EVALUATION OF DATABASE SEARCH PROGRAMS FOR ACCURATE DETECTION OF NEUROPEPTIDES IN TANDEM MASS SPECTROMETRY EXPERIMENTS

BY
MALIK NADEEM AKHTAR

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Bioinformatics
with a concentration in Animal Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2012

Urbana, Illinois

Thesis Committee:
Professor Sandra Luisa Rodriguez-Zas
Professor Jonathan V. Sweedler
Assistant Professor Alfred Roca

ABSTRACT

Programs to identify proteins in tandem mass spectrometry experiments are not optimized to identify neuropeptides and other peptides resulting from the processing of prohormones. This is due to the unique characteristics of neuropeptides including release after complex processing of prohormones, potentially intense post-translational modifications and their small size. The aims of this study were: (1) to evaluate the strengths and limitations of different tandem mass spectra search algorithms to detect neuropeptides and other peptides resulting from prohormone processing; (2) to evaluate the impact of mass spectrometry factors such as charge on the identification of these peptides; and (3) to offer guidelines to obtain the most comprehensive and accurate survey of the prohormone peptides of a sample. Three software database search programs, OMSSA, X!Tandem and Crux, were applied to identify neuropeptides from *in silico* produced mass spectra. The spectra were simulated from a database of 7850 mouse peptides from 92 prohormones. For each peptide, spectra were simulated with either +1, +2 and +3 precursor charge states, and +1 charged *b* and *y* product ions including single water and/or ammonia loss depending on amino acid composition. The spectra were searched against the mouse database and a rat database including 7647 neuropeptides. OMSSA, X!Tandem and Crux correctly detected 98.9%, 93.9% and 88.7% of the peptides, respectively, at the comparable significance *E*- or *p*-value < 1 x 10-6. Scoring only *b*- or *y*-ion series significantly reduced peptide identification for both OMSSA and X!Tandem. At E-value < 1 x 10-6, 50.8% and 55.3% of peptides were correctly identified by both algorithms using *b*- and *y*-ion series, respectively. Furthermore, availability of only *b*-ion, *y*-ion series and 50% random ions for peptide identification had in general minor influence on the scoring functions of OMSSA and Crux. The comparatively weaker performance of X!Tandem

suggests that the corresponding scoring function favors continuity of ions. The charge state had minor effect on the detection of neuropeptides. Unlike Crux and X!Tandem, OMSSA was negatively influenced by the presence of additional peaks in the spectra at higher precursor charge states. The sensitivity of either program to detect small neuropeptides (< 10 amino acids in length) was limited. This is particularly troublesome given the large number of neuropeptides that are small. Peptide identification by X!Tandem across species suggests that the position of the mismatch in the sequence is critical when using non-specific species databases. These results indicate that alternative algorithmic specifications and implementations must be developed to optimize the detection of neuropeptides.

ACKNOWLEDGMENTS

Most humble gratitude to ALLAH Almighty for blessing me with this opportunity to fulfill my degree requirements. Many thanks to my supervisor Professor Sandra Luisa Rodriguez-Zas and Dr. Bruce Robert Southey for their continuous efforts, support and guidance for me in an encouraging and polite manner. I would like to thank Professor Jonathan V. Sweedler and Professor Alfred Roca for being part of my of thesis committee. I am grateful to my parents, friends (Arshan Nasir, Iftikhar Ahmed, Tahir Shah and Tayyab Nawaz) and lab members (Ahmed Sadeque, Zeeshan Fazal, Nicola Serao and Ken Porter) for their continuous support. Also I would like to thank COMSATS for providing me with a fellowship to continue my studies.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF TABLES

CHAPTER ONE: LITERATURE REVIEW

1.1 NEUROPEPTIDES

Neuropeptides are non-tryptic endogenous peptides that play critical roles in many critical biological processes [1, 2]. Neuropeptides encompass neurotransmitters and peptide hormones and usually range in length from 3 to 40 amino acids [1]. Biosynthesis by neurons, regulated release, ability to function by acting on neural receptors [3], short or long unique primary sequences, non-tryptic nature [4], and less complex 3D structures than the regular proteins due to smaller size [5] are among characteristic features of neuropeptides. Neuropeptides are involved in cell-cell communication as peptide neurotransmitters and regulate many biological processes such as growth, learning, memory, metabolism, and neuronal differentiation by acting as peptide hormones [1]. Biological functions of neuropeptides are largely determined by their unique primary sequences, with the same neuropeptide often acting as neurotransmitter in the nervous system and as peptide hormone in the peripheral endocrine system. One such example is the neuropeptide Enkephalin (PENK), a neurotransmitter in the central nervous system that is also involved in regulating intestinal motility and immune cell functions in the peripheral endocrine system [1].

Neuropeptides are derived from larger full length precursor proteins known as proneuropeptides or prohormones through complex post-translational processing which includes cleavage at basic lysine (K) or arginine (R) amino acids, removal of C-terminal basic amino acids, and modifications such as amidation, acetylation and others [1, 5]. One or more than one copy of the same neuropeptide could be present in a single prohormone, for example Proenkephalin

contains four copies of Metenkephalin [1]. So far approximately 100 prohormone genes have been reported across species [6].

Preprohormone is a collective term used for prohormones carrying a signal peptide chain of hydrophobic residues on their N-terminals [7]. Signal peptide is characteristic of secretory proteins and directs precursor proteins through the ribosome into the lumen of rough endoplasmic reticulum (RER) in an ATP-mediated process. In the RER, biosynthesis of neuropeptides after proteolytic processing of prohormones starts co-translationally. As a first step, signal peptide is removed by the signal peptidase followed by prohormones folding, post-translational modifications or PTMs (e.g., glycosylation, disulfide bridges etc.,) and transfer to Golgi apparatus, where it is packed into newly formed secretory vesicles along with processing proteases [7, 8]. Secretory vesicles serve as primary sites for the formation of mature and biologically active neuropeptides upon maturation [1].

In the secretory vesicles prohormones undergo complex post-translational processing including cleavage of precursor proteins by proteases into shorter peptides. Cleavage can occur either at pairs or multiple basic amino acids or pairs of basic amino acids separated by *n* number of non-basic amino acids. Mostly cleavage takes place at KR or RR and in some cases at R-nX-R, where X represents any amino acid and $n = 0, 2, 4$ or 6 [9]. Additionally, few precursor prohormones are also cleaved at single basic amino acids (K or R) or occasionally at non-basic amino acids [1, 9]. The newly formed peptides undergo further post-translational processing including removal of C-terminal basic amino acids by carboxypeptidases. Bioactive peptides may undergo modifications such as N-terminal acetylation, C-terminal amidation, glycosylation,

phosphorylation and sulfation [10]. Neuropeptides participate in signal transduction pathways through cell surface receptors on their target cells.

Neuropeptides exert their effect through interaction with receptors. Neuropeptides receptors include G-protein coupled receptors or enzyme-linked receptors. In most cases, receptors are G-protein coupled receptors which have seven membrane spanning alpha helices. Recent studies have shown that more than 100 G-protein coupled receptors exist in the human genome, which likely have peptide ligands highlighting the importance of peptide identification present in tissue or cells [11].

1.2 PREDICTION OF CLEAVAGE SITES IN PROHORMONES

Identification and characterization of neuropeptides using experimental procedure has been mostly limited to few model species (e.g., human, mouse, rat, honey bee, zebra finch, tribolium). This is because neuropeptide identification is resource intensive and challenging as cleavage patterns of prohormones vary across species, tissues, pH levels and developmental stages among other conditions [12]. Additionally, conventional sequence homology based annotation approaches may be ineffective due to the diversity in prohormone sequences across species and small sizes of neuropeptides, leading to inaccurate neuropeptide identification [6, 12].

Neuropeptides are formed mostly as a result of proteolytic cleavage of prohormones either at single or pairs of basic amino acids by proteases [1, 9]. The identification of much conserved cleavage patterns can surmount limitations of homology-based annotations by considering multiple

cleavage motifs in prohormones [13]. Various cleavage prediction models have been implemented to accurately identify neuropeptide cleavage sites.

Southey et al. developed "known motif" model for the prediction of prohormone cleavage sites using training prohormone sequences from insects, birds, mammals, fish and other species [13]. The model assigns cleavage probabilities to windows containing one of these motifs: KK, KR, RR, RxxK, or RxxR, where x, R and K refers to any, arginine and lysine amino acids, respectively. The cleavage probabilities were derived from the frequency of occurrence of above motifs at the cleaved sites [13]. Amare et al. trained a binary logistic regression model on mammalian sequences. The model included training precursor sequences from human, pig, mouse, rat and cattle. The mammalian model proved to be more sensitive than known motif, Aplysia and other models [14].

Logistic regression and artificial neural network models based on relative positions and physiochemical properties of amino acids around cleavage sites have been developed using known motifs, mammalian and specie-specific datasets with correct classification rate ranging from 85% to 100% across species and models. This suggests that despite sequence divergence in prohormones, a majority of cleavage patterns remain conserved across species, although higher results of species-specific models relative to known motifs or mammalian models shows that there are species-dependent cleavage patterns in mouse, rat, human and cattle [6].

NeuroPred (http://neuroproteomics.scs.illinois.edu/neuropred.html; [15]) is a cleavage site prediction web application written in the Python programming language. It assigns cleavage

probabilities using known motif, mollusk, mammalian and insect binary logistic regression models. This system supports the selection of one or multiple predictive models and the validation of the predictions including model accuracy statistics. In addition, the masses of predicted neuropeptides including PTMs are also provided [15].

1.3 DATABASES OF PROHORMONES AND NEUROPEPTIDES SEQUENCE INFORMATION

UNIPROT

UniProt (http://www.uniprot.org) is a unified database of protein sequences and functional characterization that encompasses information from Swiss-Prot, Translated EMBL Nucleotide Sequence Data Library (TrEMBL) and Protein Information Resource Protein Sequence Database (PIR-PSD). UniProt consists of four components: (1) the UniProt Knowledgebase (UniProtKB), a database of protein sequence and functional information; (2) the UniProt Archive (UniParc), containing sequences and history of all protein sequences; (3) the UniProt Clusters (UniRef) contains clusters of closely related sequences based on sequence identity; and (4) the UniProt Metagenomic and Environmental Sequences (UniMES) which contains metagenomic data [16].

UniProtKB is a central and curated database of protein information provided by cross-references from more than 120 databases. Each entry in UniProtKB comes from a wide variety of sources including EMBL/GenBank/DDBJ, direct submissions, experimental data from the literature, Protein Data Bank (PDB), Ensembl and RefSeq [17]. Entries in UniProtKB include primary amino acid sequence, protein entry name or description, taxonomic data and citation

information (http://www.uniprot.org/help/uniprotkb). UniProtKB is comprised of two sections: UniProtKB/Swiss-Prot and UniProtKB/TrEMBL. UniProtKB/Swiss-Prot contains manually curated records based on critical reviews of experimentally validated or computationally predicted data of each protein sequence. UniProtKB/Swiss-Prot annotation includes explanation of protein or peptide functions, enzyme specificity, domains, PTMs, sub-cellular location, tissue-specific expressions, structures and associated diseases [17]. UniProtKB/TrEMBL, includes automatic annotation of six-reading frame translations of all coding DNA sequences present in EMBL/GenBank/DDBJ, TAIR, SGD and Ensembl nucleotide sequence databases [16]. Once an entry in UniProtKB/TrEMBL is manually annotated it is moved from this section to UniProt/Swiss-Prot (http://www.uniprot.org/help/uniprotkb). The information in UniProtKB Swiss-Prot and TrEMBL is well-suited to support neuropeptide and prohormone research.

NEUROPEPTIDES

Neuropeptides (http://www.neuropeptides.nl; [3]) is an online database of known neuropeptides. The database contains information about gene symbols, precursor names, chromosomal locations and organization of genes, expression in mouse brain and a list of bioactive neuropeptides. So far more than 70 human neuropeptide genes have been reported and these are grouped into more than 13 families based on structural and functional similarities. Through hyperlinked gene symbols each neuropeptide gene is directly linked to the UCSC (University of California Santa Cruz; http://genome.ucsc.edu) human genome browser containing information about genomic location, transcripts, expression, homolog genes in other species and single nucleotide polymorphisms (SNPs). The hyperlink in the precursor column leads to a pre-computed

BLAST results page with information about homologous proteins in different species and their descriptions. The mouse brain expression data was obtained from the Allen Brain Atlas or GenePaint.org.

PEPTIDEDB

PeptideDB (http://www.peptides.be; [18]) is a composite database of known bioactive peptides, peptide precursor proteins and known protein motifs. Based on UniProt, literature review and sequence alignments, the current version of PeptideDB contains 20027 peptides and 19438 precursor proteins from 2820 species, which are grouped into 373 peptide families. Of these, 178 families have known motifs in Prosite, Pfam or SMART, while the remaining 195 families are classified as novel peptide families. Furthermore, 97% of peptides in PeptideDB are below 200 amino acids in length, while 98% precursor proteins are below 500 amino acids in length. About 14358 (72%) of the peptides belong to the phylum Chordata. The database contains both peptides with known biological activities in vitro or peptides having sequence homology with known peptides in the UniProt database. Also, 2634 precursors do not contain any known bioactive peptides, therefore, it is important to use cleavage prediction program such as NeuroPred to identify peptides in these precursor proteins [18].

1.4 APPROACHES TO STUDY NEUROPEPTIDES

Several approaches are available to identify neuropeptides in samples. These include Edman degradation, radioimmunoassay (RIA), enzyme-linked immunosorbent assay (ELISA), immunocytochemistry and the mass spectrometry (MS). The Edman degradation method is used to

find the sequences of neuropeptides by labeling and the sequential cleavage of N-terminal amino acids. However, the Edman degradation method is time consuming and N-terminal post-translational modifications (for example acetylation) can hinder in step-wise degradation process [19]. Bioassays are very useful for the detection of peptides and to study their possible physiological effects [20]. The RIA approach is used to characterize and quantify peptides in complex mixtures using specific anti-peptide antibodies. The peptide antigens compete with the radioactively labeled peptide tracers for binding with the antibodies. An RIA standard curve shows quantity of peptides in complex mixtures. The ELISA approach provides a useful alternative to RIA to quantify peptides without using radioactively labeled isotopes [21]. Immunocytochemical techniques are similar to RIA in that these techniques detect the presence of neuropeptides using a peptide-antibody affinity approach to localize peptides in a target tissue or cell [7]. For RIA, ELISA and immunocytochemistry, prior information about exact peptide sequence is not required. However, known purified peptide samples are used to obtain specific anti-peptide antibodies. Furthermore, these approaches are less specific when samples contain many neuropeptides that can cross react with specific antibodies. In recent years the MS approach has gained much popularity and has largely replaced other techniques. Like the Edman degradation method it allows neuropeptide sequencing using much smaller amount of samples and also it has better sensitivity in detecting post-translationally modified peptides than other techniques [19]. The MS based approach is the focus of the current study and it will be discussed in detail.

1.5 MASS SPECTROMETRY BASED PROTEOMICS

The aim of MS based proteomics is to identify proteins using the masses of peptides (such as tryptic peptides) obtained from enzymatic processing of the protein [22]. Large-scale proteomics experiments involve analysis of complex protein mixtures (samples containing various proteins of different concentrations). Methods like two-dimensional (2D) gel electrophoresis, enzyme-linked immunosorbent assays (ELISA) and western blotting have low protein detection sensitivity and are difficult to be used in automated high-throughput proteomics experiments [23]. Mass spectrometry based proteomics has become a method of choice for large-scale protein and peptide identifications mainly due to availability of large number of genomic and proteomic sequence databases and improvements in MS ionization techniques such as matrix-assisted laser desorption ionization (MALDI) and electrospray ionization (ESI) [24]. These techniques rely on the fact that proteins and peptides are polar, nonvolatile and thermally unstable molecules that can undergo ionization into the gas phase without much degradation [25].

A mass spectrometer consists of an ion source that ionizes proteins and peptides into the gas phase, a mass analyzer to record measurements of the ionized molecules as mass-to-charge ratios (m/z) and a detector to measure intensity of ions at each m/z value [24]. Mass spectrometers are maintained under high vacuum to assist ions in reaching the detector without interference from air molecules or collisions with gas molecules. Collisions result in low resolution and sensitivity of mass spectrometer by increasing the kinetic energy of the ions, which can fragment and prevent ions from reaching the detector region [26]. MALDI and ESI are the two most commonly used ionizing methods for protein and peptides. Once ionized, the mass analyzer region of the mass

spectrometer separates and records the m/z value of each ion. Four types of mass analyzers that have different sensitivity, resolution, mass accuracy and spectrum generating ability are currently used in MS-based proteomics research. These analyzers include ion trap, time-of-flight (TOF), quadrupole and Fourier transform ion cyclotron (FT-MS). A mass spectrometer can have one or more mass analyzers in tandem separated by collision cells. The later are known as Tandem mass spectrometers [24]. In tandem MS, proteomic mass analysis is carried out either on intact precursor ions (MS scan) or on further fragmentation of precursor ions into N and C-terminal containing ions (MS/MS scan). This further fragmentation is carried out in the collision cell upon reaction of the precursor ions with inert gas molecules in a process known as Collision Induced Dissociation (CID) [25]. The detector region of a mass spectrometer records the intensities of the ions by counting the number of each m/z value. The mass-to-charge ratios (m/z) and corresponding intensity values are finally represented in the form of a MS or MS/MS spectrum of m/z values for further downstream analysis [24]. In MS-based proteomics, proteins are characterized either as enzymatically digested peptides (Bottom-Up approach) or as intact proteins (Top-Down approach). Both approaches correlate MS data with sequence information from proteomic and genomic databases.

TOP-DOWN APPROACH

The Top-Down is a peptide-protein mapping strategy used to identify proteins based on masses of the intact proteins and their fragmented ions. Fragmentation of proteins into peptides is carried out in the gaseous phase of ESI and MALDI mass spectrometers [27]. Recent techniques such as electron capture dissociation (ECD) and electron transfer dissociation (ETD) provides

more useful fragmentation of complete protein sequences and depiction of multiple PTMs. The informative fragmentation patterns are key to identify protein isoforms through correct peptide-protein mapping [25, 27]. With Top-Down approach, it's possible to directly quantify proteins by measuring protein abundances rather than characterizing them through peptide sequences. Several limitations including limited front-end separation of intact proteins as compared to the peptide samples, large sample requirements, less sensitivity and less effective fragmentation of larger proteins in the gaseous phase have restricted the application of the Top-Down approach to the analysis of single proteins and simple protein mixtures. Mostly, MS data of the Top-Down approach is analyzed using Expressed Sequence Tags (EST) and *de novo* methods [25].

BOTTOM-UP APPROACH

The Bottom-Up approach is widely used in MS-based proteomics studies. It is also known as multidimensional LC/MS/MS or multidimensional protein identification technology (MudPIT). In this approach, complex protein mixtures are enzymatically digested into peptides, masses of intact peptides are measured followed by front-end peptide separation (either using chromatography or Gel-based methods) and MS or MS/MS analysis in the gas phase using ESI and MALDI techniques of MS [28]. In the gas phase peptides are further fragmented to produce peptide fragment ion mass ladders in a process termed as Collision Induced Dissociation (CID). The measurements are than recorded either as masses of the intact peptides (MS) or fragmented ions of the peptides (MS/MS) [25]. The parent proteins are identified either using Peptide Mass Fingerprinting (for MS scan) or Shotgun proteomics strategy (MS/MS scan).

The Bottom-Up approach is well suited for large-scale proteomic studies as it allows analysis of complex protein mixtures coupled with good front-end separation techniques, has better sensitivity and requires a lesser amount of the sample as compared to Top-Down approach [25]. The peptides are more solubilized and easy to separate than parent protein molecules. Unlike the Top-Down approach, Bottom-Up is not an ideal choice to identify splice variants and PTMs of proteins due to less sequence coverage of proteins. The lower coverage is due to the fact that not all peptides of enzymatically digested proteins are detectable and only some of them give useful fragment ion ladders [27].

The Bottom-Up MS data is generally analyzed by one of these methods: database searching, spectral library search, *de novo* methods and peptide mass fingerprinting (PMF). Among these, database searching is most commonly used in which experimental MS/MS spectra is correlated with theoretical spectra generated from a set of target database peptides, satisfying a certain search criteria for database search algorithms [25].

SHOTGUN PROTEOMICS

The Bottom-Up Shotgun Proteomics approach is widely used in proteomic studies to identify proteins and their PTMs with better sensitivity. This technique is an analogous method to shotgun DNA sequencing (whole genome sequencing from short sequence reads). Figure 1 depicts the general scheme of shotgun proteomics. First proteins are enzymatically digested and resultant complex mixture of peptides is separated by one or multidimensional chromatography. After separation, proteins are subjected to further fragmentation in collision cell of tandem mass

spectrometers to produce MS/MS spectra [29]. The identification of proteins starts with the correct assignment of peptides sequences to the MS/MS spectra. Several computational methods that have complementary features have been developed for this purpose. These computational methods are classified into three major groups: (1) the database search approach; (2) the *de novo* approach; and (3) the hybrid approach. The database search is the most efficient method in which each experimental MS/MS spectrum is assigned a peptide sequence from sequence databases by correlating experimental spectrum with theoretical spectra of peptides in the database using database search algorithms [23, 30].

