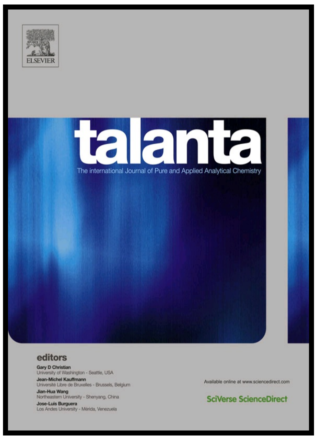

talanta
The International Journal of Pure and Applied Analytical Chemistry
editors
SciVerse ScienceDirect

www.elsevier.com/locate/talanta

Peptide identifications and false discovery rates using different mass spectrometry platforms

Krishna D.B. Anapindi, Elena V. Romanova, Bruce R. Southey, Jonathan V. Sweedler

PII: S0039-9140(18)30068-7
DOI: https://doi.org/10.1016/j.talanta.2018.01.062
Reference: TAL18285

To appear in: *Talanta*

Received date: 5 January 2018
Accepted date: 28 January 2018

Cite this article as: Krishna D.B. Anapindi, Elena V. Romanova, Bruce R. Southey and Jonathan V. Sweedler, Peptide identifications and false discovery rates using different mass spectrometry platforms, *Talanta,* https://doi.org/10.1016/j.talanta.2018.01.062

Peptide identifications and false discovery rates using different mass spectrometry platforms

Krishna D. B. Anapindi[1], Elena V. Romanova[1], Bruce R. Southey[2], Jonathan V. Sweedler[1,*]

[1]Department of Chemistry and the Beckman Institute, University of Illinois at Urbana-Champaign, Urbana 61801, IL, USA

[2]Department of Animal Sciences, University of Illinois at Urbana-Champaign, Urbana 61801, IL, USA

*To whom correspondence may be addressed: E-mail: jsweedle@illinois.edu.

ABSTRACT

Characterization of endogenous neuropeptides produced from post-translational proteolytic processing of precursor proteins is a demanding task. A variety of complex prohormone processing steps generate molecular diversity from neuropeptide prohormones, making *in silico* neuropeptide discovery difficult. In addition, the wide range of endogenous peptide concentrations as well as significant peptide complexity further challenge the structural characterization of neuropeptides. Liquid chromatography-mass spectrometry (MS), performed in conjunction with bioinformatics, allows for high-throughput characterization of peptides. Mass analyzers and molecular dissociation techniques render specific characteristics to the acquired data and thus, influence the analysis of the MS data using bioinformatic algorithms for follow-up peptide identification. Here we evaluated the efficacy of several distinct peptidomic workflows using two mass spectrometers for confident peptide discovery and characterization, the Thermo Orbitrap Fusion Tribrid and Bruker Impact HD UHR-QqTOF. We compared the results in several categories, including the numbers of identified peptides, full-length mature neuropeptides among all identifications, and precursor proteins mapped by the identified peptides. We also characterized the peptide false discovery rate (FDR) based on the occurrence of amidation, a known post-translational modification (PTM) that has been shown to require the presence of a C-terminal glycine. Thus, amidation events without a preceding glycine were considered false-positive amidation assignments. We compared the FDR calculated by the search engines used here to the minimum FDR estimated via false amidation assignments. The search engines severely underestimated the rate of false PTM assignments among the identified peptides, regardless of the specific MS platform used.

Keywords: neuropeptides, amidation, *de novo* sequencing, peptidomics, proteomics

INTRODUCTION

Neuropeptides are expressed and secreted by neurons and neuroendocrine organs, act as cell-to-cell signaling molecules, and are involved in a range of physiological processes, e.g., feeding, reproduction, locomotion, memory, and learning [1-6]. As outlined in **Fig. 1**, neuropeptides are produced by post-translational prohormone processing of larger precursor proteins via multiple steps of enzymatic cleavage, followed by additional modifications [7]. Since the 1990s, mass spectrometry (MS)-based characterization of peptides and proteins has played a vital role in understanding numerous physiological processes and disease states in models ranging from unicellular organisms to complex mammalian systems, with hundreds of peptides identified and characterized [8-12]. This progress has been made possible due to advancements in instrumental capabilities and computational tools for peptide sequencing and identification, as well as the development of robust workflows and peptide discovery strategies [13]. The speed, sensitivity, resolution, and dynamic range capabilities of modern mass spectrometers make them effective tools for peptide discovery and characterization.

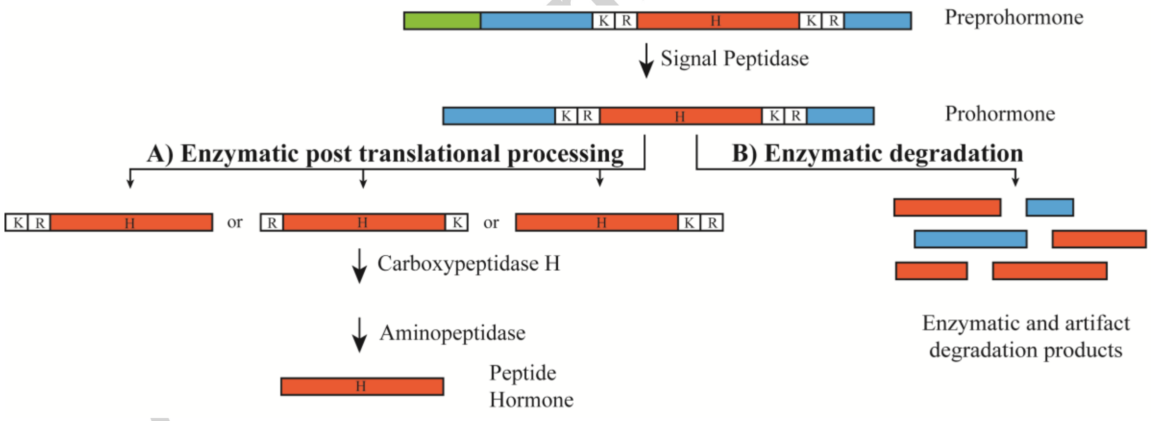


Fig. 1. Schematic of post-translational prohormone processing. (**A**) The prohormone is acted upon by

Given the distinct operational mechanisms and performance specifications of the mass analyzers available today, the analytical platform selected for peptide and neuropeptide identification is important, with decisions made according to experimental goals. Because mass spectrometers differ in resolution, sensitivity, accuracy, means of ion generation, ion focusing, transfer, accumulation, fragmentation, and detection, the produced tandem MS (MS^2) data may differ in ways that ultimately affect how these data integrate with bioinformatic sequencing and identification algorithms. Elias et al. [14] have shown that peptides identified exclusively by ion trap (IT)-based mass spectrometers are on an average twice as long as peptides identified by quadrupole (Q) time-of-flight (TOF) instruments. They also reported the percentage of confidently assigned MS^2 spectra to be 50% higher for IT compared to QTOF analyzers. Thus, different platforms may be biased toward preferential detection of molecules with specific physiochemical properties, even from the same sample. In addition to variations in mass spectrometer configurations and technical aspects of MS^2 data acquisition, bioinformatic requirements play a significant role in successful peptide identifications that drive discovery.

