Published in final edited form as:
J Proteome Res. 2023 February 03; 22(2): 420–431. doi:10.1021/acs.jproteome.2c00597.

HyPep: An Open-Source Software for Identification and Discovery of Neuropeptides using Sequence Homology Search

Nhu Q. Vu[1,¥], **Hsu-Ching Yen**[2,¥], **Lauren Fields**[1], **Weifeng Cao**[1], **Lingjun Li**[1,3,*]

[1]Department of Chemistry, University of Wisconsin-Madison, 1101 University Avenue, Madison, WI 53706, USA

[2]Department of Biochemistry, University of Wisconsin-Madison, 433 Babcock Drive, Madison, WI 53706, USA

[3]School of Pharmacy, University of Wisconsin-Madison, 777 Highland Avenue, Madison, WI 53705, USA

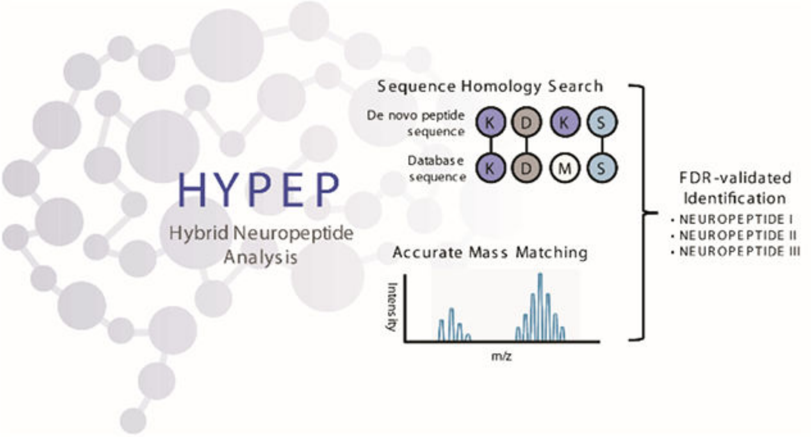
Abstract

Neuropeptides are a class of endogenous peptides that have key regulatory roles in biochemical, physiological, and behavioral processes. Mass spectrometry analyses of neuropeptides often rely on protein informatics tools for database searching and peptide identification. As neuropeptide databases are typically experimentally built and comprised of short sequences with high sequence similarity to each other, we developed a novel database searching tool, HyPep, which utilizes sequence homology searching for peptide identification. HyPep aligns database-free, *de novo* sequenced peptide sequences generated through PEAKS software with neuropeptide database sequences to identify neuropeptides based on an alignment score. HyPep performance was optimized using LC-MS/MS measurements of peptide extracts from various *C. sapidus* neuronal tissue types and compared with a commercial database searching software, PEAKS DB. HyPep identified more neuropeptides from each tissue type than PEAKS DB at 1% false discovery rate and the false match rate from both programs was 2%. In addition to identification, this report describes how HyPep can aid in the discovery of novel neuropeptides.

Graphical Abstract

[*]To whom correspondence should be addressed. lingjun.li@wisc.edu. Phone: (608)-265-8491, Fax: (608)-262-5345. Mailing Address: 5125 Rennebohm Hall, 777 Highland Avenue, Madison, WI 53705, USA.
[¥]These authors contributed equally

Supporting Information:
- Excel. (.XLSX): Tables of the neuropeptide identifications for the brain, SG, PO, and CoG tissue types from HyPep using the original database (Supplemental File 1), from PEAKS DB using the original database (Supplemental File 2), from HyPep using the entrapment database (Supplemental File 3) and from PEAKS DB using the entrapment database (Supplemental File 4).
- Word (.DOCX): HyPep user manual (Supplemental File 5).
- Word (.DOCX): (Supplemental Figures) - Distribution of sequence lengths from target database (Figure S1), description of decoy database generation methods (Figure S2), distribution of sequence lengths from decoy databases (Figure S3), and plot showing % sequence identity between members of the crustacean hyperglycemic hormone neuropeptide superfamily (Figure S4).
- Excel. (.CSV): In-house crustacean database (Supplemental File 6) and entrapment (crustacean and non-crustacean/trap) database (Supplemental File 7)



Keywords

Neuropeptide; HyPep; homology; peptide; mass spectrometry; *de novo* sequencing; FDR; peptidomics

Introduction:

Neuropeptides are signaling molecules that are expressed by neurons and have been implicated in regulation of numerous biochemical pathways, [1,2] physiological processes,[3,4] and behaviors.[5,6] Precursor peptides (>90 amino acids) originating from nuclear DNA undergo selective cleavage to form mature, bioactive neuropeptides.[7,8] The resulting neuropeptides are typically 3-36 amino acids long, and the function of a neuropeptide sequence can be altered by different post-translational modifications, or by truncations or cyclization.[9–11] Identification and discovery of neuropeptides is important for elucidation of their functional mechanism, as well as establishing a foundation for disease therapeutics research.[12–17] A crustacean animal model is often utilized for neurobiology research due to their well-characterized nervous system, which facilitates the elucidation of neuropeptide function at the neuronal circuit and systems levels.[18] Mass spectrometry (MS) is well-suited for characterizing the full complement of neuropeptides due to its high sensitivity and ability to capture the expression profile of many neuropeptides simultaneously.[19–22] Since neuropeptides undergo multiple post-transcriptional and post-translational processing events before arriving at their mature form, neuropeptide databases constructed from nucleotide sequence predictions capture only a fraction of the bioactive neuropeptide sequences.[2,7,23] As a result, many neuropeptide databases utilized within proteomics software are experimentally built to contain *de novo* sequenced neuropeptides as well as neuropeptide sequences predicted from genomic or transcriptomic information.[20,24,25]

Although most commercial proteomics software have been developed and optimized for bottom-up proteomics, they are often used for performing neuropeptide identification.[26–28] The two main software approaches include a peptide-spectrum matching (PSM) approach where raw MS and MS/MS data are compared with *in silico* MS and MS/MS values generated from a known peptide database.[26,29–32] This may involve utilizing a spectral library from reference MS/MS datasets instead of theoretically generated MS/MS data.[33,34] The other approach involves processing the MS dataset with a *de novo* sequencing algorithm, such as PEAKS, and comparing the *de novo* sequenced peptides with database sequences.[35–37] One of the main differences between neuropeptide databases and typical protein databases is that mature neuropeptide sequences are shorter than protein sequences due to the several proteolytic processing steps involved during neuropeptide biosynthesis.[38–41] This is observed within our in-house crustacean neuropeptide database where most sequences are less than 20 amino acids long (Figure S1).[6] Another unique database feature arises from the alternative splicing that occurs during neuropeptide biosynthesis, resulting in a neuropeptide database comprised of sequences with high similarity to each other.[38–40,42–44] Additionally, due to the low neuropeptide expression levels *in vivo*, it is common for the neuropeptide fragment ion abundance within MS/MS spectra to be lower than that observed within bottom-up proteomics datasets.[9,45] Considering these shortcomings, low neuropeptide identification rates are observed when proteomics software is used for neuropeptide data analysis.[20,46–48]

Development of informatic methodologies optimized for neuropeptides and equipped to address these shortcomings include IggyPep, PRESNovo and NeuroPedia.[38,39,50,51] A review of these methods was presented by Phetsanthad, et al.[39] IggyPep and PRESNovo both strategically leverage neuropeptide homology or motif searching to assist in peptide identification, as consideration of these features has been reported to result in higher success rates when identifying homologs using smaller databases.[40,51] HyPep is a novel database searching software that utilizes a sequence homology search approach for neuropeptide identification (Figure 1). First, LC-MS/MS data is processed through the PEAKS software *de novo* sequencing algorithm for database-free *de novo* sequencing. These *de novo* sequenced peptide queries are matched with database sequences and the overall match is scored based on the sum of subscores from four local alignment strategies. These local alignments consist of fixed and varying local alignments (Figure 2), and users may control the level of alignment stringency by changing the sliding window size (SWS) parameter within the varying-based local alignments.[52] It is worth noting that this scoring system treats all mismatches equally regardless of structural similarity or amino acid related-ness (*i.e.*, glutamine and glutamic acid). Weighted scoring systems, such as BLOSUM62, are robust because they were built upon empirically determined probabilities of amino acid substitutions using a comprehensive protein database.[53] Until this breadth of information is available for neuropeptides, it is premature to implement an identical scoring system.

After sequence alignment, matching and scoring occurs within the sequence homology search module, matches between the *de novo* sequenced peptide query and neuropeptide database sequences are subsequently filtered based on its match score according to the user-defined false discovery rate (FDR) threshold. Within HyPep, there are four target-decoy options for calculating the FDR, which include the reverse, shuffle, random, and a novel target-decoy method, hybrid, which contains characteristics from both shuffle and random methods. Naturally, all perfect matches between *de novo* sequenced query and database sequences are reported in the final HyPep output. Imperfect sequence matches above the specified FDR threshold are subsequently verified within HyPep by searching the raw MS/MS data file and reporting the scan(s) where the isolated precursor mass and fragment masses match with the theoretical database sequence mass. Since PEAKS software was used for generation of the *de novo* sequenced peptide queries used as the input for HyPep, the PEAKS database searching (PEAKS DB) algorithm was also used to compare neuropeptide identifications from *Callinectes sapidus* (blue crab) neuronal tissue types with HyPep, as they use the same input information. It is worth noting that the genome for this species was recently assembled, but the information was not able to be utilized in this study.[54] Although the Comparisons between HyPep and PEAKS DB identifications at 1% FDR showed that HyPep identified sequences that were shorter, contained more neuropeptides originating from the same neuropeptide family, and overall greater number of identifications at the same false match rate (FMR).

Experimental Section:

Materials

Methanol (MeOH), acetonitrile (ACN), glacial acetic acid (GAA), ammonium bicarbonate, formic acid (FA), and all crab saline components were purchased from Fisher Scientific (Pittsburgh, PA). All water (H_2O) used in this study was either of HPLC grade or doubly distilled on a Millipore filtration system (Burlington, MA), and C18 Ziptips were purchased from Millipore (Burlington, MA). All LC solvents were of Fisher Optima Grade.

