UNIVERSITY OF CALIFORNIA, SAN DIEGO

# NeuroPedia: Neuropeptide database and spectral library

A thesis submitted in partial satisfaction of the requirements

for the degree Master of Science

in

Electrical Engineering (Computer Engineering)

by

Yoona Kim

Committee in Charge:

Professor Nuno Bandeir, Chair
Professor Robert Hecht-Nielsen, Co- Chair
Professor William Hodgkiss
Professor Young-Han Kim

2011

The Thesis of Yoona Kim is approved and it is acceptable in quality and form for publication on microfilm and electronically:

Co- Chair

Chair

University of California, San Diego

2011

TABLE OF CONTENTS

# LIST OF ABBRIVIATIONS

| | |
|---|---|
| CID | Collision-induced dissociation |
| ESI | Electrospray ionization |
| FDR | False discovery rate |
| HQ | High quality |
| IT | Ion trap |
| LQ | Low quality |
| LTQ | Linear Trap Quadrupole |
| MALDI | Matrix-assisted laser desorption/ionization |
| *m/z* | mass-to-charge ratio |
| MS | Mass spectrometry |
| MS/MS | Tandem mass spectrometry |
| MS1 | Single-stage mass spectrometry |
| MS2 | tandem mass spectrometry |
| NCBI | National Center for Biotechnology Information |
| NIST | National Institute of Standards and Technology |
| PSM | Peptide to spectrum match |
| QTOF | Quadrupole time-of-flight |
| UniProt | Universal Protein Resource |

LIST OF FIGURES

LIST OF TABLES

# ACKNOWLEDGEMENTS

Foremost, I would like to gratefully acknowledge Professor Nuno Bandeir for his support as the chair of my committee. His continuing guidance, patience, enthusiasm, and encouragement helped me in all the research and writing of thesis. I could not have imagined having a better advisor and mentor for my MS study. Also I gave special thanks to Professor Robert Hecht-Nielsen for his support as the co- chair of my committee. His all advice, and his immense knowledge motivated me a lot to study in Bioinformatics. Besides my advisor and co-chair of the committee, I would also like to acknowledge other fine scientists on my committee: Young-Han Kim, and William Hodgkiss, for their insightful comments and suggestions.

I would also like to thank to Nitin Gupta, Jocelyn Bruand, and Jian Wang for their help on data collection and information. I owe my deepest gratitude to all lab members. Also I am grateful to Professor Vivian Hook and Dr. Steven Bark for collaborating to publish a research paper: Kim., Y., Bark., SJ., Hook, V., Bandeira, N. (2011). "NeuroPeida: Neuropeptide database and spectral library" in Bioinformatics journal.

My sincere thanks also go to my friends: Minjeong Kwon, Misuk shin, Kyunghwa Kim, Eunjeong Jang, Jungwook Lee, Jungik Kim, Mingyu Kim, Jimin Ban and UCSD friends for always helping me by giving me friendship and cheering me a lot. Also I would like to show my gratitude to alumni Sungmo Seo, JoAnn M. Clark and her

husband Mr. Clark, my mentor Kannan Kamarjan, and Professor Eon-kyeong Joo for supporting me immensely by cheering me and giving me endless warmth.

My deepest gratitude goes to my family: my aunt Chong-hui Johnson, uncle Rodney Johnson and Wonshik Choi who has always been supportive and affectionate, my sister Jina Ram, and brother Sang-wan Kim who have always been there with me and giving me joy and happiness, my beloved grandma Gye-hwa Kim for her bondless love, and last but not least my parents Kyung-tae Kim, and Mae-hee Choi for their loving consideration and understanding and so much more.

ABSTRACT OF THE THESIS

NeuroPedia: Neuropeptide database and spectral library

by

Yoona Kim

Master of Science in Electrical Engineering (Computer Engineering)

University of California, San Diego, 2011

Professor Nuno Bandeira, Chair

Professor Robert Hecht-Nielsen, Co- Chair

Neuropeptides are essential for cell-cell communication in neurological and endocrine physiological processes in health and disease. While many neuropeptides have been identified in previous studies, the resulting data has not been structured to facilitate further analysis by tandem mass spectrometry (MS/MS), the main technology for high throughput neuropeptide identification. Many neuropeptides are difficult to identify when

searching MS/MS spectra against large protein databases because of their atypical lengths (e.g., shorter/longer than common peptides) and lack of appropriate residues to facilitate peptide ionization/fragmentation.

NeuroPedia is a neuropeptide encyclopedia of peptide sequences (including genomic and taxonomic information) and spectral libraries of identified MS/MS spectra of homolog neuropeptides from multiple species. Searching neuropeptide MS/MS data against known NeuroPedia sequences improves the sensitivity of database search tools. Moreover, the availability of neuropeptide spectral libraries also enables the utilization of spectral library search tools, and further improves the sensitivity of peptide identification. These will also reinforce the confidence in peptide identifications by enabling visual comparisons between new and previously identified neuropeptide MS/MS spectra.

# Chapter 1 INTRODUCTION

## 1.1 Overviews of neuropeptide and neuropeptidomics

Neuropeptides are peptide neurotransmitters and hormones that mediate cell-cell communication for regulation of physiological functions and biological processes. They are present throughout the central nervous system as well as in peripheral organs such as the pancreas, the adrenal glands, and the cells of the immune system (Hokfelt, T., Bartfai, T., et al. 2003; Ubink, R., Calza, L., et al. 2003; Svensson, M., Sköld, K., et al. 2007). Understanding the role and regulation of neuropeptide forms in health, disease, and drug treatments requires the ability to globally analyze neuropeptide expression in an unbiased form. Multiple neuropeptides are secreted and utilized to coordinate regulation of physiological functions and it is important to understand the global knowledge of the neuropeptide profiles utilized in neuroendocrine control of cellular and biological functions (Hook, V., Bark, SJ., et al. 2010). For example, stress induces the secretion of enkephalins, catestain, NPY, VIP, galanin, and other neuropeptides (Goldstein, DS., and Kopin, IG. 2008; Hook, V., Toneff, T., et al. 2008; Whitworth, EJ., Kosti, O., et al. 2003; Nankova, BB., and Sabban, EL. 1999).

Neuropeptidomics is the systematic, comprehensive, qualitative and quantitative multiplex analysis of neuroendocrine peptides (Fälth, M., Sköld, F., et al. 2007). It is

helpful to detect and quantify peptide patterns in the sample, compare and select peptides that differ in abundance by more than normal biological variation, and identify and further characterize the selected peptides (Svensson, M., Sköld, K., et al. 2007). Mass spectrometry (MS) based neuropeptidomics is highly suited for untargeted, global neuropeptides studies and also it helps to understand how multiple neuropeptides, rather than a single neuropeptide, are cosecreted for coregulation of key biological functions (Svensson, M., Sköld, K., et al. 2007; Fricker, L. D. et al. 2007; Bora, A., Annanqudi, SP., et al. 2008; Li, L. and Sweedler, JV. 2008; Hook, V., Bark, SJ., et al. 2010).

## 1.2 Mass spectrometry

Mass spectrometry (MS) is a highly sensitive analytical technique that measures the mass or mass-to-charge ratio (*m/z*) of charged particles, such as protein, peptide and other ionizable molecules of biological interest. It is used for determining masses of particles, for quantitating the elemental compositions, and for elucidating chemical structures.

## 1.2.1 Mass spectrometer

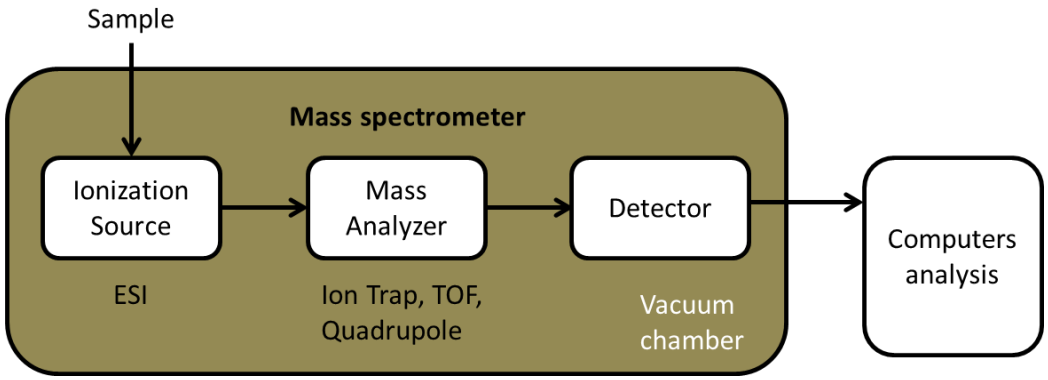



Figure 1.1: A simplified schematic of Mass spectrometer. Sample is introduced to mass spectrometer. Ionization source makes gas phase ions. Mass analyzer separates ions by *m/z* and fragments Ions. Finally, detector counts number of ions for each *m/z* and a mass spectrum is produced.

Mass spectrometry uses instrument called mass spectrometer consisting of three fundamental parts: Ionization Source, Mass Analyzer, and Detector. A simplified schematic of a mass spectrometer system is given in Figure1.1.

Before mass spectrometry analysis, protein samples are usually treated with a specific protease (e.g. trypsin, v8, and etc.), which cleaves the protein in a predictable way. The most commonly used protease is trypsin, which results in so-called “tryptic” peptides. The trypsin proteolysis of sample digestion leads to fragments according to the desirability of placing basic residues, notable arginine (R) and lysine (K) at the C-terminus of a peptide (Svensson, M., Sköld, K., et al. 2007).

The samples can be introduced to the ionization source of the mass spectrometer directly via an intermediary chromatography device (e.g. Liquid chromatography, and

Gas chromatography) according to the ionization method being used, as well as the type and complexity of the sample. For soluble biological samples, Electrospray Ionization (ESI) is one of the most used ionization methods (Carlton, DD., and Schug, KA. 2011).

In ESI, the sample is dissolved and a liquid of sample is introduced at high voltage. This creates a spray of charged droplets (sample) which leads to very small highly charged droplets capable of producing gas phase ions (Andersen, JS., Svensson, B., et al 1996). The sample is thus ionized by the addition or removal of hydrogen ions. Ionized molecules in the ionization source acquire energy to leave the source. According to *m/z* of ionized molecules, the mass accuracy, and the mass resolution, a mass analyzer separates the ions. There are several types of mass analyzers currently used in mass spectrometry including ion trap, time-of-flight (TOF), quadrupoles and Fourier transform ion cyclotron analyzers. The compatibility of different analyzers with different ionization methods varies but ESI can use all of the analyzers listed above.

Finally, the detector records either the charge induced or the current produced when an ion passes by or hits a surface. A scanning instrument in the detector will produce a mass spectrum plotted *m/z* values of the ions against their intensities to show the number of components in the sample, the molecular mass of each component, and the relative abundance of the various components in the sample (Robert, K. 1994).