The goal of this work was to assess the technical advantages of several common instrumental platforms and mass spectrometric methodologies targeting neuropeptidomic applications. We analyzed peptide extracts from the abdominal ganglion of the mollusk *Aplysia californica*, a relatively simple animal model with a ganglionic nervous system comprised of ~20,000 neurons, which can be sampled selectively and reproducibly. Hundreds of *Aplysia* neuropeptides from numerous prohormones have been characterized by MS, with many localized to the abdominal ganglion. Moreover, a wealth of neuropeptide expression data are available for *Aplysia* [15-18], allowing for an informed assessment of the neuropeptide identification results collected from the various MS platforms tested here.

The criteria used to assess platform performance are based on bioinformatic outcomes when using automatic interrogation of the MS^2 data obtained from each instrumental platform and compared against the *Aplysia* protein database from UniProt [19] (https://www.uniprot.org/). We tabulated metrics such as the number of unique peptides and more specifically, mature, full-length neuropeptides, neuropeptide precursor protein coverage by detected peptides, mass range of the peptides detected, and percentage of peptide false-positive hits judged by the validity of a well-understood post-translational modification (PTM).

The term PTM usually refers to a covalent chemical change on proteins and peptides, which may turn the peptide molecule bioactive by improving its receptor binding or lifetime [20-22]. Molecular mechanisms of PTM formation are often highly conserved across different species. Here we evaluate the validity of one such PTM, C-terminal amidation, widely represented among known *Aplysia* neuropeptides and other animals, and identified in our experiments across different platforms. Significant experimental evidence on the *in-vivo* mechanism of peptide amidation indicates the only known mechanism for C-terminal amidation of polypeptides requires the presence of glycine on the C-termini. This glycine is acted upon by two enzymes, peptidylglycine alpha-hydroxylating monooxygenase and peptidyl-alpha-hydroxyglycine alpha-amidating lyase, in tandem, or a single combined enzyme, peptidyl-glycine alpha-amidating monooxygenase. Both processes result in the removal of glycine and the formation of a C-terminal amidated peptide with the loss of a glyoxylate anion [23].

Unfortunately, virtually all peptide-sequencing software packages consider only the mass shifts associated with substitution of a carboxyl group by an amine group amino residue amidation, regardless of whether this residue is preceded by glycine. Manual curation of automatically generated data for false PTM assignments, as reported in the current work, changes statistically reported false discovery rates (FDRs) and illuminates the advantages of using such biological information to evaluate peptidomic results. This PTM-based evaluation of results is one unique aspect of the current work. There have been studies to evaluate the performance of different mass spectrometric platforms and bioinformatic search algorithms [24-26]; however, there has been little effort to assess the spectral characteristics of data acquired by different platforms based on the actual identity of detected peptides that also used in-depth biological information on peptide formation. The PTM-based approach presented here evaluates whether the automatically deduced peptide structures are feasible from a biological standpoint.

In addition to the FDR estimation via known PTMs, we tested the fidelity of automatic peptide identification by searching the MS^2 data against a mixed species database containing protein entries from *Homo sapiens* in addition to *A. californica.* The *H. sapiens* database serves as a ‘dummy database,’ as described by Jeong et al. [27], from which no significant peptide spectrum match (PSM) is expected; however, the total number of PSM matches crossing a specific threshold are now reduced due to the increase in the size of the database. This reduction in the number of identified PSMs in part depends on the spectral quality acquired by the mass spectrometer, and may vary for different MS platforms. Data sets with high-quality spectra are likely to have fewer reductions compared to low-quality spectra. Moreover, the probability of a non-random match between the MS^2 spectra and peptide sequence within a database depends on the size of the database [27]. Hence, a higher quality MS^2 spectrum would be required to selectively identify *A. californica* peptides from a list of predominantly irrelevant proteins.

The first of two instruments used was an Orbitrap Fusion Tribrid mass spectrometer (Thermo Fisher Scientific, Waltham, MA); the flexibility offered by this system allowed us to test different combinations of molecular fragmentation modes and analyzer types for fragment ion analysis (MS^2). Three combinations were used with the Fusion Tribrid for MS^2: (1) high-energy collisional dissociation with Orbitrap (HCD-OT); (2) high-energy collisional dissociation with ion trap (HCD-IT); and (3) collision-induced dissociation with ion trap (CID-IT). Collision-induced dissociation with the Orbitrap (CID-OT) for MS^2 was not evaluated as it has been shown that the ion routing mechanism for this method leads to suboptimal performance of the instrument [25]. The second instrument was an Impact HD UHR-QqTOFmass spectrometer (Bruker, Billerica, MA), used to test collision-induced dissociation time-of-flight (CID-TOF) for MS^2. We demonstrate that the identification results, as judged by FDRs and biological merit, are influenced by the instrumental platform used for data acquisition.

EXPERIMENTAL

Animals

Aplysia californica were obtained from the NIH/University of Miami National Resource for *Aplysia* (Miami, FL) and housed in a tank with aerated, circulated, and filtered artificial seawater (and chilled to 15 ºC) prepared from Instant Ocean Sea Salt (Instant Ocean, Aquarium Systems Inc., Mentor, OH), dissolved in purified water. Four animals weighing 120–140 g were used for the current study.

Peptide extraction

Animals were anesthetized by injection of isotonic $MgCl_2$ (~50% of body weight) into the body cavity. Abdominal ganglia were quickly dissected, incubated for 30 min at 34 ºC in 10 mg/mL protease IX solution in artificial seawater (ASW) to soften the connective tissue (ASW: 460 mM NaCl, 10 mM KCl, 10 mM $CaCl_2$, 22 mM $MgCl_2$, 26 mM $MgSO_4$, and 10 mM HEPES in Milli-Q water (Millipore, Billerica, MA), pH adjusted to 7.8). Treated ganglia were rinsed in ASW supplemented with 100 units/mL penicillin G, 100 µg/mL streptomycin, and 100 µg/mL gentamicin, transferred into a vial with 100 µL of ice-cold acidified acetone (acetone:water:formic acid (FA) 40:5:5), and homogenized using a mechanical pestle (Kontes Pellet Pestle Motor, Thermo Fisher Scientific). The homogenate extraction was placed on ice for 30 min followed by centrifugation at 14,000 × g. The supernatant was collected, vacuum dried at room temperature (24 ºC) and stored at –20 ºC until further analysis. For liquid chromatography (LC)-MS analysis, the dry sample was reconstituted in 100 µL of 0.1 % FA in LC-MS grade water; 5 µL of this reconstituted sample was used for each of the technical replicates.

Peptide extract separation with nanoLC

LC was performed with a Dionex Ultimate 3000 RSLC with a nanoflow selector (Thermo Fisher Scientific). The separation method was kept consistent across the different MS instruments and configurations to ensure reproducible separation. The sample was loaded onto a C18 Acclaim PepMap µ-Precolumn trap (5 µm; Thermo Fisher Scientific) with a loading solvent (99% water, 1% acetonitrile (ACN), 0.1% FA, 0.01% trifluoroacetic acid) at 15 µL/min for 3 min. The trap was switched in line with an Acclaim PepMap RSLC column (C18, 75 µm × 150 mm, 2 µm, 100Å; Thermo Fisher Scientific), and sample separated at a uniform flow rate of 300 nL/min using 0.1% FA in LC-MS grade water (solvent A) and 0.1% FA in LC-MS grade ACN (solvent B) as the mobile phase. The flow gradient conditions were: 0–3 min, 1–1% B; 3–6 min 1–10% B; 6–90 min, 10–70% B; 90–100 min, 70–99 % B; 100–110 min 99–1% B; 110–120 min, 1–1% B.

