Published in final edited form as:

J Proteome Res. 2023 October 06; 22(10): 3123–3134. doi:10.1021/acs.jproteome.2c00307.

Assessment and Comparison of Database Search Engines for Peptidomic Applications

Eduardo A. De La Toba[†,1,2], **Krishna D. B. Anapindi**[†,1,2], **Jonathan V. Sweedler**[*,1,2]

[1]Beckman Institute of Advanced Science and Technology, University of Illinois at Urbana-Champaign, 61801

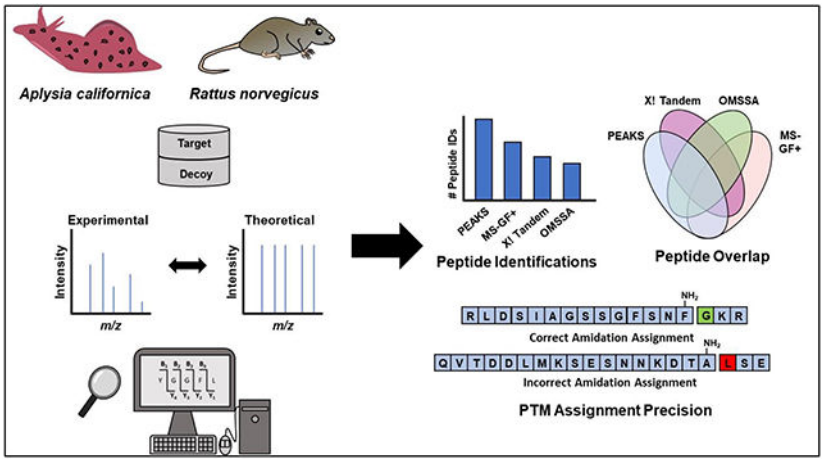
[2]Department of Chemistry, University of Illinois at Urbana-Champaign, 61801

Abstract

Protein database search engines are an integral component of mass spectrometry-based peptidomic analyses. Given the unique computational challenges of peptidomics, many factors must be taken into consideration when optimizing search engine selection, as each platform has different algorithms by which tandem mass spectra are scored for subsequent peptide identifications. In this study, four different database search engines, PEAKS, MS-GF+, OMSSA, and X! Tandem, were compared with *Aplysia californica* and *Rattus norvegicus* peptidomics datasets, and various metrics were assessed such as the number of unique peptide identifications, unique protein identifications, and peptide length distributions. Given the tested conditions, PEAKS was found to have the highest number of peptide, neuropeptide, and protein identifications out of the four search engines in both datasets. Furthermore, principal component analysis and multivariate logistic regression were employed to identify whether specific spectral features contribute to false C-terminal amidation assignments by each search engine. From this analysis, it was found that the primary feature influencing incorrect peptide assignments were the precursor and fragment ion *m/z* error tolerances. Finally, an assessment employing a mixed species database was also performed to evaluate search engine precision and sensitivity when searched against an enlarged search space containing human proteins.

Graphical Abstract

[*]Corresponding author: Jonathan V. Sweedler: (217) 244-7359, jsweedle@illinois.edu.
[†]Co-first author



Keywords

Peptidomics; Neuropeptides; Search Engines; Post-Translational Modifications; Database Searching; C-terminal Amidation

Introduction

Peptidomics is the study of characterizing the suite of endogenous peptides in a biological system, and these analyses are typically performed with the goal of understanding which peptides may be involved in a specific biological process or disease, and also identifying novel peptides as potential biomarkers.[1] In peptidomics, peptide identifications are performed in an untargeted manner, allowing for the identification of potentially hundreds to thousands of peptides within a sample. Liquid chromatography coupled to mass spectrometry (LC-MS) is typically the method by which peptidomic experiments are conducted. The peptides identified in these analyses comprise not only intact peptides with known biological functions, such as neuropeptides and peptide hormones, but also modified and truncated forms of these peptides along with fragments of more abundant cytosolic and structural proteins.[2] The basic workflow of most peptidomics workflows include sample stabilization, peptide extraction from the biological specimen and sample cleanup, chromatographic separation of the peptides, analysis with tandem mass spectrometry, and peptide identification from the acquired mass spectra for subsequent statistical and bioinformatic analyses.

A key feature of peptidomics is the computational step of inferring a peptide's amino acid sequence from its tandem mass spectrum. During this process, the obtained MS/MS spectra along with the target organism's protein database are imported into a database search engine, whereby the search engine performs an *in-silico* digestion of each protein in the protein database depending upon the user-selected enzyme cleavage parameters to generate a theoretical spectrum of the potential peptides from the database. However, with peptidomics, the enzyme cleavage parameter is typically not selected, enabling an *in-silico* search of the protein database without the constraint of enzyme-specific cleavage parameters to identify potential endogenous peptides from the spectra. However, this has the consequence of

increasing the search space, since all potential peptide combinations and fragments in the protein database must be considered, usually resulting in longer search times and a higher probability of false identifications compared to shotgun proteomics. The MS/MS spectra are then compared against candidate peptides from the database to generate potential peptide-spectrum matches (PSMs), and depending upon the search engine's scoring algorithm and the quality of the spectrum, these PSMs are assigned a score to reflect the confidence in which the spectra match the theoretical spectra.[3] These identification scores reflect various features of the mass spectra such as the fragment ion coverage of the peptide, the mass error for both precursor and fragment ions, and low levels of noise peaks in the spectrum.[4] Various open-source and commercial database search engines exist for this purpose and they share certain features that allow users to select for optimization of their analysis based on the instrument type used for the experiment. These features include selection of the *m/z* error tolerances for the precursor and fragment ions, selection of fixed and variable peptide modifications to search against, and selection of the fragmentation type used by the mass spectrometer (e.g., collision-induced dissociation or electron-transfer dissociation).

Most studies comparing the performance of database search engines have been applied for bottom-up proteomics applications, while few comparisons have been applied at the peptidomics level.[5–8] For instance, Akhtar *et al.* compared the performance of three different search engines to assess neuropeptide identification rates, but this study was focused on analyzing simulated peptide spectra which represents idealized MS/MS spectra which may not capture the spectral complexity of experimental spectra, nor did it assess the search engines' ability to identify peptides with various PTMs, which may comprise a significant portion of the endogenous peptides in a sample.[9] Peptidomics presents specific computational challenges that may not be as apparent in proteomics, such as the fact that endogenous peptides have various residue cleavage sites that cannot be completely captured by user-specific enzyme parameters, structures that may make them difficult to fragment by MS, and a multitude of possible PTMs. Furthermore some peptides have unknown protein precursors (e.g. endomorphin 1 and 2), and may only be identified through *de novo* approaches, placing additional computational demands on the software.[10–12] The wide dynamic range of endogenous bioactive peptides (some may be present at high levels and other will be present below the detection limits of the system, even for peptides from the same precursor) introduce another issue. Some bioactive peptides are short (for example 4 – 6 amino acids long), hindering the scoring algorithms used by some search engines. Lastly, the goal is not to identify the presence of the precursor protein but each individual peptide so that missing values become problematic. In the current study, four commonly used database search engines, PEAKS,[13] MS-GF+,[14] X! Tandem,[15] and OMSSA,[16] were compared to assess their performance for peptidomics from experimental tandem mass spectra. The advantage of comparing multiple database search engines lies in the fact that since search engines have different peptide scoring algorithms, some may be better suited towards identifying peptides with certain features, such as smaller or larger peptides, or peptides with specific PTMs. Finally, there is also the issue of the search engine's precision with respect to its ability to discern true hits from false hits. Thus, an assessment comparing the precision of the four search engines was also performed.

For this study, two previously published peptidomics datasets, one from the *Aplysia californica* abdominal ganglia (*Aplysia*)[17] and the other from *Rattus norvegicus* hypothalamus (rat)[18], were analyzed with each of the four search engines. Additionally, peptides that were C-terminally amidated, were subjected to further analysis to assess the occurrence of false amidation assignments. This PTM occurs in the secretory vesicles of endocrine cells by the action of an enzyme called peptidylglycine α-amidating monooxygenase (PAM), and is critical for imparting enhanced biological activity and stability to various endogenous peptides.[19] However, a requirement for enzymatic C-terminal amidation by PAM is that prior to amidation, the peptide must be C-terminally extended by a glycine residue in order to act as a substrate for this enzyme.[20] After recognition by PAM, the C-terminal glycine on the peptide is oxidatively cleaved to yield the α-amidated peptide and glyoxylic acid.[21] Despite this requirement though, some peptides identified by database search engines in peptidomic analyses are still assigned with this PTM, even if the C-terminally extended version of this peptide does not contain a glycine residue. This discrepancy therefore provides an efficient method to screen against peptides that have been incorrectly assigned with this PTM by checking whether these peptides possess the required glycine residue, which is an approach utilized by Anapindi et al.[17] Though there is a provision in certain search engines to input this PTM as a variable modification parameter that does take into the C-terminal Glycine requirement, this variable modification was not selected here, as the main purpose of this analysis was to compare the search engines in terms of their selectivity for discerning true biological amidated peptides. Since C-terminal amidation followed by glycine serves as a proxy for true biological peptides, we chose to not include it as a pre-defined criteria for the modification, as this provides a method by which known false positives can be accounted for. A multivariate analysis with principal component analysis (PCA) and multiple logistic regression was performed to determine whether certain MS/MS spectral features contribute to incorrect amidation assignments.