Animals

All female blue crabs, *Callinectes sapidus*, were purchased from Midway Asian Foods (Madison, WI) and housed in artificial seawater at 35 parts per thousand (ppt) salinity, 13-16 °C, and 8-10 parts per million (ppm) (~80-100%) O_2. Crabs were anesthetized on ice for 20 minutes and sacrificed for the collection of brain, sinus glands (SG), pericardial organs (PO), and commissural ganglia (CoG) as previously described.[55] All dissections were performed in chilled (~10 °C) physiological saline (composition: 440 mM NaCl; 11 mM KCl; 13 mM $CaCl_2$; 26 mM $MgCl_2$; 10 mM Trizma acid; and pH 7.4, adjusted with NaOH).

NanoLC-ESI-Orbitrap Analysis of Tissue Samples

For each tissue type, 3 tissue samples were pooled together before sample processing. Tissues were extracted for neuropeptides using 90/9/1 (v/v/v) MeOH/H_2O/GAA and desalted using Millipore Ziptips. Peptide extract was reconstituted in 0.1% FA in water and loaded onto a 15 cm capillary (75 μm i.d.) packed using 1.7 μm diameter Ethylene Bridged Hybrid C18 material with the integrated emitter tip in line with the instrument inlet. Untargeted neuropeptide profiling LC-MS/MS measurements were conducted on Thermo Q Exactive HF equipped with a Dionex Ultimate 3000 system. Mobile phase A was 0.1% FA in H_2O and mobile phase B was 0.1% FA in can. Peptides were separated with a gradient elution of 10 to 20% B over 70 min and 20 to 95% B over 20 min at a flow rate of 300 nL/min. Full MS scans were acquired in profile mode ranging from *m/z* 200 to 2000 at a resolution of 60 K. Automatic gain control (AGC) target was 1×10^6, and maximum injection time was 250 ms. Tandem mass spectra were acquired in centroid mode. The top 10 most abundant precursor ions were selected for higher-energy collisional dissociation (HCD) fragmentation with a dynamic exclusion of 30 s. Data-dependent acquisition (DDA) parameters were set as resolution power of 15 K, isolation window of 2.0 Th, normalized collision energy (NCE) of 30, the maximum injection time of 120 ms, AGC target of 2 × 10^5, and fixed first mass of *m/z* 100. Each sample was injected in triplicate.

Peptide Identification and Discovery using HyPep

The algorithm and GUI were written in Python. The program is compatible with Python 3 and was validated with Python v.3.10. HyPep is open-source and freely available at https://github.com/lingjunli-research/HyPep-v1.0, with a user manual and tutorial included. A schematic of the HyPep workflow is shown in Figure 1. Thermo RAW files from LC-MS/MS measurements were *de novo* sequenced using PEAKS software (Bioinformatics Solutions Inc). PEAKS parameters were parent mass error tolerance = 20.0 ppm, fragment mass error tolerance 0.02 Da, enzyme = no digestion, variable modifications: amidation (−0.98 Da), oxidation (M) (+15.99 Da), pyro-Glu from E (−18.01 Da), pyro-Glu from Q (−17.03 Da), and max variable post-translational modifications (PTM) per peptide = 3. *De novo* sequenced peptides were filtered for average local confidence (ALC) > 50 and exported as *de novo peptides.csv*. This .csv file was loaded along with the neuropeptide database files into HyPep for processing. The first neuropeptide database input is a .csv file containing each neuropeptide sequence annotated with known PTMs and monoisotopic $[M+H]^+$ mass. The second neuropeptide database input includes a folder that contains a .txt file of the theoretical b- and y-series fragment ions for each neuropeptide database sequence. Theoretical masses were generated from ProteinProspector (https://prospector.ucsf.edu/prospector/mshome.htm). A sequence homology search (SHS) was performed by matching *de novo* sequenced peptides with neuropeptide database sequences using a local alignment strategy. Matches, or identifications, were scored using the SHS scoring function (Figure 2). Theoretical false positives below a user-defined FDR were removed from the identifications list by implementing a target-decoy method. As the SHS database searches of the target and decoy databases occur separately, the FDR is calculated as the number of hits to the decoy database (N_{decoy}) divided by the total number of hits to the target (N_{target}) (Equation 1).[56]

$$FDR = \frac{N_{decoy}}{N_{target}} \quad \text{Equation 1}$$

Identifications above the user-defined % FDR threshold from the SHS module were then verified through the included accurate mass matching (AMM) module. Both SHS and AMM modules are automatically performed in each HyPep run. AMM was performed at the peptide precursor and fragment level at the same error tolerances as those used for PEAKS *de novo* sequencing (20 ppm for precursor mass and 0.02 Da for fragment mass) to verify that HyPep identifications with a score less than 4 (*i.e.*, an imperfect sequence alignment) were present in at least one MS/MS spectrum within the Thermo RAW file.

To determine the optimal (*i.e.*, most sensitive) sliding window size (SWS), neuropeptide identifications from all SWS values (1-10) were compared. To determine the optimal target-decoy method, neuropeptide identifications using each target-decoy method (reversed, shuffled, randomized, and hybrid) method were compared.

Optimal HyPep parameters (SWS = 2 and target-decoy method = shuffled) were used for comparisons between HyPep and PEAKS DB. PEAKS DB parameters were the same as those used for PEAKS *de novo* sequencing parameters. The number of identifications at 1% FDR threshold from both programs were compared. To assess accuracy of both programs, an entrapment database was constructed consisting of all sequences from the original target crustacean neuropeptide database appended to a database of non-crustacean neuropeptides that were obtained from NeuroPep.[24] There were no overlapping sequences between the target and trap databases and there were the same number of target and trap sequences in the final entrapment database. Hits to the non-crustacean neuropeptides (N_{trap}) divided by the number of hits to the crustacean neuropeptides (N_{target}) were used to calculate the FMR (Equation 2).[57] The number of identifications assigned to the correct crustacean or incorrect non-crustacean sequences at 1% FDR threshold from both programs were compared.

$$FMR = \frac{N_{trap}}{N_{target}} \qquad \text{Equation 2}$$

For discovery of novel neuropeptides, identifications from HyPep using optimal parameters (SWS = 2 and target-decoy method = shuffled) at 1% FDR threshold were scanned for *de novo* sequences originating from 1) multiple MS/MS scans, 2) MS/MS scans with ALC = 99 and 3) HyPep score < 4. Sequences and MS/MS scans were manually evaluated, and novel neuropeptides were reported.

Results and Discussion:

Optimization of HyPep Sliding Window Size

To assess the performance of HyPep, determination of HyPep parameters that produce optimal (*i.e*., most sensitive) results is necessary. HyPep identifications are scored based on the sum of the subscores from four local alignment strategies: forward-fixed, backward-fixed, forward-varying, and backward-varying (Figure 2). The latter two alignment types were specifically designed to enable flexibility during the amino acid alignment process between the *de novo* sequence query and the neuropeptide database sequence, with flexibility regulated by the SWS.[52,58] For example, when matching a query sequence, sequence A, to a database sequence, sequence B, the SWS value is the maximum number of consecutive amino acids that will be scanned in sequence B to search for a matching residue. HyPep allows SWS values from one to ten. Intuitively, the smallest SWS value should result in the fewest number of identifications because it offers the least flexibility and the most stringent matching process, though this was not observed in actuality.

Neuropeptide extracts from crustacean brain tissues were analyzed with LC-MS/MS and processed using HyPep. Figure 3 shows number of identifications as a function of % FDR at all SWS values 1-10. Overall, there is agreement in number of neuropeptide identifications regardless of SWS value used. The number of identifications between 1-5% FDR were closely examined, as these are the thresholds typically used within the proteomics community. The greatest number of identifications occurred when smaller SWS values (*i.e.*, 1-3) were used, which was unexpected as a larger SWS offers more opportunities for a match to occur between a *de novo* sequence and a database sequence. This observation was true in both the original, target database as well as the decoy database, implying that the decrease in neuropeptide identifications as SWS increases is likely due to an increasing number of hits to the decoy database (*i.e.*, decoy-hits). Since sequence query alignment and scoring to both target and decoy database sequences occurs prior to the FDR calculation, increased decoy-hits raises the score that corresponds to 1% FDR. As this score is raised, fewer target-hits meet this score threshold, resulting in fewer neuropeptide identifications. Figure 4 shows a broader comparison of the SWS values, as it includes the number of identifications at 1% FDR from all four neuroendocrine tissue types (Brain, CoG, SG, and PO) at SWS values 1-10. Both Figures 3 and 4 show that an SWS value of 2 produces a higher number of identifications in all tissues. Therefore, a SWS value of 2 was chosen as the optimal value.

Evaluation of HyPep Target-Decoy Methods

The selection of an appropriate target-decoy method for calculating the FDR in proteomics is an ongoing discussion and it is beyond the scope of this report to comprehensively compare the benefits and pitfalls of each method.[56,59–65] It is unanimously agreed upon that the decoy database generation method and FDR calculation must be transparently described for each informatics tool, so that the user may adapt the most suitable target-decoy method for their dataset. Within HyPep, three common methods, reverse, shuffle, and random, as well as one novel hybrid method for decoy database generation exists (Figure S2).[66,67] Figure S3 compares the characteristics of the target (original) crustacean neuropeptide database with the four decoy database options. All decoy databases, including the novel hybrid method described in this report, meet the criteria set forth by Gygi et al., which states that decoy databases must contain 1) similar amino acid distribution, 2) similar sequence length distribution, and 3) similar number of sequences between target and decoy database, and 4) no peptides in common between target and decoy databases.[56] Each HyPep decoy database sequence is identical in length to its corresponding target database sequence. Each reverse and shuffle decoy sequence mass is identical to its corresponding target sequence, and the distribution of sequence masses from the random and hybrid decoy databases are in good agreement with the target database (Figure S3). Based on these characteristics, as well as the increasing level of stochasticity inherent to the reverse, shuffle, hybrid, and random decoy databases, respectively, sequences from the reverse and shuffle decoy databases contain decoy sequences highly similar to the target sequences, and therefore should result in more matches between *de novo* sequenced queries and the decoy database sequences than the hybrid and randomized decoy databases. We indeed observed that the reverse target-decoy method produced the lowest number of identifications at 1% FDR, yet the shuffle decoy method produced the largest number (Figures 3, 4, and 5). Although the number of identifications resulting from each target-decoy method were not the same, all the identifications from each method are valid due to the verification step that occurs after the FDR calculation by the HyPep AMM module.