The assignment of peptides to the spectra is assessed either manually or based on statistical measures such as *E*-value or *p*-value. Unlikely or incorrect assignments are removed. Peptide assignments that have strong supporting evidence are used to identify the original proteins by peptide-protein mapping. For this purpose peptides are grouped on the basis of their parent proteins and statistical confidence scores are assigned to peptide-protein mappings [23].

1.6 PEPTIDOMICS

Peptidomics is the study of endogenous peptides (molecular mass < 20 KDa) present in a cell, tissue or organism. Earlier peptide identification studies relied on the Edman degradation method, and currently MS-based peptidomics is the method of choice. With the advent of MS-based peptidomics it became possible to identify large number of peptides and their PTMs in a single experiment [10]. MS-based peptidomics is a two step process: first separation of complex mixtures of peptides by multidimensional liquid chromatography or gel or liquid-based isoelectric focusing; and second MS-based peptide identification [5].

Peptidomics is an analogous term for proteomics but has some obvious differences. First, unlike proteomics which aims to identify proteins, the peptidomics refers to peptide identification from sequence information [22]. Second, in proteomics identification of only few peptides (in most cases not all peptides of precursor are required) can lead to identification of precursor proteins, while in peptidomics the primary aim is to identify as many peptides as possible from sample considering each peptide as a bioactive molecule performing specific functions. Third, proteomics uses enzymatically digested peptides to characterize parent protein, while peptidomics does not involve enzymatic digestion in order to identify original peptides along with their PTMs [10].

1.7 IDENTIFICATION OF PEPTIDES VIA TANDEM MASS SPECTROMETRY

Peptide identification using MS/MS spectra is a highly efficient and sensitive method that is well-suited for large scale peptidomics studies [23]. The peptide identification process starts after digestion of proteins with enzymes (such as trypsin) resulting in a complex mixture of peptides, which are separated by one dimension (e.g. High Performance Liquid Chromatography) or multidimensional (MudPIT) chromatography columns. The peptides undergo ionization at the end of a column with MALDI or ESI ionization techniques. Once ionized, peptides travel through mass spectrometer vacuum and m/z and intensity of peptide ions (precursor ions) are recorded as MS spectrum by mass analyzers. The selected peptide ions are transferred to collision chambers of tandem mass spectrometers (spectrometer with more than one mass analyzer separated by collision chambers), where the peptide ions react with inert gas (such as nitrogen, argon and helium) and peptide bond is broken to generate fragment ions. This fragmentation process is called collision induced dissociation (CID). The breakage of the peptide bond results in N-terminal containing ions

and C-terminal containing ions. The mass spectrometers then generate MS/MS spectra using m/z ratio and intensity pairs of fragmented ions. The fragmentation pattern of MS/MS spectrum is used to identify peptide sequence [31].

Tandem mass spectrometers differ in the CID conditions and this in turn affects the fragmentation patterns of the peptides. Mass spectrometers such as ion trap, triple quadrupole and quadrupole time-of-flight (TOF) have low energy CID conditions generating mostly *b*- and *y*-ions. However, high energy CID conditions in mass spectrometers such as TOF-TOF can also result in *a*-, *c*-, *x*- , *z*-ions. Moreover, other ions can result from side chain cleavages in addition to primarily occurring *b*- and *y*-ions. Additionally, fragmentation patterns in ion trap gives both *b*- and *y*-ions, while in quadrupole/quadrupole-TOF *y*-ions dominate the resulting spectrum [31].

The fragmentation produces different types of ions depending upon type of the broken bonds between two adjacent amino acids in a peptide sequence. However, in MS product ions should carry charge state one or higher in order to be detected. A product ion is categorized as N-terminal ion or C-terminal ion if charge is retained on the N-terminal or C-terminal, respectively [32]. Most common ions are *b*- and *y*-series ions, generated as a result of the breakage between carbonyl carbon and the amide nitrogen. A breakage of the bond connecting alpha carbon and the carbonyl carbon gives *a*- and *x*-ion series. The third breakable bond is amide nitrogen and alpha carbon bond, which yields *c*- and *z*-ion series. Additional peaks due to neutral mass loss such as loss of ammonia (nh_3, -17 Da) or water molecule (h_2o, -18 Da) from either *b*- or *y*-ion series are also common in MS/MS spectrum [33]. The *a*-, *b*- and *c*-ions are classified as N-terminal ions,

while *x*-, *y*- and *z*-ions are classified as C-terminal ions [32]. Molecular masses of different types of ions are listed in Table 1.

The matching of MS/MS spectra to peptides is complicated by various factors. First, an ideal MS/MS spectrum should contain only *b*- and *y*-ion series , no noise/rare ion peaks, and all *b*- and *y*-ions (complete fragmentation of peptides) with same intensity values [34]. However, a real MS/MS spectrum contains incomplete fragmentation patterns, noise peaks, other types of ions and variable intensity values depending upon types of ions [34, 35]. Under low energy CID conditions peptides undergo incomplete fragmentation (unbroken bonds between amino acids) in instruments such as triple quadrupole, hybrid quadrupole-TOF and ion traps, resulting in tandem MS/MS spectra with many ions missing from common *b*- and *y*-ion series (discontinuous series). High energy CID conditions (such as in hybrid sector/TOF) results in complete fragmentation of the peptides but are expensive to use, generate many rare ion peaks, and must be used by highly trained professionals [35]. Second, discrepancies between the m/z of precursor and fragment ions depend on the instrument, temperature, and instrument parameter settings [29, 36]. Generally, the instrumental mass errors range from few parts per million (high mass accuracy instruments) to 500 parts per million (low mass accuracy instruments) [29]. In mass spectrometers, mass calculations are based either on monoisotopic masses of amino acids (with ^{12}C atoms only) or average mass of all ions (including ^{13}C atoms). Third, the determination of the ion charge state (especially in multiple charged ions) is challenging. The assessment of the charge state partially depends on the accuracy of the mass measurements and could be determined from isotopic patterns observed in the MS/MS spectrum. Fourth, the presence of isobaric amino acids, amino acids with same masses for example leucine (L) and isoleucine (I). Low accuracy instruments cannot resolve masses

difference between such amino acids [23]. Fifth, the presence of static and variable PTMs challenges the identification of peptides from ion fragments.

POST-TRANSLATIONAL MODIFICATIONS

Post-translational modifications are covalent changes in proteins due to the proteolytic cleavage or addition of modifying groups [37]. Most proteins are modified and so far approximately 200 different types of modifications have been reported for these proteins [38]. Common PTMs for neuropeptides are glycosylation, amidation, acetylation, phosphorylation and sulfation. These modifications occur in secretory granules and are species- or tissue-specific [7]. Each modification causes mass shifts in precursor or fragment ion m/z values containing that modification thus further complicating interpretation of MS/MS spectra. However with MS, it's possible to detect PTMs on proteins and peptides with better sensitivity by considering the possible mass shifts due to modifications [37].

1.8 DATABASES WITH MASS SPECTRAL DATA ON NEUROPEPTIDES

SWEPEP

SwePep (http://www.swepep.org; [39]) is a database of annotated endogenous peptides intended to facilitate peptide identification from samples using mass spectrometry. As of April 2012, SwePep contains 4180 endogenous peptides from 394 species. The neuropeptides in SwePep have been derived from 1643 precursor proteins with more than 100 neuropeptides having been confirmed experimentally [39]. Additionally, literature supported information about peptides,

precursor proteins, PI, and PTMs are also provided. In this database, the peptides are grouped into three classes: (1) biologically active peptides containing characteristics of neuropeptides and hormones such as convertase specific processing sites, PTMs, and well-known biological functions; (2) potentially biologically active peptides that share characteristics of biologically active peptides but become proteolytically inactivated after sampling; and (3) uncharacterized peptides that are identified but do not share characteristics of above two classes [39, 40].

In the peptide identification process using SwePep, m/z values in experimental spectra are matched against m/z values calculated from database peptides both with and without PTMs and scores are assigned to matches. This search procedure is less time consuming than spectral search against whole proteomes of species, as searches against whole proteomes returns more false positives due to large search space [39, 40]. The availability of MS/MS data on 219 unique peptides in SwePep database enables the verification of low quality spectra with low signal to noise ratio by correlating this spectra to the high quality spectra in SwePep. Spectrum validation is required when a peptide was identified with both high and low confidence scores in MS/MS across different experiments. In the SwePep MS/MS database, spectrum-spectrum identification starts by assigning a Pearson correlation coefficient as a measure of similarity between two the spectra that results from comparing their fragmentation patterns (intensities of *b*- and *y*-ions) [40].

NEUROPEDIA

NeuroPedia (http://proteomics.uscd.edu/Software/NeuroPedia.html; [4]) is a sequence database and MS/MS library of neuropeptides. NeuroPedia contains sequences, taxonomic and genomic information of 847 neuropeptides from human, rat, mouse, bovine, rhesus macaque, chimpanzee, California sea hare and leech. All possible pairwise alignments of 847 neuropeptides resulted in 340725 pairs of sequences, which are further grouped into 531 identical sequences, 5020 overlapping sequences and 9185 homolog sequences. Using MS/MS database search algorithms, new MS/MS data can be searched against the NeuroPedia sequence database. The NeuroPedia spectral library contains 3401 spectra from human, mouse, rat, bovine and leech. The spectral data is arranged in ten Mascot generalized format (MGF) files depending upon type of species, instrument and enzyme. Searching MS data against the NeuroPedia MS is less time consuming and gives more accurate results for small sized or non-tryptic neuropeptides as compared to searching against whole UniProt database [4] because NeuroPedia is a smaller and targeted database.

NIST

National Institute of Standards and Technology (NIST; http://peptide.nist.gov) is a public MS/MS library of tryptic peptides generated in LC-MS/MS experiments using ESI. The library mostly contains spectra from high energy ion trap mass spectrometers and some spectral data from low energy quadrupole-TOF mass spectrometers. Spectra from ten different organisms in NIST are classified into three types: (1) the consensus spectra derived from multiple identifications of the

same peptide ion; (2) the best replicate spectra; and (3) the high confidence single spectrum identifications. Using four database search engines, peptides were assigned to each spectrum. The aim of the spectral library is to collect MS/MS data and assign peptides to new spectra by spectrum-spectrum matching. This resource can be used for spectrum-spectrum matching that can be better than spectrum-whole proteome sequence matching, by reducing the time involved and false positive identifications.

1.9 PEPTIDE IDENTIFICATION APPROACHES FROM TANDEM MASS SPECTROMETRY

Many computational approaches have been developed to assign peptide sequences to MS/MS spectra. These approaches are classified into four major categories. First is the sequence database approach, in which the experimental spectrum is scored against theoretically simulated spectra of peptides present in proteomic and genomic sequence databases. In most cases only the top hit (best hit) based on assigned score is considered for further analysis. This approach is useful for only those species that have fully sequenced genomic or proteomic information [23, 29, 30]. Second, in the spectral library approach, the experimental MS/MS spectrum is searched against libraries of previously annotated MS/MS spectra. The spectral library approach is based on the idea that the same MS/MS spectra are repeatedly found in many experimental proteomic studies. Peptides with no prior representation in spectral libraries cannot be identified by this approach [41]. Third, in the *de novo* approach, peptide sequences are directly extracted from the MS/MS spectra without database searching. The basic concept behind this approach is that two adjacent peaks in a spectrum differ by a single amino acid. Novel peptides can be identified by this

approach but at the same time the error rate is high due to incomplete fragmentation patterns in MS/MS spectra [34, 35]. Forth, the Hybrid approach is a combination of the sequence database search and *de novo* method. In this approach, short sequence tags (3-5 amino acids) are extracted from the MS/MS spectra followed by an error-tolerant (allowing mismatches) search against sequence databases. The Hybrid approach is useful in detecting sequence polymorphisms and PTMs without significantly increasing database size [23, 29, 30]. The sequence database search approach has been successful in the detection of neuropeptides.

1.10 DATABASE SEARCH APPROACHES

Database searching remains a suitable, robust, and reliable method to interpret MS/MS spectra. Unlike the *de novo* method which requires high quality spectra (good fragmentation patterns and signal to noise ratio), the database search approach can also identify peptides for low quality spectra by counting shared peaks between the experimental and theoretical spectra of peptides. However, identification of novel peptides and peptides with sequence polymorphisms by performing a search against database of known peptides is challenging. The database search against translated genomic databases provides a useful alternative to identify novel peptides which are not represented in a protein databases (e.g., novel alternative splice variants and single nucleotide polymorphisms) [23, 29, 30].

Free commonly used open source proteomic database search methods include OMSSA [42], X!Tandem [43] and Crux [44] (an alternative implementation of the SEQUEST algorithm [45]), MyriMatch [46], and Tide [47]. These methods share the basic working principle shown in Figure 2. Each experimental MS/MS spectrum is scored against a theoretical spectrum (following

common peptide fragmentation rules) generated from a database comprised of enzymatically digested protein sequences. Each experimental spectrum is matched against only a subset of database peptides satisfying certain search specific parameters. The parameters which reduce the search space are: choice of an enzyme, precursor mass tolerance and PTMs. Once the candidate peptides are identified, fragment ions in experimental spectra are compared against fragment ions in theoretical spectra within a certain fragment ion tolerance, and a correlation score is computed. The score reflects the degree of similarity between experimental and theoretical spectra [48, 49].

The MS/MS database search programs mainly differ in the heuristic search algorithm used and in the calculation of a score for the match between the experimentally observed spectra and theoretical spectra. The program either reports an *E*-value or *p*-value by assuming a distribution (e.g. Poisson) of matches for a given experimental spectrum. The *E*-value denotes the number of matches with score equal or higher than the observed one that could arise by chance. Similarly, the *p*-value denotes the probability of finding a match between experimental and theoretical spectra with score than the one observed by chance under a null hypothesis. The output from these programs is a list of peptides ranked by the score for the match. In most cases, only the best match is considered for further analysis, while remaining hits are considered incorrect identifications [30]. A brief overview about these open source algorithms, their scoring schemes and conversion of scores to *E*-value or *p*-values is given below.

OMSSA

The Open Mass Spectrometry Search algorithm (OMSSA) is an open source database search algorithm implemented in C++ and developed by the National Center for Biotechnology Information (NCBI; [42]). This algorithm is available as part of the NCBI C++ toolkit, which can be downloaded and installed across different platforms and can also be used online [42]. An OMSSAAdapter is also available for OMSSA in the OpenMS project [50]. The OpenMS framework includes many tools commonly used in MS for peptide and protein identification. However, OMSSA software is not included in OpenMS toolkit, rather a user has to provide location of OMSSA binary executables (separately installed) through OMSSAAdapter to analyze MS data. The OMSSA algorithm takes an experimental spectrum as input, determines charge, filters nose, calculates mass ladders and compares the processed spectra against theoretical spectra generated from the sequences present in enzymatically digested protein database. To improve algorithmic speed, all masses are converted to integer values by using 100 as a scaling factor, spectra is sorted and indexed using precursor mass and a binary formatted sequence database is used [42].

OMSSA determines the precursor charge states by counting the number of peaks below the precursor m/z value in the spectra. A spectrum with more than 95% of its peaks below precursor m/z is considered in +1 charge state, while a spectrum is searched both as +2 and +3 when less than 95% of spectral peaks are below precursor m/z. The experimental spectra is preprocessed to remove noise peaks in multiple steps. First, the peaks with intensity below 2.5% of the maximum intensity value are removed. Second, only the most intense peak in a region of ±27 Da (around the

examined peak in a non-overlapping window) for the charge states +1 and +2 is retained, while for the charge state +3 a region of ±14 Da is used. OMSSA assumes +1 product ions for spectra with +1 and +2 precursor charge states. For precursor charge states +3, product ions are assumed to be in a +1 charge state and +1 and +2 charge state if they are above m/2 and below m/2, respectively, where m is precursor mass. Third, ion peaks that are 17 Da (ammonia loss) or 18 Da (water loss) or 1 Da (isotopic peaks) below the selected peaks are removed. After these noise removal steps, the theoretical spectra in the database peptides are compared against the processed experimental spectra for precursor charge states +1, +2, and +3 [42].

The scoring function of OMSSA is based on the number of matched ions or shared peak counts (SPC) between the experimental and the theoretical spectra. The SPC is calculated by correlating the experimental and theoretical spectra through counting common peaks that are within a certain fragment mass tolerance range of each other. To improve the sensitivity of the algorithm only those theoretical spectra that have at least one peak matched with any of the top three peaks (or other number selected by the user) in processed experimental spectra are considered for further scoring. The scoring is based on the assumption that distribution of the number of matches follows a Poisson distribution. The Poisson mean is a function of the fragment ion tolerance, number of peaks in the experimental and theoretical spectra, and the precursor mass. The calculation of the mean for spectra containing only a +1 fragment ion is different than for spectra containing both +1 and +2 fragment ions. The OMSSA noise filter does not remove all noise peaks during preprocess steps, so while considering the total number of peaks in experimental spectrum, the algorithm includes the effect of random noise in the calculation of the Poisson mean. The

probability of a match for a given number of matches in the Poisson distribution is calculated using the mean (μ) and number of matches (x) by [42]:

$$P(x,\mu) = \frac{\mu^x}{x!}e^{-\mu}$$

OMSSA reports an *E*-value which is a measure of expected number of peptides having a score equal or better than this score by chance matching. The *E*-value is a function of the number of candidate theoretical spectra (*N*) and the probability (based on number of matching ions following a Poisson distribution) that the search of a given spectra against *N* theoretical spectra is a random event.

$$E(y,\mu) = \mathrm{N}(1 - (\sum_{x=0}^{y-1} P'(x,\mu_Z))^N)$$

where *N*, *y* and *z* are the number of theoretical candidate spectra, the number of matched ions and the searched ions series, respectively [42].

SEQUEST/CRUX/TIDE

The Crux algorithm is a free open source reimplementation of the Sequest algorithm, a first commercial database search algorithm for identification of peptide sequences from a tandem mass spectra [44]. Like Crux, Tide is an alternative implementation of Sequest. Both Tide and Crux provide an alternative method to identify peptides with improved speed by using a database

indexing approach. Crux was written in the C programming language and can be used across platforms. Crux can be executed from the command line by providing a user changeable parameter file containing a description of all parameters. The default parameter specification is used when no parameter file is provided.

The general working algorithm of Crux is shown in Figure 3. The scoring functions of Crux are based on SPC. Crux computes a series of scores and ranks for a given experimental spectra and precursor m/z ratio that starts with the identification of all candidate peptides having precursor m/z ratio within a precursor mass tolerance of the experimental spectra. Candidate peptides are ranked by a preliminary score called Sequest score (SP) and the top 500 peptides are scored and reranked using cross-correlation scores (Xcorr) [44].

Prior to calculating scores, the experimental spectra is preprocessed by taking the square root of each intensity value, normalization of square rooted intensity values and rounding each m/z value to its nearest integer. The SP score is a function of the sums of the intensities of matched ions adjusted by the sequential series ions, the matched immonium ions and the total number of matched ions relative to the total number of ions. After scoring, the candidate peptides are ranked by their SP score and the top 500 are selected for further scoring [44].

The cross-correlation score (Xcorr) is a primary score for determining other scores and serves as a measure of the similarity between two spectra based on shared peak count. This score is computed by matching the processed spectrum (X) to the theoretical spectrum (Y), shifted with

respect to spectrum *X* along m/z axis by *t* mass units. The cross-correlation between spectra *X* and *Y* is:

$$\mathrm{Corr(t)} = \sum_i x_i y_{i+t}$$

where x_i and y_{i+t} are the peaks in the *X* and *Y* spectra, respectively [23]. For each experimental spectrum, the highest value of Xcorr with the database peptides represents the best match. Crux also computes a relative score (delta Cn) and a relative ranking of peptide-spectrum matches for a given input spectra. The relative score is calculated from the difference between the Xcorr scores of a best match and all other peptide-spectra matches of an experimental spectrum. Like Xcorr, higher value of delta Cn corresponds to a correct match, while low score represents likely incorrect peptide identifications. Usually in proteomics studies both Xcorr and delta Cn are considered while characterizing a peptide-spectrum match [23].