Materials and Methods

Tandem MS Data

Both the *Aplysia* and the rat datasets were analyzed from three different biological replicates. The three *Aplysia* abdominal ganglia mass spectra datasets analyzed by orbitrap-ion trap were the same datasets used in a previous study in our group, done by Anapindi et al.[17] More experimental details regarding the sample preparation and the LC-MS parameters can be seen there.

The three rat hypothalami mass spectra datasets were downloaded from the UCSD MassIVE repository, and the datasets used were from Ye et al,[18] specifically datasets titled “041213_HT_A2.raw”, “041213_HT_A3.raw”, and “041213_HT_A5.raw” (MassIVE MSV000080106).

Protein Databases

The *Aplysia* and rat protein databases used in this study were downloaded from UniProt. The *Aplysia californica* protein database contained 443 protein entries consisting of both

reviewed (Swiss-Prot) and unreviewed (TrEMBL) proteins. The *Rattus norvegicus* protein database contained only reviewed (Swiss-Prot) proteins and initially consisted of 8094 protein entries; however, to reduce the search times, this database was shortened to only contain proteins with a predicted signal peptide. In order to separate the secretory proteins from the non-secretory proteins, SignalP-5.0 Server[22] was employed, and this reduced the database size to 1566 proteins.

Database Searching with SearchGUI and PEAKS

SearchGUI (v. 4.1.1, http://compomics.github.io/projects/searchgui)[23] was used for searching the MS/MS spectra from both organism datasets with MS-GF+, OMSSA, and X! Tandem. The raw MS/MS files were converted to MGF format upon importing into SearchGUI. When the protein databases were imported into SearchGUI, the reversed decoy sequences were generated by the software and automatically appended to the end of the target protein database for false discovery rate (FDR) calculations by the target-decoy method. The precursor charge range was set to 1–7, isotopes 0–3 were selected, the digestion parameter was set to "Unspecific", the fragment ion types selected were "b" and "y", the FDR cutoff was set to 1%, and the peptide length range was set to 4 to 65 residues. The results for the SearchGUI analyses were visualized with PeptideShaker (v 2.2.0),[24] downloaded from http://compomics.github.io/projects/peptide-shaker. The search results were set to display b, b-H_2O, b-NH_3, y, y-H_2O, and y-NH_3 ions, and were exported as Excel files for further analysis.

PEAKS searches were performed with PEAKS Studio (v. 8), where a *de novo* search of the spectra was performed prior to the database search. The same MGF files used for the SearchGUI searches were also employed in the PEAKS searches. The following instrument parameters were selected: the ion source parameter was set to "ESI (nano-spray)," the fragmentation mode was set to "high-energy CID (y and b ions)," the MS scan mode was set to "FT-ICR (Orbitrap), and the MS/MS scan mode was set to "Linear Ion Trap." For the database search parameters, "none" was selected for the enzyme option, the maximum allowed variable PTMs per peptide was set to 3, and the option to estimate the FDR with decoy-fusion was selected, and an FDR cutoff of 1% was applied. Though the search engine suggested option of FDR was used to filter the initial results, a different FDR was empirically calculated from the percentage of false amidations in all the identified peptide. Additionally, PEAKS automatically searches for peptides up to 65 residues in length. The search results were set to display b, b^{2+}, b-H_2O, b-NH_3, y, y^{2+}, y-H_2O, and y-NH_3 ions. Search results were exported to excel files for further analysis.

For the both the SearchGUI and the PEAKS analyses, the following variable PTMs that that are commonly encountered in peptidomic analyses were selected: amidation of the peptide C-terminus (−0.98 Da), methionine oxidation (+15.99 Da), phosphorylation of tyrosine, threonine, and serine (+79.97 Da), and pyrrolidone from glutamic acid and glutamine (−18.01 Da and −17.03 Da, respectively). The precursor *m/z* tolerance was set to 10 ppm and the fragment *m/z* tolerance was set to 0.1 Da.

Data Processing and Statistical Analysis

OriginPro 2021 (v. 2021b, OriginLab Corporation, Northampton, MA, USA) was used to perform the statistical analysis of the collected search results and to make the figures. Results are presented as the mean and the standard deviation, where a one-way ANOVA was used to determine significance between groups by comparing the means with the Bonferroni test. For the principal component analysis, the correlation matrix of the different values for each tested variables were calculated, in which the column values were standardized. The logistic regression analysis was also performed with Origin Pro whereby multiple spectral features were used as covariates to create a binary prediction model with Boolean output. The maximum number of iterations was set to 100 and the Wald test was used to determine which independent variables were significantly different, while the Hosmer and Lemeshow test was used to determine the goodness of fit for the observed data relative to the expected results of the logistic regression model.

Results and Discussion

Peptide Identification Comparison Across Search Engines

The first metric that was evaluated to assess search engine performance was to compare the number of identified peptides across the different programs. For both the *Aplysia* and rat searches, the following numbers of unique peptides were identified by each search engine, respectively: PEAKS (517 ± 19.3 and 919 ± 103.3), MS-GF+ (238 ± 9.0 and 748 ± 100.9), OMSSA (66 ± 29.7 and 417 ± 55.6), and X! Tandem (83 ± 22.9 and 533 ± 41.8) (Figure 1A and 1B). Furthermore, given that in peptidomics, many identified peptides comprise truncated forms of bioactive peptides or fragments of more abundant structural proteins, a comparison between the numbers of identified neuropeptides (NPs) was also assessed to determine which search engine had the highest sensitivity for identifying these this class of peptides. To perform this comparison, previously reported NPs for both organisms were downloaded from the NP database NeuroPep (v. 1.0),[25] and the identified peptides were searched against this list to find NP hits. When performing this comparison, only unique NPs were considered; in other words, if a NP was identified in multiple forms with different PTMs, it would only be considered as one NP. A similar trend for NP identifications in terms of sensitivity was observed as when comparing total peptide identifications, with PEAKS identifying the following number of NPs (*Aplysia*: 63 ± 3.3, rat: 59 ± 3.1), followed by MS-GF+ (*Aplysia*: 26 ± 2.4, rat: 49 ± 1.3), X! Tandem (*Aplysia*: 13 ± 2.9, rat: 36 ± 2.2) and OMSSA (*Aplysia*: 11 ± 2.4, rat: 29 ± 2.2) (Figure 1C and 1D).

Regarding the choices of PTMs that were included in the search parameters, the most commonly reported for endogenous peptides were selected. For instance, C-terminal amidation and N-terminal pyrrolidone formation from glutamine and glutamic acid are commonly observed modifications in peptidomics and they possess biological significance due to their role in enhancing the bioactivity of many endogenous peptides, while also enhancing the resistance of these peptides against degradation by proteolytic enzymes.[19,26] Phosphorylation is also a widely-reported PTM on serine, tyrosine, and threonine residues in many endogenous peptides, and since the original mass spectra that were analyzed here were acquired by fragmentation by higher energy collisional dissociation, this fragmentation

mode has been shown to be suitable for confidently identifying phosphorylated peptides by MS/MS.[27]

Among the search engines that were evaluated here, PEAKS is unique out of the four in that it employs a *de novo* sequencing step of the spectra prior to performing the database search. This step generates *de novo* sequence tags of the amino acids in a spectrum based on the spectral quality, followed by a protein shortlisting step whereby only the proteins in the database that contain the *de novo* sequence tags will be searched against. Finally, a peptide shortlisting step is performed whereby the proteins in the previously generated protein shortlist are *in silico* digested, and the MS/MS spectra are searched against these theoretical peptides to generate PSMs, and the PSMs are scored according to their spectral quality.[13] Thus, it is likely that the higher identification number of peptides and neuropeptides for both datasets by PEAKS can partly be attributed to the incorporation of this *de novo* sequencing step, as it facilitates peptide identifications by filtering out protein candidates from spectra that did not possess high-quality *de novo* sequence tags, thereby reducing the search space. Interestingly, each of the search engines identified significantly more peptides in the rat datasets than in the *Aplysia* datasets, though this observation can likely be attributed to the fact that the rat protein database was larger and hence more MS/MS spectra could be matched to more proteins in the database, and could also be due to differences in spectral quality since the datasets were from different sources.

Peptide Overlap Among Search Engines

The percentage of peptide overlap between the four search engines was also assessed by combining the peptides from each replicate and analyzing how many peptides were uniquely identified by each search engine and shared across all four. This analysis could be beneficial to assess whether multiple search engines could be used for peptidomics to improve peptide identification coverage. For the number of identified peptides unique to a particular search engine, PEAKS identified 431 unique peptides in the *Aplysia* data (Figure 2A) and 577 peptides in the rat data (Figure 2B), comprising 54.9% and 27.9%, respectively, of the total number of identified peptides across the four search engines. On the other hand, in both the *Aplysia* and rat datasets, X! Tandem had the lowest number of unique peptide identifications, comprising 2.0% and 0.3%, respectively, of the total number of peptides identified. Additionally, there was a moderate level of peptide overlap shared between all search engines for the *Aplysia* and rat data, with 8.7% and 18.5% overlap, respectively.