Shuffle or pseudo-shuffle decoy methods have been reported to overestimate the % FDR for analysis of tryptic-digested peptides.[63,66] This speculation is amplified for neuropeptide analysis when considering the likelihood that a sequence from a shuffled decoy database not only acts as a decoy to its originating target sequence but also to multiple target sequences due to the highly homologous nature of neuropeptides within a family. An example of the high homology within a family is shown in Figure S4, where roughly half of neuropeptides from the crustacean hyperglycemic hormone (CHH) neuropeptide superfamily contain at least 50% sequence identity to another neuropeptide. Regardless, because it produced the highest number of neuropeptides identified at 1% FDR, the shuffle target-decoy method was selected as the optimal target-decoy method. It is worth noting that, although the novel hybrid method for decoy database generation was not selected as the optimal method, we believe it contains merit for potential future informatics tools because it met the aforementioned criteria and therefore worth retaining within the HyPep toolkit.[63]

Performance of HyPep

To assess the performance of HyPep compared to other available software, the LC-MS/MS data from peptide extracts of brain, SG, PO, and CoG tissue types were analyzed separately using HyPep and PEAKS DB algorithms. PEAKS DB was selected as both HyPep and PEAKS DB use the same *de novo* sequenced peptides list as the input file. The data were processed with HyPep using a SWS value of 2 and the shuffle target-decoy method, while the PEAKS DB program uses a decoy-fusion target-decoy method.[68] Since HyPep considers all *de novo* peptide sequence queries as potential full-length neuropeptide sequences, the decoy-fusion target-decoy method was not made available within HyPep because decoy-fusion sequences are twice as long as sequences within the original neuropeptide database.

A threshold of 1% FDR was used for both software, with precise % FDR values from HyPep of 0.76% for SG, 0.79% for PO, and 1.41% for CoG, and 1.3% for brain tissues. All identifications from HyPep and PEAKS DB are listed in Supplemental Files 1 and 2, respectively. HyPep resulted in more identifications than PEAKS DB software (Figure 5A). There is relatively low agreement in identifications between the two software, as 13-24% of PEAKS DB identifications were not found by HyPep. In order to understand this discrepancy, the neuropeptide family and sequence length of the identifications from both software were compared (Figures 5B and 5C). Across all tissues, one neuropeptide family, calcitonin/diapause hormone (CT/DH), was uniquely identified using PEAKS DB, while neuropeptides from the C-type allatostatin (AST-C), GSEFLamide, and proctolin families were only identified using HyPep. HyPep identified more neuropeptides within each family than PEAKS DB with the exception of the orcokinin family. Since one of the technical challenges of neuropeptide identification we sought to overcome was the ability to detect and differentiate between homologous neuropeptides, this observation implies that HyPep is better equipped to overcome that obstacle than PEAKS DB.

HyPep identified sequences with up to 18 amino acid residues while PEAKS DB identified sequences with up to 36 amino acid residues (Figure 5C). Considering the development of HyPep was motivated by improving identification of shorter neuropeptides, the increase in identification of shorter neuropeptides was promising. Since the disparity in identifications between HyPep and PEAKS DB would decrease if HyPep had been able to identify longer neuropeptides, this discrepancy is expected to decrease as algorithms used for database-free *de novo* sequencing improves their capabilities for detecting longer sequences, thus providing a higher quality input list for HyPep. HCD fragmentation was found to be most compatible with current *de novo* sequencing software, especially regarding accurate sequencing of longer peptides, so only datasets collected with HCD fragmentation were included in this report and compatible with HyPep.[69]

The accuracy of peptide identification software is an important metric that is nearly impossible to assess using a biological sample alone. A previously reported method for estimating the accuracy of database search-based identification algorithms involves generating an entrapment database containing sequences that are unlikely to be found in the sample, which are then appended to the original target database, enabling FMR measurements[57,70]. Entrapment sequences for this study included non-crustacean neuropeptides from the NeuroPep online database (http://isyslab.info/NeuroPep/). Figure 6 shows the number of identifications from PEAKS DB and HyPep at 1% FDR using the entrapment database. HyPep produced 85, 96, 79, and 47 target identifications with 2.4, 2.1, 2.5, and 2.1% FMR for the brain, SG, PO, and CoG tissues, respectively, while PEAKS DB produced 54, 50, 48, and 42 target identifications with 1.8, 2, 0, and 2.4% FMR for the same tissues. All identifications using the entrapment database are listed in Supplemental Files 3 and 4 for HyPep and PEAKS DB, respectively. The approximately 2% FMR observed for both software aligns with the 1% FDR threshold that was used as a processing parameter for both software. HyPep identified more neuropeptides regardless of which database was used (original database or entrapment database) at comparable FDR and FMR accuracy as PEAKS DB.

Neuropeptide Discovery using HyPep

Neuropeptide prohormones undergo splicing at the post-transcriptional and post-translational stages, which result in neuropeptide families containing sequences with high sequence similarity to one another.[71] Although knowledge of genomic or transcriptomic sequence coupled with confirmed biosynthesis pathway and bioactivity information is ideal for characterization of a *de novo* sequenced peptide query as a neuropeptide, *de novo* sequencing can be used as a standalone tool as a starting point for discovering neuropeptides provided the peptide query contains a highly conserved neuropeptide sequence motif.[50] It is important to note that *de novo* sequencing-based discovery is only applicable for peptide extracts from neuropeptide-relevant samples to increase confidence. Since *de novo* sequencing algorithms typically produce tens of thousands of peptide sequences from a single untargeted LC-MS/MS dataset and thousands remain even after a neuropeptide motif filter is applied, it is necessary to narrow down the list before manual examination of the spectra. HyPep's ability to find *de novo* sequenced peptide queries that partially match with entire database sequences can be utilized for neuropeptide discovery, given that the query sequence is associated with more than one PSM.

HyPep was used to discover neuropeptides from brain, PO, CoG, and SG tissue types. PEAKS conveys confidence in *de novo* sequence outcomes as an average local confidence (ALC) score, where an ALC of 99 represents a sequence with the highest level of certainty. As such, an ALC score threshold of 99 was used for this analysis. It is worth mentioning that previous reports have used lower ALC thresholds for neuropeptide discovery, thus, future usage of this discovery strategy may also apply less stringent ALC thresholds.[48,67,72] Table 1 lists 11 novel neuropeptides and reports their corresponding HyPep score to the neuropeptide database sequences, which ranges from 2.8-3.6. The upper limit of 3.6 was chosen because it was the highest score that was less than 4. The lower limit of 2.8 was a result of processing the data through HyPep at a 1% FDR threshold during the database search, where 2.8 was the highest score above a 1% FDR cutoff. Most of these putative novel neuropeptides contain conserved sequence motifs such as carboxy-terminal arginine and an amidated phenylalanine (RFamide)[73]. One of the novel sequences, SSFSRPPamide, is almost a perfect match with the database sequence SSFSPRPamide, except where the fifth and sixth amino acids are switched. If the genomic or transcriptomic information was available, the variation between these two sequences could be investigated for potential amino acid insertion, substitution, deletion, or replacement incidences.[74] Considering the high homology to known neuropeptide sequences and high spectral quality, these sequences may confidently move forward for examination of neuropeptide function.

De novo sequenced peptides were processed through HyPep and the output was filtered for peptides that 1) were assigned an average local confidence (ALC) score of 99, 2) peptide was *de novo* sequenced in more than one MS/MS spectrum and 3) the HyPep alignment score to a neuropeptide database sequence is equal to or greater than the score required for neuropeptide identification at 1% FDR identification.

Conclusions:

We present a novel strategy for neuropeptide identification, HyPep, that utilizes sequence alignment and sequence homology between *de novo* sequenced peptides and peptide database sequences. In this report, PEAKS software was used for database-free *de novo* sequencing of peptides from LC-MS/MS datasets of neural tissue extracts and processed with either HyPep or PEAKS DB for database searching and peptide identification. Overall, HyPep identified more neuropeptides than PEAKS DB from each tissue type at 1% FDR and at similar FMR from four neuronal tissue types within *C. sapidus*. The HyPep neuropeptide identifications contained shorter sequences and more sequences from each neuropeptide family than PEAKS DB. We also report our strategy for leveraging HyPep to select high quality *de novo* sequenced peptide candidates for discovery of novel neuropeptides. HyPep's methodology for peptide identification and discovery may be expanded to other classes of endogenous peptides. Although an external limitation stems from the performance of the database-free *de novo* sequencing algorithm that is used to generate the input for HyPep, future directions for HyPep include optimization of the peptide feature detection algorithm to detect longer neuropeptide sequences, as well as capabilities for detection of a/x and c/z fragment ions. HyPep is open source, and the software GUI and instructions can be downloaded at https://github.com/lingjunli-research/HyPep-v1.0.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements:

This work was supported in part by National Science Foundation (CHE-2108223) and National Institutes of Health (NIH) through grant R01DK071801. The Orbitrap instruments were purchased through the support of an NIH shared instrument grant (S10RR029531) and Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin-Madison. NQV was supported in part by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number T32GM008349 (Biotechnology Training Program) and by the National Heart Lung and Blood Institute of the National Institutes of Health under Award Number T32HL007936 (Training Program in Translational Cardiovascular Science). LF was supported in part by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number T32GM008505 (Chemistry–Biology Interface Training Program). L.L. would like to acknowledge NIH grants RF1AG052324, R01AG078794, R21AG065728, S10OD028473, and S10OD025084, as well as funding support from 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. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or National Science Foundation. The TOC figure was created with BioRender.com

Data deposition

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE [1] partner repository with the dataset identifier PXD037058.