Absolute ranking or the ranking all experimental spectra-peptide matches with respect to each other is more challenging than relative ranking. To address this problem, Crux computes a *p*-value from a Weibull distribution fitted using cross-correlation scores from all the theoretical matches of an experimental spectrum [44]. The *p*-value reported by Crux is a Bonferroni adjusted *p*-value that takes into consideration the total number of candidate peptides. Furthermore, Crux includes a machine learning method termed Percolator that uses both target peptide-spectra matches and decoy peptide-spectra matches to assign absolute ranking to identifications. A q-value which is an estimation of false discovery rate can also be computed. The *p*-value based search is

useful because this approach does not require separate decoy files (as in case of Percolator), and this saves computational resources, thus speeding up the search [44].

X!TANDEM

X!Tandem is free open source algorithm distributed by the Global Proteome Machine Organization [43, 51]. This algorithm is implemented in the C++ programming language thus augmenting the portability across platforms. X!Tandem can be executed from the command line when provided with .xml file containing location of the input spectra, parameter file and the protein or peptide database [43]. X!Tandem can also be executed through OpenMS X!TandemAdapter. X!TandemAdapter provides comparable parameter settings with other Adapters in OpenMS (such as OMSSA and Mascot) [50].

The scoring function in X!Tandem also uses the principle of SPC between the experimental and theoretical spectra. For each input spectra X!Tandem assumes that at least one peptide exists in the sequence database. The primary score of X!Tandem is known as the hyperscore and this differs from the other algorithms using similar SPC methods. X!Tandem uses dot-product to score the matches between the observed and predicted spectra [49]. Dot-product scores are converted to hyperscores by multiplying the score by the factorial number of matching *b*- and *y*-ions. By default, the algorithm uses the factorial of the matching *b*- and *y*-ions yet users can also specify other ions such as *a*-, *c*- and *z*-ions to be considered in scoring. The formula for the hyperscore is [23]:

$$S_{hyperscore} = (n_b!\, n_b!) \sum_i x_i\, y_i$$

where x_i and y_i are matched peaks between the experimental and theoretical spectra, respectively and n_b and n_y are number of assigned *b*- and *y*-ions, respectively [23]. The upper 50 percentile of the hyperscores are natural log-transformed and any scores higher than the intersection between the log transformation of the number of results (natural log of *E*-value) and the hyperscore with zero are assumed to be significant. Extrapolation of the linear regression of the natural log of the *E*-value (y-axis) relative to the hyperscore (x-axis) is used to assign *E*-values to high-scoring peptide matches. For example, if a significant hyperscore of 80 correspond to the log *E*-value of -8.3, then the *E*-value for that match is $e^{-8.3}$. The distribution of the scores is assumed to follow a hypergeometric distribution with parameter estimates obtained of the scores for random or false identifications is obtained from the database search. This distribution is used to translate the score of each match into an *Expected* or *E*-value.

MYRIMATCH

MyriMatch is a multi-threaded free open source tandem mass spectral matching algorithm implemented in C++ programming language available under Mozilla Public License. This method is based on an idea of considering user-defined intensity classes in calculating the score, because a typical MS/MS contains peaks of different intensities with few peaks having high intensity and large number of peaks in spectra having low intensity. The first step involves the preprocessing of the experimental spectra to remove noise peaks. This processing is done by computing total ion

current (TIC), total intensity of all peaks in the MS/MS spectra. The TIC that is below a user-specified TIC threshold (representing low intensity peaks) is removed and final processed spectra contain peaks with the highest intensity values. However, a spectrum is not consider for further analysis if a low number of MS/MS peaks prevents the division of the spectrum into intensity classes [46].

MyriMatch generates +1 product b- and y-ions for +1 and +2 precursor charge states to represent the theoretical spectra corresponding to the database peptides. Product ions that have +1 and +2 charge state are generated for precursors that have +3 charge state, depending on the weight of the ions [46]. Each amino acid is assigned a weight, which is the sum of weights of individual amino acids present in that ion. With each broken peptide bond, ions with smaller weight are assigned a +1 charge while ions with larger weight are assigned a +2 charge.

Each spectrum is assigned a *p*-value (probability of randomly matching) based on a multivariate hypergeometric distribution of the matches. The *p*-value calculation considers the number of peaks in a particular class, the number of peaks matched to theoretical fragment peaks from particular intensity class, the total number of *b*- and *y*-ions present in the processed experimental spectrum, and the total number of peaks predicted from the database sequence [46].

1.11 LIMITATIONS OF DATABASE SEARCH ALGORITHMS TO ASSIGN CORRECT PEPTIDE TO SPECTRA

The correct identification of peptides from MS/MS spectra depends on quality of MS/MS data, type of instrument and choice of database search engine. Many different database search

engines are available for peptide identification. The best peptide returned by these algorithms against a spectra is often considered correct, however, best match may not be correct in all cases [52]. There are several reasons for the incorrect assignments by database search engines: (1) low quality of MS/MS spectra due to the presence of many noise peaks, missing ion peaks due to incomplete fragmentation and contaminations in samples; (2) simultaneous fragmentation of more than one peptide ions having similar m/z values in MS. The database search methods often fail to correctly identify all peptide ions present in a given spectrum; (3) scoring schemes of most database search engines are based on simplified representation of the peptide fragmentation process with assumption that all ions are present in MS/MS spectra with same intensity values. Using generalized fragmentation rules to generate theoretical spectra and subsequently using it to score experimental spectra (with incomplete fragmentation and different intensity values) often leads to incorrect peptide assignment for the MS/MS spectra; (4) presence of several true homologous peptides in target sequence database leads to incorrect interpretation of biological data; (5) the database search algorithms are not suitable for the identifications of novel peptides or peptide variants, which have no prior representation in the target sequence databases [23]; and (6) characterization of non-enzymatic post-translational modifications such as isomerization, deamidation and racemization using current database search engines is challenging. The new fragmentation methods (ECD or ETD) provides useful fragmentation patterns to determine such modifications [53]. However, current database algorithms are mostly designed for CID-based fragmentation method and can lead to incorrect peptide identifications.

1.12 COMPARISON OF TANDEM MS SEARCH ALGORITHMS

Database search approach is a robust choice to identify peptides from the MS/MS because of improvements in the experimental procedures and the availability of multiple database search engines and sequence databases. However, the identification scores or statistical values reported by these algorithms depend on the quality of the spectra (signal to noise ratio, intensity, etc.,), the dissociation mechanism (such as CID, ECD and ETD), the search parameters specified (precursor and fragment mass tolerance, enzyme, etc.,), the sequence databases, and the database sizes. The selection of suitable database search algorithm for peptide identification should consider the strengths, weaknesses, and the working principle of different algorithms to correctly assess the significance of the match [48]. Several attempts have been made to benchmark these algorithms by evaluating the performance of these algorithms on common datasets using comparable search parameters.

Balgley et al. [48] obtained 155973 MS/MS from a complex mixture of 18 proteins using an ion trap mass spectrometer and ESI as ionization source. Four commonly used search algorithms (OMSSA, X!Tandem, Mascot and Sequest) were compared and finally an False Discovery Rate (FDR) adjustment of the Sequest thresholds was proposed. Comparable search parameters and the same target-decoy sequence database search strategy was used to evaluate the variations in the scoring functions of these algorithms in an unbiased manner. However, the results and criteria may not be applicable across different mass spectrometers, dissociation mechanisms (such as CID, ECD and ETD) and search parameters. Common measures of comparison MS/MS such as sensitivity, specificity, and number of proteins identified at a 1% FDR adjusted threshold

in a shotgun proteomics approach. OMSSA outperformed all other algorithms at 1% FDR returning 35% more hits than X!Tandem (second best in comparison). Overall the OMSSA, X!Tandem, Sequest and Mascot correctly identified 48328, 31367, 29463 and 24575 peptides, respectively. OMSSA outperformed X!Tandem in terms of sensitivity (97.9% vs. 74.2%, respectively) and MS/MS hits per protein (17.7 vs. 15.0, respectively) and was comparable in terms of specificity (98.7 vs. 98.8, respectively) [48].

Kapp et al. [54] compared five database search algorithms commonly used by research laboratories participating in the HUPO Plasma Proteome Project. These algorithms included Spectrum Mill, Sonar, Sequest, Mascot and X!Tandem and were compared in terms of sensitivity and specificity at a specified false discovery rate (FDR). Furthermore, the effects of database size and enzymatic digestion (tryptic vs. non-tryptic) on peptide identification were investigated using forward and reverse sequence databases. The samples were prepared using an ion trap mass spectrometer with ESI as ionization source and peptide ions are fragmented using CID. Four research groups independently selected search parameters (such as precursor and fragment ion tolerances) and search strategies based on their experience to optimize results of these algorithms. This suggests that possibly a biased comparison was made among these database search algorithms due to the selection of different search parameters and search strategies optimized for each algorithm. According to Kapp et al. [54], Sequest and Spectrum Mill are more sensitive while Mascot and X!Tandem are more specific. Overall, Mascot performed better than the other algorithms. In addition, although the scores of all the algorithms are precursor charge state-dependent, X!Tandem showed relatively constant thresholds across singly, doubly and triply

charged precursor states. Sequest exhibited higher sensitivity than Mascot with higher search space (database size indicating a negative effect of database size on the Mascot scoring function [54].

Yadav et al. [55] proposed a new method (MassWiz) for peptide identification from MS/MS and compared this approach to OMSSA, Mascot, Sequest and X!Tandem. For the validation of peptide assignments two different datasets were used: (1) a mixture of 18 proteins with known contaminations from six different instruments (AGILENT XCT, LCQ_Deca, LTQ, LTQ-FT, QTOF and ABI-4700); and (2) a complex mid-log phase yeast dataset. Mascot outperforms other algorithms on the standard mixture of 18 proteins both in terms of spectral and peptide assignments, except for spectra from AGILENT XCT that was better identified by MassWiz). Across the six instrument used, OMSSA performed better than Sequest and X!Tandem on four instruments in assigning peptides and spectra. Sequest performed better than OMSSA and X!Tandem on AGLENT XCT spectra, while X!Tandem assigned more peptides than Sequest and OMSSA but low number of spectra than OMSSA on QTOF. For the yeast dataset, OMSSA assigned higher number of spectra and peptides than any other algorithm while X!Tandem assigned the lowest number of all algorithms at 1% FDR [55].

Kandasamy et al. [56] compared OMSSA, Mascot, X!Tandem and Spectrum Mill using approximately 170000 ETD derived MS/MS. For comparison purposes, tryptic peptides and the same search specifications were used for precursor ion tolerance, fragment ion tolerance, number of missed cleavages and fixed and variable post-translational modifications. However, for OMSSA *c*- and *z*-ion series and for the other three algorithms *c*-, *z*- and *y*-ion series were included in the search process. Overall, Spectrum Mill performed better than the other algorithms at 0.1%, 1%,

and 5% FDR adjusted thresholds. OMSSA identified 477 more peptides than X!Tandem at 0.1% FDR, while X!Tandem detected 639 more peptides than OMSSA at the 1% FDR adjusted threshold. Furthermore both OMSSA and X!Tandem performed poorly on doubly charged peptides identifying only less than 1%, and 1 to 12% of peptides across the FDR thresholds, respectively. The poor performance of OMSSA, X!Tandem on doubly charge MS/MS spectrum and the choice of different search parameters could be the possible reasons for the low results of these algorithms as compared to Mascot and Spectrum Mill. Correlation analysis of peptide assignments between the four algorithms showed that only 1/6 of the peptides were identified by all four algorithms in the ETD dataset [56]. In contrast, an approximately 50% correlation between these algorithms was found for the CID dataset in previous studies [48, 54].

Li et al. [57] proposed an intensity based algorithm (SQID) for identification of peptides from MS/MS spectra and compared the performance of this algorithm against Sequest and X!Tandem. Three different datasets were used: (1) a Pacific Northwest National Laboratories (PNNL) data containing 28311 spectra from LCQ ion trap mass spectrometer; (2) a mixture of 18 proteins, resulting on 37044 spectra from ESI ion trap MS; and (3) a yeast data, containing 54799 spectra collected with Thermo LTQ ion trap MS. Similar search specifications were used across algorithms. Overall, SQID performed better than Sequest (the second best performer) and X!Tandem. For the PNNL data, SQID, Sequest and X!Tandem correctly identified 22135, 19678 and 14878 peptides, respectively. Similar performance was observed on the yeast data for SQID and Sequest (4355 vs. 3319 peptides). While X!Tandem performed better than Sequest (3501 vs. 3319 peptides) on the yeast data probably due to lower signal to noise ratio in these spectra. In

terms of unique peptide identifications in the protein mixture data SQID, Sequest and X!Tandem correctly matched 292, 273 and 241 peptides, respectively [57].

1.13 THESIS RESEARCH MOTIVATION

Mass spectrometry experiments allow the identification of peptides in complex mixtures [19]. Spectra from MS/MS can be annotated by spectral search against databases of known sequence or *de novo* sequencing. The database search approach is well suited even in situations characterized by low signal to noise ratio or incomplete fragmentation patterns under low energy CID conditions [23]. Three commonly used open source database search algorithms are: OMSSA [42], X!Tandem [43] and Crux [44]. The database search algorithms mainly work by correlating experimental MS/MS with theoretical spectra generated from known peptides in the databases [4, 48]. These algorithms use different scoring schemes and are optimized for protein identification based on the spectra from tryptic peptides. Several studies have provided insights into the relative advantages of the algorithms to identify peptides. However, no study has focused on understanding the comparative strengths of these algorithms to identify neuropeptides. Neuropeptides have unique features stemming from the complex processing, non-tryptic cleavage and typical small size. The aims of this research project are: (1) to study the strengths and weakness of three open source database search algorithms to identify neuropeptides; (2) to evaluate the effect of peptide charge, length and neutral losses on scoring functions of OMSSA, X!Tandem and Crux; and (3) to provide a comprehensive guidelines for future neuropeptidomics studies.

CHAPTER TWO: EVALUATION OF DATABASE SEARCH PROGRAMS FOR ACCURATE DETECTION OF NEUROPEPTIDES IN TANDEM MASS SPECTROMETRY EXPERIMENTS

2.1 INTRODUCTION

Neuropeptides are involved in intercellular communication, mediate neurotransmission, and regulate many biological processes such as growth, learning, memory, metabolism and neuronal differentiation [1]. Neuropeptides encompass neurotransmitters and peptide hormones and have a critical role in many disorders such as depression, Parkinson's disease, and eating and sleeping disorders [58]. Most neuropeptides range in length from 3 to 40 amino acids and are produced by a complex post-translational processing that includes cleavage of precursor prohormones at basic amino acids (K and R) and removal of C-terminal basic amino acids by carboxypeptidases [1, 58]. In addition, neuropeptide sequences can experience multiple post-translational modifications (PTMs) including pyroglutamination, acetylation, amidation, phosphorylation, and sulfation.

Mass spectrometry (MS) is a well-established technology to identify proteins and peptides. The shotgun proteomics implementation of the bottom-up approach relies on the direct protease digestion (typically with trypsin) with subsequent separation of the peptides. The resulting digested peptides are subjected to tandem mass spectrometry (MS/MS) for identification and assignment to proteins. Database searching is a common approach to identify MS/MS spectra. The overall strategy of database searches is to pair observed and theoretical or predicted spectra. The observed spectra come from MS/MS experiments and the theoretical spectra is the result of *in silico*

prediction based on the known sequence of peptides in a database. Most databases include peptides that are empirically confirmed or predicted from genome sequence assemblies or EST libraries.

Differences between the database search programs in the ability to identify proteins have been reported [49, 54, 59]. These studies offer some pointers to the algorithmic components that may be responsible for the difference in performance among programs. However, these studies evaluated the programs based on detection of fewer than 600 peptides. The inconsistencies have been attributed without proof to the different matching algorithms used as authors have not compared ideal spectra.

The same programs that match experimental to theoretical spectra from the database are used to identify neuropeptides in MS/MS experiments [4, 58, 60, 61]. However, the extrapolation of the strengths and weaknesses of each program to the identification of neuropeptides is not straightforward. This is because there are major differences between protein and neuropeptide detection using tandem mass spectrometry. First, neuropeptides already exist in the sample as endogenous peptides prior to any sample preparation or enzymatic degradation. Meanwhile, the identification of a protein can be inferred by the presence of a component peptide, on the other hand the detection of a neuropeptide requires the precise identification of the exact neuropeptide in the sample and cannot be inferred from the detection of other prohormone peptides. The second distinctive feature is that neuropeptides tend to be small, on average between 20 and 40 amino acid long. This length limits the statistical significance of the match of the peptide to a database and thus the capability to detect peptide matches beyond a user-defined statistical threshold. The third distinctive feature is that neuropeptides may result from cleavages by multiple proteases. Thus, the

digestion model to generate peptides from proteins is not applicable to neuropeptides as many neuropeptides may lack additional basic amino acids or result in smaller peptides. Also, neuropeptides are formed by cleavage at basic amino acid sites that are also cleavage targets of proteases such as trypsin. Consequently, the digestion model may not identify the correct peptide and fail to distinguish between shorter and longer forms of the same peptide. The fourth distinctive feature is that neuropeptides tend to undergo more post-translational modifications than other peptides, on a per-peptide basis. Samples of high complexity or dynamic range are particularly challenging for MS/MS and peptide search algorithms. These conditions are commonly present in samples analyzed for neuropeptide identification and quantification.

Previous comparisons of database search programs [48, 49, 59] have demonstrated the failure of some programs to identify peptides, even under carefully parameter specifications and expert usage of the programs. Kapp et al. [54] found that 15% of human serum and plasma MS/MS spectra were identified by at least one program. Similarly, Balgley et al. [48] reported an average of 34% normal human ovarian epithelium MS/MS matches to distinct peptides, and Xu and Freitas [59] reported the identification of 5% out of 1837 human histone MS/MS spectra. No large-scale, systematic study of the strengths and weaknesses of different database search programs to identify neuropeptides and other potential peptides resulting from prohormone processing have been reported. The unique characteristics of these peptides, compared to all peptides in general, call for the evaluation of database programs and algorithms that best support the identification of neuropeptides. Therefore, an assessment of the peptide database search programs and scoring schemes in the context of prohormone peptide identification is warranted. The neuropeptide research community and the proteomic community will benefit from a better understanding of the

strengths, weaknesses and limitations of the search algorithms available for protein and general peptide identification. The aims of this study were: (1) to compare the relative advantages of three complementary open-source search methods: OMSSA, X!Tandem and Crux to accurately identify prohormone peptides including neuropeptides; (2) to evaluate the impact of mass spectrometry factors such as charge on neuropeptide identification; and (3) to offer guidelines to obtain the most comprehensive and accurate survey of the peptides in a sample.

2.2 MATERIALS AND METHODS

A database of neuropeptides was assembled from neuropeptides available in the SwePep (http://www.swepep.org) and UniProt (http://www.uniprot.org; release 2011_01) databases and potentially cleaved peptides predicted from 92 mouse prohormones using the NeuroPred program (http://neuroproteomics.scs.illinois.edu/neuropred.html; [15]). The final database consisted of 7850 peptides that ranged in length from 5 to 255 amino acids including experimental confirmed neuropeptides and peptides resulting from predicted cleavages of prohormones (Table 2).

Peptides from mouse prohormones were used to simulate the observed or query spectra. Two target databases were used to identify the query neuropeptides: (a) the mouse database; and (b) a rat neuropeptide database including 7647 peptides. The rationale for matching the observed data to the same counterpart in the database without the addition of a decoy database is three-fold. First, a decoy database does not assist in determining if the algorithms can correctly match the spectra to the correct target. Rather a decoy database provides a general measure of confidence of all the matches. Second, the simulated data share the same quality and thus the addition of decoy mass spectra does not aid in addressing quality differentials in the present study. Third,

neuropeptides tend to be short and span a few residues. A reverse decoy spectra of a short peptide has higher likelihood to be present in nature than that of a longer peptide, thus biasing the objective of these spectra to help assess the probability of a random match. The lack of known spectra with no target database entry prevented the comparison of performance across programs using receiver operating characteristics curves.

The rationale for undertaking a cross-species search was three fold. First, the cross-species strategy permitted the assessment of the robustness of the programs to detect prohormone peptides in databases of peptides including all common variants such as single amino acid polymorphisms between mouse strains. This is particularly relevant in light of the multiple mouse genome projects such that the rat sequence is expected to be more different from all mouse species than the differences between mouse species. Second, the strategy allowed the evaluation of the performance of the programs to detect neuropeptides using databases from other species. This is important when considering the large number of species that have not been sequenced and the increasing number of species with sequenced genomes that lack of proteomic verification of proteins and peptides. Lastly, the spectra match is performed on an independent data set.

Among the database search programs available, three public, open source software were considered: X!Tandem [43], OMSSA [42] and Crux [44]). These programs were selected because they are open source and this allows the investigation of the code, computation of matching scores and the algorithmic specifications.