## 1.2.2. Tandem mass spectrometry (MS/MS)

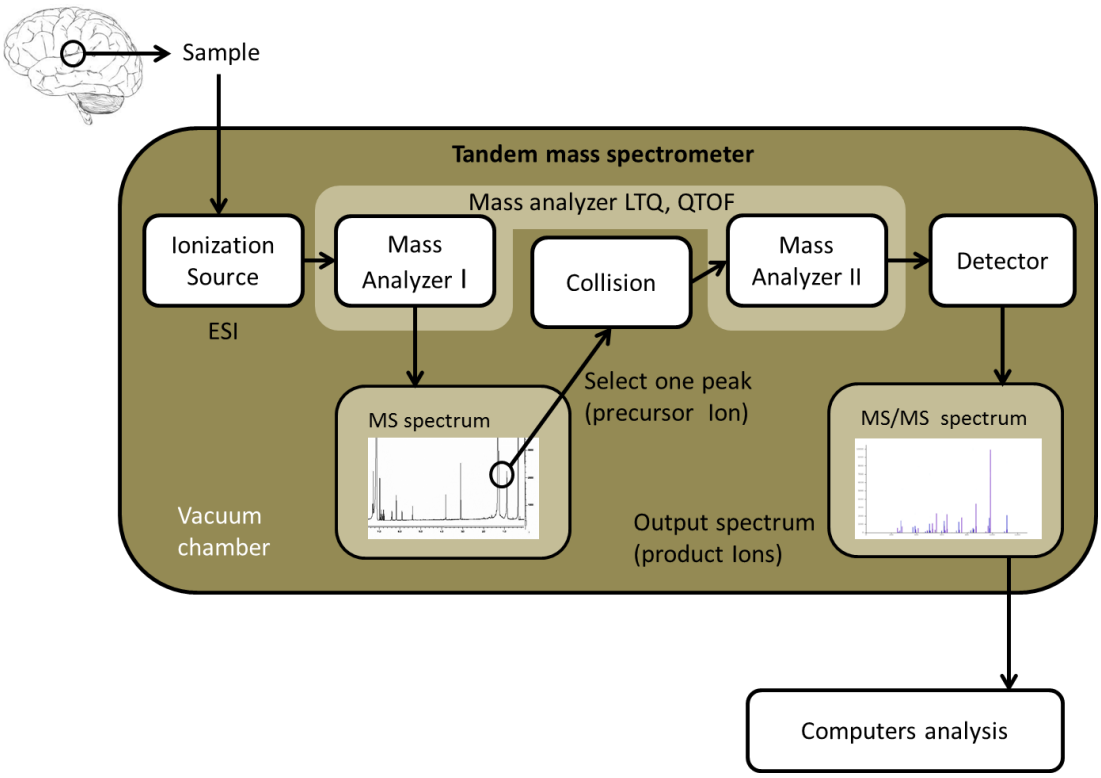



Figure 1.2: A simplified schematic of Tandem mass spectrometer (MS/MS). In difference from single-stage MS, there are two mass analyzers and a peptide collision system. First mass analyzer separates *m/z* of MS spectrum and selects one peak (precursor ion) for sequencing. The collision system fragments the ion to smaller product ions. Second mass analyzer separates *m/z* of ions and the detector determines their relative abundances in the MS/MS spectrum.

Tandem mass spectrometry is a technique using two or more stages of mass spectrometry for structural and sequencing of peptides and other biochemical samples. The principle of tandem mass spectrometry is to use two mass analyzers as illustrated in Figure 1.2 (Robert K. Boyd 1994). The first mass analyzer subjects a selection of

precursor ions formed in the ion source to fragmentation usually by collision-induced dissociation (CID, collision with inert gas molecules). The resultant *m/z* of product ions scanned in the second mass analyzer are then detected and recorded in a mass spectrum (Wells, JM. and McLuckey, SA. 2005). This is a powerful way of confirming the identity of certain compounds and of determining the structure of unknown molecular species (Sleno, L., and Volmer, DA. 2004).

In tandem mass spectrometry, mass analyzers can be used in combination as a tandem system to generate different types of data and take advantage of the strengths of each. The two most used hybrid mass analyzer system such as Linear Trap Quadrupole (LTQ), and Quadrupole-TOF (QTOF).

Thermo LTQ is an example of an ion trap instruments which detect ions by trapping them and it usually used in conjunction with ESI. It is robust, highly sensitive and has a very good fragmentation but it has relatively low mass accuracy and resolution (Schwartz, JC., Michael, W., et al. 2002). However, the hybrid QTOF instruments have high sensitivity, resolution and mass accuracy, and the resulting fragment ion spectra are often more extensive and informative than those generated in ion trap instruments (Steen, H., Küster, B., et al. 2001).

# 1.3 Peptide Identification using MS2 Analysis

## 1.3.1 Peptide fragmentation

A precursor ion mass is selected and allowed to undergo low energy collisions with neutral gas to product fragments at three different types of bonds along the amino acid backbone: the NH-CH (N-C), CH-CO (C-C), and CO-NH (C-N) bonds. Figure 1.3 shows the accepted nomenclature for fragment ions (Roepstorff, P. and Fohlman, J. 1984).

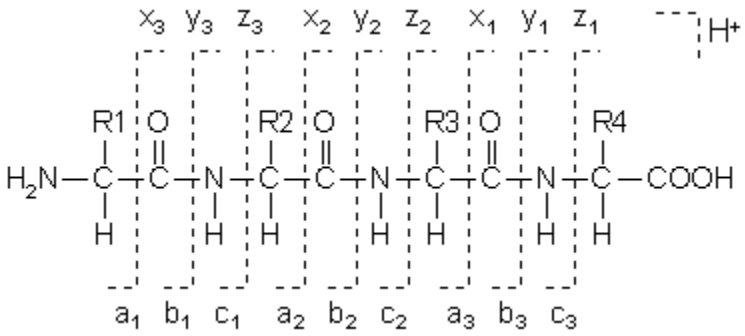



Figure 1.3: The accepted nomenclature for fragment ions (Roepstorff, P. and Fohlman, J. 1984). If a charge is retained on the N-terminal fragment, the fragment ion is *a*, *b*, or *c*. If a charge is retained on the C-terminal fragment, the fragment ion is *x*, *y* or *z*.

The subscripts of ions annotation indicate the number of residues in them. Also, low energy CID of peptides results in a limited number of fragment ions. The key sequence-specific fragment ions are the *y*-type and *b*-type ions, and both can lose water or ammonia. Figure 1.4 shows the structure of singly charged $b_2$ and $y_2$ ions according to Figure 1.3.

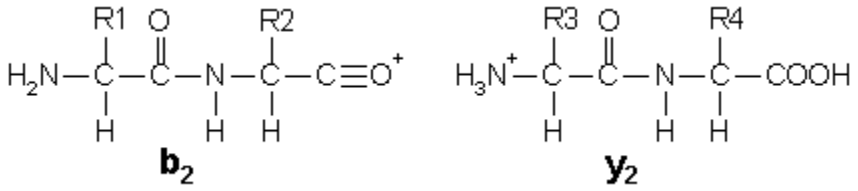



Figure 1.4: The structure of singly charged $b_2$ and $y_2$ ions according to the accepted nomenclature for fragmentations in Figure 1.3.

The second mass analyzer measures the production ions such as *b*-ions and *y*-ions and the resulting (*m/z*, intensity) pairs are represented as a spectrum. Figure 1.5 illustrates a spectrum of peptide FKLDDDLEHQ and numbers above the spectrum indicate the theoretical unit *m/z* of each *b*-ions and *y*-ions. $b_1$ and $y_1$ ions are hardly observed in the spectra. The highest intensity peak is $y_8$ (LDDDLEHQ) as the dominant ion in the analyzer and its complementary *b*-ion is $b_2$ (FK) in Figure 1.5.

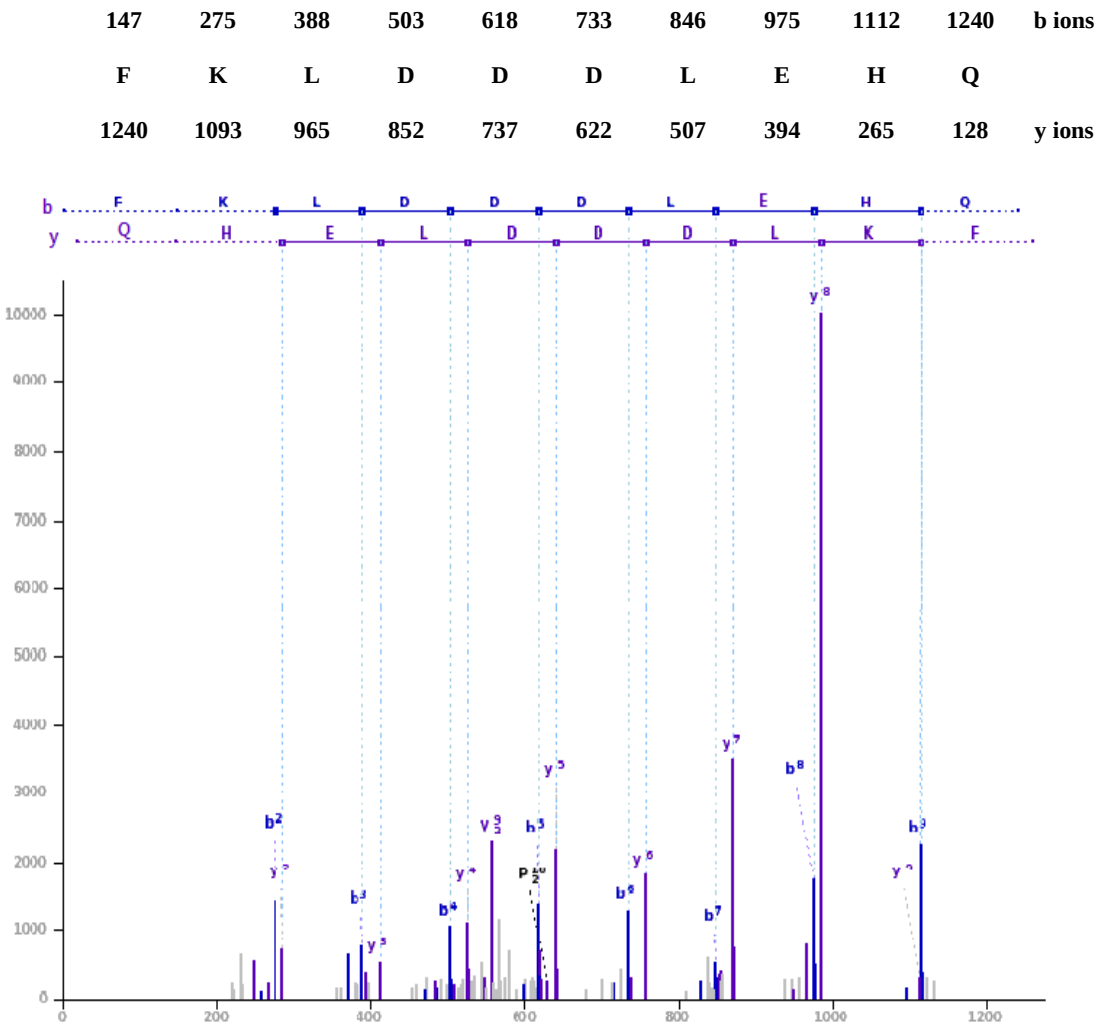



Figure 1.5: An example of tandem mass spectrum identification. A peptide FKLDDDLEHQ has predominant fragmentations such as $b_9$ (FKLDDDLEH) and $y_8$ (LDDDLEHQ) ions.

Resultant spectra from MS/MS need to be sequenced by peptide identification methods. There are two most used methods introduced in next chapter.

## 1.3.2 Database search

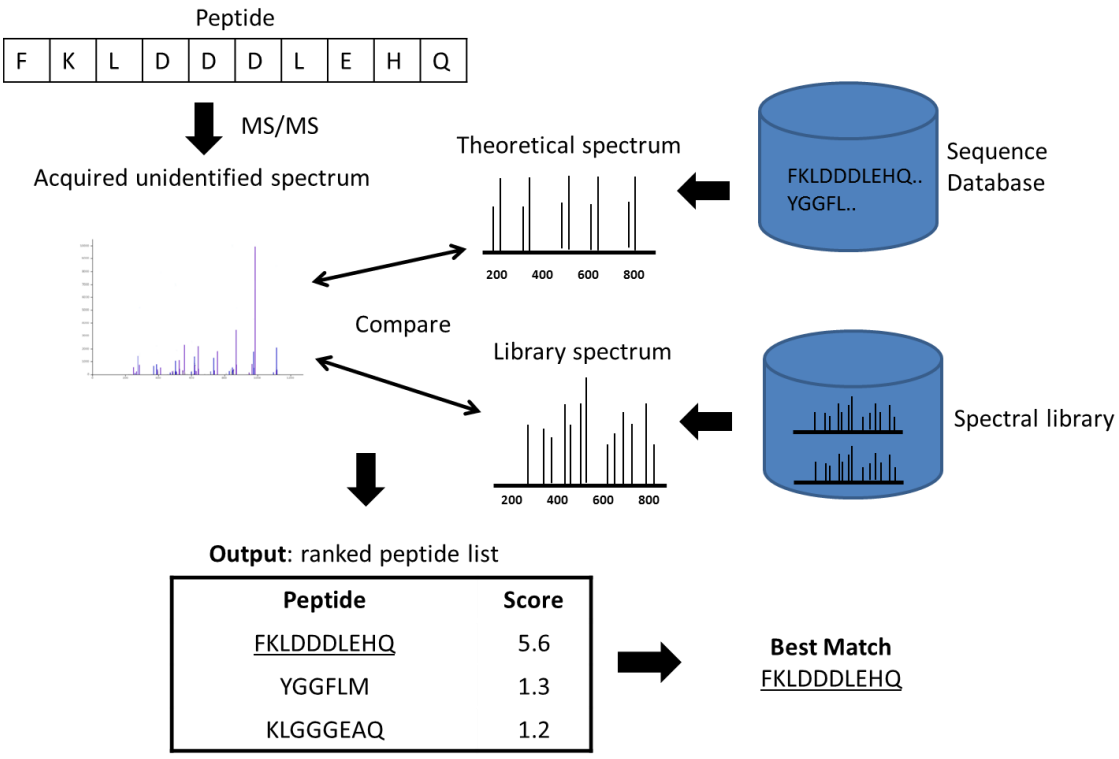



Figure 1.6: Peptide identification strategies. It can be performed by correlating acquired experimental MS/MS spectra with theoretical spectra predicted for each peptide contained in a protein sequences database (database search approach), or against spectra from a spectral library containing previously identified MS/MS spectra (spectral library search).