Abbreviations:

MS	Mass Spectrometry
MS/MS	Tandem Mass Spectrometry
PSM	Peptide-Spectrum Match
LC	Liquid Chromatography
SWS	Sliding Window Size
FDR	False Discovery Rate
PEAKS DB	Peaks Database Search
FMR	False Match Rate
MeOH	Methanol
ACN	Acetonitrile
GAA	Glacial Acetic Acid
FA	Formic Acid
H_2O	Water
$[M+H]^+$	Protonated Monoisotopic Mass
SG	Sinus Glands
PO	Pericardial Organs
CoG	Commissural Ganglions
NaCl	Sodium Chloride
$CaCl_2$	Calcium Chloride

$MgCl_2$	Magnesium Chloride
NaOH	Sodium Hydroxide
AGC	Automatic Gain Control
HCD	Higher-energy Collisional Dissociation
DDA	Data-dependent Acquisition
NCE	Normalized Collision Energy
GUI	Graphical User Interface
PTM	Post-Translational Modification
SHS	Sequence Homology Search
N_{target}	Hits to Target Crustacean Database
N_{decoy}	Hits to Decoy Database
AMM	Accurate Mass Matching
Da	Dalton
N_{trap}	Hits to Trap Non-Crustacean Database
ALC	Average Local Confidence
CHH	Crustacean Hyperglycemic Hormone
CT/DH	Calcitonin/Diapause Hormone
AST-C	Allatostatin C-Type

References:

(1). Grimmelikhuijzen CJP; Hauser F Mini-Review: The Evolution of Neuropeptide Signaling. Regul Pept 2012, 177, S6–S9. 10.1016/j.regpep.2012.05.001. [PubMed: 22726357]

(2). Sahbaz BD; Iyison NB Neuropeptides as Ligands for GPCRs. In *IntechOpen*; InTech, 2018; pp 77–102. 10.5772/intechopen.73504.

(3). Dickinson P; Calkins A; Stevens J Related Neuropeptides Use Different Balances of Unitary Mechanisms to Modulate the Cardiac Neuromuscular System in the American Lobster, Homarus Americanus. J Neurophysiol 2014, 113, 856–870. 10.1152/jn.00585.2014. [PubMed: 25392168]

(4). Dickinson PS; Qu X; Stanhope ME Neuropeptide Modulation of Pattern-Generating Systems in Crustaceans: Comparative Studies and Approaches. Curr Opin Neurobiol 2016, 41, 149–157. 10.1016/j.conb.2016.09.010. [PubMed: 27693928]

(5). Nässel DR; Pauls D; Huetteroth W Neuropeptides in Modulation of Drosophila Behavior: How to Get a Grip on Their Pleiotropic Actions. Curr Opin Insect Sci 2019, 36, 1–8. 10.1016/j.cois.2019.03.002. [PubMed: 31280184]

(6). Christie AE; Stemmler EA; Dickinson PS Crustacean Neuropeptides. Cellular and Molecular Life Sciences 2010, 67 (24), 4135–4169. 10.1007/s00018-010-0482-8. [PubMed: 20725764]

(7). Burbach JP What Are Neuropeptides? In Neuropeptides: Methods and Protocols; Merighi A, Ed.; Humana Press: New York, 2011; Vol. 789, pp 1–36. 10.1007/978-1-61779-310-3_1.

(8). Corbière A; Vaudry H; Chan P; Walet-Balieu ML; Lecroq T; Lefebvre A; Pineau C; Vaudry D Strategies for the Identification of Bioactive Neuropeptides in Vertebrates. Front Neurosci 2019, 13, 1–12. 10.3389/fnins.2019.00948. [PubMed: 30740042]

(9). Mains R; Eipper B The Neuropeptides. In Basic Neurochemistry: Molecular, Cellular and Medical Aspects; Siegel G, Agranoff B, Albers R, Eds.; Lippincott-Raven: Philadelphia, 1999. 10.1007/978-1-61779-310-3_1.

(10). Abdulganiyyu IA; Kaczmarek K; Zabrocki J; Nachman RJ; Marchal E; Schellens S; Verlinden H; Broeck J. vanden; Marco H; Jackson GE Conformational Analysis of a Cyclic AKH Neuropeptide Analog That Elicits Selective Activity on Locust versus Honeybee Receptor. Insect Biochem Mol Biol 2020, 125, 1–14. 10.1016/j.ibmb.2020.103362.

(11). Reymond T; Delmas L; Koerber SC; Brown MR; Rivier JE Truncated, Branched, and/or Cyclic Analogues of Neuropeptide Y: Importance of the Pancreatic Peptide Fold in the Design of Specific Y2 Receptor Ligandst. J. Med. Chem 1992, 35, 3653–3659. 10.1021/jm00098a009. [PubMed: 1433176]

(12). Quillet R; Ayachi S; Bihel F; Elhabazi K; Ilien B; Simonin F RF-Amide Neuropeptides and Their Receptors in Mammals: Pharmacological Properties, Drug Development and Main Physiological Functions. Pharmacol Ther 2016, 160, 84–132. 10.1016/j.pharmthera.2016.02.005. [PubMed: 26896564]

(13). Qin YY; Huang XR; Zhang J; Wu W; Chen J; Wan S; Yu XY; Lan HY Neuropeptide Y Attenuates Cardiac Remodeling and Deterioration of Function Following Myocardial Infarction. Molecular Therapy 2022, 30 (2), 881–897. 10.1016/j.ymthe.2021.10.005. [PubMed: 34628054]

(14). Brothers SP; Wahlestedt C Therapeutic Potential of Neuropeptide y (NPY) Receptor Ligands. EMBO Mol Med 2010, 2 (11), 429–439. 10.1002/emmm.201000100. [PubMed: 20972986]

(15). Chen XY; Du YF; Chen L Neuropeptides Exert Neuroprotective Effects in Alzheimer’s Disease. Front Mol Neurosci 2019, 11. 10.3389/fnmol.2018.00493. [PubMed: 30787866]

(16). Yeo XY; Cunliffe G; Ho RC; Lee SS; Jung S Potentials of Neuropeptides as Therapeutic Agents for Neurological Diseases. Biomedicines 2022, 10 (2). 10.3390/biomedicines10020343.

(17). Wei P; Keller C; Li L Neuropeptides in Gut-Brain Axis and Their Influence on Host Immunity and Stress. Comput Struct Biotechnol J 2020, 18, 843–851. 10.1016/j.csbj.2020.02.018. [PubMed: 32322366]

(18). Ache B; Atwood H; Barnes J; Beltz B; Clarac F; Cooke I; Edwards D; Fraser P; Glantz R; Harris-Warrick R; Johnson B; Krasne F; Kravitz E; Macmillan D; Paul D; Rathmayer W; Takahata M Crustacean Experimental Systems in Neurobiology; Wiese, Ed.; Springer Berlin Heidelberg: Berlin, 2002. 10.1007/978-3-642-56092-7.

(19). Li L; Sweedler J. v. Peptides in the Brain: Mass Spectrometry-Based Measurement Approaches and Challenges. Annual Review of Analytical Chemistry. 2008, pp 451–483. 10.1146/annurev.anchem.1.031207.113053.

(20). Romanova E. v.; Sweedler J. v. Peptidomics for the Discovery and Characterization of Neuropeptides and Hormones. Trends Pharmacol Sci 2015, 36 (9), 579–586. 10.1016/j.tips.2015.05.009. [PubMed: 26143240]

(21). Hummon AB; Amare A; Sweedler J. v. Discovering New Invertebrate Neuropeptides Using Mass Spectrometry. Mass Spectrom Rev 2006, 25 (1), 77–98. 10.1002/mas.20055. [PubMed: 15937922]

(22). Rubakhin SS; Churchill JD; Greenough WT; Sweedler J. v. Profiling Signaling Peptides in Single Mammalian Cells Using Mass Spectrometry. Anal Chem 2006, 78 (20), 7267–7272. 10.1021/ac0607010. [PubMed: 17037931]

(23). Hökfelt T; Broberger C; David Xu Z-Q; Sergeyev V; Ubink R; Diez M Neuropeptides-an Overview; 2000; Vol. 39. 10.1016/S0028-3908(00)00010-1.

(24). Bin Y; Zhang W; Tang W; Dai R; Li M; Zhu Q; Xia J Prediction of Neuropeptides from Sequence Information Using Ensemble Classifier and Hybrid Features. J Proteome Res 2020, 19 (9), 3732–3740. 10.1021/acs.jproteome.0c00276. [PubMed: 32786686]

(25). Wang Y; Wang M; Yin S; Jang R; Wang J; Xue Z; Xu T NeuroPep: A Comprehensive Resource of Neuropeptides. Database 2015, 2015. 10.1093/database/bav038.