The three programs can be classified as descriptive, heuristic or un-interpreted database searching based on the matching scoring algorithm [62]. Descriptive database search programs are commonly used to identify peptides in MS/MS experiments because they do not require a high-quality spectrum, although low quality spectra tend to result in low matching scores and thus may fail to lead to peptide identification. These algorithms correlate the observed or query mass spectra to the predicted mass spectra in the database. From this correlation, a score indicator of the similarity between the query and database spectra is produced and a probability that a particular peptide sequence generated the observed spectrum is obtained by chance [54]. The score is then used to compute an *E*-value (*Expected*-value) or *p*-value. The first indicator is the expected number of database matches by chance with scores equal or higher than the one observed. The second is the probability that the match between the query and target sequences is due to chance. A brief description of the three database search programs follows.

Crux (Version 1.37 released on December 22, 2011) [44] is an alternative implementation of the SEQUEST algorithm [45]. Peptide identification relies on searching a collection of spectra against an indexed sequence database, and returning a collection of peptide-spectrum matches (PSMs). Crux option to calculate *p*-values from a Weibull distribution of the cross-correlation scores [49] was used in this study. Although this approach is computationally intensive, this strategy maximizes sensitivity or true positive rate through the ability to identify peptides regardless of the quality of the spectra at the expense of higher rates of false positives or mismatches.

X!Tandem (http://www.thegpm.org/tandem; Version 2010.12.01.1 released on December 01, 2010) [43] was developed to optimize speed and to minimize the computational requirements. The algorithm includes preprocessing of the observed spectra to remove noise and technical artifacts, process database peptide sequences with cleavage reagents, post-translational and chemical modifications and scores the peptide matches between the observed and predicted spectra [49]. The scores are converted to hyperscores and the distribution of hyperscores of all matches is used to translate the hyperscore of each match into an *E*-value that indicates the number of peptides in the database that are expected to exhibit matching scores equal or higher than the one under consideration by chance alone.

The Open Mass Spectrometry Search Algorithm (OMSSA; Version 2.1.7 released on June 15, 2010; http://pubchem.ncbi.nlm.nih.gov/omssa) attempts to optimize the speed of the database searching approach [42]. The scoring of each match assumes that the number of matches between observed and predicted peaks for a peptide sequence follows a Poisson distribution. The lambda (or average) parameter of the Poisson distribution is calculated as a function of the fragment ion tolerance, the number of predicted and observed peaks and the neutral mass of the precursor ion. OMSSA provides the probability that the match between the observed and predicted spectrum is the result of chance and corresponding *E*-value based on the dimensions of the target database.

Simulated spectra were used to compare the performance of the three database search programs. There are three advantages of simulating the observed peptides to be queried against a database. First, the use of simulated mass spectra overcomes the limited number of neuropeptides with mass spectra information of comparable quality obtained using the same or similar

technologies. Second, the analysis of simulated mass spectra that share the same quality level allows benchmarking the database search programs irrespectively of sample or data quality issues including low mass accuracy, noise and low signal to noise ratio. Third, simulated mass spectra offers an absolute control of the peptides that should be detected and accurate evaluation of the number of true positives (detected and correctly identified peptides), false positives (detected but incorrectly identified peptides) and false negatives (missed peptides).

Ideal uniform spectra that have either +1, +2, and +3 charge states were simulated for each peptide precursor ion in the target database. For each precursor charge status, only +1 charged *b* and *y* product ions were simulated with equal intensity. The product ions with equal intensity values were simulated to avoid selection of only high intensity peaks for scoring and thus to eliminate the effect of intensity on scoring functions of database search methods. Neutral losses of a water and/or ammonia were simulated if the ion contained either one of four water losing amino acids (S, T, E, D) or ammonia losing amino acids (R, K, Q, N). Neutral losses from *b* and *y* product ions occurred regardless of position of these amino acids in the ions. Complementary scenarios of neutral mass loss and ion availability conditions were simulated across the three precursor charge states and searched against the database to investigate the impact of these situations on the identification of neuropeptides.

1) All *b*- and *y*-ion series with all neutral mass losses due to water and ammonia,

2) Only the possible *b*- and *y*-ion series,

3) Only the possible *b*-ion series with all neutral mass losses,

4) Only the possible *y*-ion series with all neutral mass losses,

5) Random 50% of *b*- and *y*-ion series with all neutral mass losses,

6) Random 25% of *b*- and *y*-ion series with all neutral mass losses,

7) Only scoring the *b*-ion series from *b*- and *y*-ion spectra with all neutral mass losses,

8) Only scoring the *y*-ion series from *b*- and *y*-ion spectra with all neutral mass losses.

Under low energy CID conditions in MS not all product ions are detected in common *b*- and *y*-ion series. The simulation of the random 25% and 50% of ions represents one type of incomplete fragmentation. The aim of simulating spectra with missing ions is to determine which ions are sufficient for the database search algorithms to accurately identify peptides.

In addition, charge state +1 precursor ions from mouse spectra were searched against the rat database using X!Tandem. For the latter evaluation, only SwePep and UniProt identified peptides were used to represent the experimentally known mouse peptides and avoid ambiguous annotation of corresponding rat peptides. This strategy minimized the likelihood of matches between small and large peptides where the smaller peptide is a known cleavage product of the larger peptide. X!Tandem was selected for this evaluation because this program was found to the more conservative of the three programs.

A set of composite spectra is simulated by combining product ions from more than one peptides having similar precursor m/z values (mass error ± 0.4 Da). To create composite spectra

that has minimum biasness towards any database search method, only those peptides were selected which were individually identified by all database search tools at an *E*- or *p*-value < 1 x 10^{-2}. The 945 composite spectra from 2049 peptides were grouped into four classes based upon number of peptide ions used to produce composite spectra. The purpose of this strategy is to assess the ability of OMSSA, X!Tandem and Crux to correctly identify peptides from composite spectra, representing simultaneous fragmentation of different peptide ions with same precursor m/z values in mass spectrometry conditions. Ideal uniform spectra with all *b*- and *y*-ions with neutral loss at precursor charge state +1 were used in this study.

The peptide identification search programs OMSSA, X!Tandem and Crux were evaluated using comparable algorithmic specifications and excluding PTMs. The default values of the programs were used in addition to the following specifications: (1) precursor ion tolerance: 1.5 Da; (2) product or fragment ion tolerance: 0.3 Da; (3) no fixed or variable modifications; (4) "whole protein" (OMSSA) or "enzyme: custom cleavage site" (X!Tandem and Crux) to prevent cleavage since the detection of neuropeptides does not involve protease cleavage; (5) peptide length: 5-255 residues; (6) precursor ion charge: 1+, 2+, 3+; (7) product ion charge: default values; (8) no complete or partial modifications; and (9) peptide mass: monoisotopic.

For comparison purposes, Crux probability scores (ranging from 0 to 1), X!Tandem *E*-values (ranging from 1 x 10^{-45} to 1 x 10^{+3}) and OMSSA *E*-values (ranging from 1 x 10^{-15} to 1 x 10^{+4}) were transformed using a base 10 logarithm. The match or hit with lowest *E*- or *p*-value among all hits per input spectrum was analyzed. At 1 x 10^{-6} threshold based on a 1% Bonferroni

correction ($0.01/7850 = 1.27 \times 10^{-6} \approx 1 \times 10^{-6}$) was used to determine if the match was significant while accounting for multiple testing.

2.3 RESULTS AND DISCUSSION

The overall significance and the correctness of the matched sequence of the query-to-target matches were used to assess the capability of each search algorithm to detect neuropeptides and other prohormone peptides. This evaluation step allowed discrimination between obvious and dubious, yet correct, peptide identifications. The decision tree used to assess the performance of each database search program is presented in Figure 4. A peptide match was deemed to be significant if the detection signal (e.g. *E*- or *p*-value) was lower (more significant) than a threshold $< 1 \times 10^{-6}$. This stringent threshold aimed to minimize the number of false peptide identifications because the percentage of matches that could be considered by chance (false positives) is less than a 1% Bonferroni corrected significance threshold. There were three outcomes for each simulated spectra: the neuropeptide correctly matched the simulated peptide (true positive), incorrectly matched (false positive) or failed to match (false negative).

Table 3 summarizes the results from the three search methods across three precursor charge states. From a total of 23550 simulated spectra (7850 peptides x 3 precursor charge states), OMSSA, X!Tandem and Crux had 23281 (98.9%), 22117 (93.9%) and 20890 (88.7%) true positive results (correct spectrum-peptide matches at strong *E*- or *p*-value $< 1 \times 10^{-6}$), respectively. Our results are consistent with previous reports of a higher number of spectra matched by OMSSA than by X!Tandem [55, 63]. At an unadjusted 1% threshold $< 1 \times 10^{-2}$ (i.e., no multiple test adjustment), OMSSA, X!Tandem and Crux had 23548 (99.9%), 22932 (97.4%) and 23139

(98.3%) true positive identifications. The consensus among programs suggested that the overlapping peptides are less susceptible to the assumptions and models used by each database search algorithm. In total, 20740 (88%) peptides were correctly identified by all three programs. The majority of the 267 (1%) peptides that were not identified by any one program were five amino acids in length (Table 3). The remaining peptides were found by OMSSA (4%) or OMSSA and Crux (1%) or OMSSA and X!Tandem (6%). X!Tandem provided no significant peptides that were not detected by at least one other program. This suggest that X!Tandem may be the most conservative program evaluated or that the algorithm is less sensitive.

Selected shared and distinct identifications among all three database search programs are highlighted using Venn diagrams. A Venn diagram depicting the common and distinct true positive peptides identified from the three database search programs, X!Tandem, OMSSA, and Crux using information from *b*- + *y*-ion series and peptide charge state 3 is depicted in Figure 5. This diagram underlines the substantial overlap between all three programs, between OMSSA and X!Tandem and the ability of OMSSA to identify additional peptides at charge state 3. Figures 6 and 7 present Venn diagrams depicting the common and distinct peptides identified by all three programs using only *y*- and *b*-ion series information and peptide charge state 3, respectively. These figures stress the relative advantage of Crux when only *b*-ion series are available and of OMSSA and X!Tandem when only *y*-ion series are available for peptide identification. Figures 8 and 9 present Venn diagrams depicting the common and distinct peptides identified by all three database search programs using only 50% or 25% of all ion information and peptide charge state 3 are available, respectively. These figures stress the increasing detrimental impact of missing ions on the performance of X!Tandem.

The length of the peptide had an impact on the statistical significance of the match for each program. Overall the correlation between the length of the query sequence and $\log_{10}$ transformation of the *E* or *p*-values for OMSSA, Crux and X!Tandem was 0.1%, 86.8% and 46.7%, respectively. However, the relationship was non-linear. Figure 10 depicts the relationship between the $\log_{10}$ transformed *E*- or *p*-values on peptides across peptide length. Examination of the relation between query length and log-transformed *E*-values showed rapid increase up to 11 and 15 amino acid long peptides for OMSSA and X!Tandem, respectively, before the log-transformed *E*-values stabilized. In contrast, the Crux log-transformed *p*-values showed a gradual increase to approximately 50 amino acids before the log-transformed *p*-values stabilized. Kapp et al. [49] also noted that small peptides between 600 and 700 Da were factor in peptides not identified across programs. A similar effect of peptide length on the distribution of the MaxQuant program p-scores between target and decoy database was observed [64]. In that study, peptides smaller than 15 amino acids long had a higher likelihood of being incorrectly matched than larger peptides.

The increase in the *E*-values with decreasing peptide length is due to the corresponding increase in the number of expected matches by chance. The mean of the underlying Poisson distribution used by OMSSA decreases with smaller peptides resulting in larger *E*-values due to the increased probability of a random match. In particular, the detection of short peptides by OMSSA is negatively influenced by the tendency of small peptides to exhibit neutral mass losses. Similarly with X!Tandem, the observed reduction in *E*-value significance is associated with a lower number of unique peptides that can be matched relative to larger peptides. For peptides less than 12 amino acids, the correlations between OMSSA and X!Tandem, OMSSA and Crux, and

X!Tandem and Crux were 78%, 63% and 52%, respectively. This result also indicates that the selected threshold was more stringent in Crux and X!Tandem than for OMSSA.

The statistical significance of the X!Tandem matches was inferred using the lowest scores from the matches. Consequently, the significance values assigned by X!Tandem is negatively influenced when there are insufficient matches to provide an accurate estimate of the X!Tandem score. At a computational extreme, Crux uses a resampling test where random permutations of matching sequences are generated and scored. The implementation of this procedure in Crux is flawed because resampling with replacement is permitted and this potentially allows the same sequence to be repeatedly sampled. Unlike Crux and X!Tandem, the OMSSA *E*-value is derived from the assumption that the number of matches can be represented by Poisson distribution does not depend on the matches or generated sequences although it relies on the database size. The OMSSA formulation is also dependent on peptide size so that small peptides tend to be on the low bound of significance due to the smaller proportion of ion matches than larger peptides. For example, if the Poisson mean is equal to one, then the probability of zero ion matches is 0.37%.

The search time of the three database search programs was a function of the number of neuropeptides in the search database. This computational comparison is empirical and that the database search can be easily computed in parallel because the experimental spectra can be independently analyzed. The computational speed was measured on a 3.00 GHz Intel X9650 processor to evaluate all 7850 peptides for all programs. X!Tandem returned results the fastest (averaged 23 cpu seconds), followed by Crux with no *p*-value calculation (3.8 x more time than X!Tandem; averaged 89 cpu seconds) followed by OMSSA (5.3 x more than X!Tandem; averaged

123 cpu seconds). The Crux *p*-value calculation adds considerable time due to the permutation-based approach to assess the statistical significance of the database match. This approach requires the generation and scoring of dummy sequences to obtain the Weibull density for each match. Consequently, the computation of *p*-values for 100 and 1000 dummy sequences required over 1 hour and 13 hours of cpu time, respectively. The increase in time is linear on the number of sequences evaluated such that each sequence took approximately 47 cpu seconds. This resampling test approach is not limited to Crux so a similar increase in time would occur when this approach is used with X!Tandem and OMSSA.

Crux was the only program that was able to correctly match all peptides although only 12 peptides with less than ten amino acids had *p*-values < 1×10^{-6} threshold. At unadjusted *p*-value < 1×10^{-2} threshold 33%, 64%, 75%, 87%, 98%, 99% and 100% of the 5, 6, 7, 8, 9, 10, and 11 amino acid peptides were detected with Crux. This increase is partly due to the number of Weibull samples because with 100 permutations only 61 of the peptides with charge state 1 had *p*-value < 1×10^{-5} threshold (results not shown). Consequently, adding further Weibull samples especially for small peptides may increase the significance levels by providing more accurate density estimation. X!Tandem was not able to correctly detect the 85 peptides with length of five amino acids which accounted for 1% of all peptides. Across the different scenarios, most of these peptides were not detected (80%) and the rest were incorrectly identified (mismatched). X!Tandem was able to correctly match at *E*-value < 1×10^{-6} peptides ten amino acids in length or larger, although 94% of seven amino acid long peptides and all eight and higher amino acid long peptides surpassed the unadjusted *E*-value < 1×10^{-2} threshold. OMSSA was also influenced by peptide size as most peptides larger than nine amino acids surpassed the *E*-value < 1×10^{-6} threshold. However, 100%,

27%, 6%, and 1% of the peptides with five, six, seven and eight amino acids, respectively did not reach significance with OMSSA.

There were five unique neuropeptide sequences that were not first ranked peptides in OMSSA across all simulated conditions. Four of the peptides were associated with the highly homologous Oxytocin-neurophysin 1 (NEU1) and Vasopressin-neurophysin 2-copeptin (NEU2) prohomones. These peptides were further reduced to two sets of peptides after consideration of ambiguous cleavage site that leads to two possible peptides within homolog. The simulated spectra of the *b*- and *y*-ions without neutral mass loss were very similar between these peptides with the largest difference of 19.9 m/z occurring at the b_6 ion. Another mismatch occurred with a PENK (UniProt id P22005) peptide due to the multiple occurrences of the Met-enkephalin in a longer peptide. For simulated charge state 1 and 2 including neutral mass loss, a mismatch occurred between the two Met-enkephalin peptides located at the C- and N-terminal. Due to the similarity in sequence and *E*-values, these peptides were treated as "homeometric peptides" [65] and were considered as correct matches.

OMSSA failed to detect one peptide (a SCG1 peptide; P16014[592-652] predicted by the NeuroPred mouse model) in all three charge states. Also, OMSSA had one mismatch (a SCG2 peptide Q03517[475-547] from a non-mammalian model match to NEUT; Q9D3P9[87-156]) with charge state 3 with an *E*-value > 60. This peptide was detected by OMSSA in the other two charge states. Both peptides were correctly detected when the simulation excluded neutral mass losses. This suggests a weakness (or lower sensitivity) of the algorithm to accommodate neutral mass losses. Examination of both peptides indicated that 54% of the amino acids in each sequence were

prone to lose water (28% of the amino acids) and ammonia (25% of the amino acids). As a result approximately 2/3 of the ions can include neutral mass losses and OMSSA apparently failed to distinguish the series with and without neutral losses.

Combining identifications that were significant in at least two programs improved the average identification rate across all three charge states from 89% to 94% when all ions were available for scoring and including neutral mass loss. Using a consensus approach, as has been advocated in the identification of proteins [49], can improve peptide identification because the probability of all programs incorrectly identifying a peptide is equal to or less than probability of the least accurate program being incorrect. While this consensus approach assists in the correct identification of peptides, it is less suitable to the goals of the present study because the individual programs helps us to understand the particular distributional features of the prohormone peptide population of mass spectra relative to protein database searches and recommend the best tools for particular neuropeptides. For example, a closer inspection of the few peptides (5%) that were not consistently identified across programs revealed that these peptides were correctly matched by at least Crux and OMSSA although exhibited low scores, irrespectively of the programs, due to the small size of these peptides (ranged between 5 and 11 amino acids in length). This result suggests that for these few neuropeptides all three programs have comparable disadvantages but have complementary strengths and weaknesses to detect neuropeptides.

NEUTRAL MASS LOSSES

Table 4 and Table 5 summarizes the performance of the three programs in the identification of peptides when all ions from both series are available including and excluding neutral mass losses, respectively. The inclusion of neutral mass losses had minor influence on the overall detection of peptides across the programs (Table 3). The average percentage of detected peptides at *E*- or *p*-value < 1 x 10^{-6} with neutral mass loss over all charge states was 98%, 94.4% and 89.6% in OMSSA, X!Tandem, and Crux, respectively. Slightly more peptides were detected for all three programs (89%) when neutral loss was included in the simulated query. However, the percentage of undetected peptides increased to 2%. This was mainly due to a decrease in the number of peptides identified by OMSSA either alone (111 peptides over the three charge states) or with X!Tandem (223 peptides over the three charge states) or with Crux (128 peptides over the three charge states).

Peptide detection by Crux was largely not affected by neutral mass loss scenarios. Slightly more peptides were correctly identified with neutral mass loss than without neutral mass losses at higher significance levels (*E*- or *p*-values < 1 x 10^{-9}). The inclusion of neutral mass loss noticeably influenced the significance levels of X!Tandem matches that were already highly significant (*E*-value < 1 x 10^{-10}). The *E*-values of the correct matches decreased in significance from a median of *E*-value < 1 x 10^{-25} to *E*-value < 1 x 10^{-14} when neutral mass losses were added to the same queries.

OMSSA identified 99% of the queries without neutral mass loss across charge states. Peptides simulated without neutral mass losses had more significant *E*-values than peptides simulated with neutral mass losses. This trend was reflected by the median *E*-value of peptides

with *E*-value < 1 x 10^{-6} decreasing from 8.8 x 10^{-12} to 5.6 x 10^{-13} for simulations with and without neutral mass loss, respectively. However, overall the impact of neutral mass loss is low considering that at *E*-value < 1 x 10^{-6}, more than 98.3% of the peptides were correctly matched across both neutral mass loss scenarios.

Comparison of OMSSA peptide detection across charge states with neutral loss showed a difference between charge states that was absent when no neutral losses were simulated. At the stringent threshold *E*-value < 1 x 10^{-10}, peptides that have precursor charge state 1 had more significant matches (93%) than precursor charge state 3 (81%). This difference decreased with less stringent thresholds and at the *E*-value < 1 x 10^{-6} threshold the difference in detection was only 1% between charge states 1 and 3. This may be partially explained by the assumption that +2 product ions are present in charge state +3 and higher spectra but not present in charge state +2 spectra [42]. This assumption results in a higher number of possible ions and a consequently a lower *E*-value even if the spectra lacks these highly charged product ions. These results suggest that the algorithm of OMSSA has a high risk to fail the identification of peptides that contain numerous amino acids prone to neutral mass losses, and this risk is higher at higher charge states.

Peptide size influenced the number of true positive matches across all database search programs, even when all ions were present and there were no neutral mass loss simulated. No program correctly detected five amino acid peptides. However, OMSSA and X!Tandem correctly detected all peptides longer than 7 and 11 amino acids, respectively. There was a gradual increase in the number of peptide matches that have *E*-value < 1 x 10^{-6} with Crux although only peptides 46 amino acids long and higher surpassed this threshold.

