospector/cgi-bin/msform.cgi?form=msproduct). Finally, the motif and the corresponding b- or y- series fragments are compiled into a motif database.

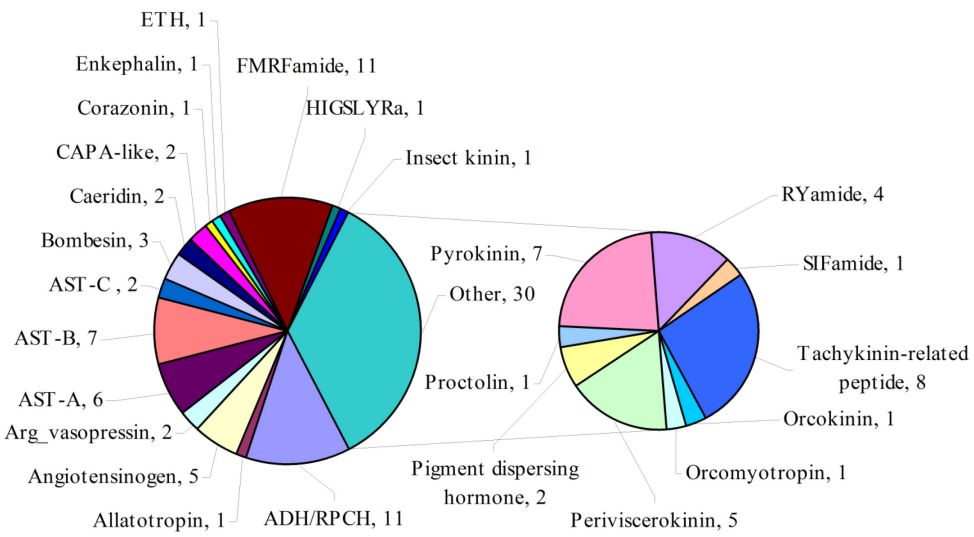

Figure 2. The distribution of motifs according to their families. The number represents the total number of motifs associated with each family.

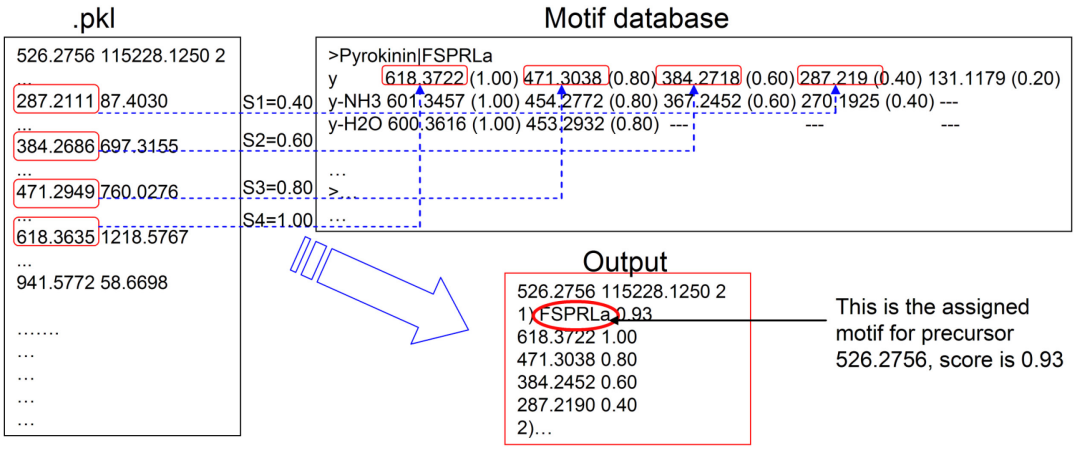

Figure 3. The schematic representation of motif assignment. The experimental fragments associated with the precursor (in .pkl file) are searched against a predefined motif database in which each motif contains y-series ions (y, y-NH_3, y-H_2O) or b-series ions (b, b-NH_3, b-H_2O). The scores from the matched experimental and theoretical fragments (S1, S2, S3, S4) are used to calculate the overall score for motif assignment as shown in Output file. The motif with the highest score is assigned to the precursor.