(26). Akhtar MN; Southey BR; Andrén PE; Sweedler J. v.; Rodriguez-Zas SL Evaluation of Database Search Programs for Accurate Detection of Neuropeptides in Tandem Mass Spectrometry Experiments. J Proteome Res 2012, 11 (12), 6044–6055. 10.1021/pr3007123. [PubMed: 23082934]

(27). Fu Q; Li L De Novo Sequencing of Neuropeptides Using Reductive Isotopic Methylation and Investigation of ESI QTOF MS/MS Fragmentation Pattern of Neuropeptides with N-Terminal Dimethylation. Anal Chem 2005, 77 (23), 7783–7795. 10.1021/ac051324e. [PubMed: 16316189]

(28). 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 Communications in Mass Spectrometry 2003, 17 (20), 2337–2342. 10.1002/rcm.1196. [PubMed: 14558135]

(29). Eng JK; Jahan TA; Hoopmann MR Comet: An Open-Source MS/MS Sequence Database Search Tool. Proteomics 2013, 13 (1), 22–24. 10.1002/pmic.201200439. [PubMed: 23148064]

(30). Eng JK; McCormack AL; Yates JR An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. J Am Soc Mass Spectrom 1994, 5 (11), 976–989. 10.1016/1044-0305(94)80016-2. [PubMed: 24226387]

(31). Eng JK; Fischer B; Grossmann J; MacCoss MJ A Fast SEQUEST Cross Correlation Algorithm. J Proteome Res 2008, 7 (10), 4598–4602. 10.1021/pr800420s. [PubMed: 18774840]

(32). May D; Fitzgibbon M; Liu Y; Holzman T; Eng J; Kemp CJ; Whiteaker J; Paulovich A; McIntosh M A Platform for Accurate Mass and Time Analyses of Mass Spectrometry Data. J Proteome Res 2007, 6 (7), 2685–2694. 10.1021/pr070146y. [PubMed: 17559252]

(33). Lam H; Deutsch EW; Eddes JS; Eng JK; King N; Stein SE; Aebersold R Development and Validation of a Spectral Library Searching Method for Peptide Identification from MS/MS. Proteomics 2007, 7 (5), 655–667. 10.1002/pmic.200600625. [PubMed: 17295354]

(34). Ulanga U; Russell M; Patassini S; Brazzatti J; Graham C; Whetton AD; Graham RLJ Generation of a Mouse SWATH-MS Spectral Library to Quantify 10148 Proteins Involved in Cell Reprogramming. Sci Data 2021, 8 (1), 118. 10.1038/s41597-021-00896-w. [PubMed: 33903600]

(35). Ma B; Johnson R De Novo Sequencing and Homology Searching. Molecular and Cellular Proteomics. February 2012. 10.1074/mcp.O111.014902.

(36). Tran NH; Zhang X; Xin L; Shan B; Li M De Novo Peptide Sequencing by Deep Learning. Proc Natl Acad Sci U S A 2017, 114 (31), 8247–8252. 10.1073/pnas.1705691114. [PubMed: 28720701]

(37). Medzihradszky KF; Chalkley RJ Lessons in de Novo Peptide Sequencing by Tandem Mass Spectrometry. Mass Spectrom Rev 2015, 34 (1), 43–63. 10.1002/mas.21406. [PubMed: 25667941]

(38). Kim Y; Bark S; Hook V; Bandeira N NeuroPedia: Neuropeptide Database and Spectral Library. Bioinformatics 2011, 27 (19), 2772–2773. 10.1093/bioinformatics/btr445. [PubMed: 21821666]

(39). Phetsanthad A; Vu NQ; Yu Q; Buchberger AR; Chen Z; Keller C; Li L Recent Advances in Mass Spectrometry Analysis of Neuropeptides. Mass Spectrom Rev 2021. 10.1002/mas.21734.

(40). Pearson WR An Introduction to Sequence Similarity (“homology”) Searching. Curr Protoc Bioinformatics 2013, No. SUPPL.42. 10.1002/0471250953.bi0301s42.

(41). Hook V; Funkelstein L; Lu D; Bark S; Wegrzyn J; Hwang SR Proteases for Processing Proneuropeptides into Peptide Neurotransmitters and Hormones. Annu Rev Pharmacol Toxicol 2008, 48, 393–423. 10.1146/annurev.pharmtox.48.113006.094812. [PubMed: 18184105]

(42). Fu Q; Christie AE; Li L Mass Spectrometric Characterization of Crustacean Hyperglycemic Hormone Precursor-Related Peptides (CPRPs) from the Sinus Gland of the Crab, Cancer Productus. Peptides (N.Y.) 2005, 26 (11), 2137–2150. 10.1016/j.peptides.2005.03.040.

(43). Dircksen H; Tesfai LK; Albus C; Nässel DR Ion Transport Peptide Splice Forms in Central and Peripheral Neurons throughout Postembryogenesis of Drosophila Melanogaster. Journal of Comparative Neurology 2008, 509 (1), 23–41. 10.1002/cne.21715. [PubMed: 18418898]

(44). Buck LB; Bigelow JM; Axel R Alternative Splicing in Individual Aplysia Neurons Generates Neuropepetide Diversity. Cell 1987, 51, 127–133. 10.1016/0092-86748790017-1. [PubMed: 3652207]

(45). Russo AF Overview of Neuropeptides: Awakening the Senses? Headache 2017, 57, 37–46. 10.1111/head.13084. [PubMed: 28485842]

(46). Hui L; Cunningham R; Zhang Z; Cao W; Jia C; Li L Discovery and Characterization of the Crustacean Hyperglycemic Hormone Precursor Related Peptides (CPRP) and Orcokinin Neuropeptides in the Sinus Glands of the Blue Crab Callinectes Sapidus Using Multiple Tandem Mass Spectrometry Techniques. J Proteome Res 2011, 10 (9), 4219–4229. 10.1021/pr200391g. [PubMed: 21740068]

(47). Stemmler EA; Bruns EA; Cashman CR; Dickinson PS; Christie AE Molecular and Mass Spectral Identification of the Broadly Conserved Decapod Crustacean Neuropeptide PQIRYHQCYFNPISCF: The First PISCF-Allatostatin (Manduca Sexta- or C-Type Allatostatin) from a Non-Insect. Gen Comp Endocrinol 2010, 165 (1), 1–10. 10.1016/j.ygcen.2009.05.010. [PubMed: 19467234]

(48). Phetsanthad A; Roycroft C; Li L Enrichment and Fragmentation Approaches for Enhanced Detection and Characterization of Endogenous Glycosylated Neuropeptides. Proteomics 2022. 10.1002/pmic.202100375.

(49). Kou Q; Xun L; Liu X TopPIC: A Software Tool for Top-down Mass Spectrometry-Based Proteoform Identification and Characterization. Bioinformatics 2016, 32 (22), 3495–3497. 10.1093/bioinformatics/btw398. [PubMed: 27423895]

(50). DeLaney K; Cao W; Ma Y; Ma M; Zhang Y; Li L PRESnovo: Prescreening Prior to de Novo Sequencing to Improve Accuracy and Sensitivity of Neuropeptide Identification. J Am Soc Mass Spectrom 2020, 31 (7), 1358–1371. 10.1021/jasms.0c00013. [PubMed: 32266812]

(51). Menschaert G; Vandekerckhove TTM; Baggerman G; Landuyt B; Sweedler J. v.; Schoofs L; Luyten W; van Criekinge W A Hybrid, de Novo Based, Genome-Wide Database Search Approach Applied to the Sea Urchin Neuropeptidome. J Proteome Res 2010, 9 (2), 990–996. 10.1021/pr900885k. [PubMed: 20000637]

(52). Zhang X; Kahveci T A New Approach for Alignment of Multiple Proteins. Pacific Symposium on Biocomputing 2006, 11, 339–350. 10.1142/9789812701626_0031.

(53). Eddy S Where Did the BLOSUM62 Alignment Scorematrix Come From? Nat Biotechnol 2004, 22 (8), 1035–1036. 10.1038/nbt0804-1035. [PubMed: 15286655]

(54). Bachvaroff TR; McDonald RC; Plough L. v.; Chung JS Chromosome-Level Genome Assembly of the Blue Crab, Callinectes Sapidus. G3 2021, 11 (9), 1–11. 10.1093/g3journal/jkab212.

(55). Gutierrez GJ; Grashow RG Cancer Borealis Stomatogastric Nervous System Dissection. Journal of Visualized Experiments 2009, No. 25. 10.3791/1207.

(56). Elias JE; Gygi SP Target-Decoy Search Strategy for Increased Confidence in Large-Scale Protein Identifications by Mass Spectrometry. Nat Methods 2007, 4 (3), 207–214. 10.1038/nmeth1019. [PubMed: 17327847]

(57). Feng X. dong; Li L. wei; Zhang J. hong; Zhu Y. ping; Chang C; Shu K. xian; Ma J Using the Entrapment Sequence Method as a Standard to Evaluate Key Steps of Proteomics Data Analysis Process. BMC Genomics 2017, 18. 10.1186/s12864-017-3491-2. [PubMed: 28056769]

(58). Proutski V; Holmes EC SWAN: Sliding Window Analysis of Nucleotide Sequence Variability. BIOINFORMATICS 1997, 14 (1998), 467–468. 10.1093/bioinformatics/14.5.467.

(59). Elias JE; Gygi SP Target-Decoy Search Strategy for Mass Spectrometry-Based Proteomics. Methods Mol Biol 2010, 604, 55–71. 10.1007/978-1-60761-444-9_5. [PubMed: 20013364]

(60). Gupta N; Bandeira N; Keich U; Pevzner PA Target-Decoy Approach and False Discovery Rate: When Things May Go Wrong. J Am Soc Mass Spectrom 2011, 22 (7), 1111–1120. 10.1007/s13361-011-0139-3. [PubMed: 21953092]

(61). Lam H; Deutsch EW; Aebersold R Artificial Decoy Spectral Libraries for False Discovery Rate Estimation in Spectral Library Searching in Proteomics. J Proteome Res 2010, 9 (1), 605–610. 10.1021/pr900947u. [PubMed: 19916561]

(62). Keich U; Tamura K; Noble WS Averaging Strategy To Reduce Variability in Target-Decoy Estimates of False Discovery Rate. J Proteome Res 2019, 18 (2), 585–593. 10.1021/acs.jproteome.8b00802. [PubMed: 30560673]

(63). Kim H; Lee S; Park H Target-Small Decoy Search Strategy for False Discovery Rate Estimation. BMC Bioinformatics 2019, 20(1). 10.1186/s12859-019-3034-8.

(64). Keich U; Noble WS Progressive Calibration and Averaging for Tandem Mass Spectrometry Statistical Confidence Estimation: Why Settle for a Single Decoy? In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer Verlag, 2017; Vol. 10229 LNCS, pp 99–116. 10.1007/978-3-319-56970-3_7.

(65). Levitsky LI; Ivanov M. v.; Lobas AA; Gorshkov M. v. Unbiased False Discovery Rate Estimation for Shotgun Proteomics Based on the Target-Decoy Approach. J Proteome Res 2017, 16 (2), 393–397. 10.1021/acs.jproteome.6b00144. [PubMed: 27959540]

(66). Lee S; Park H; Kim H Comparison of False-Discovery Rates of Various Decoy Databases. Proteome Sci 2021, 19 (1). 10.1186/s12953-021-00179-7.