Additionally, when comparing the percentages of the total peptides, instead of unique peptides, identified by a search engine in relation to the combined number of peptide identifications across all four search engines in the *Aplysia* data, the percentages of the identified peptides were 91%, 41%, 16%, and 14% for PEAKS, MS-GF+, X! Tandem, and OMSSA, respectively. Similarly in the rat datasets, the percentages of identified peptides were 79%, 59%, 43%, and 34% for PEAKS, MS-GF+, X! Tandem, and OMSSA, respectively.

When comparing peptide overlap between two combined search engine results, in both organism datasets, PEAKS and MS-GF+ combined searches had the highest total number of identified peptides out of the different combinations of two search engines. These combined

searches yielded 99.0% and 94.4% of the total identified peptides out of the four search engines in the *Aplysia* and rat datasets, respectively, confirming the notion that multiple search engines can be used to improve peptide coverage.

Peptide Length Distribution Among Search Engines

The peptide length distribution for the search results from each of the search engines was evaluated in order to assess whether the search engines were biased towards longer or shorter peptides. Endogenous peptides may have wide differences in peptide lengths, ranging from approximately 2 residues up to over 90 residues. Therefore, if different search engines are skewed towards identifying longer or shorter peptides, certain peptides may be missed, compromising peptide coverage. For this analysis, the lengths of the peptides in each replicate were combined and duplicates were removed. For each of the search engines, the largest density of peptides for both organisms was in the range of 10 to 20 residues, and the median peptide lengths were between 14 and 22 residues. For both organisms, X! Tandem had the shortest range of peptide lengths, with the peptides ranging from 9 to 46 residues in the *Aplysia* results (Figure 3A), and 7 to 51 residues in the rat results (Figure 3B). Conversely, MS-GF+ had the largest range of peptide lengths in the rat results, with a range of 6 to 65 residues, and in the *Aplysia* results, PEAKS had the largest range of peptide lengths, with a range of 5 to 60 residues. These observations are not surprising, as fragmentation efficiency is typically less efficient for larger peptides, resulting in less complete ion series in the spectra and subsequently lower identification rates due to reduced peptide fragment ion coverage.[28] The decreasing frequency of peptide identifications with shorter sequences (e.g. 5–9 residues) can partly be attributed to the fact that smaller peptides are more likely to be singly charged, reducing fragmentation as compared to multiply charged peptides especially if basic residues such as arginine or lysine are present, resulting in low ion series production for confident peptide assignments.[29] Furthermore, while some short peptides, such as enkephalins, are known to exert potent biological effects, various forms of peptide processing are often observed in peptidomics, where the extended forms of these peptides are observed along with peptide fragments of more abundant structural proteins, thereby potentially skewing the peptide length distribution towards longer peptides in the range of 10–20 residues.[30,31]

Assessment of Falsely Assigned Amidation to Peptides

Another main aspect of this work was the evaluation of the occurrence of falsely assigned C-terminal amidation of the peptides. This PTM requires the presence of a glycine residue on the C-terminus of the immature peptide for conversion into the amidated peptide by the enzyme PAM; however, a common occurrence in peptidomic studies is the observation that some peptides that do not follow this requirement are still assigned with this PTM. This observation was described in detail by Anapindi et al, whereby the rate of occurrence of these false amidations was evaluated across different MS analyzers.[17] Here, amidated peptides were classified as incorrectly assigned if the amidated peptide was not flanked by a C-terminal glycine residue, and the percentage of falsely assigned amidated peptides relative to the total number of assigned amidated peptides was calculated.

For the *Aplysia* search results, MS-GF+ and PEAKS had the same average percentage of falsely amidated peptides relative to the total number of amidated peptides (12.4 ± 6.9% and 12.4 ± 1.7%, respectively, Figure 4A). X! Tandem and OMSSA had the lowest and second lowest average percentage of falsely amidated peptides, respectively; however, with both search engines, only one sample contained false amidation assignments. For the rat search results, X! Tandem and OMSSA had similar percentages of falsely assigned amidated peptides relative to the total number of amidated peptides (20.3 ± 6.0%, and 20.2 ± 2.9%, respectively) and PEAKS had the lowest average percentage of false amidation assignments (8.2 ± 3.2%), while MS-GF+ had an average percentage of 10.2 ± 4.4 (Figure 4B). The total amidated peptides for each of the search engine and sample type are included in the supplementary information (Supplementary Table 1).

While we state that a peptide assigned as amidated without its required glycine-containing precursor is not possible, a few caveats must be addressed. For instance, peptide amidation has been reported to occur in bacteria through enzymes other than PAM which do not require the C-terminal glycine residue on the extended peptide.[32] Furthermore, non-enzymatic routes of peptide amidation have been reported *in vitro* for peptides extended by serine, threonine, or cysteine residues.[33,34] However, these biosynthetic routes for peptide amidation have not been reported for mammals or mollusks in the intervening decades. In support of chemical modification not being a significant pathway in this study, the majority of peptides assigned as amidated that were not extended by glycine, were mostly extended by residues other than serine, threonine, or cysteine, suggesting that these assignments were likely incorrect, as indicated by the sequence logos for both true and false amidated peptides identified by each search engine in both animal datasets (Supplementary Figures 1 and 2).

While the average percentage of falsely amidated peptides between PEAKS and MS-GF+ was the same in the *Aplysia* set, PEAKS had the lowest false identification rate in the rat dataset. This may be since PEAKS employs a different method of calculating the FDR compared to MS-GF+, OMSSA, and X! Tandem, all of which employ the target-decoy method in SearchGUI to calculate the FDR, whereas PEAKS employs the decoy-fusion method for FDR calculations. The target-decoy method relies on the concatenation of either the reversed or shuffled protein sequences (decoys) to the end of the target protein database, effectively doubling the size of the database.[35] On the other hand, the decoy-fusion technique adds the decoy protein sequence to the end of the target protein sequence, keeping the number of entries in the database the same. The decoy-fusion approach has been shown to be a more accurate method of FDR estimation, due to the claim that the target-decoy strategy can potentially underestimate the FDR.[13,36] Therefore, the implementation of the more accurate decoy-fusion strategy in PEAKS may contribute to fewer poor quality spectra crossing the FDR threshold, reducing the overall percentage of false amidation assignments.

Principal Component Analysis of Amidated Peptides

While these results indicate that there are differences between the search engines with regards to their ability to accurately assign peptides with a given PTM, they do not explain which spectral features are responsible for these false positive hits. Therefore, to assess which factors in the mass spectra may contribute to false amidation assignments, PCA was

employed to determine if true and false amidated peptides could be distinguished based on various spectral features. This assessment was performed for the rat datasets only due to having a higher number of amidated peptides identified overall, whereby the search results for all three rat datasets were combined for each search engine dataset. For performing these analyses, both singly and doubly charged b and y ions were included for analysis, along with singly charged b and y ions containing water and ammonia losses for the MS-GF+, OMSSA, and X! Tandem analyses. However, with PEAKS, only singly and double charged b and y ions without neutral losses were used in the analyses due to better visualization of the true and false amidated peptides by PCA. PCA was also performed on the individual rat datasets with these same parameters, and the biplots for those analyses can be seen in the supplementary information (Supplementary Figures 3A–D). The spectrum features that were evaluated for PCA can be seen in Table 1. Though the extent of the impact of these features is workflow dependent, the list itself comprises of factors that would have the most impact on peptide identification by search engines. Hence, the importance of these factors might vary based on the workflow; however, the list of features still represents the most impactful factors towards peptide identifications. We also discussed the impact of MS platforms on false peptide identifications in one of our previous works (Anapindi et al).[17] The PCA biplots indicate that true amidated peptides can largely be distinguished from the false amidated peptides, with over 50% combined variation in PCs 1 and 2 for MS-GF+, OMSSA, and X! tandem, and PCs 1 and 3 for PEAKS (Figure 5A–D). Additionally, the PCA loadings vectors indicate that for each search engine, the absolute median and mean fragment ion *m/z* errors contribute highly to the false amidation assignments, with the incorrect assignments having a larger magnitude of *m/z* error (ppm). Additionally, for each of the search engines except PEAKS, the peptide precursor *m/z* error also plays a role in distinguishing the true and false amidated peptides

Multivariate Logistic Regression Analysis of Spectral Features to Predict True and False Amidation Assignments

Due to the dichotomous classification of correct or incorrect amidation assignments, multivariate logistic regression was also performed in addition to PCA, to determine if the status of amidated peptides, either true or false, could be predicted based on the evaluation of various spectral features seen in Table 1. To perform the logistic regression analysis, all features were originally included in the model as the independent variables, with the amidation status of the peptide (true or false) serving as the dependent variable. However, to avoid issues of multicollinearity such as reduced precision and wider confidence intervals of the variable coefficients, the features ‘B Ion Intensity %’ and ‘Y Ion Intensity %’ were excluded from this analysis.[37,38] Furthermore, if other highly correlated variables (Pearson correlation ≥ |0.70|) were present in the correlation matrix between the different variables, the variable that had the higher p value between the two highly correlated variables was also excluded from the analysis. The variables included in the logistic regression model for each search engine can be found in Supplementary Tables 2–5.