Mass spectrometric measurements

Orbitrap. Top speed data-dependent precursor ion selection was used for all of the three fragmentation modes on the Orbitrap Fusion Tribrid mass spectrometer (Thermo Fisher Scientific) with a cycle time of 3 s. The parent ions were scanned with an Orbitrap (OT) resolution of 120 K and an automatic gain control (AGC) target of 200,000. Dynamic exclusion was turned on with the following settings: exclusion time = 60 *s*; mass tolerance = +/- 10 ppm; repeat count = 3. For OT detection, the parent ions were scanned in the range of 300–1500 *m/z*, the fragment ions were scanned with an OT resolution of 30 K, maximum injection time of 60 ms, and AGC target of 50,000. Precursor ions with a charge ranging from +2 to +7 were considered. A higher range of charge states has been used for OT as a majority of peptides analyzed via electrospray ionization (ESI)-OT-MS are multiply charged. Additionally, several contaminant ions present in the sample are usually singly charged. So, to avoid the background noise caused by those contaminants, a charge state of +2 to +7 was chosen. However, in ESI-QTOF-MS, a significant number of singly charged peptides are present; hence, a lower charge state range of +1 to +4 was chosen. A normalized collision energy value of 35% was used for the HCD fragmentation. For the IT detection, a maximum MS^2 injection time of 35 ms and an AGC target of 10,000 was chosen. A normalized collision energy of 35% was used for both CID and HCD fragmentation. Each of the four MS methods were analyzed in triplicate (*n* = 3).

QTOF. The Bruker Impact HD UHR-QqTOF mass spectrometer, outfitted with a CaptiveSpray nanosource, was used in MS^2 mode with CID fragmentation. The data were acquired over a range of 300–3000 *m/z* in a top speed data-dependent mode with a cycle time of 3 sec. Precursor ions in the range of +1 to +4 were considered. A fixed MS^1 scan rate of 4 Hz, and a variable MS^2 scan rate of 8 Hz for low intensity (5000 per 1000 sum) and 16 Hz for high intensity (100,000 per 1000 sum), were used. Collision energy was set at 10 eV with the stepping feature turned on. MS^2 collision energy was set at 100% for 70% of the time, and 200% for the remaining 30%. Dynamic exclusion was turned on, with an exclusion after 3 spectra per precursor ion for a duration of 60 s. Spectra corresponding to the same precursor ion were reconsidered for analysis if the new spectral intensity was more than 2.5 times the previous intensity.

MS^2 data characteristics

The average numbers of the MS^1 and MS^2 spectra acquired by each of the instrumental platforms are as follows. For MS^1 spectra—HCD-OT: 6396.67 (+/- 123.86 SD); CID-IT: 10528.67 (+/- 25.65 SD); HCD-IT: 12535.67 (+/- 258.12 SD); and QTOF-CID: 6041.67 (+/- 704.92 SD). For MS^2 spectra—HCD-OT: 55531 (+/- 239.22 SD); CID-IT: 77507.33 (+/- 225.63 SD); HCD-IT: 89481.33 (+/- 1344.339 SD); and QTOF-CID: 41332.67 (+/- 1466.04 SD). The OT data were acquired in .raw format and the QTOF data in .d format.

Bioinformatic peptide sequencing and identification, and post-search filtering criteria

The raw spectra from the QTOF instrument were converted into. mzxml format and loaded into the *de novo*-based peptide identification search engine, PEAKS Studio (Version 8.0, Bioinformatics Solutions Inc., Canada). The .raw spectra from the OT were directly loaded into PEAKS. The *A. californica* database (total entries 434) was used individually and merged with the *H. sapiens* database (total entries 139,331) from UniProt [19] for all of the searches. The search parameters were consistent across all four datasets from the four instrumental methods tested, and included no enzymatic digestion and variable PTMs of up to 3 per peptide: acetylation (K- and N-terminus), amidation, phosphorylation (S,T, and Y), half-disulfide bond per cysteine residue, pyroglutamination from E and Q, and Met oxidation. For the QTOF detection method, a precursor ion tolerance of 50 ppm and fragment ion tolerance of 0.1 Da were used.

Different precursor and fragment ion tolerance settings were used to search the data obtained by the Tribrid instrument because of the difference in mechanisms behind the ion acquisition for its different mass analyzers. OT is a high-resolution detector with greater mass accuracy, whereas IT can analyze ions at a much faster rate but with lower mass accuracy. To accommodate the differences in the operational mechanism of these two analyzers, separate precursor and product ion filtering criteria were used for different analyzer combinations when searching the database with the respective data: 20 ppm precursor ion tolerance with 0.02 Da fragment ion tolerance for HCD-OT, and 20 ppm precursor ion tolerance with 0.3 Da fragment ion tolerance for CID-IT and HCD-IT. Alternate search criteria with a 0.1 Da mass tolerance for both the OT and IT detectors were used but resulted in fewer chemically unique peptide hits. A filtering criterion of a 5% peptide-spectrum match (PSM) FDR was used for calculating the total number of peptide, neuropeptide, and precursor protein identifications. Additionally, filtering criteria with four different FDR percentages *viz.*, 0.1%, 0.5%, 1%, and 2%, and database searches with amidation as the only allowable variable PTM, were performed to evaluate the effect of these parameters on the FDR trends using the different instrumental methods.

Search engines, the software tools used to predict the peptide sequence from an MS^2 spectrum, are broadly classified into two categories: *de novo* sequencing and database searches. *De novo* sequencing algorithms predict the peptide sequences purely based on the pattern of MS^2 fragmentation, whereas the database search algorithms try to match the generated MS^2 spectrum to a sequence within the database. Modern search engines like PEAKS DB use a hybrid approach that implements both *de novo* sequencing and database search strategies to improve the accuracy and sensitivity of peptide identifications, as described by Zhang et al. [28]. Briefly, PEAKS DB first performs a *de novo* sequencing for each input spectrum followed by a peptide shortlisting. The shortlisted peptides are then assigned a score based on the match between a database sequence and an experimentally acquired MS^2 spectrum, and referred to as a PSM. A peptide is scored higher if there are multiple high-quality PSMs, all mapping to the same sequence in the database, and scored lower if there is just one low-quality PSM that maps to the sequence. Several factors influence the scoring of peptides during the database search with MS^2 data: peptide length, search space, number of fragment ions in the spectrum, and quality of the spectrum due to the mass accuracy of precursor and fragment ions, resolution, and signal-to-noise ratio. Peptide scores reported by a database search engine, however, ignore the biological feasibility of a peptide and calculate the FDR on an exclusively statistical basis [29].