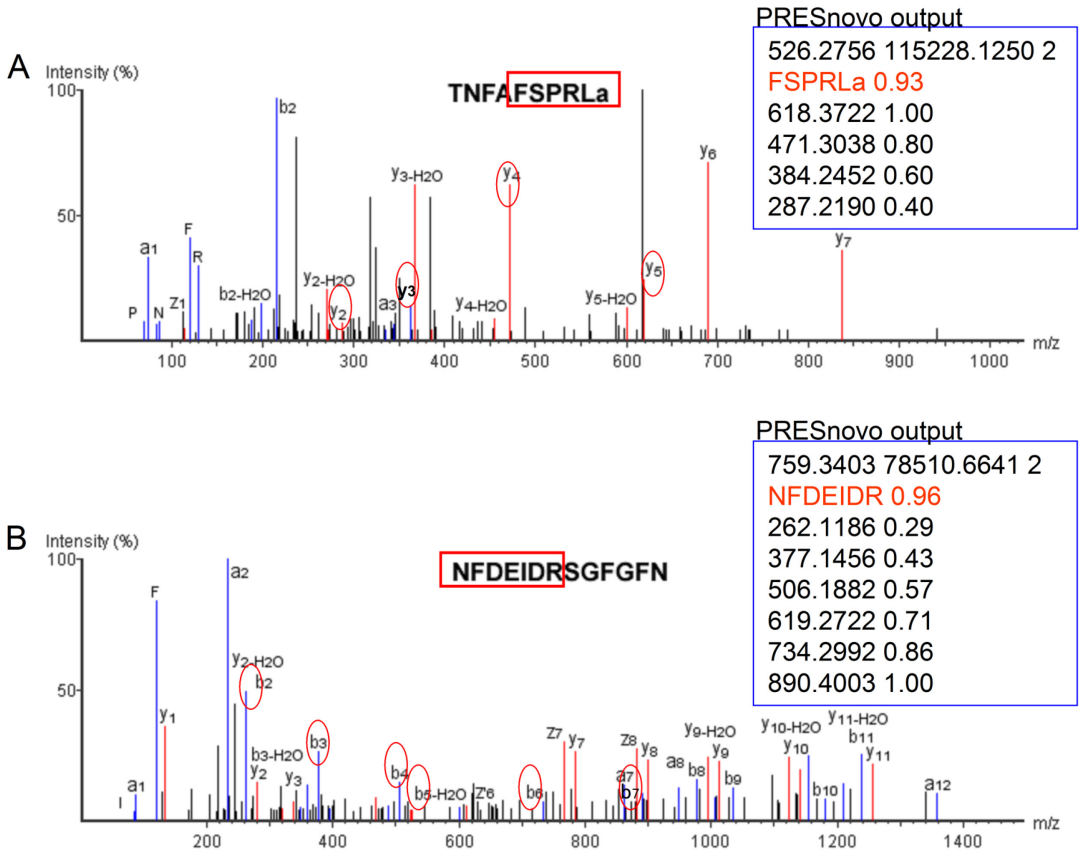

Figure 4. The representative tandem MS spectra and PRESnovo output. A) MS/MS spectrum of TNFAFSPRLa. PRESnovo found four motif-associated fragments (y2, y3, y4 and y5) and then assigned C-terminal motif of FSPRLa to this precursor; B) MS/MS spectrum of NFDEIDRSGFGFN. PRESnovo found six motif-associated fragments (b2, b3, b4, b5, b6 and b7) and then assigned N-terminal motif of NFDEIDR to this precursor.

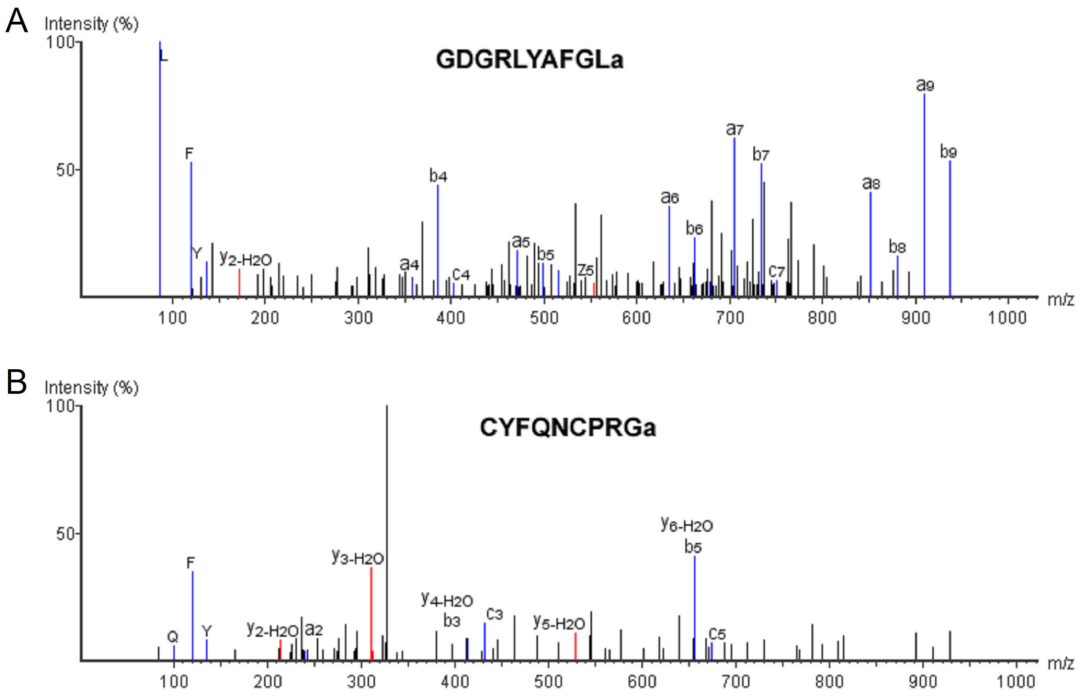

Figure 5. PRESnovo fails to interpret tandem MS spectra of AST-A peptide and disulfide bond bridging peptide. A) a/b/c ions dominate in MS/MS spectrum of GDGRLYAFGLa; B) incomplete motif-associated fragments in MS/MS spectrum of CYFQNCPRGa.