The most predominant identification method is database search. Several MS/MS database search tools are currently available, such as SEQUEST (Eng, JK., McCormack AL., et al. 1994), Mascot (Perkins, DN., Pappin, DJ., et al. 1999), X!Tandem (Craig, R., and Beavis, RC. 2004) and InsPecT (Tanner, S., Shu, H., et al. 2005). All these tools

operate in similar manner described in Figure 1.6 such as finding best match score resulting from comparison of query spectra (experimental spectra) with theoretical spectra from protein database. However each tool has slightly different ways for assigning peptides to MS/MS spectra and for statistical validation of peptide identification. Also it is very time and space intensive for identification if the input spectrum is searched against all possible protein databases. Therefore, each database search tool provides options for making a smaller set of candidate peptide according to the parent ion mass tolerance, and enzyme digestion constraint (trypsin, chymotrypsin, Lys-C, non-enzyme) (Nesvizhskii, AI. 2010).

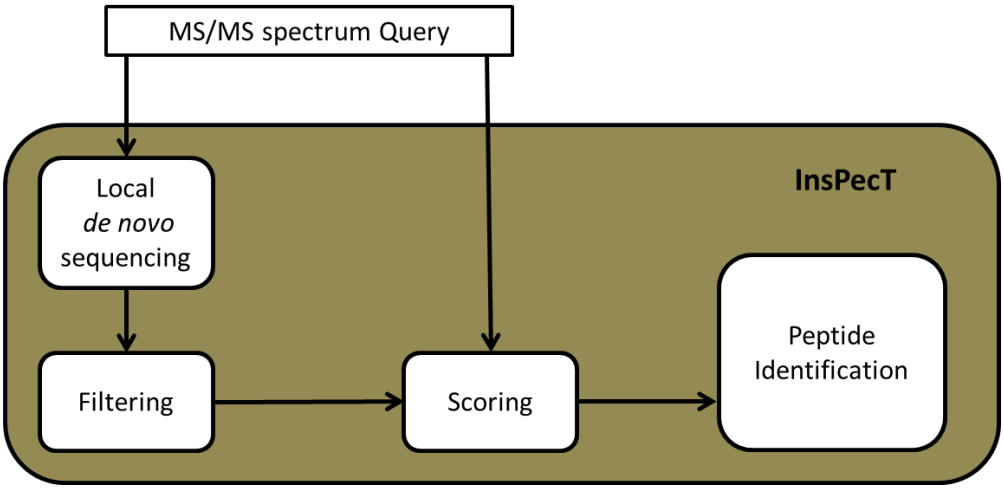



Figure 1.7: The flow of InsPecT from Tanner, S., Shu, H., et al. 2005.

InsPecT is one of database searching tool but its filtering stage performs a partial *de novo* interpretation which generates short peptide sequence tags from the spectrum. A peptide sequence tag (3-5 residues long) obtained by MS/MS can be used to filter peptides in a protein database. Tag-based search is not only orders of magnitude more efficient than other filters, but also accurate tagging strategy is important to restrict candidate peptides for scoring. Following the scoring stage, a validation stage calculates the probability that the top scoring peptide is the correct one.

In this project, we use InsPecT for peptide identification from experimental spectra in order to construct spectral library.

### 1.3.3 Spectral library search

Instead of identifying against theoretically predicted spectra, MS/MS spectrum can be assigned peptides by matching against a spectral library - just a large collection of experimentally observed MS/MS spectra identified in previous experiments (Yates, JR., Morgan, SF., et al. 1998; Nesvizhskii, AI. 2010). SpectraST (Lam, H., Deutsch, EW., et al. 2007), Bibliospec (Frewen, BE., Merrihew, GE,. et al. 2006), X!Hunter (Craig, R., Contens, JC., et al. 2006) and M-SPLIT (Wang, J., Perez-Santiago, J., et al. 2010) are existing library search tools which decide the best match between a MS/MS spectrum and library spectra, as illustrated in Figure 1.6.

In addition, the spectral library search improves efficiency, sensitivity and reliability of peptide identification by considering all spectral features, including actual

fragment intensities, neutral losses from fragments, and various uncommon or even unknown fragments to determine the best matches.

The main disadvantage of spectral library search is to require a collection of identified spectra instead of just sequences. However, in the context of proteomics, spectral libraries are typically compiled from peptide MS/MS spectra obtained from the analysis of complex biological samples and identified confidently by traditionally sequence-database searching (Lam, H., Deutsch, EW., et al. 2008).

### 1.3.4 False Discovery Rate

Database search identifies peptide to spectrum match (PSM) with highest PSM score for each spectrum. The correct interpretation of the spectrum may not be among the candidates considered by the search engines because of peptides not contained in sequence databases or spectra libraries, or even nonpeptide species that happened to be selected for fragmentation (Lam, H., Deutsch, EW., et al. 2009). It is difficult to determine manually whether each identification is correct or not because of the tens of thousands of spectra.

False discovery rate (FDR) is defined as the expected proportion of incorrect PSMs among all accepted PSMs (see Figure 1.8). It is essential for statistical confidence of peptide identification data for known error rates (Benajmini, Y., and Hochberg, Y. 1995). One of most used FDR assessment strategies is Target-decoy strategy. In the context of sequence database searching, one can calculate the FDR from the number of positive decoy identification and this is often achieved by concatenating a decoy protein

database, typically consisting of reversed, or randomized, or shuffled sequences from real protieins, to the tartet database before searching (Lam, H., Deutsch, EW., et al. 2009). The basic assumption is that matches to decoy PSMs and false matches to sequences from the target database follow the same distribution (Elias, JE., and Gygi, SP. 2007). At score threshold $S_T$, we calculate the FDR as follows:

$$\text{FDR } (S_T) = \frac{N_d(S_T)}{N_t(S_T)}$$

, where $N_t$ is the number of target PSMs with scores above the threshold, and $N_d$ is the number of decoy PSMs among them.

Figure 1.8 illustrates an example of filtering using the target-decoy strategy. Given FDR consider as an associated error rate estimate and it is predefined before the search. PSMs are filtered using the score threshold calculated from FDR.

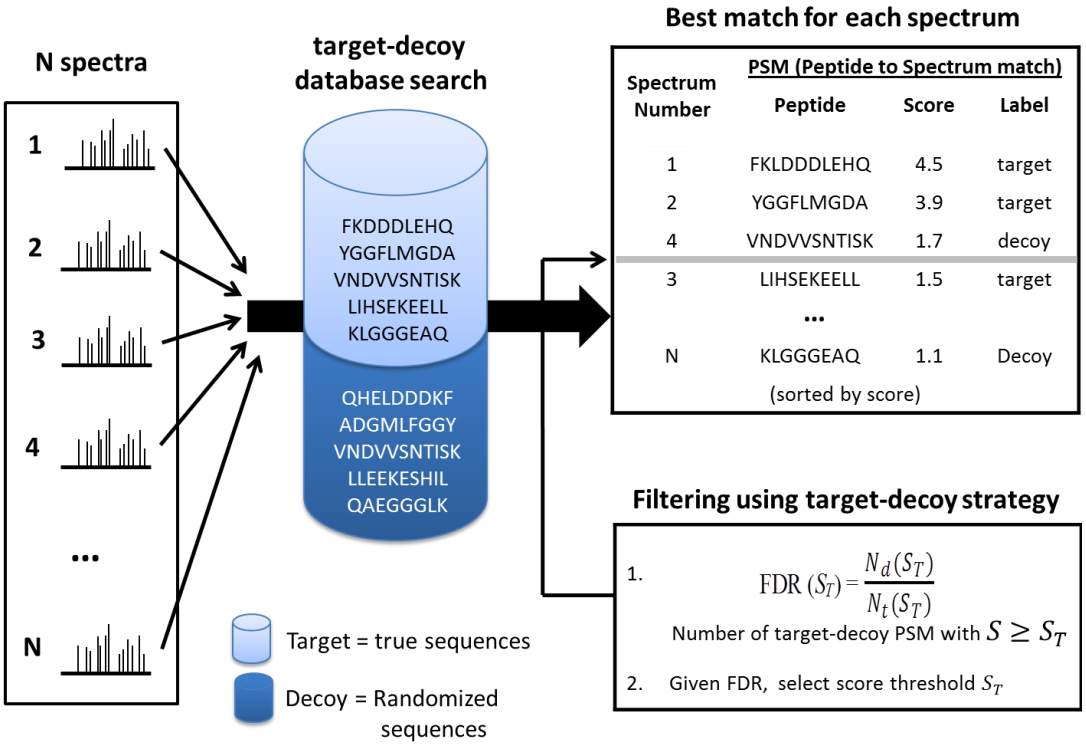



Figure 1.8: Target-decoy strategy for FDR assessment. The best peptide match for each spectrum is selected for further analysis based on score threshold calculated from FDR.

# 1.4 Neuropeptide identification using MS2

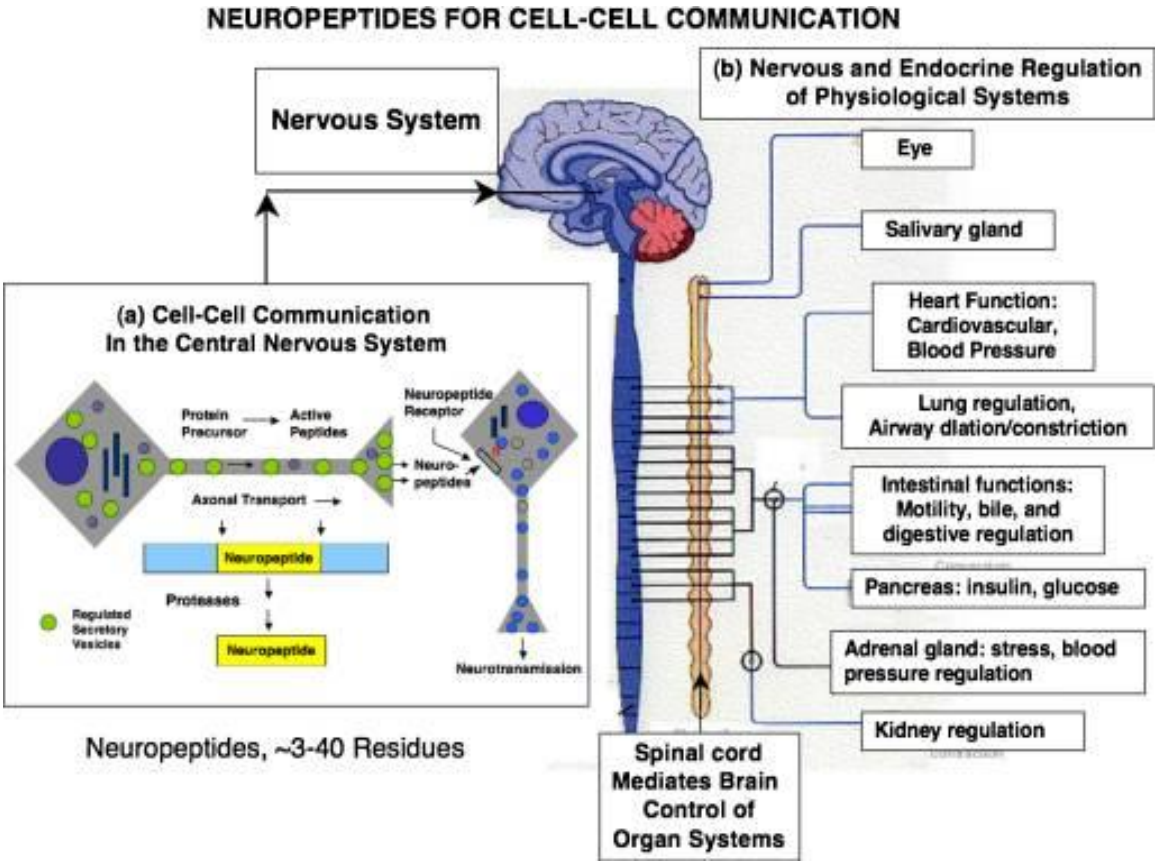



Figure 1.9: Neuropeptides for neuronal and endocrine cell-cell communication. (a) Neuropeptide, neurotransmitters, in the central nervous system of brain. (b) Neuropeptide, neurotransmitters and peptide hormones in the peripheral nervous system and endocrine systems for regulation of physiological organ functions. (Hook, V., Bark, SJ., et al. 2010)