RESULTS

Peptide and neuropeptide identification rates

The identification rates from data acquired using four different parameters (one for the IMPACT QTOF and three for the Orbitrap Fusion Tribrid) were evaluated based on bioinformatic metrics, including the confident identification of the total number of peptides, mature neuropeptides originating from known or predicted cleavage sites on the prohormone, percentage of the precursor protein sequence coverage by the neuropeptides, and percentage of false positives via known PTM sites. For simplicity, only mono- or di-basic cleavage sites on the prohormone were

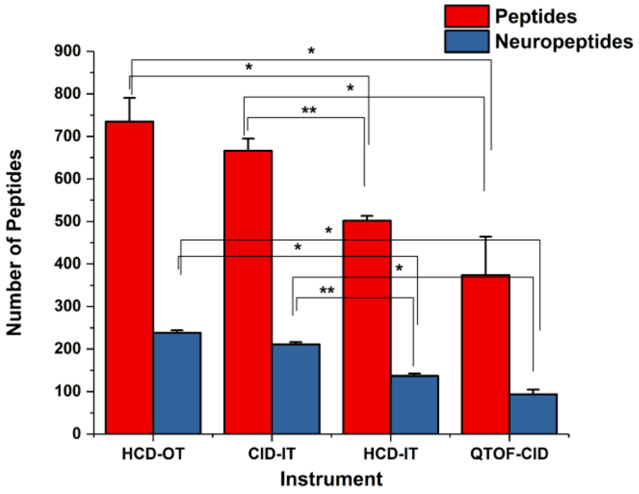


Fig. 2. Average number of unique peptide and neuropeptide sequences identified over *n* = 3 technical replicates for four different instrumental methods, where neuropeptides are defined as peptides derived from prohormones. Error bars represent the standard deviation. Complete lists of the peptides and proteins are provided in the supplementary material (Figs. S3–S5);*p <0.05, **p <0.01.

assessed to compare mature neuropeptides with potential peptide fragments from post-mortem degradation. The HCD-OT method resulted in significantly higher (p <0.05) numbers of peptide and neuropeptide identifications of 735.0 (+/- 37.6 SD) and 238.3 (+/- 7.2 SD), respectively, compared to the HCD-IT and CID-QTOF. With the HCD-IT, 501.7 (+/- 49.6 SD) peptides and 136.7 (+/- 13.1 SD) neuropeptides were identified, whereas the CID-QTOF dataset generated 373.7 (+/- 91.7 SD) and 93.7 (+/- 27.5 SD) peptides and neuropeptides, respectively. Using CID-IT, 666.3 (+/- 42.9 SD) peptides and 211.0 (+/- 14.9 SD) neuropeptides were identified (**Fig. 2**).

Prohormone identification rates from the A. californica protein database

Proteins with at least two chemically unique peptides were considered as a hit for the peptide precursor protein identification (**Fig. 3).** Using these criteria, the three Orbitrap methods—HCD-OT, CID-IT, HCD-IT—allowed identification of 61 (+/- 3.6 SD), 52.7 (+/- 7.1 SD), and 56 (+/- 7.2 SD) proteins, respectively; CID-QTOF resulted in identification of 25 proteins (+/- 3 SD).

Additionally, precursor protein sequence coverage by the individually identified peptides was evaluated in our study on an example of one unique prohormone, the egg-laying hormone (ELH). The ELH prohormone is highly expressed in the abdominal ganglion. Unlike many other prohormones detected in abdominal ganglion extracts, ELH encodes about 20 mature peptides, which have been previously characterized by MS [30-32]. We looked at the peptides with endogenous mono/dibasic cleavage sites to differentiate between the full-length mature neuropeptides from degradation products and sequentially cleaved ladder peptide sequences. All three methods employing the OT mass analyzer showed a consistent 21–22% neuropeptide detection among the ELH-mapped peptides, whereas only 15.5% of the mapped peptides turned out to be endogenously cleaved neuropeptides in the QTOF dataset (**Fig. S1**).

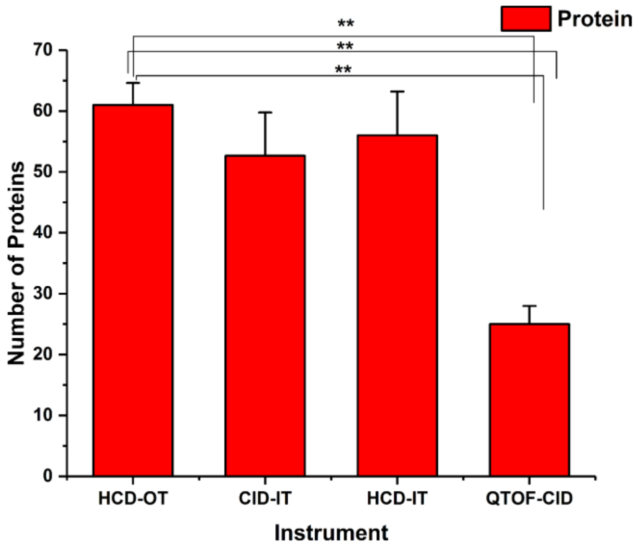


Fig. 3. Average number of neuropeptide precursor proteins identified over *n* = 3 technical replicates for four different instrumental methods. Error bars represent the standard deviation;**p <0.01.

Evaluating MS^2 spectral quality via searching against a mixed species database

To evaluate the quality of the spectral data acquired by the different platforms, we constructed a mixed species database consisting of the *A. californica* proteins and predominantly irrelevant proteins from *H. sapiens* by appending both the proteomes. This inflated database approach of results validation should help illuminate the quality differences among MS^2 datasets from different platforms. Because of the increased number of high-scoring random hits from an inflated decoy database [33], datasets of lower quality are expected to have a greater drop in the number of confident peptide identifications. As expected, a database size-dependent decrease in peptide identification rates was consistently observed across all of the platforms (**Fig. 4**). The HCD-IT method showed the largest decrease in the number of confident peptide identifications, which dropped 84%, from 501.7 (+/- 49.6 SD) to 81.3 (+/- 7.1 SD) when using the multi-species database. Other tested instrumental methods performed similarly, with a 63–67% drop in the peptide identification rate. In particular, with the CID-QTOF method, the average total peptide identifications were reduced from 373.7 (+/- 91.7 SD) to 126.3 (+/- 43.9 SD); the CID-IT and

HCD-OT methods resulted in identification of 251.0 (+/- 16.5 SD) and 323.3 (+/- 11.0 SD) *Aplysia* peptides, respectively, from the mixed-species database.

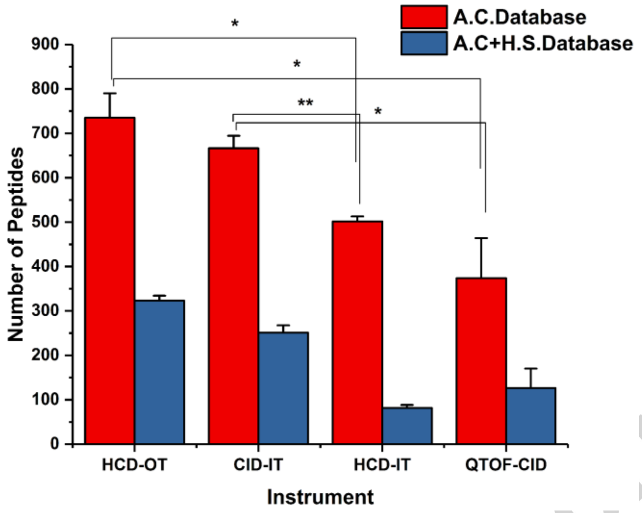


Fig. 4. Average number of peptides identified from *n* = 3 technical replicates for four different instrumental methods using two different databases: exclusively *Aplysia californica* (target, A.C.) and mixed species with precursor proteins from (A.C) + *Homo sapiens* (H.S.). *p <0.05, **p <0.01. For the peptides from the A.C. + H.S. database, all pairwise comparisons, except HCD-IT and QTOF-CID, are significantly different; p <0.05.