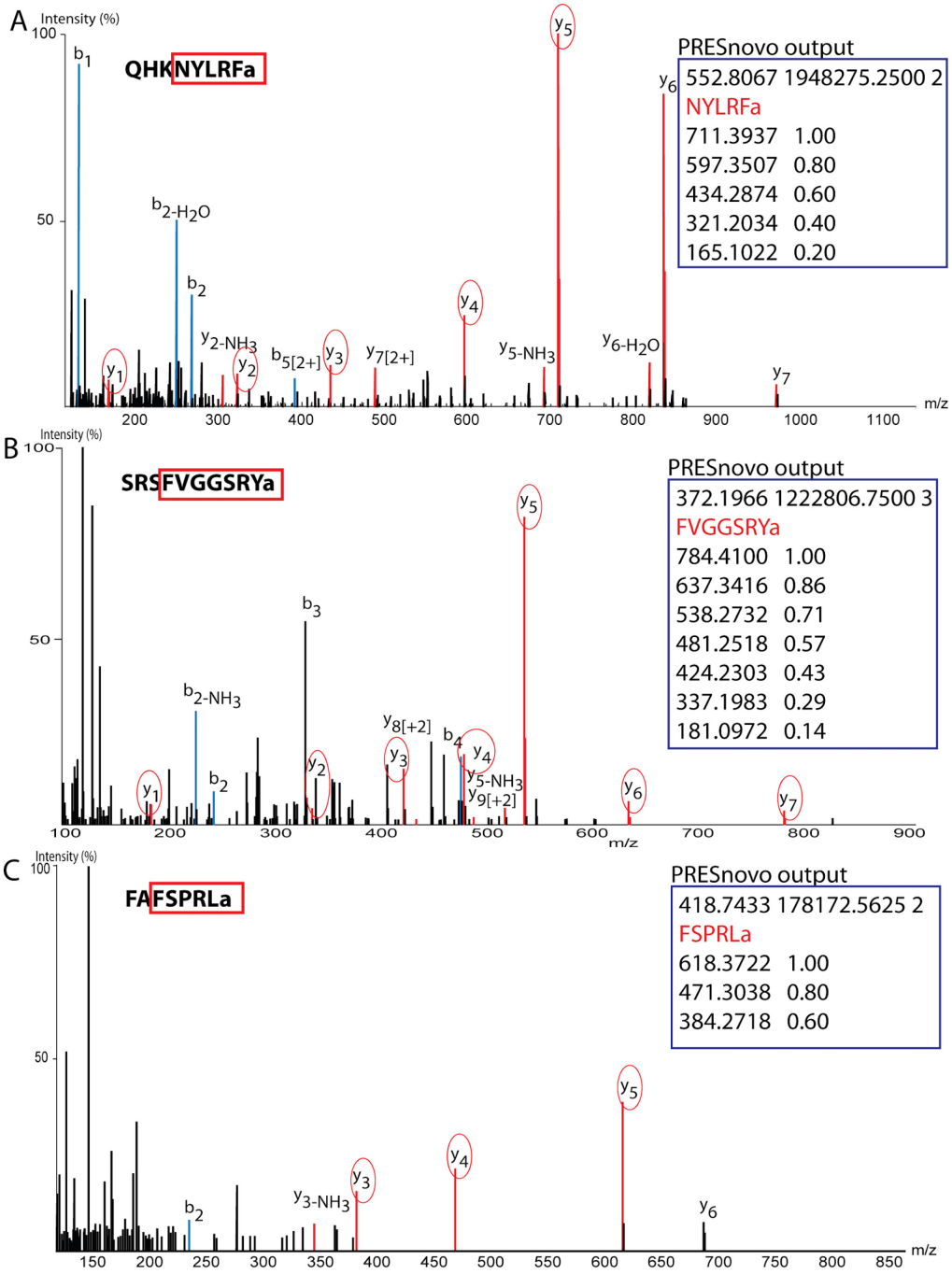

Figure 6. MS/MS spectra of putative novel neuropeptides detected in (A) brain and (B) sinus gland of Jonah crab, *C.borealis*, as well as a neuropeptide matching the crustacean neuropeptide database that PEAKS failed to assign correctly. In all 3 cases, PRESnovo was able to identify fragment ions indicative of common motifs.

For Graphical Abstract Only