MISSING IONS

Tables 6, 7, 8, and 9 summarizes the performance of the three programs in the identification of peptides when only *b*-ion series, *y*-ion series, random 50% of all ions, and random 25% of all ions are available including neutral mass losses, respectively. The average percentage of correct identifications across charge states varied from 85% to 94% when 50% of all possible ions were available and between 69% and 83% when 25% of ions were available. The proportion of correct identifications in all programs was 87%, 88%, 85% and 68% when *b*-ion series, *y*-ion series, random 50% and random 25% of the ions were available, respectively. The proportion of unidentified peptides in all programs was 8%, 6%, 7% and 14%, when *b*-ion series, *y*-ion series, random 50% and random 25% of the ions were available, respectively. The lower percentage of peptides identified in scenarios that had 50% and 25% of the ions available was mainly due to a poorer performance of X!Tandem. This conclusion is based on the percentage of peptides correctly identified by Crux and OMSSA that increased from 3% to 11%. Missing ions also impacted the detection of peptides by OMSSA because the percentage of correctly identified peptides by Crux only increased from zero in the random 50% scenario to 4% in the random 25% of the ions scenario.

Unlike for the other two programs, missing ions on the query had minor influence on the identification and significance level of the peptides in Crux. The availability of only one ion series resulted in 89% and 90% when only the *b*- and *y*-ion series were available, respectively, and was similar to the 90% when all ions were used. The percentages of peptides that had significant matches were 88% and 83% when 50% and 25% of the ions were available, respectively. On

average 119 and 480 fewer significant matches were detected across the three charge states when 50% and 25% of the ions were available, respectively.

The number of ions available for scoring affected the detection of peptides by X!Tandem, regardless of the charge state. The number of correctly detected peptides decreased from 94% when all ions were available to 89% when only the *b*- or *y*-ion series were available. Furthermore, availability of 50% or 25% of all possible ions reduced the number of correctly identified peptides to 85% and 69%, respectively. The minimum length for a peptide to be detected at significance *E*-value $< 1 \times 10^{-6}$ were 13, 15, 14 and 81 amino acids for the *y*-ion series, *b*-ion series, 50% ions and 25% ions available, respectively, compared to ten amino acids when all ions were available. These trends are likely to be related to the number of ions that can potentially be available rather than the percentage of ions available.

The absence of the *b*-ion series, *y*-ion series, or 50% of all ions at random had minor effect on the identification of peptides using OMSSA compared to the availability of all ions. On average, 95% of the peptides identified when all ions were available were also identified when only 50% of the possible ions were available. On average 65 peptides across all three charge states were undetected when only 25% of all ions were available. These peptides were between five and eight amino acids in length. Furthermore, only peptides of more than 25 amino acids in length had *E*-value $< 1 \times 10^{-6}$.

There was a slight tendency for the number of undetected peptides by OMSSA to increase with the increasing charge state when only 50% of ions were available. This result is consistent

with the observed trend in the presence of neutral loss simulation suggesting that the presence of neutral loss rather than the absence of 50% of ions was the factor driving the lower detection rate. However, when only 25% of the ions are available, charge state 1 peptides were four times more likely to be undetected relative to higher charge states. These results indicate that the absolute number of ions present is potentially more critical to the OMSSA algorithm than the relative percentage of ions available. Overall, these findings point out that there was a diminishing return on accurate identification for additional ions used by OMSSA, with the detection *E*-value threshold dependent on the precursor charge state. Longer peptides are expected to generate more ions, suggesting that the OMSSA scoring system based on the actual number of mass spectra peak matches needs to account for the overall peptide length. This adjustment is particularly critical for neuropeptides because these peptides tend to be small.

The simulation of the random proportion of ions represents one type of incomplete fragmentation that is an important component of variation between programs and a lack of peptide identification [49]. Peptides can be identified by the programs when incomplete fragmentation provided sufficient ions are present especially for large peptides. The difficultly is assigning an appropriate significance threshold since most of the peptides were correctly matched regardless of program used. The low impact on Crux is possibly due to the lack of resampled peptides that share similar ion patterns. The OMSSA *E*-values increased with fewer ions present since the *E*-value computation assumes that all possible ions are present. X!Tandem is clearly negatively influenced by the decreased number of ions available. A possible explanation is that with fewer ions present, the score of the correct match is not sufficient different from the incorrect matches as with all ions, both leading to a low score.

ION SERIES-DEPENDENT SCORING

An alternative approach to improve the speed of programs is to search only one ion series. Spectra scoring using a single ion series affected only OMSSA and X!Tandem. Table 10 summarizes the performance of the database search programs by ion series scored. Tables 11 and 12 summarize the performance of the database search programs across match significance levels when the *b*- and *y*-ion series are scored, respectively.

Scoring only one of the ion series was noticeably detrimental to peptide detection for both OMSSA and X!Tandem. The *y*-ion series provided a higher detection rate (89% and 82% for OMSSA and X!Tandem, respectively) than the *b*-ion series (87% and 78% for OMSSA and X!Tandem, respectively). Consequently, only 50.8% and 55.3% of peptides were correctly detected by both programs using the *b*- and *y*-ion series, respectively. The percentage of false positive peptide matches in both programs was 6% and 5% for the *b*- and *y*-ion series, respectively. Very few peptides were incorrectly identified by both programs (0.3%); X!Tandem and OMSSA incorrectly matched 0.8% and 3% of the peptides, respectively. This result indicated that OMSSA may be more prone to false positive results when scoring only one ion series.

The major reason for weaker performance of X!Tandem when scoring one ion series was that the peptides had less significant *E*-values than when scoring both ion series. At the *E*-value < 1 x 10^{-5} threshold, 78% and 82% of the peptides were identified by X!Tandem when only the *b*-ion series and *y*-ion series were scored, respectively. Consequently, using an unadjusted *E*-value < 1 x 10^{-2} threshold, 90% and 92% of peptides were detected in both X!Tandem and OMSSA when *b*- and *y*-ion series were scored, respectively. The major difference between the two ion series in

OMSSA was that scoring using only *y*-ion series resulted in fewer unmatched peptides (difference of 34), mismatched peptides (difference of 12) and peptides with less significant *E*-values (98% of peptides with the *E*-value < 1×10^{-2}). In both programs, higher charge states were associated with slightly poorer peptide detection with scoring based on the *y*-ion series being less affected than on the *b*-ion series.

The length of the peptide was also critical when one ion series was used to score the matches between the query and target database mass spectra. The minimum length among the detected peptides was 10 and 13 amino acids in OMSSA and X!Tandem, respectively, compared to six amino acids when both ion series were scored. Consequently, the median length of the correctly identified peptides with both programs using the *b*- and *y*-ion series was 83 and 76 amino acids, respectively. Also, the median length of the missed (false negative) peptides in both programs was seven and six amino acids when the *b*- and *y*-ion series were used, respectively. This result reflects the issues of correctly identifying small peptides at *E*-value < 1×10^{-6}.

CHIMERA SPECTRA

The presence of chimera spectra is a likely event among prohormone peptides and the performance of the programs under these circumstances was evaluated. The 2049 peptides that had at least one other peptide with theoretical mass ± 0.4 Da were split into 945 groups including at least two peptides within a theoretical mass range or tolerance within group. Of these, 804, 126, 12, and 3 groups included 2, 3, 4 and 5 peptides, respectively. The performance of each program to identify the peptides in a chimera was evaluated. Overall, Crux had the best performance and

X!Tandem generally failed to identify peptides from chimera spectra regardless of the threshold used. At a *E*-or *p*-value < 1 x 10^{-6}, OMSSA, X!Tandem and Crux, correctly identified 79%, 10% and 76% of peptides in each group, respectively (Table 13). At *E*- or *p*-value < 1 x 10^{-2} threshold, OMSSA, X!Tandem and Crux, correctly identified 81%, 43% and 99% of peptides, respectively. Further study of the peptides identified at *E*-value < 1 x 10^{-2} indicated that X!Tandem only detected one peptide in the chimera spectra resulting in 93% of spectra with at least one peptide correctly matched. At the other extreme, Crux only had three unmatched peptides at *p*-value < 10^{-2}. OMSSA had a correct match rate similar to Crux at *E*-value < 1 x 10^{-6} and lower at *E*-value < 1 x 10^{-2} threshold. The typical OMSSA correct match is at a high *E*-value or else OMSSA fails to match the peptide. A further decrease in the accuracy of peptide identification in chimeras was observed for the peptides less than 10 amino acids in length in case of X!Tandem and Crux. At *E*- or *p*-value < 10^{-2} , the correct identification by OMSSA, X!Tandem and Crux was 81.2%, 37.6% and 91.8%, respectively. Consistent with our results, Houel et al. [66] reported that MASCOT correctly identified peptide A in 87% of chimera spectra containing 50% of peptide A and 50% of peptide B.

ACROSS-SPECIES COMPARISON

The identification of peptides using information from sequence variants or a different species must consider the potential impact of non-synonymous amino acid changes, insertions and deletions. Changing the precursor tolerance is a simple approach to account for non-synonymous amino acid changes. Increasing the precursor tolerance during the database search process allows the evaluation of a wider range of peptide sequences, resulting in an increment in the potential

number of matches. The disadvantages of increasing the tolerance include the increased number of candidates to evaluate (increasing computational time) and an increased chance of an incorrect match. The evaluation of simulated spectra permitted the assessment of the impact of the precursor tolerance on the accuracy of peptide identification.

At a 100 Da precursor tolerance, equivalent to less than an "average" amino acid in difference, four and two mouse peptides had two and three matches in the rat database, respectively. None of the additional matches to the mouse peptide correctly identified the corresponding or expected rat peptide. At a 200 Da precursor tolerance that encompasses most amino acid mutations, 13 and 1 mouse peptides had two and three rat matches although there were only four second ranked matches to the expected rat peptide. At a 500 Da precursor tolerance that encompasses up to three amino acid substitution there were 16 and 1 mouse peptides that had two and three rat matches, respectively, but only seven matches identified the expected rat peptide. At a 1000 Da precursor tolerance that permits multiple amino acid changes, 17, 3, and 1 mouse peptides had two, three and four rat matches, respectively, although only seven peptides that had two matches identified the expected rat peptide. In all cases, when there was more than one correct match (when complete and incomplete prohormone sequences were used to generate rat peptides), one of the rat matches had a closer mass to the mouse peptide than the other match. Among the six mouse peptides that had second ranked matches to the expected rat homolog, three, two and one peptides had second ranked matches at precursor tolerance < 1000 Da, 1000 Da < precursor tolerance < 500 Da, and 500 Da < precursor tolerance < 200 Da, respectively.

This limited ability of programs to handle sequence differences without adjustments in the search parameter specifications was also reported in a study of the similarity between the target and an isobaric decoy database by Colaert et al. [67]. In this study, peptides that were isobaric to the correct sequences were generated by permuting a pair of amino acids or using insertions or deletions. The resulting decoys became almost homeometric because only the ions that fragment at the modified region were actually different. No obvious difference in the type of change was reported however, the importance of the location of the change within the peptide was not examined.

Irrespective of tolerance level, 91 and 110 mouse peptides were correctly matched to the expected rat peptide at *E*-value < 1×10^{-6} and *E*-value < 1×10^{-2}, respectively. Most matches were exact (80) or had a single amino acid substitution (15). Only three mouse peptides matches that had an *E*-value < 1×10^{-2} across all tolerances matched a longer or shorter form of the expected rat counterpart in at least one tolerance level. Only one peptide that differed in length by a single amino acid between the rat and mouse sequences was identified. Among the mouse queries, 45 peptides matched an incomplete or truncated rat peptide. Most of the 19 partial peptide matches (13) had an *E*-value < 1×10^{-6} in particular precursor tolerance levels and this could be due to matches with peptide lengths. Most of the 26 non-significant *E*-value < 1×10^{-2} matches (19) exhibited amino acid substitutions and ten were due to a single change.

Regardless of the precursor tolerance level, 96 mouse peptides were not correctly matched. In these cases, the difference between the mouse and homologous rat sequences averaged 202 Da and ranged from 2 Da to 3727 Da. All of these peptides had more than one difference between

species including sequence length and amino acid substitutions. All peptides that differed in sequence length also had at least one amino acid substitution.

There were 57 mouse peptides unmatched in the rat database that had the same length in both species. These included 31 peptides that are annotated in UniProt as propeptides, 12 peptides that are known to be cleaved further to produce bioactive peptides, three UniProt named peptides that are the only named peptides produced from the prohormone and additional 11 UniProt named peptides that do not correspond to any of the previous groups. Only three of these unmatched mouse peptides differed from the rat peptide by a single amino acid. The remaining 54 peptides had the same length in both species (i.e., alignment with no gaps) and included multiple amino acid substitutions.

Among the mouse peptides unmatched in the rat database, four peptides had the same length in both species and are expected to have biological function. Two large neuropeptides, adrenomedullin and osteocrin, have multiple amino acid substitutions indicating that mutations are more likely to be tolerated in non-critical regions compared to critical regions. The relaxin B chain (P47932[23-57]) has nine amino acid differences spanning over 35 amino acids, and these occur near the center of the peptide although this peptide is not well conserved across species. Neuropeptide B-29 (Q8K4P1[22-50]) only had two changes near the center of the peptide (at position 9 (P to S) and position 18 (S to A) mouse to rat) that resulted in incorrect peptide identification.

The remaining 39 unmatched mouse peptides differed in peptide length and amino acid substitutions between species. Only 11 of these peptides had at least four amino acid substitutions and differed in length by at least three amino acids. In most cases the difference in length was due to a loss of amino acids rather than sequence differences or unreported cleavages. However, for the neuropeptide Vascular endothelial growth factor D there were sequence differences because the UniProt rat version is not the complete rat prohormone sequence. The average number of gaps between the mouse and rat sequences was 3.6 indicating that X!Tandem was not able to simultaneously accommodate variations both in sequence length and amino acids.

For the unmatched mouse peptides, increasing the precursor tolerance over 100 Da added 19 peptides that had non-significant matches when the precursor tolerance was 100 Da. However, the increased precursor tolerance also resulted in the matches to longer or shorter versions of the correct match rather than the correct match of a similar size. Typically, when this occurred the correct match was detected at a lower precursor tolerance level. In many cases this was due to the precursor tolerance being too small to allow for the mass difference between the mouse and rat sequences. Although for one peptide there was a significant match that was not present in the higher precursor tolerance levels. Consistent with the increased tolerance that permitted matching of truncated peptides that shared the same sequence, all of the matches involved truncated peptides. This result was consistent with the increased tolerance that allows additional amino acids. The correct match would have been identified by the actual mass difference between the observed and theoretical masses.

With the exception of one mouse peptide, all the unmatched peptides that had a corresponding rat sequence of the same length were not considered to be functionally constrained due to inactive peptides or changes in the non-critical functions. This lack of functional constrain permits a potentially large number of non-synonymous substitutions in these peptides. This results in species differences that cannot be accounted for in the standard mass spectrometry just by using increased tolerance and a known sequence in one species. Rather, the identification of peptides with a large number of potential variation requires a suitable database or error tolerant search program [68, 69].

Further characterization of the differences between mouse and rat sequences helped to understand the impact of sequence variations on the database search methods. The differences between the mouse and rat peptide sequences were characterized using two complementary criteria, the mass differences and Levenshtein distances. The Levenshtein or edit distance is defined as the minimum number of edits needed to transform one string into the other, with the allowable edit operations being insertion, deletion, or substitution of a single character. In this study string corresponds to a peptide sequence and a character corresponds to an amino acid. The median mass difference and Levenshtein distance for the correct matches was zero, reflecting that most of these peptides had the same sequence in rat and mouse. The mouse peptides that were not matched had the largest Levenshtein distance and average mass difference with the expected rat counterpart. The Levenshtein distance and mass differences for the marginally significant (*E*-value $< 1 \times 10^{-2}$) correctly matched peptides averaged 37.4 and 1.36, respectively. The difference in *E*-value between mouse and rat peptides with partial matches in some precursor tolerance levels appeared mainly due to the higher average Levenshtein distance than mass differences.

Peptide sequence dissimilarity and mass differences were insufficient to explain the matching and significance of the match. As previously noted, the higher precursor tolerance typically accommodated some of the previous differences in peptide mass. Evaluation of the individual peptides that differ among species indicated that location of the sequence has a role in the determining match and associated significance. To further investigate this, the number of differences (gaps plus substitutions) in the aligned rat and mouse sequences was determined for the N-terminal 25%, the C-terminal 25% and the remaining middle 50% of the total length of the aligned peptides. The region with the largest impact is the C-terminal region where the average number of sequence differences increased from 0.02 in the significant correct matched peptides to 3.76 in the non-significant, non-matched peptides. A very similar difference was also observed when there were no gaps in the alignment (0.02 vs. 3.42) indicating that insertions or deletions were not factors in the identification of peptides using databases that include variants. The difference in the C-terminal region between mouse and rat affects the *y*-ion series and the large *b*-ion series. This result suggests that X!Tandem heavily relies on matching these ions. This is also consistent with the poor performance of the program when a random percentage of ions was available for peptide identification, especially when only 25% of the ions were available for scoring.

2.4 CONCLUSION AND FUTURE STUDIES

The present study demonstrates that although most neuropeptides and prohormone peptides with ideal MS/MS spectra can be identified using standard database search methods, a careful assessment of the accuracy of the match is still required. The present study evaluated the impact of

various factors including sequence size, characteristics (including charge state, neutral mass loss), variation (including cross species searches), spectra completeness and search specifications on accurate peptide identification. Our results indicate the need to optimize for the search for neuropeptides and small peptides. The database search methods must accurately assess the match significance irrespective of peptide length, especially for small prohormone peptides less than ten amino acids in length.

The results from the present study indicate that the correct identification of peptides based on a single threshold across all spectra is challenging even when provided with ideal spectra and target data. A major component of this challenge was the scoring and assignment of a single significance threshold for all peptides. This problem is exacerbated when analyzing experimental data because the quality of the data and the specifications of the program have a large impact on the accuracy of peptide identification. A more comprehensive simulation approach can be used to extend these results to assess the importance of other aspects of MS/MS on peptide identification [70].

The evaluations performed in this study assumed that the peptides had an ideal, uniform spectra. Different peptide ion fragmentation methods (e.g., CID, HCD, ETD) have different abilities to fragment; thus, resulting in different identification performance by database search tools [71]. Future studies will consider the impact of the fragmentation method and PTMs on the ability to identify neuropeptides.

In recent years, the identification of peptides using spectrum-to-spectrum search tools has been proposed [72]. Spectral library search strategies are a promising alternative for peptide identification, in which MS/MS spectra are directly compared against a reference library of confidently assigned spectra. A study of the effectiveness of spectrum-to-spectrum searches when applied to small prohormone peptide identification needs to be undertaken. Supporting resources include the development of a well-curated library of neuropeptide MS/MS spectra.

REFERENCES

[1] Hook, V., Funkelstein, L., Lu, D., Bark, S., Wegrzyn, J., Hwang, S. R., Proteases for processing proneuropeptides into peptide neurotransmitters and hormones. Annu. Rev. Pharmacol. Toxicol. 2008, *48*, 393-423. doi:10.1146/annurev.pharmtox.48.113006.094812.

[2] Hokfelt, T., Bartfai, T., Bloom, F., Neuropeptides: opportunities for drug discovery. Lancet Neurol. 2003, *2*, 463-472.

[3] Burbach, J. P., Neuropeptides from concept to online database www.neuropeptides.nl. Eur. J. Pharmacol. 2010, *626*, 27-48. doi:10.1016/j.ejphar.2009.10.015.

[4] Kim, Y., Bark, S., Hook, V., Bandeira, N., NeuroPedia: neuropeptide database and spectral library. Bioinformatics 2011, *27*, 2772-2773. doi:10.1093/bioinformatics/btr445.

[5] Svensson, M., Skold, K., Nilsson, A., Fälth, M., Nydahl, K., Svenningsson, P., Andrén, P. E., Neuropeptidomics: MS applied to the discovery of novel peptides from the brain. Anal. Chem. 2007, *79*, 15-6, 18-21.

[6] Tegge, A. N., Southey, B. R., Sweedler, J. V., Rodriguez-Zas, S. L., Comparative analysis of neuropeptide cleavage sites in human, mouse, rat, and cattle. Mamm. Genome 2008, *19*, 106-120. doi:10.1007/s00335-007-9090-9.

[7] Strand, F. L., Neuropeptides: Regulators of Physiological Processes. Cambridge, Mass. : MIT Press, ©1999 1999.

[8] von Eggelkraut-Gottanka, R., Beck-Sickinger, A. G., Biosynthesis of peptide hormones derived from precursor sequences. Curr. Med. Chem. 2004, *11*, 2651-2665.

[9] Fricker, L. D., Neuropeptide-processing enzymes: applications for drug discovery. AAPS J. 2005, *7*, E449-55. doi:10.1208/aapsj070244.