(67). Cao Q; Yu Q; Liu Y; Chen Z; Li L Signature-Ion-Triggered Mass Spectrometry Approach Enabled Discovery of N- And O-Linked Glycosylated Neuropeptides in the Crustacean Nervous System. J Proteome Res 2020, 19 (2), 634–643. 10.1021/acs.jproteome.9b00525. [PubMed: 31875397]

(68). Zhang J; Xin L; Shan B; Chen W; Xie M; Yuen D; Zhang W; Zhang Z; Lajoie GA; Ma B PEAKS DB: De Novo Sequencing Assisted Database Search for Sensitive and Accurate Peptide Identification. Molecular and Cellular Proteomics 2012, 11 (4). 10.1074/mcp.M111.010587.

(69). Muth T; Renard BY Evaluating de Novo Sequencing in Proteomics: Already an Accurate Alternative to Database-Driven Peptide Identification? Brief Bioinform 2018, 19 (5), 954–970. 10.1093/bib/bbx033. [PubMed: 28369237]

(70). Vaudel M; Burkhart JM; Breiter D; Zahedi RP; Sickmann A; Martens L A Complex Standard for Protein Identification, Designed by Evolution. J Proteome Res 2012, 11 (10), 5065–5071. 10.1021/pr300055q. [PubMed: 22489649]

(71). Southey BR; Rodriguez-Zas SL Alternative Splicing of Neuropeptide Prohormone and Receptor Genes Associated with Pain Sensitivity Was Detected with Zero-Inflated Models. Biomedicines 2022, 10 (4). 10.3390/biomedicines10040877.

(72). Delaney K; Li L Data Independent Acquisition Mass Spectrometry Method for Improved Neuropeptidomic Coverage in Crustacean Neural Tissue Extracts. Anal Chem 2019, 91 (8), 5150–5158. 10.1021/acs.analchem.8b05734. [PubMed: 30888792]

(73). Findeisen M; Rathmann D; Beck-Sickinger AG RFamide Peptides: Structure, Function, Mechanisms and Pharmaceutical Potential. Pharmaceuticals 2011, 4 (9), 1248–1280. 10.3390/ph4091248.

(74). Iengar P An Analysis of Substitution, Deletion and Insertion Mutations in Cancer Genes. Nucleic Acids Res 2012, 40 (14), 6401–6413. 10.1093/nar/gks290. [PubMed: 22492711]

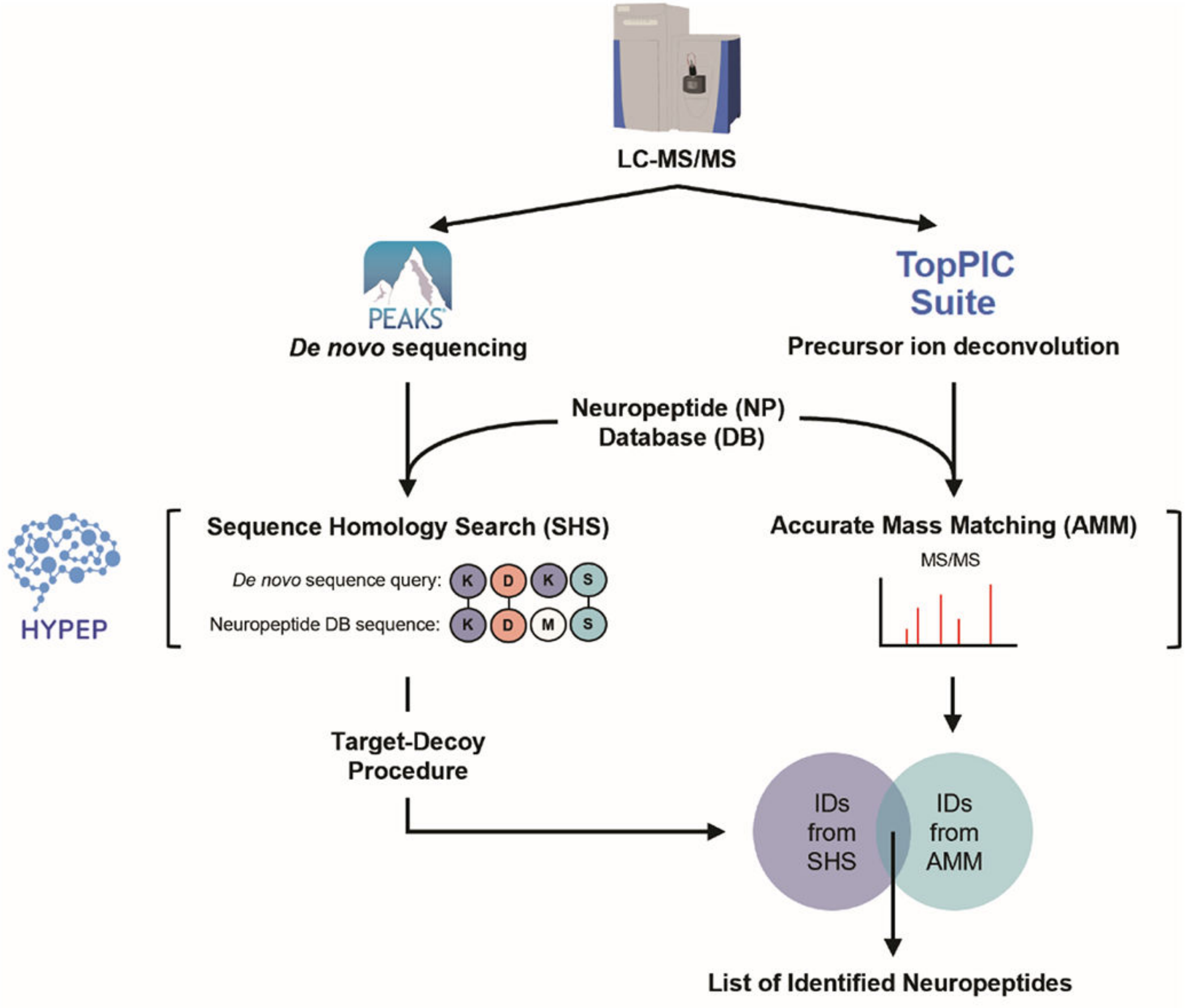

LC-MS/MS
PEAKS
De novo sequencing
TopPIC Suite
Precursor ion deconvolution
Neuropeptide (NP) Database (DB)
HYPEP
Sequence Homology Search (SHS)
De novo sequence query:
K D K S
Neuropeptide DB sequence:
K D M S
Accurate Mass Matching (AMM)
MS/MS
Target-Decoy Procedure
IDs from SHS
IDs from AMM
List of Identified Neuropeptides

Figure 1.

The workflow for HyPep analysis begins with processing LC-MS/MS data through PEAKS *de novo* sequencing program. *De novo* sequenced peptides are loaded into HyPep. HyPep performs a sequence homology search (SHS) between *de novo* sequenced query and neuropeptide database sequence and sequence matches above a false discovery rate threshold are reported. In parallel, LC-MS/MS data undergo precursor ion deconvolution through TopFD within TopPIC Suite[49]. Deconvoluted precursor masses are loaded into HyPep. HyPep performs accurate mass matching (AMM) between intact neuropeptide database sequences and deconvoluted precursor masses. Then, HyPep searches MS/MS scans where the isolated precursor corresponds to a deconvoluted precursor mass and contains fragment ions pertaining to the neuropeptide sequence. The filtered results from the SHS module are compared with the results from the AMM module and database sequences identified in both modules are reported as a final output.