The logistic regression analysis results largely agreed with the PCA results, whereby for both MS-GF+ and X! Tandem, the peptide precursor *m/z* error (ppm) was found to be significantly associated ($p < 0.05$) with incorrectly assigned amidated peptides (odds

ratio = 1.925 and 1.628 for MS-GF+ and X! Tandem, respectively). Furthermore, for X! Tandem and OMSSA, the mean or median fragment ion *m/z* error (ppm), respectively, were found to be significantly associated with false amidations (odds ratios = 1.051 and 1.161, respectively), whereas for PEAKS both the mean and median fragment ion *m/z* errors were found to be significant variables (odds ratios = 1.070 and 1.166, respectively).

With respect to the accuracy of the different logistic regression models for each search engines, where accuracy here is defined as the number of correctly predicted peptides (true or false amidation) relative to the total number of peptides with the observed outcome, all displayed an accuracy of over 95% correct predictions for true amidated peptides (100.00%, 99.19%, 98.25%, and 97.10% for PEAKS, MS-GF+, X! Tandem, and OMSSA, respectively), and accuracies over 60% for the false amidation assignments (80.00%, 61.54%, 71.43, and 82.35% for PEAKS, MS-GF+, X! Tandem, and OMSSA, respectively). For these models, a threshold for a predicted probability of 0.5 was implemented, whereby falsely amidated peptides were considered correctly predicted by the logistic regression model as false amidations if their probabilities were calculated at above 0.5, whereas true amidated peptides were classified as correctly predicted if their probabilities were calculated at under 0.5. The resulting predicted probabilities calculated for each logistic regression for each search engine were then plotted as a function of their rank, whereby the predicted probabilities were assigned a rank in order of increasing predicted probabilities (Figure 6).

While the accuracy of these logistic regression models indicates a distinction between the true and false amidated peptides, some of the regression models for some of the search engines clearly did not perform as well compared to others. One of the main limitations of this analysis is that is that while we know that certain peptides cannot be true, e.g., those without the required glycine, we do not know for certain which amidated peptides are true. Here we assume that any amidated peptide that is C-terminally flanked by glycine is a true peptide; however, this may not necessarily be true because a peptide that did possess the glycine residue and was assigned as amidated could have potentially still been a false positive with a sufficiently high score to pass the FDR threshold. However, measuring the false positive rate is difficult in this scenario since only a potential subset of all the false positives is known, and it is therefore possible that the false amidation rate is being underestimated by our assumption, which may account for some of the true amidated peptides being classified as false amidated peptides by the logistic regression model. Further accuracy of these logistic regression models could potentially be achieved if other spectral features that were not included in this analysis were evaluated such as the number of continuous ion series in a spectrum; however, this assessment indicates that a relatively high level of accuracy can be obtained when selecting these various spectral features as predictors of true and false amidation.

Assessment of Search Engine Precision with a Mixed Species Protein Database

The performance of the four database search engines was also assessed based on the precision demonstrated when identifying peptides from the target organism database using the *Aplysia* datasets. To evaluate this parameter, a mixed database search approach as previously described by Anapindi et al, was employed, which consisted of 441 *Aplysia*

proteins concatenated with 2000 randomly shuffled human proteins.[17] One of the main disadvantages of artificially expanding the protein database as was done here, is that this can often lead to an increase in the number of spectra that are incorrectly matched to a peptide, resulting in an increased number of false assignments that could potentially cross the FDR threshold, in addition to an overall reduction in the sensitivity of the search space when searching against a larger database.[3,40] These issues are also magnified in peptidomics since unspecific protein cleavage parameters are typically selected for the search parameters, resulting in an even larger combination of peptides that could be matched to the spectra.[11] Therefore, utilizing this mixed organism protein database approach provides a convenient method to assess the precision of the search engines, based on the assumption that peptides derived from human proteins should not be present in the *Aplysia* extracts, and subsequently, human peptide matches can be considered false positives. For these assessments, the *Aplysia* spectra were searched against the *Aplysia*-human mixed database and no variable PTMs were included in the search parameters to prevent an even larger increase of the search space. To estimate the precision of the different search engines, the percentage of *Aplysia* peptides identified relative to the total number of identified human and *Aplysia* peptides was calculated at various PSM FDR thresholds: 0.1, 0.5, 1, 2, 3, and 5 %, with human peptide matches considered false.

The effect of searching the *Aplysia* spectra against either the *Aplysia* or *Aplysia* + human protein database was first evaluated to determine if the larger search space would influence the total number of identified *Aplysia* peptides. When comparing the number of peptides identified, the average number of identifications across all search engines was higher when searching against only the *Aplysia* database, though the difference between the number of peptide identifications was only significant ($p < 0.01$) in the case of PEAKS (Figure 7A). The decreased number of identifications when searched against the mixed database is not surprising given the fact this database had a percent increase in the number of proteins entries of approximately 450% compared to the *Aplysia*-only database, and this decrease in sensitivity with an enlarged search space was previously demonstrated by Anapindi et al. The percent decrease for the number of peptide identifications was also varied across the search engines, with OMSSA having the highest percent decrease of 23.8% (112.3 ± 9.8 to 88.0 ± 24.1 peptides), while X! Tandem had the lowest percent decrease of 5.6% (47.3 ± 11.3 to 44.7 ± 15.1 peptides).

For the search engine precision assessments across several FDR thresholds, it was observed that MS-GF+, X! Tandem, and OMSSA performed similarly with a slow decrease in precision at higher FDR thresholds, where the precision remained above 95% with each of the tested FDR thresholds (Supplementary Table 6). In the case of X! Tandem, the highest precision was calculated, with no human peptides assigned until the 3% FDR. PEAKS, on the other hand, had the sharpest decrease in precision with a relatively linear decrease in precision as the FDR threshold increased and a precision value of 83.7 ± 3.8% at the 5% FDR threshold (Figure 7B). These results suggest that while PEAKS has the greatest overall sensitivity out of the four tested search engines, care must be taken at the higher FDR thresholds, as there will be a higher proportion of false positives relative to the theoretical FDR determined by the search engine. This was especially pronounced at the highest FDR % tested, as when the precision was tested with PEAKS at this FDR, approximately 14%

of the identified peptides were incorrectly mapped to human proteins. Such a pronounced reduction in precision was not observed with the other three search engines; however, this is likely attributed to the observation that the other search engines had a lower sensitivity and thus had an overall lower number of spectra assigned. Given this information though, one may suspect that the higher sensitivity of PEAKS, as was described in the first section of this manuscript comparing the number of peptide identifications across the search engines, may simply be due to PEAKS identifying a higher number of false positives; however, with this mixed species database search comparison, it can be noted that while the average precision of PEAKS at the FDR cutoff of 1% was the lowest (97.2 ± 0.4%) of the four search engines, it was only lower by 1.71, 2.16, and 2.81% when compared to MS-GF+, OMSSA, and X! Tandem, respectively. Furthermore, in the previous peptide identification comparisons, the search results were filtered to an FDR cutoff of 1%; therefore, if the assumption is made that the rate of human hits in the mixed database searches is representative of the false positive rate in the previous searches, the lower precision of PEAKS is not enough to suggest that the higher number of peptide identifications by PEAKS is due solely to a higher number of false identifications. However, unless there is an obvious situation where a peptide is known to be false (such as with amidated peptides), it is difficult to determine specifically which peptides are false positives, and therefore these combined-species database precision assessments provide a method to estimate how the false positive rate is influenced when a less stringent FDR is implemented. No investigation was conducted to assess specifically why certain spectra were incorrectly mapped to peptides from human proteins, though it is likely that this is due to poor spectral quality as was seen in the false amidation assessment.

Conclusions

In summary, four commonly used database search engines were compared for two different peptidomics datasets. The results obtained in this study demonstrate the advantage of incorporating database searches with multiple search engines for obtaining higher peptide coverage levels in peptidomic applications. Furthermore, a commonly observed issue regarding C-terminally amidated peptides was evaluated as well as the precision of the search engines when searching against a mixed species database. The amidation results are useful to determine true error rates when evaluating search engine performance. Overall, this information can help guide the selection of the specific search engines to employ for peptidomics experiments when considering the balance between total peptide identifications, PTM accuracy, and search precision.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was funded by the National Institute on Drug Abuse under Award No. P30 DA018310. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

References

(1). Anapindi KDB; Romanova EV; Checco JW; Sweedler JV Mass Spectrometry Approaches Empowering Neuropeptide Discovery and Therapeutics. Pharmacol. Rev 2022, 74 (3), 662–679. 10.1124/pharmrev.121.000423. [PubMed: 35710134]

(2). Romanova EV; Sweedler JV 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]

(3). Verheggen K; Ræder H; Berven FS; Martens L; Barsnes H; Vaudel M Anatomy and Evolution of Database Search Engines—a Central Component of Mass Spectrometry Based Proteomic Workflows. Mass Spectrom. Rev 2020, 39 (3), 292–306. https://doi.org/10.1002/mas.21543. [PubMed: 28902424]