Neuropeptide is essential for neuronal and endocrine cell-cell communication for biological and physiological organ functions. As illustrated in Figure 1.9, neuropeptides mediate chemical cell-cell communications among neurons and organs. Proteolytic

processing of the precursor protein (regulated in secretary vesicles) occurs during transport from the neuronal cell body via the axon to nerve terminals. Processed neuropeptides are contained within secretory vesicles at the synapse as well as neurotransmission. Moreover, for neuronal and endocrine regulation of physiological systems, neuropeptides function as hormones that regulate organ systems, linking the central nervous system with peripheral neuronal control of physiological functions.

The neuropeptide sequences consist of a combination of 20 amino acids and their length are 3-40 amino acids (Hokfelt, T., Bartfai, T., et al. 2003; Strand, FL. 2003; Svensson, M., Sköld, K., et al. 2007). Also the sequence is not only the combination of 20 amino acids but it may also have biochemical, and structural changes due to post-translational modifications – Addition/removal of chemical groups to/from specific amino acids. Posttranslational processing occurs when proteins are digested into neuropeptides and it extends the range of functions of the neuropeptides (Svensson, M., Sköld, K., et al. 2007).

As mentioned in Chapter1, multiple neuropeptides regulate biological and physiological functions in the body. New approaches utilizing mass spectrometry-based approaches (Yates, JR., Ruse, CI., et al. 2009; Leitner, A., and Lindner, W. 2006; Bantscheff, M., Shirle, M., et al. 2007; Nesvizhskii, AI. 2007; Kapp, EA., Schutz, F., et al. 2005) to simultaneously identify neuropeptide profiles in a single biological event have the potential to open our understanding of how multiple neuropeptides, rather than a single neuropeptide, are cosecreted for coregulation of key biological functions (Hook, V., Bark, SJ., et al. 2010).

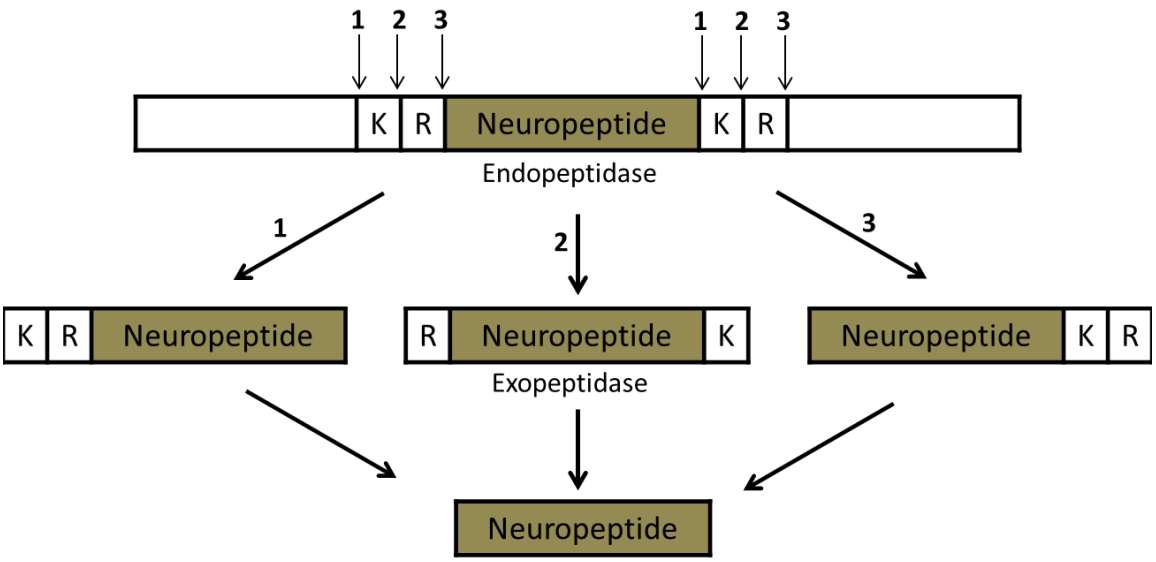



Figure 1.10: A schematic protease pathway of non-tryptic neuropeptide.

The protease pathways for neuropeptide production illustrated in Figure 1.10 are different from common tryptic protease pathways. It creates non-tryptic peptide fragments consisting of very short lengths of three to seven residues or long lengths of more than 15-20 residues. These unique characteristics of neuropeptides (e.g. short/long sequences or nontryptic) presents difficulties for identification from tandem mass spectrometry (MS/MS) with popular database search tools developed for identifications of tryptic peptides. For example, short neuropeptides can lead to inaccurate search results because the database search tools usually assign lower scores to short peptides. Conversely, long or nontryptic neuropeptides are difficult to identify because most database search tools are trained for tryptic peptides cleaved at K/R and because peptide fragmentation processes for long neuropeptides are usually not efficient.

In addition, searching larger databases takes more time because of the number of comparisons and reduces the number of resulting identifications by allowing more choices for false positives (Nesvizhskii, AI. et al., 2010). Therefore, while some neuropeptides can be identified with current bioinformatics approaches, complete neuropeptidomics will require the design of novel computational tools for identifying both short and long neuropeptides using tandem mass spectrometry (Hook V., Bark, SJ., et al. 2010).

The online neuropeptide repository at www.neuropeptides.nl provides non-searchable neuropeptide sequences, gene names, precursor names, and expected expression in the human brain. It also offers hyperlinks to bioinformatics databases on genomes, transcripts, protein structure and brain expression (Burbach, JP. 2009). Unfortunately, this resource is not designed to enable identification from MS/MS data. Users must search their data using other peptide database search tools and later compare the results against the neuropeptide list. This process is much less sensitive and requires time consuming manual matching of search results to information in existing resources.

However, the new NeuroPedia database and spectral library holds great potential for removing the bottleneck that occurs during the identification process in the field for neuropeptidomics. The aim of this research was to develop neuropeptide sequence database and spectral library for facilitating the neuropeptide identification using tandem mass spectrometry.

# Chapter 2 NeuroPedia overview

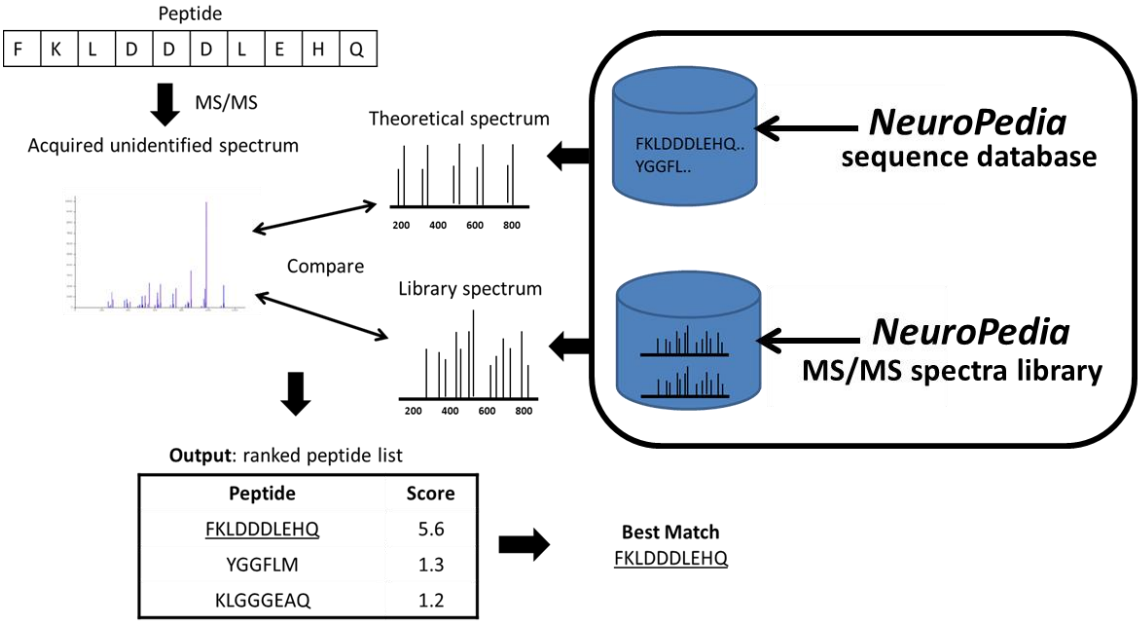



Figure 2.1: The role of NeuroPedia in peptide identification using tandem mass spectrometry.

NeuroPedia is a specialized neuropeptide database and spectral library that is directly searchable using mass spectrometry data. Figure 2.1 shows which part of 'peptide identification using MS2' NeuroPedia contributes.

# 2.1 Neuropeptide sequence databases

NeuroPedia sequence database provides genomic and taxonomic information for neuropeptides from human, chimpanzee, mouse, rat, bovine, rhesus macaque, and California sea hare. The excel file format of the database includes neuropeptide sequences, their start and end amino acidpositions on the precursor protein, species, RefSeq ID (McEntyre, J., and Ostell, J. 2002), UniProt ID (Jain, E., Bairoch, A., et al. 2009), NCBI taxonomy ID (Benson, DA., Karsch-Mizrachi, I., et al. 2009; Sayers, EW., Barrett, T., et al. 2009) and gene reference ID. Also each neuropeptide is associated with its the number of identified spectra in the spectral library and the number of identical, overlapping and contained sequences. In addition, neuropeptide sequences are downloadable in the standard FASTA searchable format including neuropeptide name, neuroprecursor name and sequence.

Our sequence databases are compatible with InsPecT as well as other database search tools and users can search their unidentified spectra against the smaller neuropeptide sequence database instead of larger protein databases, thus achieving faster, more accurate and reliable identifications. It is because that the neuropeptide sequence database is much smaller and it thus reduces both the search time and space. Also the resultant identifications only consist of neuropeptide.

## 2.2 Neuropeptide spectral libraries

NeuroPedia spectral library consists of standard MGF format files classified by five species (human, bovine, mouse, rat, and medicinal leech), instrument and enzyme. Our spectral library sources (public spectral libraries and two in-house experiments) didn't have spectra of chimpanzee, rhesus macaque, and Californian sea hare. The header information of each MGF file has TITLE (File name of original spectra library), SCANS (Scan number from original spectral library), CHARGE (Peptide charge), PEPMASS (Peptide mass), and SEQ (Amino acid sequence). TITLE and SCANS can help to trace the spectra to its original file. These files are downloadable and can be browsed in NeuroPedia's web site as described in the conclusion.