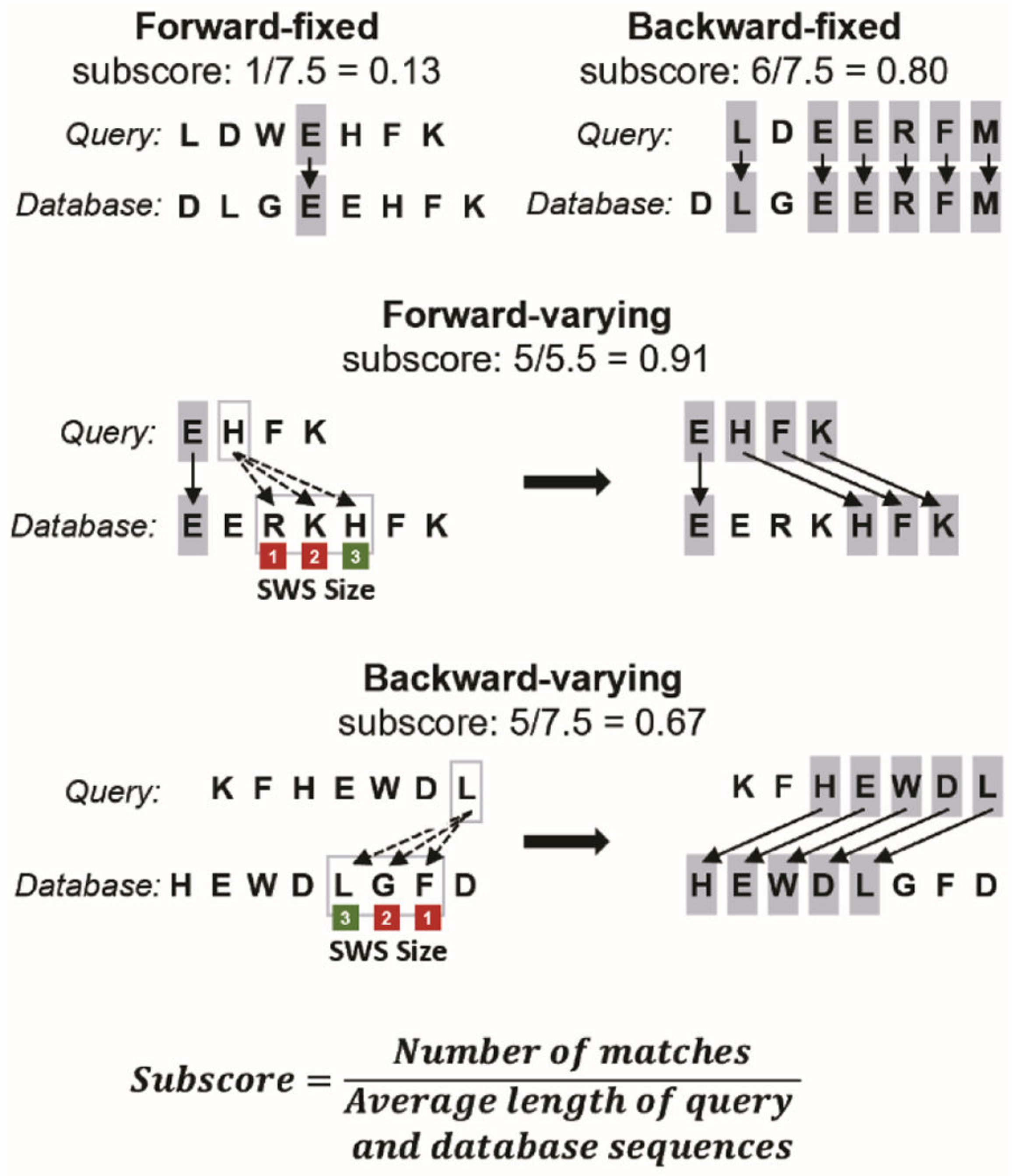


Figure 2.

HyPep's sequence homology search (SHS) algorithm contains four local alignment strategies where each produce a subscore that is calculated by taking the number of amino acid matches divided by the average amino acid length of the aligned sequences. All *de novo* sequenced peptide queries are aligned with all neuropeptide database sequences using all four local alignment strategies and the subscores from each alignment are summed to produce the final HyPep score for each identification. Forward-fixed involves aligning sequences starting on the N-terminus and backward-fixed aligns them on the C-terminus.

Forward-varying also aligns the sequences on the N-terminus, but a sliding window size (SWS) is incorporated to allow amino acids from one sequence to match with amino acids from the other sequence in cases where amino acid rearrangements have occurred. Backward-varying follows the same process but aligns the sequences on the C-terminus.

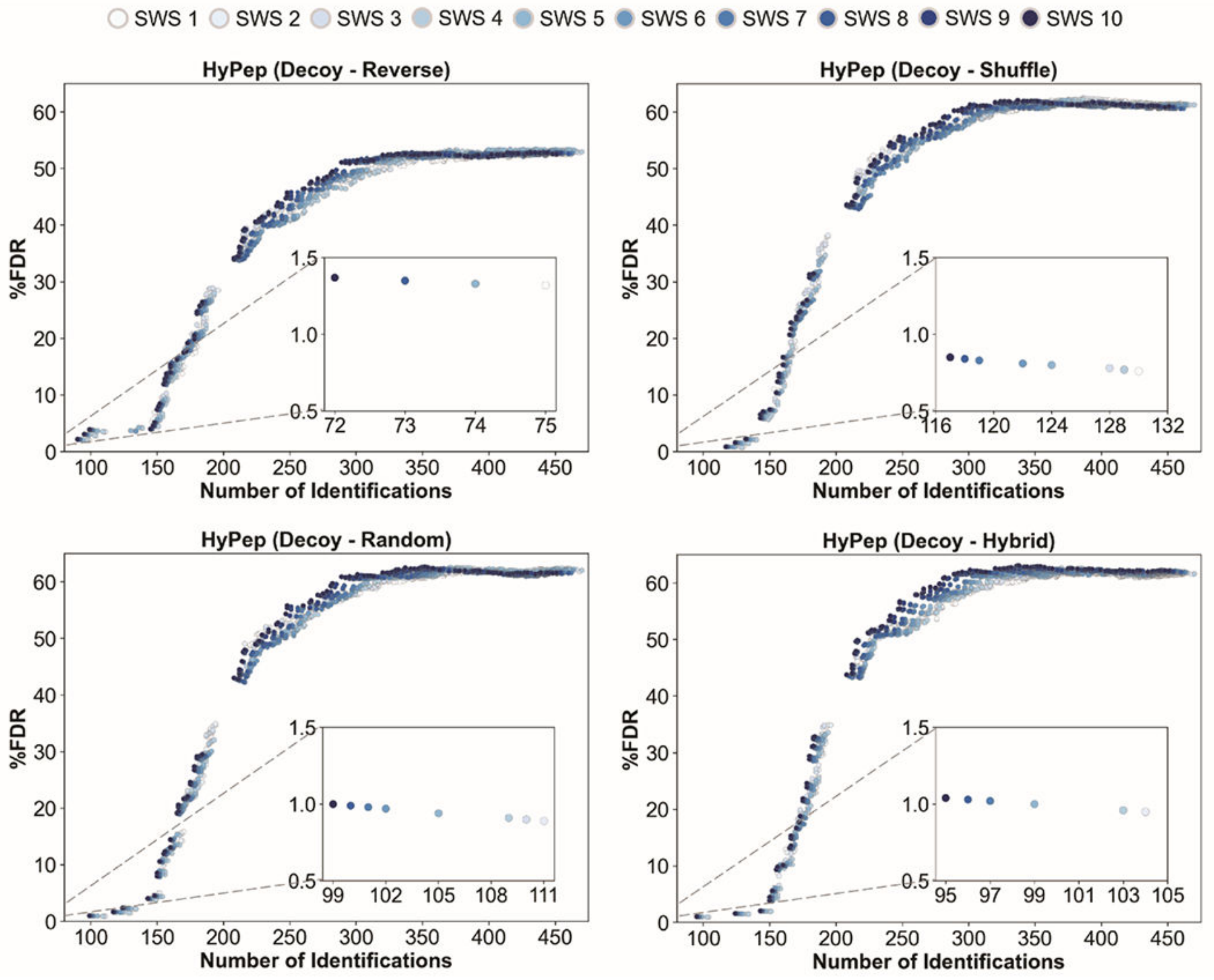


Figure 3.

Neuropeptide identifications as a function of % FDR from *C. sapidus* brain tissue processed using HyPep, which contains four target-decoy methods to calculate % FDR. Decoy databases can be generated using either shuffle, reverse, random, or hybrid decoy methods for FDR calculation (Figure S1). Insets show magnification at 0.5-1.5% FDR. Each color represents different sliding window sizes (SWS). For each target-decoy method, the same decoy database was used for neuropeptide identification using SWS values 1-10.

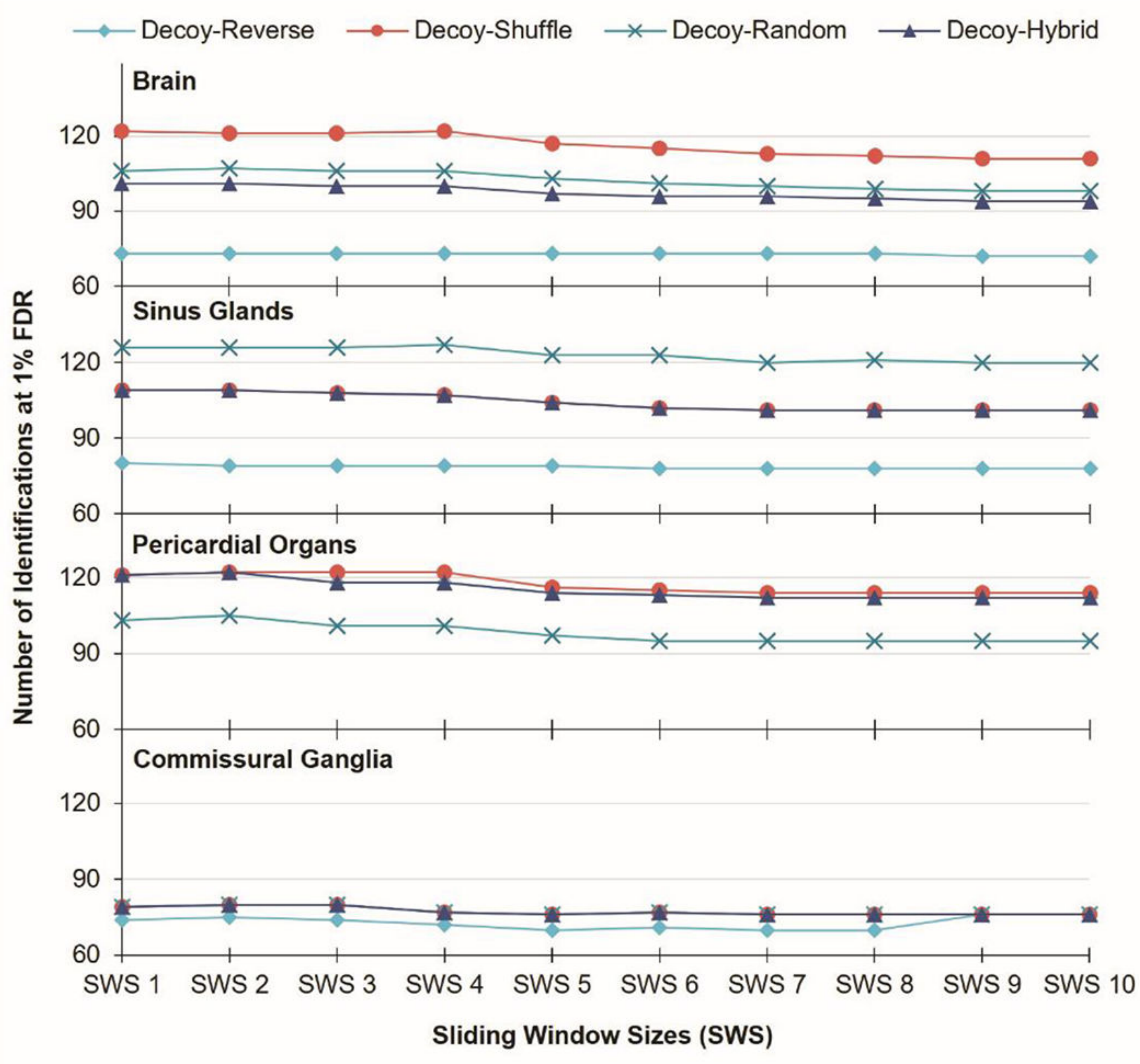


Figure 4.