(4). Révész Á; Milley MG; Nagy K; Szabó D; Kalló G; Csősz É; Vékey K; Drahos L Tailoring to Search Engines: Bottom-Up Proteomics with Collision Energies Optimized for Identification Confidence. J. Proteome Res 2021, 20 (1), 474–484. 10.1021/acs.jproteome.0c00518. [PubMed: 33284634]

(5). Audain E; Uszkoreit J; Sachsenberg T; Pfeuffer J; Liang X; Hermjakob H; Sanchez A; Eisenacher M; Reinert K; Tabb DL; Kohlbacher O; Perez-Riverol Y In-Depth Analysis of Protein Inference Algorithms Using Multiple Search Engines and Well-Defined Metrics. J. Proteomics 2017, 150, 170–182. https://doi.org/10.1016/j.jprot.2016.08.002. [PubMed: 27498275]

(6). Yuan Z-F; Lin S; Molden RC; Garcia BA Evaluation of Proteomic Search Engines for the Analysis of Histone Modifications. J. Proteome Res 2014, 13 (10), 4470–4478. 10.1021/pr5008015. [PubMed: 25167464]

(7). Amir SH; Yuswan MH; Aizat WM; Mansor MK; Desa MNM; Yusof YA; Song LK; Mustafa S Comparative Database Search Engine Analysis on Massive Tandem Mass Spectra of Pork-Based Food Products for Halal Proteomics. J. Proteomics 2021, 241, 104240. https://doi.org/10.1016/j.jprot.2021.104240. [PubMed: 33894373]

(8). Shteynberg D; Nesvizhskii AI; Moritz RL; Deutsch EW Combining Results of Multiple Search Engines in Proteomics. Mol. Cell. Proteomics 2013, 12 (9), 2383–2393. 10.1074/mcp.R113.027797. [PubMed: 23720762]

(9). Akhtar MN; Southey BR; Andrén PE; Sweedler J V; Rodriguez-Zas, S. L. 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]

(10). Hummon AB; Huang H-Q; Kelley WP; Sweedler JV A Novel Prohormone Processing Site in Aplysia Californica: The Leu–Leu Rule. J. Neurochem 2002, 82 (6), 1398–1405. https://doi.org/10.1046/j.1471-4159.2002.01070.x. [PubMed: 12354287]

(11). Maes E; Oeyen E; Boonen K; Schildermans K; Mertens I; Pauwels P; Valkenborg D; Baggerman G The Challenges of Peptidomics in Complementing Proteomics in a Clinical Context. Mass Spectrom. Rev 2019, 38 (3), 253–264. https://doi.org/10.1002/mas.21581. [PubMed: 30372792]

(12). Matsushima A; Sese J; Koyanagi KO Biosynthetic Short Neuropeptides: A Rational Theory Based on Experimental Results for the Missing Pain-Relief Opioid Endomorphin Precursor Gene. ChemBioChem 2019, 20 (16), 2054–2058. https://doi.org/10.1002/cbic.201900317. [PubMed: 31269328]

(13). 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. Mol. Cell. Proteomics 2012, 11 (4), M111.010587–M111.010587. 10.1074/mcp.M111.010587.

(14). Kim S; Pevzner PA MS-GF+ Makes Progress towards a Universal Database Search Tool for Proteomics. Nat. Commun 2014, 5 (1), 5277. 10.1038/ncomms6277. [PubMed: 25358478]

(15). Craig R; Beavis RC TANDEM: Matching Proteins with Tandem Mass Spectra. Bioinformatics 2004, 20 (9), 1466–1467. 10.1093/bioinformatics/bth092. [PubMed: 14976030]

(16). Geer LY; Markey SP; Kowalak JA; Wagner L; Xu M; Maynard DM; Yang X; Shi W; Bryant SH Open Mass Spectrometry Search Algorithm. J. Proteome Res 2004, 3 (5), 958–964. 10.1021/pr0499491. [PubMed: 15473683]

(17). Anapindi KDB; Romanova EV; Southey BR; Sweedler JV Peptide Identifications and False Discovery Rates Using Different Mass Spectrometry Platforms. Talanta 2018, 182, 456–463. https://doi.org/10.1016/j.talanta.2018.01.062. [PubMed: 29501178]

(18). Ye H; Wang J; Tian Z; Ma F; Dowell J; Bremer Q; Lu G; Baldo B; Li L Quantitative Mass Spectrometry Reveals Food Intake-Induced Neuropeptide Level Changes in Rat Brain: Functional Assessment of Selected Neuropeptides as Feeding Regulators. Mol. Cell. Proteomics 2017, 16 (11), 1922–1937. 10.1074/mcp.RA117.000057. [PubMed: 28864778]

(19). Eipper BA; Stoffers DA; Mains RE The Biosynthesis of Neuropeptides: Peptide Alpha-Amidation. Annu. Rev. Neurosci 1992, 15, 57–85. 10.1146/annurev.ne.15.030192.000421. [PubMed: 1575450]

(20). Kim K-H; Seong BL Peptide Amidation: Production of Peptide Hormonesin Vivo Andin Vitro. Biotechnol. Bioprocess Eng 2001, 6 (4), 244–251. 10.1007/BF02931985.

(21). Chufán EE; De M; Eipper BA; Mains RE; Amzel LM Amidation of Bioactive Peptides: The Structure of the Lyase Domain of the Amidating Enzyme. Structure 2009, 17 (7), 965–973. 10.1016/j.str.2009.05.008. [PubMed: 19604476]

(22). Almagro Armenteros JJ; Tsirigos KD; Sønderby CK; Petersen TN; Winther O; Brunak S; von Heijne G; Nielsen H SignalP 5.0 Improves Signal Peptide Predictions Using Deep Neural Networks. Nat. Biotechnol 2019, 37 (4), 420–423. 10.1038/s41587-019-0036-z. [PubMed: 30778233]

(23). Barsnes H; Vaudel M SearchGUI: A Highly Adaptable Common Interface for Proteomics Search and de Novo Engines. J. Proteome Res 2018, 17 (7), 2552–2555. 10.1021/acs.jproteome.8b00175. [PubMed: 29774740]

(24). Vaudel M; Burkhart JM; Zahedi RP; Oveland E; Berven FS; Sickmann A; Martens L; Barsnes H PeptideShaker Enables Reanalysis of MS-Derived Proteomics Data Sets. Nat. Biotechnol 2015, 33 (1), 22–24. 10.1038/nbt.3109. [PubMed: 25574629]

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

(26). Hook V; Lietz CB; Podvin S; Cajka T; Fiehn O Diversity of Neuropeptide Cell-Cell Signaling Molecules Generated by Proteolytic Processing Revealed by Neuropeptidomics Mass Spectrometry. J. Am. Soc. Mass Spectrom 2018, 29 (5), 807–816. 10.1007/s13361-018-1914-1. [PubMed: 29667161]

(27). Ferries S; Perkins S; Brownridge PJ; Campbell A; Eyers PA; Jones AR; Eyers CE Evaluation of Parameters for Confident Phosphorylation Site Localization Using an Orbitrap Fusion Tribrid Mass Spectrometer. J. Proteome Res 2017, 16 (9), 3448–3459. 10.1021/acs.jproteome.7b00337. [PubMed: 28741359]

(28). Fricker LD Limitations of Mass Spectrometry-Based Peptidomic Approaches. J. Am. Soc. Mass Spectrom 2015, 26 (12), 1981–1991. 10.1007/s13361-015-1231-x. [PubMed: 26305799]

(29). Paizs B; Suhai S Fragmentation Pathways of Protonated Peptides. Mass Spectrom. Rev 2005, 24 (4), 508–548. 10.1002/mas.20024. [PubMed: 15389847]

(30). De La Toba EA; Bell SE; Romanova EV; Sweedler JV Mass Spectrometry Measurements of Neuropeptides: From Identification to Quantitation. Annu. Rev. Anal. Chem. (Palo Alto. Calif) 2022, 15 (1), 83–106. 10.1146/annurev-anchem-061020-022048. [PubMed: 35324254]

(31). Hook V; Bandeira N Neuropeptidomics Mass Spectrometry Reveals Signaling Networks Generated by Distinct Protease Pathways in Human Systems. J. Am. Soc. Mass Spectrom 2015, 26 (12), 1970–1980. 10.1007/s13361-015-1251-6. [PubMed: 26483184]

(32). Leisico F; Vieira DV; Figueiredo TA; Silva MCabrita EJSobral RGLudovice AMTrincão JRomão MJde Lencastre HSantos-Silva T First Insights of Peptidoglycan Amidation in Gram-Positive Bacteria - the High-Resolution Crystal Structure of Staphylococcus Aureus Glutamine Amidotransferase GatD. Sci. Rep 2018, 8 (1), 5313. 10.1038/s41598-018-22986-3. [PubMed: 29593310]

(33). Ranganathan D; Saini S Transformation of C-Terminal Serine and Threonine Extended Precursors into C-Terminal .Alpha.-Amidated Peptides: A Possible Chemical Model for the .Alpha.-Amidating Action of Pituitary Enzymes. J. Am. Chem. Soc 1991, 113 (3), 1042–1044. 10.1021/ja00003a048.