[10] Fricker, L. D., Lim, J., Pan, H., Che, F. Y., Peptidomics: identification and quantification of endogenous peptides in neuroendocrine tissues. Mass Spectrom. Rev. 2006, *25*, 327-344. doi:10.1002/mas.20079.

[11] Robas, N. M., Fidock, M. D., Identification of orphan G protein-coupled receptor ligands using FLIPR assays. Methods Mol. Biol. 2005, *306*, 17-26. doi:10.1385/1-59259-927-3:017.

[12] Southey, B. R., Sweedler, J. V., Rodriguez-Zas, S. L., A python analytical pipeline to identify prohormone precursors and predict prohormone cleavage sites. Front. Neuroinform 2008, *2*, 7. doi:10.3389/neuro.11.007.2008.

[13] Southey, B. R., Rodriguez-Zas, S. L., Sweedler, J. V., Prediction of neuropeptide prohormone cleavages with application to RFamides. Peptides 2006, *27*, 1087-1098. doi:10.1016/j.peptides.2005.07.026.

[14] Amare, A., Hummon, A. B., Southey, B. R., Zimmerman, T. A., Rodriguez-Zas, S. L., Sweedler, J.V., Bridging neuropeptidomics and genomics with bioinformatics: Prediction of mammalian neuropeptide prohormone processing. J. Proteome Res. 2006, *5*, 1162-1167. doi:10.1021/pr0504541.

[15] Southey, B. R., Amare, A., Zimmerman, T. A., Rodriguez-Zas, S. L., Sweedler, J. V., NeuroPred: a tool to predict cleavage sites in neuropeptide precursors and provide the masses of the resulting peptides. Nucleic Acids Res. 2006, *34*, W267-72. doi:10.1093/nar/gkl161.

[16] UniProt Consortium, The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res. 2010, *38*, D142-8. doi:10.1093/nar/gkp846.

[17] Magrane, M., Consortium, U., UniProt Knowledgebase: a hub of integrated protein data. Database (Oxford) 2011, *2011*, bar009. doi:10.1093/database/bar009.

[18] Liu, F., Baggerman, G., Schoofs, L., Wets, G., The construction of a bioactive peptide database in Metazoa. J. Proteome Res. 2008, *7*, 4119-4131. doi:10.1021/pr800037n.

[19] Hummon, A. B., Amare, A., Sweedler, J. V., Discovering new invertebrate neuropeptides using mass spectrometry. Mass Spectrom. Rev. 2006, *25*, 77-98. doi:10.1002/mas.20055.

[20] Quirion, R., Bioassays in modern peptide research. Peptides 1982, *3*, 223-230.

[21] Smith, J. A., Solution radioimmunoassay of proteins and peptides. Curr. Protoc. Mol. Biol. 2006, *Chapter 10*, Unit 10.24. doi:10.1002/0471142727.mb1024s74.

[22] Boonen, K., Landuyt, B., Baggerman, G., Husson, S. J., Huybrechts, J., Schoofs, L., Peptidomics: the integrated approach of MS, hyphenated techniques and bioinformatics for neuropeptide analysis. J. Sep. Sci. 2008, *31*, 427-445. doi:10.1002/jssc.200700450.

[23] Nesvizhskii, A. I., Protein identification by tandem mass spectrometry and sequence database searching. Methods Mol. Biol. 2007, *367*, 87-119. doi:10.1385/1-59745-275-0:87.

[24] Aebersold, R., Mann, M., Mass spectrometry-based proteomics. Nature 2003, *422*, 198-207. doi:10.1038/nature01511.

[25] Yates, J. R., Ruse, C. I., Nakorchevsky, A., Proteomics by mass spectrometry: approaches, advances, and applications. Annu. Rev. Biomed. Eng. 2009, *11*, 49-79. doi:10.1146/annurev-bioeng-061008-124934.

[26] Siuzdak, G., Mass Spectrometry for Biotechnology. Academic Press 1996.

[27] Chait, B. T., Chemistry. Mass spectrometry: bottom-up or top-down? Science 2006, *314*, 65-66. doi:10.1126/science.1133987.

[28] Resing, K. A., Ahn, N. G., Proteomics strategies for protein identification. FEBS Lett. 2005, *579*, 885-889. doi:10.1016/j.febslet.2004.12.001.

[29] Nesvizhskii, A. I., A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. J. Proteomics 2010, *73*, 2092-2123. doi:10.1016/j.jprot.2010.08.009.

[30] Nesvizhskii, A. I., Vitek, O., Aebersold, R., Analysis and validation of proteomic data generated by tandem mass spectrometry. Nat. Methods 2007, *4*, 787-797. doi:10.1038/nmeth1088.

[31] Steen, H., Mann, M., The ABC's (and XYZ's) of peptide sequencing. Nat. Rev. Mol. Cell Biol. 2004, *5*, 699-711. doi:10.1038/nrm1468.

[32] Papayannopoulos, I. A., The interpretation of collision-induced dissociation tandem mass spectra of peptides. Mass Spectrometry Reviews 1995, *14*, 49-73. doi:10.1002/mas.1280140104.

[33] Marcotte, E. M., How do shotgun proteomics algorithms identify proteins? Nat. Biotechnol. 2007, *25*, 755-757. doi:10.1038/nbt0707-755.

[34] Chen, T., Kao, M. Y., Tepel, M., Rush, J., Church, G. M., A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry. J. Comput. Biol. 2001, *8*, 325-337. doi:10.1089/10665270152530872.

[35] Dancik, V., Addona, T. A., Clauser, K. R., Vath, J. E., Pevzner, P. A., De novo peptide sequencing via tandem mass spectrometry. J. Comput. Biol. 1999, *6*, 327-342. doi:10.1089/106652799318300.

[36] Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A., Lajoie, G., PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. Rapid Commun. Mass Spectrom. 2003, *17*, 2337-2342. doi:10.1002/rcm.1196.

[37] Mann, M., Jensen, O. N., Proteomic analysis of post-translational modifications. Nat. Biotechnol. 2003, *21*, 255-261. doi:10.1038/nbt0303-255.

[38] Pevzner, P. A., Mulyukov, Z., Dancik, V., Tang, C. L., Efficiency of database search for identification of mutated and modified proteins via mass spectrometry. Genome Res. 2001, *11*, 290-299. doi:10.1101/gr.154101.

[39] Fälth, M., Sköld, K., Norrman, M., Svensson, M., Fenyö, D., Andrén, P. E., SwePep, a database designed for endogenous peptides and mass spectrometry. Mol. Cell. Proteomics 2006, *5*, 998-1005. doi:10.1074/mcp.M500401-MCP200.

[40] Fälth, M., Svensson, M., Nilsson, A., Skold, K., Fenyö, D., Andrén, P. E., Validation of endogenous peptide identifications using a database of tandem mass spectra. J. Proteome Res. 2008, *7*, 3049-3053. doi:10.1021/pr800036d.

[41] Deutsch, E. W., Tandem mass spectrometry spectral libraries and library searching. Methods Mol. Biol. 2011, *696*, 225-232. doi:10.1007/978-1-60761-987-1_13.

[42] Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L., Xu, M., Maynard, D. M., Yang, X., Shi, W., Bryant, S.H., Open mass spectrometry search algorithm. J. Proteome Res. 2004, *3*, 958-964. doi:10.1021/pr0499491.

[43] Craig, R., Beavis, R. C., TANDEM: matching proteins with tandem mass spectra. Bioinformatics 2004, *20*, 1466-1467. doi:10.1093/bioinformatics/bth092.

[44] Park, C. Y., Klammer, A. A., Kall, L., MacCoss, M. J., Noble, W. S., Rapid and accurate peptide identification from tandem mass spectra. J. Proteome Res. 2008, *7*, 3022-3027. doi:10.1021/pr800127y.

[45] Eng, J. K., McCormack, A. L., Yates, J. R., An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. 1994, *5*, 976-989. doi:10.1016/1044-0305(94)80016-2.

[46] Tabb, D. L., Fernando, C. G., Chambers, M. C., MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. J. Proteome Res. 2007, *6*, 654-661. doi:10.1021/pr0604054.

[47] Diament, B. J., Noble, W. S., Faster SEQUEST searching for peptide identification from tandem mass spectra. J. Proteome Res. 2011, *10*, 3871-3879. doi:10.1021/pr101196n.

[48] Balgley, B. M., Laudeman, T., Yang, L., Song, T., Lee, C. S., Comparative evaluation of tandem MS search algorithms using a target-decoy search strategy. Mol. Cell. Proteomics 2007, *6*, 1599-1608. doi:10.1074/mcp.M600469-MCP200.

[49] Kapp, E., Schutz, F., Overview of tandem mass spectrometry (MS/MS) database search algorithms. Curr. Protoc. Protein Sci. 2007, *Chapter 25*, Unit25.2. doi:10.1002/0471140864.ps2502s49.

[50] Sturm, M., Bertsch, A., Gröpl, C., Hildebrandt, A., Hussong, R., Lange, E., Pfeifer, N., Schulz-Trieglaff, O., Zerck, A., Reinert, K., Kohlbacher, O., OpenMS - an open-source software framework for mass spectrometry. BMC Bioinformatics 2008, *9*, 163. doi:10.1186/1471-2105-9-163.

[51] MacLean, B., Eng, J. K., Beavis, R. C., McIntosh, M., General framework for developing and evaluating database scoring algorithms using the TANDEM search engine. Bioinformatics 2006, *22*, 2830-2832. doi:10.1093/bioinformatics/btl379.

[52] Carr, S., Aebersold, R., Baldwin, M., Burlingame, A., Clauser, K., Nesvizhskii, A., The need for guidelines in publication of peptide and protein identification data: Working Group on Publication Guidelines for Peptide and Protein Identification Data. Mol. Cell. Proteomics 2004, *3*, 531-533. doi:10.1074/mcp.T400006-MCP200.

[53] Zhang, J., Yip, H., Katta, V., Identification of isomerization and racemization of aspartate in the Asp-Asp motifs of a therapeutic protein. Anal. Biochem. 2011, *410*, 234-243. doi:10.1016/j.ab.2010.11.040.

[54] Kapp, E. A., Schutz, F., Connolly, L. M., Chakel, J. A., Meza, J. E., Miller, C. A., Fenyo, D., Eng, J. K., Adkins, J. N., Omenn, G. S., Simpson, R. J., An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis. Proteomics 2005, *5*, 3475-3490. doi:10.1002/pmic.200500126.

[55] Yadav, A. K., Kumar, D., Dash, D., MassWiz: a novel scoring algorithm with target-decoy based analysis pipeline for tandem mass spectrometry. J. Proteome Res. 2011, *10*, 2154-2160. doi:10.1021/pr200031z.

[56] Kandasamy, K., Pandey, A., Molina, H., Evaluation of several MS/MS search algorithms for analysis of spectra derived from electron transfer dissociation experiments. Anal. Chem. 2009, *81*, 7170-7180. doi:10.1021/ac9006107.

[57] Li, W., Ji, L., Goya, J., Tan, G., and Wysocki, V. H., SQID: An Intensity-Incorporated Protein Identification Algorithm for Tandem Mass Spectrometry. 2011, *10*, 1593-1602.

[58] Svensson, M., Skold, K., Svenningsson, P., Andren, P. E., Peptidomics-based discovery of novel neuropeptides. J. Proteome Res. 2003, *2*, 213-219.

[59] Xu, H., Freitas, M. A., MassMatrix: a database search program for rapid characterization of proteins and peptides from tandem mass spectrometry data. Proteomics 2009, *9*, 1548-1555. doi:10.1002/pmic.200700322.

[60] Yin, P., Hou, X., Romanova, E. V., Sweedler, J. V., Neuropeptidomics: mass spectrometry-based qualitative and quantitative analysis. Methods Mol. Biol. 2011, *789*, 223-236. doi:10.1007/978-1-61779-310-3_14.

[61] Lee, J. E., Atkins, N.,Jr, Hatcher, N. G., Zamdborg, L., Gillette, M. U., Sweedler, J. V., Kelleher, N. L., Endogenous peptide discovery of the rat circadian clock: a focused study of the suprachiasmatic nucleus by ultrahigh performance tandem mass spectrometry. Mol. Cell. Proteomics 2010, *9*, 285-297. doi:10.1074/mcp.M900362-MCP200.

[62] Sadygov, R. G., Cociorva, D., Yates, J. R.,3rd, Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. Nat. Methods 2004, *1*, 195-202. doi:10.1038/nmeth725.

[63] Xu, H., Freitas, M. A., A dynamic noise level algorithm for spectral screening of peptide MS/MS spectra. BMC Bioinformatics 2010, *11*, 436. doi:10.1186/1471-2105-11-436.

[64] Cox, J., Mann, M., MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. 2008, *26*, 1367-1372. doi:10.1038/nbt.1511.

[65] Frank, A. M., Savitski, M. M., Nielsen, M. L., Zubarev, R. A., Pevzner, P. A., De novo peptide sequencing and identification with precision mass spectrometry. J. Proteome Res. 2007, *6*, 114-123. doi:10.1021/pr060271u.

[66] Houel, S., Abernathy, R., Renganathan, K., Meyer-Arendt, K., Ahn, N. G., Old, W. M., Quantifying the impact of chimera MS/MS spectra on peptide identification in large-scale proteomics studies. J. Proteome Res. 2010, *9*, 4152-4160. doi:10.1021/pr1003856.

[67] Colaert, N., Degroeve, S., Helsens, K., Martens, L., Analysis of the resolution limitations of peptide identification algorithms. J. Proteome Res. 2011, *10*, 5555-5561. doi:10.1021/pr200913a.

[68] Gatlin, C. L., Eng, J. K., Cross, S. T., Detter, J. C., Yates, J. R.,3rd, Automated identification of amino acid sequence variations in proteins by HPLC/microspray tandem mass spectrometry. Anal. Chem. 2000, *72*, 757-763.

[69] Creasy, D. M., Cottrell, J. S., Error tolerant searching of uninterpreted tandem mass spectrometry data. Proteomics 2002, *2*, 1426-1434. doi:2-5.

[70] Bielow, C., Aiche, S., Andreotti, S., Reinert, K., MSSimulator: Simulation of mass spectrometry data. J. Proteome Res. 2011, *10*, 2922-2929. doi:10.1021/pr200155f.

[71] Shen, Y., Tolic, N., Xie, F., Zhao, R., Purvine, S.O., Schepmoes, A. A., Moore, R. J., Anderson, G. A., Smith, R. D., Effectiveness of CID, HCD, and ETD with FT MS/MS for degradomic-peptidomic analysis: comparison of peptide identification methods. J. Proteome Res. 2011, *10*, 3929-3943. doi:10.1021/pr200052c.

[72] Yen, C. Y., Houel, S., Ahn, N. G., Old, W. M., Spectrum-to-spectrum searching using a proteome-wide spectral library. Mol. Cell. Proteomics 2011, *10*, M111.007666. doi:10.1074/mcp.M111.007666.

[73] Clynen, E., Liu, F., Husson, S. J., Landuyt, B., Hayakawa, E., Baqqerman, G., Wets, G., Schoofs, L., Bioinformatic approaches to the identification of novel neuropeptide precursors. Methods Mol. Biol. 2010, *615*, 357-374. doi:10.1007/978-1-60761-535-4_25.

[74] PEAKS., http://www.bioinformaticssolutions.com/peaks/downloads/masstable.html. 2012, *2012*, 1.

FIGURES AND TABLES

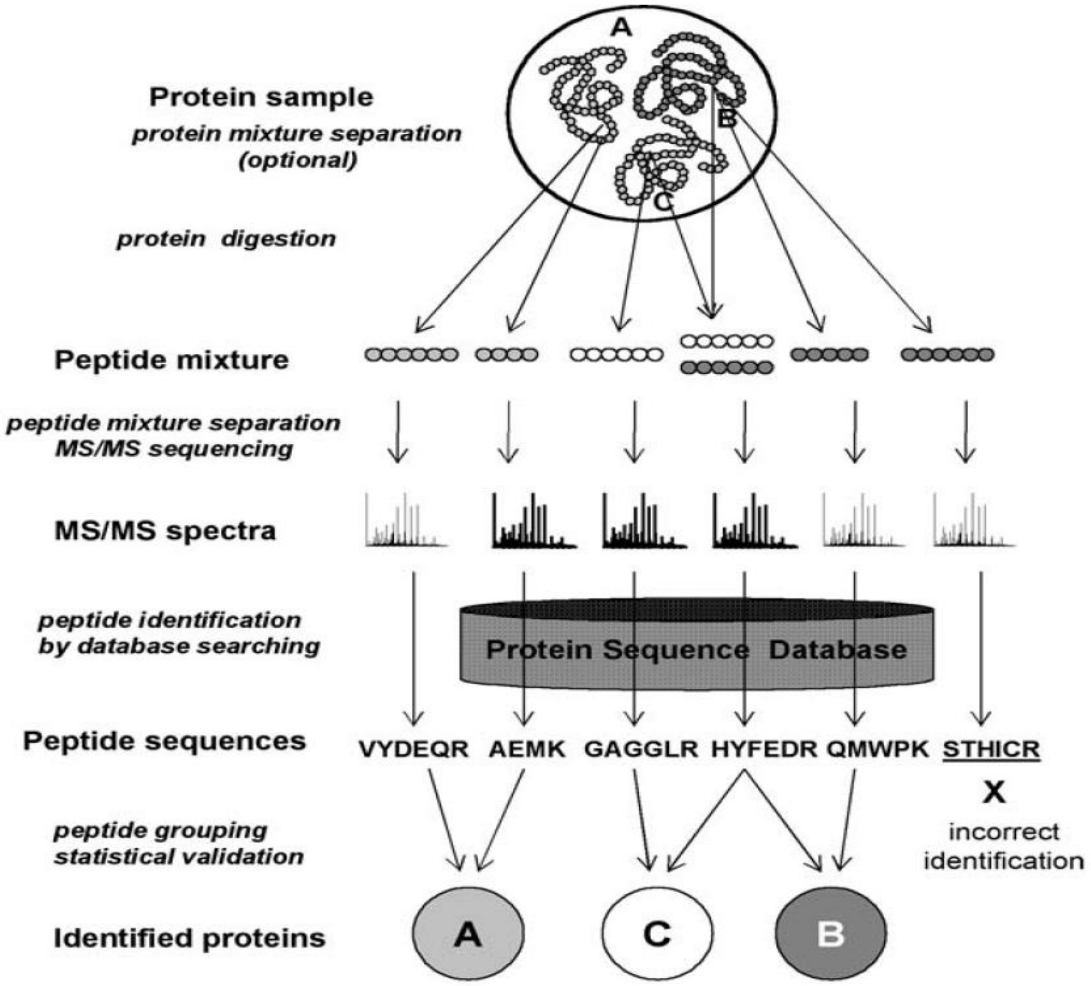

Figure 1. General view of the experimental steps and flow of the data in shotgun proteomics analysis. Sample proteins are first proteolytically cleaved into peptides. After separation using one- or multidimensional chromatography, peptides are ionized and selected ions are fragmented to produce signature tandem mass spectrometry (MS/MS) spectra. Peptides are identified from MS/MS spectra using automated database search programs. Peptide assignments are then statistically validated and incorrect identifications filtered out (peptide STHICR). Sequences of the identified peptides are used to infer which proteins are present in the original sample. Some peptides are present in more than one protein (peptide HYFEDR), which can complicate the protein inference process[1].

[1] Springer and the Methods in Molecular Biology, 367, 2007, 87-119, Protein identification by tandem mass spectrometry and sequence database searching, Nesvizhskii AI, 1; with kind permission from Springer Science and Business Media.