Peptide FDR estimation via identification of incorrectly assigned amidation

Here we assessed the biological feasibility of peptide structures deduced by PEAKS to evaluate the peptide level FDR. One advantage of using known enzymatic processing steps is that one can determine impossible PTMs. For example, amidation on residues that are not followed by a C-terminal glycine serves as a reliable way to check false-positive hits in automatic sequencing. Here we computed the percentage and an absolute number of such falsely identified peptides for five PSM-level FDR filters applied to PEAKS DB search results, i.e., 0.1%, 0.5%, 1%, 2%, and 5%. For a fixed search engine PSM level FDR cut-off of 5%, the average percentages of incorrectly identified amidated peptides were 29.3 (+/- 2.3 SD), 35.6 (+/- 7.4 SD), 36.3 (+/- 3.3

SD), and 51.0 (+/- 6.6 SD) for the HCD-OT, CID-IT, HCD-IT and CID-QTOF methods, respectively (**Fig. 5**). These percentages correspond to the peptide-level FDR values. Even for a strict search engine cutoff of 0.1% PSM FDR, the percentage of falsely identified amidated peptides ranged from 9.9 % (+/- 0.9 SD) using HCD-IT to 18.5 % (+/- 18.0 SD) using CID-QTOF, representing almost an order of magnitude range in the corresponding peptide-level FDR values reported by the search engine. Additionally, we noticed that higher the performance of the MS instrument in terms of mass resolution (OT and TOF analyzers), the greater the error in the

FDR, as observed versus calculated, compared to low resolution analyzers (IT). This trend is more apparent for the lower PSM FDR cutoff values (0.1%, 0.5%, and 1%).

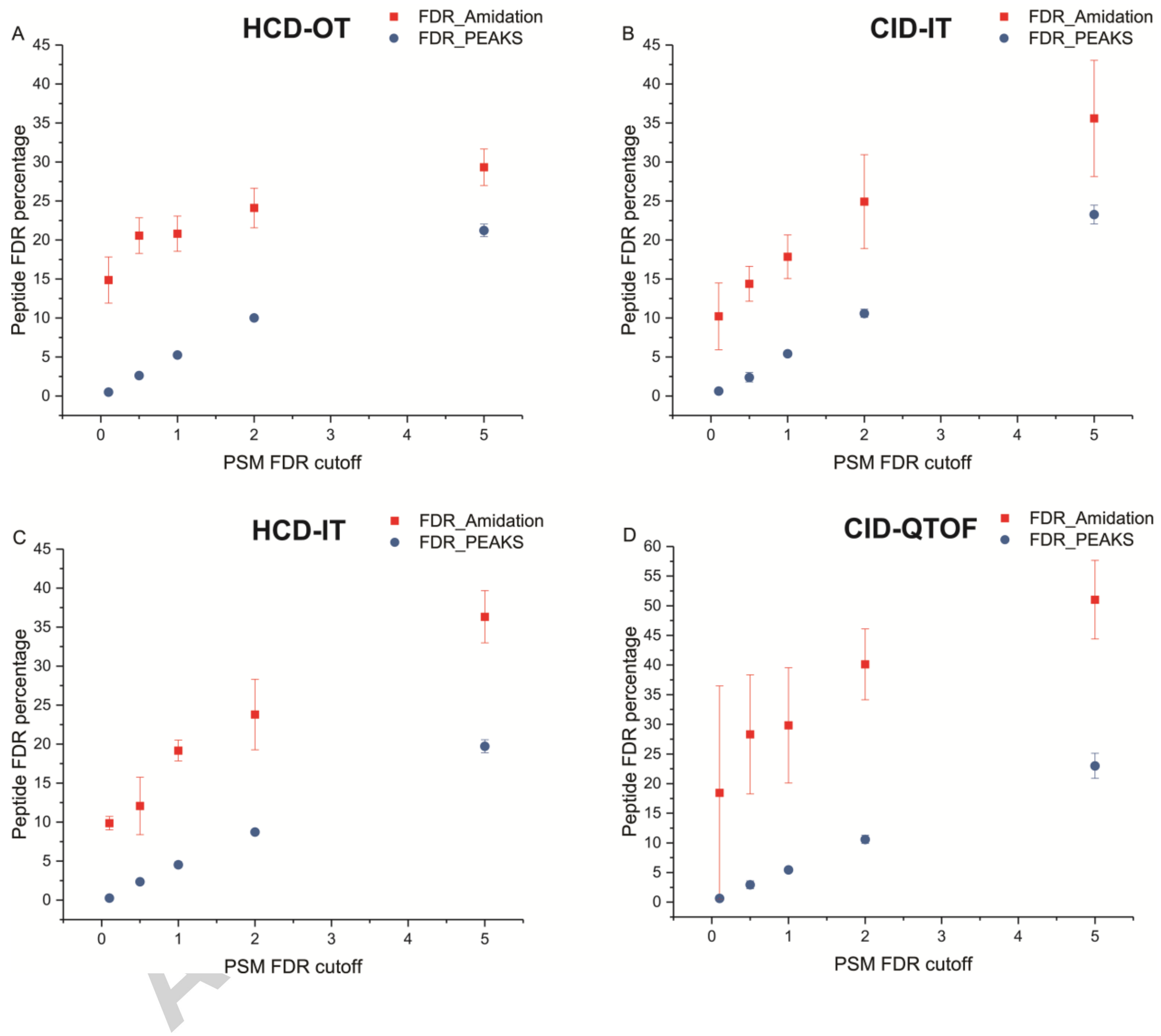


Fig. 5. Comparison of the search engine-reported peptide FDR percentages to experimentally determined percentages for the occurrence of false positives using all of the four tested platforms. The x-axis corresponds to

Occurrence of falsely amidated peptides among the technical replicates

The number of falsely amidated peptides among all peptide sequences with amidation that were identified in all three technical replicates was also determined for each of the studied methods. The false amidation results are as follows: CID-QTOF, 2 out of 15 total amidated peptides; HCD-OT, 8 out of 89; CID-IT, 6 out of 60; and HCD-IT, 1 out of 40.

Influence of search spaces on PSM assignment

An additional search with amidation as the only allowable variable PTM was performed to evaluate the effect of a reduced search space on PTM assignment. In this case, no significant difference was observed in the average number of peptides with false amidation relative to typical search space with seven other PTMs. Again, five different PSM FDR filters were tested (0.1%, 0.5%, 1%, 2%, and 5%), and the false amidation occurrence significantly increased the true FDR relative to the search engine-reported value (**Fig. S2**).