(34). Rink R; Arkema-Meter A; Baudoin I; Post E; Kuipers A; Nelemans SA; Akanbi MHJ; Moll GN To Protect Peptide Pharmaceuticals against Peptidases. J. Pharmacol. Toxicol. Methods 2010, 61 (2), 210–218. https://doi.org/10.1016/j.vascn.2010.02.010. [PubMed: 20176117]

(35). Wang G; Wu WW; Zhang Z; Masilamani S; Shen R-F Decoy Methods for Assessing False Positives and False Discovery Rates in Shotgun Proteomics. Anal. Chem 2009, 81 (1), 146–159. 10.1021/ac801664q. [PubMed: 19061407]

(36). Ivanov MV; Levitsky LI; Gorshkov MV Adaptation of Decoy Fusion Strategy for Existing Multi-Stage Search Workflows. J. Am. Soc. Mass Spectrom 2016, 27 (9), 1579–1582. 10.1007/s13361-016-1436-7. [PubMed: 27349255]

(37). Midi H; Sarkar SK; Rana S Collinearity Diagnostics of Binary Logistic Regression Model. J. Interdiscip. Math 2010, 13 (3), 253–267. 10.1080/09720502.2010.10700699.

(38). Ranganathan P; Pramesh CS; Aggarwal R Common Pitfalls in Statistical Analysis: Logistic Regression. Perspect. Clin. Res 2017, 8 (3), 148–151. 10.4103/picr.PICR_87_17. [PubMed: 28828311]

(39). Eng JK; Searle BC; Clauser KR; Tabb DL A Face in the Crowd: Recognizing Peptides through Database Search. Mol. Cell. Proteomics 2011, 10 (11), R111.009522–R111.009522. 10.1074/mcp.R111.009522.

(40). Tanca A; Palomba A; Deligios M; Cubeddu T; Fraumene C; Biosa G; Pagnozzi D; Addis MF; Uzzau S Evaluating the Impact of Different Sequence Databases on Metaproteome Analysis: Insights from a Lab-Assembled Microbial Mixture. PLoS One 2013, 8 (12), e82981. 10.1371/journal.pone.0082981. [PubMed: 24349410]

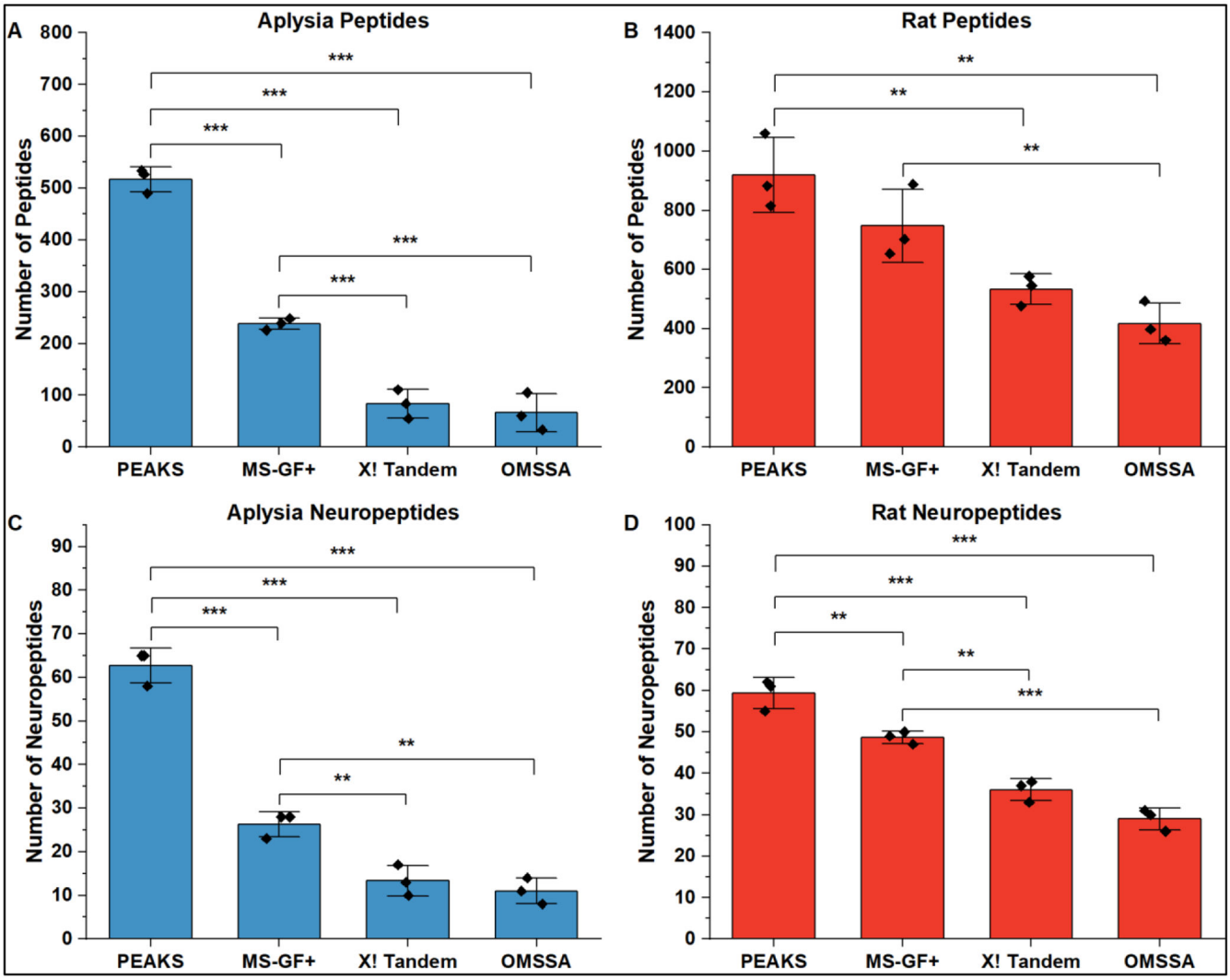


Figure 1:

Bar graphs showing the average number of (**A, B**) peptides and (**C, D**) neuropeptides in the *Aplysia* (left) and rat (right) datasets (n = 3) identified in each of the four search engines. Error bars indicate the standard deviation. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

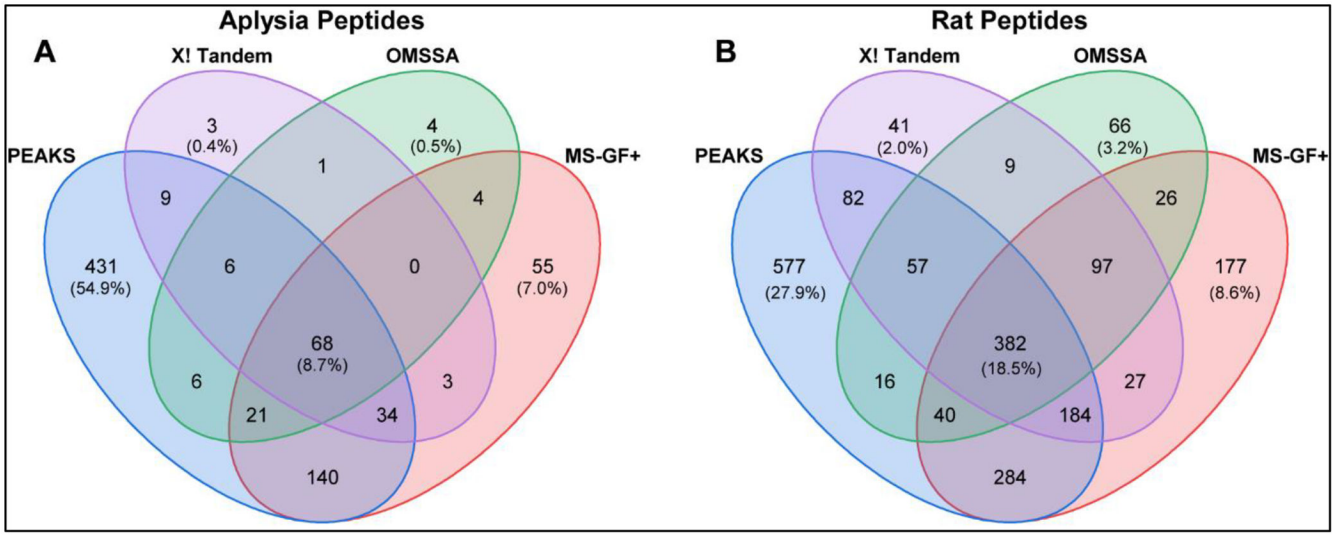


Figure 2:

Venn diagrams showing the number of unique peptides identified by each of the search engines for the *Aplysia* (**A**) and rat (**B**) datasets. The percentages indicate the percent of unique peptides identified in the search engine relative to the total number of identifications across all search engines.