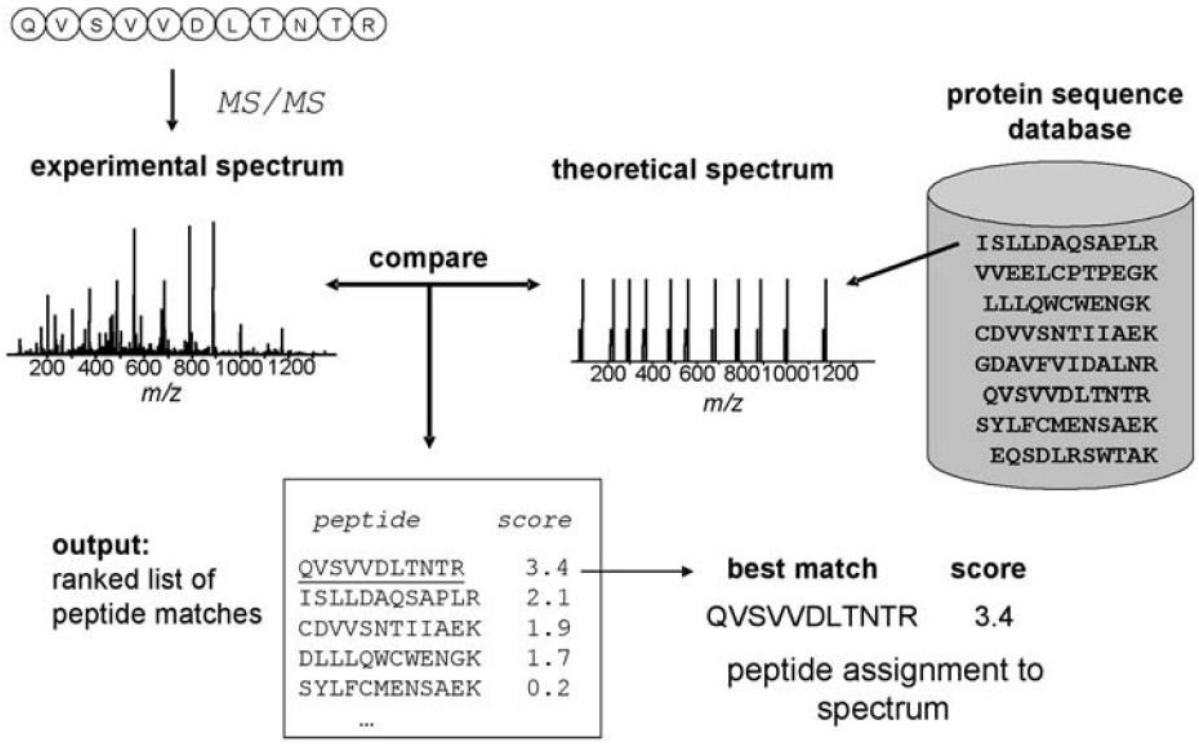

Figure 2. Tandem mass spectrometry (MS/MS) database searching. Acquired MS/MS spectra are correlated against theoretical spectra constructed for each database peptide that satisfies a certain set of database search parameters specified by the user. A scoring scheme is used to measure the degree of similarity between the spectra. Candidate peptides are ranked according to the computed score, and the highest scoring peptide sequence (best match) is selected for further analysis[2].

---

[2]Springer and the Methods in Molecular Biology, 367, 2007, 87-119, Protein identification by tandem mass spectrometry and sequence database searching, Nesvizhskii AI, 3; with kind permission from Springer Science and Business Media.

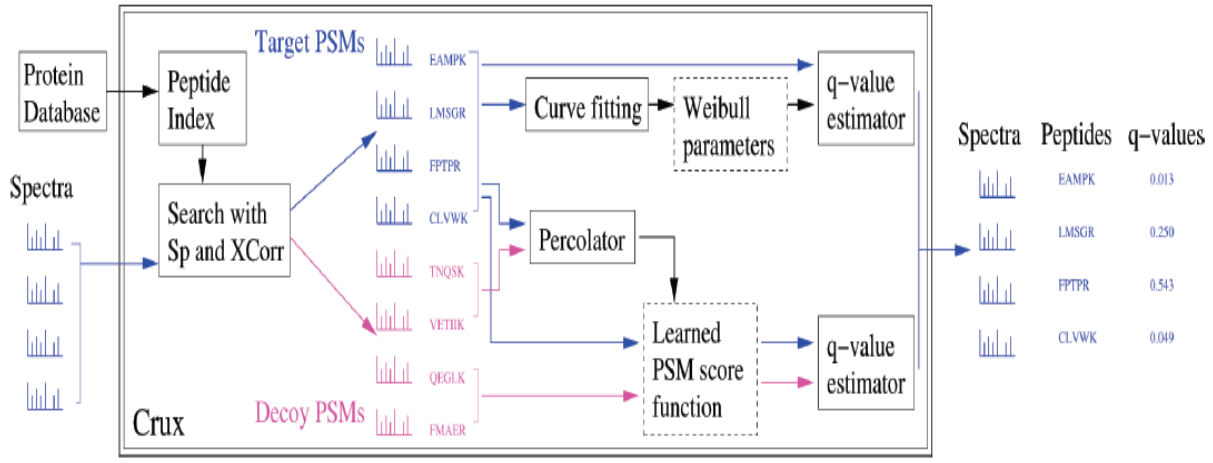

Figure 3. The Crux algorithm[3].

[3] Adapted with permission from (Park, C. Y., Klammer, A. A., Kall, L., MacCoss, M. J., Noble, W. S., Rapid and accurate peptide identification from tandem mass spectra. J. Proteome Res. 2008, *7*, 3022-3027. doi:10.1021/pr800127y). Copyright (2008) American Chemical Society.

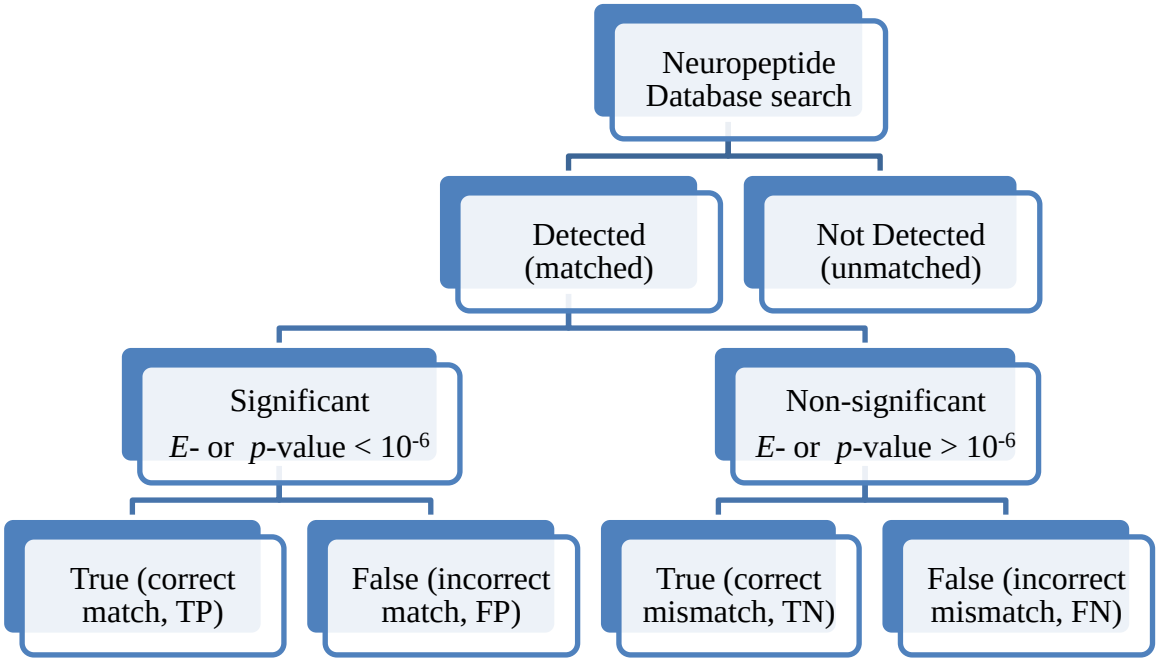

Figure 4. Decision tree depicting the flow of criteria used to evaluate the performance of the three tandem mass spectrometry database search programs.

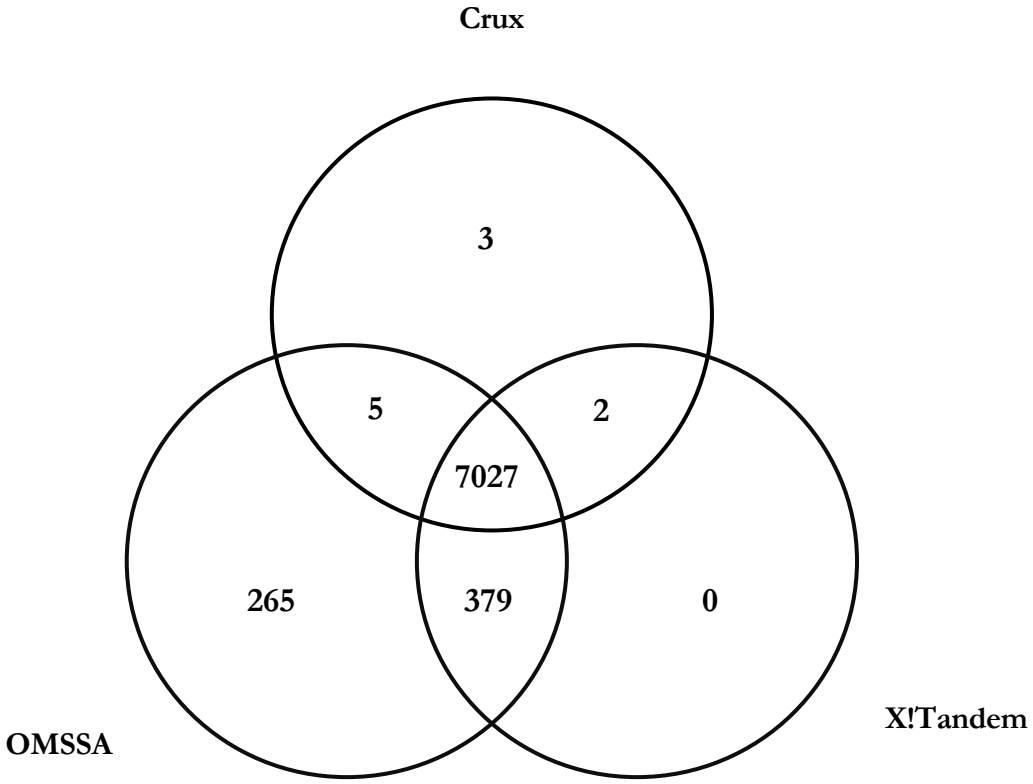

Figure 5. Venn diagram depicting the common and distinct true positive peptides identified from the three database search programs, X!Tandem, OMSSA, and Crux using all ion information and peptide charge state 3.

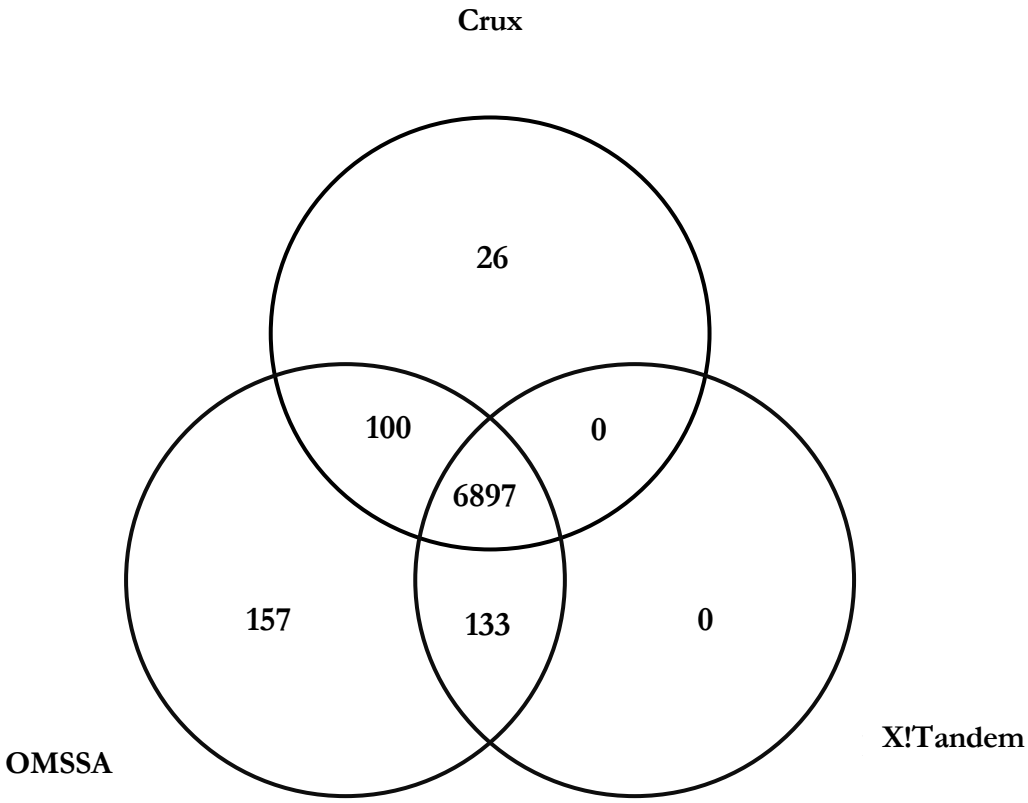

Figure 6. Venn diagram depicting the common and distinct peptides identified by all three programs (X!Tandem, OMSSA, and Crux) using only *y*-ion series information and peptide charge state 3.

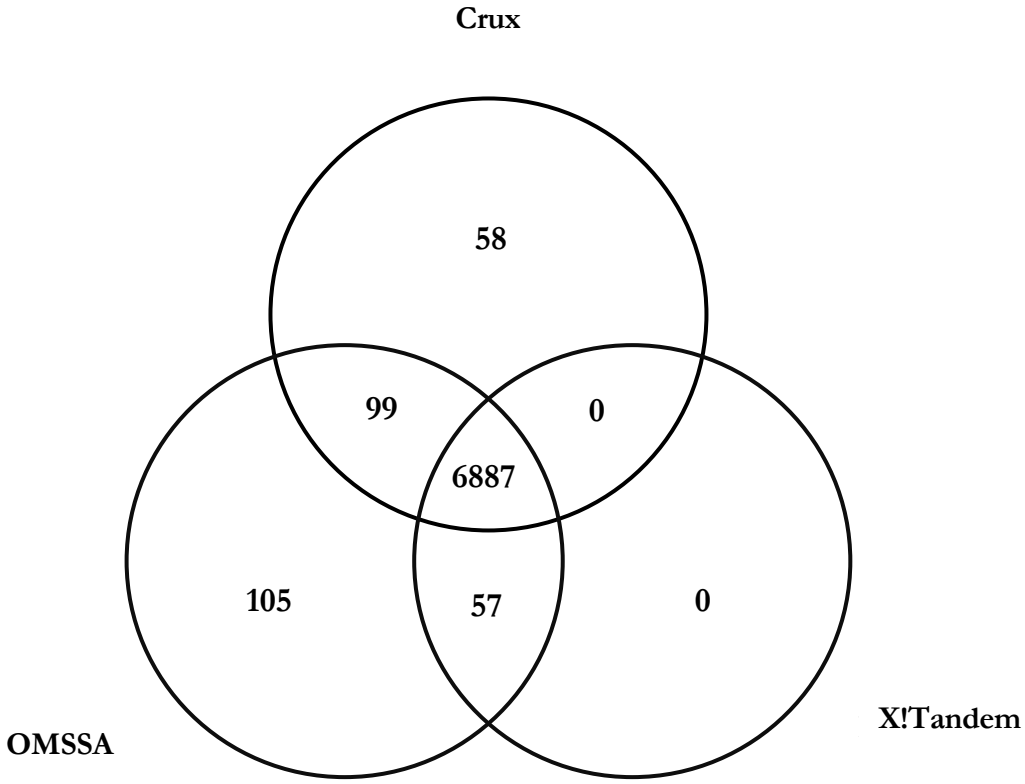

Figure 7. Venn diagram depicting the common and distinct peptides identified by all three database search programs using only *b*-ion series information and peptide charge state 3.

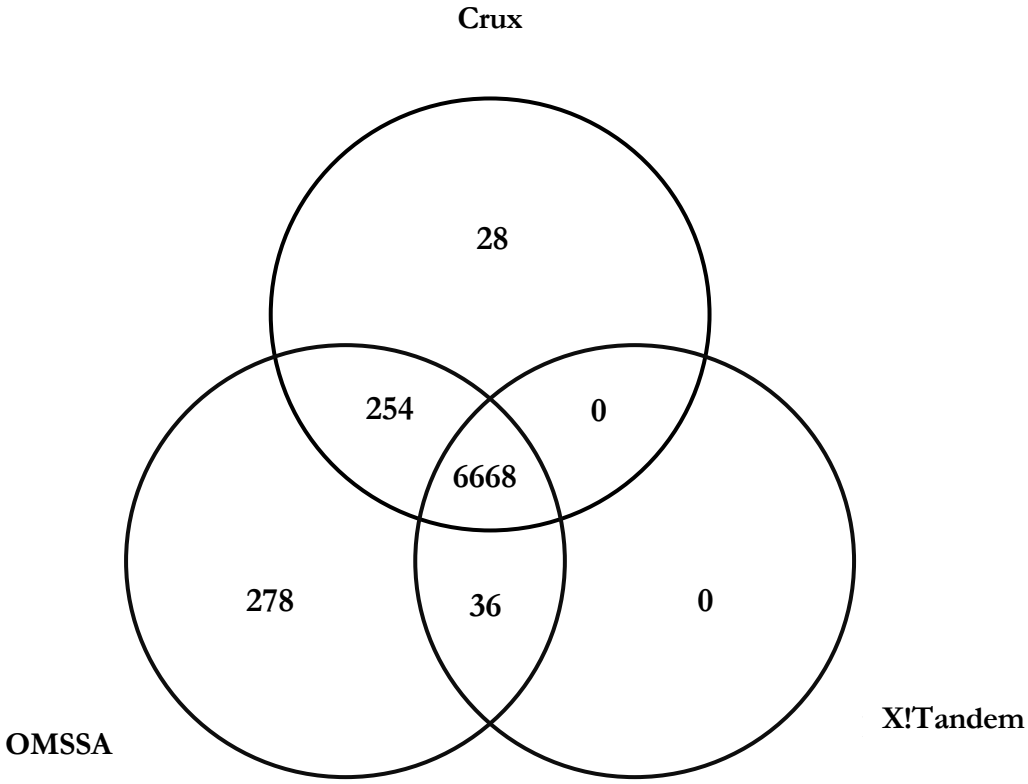

Figure 8. Venn diagram depicting the common and distinct peptides identified by all three database search programs using only 50% of all ion information and peptide charge state 3.

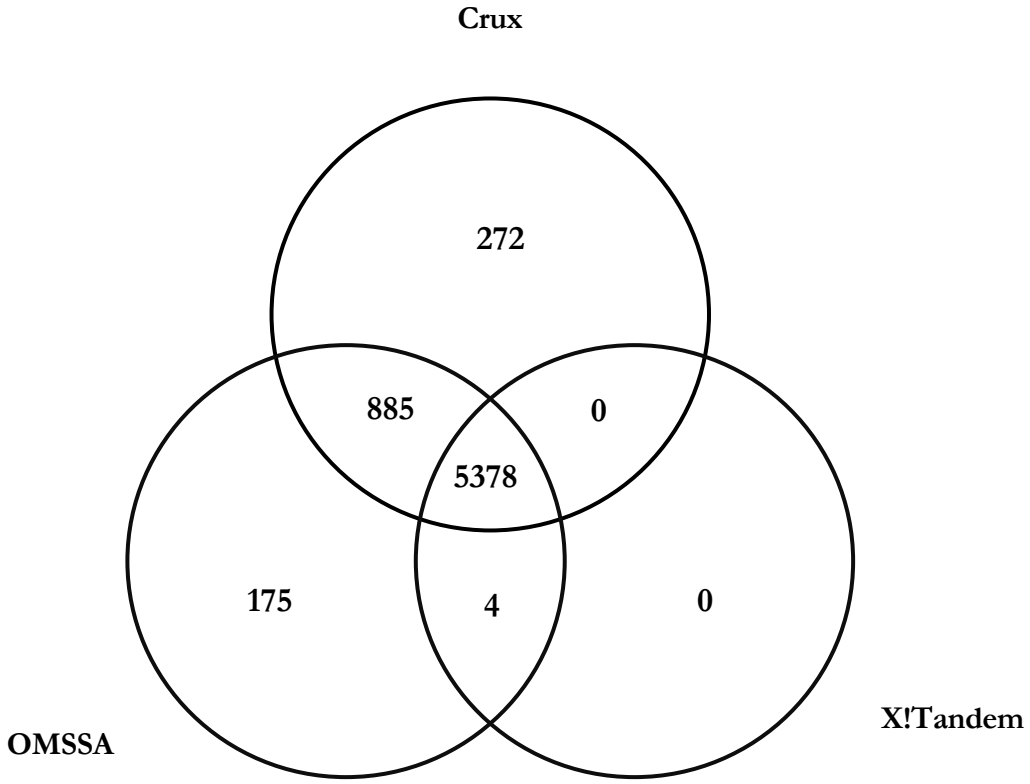

Figure 9. Venn diagram depicting the common and distinct peptides identified by all three database search programs using only 25% of all ion information and peptide charge state 3.