Fragmentation efficiency as a likely basis for false amidation identification in sequencing tools

The quality of an MS^2 spectrum plays a crucial role in peptide sequence determination using bioinformatics tools. PEAKS only reports a candidate peptide sequence with the best match to a spectrum as a highest score hit. The match is determined by mass accuracy and the number of assigned fragment ions. For example, peptide R.GIFTQSAYGSYPRV(a).G (-10LogP score of 98.98) is C-terminally amidated with a glycine residue following valine (**Fig. S3**). This peptide has a complete list of high-intensity b and y fragment ions that facilitate a confident and accurate estimation of the peptide sequence. In contrast, L.FGLTISDMGCAITLF(a).W (-10LogP score of 20.07) is one of the false PTM identifications where the C-terminal phenylalanine is followed by a tryptophan and not a glycine. The MS^2 fragmentation pattern (**Fig. S4**), and distribution of b and y ions for this peptide, reveal that the ions in the *m/z* range of 800–1800 are mostly absent, so essentially the sequence has been determined based on fewer fragment ions. Though poor fragmentation and insufficient confirmation of the proposed sequence via experimental fragment ions are the reasons for most false amidation assignments, the converse, however, is not true. The peptide R. GGSLDALRSGHQVPMLRA(a).GR (-10LogP score of 19.66) has a similar length, fragmentation efficiency, ion coverage, and -10LogP score (**Fig. S5**) as the falsely amidated peptide L.FGLTISDMGCAITLF(a).W. However, this is most likely a low-abundant true positive as it has basic cleavage sites on both of the termini, and the amidation occurs on the residue preceded by glycine.

DISCUSSION

Automated database searches are an integral tool for high-throughput discovery and characterization of neuropeptides from MS^2 data. However, a growing body of research has drawn attention to unrealistic FDRs reported by the majority of database search engines for MS^2 data interpretation, with proposed solutions ranging from improved algorithms and their combinatorial strategies, to the possibility of biological validation of automatically generated peptide identifications [34-38]. Here we investigated how spectral characteristics from different MS platforms work with a commercial bioinformatics package for effective peptide identification, and show that screening for C-terminus amidation, which requires a following glycine, allows one category of false-positive identifications to be determined. The screening approach was found effective for datasets obtained with different mass analyzers and molecular dissociation methods.

The details of the specific MS platform used influence automatic peptide identification outcomes. As evident from the average total number of unique peptide sequences obtained, HCD-OT resulted in a significantly higher number of total peptide identifications compared to HCD-IT and CID-QTOF. Additionally, datasets acquired with the OT- and IT-based methods contained 20–50% more MS^2 spectra compared to the TOF dataset. Although the average total number of unique peptides identified by CID-IT was lower than HCD-OT, the difference was not significant (p >0.05). The number of neuropeptide identifications also followed the same trend as the number of total peptides identified. Identifying the mature neuropeptides aids in distinguishing between the endogenous peptides and the peptides produced as an artifact of sampling, measurement, or sequencing errors. This difference distinguishes the redundant peptide forms that are usually the sequential degradation products of a mature full-length peptide. In most cases, the full-length mature neuropeptides are the biologically active compounds that bind to specific receptors and modulate various physiological functions. Loss or substitution of even a single amino acid residue in a peptide sequence may compromise G-protein receptor coupling [39].

Although a search engine estimates PSM-level FDRs from a statistical viewpoint, a more practical peptide-level FDR is appropriate for studies that rely on peptide identification results. Typically, multiple correct PSMs for each sequence tag could be produced from an MS^2 dataset, and the best scoring PSM is reported by the search engine used as a representative of the peptide. In contrast, the false identifications are usually supported by a single PSM to a low quality MS^2 spectrum. This results in a significant difference between a search engine-reported PSM-level FDR value and the peptide-level FDR. Although few search engines evaluate the peptide-level FDR, they do report the empirical FDR, which makes them prone to biases in the datasets. Jeong et al. [27] evaluated both factual and empirical peptide-level FDRs, but their reported factual peptide-level FDRs still depended on the results from the target-decoy approach, and may have introduced bias. Individual evaluation of mass spectra to confirm the identity of the PSMs is a plausible approach to address the above issue, but doing a manual inspection of all spectra in a typical MS experiment containing several thousand spectra is time consuming. Moreover, it is often not possible to determine whether a PSM is a true positive or a false positive.

Taking a different approach, we took advantage of the fact that C-terminal amidation cannot occur without the loss of glycine as a glyoxylate ion. In mammals, amidated peptides play roles in neuropeptide signaling pathways by mediating water balance (antidiuretic hormone), pregnancy and lactation in females (oxytocin), and positive regulation of cytosolic calcium ion concentration, among many other functions [40-42]. The results from the current study indicate that for any given PSM-level FDR cutoff, the search engine-estimated FDR percentage of peptides is always lower compared to the percentage of false positives estimated via evaluation of false amidation. While high quality mass spectra should minimize the number of false positives from a database search, our data suggest that the issue may be related to the method used to characterize the FDR because the percentage errors in FDRs are nearly as high in the platform with the best MS figures of merit, and are larger for more strict FDR values. Also, the fact that we observed a greater error in observed versus calculated peptide FDRs for higher resolution MS instruments suggests that the problem could lie in the informatics routines used.

However, there are limitations to using amidated peptides at a constant FDR filter as a benchmark to evaluate the false positives. Firstly, they represent only 15–20% of the total peptide identifications; hence, the sample set is smaller. Secondly, amidation is known to occur only in the secretory pathway in acidified vesicles, and so is not applicable to datasets outside of secretory products. Lastly, though a constant FDR filter ensures that the ratio of decoy PSMs to total PSMs remains consistent across all the platforms as calculated by the search engine used, the actual ratio could vary from platform to platform, depending on the spectral quality and the database used to search. Since there is no real way to verify the identity of all peptides, benchmarking amidated peptides provides a fairly simple and reasonably accurate way to estimate false positives, despite the caveats mentioned above.

It is important to note that the MS^2 fragmentation pattern of a precursor ion has a significant impact on peptide identification. By manually inspecting the data, we found that false identifications usually associate with low fragmentation spectrum quality and/or disparate peptide sequence coverage by assigned fragment ions. PEAKS identifies the maximum discriminating fragment ions, and attributes the amidation site to the last residue, if subtracting 0.98 Da leads to a sequence with a higher number of fragment ion matches to the experimental spectrum, which results in the false identification. In contrast, with an efficient fragmentation spectrum, the correct sequence is easily fitted to the spectrum, which leads to a true identification with an amidation followed by a C-terminal glycine. Although there are falsely amidated peptides common to all the three replicates for all platforms, the ratio of the number of false identifications over total identifications among the peptides common to all the three replicates is much lower than the same ratio for individual technical replicates. These low percentages of false positives that are common to the three technical replicates for a given experimental platform, suggest that the PSMs that result in factual false positives do not consistently occur in different technical replicates, i.e., the wrongly assigned *m/z* are not detected in all of the replicates. We observed similar peptide FDR values when using a reduced search space with amidation as the only allowable PTM, suggesting that the percentage of peptides identified to be incorrectly amidated is not likely due to the search parameters, such as search space and number of PTMs chosen, but is more dependent on the quality of the MS^2 spectra acquired. Oftentimes spectra with poor/incomplete lower mass ions are assigned to a false amidation.