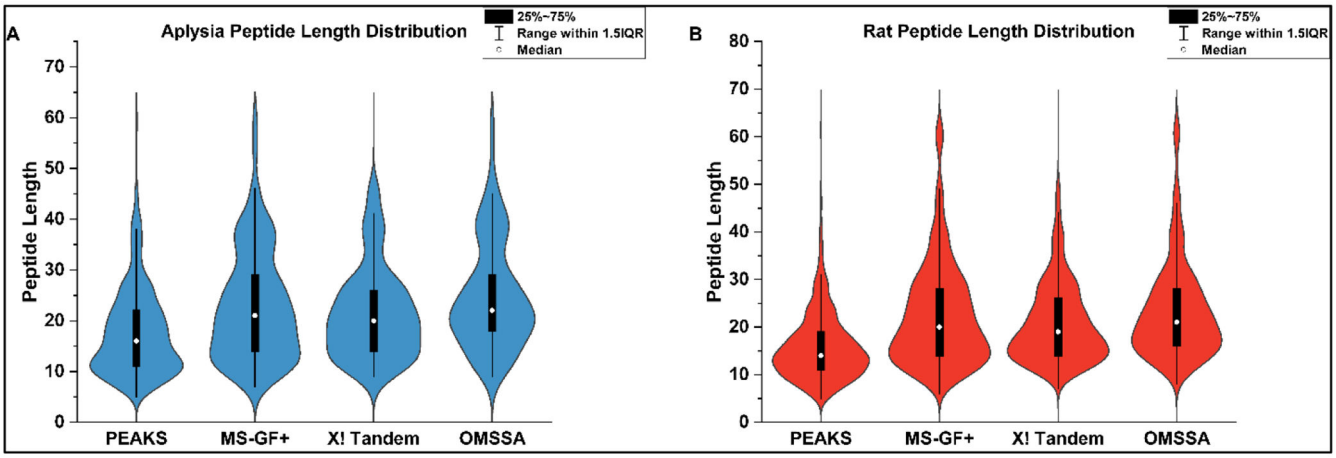


Figure 3:

The violin plots illustrate the peptide length distributions among the search engines for the *Aplysia* (**A**) and rat (**B**) datasets. The white circle indicates the median length, the black boxes represent the interquartile range, and the whiskers indicate the outliers at a factor of 1.5 from the interquartile range.

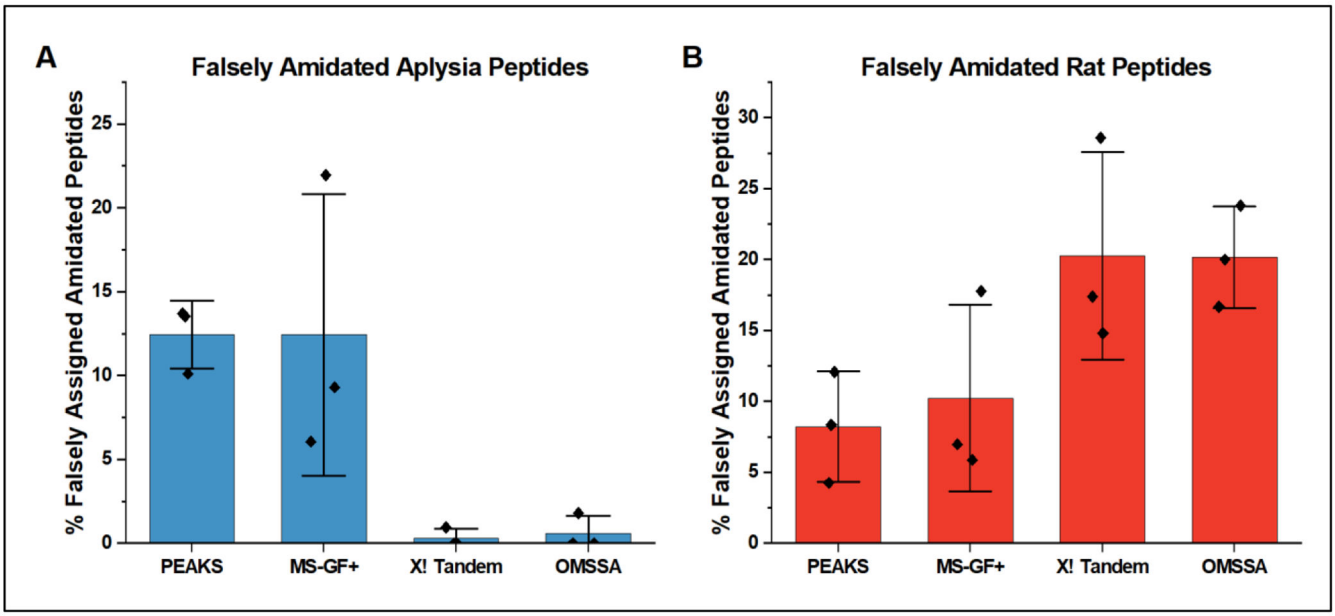


Figure 4:

Bar plots illustrating the average percentage of falsely assigned amidated peptides relative to the total number of amidated peptides in the *Aplysia* datasets (**A**), and the rat datasets (**B**). Error bars indicate the standard deviation (n = 3).

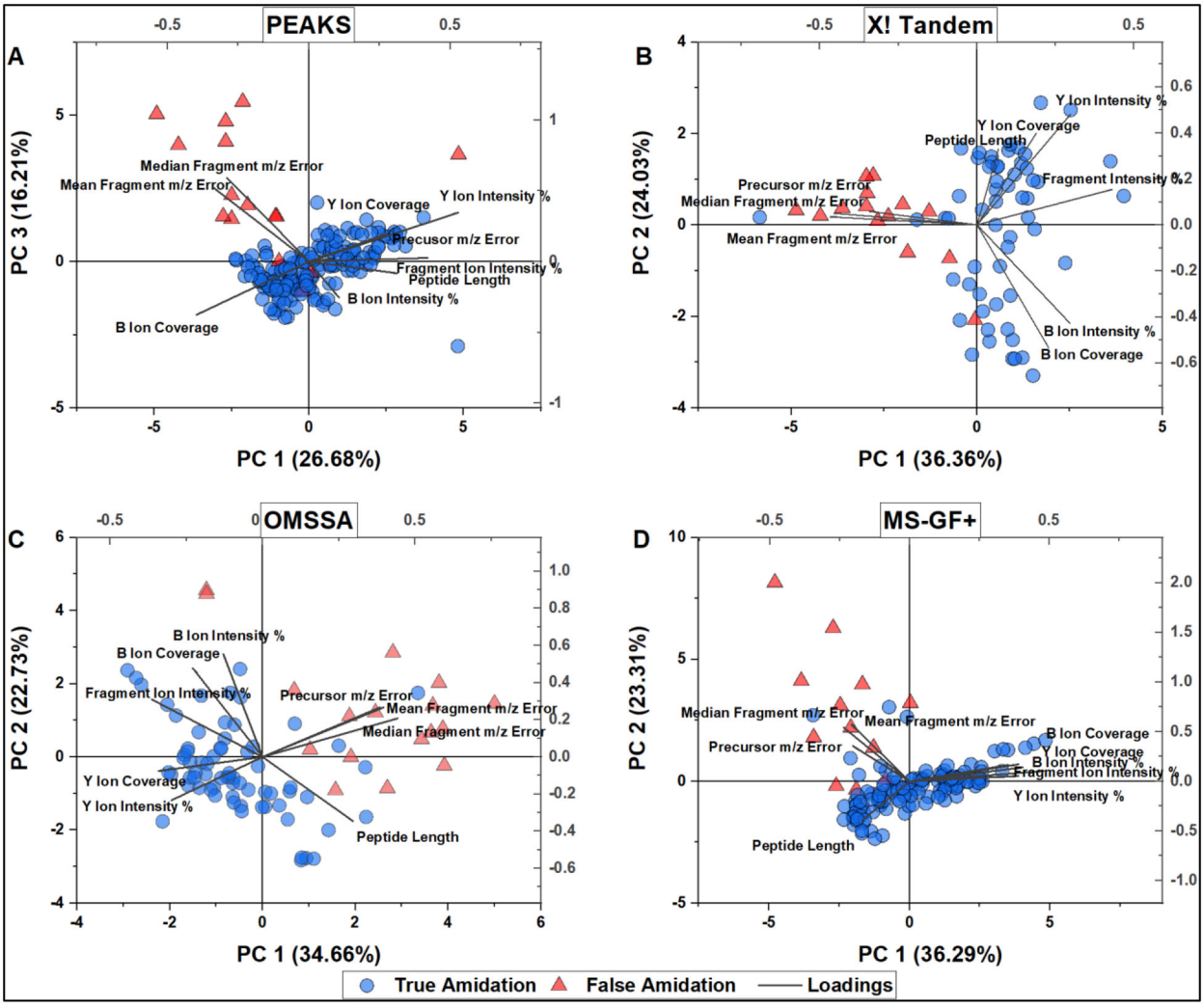


Figure 5:

PCA biplots for the amidated peptides in each of the four search engines (**A-D**). Blue circles represent true amidated peptides, and the red triangles correspond to false amidated peptides. The left and bottom axes correspond to the score plot and the top and right axes correspond to the loadings plot.