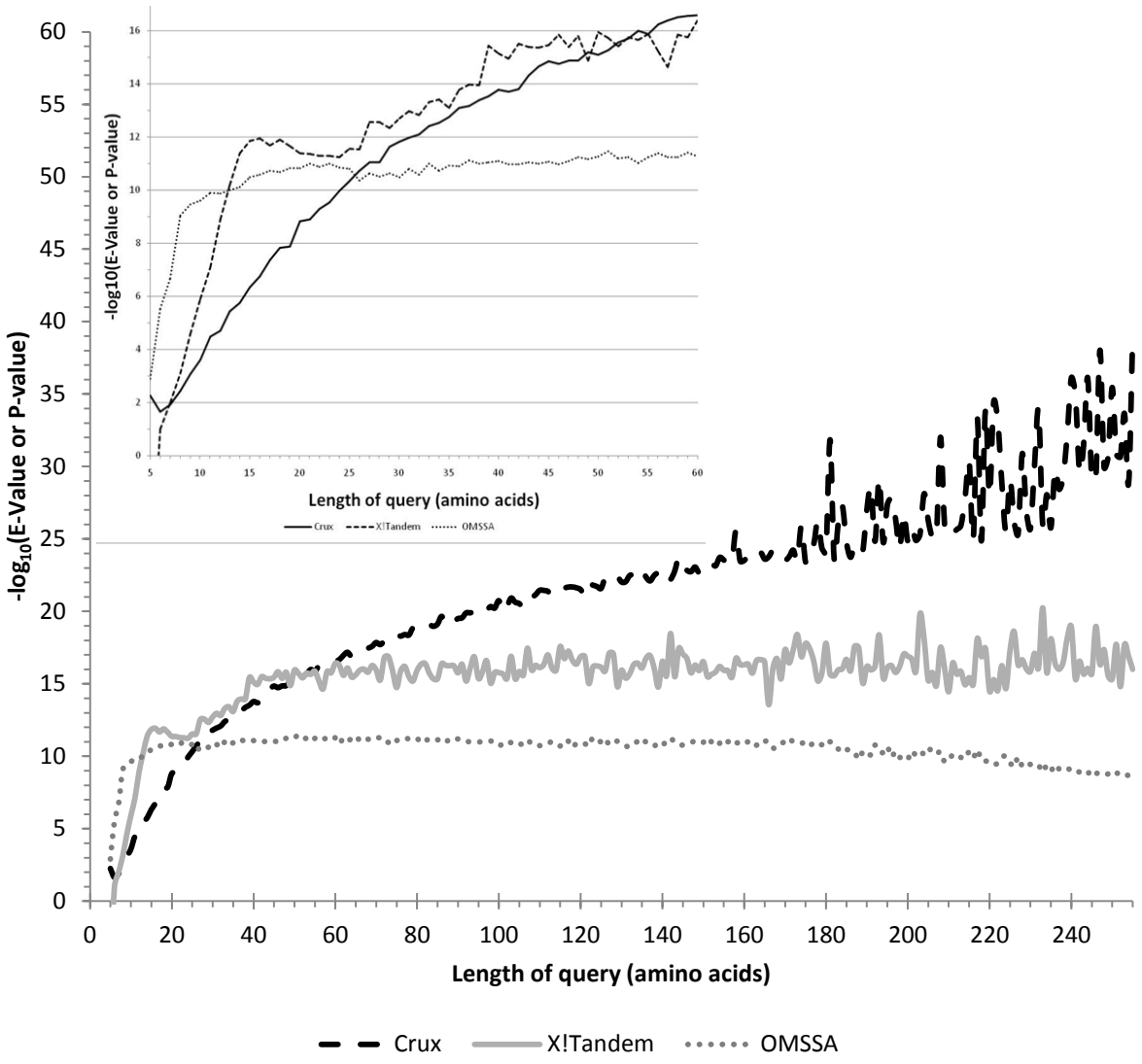

Figure 10. Comparison of OMSSA, Crux, and X!Tandem $\log_{10}$(*E*- or *p*-values) averaged across peptide length and precursor charge states for all peptides (main plot) and magnified for peptides up to 60 amino acids in length (insert).

Table 1. Masses of different ion types[4].

Ion type	Mass
a	σ[a)] - 26.9871
b	σ + 1.0078
c	σ + 18.0344
x	σ + 44.9977
y	σ + 19.0184
neutral loss (nh3 or h2o)	*b* - 17 or *y* - 17 or *b* - 18 or *y* - 18

a) Total residue mass of ion.

[4] http://www.bioinformaticssolutions.com/peaks/downloads/masstable.html

Table 2. Summary of the mouse peptides used to simulate the query spectra and of the mouse and rat peptides used to populate the search database.

	Mouse	Rat
Number of prohormones	92	90
Number of peptides	7850	7647
Average (min, max) number of peptides/prohormones	74.06 (1, 1139)	76.47 (1, 1172)
Average (min, max) peptide size (amino acids)	75.23 (5, 255)	76.00 (5, 255)
Percentage of peptides from UniProt	3.35	3.45
Percentage of peptides not from UniProt	96.65	96.55
Percentage of peptides less than 10 amino acids in length	5.45	5.07
Percentage of peptides less than 20 amino acids in length	16.60	15.72
Percentage of peptides less than 30 amino acids in length	26.52	25.77

Table 3. Comparison of peptide detection among database search programs.

Correctly Matched		X!Tandem+OMSSA+Crux						OMSSA+Crux				X!Tandem+Crux			Crux
Scenario[b)]	Charge	All[a)]	OC	OX	O	C	N	OC	O	C	N	XC	C	N	N
All	+1	7028	8	378	327	0	23	0	0	1	84	1	0	0	0
	+2	7012	7	397	313	0	35	0	0	0	85	1	0	0	0
	+3	7027	5	379	265	0	87	0	0	3	82	2	0	0	0
b + *y* ions	+1	6874	5	503	339	0	3	41	0	1	84	0	0	0	0
	+2	6888	5	485	340	0	3	44	0	0	85	0	0	0	0
	+3	6978	3	389	337	0	8	50	0	1	84	0	0	0	0
b ions	+1	6837	109	105	184	46	484	0	0	1	84	0	0	0	0
	+2	6831	99	109	175	60	491	0	0	0	85	0	0	0	0
	+3	6887	99	57	105	57	560	0	0	1	84	0	0	0	0
y ions	+1	6911	126	118	221	17	370	2	0	1	84	0	0	0	0
	+2	6905	116	133	202	11	397	1	0	1	84	0	0	0	0
	+3	6897	99	133	157	24	454	1	0	2	83	0	0	0	0
R50	+1	6646	230	69	394	14	410	1	0	0	85	1	0	0	0
	+2	6638	244	69	382	10	421	1	0	2	83	0	0	0	0
	+3	6668	254	36	278	27	502	0	0	1	84	0	0	0	0
R25	+1	5370	661	21	156	318	989	198	4	8	52	0	1	29	43
	+2	5358	673	25	160	295	987	200	4	6	82	0	0	30	30
	+3	5378	675	4	170	267	1010	210	5	5	63	0	0	22	41

a):
All =OMSSA, X!Tandem and Crux *E*- or *p*-value < 1×10^{-6}.
OC = only OMSSA and Crux *E*- or *p*-value < 1×10^{-6}.
OX= only OMSSA and X!Tandem *E*-value < 1×10^{-6}.
XC= only X!Tandem and Crux *E*- or *p*-value < 1×10^{-6}.
O= only OMSSA *E*-value < 1×10^{-6}.
C= only Crux *p*-value < 1×10^{-6}.
N= No program *E*- or *p*-value < 1×10^{-6}.
b):
All = Match using all *b*- and *y*-ion series including neutral mass losses.
b + *y* ions = Match using all *b*- and *y*-ion series excluding neutral mass losses.
b ions = Match only using the *b*-ion series including neutral mass losses.
y ions = Match only using the *y*-ion series including neutral mass losses.
R50 = Match only using random 50% of all ions including neutral mass losses.
R25 = Match only using random 25% of all ions including neutral mass losses.

Table 4. Performance of the three programs in the identification of peptides with precursor ion charge states +1, +2, and +3 when all ions from both series are available including neutral mass losses.

Significance[a]	OMSSA			X!Tandem			Crux		
	+1[e]	+2	+3	+1	+2	+3	+1	+2	+3
Unmatched[b]	1	1	1	69	67	63	0	0	0
Mismatch[c]	0	0	1	16	18	22	0	0	0
0	0	0	0	4	6	5	1	5	1
1	1	2	4	73	67	72	118	131	109
2	11	19	42	91	88	91	214	209	237
3	48	52	34	82	79	82	171	165	146
4	24	13	33	33	41	33	160	150	151
5	24	34	59	75	74	74	151	172	170
6	49	57	54	91	85	89	172	170	170
7	73	58	40	83	78	77	171	179	182
8	28	27	336	47	59	51	194	179	189
>=9	7591	7587	7246	7186	7188	7191	6498	6490	6495
Prop >6[d]	98.6%	98.5%	97.8%	94.4%	94.4%	94.4%	89.6%	89.4%	89.6%

a) Significance threshold (t) for matched to be considered significant at *E*- or *p*-value < 1×10^{-t}.
b) Unmatched: the program does not provide a match with the program setting.
c) Mismatched: the program provided an incorrect match.
d) Percentage of the matches that have *E*- or *p*-value < 1×10^{-6}.
e) Peptide charge state.

Table 5. Performance of the three programs in the identification of peptides with precursor ion charge states +1, +2, and +3 when all ions from both series are available excluding neutral mass losses.

Significance[a]	OMSSA			X!Tandem			Crux		
	+1[e]	+2	+3	+1	+2	+3	+1	+2	+3
Unmatched[b]	0	0	0	115	116	118	0	0	0
Mismatched[c]	0	0	0	11	13	17	0	0	0
0	0	0	0	2	4	3	1	5	2
1	0	1	1	73	70	72	129	153	124
2	1	0	2	93	90	93	236	228	233
3	2	5	6	80	81	80	226	193	171
4	10	30	78	30	32	33	178	175	140
5	75	52	6	69	71	67	170	163	151
6	4	7	4	95	90	94	172	186	159
7	5	3	63	74	76	74	213	190	198
8	63	68	79	13	17	13	200	245	207
>=9	7690	7684	7611	7195	7190	7186	6325	6312	6465
Prop >6[d]	98.9%	98.9%	98.8%	94.0%	93.9%	93.8%	88.0%	88.3%	89.5%

a) Significance threshold (t) for matched to be considered significant at *E*- or *p*-value < 1×10^{-t}.
b) Unmatched: the program does not provide a match with the program setting.
c) Mismatched: the program provided an incorrect match.
d) Percentage of the matches that have *E*- or *p*-value < 1×10^{-6}.
e) Peptide charge state.

Table 6. Performance of the three programs in the identification of peptides with precursor charge states +1, +2, and +3 when only the *b*-ion series is available including neutral mass losses.

Significance[a]	OMSSA			X!Tandem			Crux		
	+1[e]	+2	+3	+1	+2	+3	+1	+2	+3
Unmatched[b]	0	0	0	79	76	75	0	0	0
Mismatched[c]	0	0	0	6	9	10	0	0	0
0	160	179	234	237	237	240	0	4	2
1	84	94	100	109	110	107	93	99	105
2	87	82	89	149	147	145	229	249	214
3	100	100	112	122	123	125	215	175	175
4	94	92	83	96	97	97	154	161	157
5	90	89	84	105	105	104	167	173	153
6	64	57	119	137	137	134	188	195	166
7	94	111	83	104	101	104	167	160	195
8	93	75	95	89	93	90	168	170	220
>=9	6984	6971	6851	6617	6615	6619	6469	6464	6463
Prop >6[d]	92.2%	91.9%	91.1%	88.5%	88.5%	88.5%	89.1%	89.0%	89.7%

a) Significance threshold (t) for matched to be considered significant at *E*- or *p*-value < 1×10^{-t}.
b) Unmatched: the program does not provide a match with the program setting.
c) Mismatched: the program provided an incorrect match.
d) Percentage of the matches that have *E*- or *p*-value < 1×10^{-6}.
e) Peptide charge state.

Table 7. Performance of the three programs in the identification of peptides with precursor charge states +1, +2, and +3 when only the *y*-ion series is available including neutral mass losses.

Significance[a]	OMSSA			X!Tandem			Crux		
	+1[e]	+2	+3	+1	+2	+3	+1	+2	+3
Unmatched[b]	0	0	0	72	69	66	0	0	0
Mismatched[c]	0	0	1	15	17	20	0	0	0
0	48	55	95	138	135	138	2	8	7
1	62	76	96	113	109	112	131	140	135
2	86	86	104	156	161	155	196	190	186
3	99	99	95	113	112	116	155	171	161
4	88	86	98	98	94	99	169	149	171
5	89	91	74	109	108	106	140	158	167
6	77	79	90	139	143	138	173	157	193
7	73	70	96	105	109	106	196	185	228
8	90	99	110	103	103	104	226	226	257
>=9	7138	7109	6991	6689	6690	6690	6462	6466	6345
Prop >6[b]	94.0%	93.7%	92.8%	89.6%	89.7%	89.7%	89.9%	89.6%	89.5%

a) Significance threshold (t) for matched to be considered significant at *E*- or *p*-value < 1×10^{-t}.
b) Unmatched: the program does not provide a match with the program setting.
c) Mismatched: the program provided an incorrect match.
d) Percentage of the matches that have *E*- or *p*-value < 1×10^{-6}.
e) Peptide charge state.

Table 8. Performance of the three programs in the identification of peptides with precursor charge states +1, +2, and +3 when only random 50% of all ions are available including neutral mass losses.

Significance[a]	OMSSA			X!Tandem			Crux		
	+1[e]	+2	+3	+1	+2	+3	+1	+2	+3
Unmatched[b]	1	0	0	73	75	66	0	0	0
Mismatched[c]	0	0	0	13	11	19	0	0	0
0	71	88	136	316	296	313	9	10	11
1	72	77	86	170	166	159	151	169	167
2	85	69	98	133	147	145	243	241	208
3	87	91	101	166	181	172	188	188	187
4	106	99	96	104	107	109	180	151	148
5	88	92	97	157	157	160	188	196	179
6	81	95	106	109	96	95	190	198	213
7	86	83	89	122	144	139	209	208	244
8	74	92	92	127	117	111	244	230	235
>=9	7099	7064	6949	6360	6353	6362	6248	6259	6258
Prop >6[d]	93.5%	93.4%	92.2%	85.6%	85.5%	85.4%	87.8%	87.8%	88.5%

a) Significance threshold (t) for matched to be considered significant at *E*- or *p*-value < 1×10^{-t}.
b) Unmatched: the program does not provide a match with the program setting.
c) Mismatched: the program provided an incorrect match.
d) Percentage of the matches that have *E*- or *p*-value < 1×10^{-6}.
e) Peptide charge state.

Table 9. Performance of the three programs in the identification of peptides with precursor charge states +1, +2, and +3 when random 25% of all ions are available including neutral mass losses.

Significance[a]	OMSSA			X!Tandem			Crux		
	+1[e]	+2	+3	+1	+2	+3	+1	+2	+3
Unmatched[b]	73	60	63	295	312	311	0	0	0
Mismatched[c]	4	0	1	10	10	13	0	0	0
0	492	512	506	1133	1113	1115	60	71	80
1	182	177	170	284	292	280	322	322	298
2	184	194	189	228	228	250	302	306	244
3	178	188	169	229	227	225	218	226	183
4	160	143	171	140	128	131	183	182	236
5	167	156	139	136	153	139	209	211	274
6	133	147	154	120	110	130	273	236	285
7	113	124	124	146	130	132	326	318	312
8	106	100	125	131	126	112	365	378	434
>=9	6058	6049	6039	4998	5021	5012	5592	5600	5504
Prop >6[d]	81.7%	81.8%	82.1%	68.7%	68.6%	68.6%	83.5%	83.2%	83.2%

a) Significance threshold (t) for matched to be considered significant at *E*- or *p*-value < 1×10^{-t}.
b) Unmatched: the program does not provide a match with the program setting.
c) Mismatched: the program provided an incorrect match.
d) Percentage of the matches that have *E*- or *p*-value < 1×10^{-6}.
e) Peptide charge state.

Table 10. Performance of OMSSA and X!Tandem by ion series scored for precursor charge states +1, +2, and +3.

Correctly matched[a)]	Significance Threshold[b)]	*b*-ion series scored			*y*-ion series scored		
		+1[c)]	+2	+3	+1	+2	+3
Both	Both	4003	3974	3980	4459	4288	4270
Both	OMSSA	2919	2926	2791	2615	2760	2678
Both	X!Tandem	0	0	3	0	0	1
Both	None	438	460	587	331	357	443
X!Tandem	X!Tandem	177	186	184	156	147	151
X!Tandem	None	228	219	220	204	213	222
OMSSA	None	64	64	64	69	69	69
None	None	21	21	21	16	16	16

a) Both: OMSSA and X!Tandem both correctly identified the peptide; OMSSA: only OMSSA correctly identified the peptide; X!Tandem: only X!Tandem correctly identified the peptide; None: Neither OMSSA and X!Tandem correctly identified the peptide.
b) Both: OMSSA and X!Tandem *E*-values were both $< 1 \times 10^{-6}$; OMSSA: only OMSSA *E*-value was $< 1 \times 10^{-6}$; X!Tandem: only X!Tandem *E*-value was $< 1 \times 10^{-6}$; None: Neither OMSSA and X!Tandem *E*-value was $< 1 \times 10^{-6}$.
c) Peptide charge state.

Table 11. Performance of OMSSA and X!Tandem across match significance levels and precursor charge states when the *b*-ion series is scored.

Significance[a]	OMSSA			X!Tandem		
	+1[e]	+2	+3	+1	+2	+3
Unmatched[b]	415	396	373	75	75	71
Mismatched[c]	11	30	54	10	10	14
0	122	141	171	248	249	249
1	63	64	89	113	113	115
2	76	65	80	187	195	190
3	76	89	106	270	292	298
4	103	95	88	746	851	854
5	62	70	118	1902	1800	1788
6	102	118	110	1536	1471	1471
7	108	105	99	842	868	878
8	116	95	114	690	699	691
>=9	6596	6582	6448	1231	1227	1231
Prop >6[d]	88.2%	87.9%	86.3%	54.8%	54.3%	54.4%

a) Significance threshold (t) for matched to be considered significant at *E*- or *p*-value < 1×10^{-t}.
b) Unmatched: the program does not provide a match with the program setting.
c) Mismatched: the program provided an incorrect match.
d) Percentage of the matches that have *E*- or *p*-value < 1×10^{-6}.
e) Peptide charge state.

Table 12. Performance of OMSSA and X!Tandem across match significance levels and precursor charge states when the *y*-ion series is scored.

Significance[a]	OMSSA			X!Tandem		
	+1[e]	+2	+3	+1	+2	+3
Unmatched[b]	365	361	355	74	70	70
Mismatched[c]	11	15	34	11	15	15
0	47	60	79	151	145	152
1	50	51	69	108	108	110
2	69	67	77	179	186	183
3	66	73	92	214	234	241
4	87	86	103	591	688	693
5	81	89	93	1785	1857	1853
6	90	84	113	1948	1794	1786
7	80	95	118	861	841	830
8	126	122	97	648	613	620
>=9	6778	6747	6620	1280	1299	1297
Prop >6[d]	90.1%	89.8%	88.5%	60.3%	57.9%	57.7%

a) Significance threshold (t) for matched to be considered significant at *E*- or *p*-value < 1 x 10^{-t}.
b) Unmatched: the program does not provide a match with the program setting.
c) Mismatched: the program provided an incorrect match.
d) Percentage of the matches that have *E*- or *p*-value < 1 x 10^{-6}.
e) Peptide charge state.

Table 13. Performance of X!Tandem, OMSSA and Crux in the number of spectra and percentage of peptides identified from chimera spectra with precursor charge state +1 with all ions are available and including neutral mass losses.

		Number of peptides correctly matched in a spectra with an *E*- or *p*-value $< 1 \times 10^{-2}$						Percentage of peptides detected	
Program	N pep[a)]	0	1	2	3	4	5	>2[b)]	>6[c)]
OMSSA	2	11	213	580				85.4	84.1
	3	3	25	64	34			67.5	61.9
	4	1	3	5	2	1		47.9	33.3
	5	0	0	1	0	1	1	73.3	66.7
	Total	15	241	650	36	2	1	81.1	78.7
X!Tandem	2	0	799	5				50.3	12.8
	3	59	67	0	0			17.7	0.5
	4	11	1	0	0	0		2.1	0.0
	5	3	0	0	0	0	0	0.0	0.0
	Total	73	867	5	0	0	0	42.8	10.2
Crux	2	0	10	794				99.4	81.3
	3	0	0	3	123			99.2	61.6
	4	0	0	0	0	12		100.0	20.8
	5	0	0	0	0	0	3	100.0	13.3
	Total	0	10	797	123	12	3	99.4	75.8

a) Number of peptides simulated in a spectra.
b) Percentage of correctly matched peptides with an *E*- or *p*-value $< 1 \times 10^{-2}$.
c) Percentage of correctly matched peptides with an *E*- or *p*-value $< 1 \times 10^{-6}$.