NeuroPedia spectral libraries can be used by any spectral library search tools with support for MGF spectral libraries. Compared to larger spectral libraries including many more peptide spectra, NeuroPedia spectral libraries can lead to more sensitive, accurate and faster identifications of neuropeptides at the same FDR. It is because the smaller size of our spectral library only consists of neuropeptide spectra.

# Chapter 3 Methods

## 3.1 Developing sequence databases

The online neuropeptide repository www.neuropeptide.nl provided neuropeptide sequences, names and gene families from human brain and it helped to find neuropeptides of other species which the online repository didn't include. Using python HTML parsing of the neuropeptide repository, we initially gathered neuropeptide names, gene families, gene names and their protein names.

The collected information as initial set helped to distinguish neuropeptides from other proteins, and peptides database. We manually typed all neuropeptide sequences and names from of the Handbook of Biologically Active Peptides (Kastin, A. 2006), sections X: Brain Peptides Section, XI: Endocrine Peptides Section, XII: Ingestive Peptides Section, XIII: Gastrointestinal Peptide Section and XVII: Opioid Peptide Section. In addition, we collected NCBI taxonomy ID and gene reference ID from NCBI (http://www.ncbi.nlm.nih.gov/). From UniProt (http://www.uniprot.org/), a comprehensive public repository for protein sequences and annotation data, we obtained also neuropeptide sequences, their start and end positions on the precursor protein, species, RefSeq ID, and UniProt ID. Using cluster searching at 50% protein homolog in UniProt, we expanded the catalog of species from human into chimpanzee, mouse, rat, bovine, rhesus macaque, and California sea hare.

We further analyzed the collected neuropeptide sequences to classify sequence similarities between neuropeptides into three match types: a) identical if the sequences are exactly the same, b) overlapping if the prefix of one sequence exactly matches the suffix of the other sequence for at least k characters, where k is half the length of longest sequence and c) homolog if overlapping as in b) but allowing up to two amino acid substitutions. For collected sequences longer than four amino acids, it is necessary to apply efficient algorithm for finding relations between all possible 340,725 pairs. First, we constructed an index dictionary including 104,976 subsequences of length four on 18 amino acids (ARNDCEGHIKMFPSTWYV). The reason why we didn't consider 20 amino acids is that we replaced all Leucines (L, 113 Da) with Isoleucines (I, 113 Da) and all Glutamines (Q, 128 Da) with Lysines (K, 128 Da) because of low mass accuracy of their indistinguishability at the observed CID spectra. For speeding up their comparisons, length four subsequences were converted to 18-nary numbers and we assigned index numbers for all candidate sequences. For every subsequence, its index is saved in the dictionary. For example, a peptide YGGMF can have two subsequences YGGM and GGMF. This peptide will be saved in two subsequence indexes 95,374 (YGGM, $18^3 \times 16 + 18^2 \times 6 + 18^1 \times 6 + 18^0 \times 10$) and 37,127 (GGMF, $18^3 \times 6 + 18^2 \times 6 + 18^1 \times 10 + 18^0 \times 11$). There are comparisons only between sequences sharing at least one subsequence index thus meaning that at least four characters should be exactly matched for finding any of three relations described above.

# 3.2 Collecting Spectral libraries

## 3.2.1 NIST spectral library and In-house spectral library

Most of spectra in the NeuroPedia library were obratined from Gupta N., Bark, SJ., et al. 2010. For human, there are three digestions using trypsin, v8 and non-enzyme (non-digested peptide extracts) and two types of instruments: ion trap (IT) and QTOF. For bovine, we had spectra of ion trap and non-enzyme. Gupta N., Bark, SJ., et al. 2010 also provides information of identified neuropeptides from their experimental spectra. All spectra were also searched using InsPecT search. Also Bruand, J., Sistila S., et al. 2011 provided a recently discovered medicinal leech neuropeptide from MALDI (Matrix-assisted laser desorption/ionizaition) imaging analysis.

Moreover, neuropeptide spectra were collected from NIST (National Institute of Standards and Technology) spectral libraries (Stein, S. E. and Rudnick, P.A. 2009). Because all spectra in NIST were digested with trypsin, more than half part of each resultant spectral library consists of tryptic peptide and other part are N-semitryptic, and C-semitryptic.

## 3.2.2 InsPecT search

All collected MS/MS spectra were searched against the NeuroPedia sequences database using InsPecT (Tanner, S., Shu, H., et al. 2005) at http://proteomics.ucsd.edu with search parameters: Instrument (ESI-ION-TRAP or QTOF), Cysteine protecting group (Carbamidomethylation +57), Protease (Trypsin, None), 2Da Parent mass tolerance,

0.5Da Ion tolerance, no post-translational modifications and including common contaminants (digestion enzymes and Human Keratins). V8 digested runs were searched as above but with the protease parameter set to 'None'.

### 3.2.3 Analyze identified spectrum

For identified spectra from InsPecT searches, we generated nine spectral libraries corresponding to species, instruments, and enzymes. In each spectral library, we saw repeated MS/MS acquisitions from the same peptides such that some spectra have exactly same sequence but their MQScores (Match quality score, the main measure of match quality) are different. We chose the highest MQScore spectra among spectra having the exactly same sequences. The collected sets of spectra consist of all different sequence corresponding to species, instruments, and enzymes. We plotted spectra using Specplot (Bouchard, P., and Bandeira, N. 2010) for visualization them and assessed the quality of the peptide/spectra matches. There are several factors of quality classification including poor quality of fragmentation/ionization.

We manually classified every spectrum as High quality (HQ) and Low quality (LQ), as illustrated in Figure 3.1. The light grey part of peaks shows unidentified sequence. Figure 3.1 has less grey part and it is easy to observe good series of sequence depending on peaks of *b*-ions and *y*-ions. Also we can observe clusters of ions according to mass addition of isotopic elements (C, H, N, O), and loss of water or ammonia. However, Figure 3.2 has a lot of grey part and also a lot of unexplained peaks located

between amino acid residues. It is difficult to figure out *b*-ions, *y*-ions, and their clusters because of a lot of unannotated noise peaks.

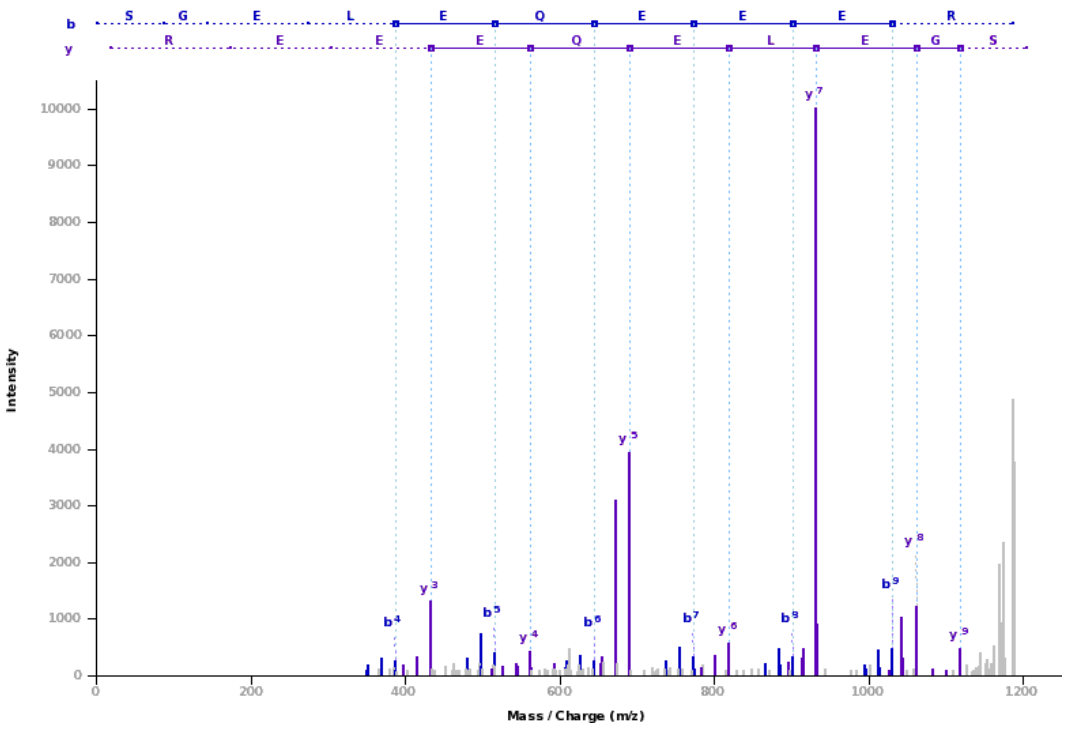



Figure 3.1: An example of Low quality spectra of SGELEQEEER from the human, IT, and trypsin database.

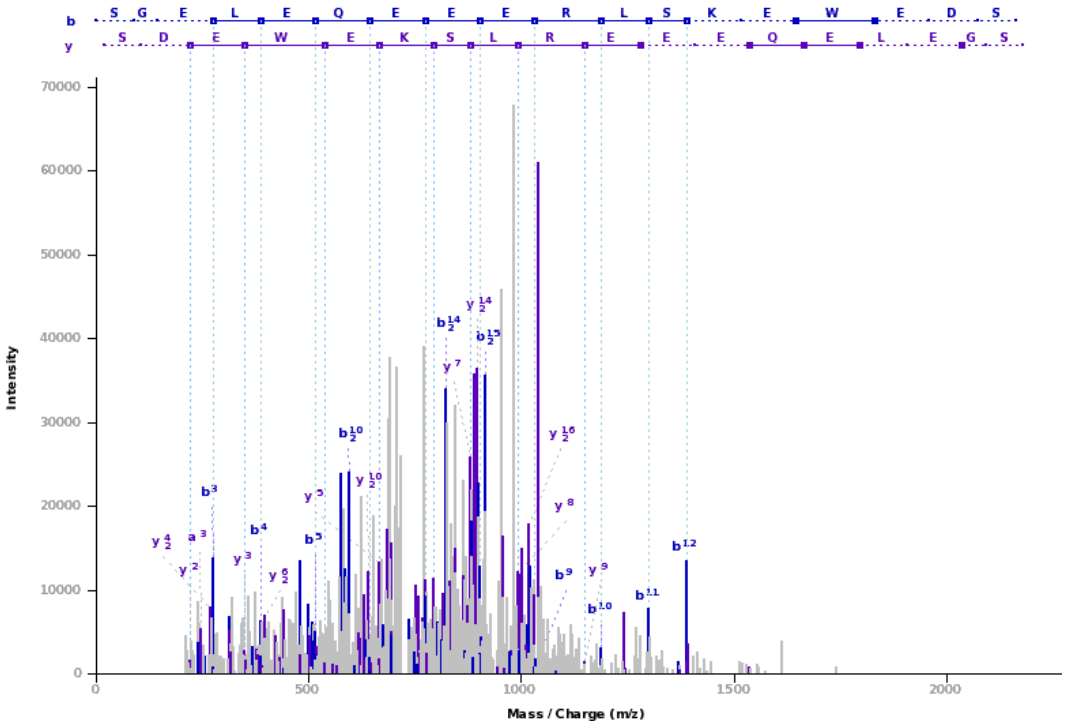



Figure 3.2: An example of Low quality spectra of SGELEQEEERLSKEWEDS from the human, IT, and trypsin database.

# Chapter 4 RESULTS

The NeuroPedia spectral library contains a total of 3,401 identified spectra in ten MGF files as described in Table 4.1. In addition to providing libraries for all identified spectra, NeuroPedia also contains libraries of manually validated high/low quality spectra for unique combinations of peptide sequence and precursor charge states (see Table 4.2).

As shown in “In-house”/”UniProt” in Table 4.1, searching against NeuroPedia identifies many more spectra than the UniProt database. “Uniprot (in-house)” numbers should be contrasted with the “In-house” column showing database search results of in-house spectra against NeuroPedia sequences, not against “NeuroPeida” (which shows the size of the processed spectra library). In all cases, the number of identified spectra based on NeuroPedia database is larger than the number based on UniProt.