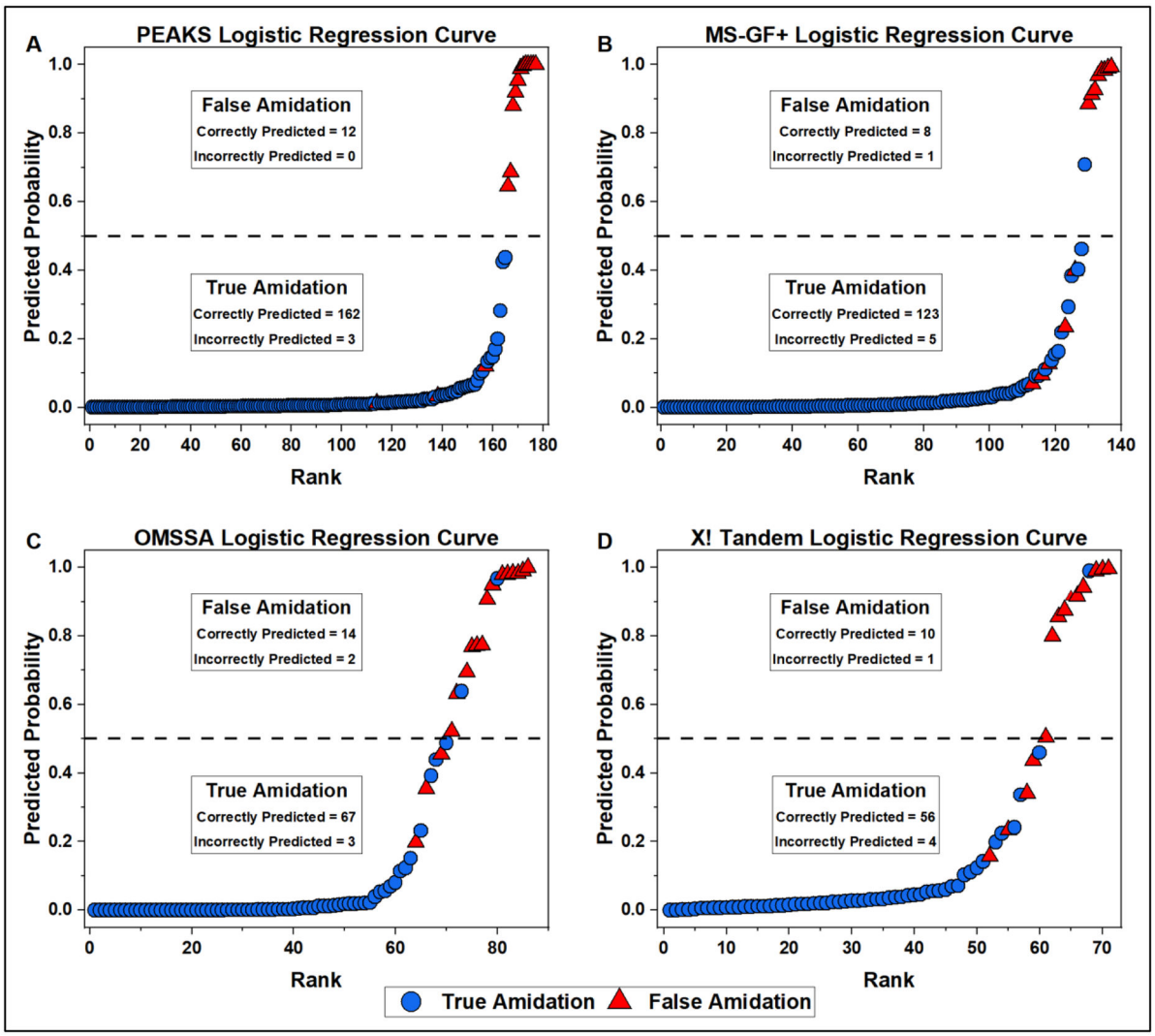


Figure 6:

Logistic regression scatterplots for each search engine. The blue circles indicate true amidated peptides and the red triangles indicate false amidated peptides. The dashed line indicates the predicted probability cutoff at 0.5, whereby false amidated peptides were considered correctly predicted by the logistic regression models as false amidations if their predicted probabilities were above this threshold, and true amidated peptides were considered correctly predicted as true amidations if they had a predicted probability below this threshold.

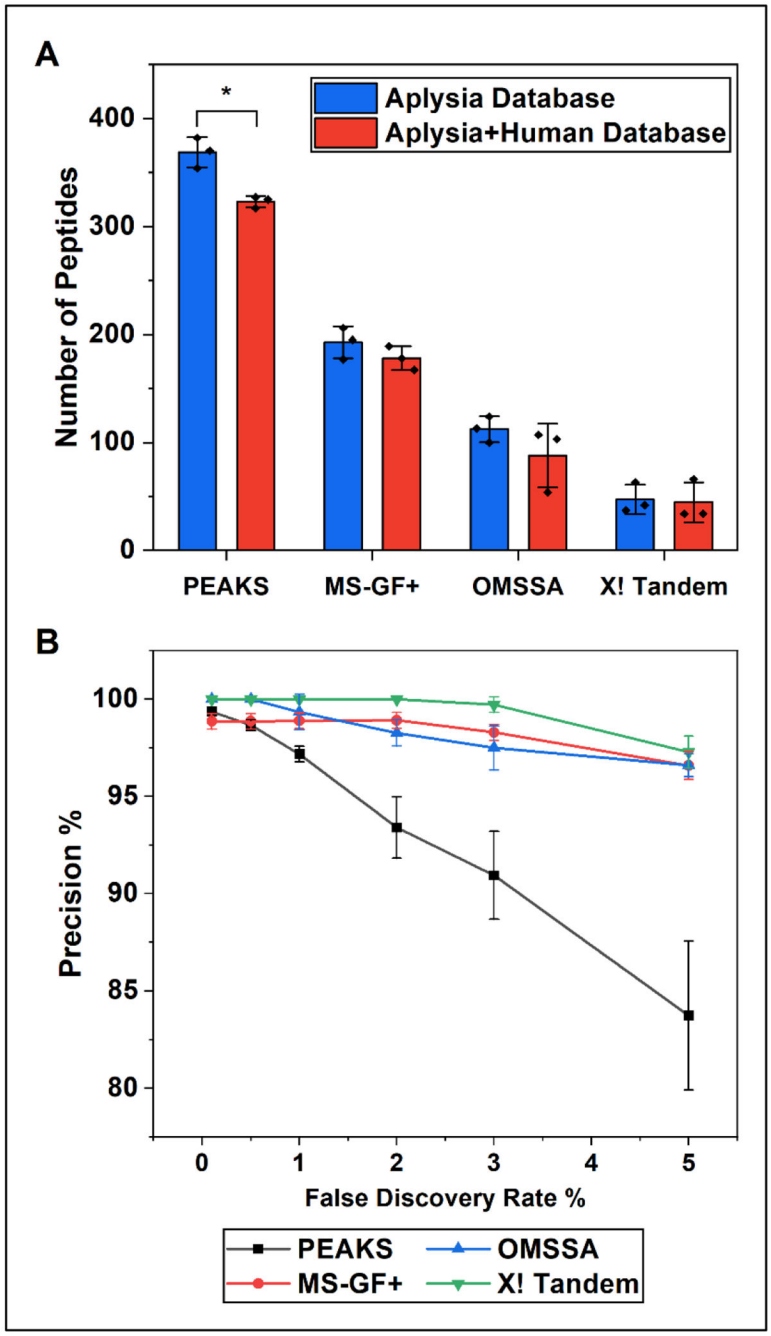


Figure 7:

(**A**) Bar graph showing the average number of Aplysia peptides identified by each search engine when searching against the *Aplysia* or mixed *Aplysia*-human database. (**B**) Plot showing the average precision of the different search engines at various FDR thresholds (n = 3). The error bars indicate the standard deviations and * indicates $p < 0.01$.

Table 1:

List of the different factors that were evaluated for PCA when assessing false assigned amidated peptides in the four search engines.

Spectrum Features	Description
Mean Fragment *m/z* Error	The average absolute fragment *m/z* error for the b and y ions in a peptide (ppm).
Median Fragment *m/z* Error	The median absolute fragment *m/z* error for the b and y ions in a peptide (ppm).
Precursor *m/z* Error	The absolute peptide precursor ion *m/z* error (ppm).
Peptide Length	The number of amino acids in an assigned peptide.
Y Ion Coverage	The percentage of y ions annotated for the peptide sequence in the spectrum[a]
B Ion Coverage	The percentage of b ions annotated for the peptide sequence in the spectrum[a]
B Ion Intensity %	The percentage of the sum of the annotated B ions relative to the sum of all ions in a spectrum[b]
Y Ion Intensity %	The percentage of the sum of the annotated Y ions relative to the sum of all ions in a spectrum[b]

[a]When calculating percentages of b or y ion coverage for a peptide, only the position of the ion was counted, such that if the multiple forms of an ion type position were assigned in the spectrum (e.g., y_2 and y_2^{2+}, and y_2-NH_3), only one instance of that ion position would be used for the percentage calculation.

[b]When calculating the percentages of b and y ion intensity coverage, all the assigned ions were counted separately, such that ions of the same type (b or y) and position with different charge states or neutral losses (e.g., y_2, y_2^{2+}, and y_2-NH3) were counted as three separate ions if they were all present in the spectrum, and were each counted separately for the percentage calculation.