Searches using the multispecies database of *H. sapiens* + *A. californica* understandably resulted in significant decreases in the total peptide identifications across all platforms. This can be attributed to the well-known fact that fewer PSMs would cross the search engine threshold when searched against a larger database compared to a more compact database [27, 43]. The smallest decrease (56% compared to the *A. californica*-only database) in unique peptide identifications in the mixed-species database was noticed with the method employing HCD fragmentation with the OT analyzer. HCD-OT offers high-resolution ion detection, no low-mass cutoff, and increased parent ion fragmentation, which leads to an overall good MS^2 spectral quality. Because of these reasons, around half of the peptides from the *A. californica*-only database were also confidently matched to a protein in the *A. californica* database, despite the presence of an overwhelmingly large number of proteins from *H. sapiens* in the mixed-species database. In contrast, HCD-IT resulted in the largest percentage decrease in peptide identifications (84% decrease in peptide identifications compared to the results obtained from the *A. californica*-only database). After adding the human proteins, the increased size of the database resulted in many PSMs not crossing a set threshold, and the decrease in identifications. Hence, though HCD-IT has an edge over other methods in terms of its high speed and fragmentation efficiency, the benefits may be offset when searching against a large protein database, which is often the case when using models with unsequenced or poorly annotated genomes.

CONCLUSIONS

Our data indicate that the goals of a peptidomics study can be successfully achieved using a range of instrumental platforms. Although the multiple platforms tested here performed well, the OT and IT analyzers allowed the identification of the most neuropeptides from complex samples, in agreement with a prior report on the Tribrid mass spectrometer [24]. The IT advantage can be best utilized in a targeted study where data are searched against a smaller, targeted database. For discovery efforts where a targeted database is not available or is too large, such as cross-species homology searches, either OT or QTOF can be effective as they generate data that are equally less affected by the database inflation.

Regardless of the MS platform and method used, the search engine-reported peptide FDR levels are consistently underestimated for any given PSM FDR cutoff, with larger errors obtained for the more stringent FDR cutoffs. Establishing unique, model-specific criteria for biological validation of automatically generated interpretation of MS^2 spectra and peptide/protein assignment can improve the outcomes of a peptidomics experiment. At a minimum, we recommend a careful manual inspection of lower-scoring PSMs for quality and meaningful fragment ion assignments before considering the proposed sequences as true peptide identifications.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version.

ACKNOWLEDGEMENTS

The project described was supported by Award No. P30 DA018310 from the National Institute on Drug Abuse, Award No. R01 NS031609 from the National Institute of Neurological Disorders and Stroke, and Award No. CHE-16-06791 from the National Science Foundation. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

REFERENCES

[1] A. Proekt, F.S. Vilim, V. Alexeeva, V. Brezina, A. Friedman, J. Jing, L. Li, Y. Zhurov, J.V. Sweedler, K.R. Weiss, Identification of a new neuropeptide precursor reveals a novel source of extrinsic modulation in the feeding system of Aplysia, J. Neurosci. 25 (2005) 9637-9648.

[2] J. Jing, F.S. Vilim, C.C. Horn, V. Alexeeva, N.G. Hatcher, K. Sasaki, I. Yashina, Y. Zhurov, I. Kupfermann, J.V. Sweedler, K.R. Weiss, From hunger to satiety: reconfiguration of a feeding network by Aplysia neuropeptide Y, J. Neurosci. 27 (2007) 3490-3502.

[3] S. Arora, Anubhuti, Role of neuropeptides in appetite regulation and obesity – A review, Neuropeptides 40 (2006) 375-401.

[4] P. Trayhurn, J.H. Beattie, Physiological role of adipose tissue: white adipose tissue as an endocrine and secretory organ, Proc. Nutr. Soc. 60 (2001) 329-339.

[5] E.B. De Souza, Corticotropin-releasing factor receptors: Physiology, pharmacology, biochemistry and role in central nervous system and immune disorders, Psychoneuroendocrinology 20 (1995) 789-819.

[6] J.M. Friedman, J.L. Halaas, Leptin and the regulation of body weight in mammals, Nature 395 (1998) 763-770.

[7] N.G. Seidah, M. Chrétien, Proprotein and prohormone convertases: a family of subtilases generating diverse bioactive polypeptides, Brain Res. 848 (1999) 45-62.

[8] A. Brockmann, S.P. Annangudi, T.A. Richmond, S.A. Ament, F. Xie, B.R. Southey, S.R. Rodriguez-Zas, G.E. Robinson, J.V. Sweedler, Quantitative peptidomics reveal brain peptide signatures of behavior., Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 2383-2388.

[9] N.G. Hatcher, N. Atkins, S.P. Annangudi, A.J. Forbes, N.L. Kelleher, M.U. Gillette, J.V. Sweedler, Mass spectrometry-based discovery of circadian peptides., Proc. Natl. Acad. Sci. U. S. A. 105 (2008) 12527-12532.

[10] J.E. Lee, N. Atkins, N.G. Hatcher, L. Zamdborg, M.U. Gillette, J.V. Sweedler, N.L. Kelleher, Endogenous peptide discovery of the rat circadian clock: a focused study of the suprachiasmatic nucleus by ultrahigh performance tandem mass spectrometry, Mol. Cell. Proteomics 9 (2010) 285-297.

[11] M. Svensson, K. Sköld, P. Svenningsson, P.E. Andren, Peptidomics-based discovery of novel neuropeptides, J. Proteome Res. 2 (2003) 213-219.

[12] L.D. Fricker, J. Lim, H. Pan, F.-Y. Che, Peptidomics: Identification and quantification of endogenous peptides in neuroendocrine tissues, Mass Spectrom. Rev. 25 (2006) 327-344.

[13] E.V. Romanova, J.V. Sweedler, Peptidomics for the discovery and characterization of neuropeptides and hormones, Trends Pharmacol. Sci. 36 (2015) 579-586.

[14] J.E. Elias, W. Haas, B.K. Faherty, S.P. Gygi, Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations., Nat. Methods 2 (2005) 667-675.

[15] E.V. Romanova, K. Sasaki, V. Alexeeva, F.S. Vilim, J. Jing, T.A. Richmond, K.R. Weiss, J.V. Sweedler, Urotensin II in invertebrates: from structure to function in Aplysia californica, PLoS One 7 (2012) e48764.

[16] S.S. Rubakhin, R.W. Garden, R.R. Fuller, J.V. Sweedler, Measuring the peptides in individual organelles with mass spectrometry, Nat. Biotechnol. 18 (2000) 172-175.

[17] A.B. Hummon, N.P. Hummon, R.W. Corbin, L. Li, F.S. Vilim, K.R. Weiss, J.V. Sweedler, From precursor to final peptides: a statistical sequence-based approach to predicting prohormone processing, J. Proteome Res. 2 (2003) 650-656.

[18] J.M. Fisher, W. Sossin, R. Newcomb, R.H. Scheller, Multiple neuropeptides derived from a common precursor are differentially packaged and transported, Cell 54 (1988) 813-822.

[19] The UniProt Consortium, UniProt: the universal protein knowledgebase, Nucleic Acids Res. 45(Database issue) (2017) D158-D169.

[20] D. Houde, Y. Peng, S.A. Berkowitz, J.R. Engen, Post-translational modifications differentially affect IgG1 conformation and receptor binding., Mol. Cell. Proteomics 9 (2010) 1716-1728.

[21] L.N. Rahman, G.S.T. Smith, V.V. Bamm, J.A.M. Voyer-Grant, B.A. Moffatt, J.R. Dutcher, G. Harauz, Phosphorylation of Thellungiella salsuginea dehydrins TsDHN-1 and TsDHN-2 facilitates cation-induced conformational changes and actin assembly, Biochemistry 50 (2011) 9587-9604.