Table 4.1: NeuroPedia spectral libraries (1) (including repeated MS/MS acquisitions from the same peptides).

| Species | Type[a] | Enzyme | NIST[b] | In-house[c] | NeuroPedia[d] | UniProt[e] |
|---|---|---|---|---|---|---|
| Human | IT[g] | Trypsin | 385 | 221 | 606 | 89 |
| Human | IT | v8 | 0 | 91 | 91 | 31 |
| Human | IT | none[i] | 0 | 1,630 | 1,630 | 335 |
| Human | QTOF[h] | Trypsin | 41 | 454 | 495 | 41 |
| Human | QTOF | v8 | 0 | 202 | 202 | 39 |
| Human | QTOF | None | 0 | 160 | 160 | 22 |
| Mouse | IT | Trypsin | 67 | 0 | 67 | 0 |
| Rat | IT | Trypsin | 4 | 0 | 4 | 0 |
| Bovine | IT | None | 0 | 145 | 145 | 0 |
| Leech[f] | QTOF | None | 0 | 1 | 1 | 0 |
| | Total | | 497 | 2,904 | 3,401 | 571 |

[a] Instrument type
[b] Spectra from NIST spectral libraries at http://peptide.nist.gov, accessed on July 30, 2010
[c] Spectra from Gupta N., Bark, SJ., et al. 2010 (searched against NeuroPedia peptide sequences) and Bruand, J., Sistila S., et al. 2011
[d] NeuroPedia spectral library
[e] Total number of identified spectra when searching In-house spectra against all UniProt Human sequences (comparable with search results in the "In-house" column)
[f] Medicinal leech
[g] Ion Trap
[h] Quadrupole Time-Of-Flight
[i] Undigested low molecular weight (≤ 10kDa)

Table 4.2: Neuropeptide spectral libraries (2) (including only best spectrum per peptide).

| Species | Type | Enzyme | NIST[a] | In-house[b] | Total[c] | HQ[d] | LQ[e] |
|---|---|---|---|---|---|---|---|
| Human | IT | Trypsin | 296 | 68 | 364 | 303 | 61 |
| Human | IT | v8 | 0 | 53 | 53 | 24 | 29 |
| Human | IT | None | 0 | 121 | 121 | 54 | 67 |
| Human | QTOF | Trypsin | 37 | 109 | 146 | 41 | 96 |
| Human | QTOF | v8 | 0 | 69 | 69 | 39 | 55 |
| Human | QTOF | None | 0 | 44 | 44 | 22 | 31 |
| Mouse | IT | Trypsin | 60 | 0 | 60 | 0 | 18 |
| Rat | IT | Trypsin | 4 | 0 | 4 | 2 | 2 |
| Bovine | IT | None | 0 | 33 | 33 | 24 | 9 |
| Leech | QTOF | None | 0 | 1 | 1 | 1 | 0 |
| All[f] | IT | All[q] | 360 | 275 | 635 | 449 | 186 |
| All | QTOF | All | 37 | 223 | 260 | 78 | 182 |
| | Total | | 397 | 498 | 895 | 527 | 368 |

[a] Number of spectra from NIST for unique peptide/charge-state pairs
[b] Number of spectra from Gupta N., Bark, SJ., et al. 2010 and Bruand, J., Sistila S., et al. 2011 for unique peptide/charge-state pairs
[c]Total Number of spectra for unique peptide/charge-state pairs 527 neuropeptide spectra
[d] Total Number of spectra in High Quality spectral library
[e] Total number of identified spectra when searching In-house spectra against
[f] Collection of species (human, mouse, rat, bovine, and leech) according to type of instruments
[g] Collection of enzymes (trypsin, v8, and no enzyme) according to type of instruments

Table 4.3: NeuroPedia sequence databases.

| Species | Number of sequences |
|---|---|
| Human | 270 |
| Rat | 195 |
| Mouse | 188 |
| Bovine | 154 |
| Rhesus macaque | 20 |
| Chimpanzee | 17 |
| California sea hare | 2 |
| Medicinal Leech | 1 |
| Total | 847 |

The NeuroPedia sequence database contains 847 neuropeptides from human, chimpanzee, mouse, rat, cow, California sea hare, rhesus macaque, and medicinal leech (see Table 4.3). Using InsPecT or any other database search tool, new MS/MS data can be searched against this sequence database. Figure 4.1 describes the excel format of database.

Out of all possible 340,725 pairs of neuropeptides sequences (without considering species), there are 531 pairs with identical sequences (type a), 5,020 pairs with overlapping sequences (type b), and 9,185 pairs with homolog sequences (type c). It can help interpret search results with relate missed-cleavage versions of such peptides. The database of sequence relation describes each pair in a row including neuropeptide sequence, neuropeptide name and species in alphabetical order, as illustrated in Figure 4.2.

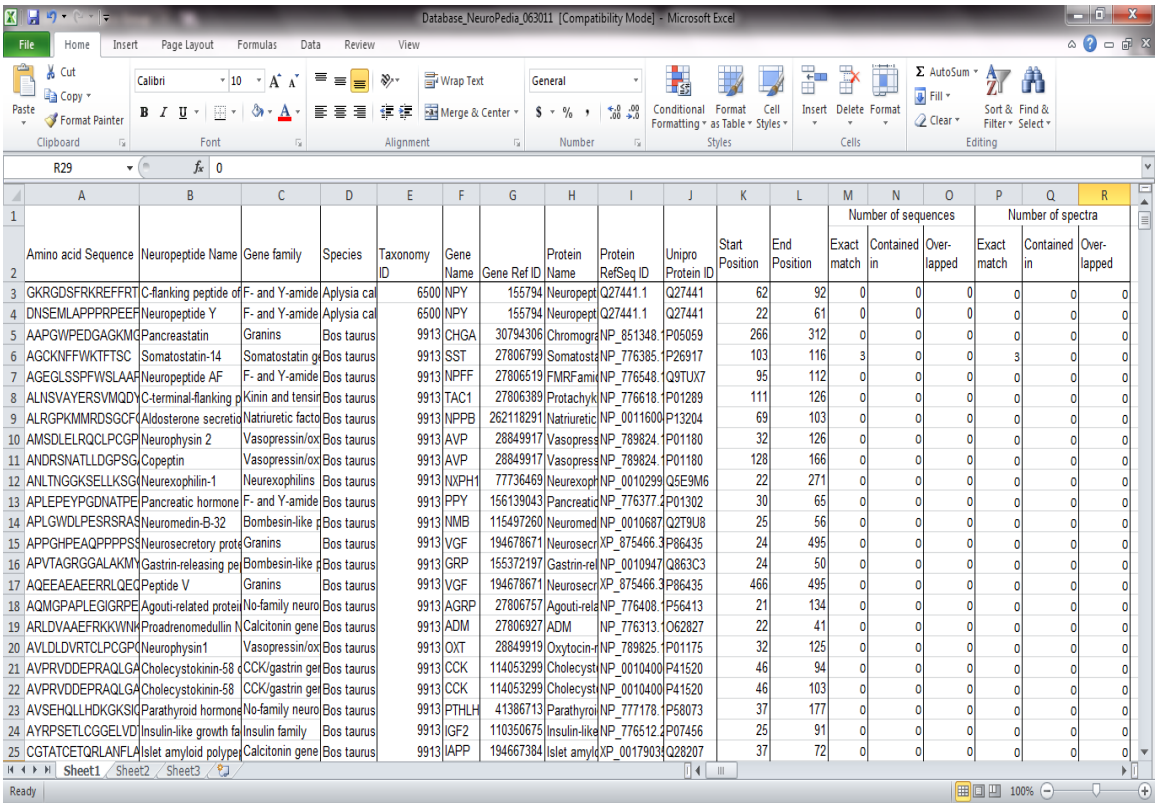
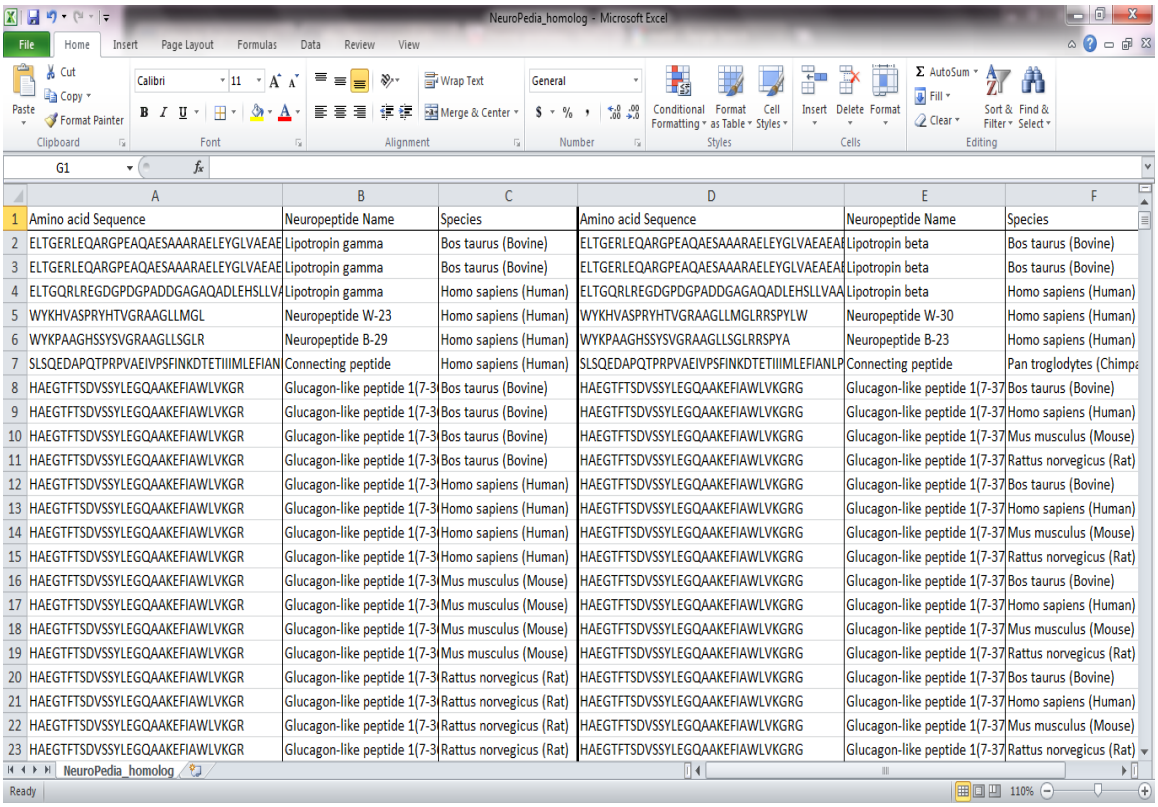



Figure 4.1: NeuroPedia sequence database excel screen shot.




| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | Amino acid Sequence | Neuropeptide Name | Species | Amino acid Sequence | Neuropeptide Name | Species |
| 2 | ELTGERLEQARGPEAQAESAAARAELEYGLVAEAE | Lipotropin gamma | Bos taurus (Bovine) | ELTGERLEQARGPEAQAESAAARAELEYGLVAEAEA | Lipotropin beta | Bos taurus (Bovine) |
| 3 | ELTGERLEQARGPEAQAESAAARAELEYGLVAEAE | Lipotropin gamma | Bos taurus (Bovine) | ELTGERLEQARGPEAQAESAAARAELEYGLVAEAEA | Lipotropin beta | Bos taurus (Bovine) |
| 4 | ELTGQRLREGDGPDGPADDGAGAQADLEHSLLVA | Lipotropin gamma | Homo sapiens (Human) | ELTGQRLREGDGPDGPADDGAGAQADLEHSLLVAA | Lipotropin beta | Homo sapiens (Human) |
| 5 | WYKHVASPRYHTVGRAAGLLMGL | Neuropeptide W-23 | Homo sapiens (Human) | WYKHVASPRYHTVGRAAGLLMGLRRSPYLW | Neuropeptide W-30 | Homo sapiens (Human) |
| 6 | WYKPAAGHSSYSVGRAAGLLSGLR | Neuropeptide B-29 | Homo sapiens (Human) | WYKPAAGHSSYSVGRAAGLLSGLRRSPYA | Neuropeptide B-23 | Homo sapiens (Human) |
| 7 | SLSQEDAPQTPRPVAEIVPSFINKDTETIIIMLEFIAN | Connecting peptide | Homo sapiens (Human) | SLSQEDAPQTPRPVAEIVPSFINKDTETIIIMLEFIANLP | Connecting peptide | Pan troglodytes (Chimpa |
| 8 | HAEGTFTSDVSSYLEGQAAKEFIAWLVKGR | Glucagon-like peptide 1(7-3 | Bos taurus (Bovine) | HAEGTFTSDVSSYLEGQAAKEFIAWLVKGRG | Glucagon-like peptide 1(7-37 | Bos taurus (Bovine) |
| 9 | HAEGTFTSDVSSYLEGQAAKEFIAWLVKGR | Glucagon-like peptide 1(7-3 | Bos taurus (Bovine) | HAEGTFTSDVSSYLEGQAAKEFIAWLVKGRG | Glucagon-like peptide 1(7-37 | Homo sapiens (Human) |
| 10 | HAEGTFTSDVSSYLEGQAAKEFIAWLVKGR | Glucagon-like peptide 1(7-3 | Bos taurus (Bovine) | HAEGTFTSDVSSYLEGQAAKEFIAWLVKGRG | Glucagon-like peptide 1(7-37 | Mus musculus (Mouse) |
| 11 | HAEGTFTSDVSSYLEGQAAKEFIAWLVKGR | Glucagon-like peptide 1(7-3 | Bos taurus (Bovine) | HAEGTFTSDVSSYLEGQAAKEFIAWLVKGRG | Glucagon-like peptide 1(7-37 | Rattus norvegicus (Rat) |
| 12 | HAEGTFTSDVSSYLEGQAAKEFIAWLVKGR | Glucagon-like peptide 1(7-3 | Homo sapiens (Human) | HAEGTFTSDVSSYLEGQAAKEFIAWLVKGRG | Glucagon-like peptide 1(7-37 | Bos taurus (Bovine) |
| 13 | HAEGTFTSDVSSYLEGQAAKEFIAWLVKGR | Glucagon-like peptide 1(7-3 | Homo sapiens (Human) | HAEGTFTSDVSSYLEGQAAKEFIAWLVKGRG | Glucagon-like peptide 1(7-37 | Homo sapiens (Human) |
| 14 | HAEGTFTSDVSSYLEGQAAKEFIAWLVKGR | Glucagon-like peptide 1(7-3 | Homo sapiens (Human) | HAEGTFTSDVSSYLEGQAAKEFIAWLVKGRG | Glucagon-like peptide 1(7-37 | Mus musculus (Mouse) |
| 15 | HAEGTFTSDVSSYLEGQAAKEFIAWLVKGR | Glucagon-like peptide 1(7-3 | Homo sapiens (Human) | HAEGTFTSDVSSYLEGQAAKEFIAWLVKGRG | Glucagon-like peptide 1(7-37 | Rattus norvegicus (Rat) |
| 16 | HAEGTFTSDVSSYLEGQAAKEFIAWLVKGR | Glucagon-like peptide 1(7-3 | Mus musculus (Mouse) | HAEGTFTSDVSSYLEGQAAKEFIAWLVKGRG | Glucagon-like peptide 1(7-37 | Bos taurus (Bovine) |
| 17 | HAEGTFTSDVSSYLEGQAAKEFIAWLVKGR | Glucagon-like peptide 1(7-3 | Mus musculus (Mouse) | HAEGTFTSDVSSYLEGQAAKEFIAWLVKGRG | Glucagon-like peptide 1(7-37 | Homo sapiens (Human) |
| 18 | HAEGTFTSDVSSYLEGQAAKEFIAWLVKGR | Glucagon-like peptide 1(7-3 | Mus musculus (Mouse) | HAEGTFTSDVSSYLEGQAAKEFIAWLVKGRG | Glucagon-like peptide 1(7-37 | Mus musculus (Mouse) |
| 19 | HAEGTFTSDVSSYLEGQAAKEFIAWLVKGR | Glucagon-like peptide 1(7-3 | Mus musculus (Mouse) | HAEGTFTSDVSSYLEGQAAKEFIAWLVKGRG | Glucagon-like peptide 1(7-37 | Rattus norvegicus (Rat) |
| 20 | HAEGTFTSDVSSYLEGQAAKEFIAWLVKGR | Glucagon-like peptide 1(7-3 | Rattus norvegicus (Rat) | HAEGTFTSDVSSYLEGQAAKEFIAWLVKGRG | Glucagon-like peptide 1(7-37 | Bos taurus (Bovine) |
| 21 | HAEGTFTSDVSSYLEGQAAKEFIAWLVKGR | Glucagon-like peptide 1(7-3 | Rattus norvegicus (Rat) | HAEGTFTSDVSSYLEGQAAKEFIAWLVKGRG | Glucagon-like peptide 1(7-37 | Homo sapiens (Human) |
| 22 | HAEGTFTSDVSSYLEGQAAKEFIAWLVKGR | Glucagon-like peptide 1(7-3 | Rattus norvegicus (Rat) | HAEGTFTSDVSSYLEGQAAKEFIAWLVKGRG | Glucagon-like peptide 1(7-37 | Mus musculus (Mouse) |
| 23 | HAEGTFTSDVSSYLEGQAAKEFIAWLVKGR | Glucagon-like peptide 1(7-3 | Rattus norvegicus (Rat) | HAEGTFTSDVSSYLEGQAAKEFIAWLVKGRG | Glucagon-like peptide 1(7-37 | Rattus norvegicus (Rat) |

Figure 4.2: NeuroPedia sequence match types of pairs of sequences screen shot.

NeuroPedia spectral libraries are compatible with the publicly available spectral library search tool M-SPLIT (Wang, J., Perez-Santiago, J., et al. 2010) and can be easily converted to other spectral library formats. To further facilitate visual evaluation of neuropeptide MS/MS spectra, NeuroPedia provides annotated spectrum images for every library spectrum (see Figure 4.3) and further separates spectral libraries by species, digestion enzyme, and instrument type in NeuroPedia webpage shown in Figure 4.4.

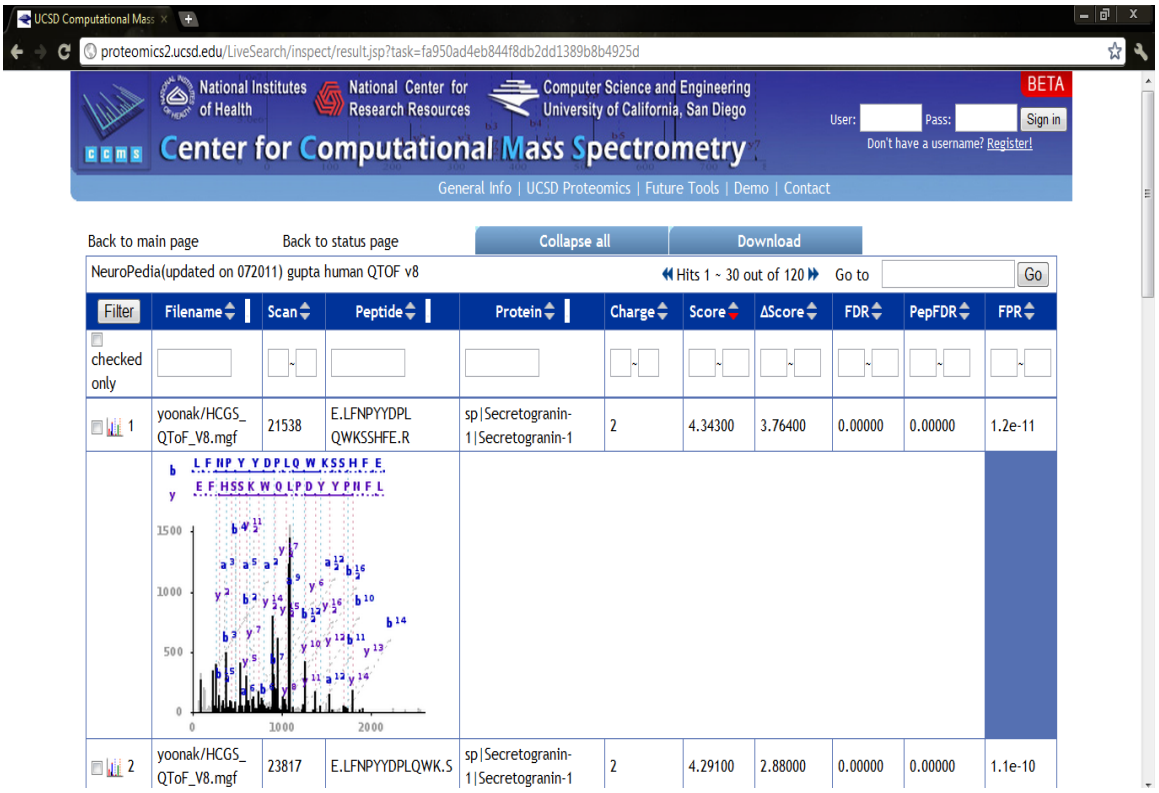



Figure 4.3: One example of browsing identified spectra. Description of each column in the table: "Filename" is the name of query spectra file. "Scan" is the number of scan number in the query file. "Peptide" is the identified peptide sequence. "Charge" is the charge of identified spectrum. "Score" is matching quality score of identification. "FDR" is false discovery rate. "PepFDR" is peptide false discovery rate. "FPR" is false positive rate of peptide identification and fraction of incorrect assignments above score threshold.

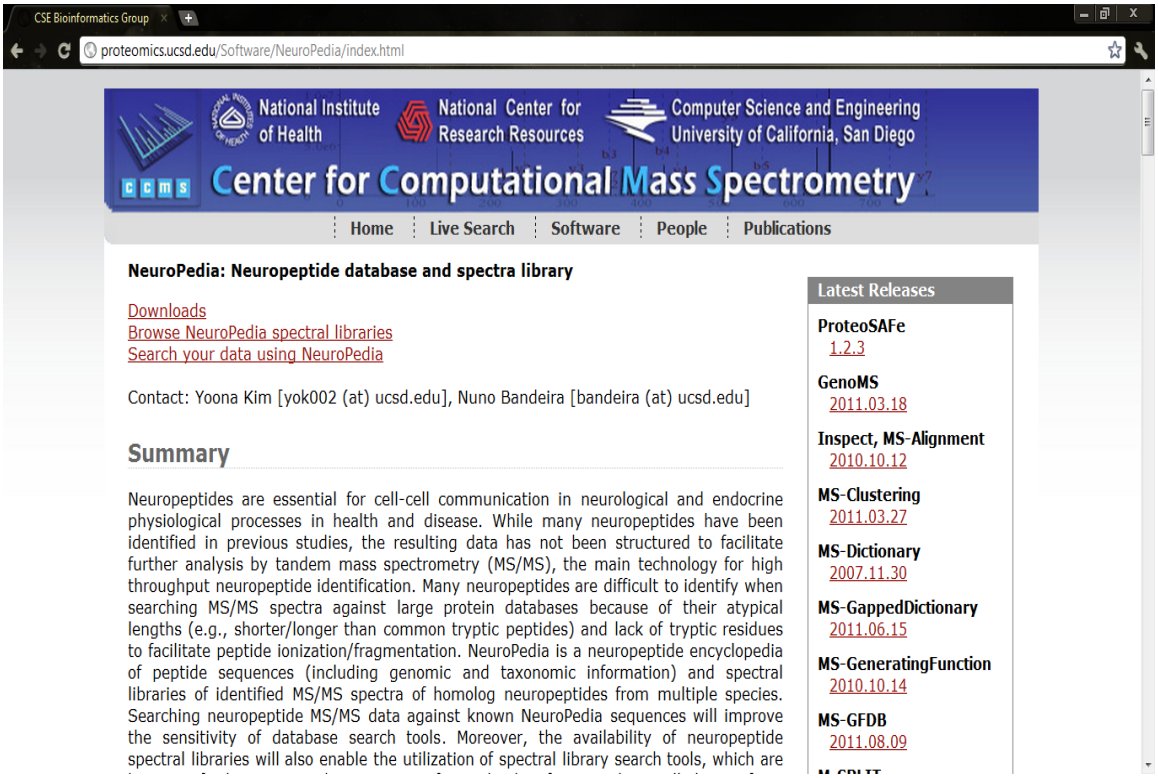



Figure 4.4: NeuroPedia web page screen shot.