Four tissues from *C. sapidus* (brain, sinus glands, pericardial organs, and commissural ganglia) were processed using HyPep parameters at SWS values 1-10 and four target-decoy methods (reverse = diamond, circle = shuffle, x = random, and triangle = hybrid). The number of identifications for each tissue using each SWS value and each target-decoy method are reported at threshold of 1% FDR.

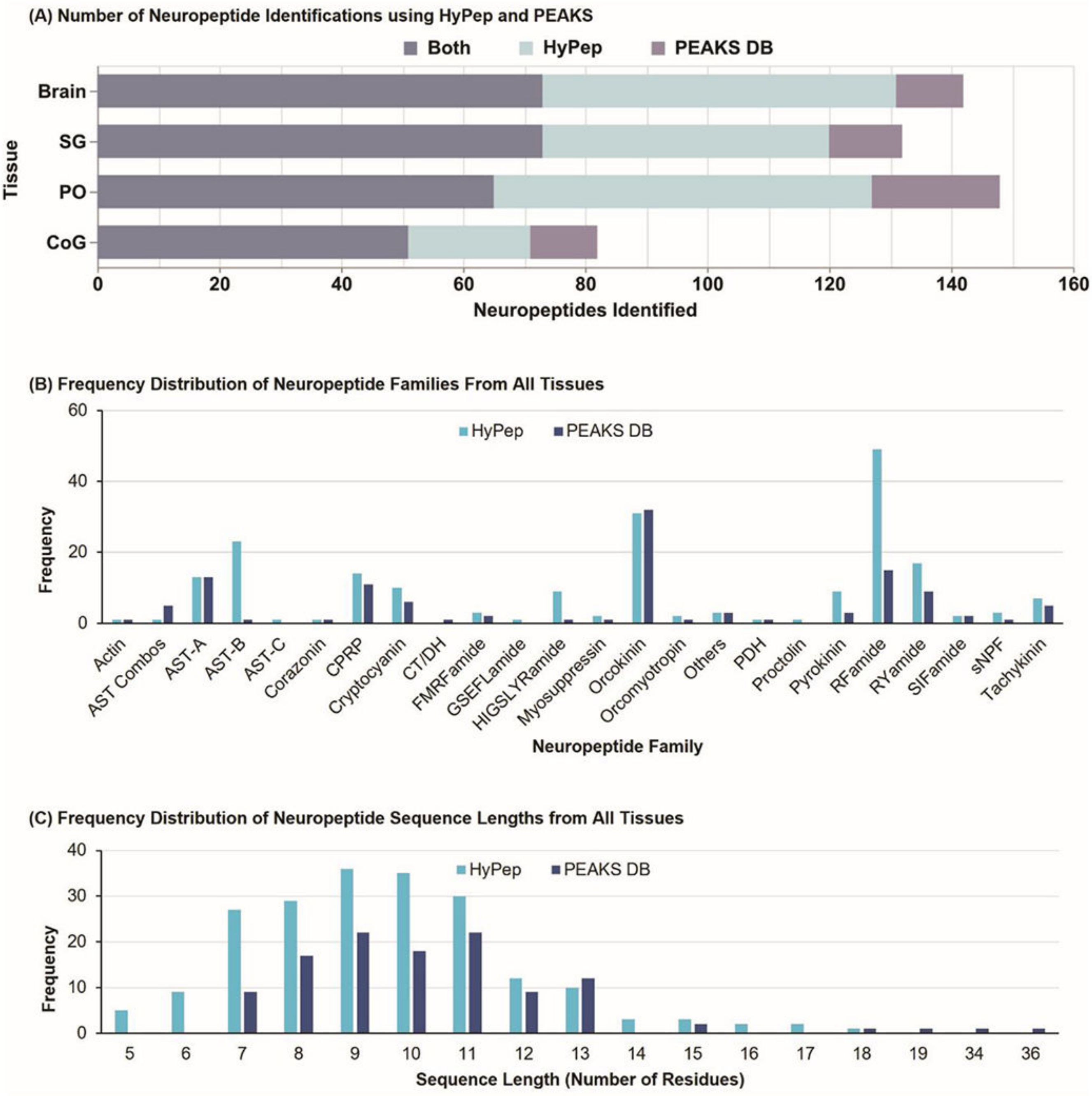


Figure 5.

Comparison of the number of neuropeptides identified using PEAKS DB and HyPep from four tissue types (brain, sinus glands (SG), pericardial organs (PO), and commissural ganglia (CoG)) from *C. sapidus*. 1% FDR threshold was set for both software. Neuropeptides were examined for (A) overlapping and unique sequences, (B) sequences from different neuropeptide families, and (C) lengths of neuropeptide sequences. AST: allatostatin; CPRP: crustacean hyperglycemic hormone precursor-related peptide; CT/DH: calcitonin/diapause hormone; PDH: pigment dispersing hormone; sNPF: short neuropeptide F.

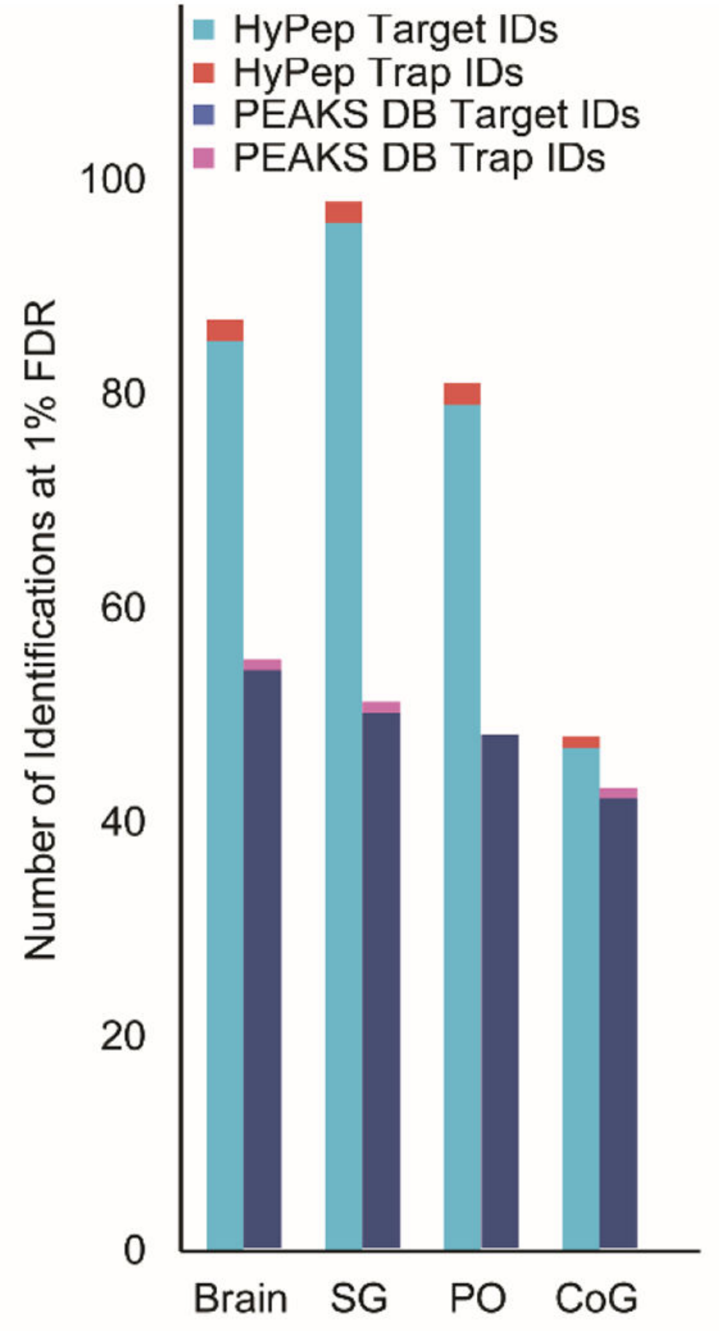


Figure 6.

Accuracy of neuropeptide identification from HyPep and PEAKS DB was evaluated by processing four tissue types from *C. sapidus* (brain, sinus glands (SG), pericardial organs (PO), and commissural ganglia (CoG)) through both programs. An entrapment database containing both crustacean neuropeptides and non-crustacean neuropeptides was used, and the number of total identifications and false identifications were reported at 1% FDR.

Table 1.

Putative novel neuropeptide sequences that were discovered in four tissue types from *C. sapidus*.

Novel Neuropeptide	Tissue Type	Homologous Neuropeptides from Database	HyPep Score
AGHKNYLRF(Amidated)	Brain, PO, SG, CoG	GAHKNYLRF (Amidated)	3.333
DARTPALRLRF(Amidated)	Brain, SG, CoG	DGRTPALRLRF(Amidated)	3.636
DARTPALRLRF	SG	DGRTPALRLR F(Amidated)	3.636
(Pyro-glu)ERNFLRF(Amidated)	PO	ELNFLRF(Amidated)	3.429
HLSSLLR	Brain, PO, SG	HYSSLLR(Amidated)	3.429
HLSSLLR(Amidated)	SG	HYSSLLR(Amidated)	3.429
HYGSLLR	SG	HYSSLLR(Amidated)	3.429
IMFDELDRS	CoG	NFDEIDRSA	2.824
QHKNYLRF(Amidated)	CoG	KHKNYLRF(Amidated)	3.5
RNNFLRF(Amidated)	SG, PO	GNNFLRL(Amidated) [SG] KNEFIRF(Amidated) [PO]	3.429 [SG] 2.857 [PO]
SSFSRPP(Amidated)	SG	SSFSPRP(Amidated)	3.143