[22] A. Zhang, P. Hu, P. MacGregor, Y. Xue, H. Fan, P. Suchecki, L. Olszewski, A. Liu, Understanding the conformational impact of chemical modifications on monoclonal antibodies with diverse sequence variation using hydrogen/deuterium exchange mass spectrometry and structural modeling, Anal. Chem. 86 (2014) 3468-3475.

[23] B.A. Eipper, D.A. Stoffers, R.E. Mains, The biosynthesis of neuropeptides: peptide α-amidation, Annu. Rev. Neurosci. 15 (1992) 57-85.

[24] G. Espadas, E. Borràs, C. Chiva, E. Sabidó, Evaluation of different peptide fragmentation types and mass analyzers in data-dependent methods using an Orbitrap Fusion Lumos Tribrid mass spectrometer, Proteomics 17 (2017) 1600416.

[25] C. Tu, J. Li, S. Shen, Q. Sheng, Y. Shyr, J. Qu, Performance investigation of proteomic Identification by HCD/CID fragmentations in combination with high/low-resolution detectors on a Tribrid, high-field Orbitrap instrument, PLoS One 11 (2016) e0160160.

[26] B.M. Balgley, T. Laudeman, L. Yang, T. Song, C.S. Lee, Comparative evaluation of tandem MS search algorithms using a target-decoy search strategy., Mol. Cell. Proteomics 6 (2007) 1599-1608.

[27] K. Jeong, S. Kim, N. Bandeira, False discovery rates in spectral identification, BMC Bioinformatics 13 Suppl 1 (2012) S2.

[28] J. Zhang, L. Xin, B. Shan, W. Chen, M. Xie, D. Yuen, W. Zhang, Z. Zhang, G.A. Lajoie, B. Ma, PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification, Mol. Cell. Proteomics 11 (2012) M111.010587-M010111.010587.

[29] J.E. Elias, S.P. Gygi, Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry, Nat. Methods 4 (2007) 207-214.

[30] R.W. Garden, S.A. Shippy, L. Li, T.P. Moroz, J.V. Sweedler, Proteolytic processing of the Aplysia egg-laying hormone prohormone., Proc. Natl. Acad. Sci. U. S. A. 95 (1998) 3972-3977.

[31] L. Li, R.W. Garden, P.D. Floyd, T.P. Moroz, J.M. Gleeson, J.V. Sweedler, L. Pasa-Tolic, R.D. Smith, Egg-laying hormone peptides in the Aplysiidae family, J. Exp. Biol. 202 (1999).

[32] S.S. Rubakhin, J.S. Page, B.R. Monroe, J.V. Sweedler, Analysis of cellular release using capillary electrophoresis and matrix assisted laser desorption/ionization-time of flight-mass spectrometry, Electrophoresis 22 (2001) 3752-3758.

[33] K. Ma, O. Vitek, A.I. Nesvizhskii, E. Paek, S.-W. Lee, K.-B. Hwang, M. Chambers, L. Zimmerman, K. Shaddox, S. Kim, A statistical model-building perspective to identification of MS/MS spectra with PeptideProphet, BMC Bioinformatics 13(Suppl 16) (2012) S1.

[34] B. Blank-Landeshammer, L. Kollipara, K. Biß, M. Pfenninger, S. Malchow, K. Shuvaev, R.P. Zahedi, A. Sickmann, Combining de novo peptide sequencing algorithms, a synergistic approach to boost both identifications and confidence in bottom-up proteomics, J. Proteome Res. 16 (2017) 3209-3218.

[35] G. Hart-Smith, D. Yagoub, A.P. Tay, R. Pickford, M.R. Wilkins, Large scale mass spectrometry-based identifications of enzyme-mediated protein methylation are subject to high false discovery rates, Mol. Cell. Proteomics 15 (2016) 989-1006.

[36] Y. Fu, X. Qian, Transferred subgroup false discovery rate for rare post-translational modifications detected by mass spectrometry, Mol. Cell. Proteomics 13 (2014) 1359-1368.

[37] H. Li, J. Park, H. Kim, K.-B. Hwang, E. Paek, Systematic comparison of false-discovery-rate-controlling strategies for proteogenomic search using spike-in experiments, J. Proteome Res. 16 (2017) 2231-2239.

[38] C. Ma, S. Xu, G. Liu, X. Liu, X. Xu, B. Wen, S. Liu, Improvement of peptide identification with considering the abundance of mRNA and peptide, BMC Bioinformatics 18 (2017) 109.

[39] A.S. Edison, E. Espinoza, C. Zachariah, Conformational ensembles: the role of neuropeptide structures in receptor binding., J. Neurosci. 19 (1999) 6318-6326.

[40] G. Duan, D. Walther, The roles of post-translational modifications in the context of protein interaction networks., PLoS Comp. Biol. 11 (2015) e1004049.

[41] G.k. Mutlu, P. Factor, Role of vasopressin in the management of septic shock, Intensive Care Med. 30 (2004) 1276-1291.

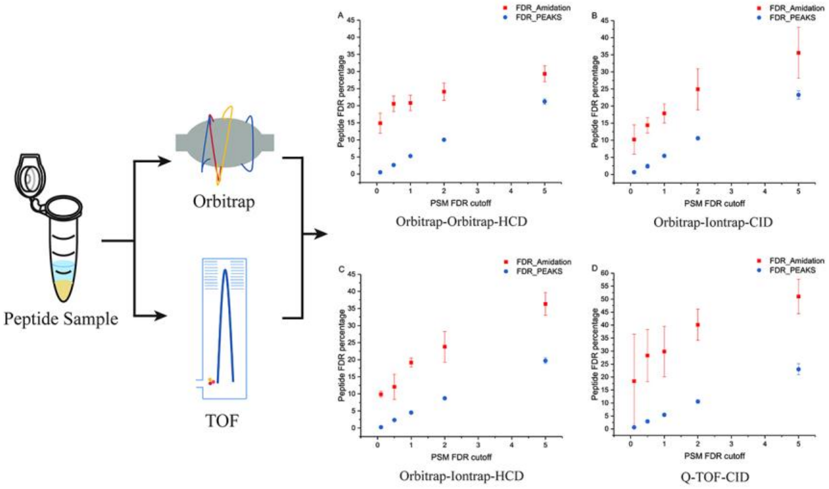
[42] A.P. Borrow, N.M. Cameron, The role of oxytocin in mating and pregnancy, Horm. Behav. 61 (2012) 266-276.

[43] G.M. Knudsen, R.J. Chalkley, The effect of using an inappropriate protein database for proteomic data analysis, PLoS One 6 (2011) e20873.

Highlights

- Evaluated neuropeptide identification approaches based on MS platform and software.
- Compared QTOF and Orbitrap platforms for confident neuropeptide identification.
- Determined reported false-discovery rates underestimate the identification errors.
- Evaluated the effects of protein database size on peptide identifications.
- Assessed the strengths and weaknesses of the workflows tested.

Graphical abstract


Peptide Sample
Orbitrap
TOF
A
B
C
D
FDR_Amidation
FDR_PEAKS
Peptide FDR percentage
PSM FDR cutoff
0
1
2
3
4
5
Orbitrap-Orbitrap-HCD
Orbitrap-Iontrap-CID
Orbitrap-Iontrap-HCD
Q-TOF-CID