# Chapter 5 CONCLUSIONS

NeuroPedia is a convenient and accessible repository of neuropeptide sequences and MS/MS spectral libraries. It offers advantages in terms of faster and more precise identification of small or nontryptic neuropeptides. We anticipate that NeuroPedia will continue to grow as data from more laboratories and experiments is contributed directly to NeuroPedia or otherwise becomes publicly available in mass spectrometry data repositories. In particular, it is expected that NeuroPedia will expand to include neuropeptide information for more species and mass spectrometry data of post-translationally modified neuropeptides. NeuroPedia can be accessed at http://proteomics.ucsd.edu/Software/NeuroPedia.html.

# REFERENCES


Andersen, JS., Svensson, B., Roepstorff, P. (1996) "Electrospray ionization and matrix assisted laser desorption/ionization mass spectrometry: powerful analytical tools in recombinant protein chemistry" Nature Biotechnology **14**: 449-57.

Bantscheff, M., Shirle, M., Sweetman, G., Rick, J., Kuster, B. (2007) "Quantitative mass spectrometry in proteomics: a critical review." Anal Bioanal Chem **389**(4):1017-31.

Benson, DA., Karsch-Mizrachi, I., Lipman, DJ., Ostell, J., Wheeler. DL. (2009). "GenBank." Nucleic Acids Res **37**: 23-7.

Benajmini, Y., and Hochberg, Y. (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing. " J R Stat Soc B Methodol **57**: 289-300.

Bora, A., Annanqudi, SP., Millet, LJ., Rubakhin, SS., Forbes, AJ., Kelleher, NL., Gillette, MU., Sweedler, JV. (2008) "Neuropeptidomics of the supraoptic rat nucleus." J Proteome Res **7**(11): 4992-5003.

Bouchard, P., and Bandeira, N. (2010) "SpecPlot – a versatile tool for spectrum visualization" The ASMS *processing*.

Bruand, J., Sistila S., Mériaux C, Dorrestein, PC., Gaasterland, T., Ghassemian, M., Wisztorski, M., Fournier, I., Salzet, M., Macagno, E., Bafna, V. (2011) "Automated Querying and Identification of Novel Peptides using MALDI Mass Spectrometric Imaging." J. Proteome Res **10**(4): 1915-28.

Burbach, JP. (2009) "Neuropeptides from concept to online database www.neuropeptides.nl." European Journal of Pharmacology **626**(1): 27-48.

Carlton, DD., and Schug, KA. (2011). "A review on the interrogation of peptide-metal interactions using electrospray ionization-mass spectrometry." Anal Chim Acta **686**(1-2): 19-39.

Craig, R., and Beavis, RC. (2004) "TANDEM: matching proteins with tandem mass spectra." Bioinformatics **20**(9): 1466-7.

Craig, R., Contens, JC., Fenyo, D., Beavis, RC. (2006) "Using annotated peptide mass spectrum libraries for protein identification." J Proteome Res **5**: 1843-9.

Elias, JE., and Gygi, SP. (2007). "Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry." Nat Methods **4**(3): 207-14

Eng, JK., McCormack AL., Yates, JR. (1994). "An approach to correlate tandem mass-spectral data of peptides with amino acid sequences in a protein database." J. Am. Soc. Mass Spectrometry **5**(11): 976-89.

Fälth, M., Sköld, F., Svensson, M., Nilsson, A., Fenyö, D., Andren, PE. (2007). "Neuropeptidomics Strategies for Specific and Sensitive Identification of Endogenous Peptides." Mol Cell Proteomics **6**(7): 1188-97.

Fricker, L. (2007). "Neuropeptidomics to study peptide processing in animal models of obesity." Endocrinology **148**(9): 4185-90.

Frewen, BE., Merrihew, GE,. Wu, CC., Noble, WS., MacCoss, MJ. (2006). "Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries." Anal Chem **78**: 5678-84.

Goldstein, DS., and Kopin, IG. (2008). "Adrenomedullary, adrenocortical, and sympathoneural responses to stressors: a meta-analysis." Endocr Regul **42**: 111-9.

Gupta, N., Bark, SJ., Lu, WD., Taupenot, L., O'Connor, DT., Pevzner, P., Hook, V. (2010). "Mass spectrometry-based neuropeptidomics of secretory vesicles from human adrenal medullary pheochromocytoma reveals novel peptide products of prohormone processing." J Proteome Res **9**(10): 5065-5075.

Hokfelt, T., Vartfai, T., Bloom, F. (2003). "Neuropeptides: opportunities for drug discovery." Lancet Neurol **2**(8): 463-72.

Hook, V., Toneff, T., Baylon, S., Sei, C. (2008). "Differential activation of enkephalin, galanin, somatostatin, NPY, and VIP neuropeptide production by stimulators of protein kinases A and C in neuroendocrine chromaffin cells." Neuropeptides **42**(5-6):503-11.

Hook, V., Bark, SJ., Gupta, N., Lortie, M., Lu, WD., Bandeira, N., Funkelstein, L., Wegrzyn, J., O'Connor, DT., Pevzner, P. (2010). "Neuropeptidomic Components Generated by Proteomic Functions in Secretory Vesicles for Cell–Cell Communication." The AAPS Journal **12**(4): 635-45.

Jain, E., Bairoch, A., Duvaud, S., Phan, I., Redaschi, N., Suzek, BE., Martin, MJ., McGarvey, P., Gasteiger, E. (2009). "Infrastructure for the life sciences: design and implementation of the UniProt website." BMC Bioinformatics **10**: 136.

Kapp, EA., Schutz, F., Connolly, LM., Chakel, JA., Meza, JE., Miller, CA., Fenyo, D., Eng, JK., Adkins, JN., Omenn, GS., Simpson, RJ. (2005) "An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis." Proteomics **5**(13): 3475-90.

Kastin, A. (2006) "Handbook of Biologically Active Peptides." Elsevier

Lam, H., Deutsch, EW., Eddes, JS., Eng, JK., King, N., Stein, SE., Aebersold, R. (2007). "Development and validation of a spectral library searching method for peptide identification from MS/MS." Proteomics **7**(5): 655-76.

Lam, H., Deutsch, EW., Eddes, JS., Eng, JK., Stein, SE., Aebersold, R. (2008). "Building consensus spectral libraries for peptide identification in proteomics." Nat Methods **5**(10): 873-5.

Lam, H., Deutsch, EW., Aebersold, R. (2009). "Artificial decoy spectral libraries for false discovery rate estimation in spectral library searching in proteomics." J Proteome Res **9**: 605-10.

Li, L. and Sweedler, JV. (2008). "Peptides in the brain: mass spectrometry-based measurement approaches and challenges." Annu Rev Anal Chem **1**: 451-483.

McEntyre, J., and Ostell, J. (2002). "The NCBI Handbook." National Center for Biotechnology Information (US).

Nankova, BB., and Sabban, EL. (1999). "Multiple signalling pathways exist in the stress-triggered regulation of gene expression for catecholamine biosynthetic enzymes and several neuropeptides in the rat adrenal medulla." Acta Physiol Scand **167**(1): 1-9.

Nesvizhskii, AI. (2007) "Protein identification by tandem mass spectrometry and sequence database searching." Methods Mol Biol **367**: 87-119.

Nesvizhskii, AI. (2010). "A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics." J Proteomics **73**(11): 2092-2123.

Perkins, DN., Pappin, DJ., Creasy, DM., Cottrell, JS. (1999). "Probability based protein identification by searching sequence databases using mass spectrometry data." Electrophoresis **20**(18): 3551-3567.

Robert, K. (1994). "Linked-scan techniques for MS/MS using tandem-in-space instruments." Mass Spectrometry Reviews **13**(5–6): 359-410.

Roepstorff, P. and Fohlman, J. (1984). "Proposal for a common nomenclature for sequence ions in mass spectra of peptides." Biomed Mass Spectrum **11**(11): 601.

Sayers, EW., Barrett, T., Benson, DA,, Bolton, E., Bryant, SH., Canese, K., Chetvernin, V., Church, DM., DiCucci, M., Federhen, S., Feolo, M., Fingerman, IM., Geer, LY., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, DJ., Lu, Z., Madden, TL., Madej, T., Maglott, DR., Marchler-Bauer, A., Miller, V., Mizrachi, I., Ostell, J., Panchenko, A., Phan, L., Pruitt, KD., Schuler, GD., Sequeira, E., Sherry, ST., Shumway, M., Sirotkin, K., Slotta, D., Souvorov, A., Starchenko, G., Tatusova, TA., Wagner, L., Wang, Y., Wilbur, WJ., Yaschenko, E., Ye, J. (2009). "Database resources of the National Center for Biotechnology Information." Nucleic Acids Res **37**: 5-15.

Schwartz, JC., Michael, W., Syka, JE. (2002). "A two-dimensional quadrupole ion trap mass spectrometer." J Am Soc Mass Spectrom **13**(6): 659-69.

Sleno, L., and Volmer, DA. (2004). "Ion activation methods for tandem mass spectrometry." J Mass Spectrom **39**(10): 1091-112.

Steen, H., Küster, B., Mann, M. (2001). "Quadrupole time-of-flight versus triple-quadrupole mass spectrometry for the determination of phosphopeptides by precursor ion scanning." J Mass Spectrom 36(7): 782–90.

Strand, FL. (2003) "Neuropeptides: general characteristics and neuropharmaceutical potential in treating CNS disorders." Prog Drug Res **61**: 1-37.

Svensson, M., Sköld, K., Nilsson, A., Fälth, M., Svenningsson, P., Andrén, PE. (2007). "Neuropeptidomics: expanding proteomics downwards." Biochem Soc Trans **35**(3): 588-593.

Svensson, M., Sköld, K., Nilsson A, Fälth, M., Nydahl, K., Svenningsson, P., Andrén, PE. (2007). "Neuropeptidomics: MS applied to the discovery of novel peptides from the brain." Anal Chem **79**(1):15-6, 18-21

Stein, SE. and Rudnick, PA. (2009). "NIST Peptide Tandem Mass Spectral Libraries. Human Pep-tide Mass Spectral Reference Data, H. sapiens, ion trap" Downloaded from http://peptide.nist.gov on July 25, 2011

Tanner, S., Shu, H., Frank, A., Wang, LC., Zandi, E., Mumby, M., Pevzner, PA., Bafna, V. (2005). "InsPecT: identification of posttranslationally modified peptides from tandem mass spectra." Analytical Chemistry **77**(14): 4626–4639.

Ubink, R., Calza, L., Hökfelt, T. (2003). "'Neuro'-peptides in glia: focus on NPY and galanin." Trends Neurosci **26**(11): 604-9

Wang, J., Perez-Santiago, J., Katz, JE., Mallick, P., Bandeira, N. (2010). "Peptide identification from mixture tandem mass spectra." Mol Cell Proteomics **9**(7): 1476-1485.

Wells, JM. and McLuckey, SA. (2005). "Collision-induced dissociation (CID) of peptides and proteins." Meth Enzymol **402**: 148-85.

Whitworth, EJ., Kosti, O., Renshaw, D., Hinson, JP. (2003). "Adrenal neuropeptides: regulation and interaction with ACTH and other adrenal regulators." Microsc Res Tech **61**(3): 259-67.

Yates, JR., Morgan, SF., Gatlin, CL, Griffin, PR, Eng, JK. (1998) "Method to compare collision-induced dissociation spectra of spectra of peptides: potential for library searching and subtractive analysis." Anal Chem **70**(17): 3557-65.

Yates, JR., Ruse, CI., Nakorchevsky, A. (2009) "Proteomics by mass spectrometry: approaches, advances, and applications." Annu Rev Biomed Eng **11**: 49-79.