Neuropeptidomics Strategies for Specific and Sensitive Identification of Endogenous Peptides*

Maria Fälth‡§, Karl Sköld‡§, Marcus Svensson‡§, Anna Nilsson‡§, David Fenyö¶, and Per E. Andren‡§‖

A new approach using targeted sequence collections has been developed for identifying endogenous peptides. This approach enables a fast, specific, and sensitive identification of endogenous peptides. Three different sequence collections were constituted in this study to mimic the peptidomic samples: SwePep precursors, SwePep peptides, and SwePep predicted. The searches for neuropeptides performed against these three sequence collections were compared with searches performed against the entire mouse proteome, which is commonly used to identify neuropeptides. These four sequence collections were searched with both Mascot and X! Tandem. Evaluation of the sequence collections was achieved using a set of manually identified and previously verified peptides. By using the three new sequence collections, which more accurately mimic the sample, 3 times as many peptides were significantly identified, with a false-positive rate below 1%, in comparison with the mouse proteome. The new sequence collections were also used to identify previously uncharacterized peptides from brain tissue; 27 previously uncharacterized peptides and potentially bioactive neuropeptides were identified. These novel peptides are cleaved from the peptide precursors at sites that are characteristic for prohormone convertases, and some of them have post-translational modifications that are characteristic for neuropeptides. The targeted protein sequence collections for different species are publicly available for download from SwePep. ***Molecular & Cellular Proteomics 6:1188–1197, 2007.***

Neuropeptidomics is the technology approach for detailed analysis of endogenous peptides from the brain and the central nervous system (1–6). In contrast to proteomics, which is focused on studying proteins (>10 kDa) and their interactions, peptidomics is focused on studying endogenous peptides (<10 kDa), such as peptide hormones and neuropeptides. Neuropeptides are involved in many physiological processes including pain, hunger, and growth (7). They often function as messengers, and some of them coexist with and complement the classical neurotransmitters (8).

MS is a powerful tool utilized for thorough analytical profiling of a large number of neuropeptides (5). The MS methodology in combination with either ESI (9, 10) or MALDI (11) permits sensitive detection of peptide changes in complex mixtures of hundreds of different peptides simultaneously (5). The resolution and specificity of a neuropeptide analysis is further enhanced by coupling MS to LC or other high resolution separation techniques.

Neuropeptidomics MS experiments, aimed at understanding the healthy and diseased mammalian brain, generate a large amount of data. To efficiently analyze these large datasets, reliable tools for automatic identification are needed. Such tools should be fast, yield few false peptide identifications (false positives), and leave few correct peptides unidentified (false negatives). So far, the main focus of the proteomics field has been on developing tools for identification of proteins, which are typically digested with trypsin, *i.e.* an enzyme with high specificity (12), limiting the search space of possible peptides. In contrast, endogenous peptide precursors are often processed by several enzymes (13), and some of these have unknown specificity, making it difficult to accurately predict the sequence of mature endogenous peptides. Therefore, when searching for endogenous peptides, the entire proteome is often cleaved assuming an enzyme with no specificity (*i.e.* cleaving between any pair of amino acids). This creates a very large search space and yields poor results because only peptides that have strong experimental support can be identified. In a typical peptidomics experiment many hundreds of peptides are detected (5), but about an order of magnitude less are identified confidently.

Many bioactive endogenous peptides are post-translationally modified, and it is common that a peptide contains more than one modification, further complicating the identification process. Important peptide modifications include acetylation, amidation, phosphorylation, and sulfation (7). Approximately 300 different modifications have so far been reported for proteins (14–17). For example, 30% of the mammalian proteins are believed to be phosphorylated at one time or another

From the ‡Laboratory for Biological and Medical Mass Spectrometry, Biomedical Centre, Box 583, Uppsala University, SE-75123 Uppsala, Sweden, §Department of Pharmaceutical Biosciences, Biomedical Centre, Uppsala University, SE-75124 Uppsala, Sweden, and ¶The Rockefeller University, New York, New York 10021

Received, January 18, 2007, and in revised form, March 28, 2007
Published, MCP Papers in Press, March 30, 2007, DOI 10.1074/mcp.M700016-MCP200

(18). The C-terminal amidation, a common neuropeptide modification, seems to modify 50% of all bioactive peptides (19, 20). Briefly the unknown specificity of the processing enzymes and the numbers of possible modifications make the identification of endogenous peptides difficult. Another difficulty stems from the less informative and inadequately understood fragmentation patterns for endogenous peptides compared with that of tryptic peptides.

The aim of this study was to investigate how to optimize the identification process for endogenous peptides analyzed by tandem mass spectrometry by improving the sequence collections used by the search engines. During this study, several previously uncharacterized peptides were discovered from mouse brain tissue. Some of these peptides are potential novel neuropeptides as they are processed from proteins, known to contain neuropeptides, at sites that are characteristic for neuropeptides. Identifying novel neuropeptides is important for the understanding of the biochemical processes in the mammalian brain. This study demonstrates the importance of using optimized sequence collections when identifying endogenous peptides.

EXPERIMENTAL PROCEDURES

Sequence Collections

SwePep is a database constructed for endogenous peptides and mass spectrometry (21). This is a relatively new database specifically designed to speed up the identification process of endogenous peptides from complex tissue samples utilizing mass spectrometry. To create sequence collections that mimic the mouse peptidome rather than the mouse proteome, sequence information about peptides and their precursors were extracted from SwePep (updated February 15, 2006, containing 4,180 non-redundant peptide sequences). Four sequence collections were used in this study: 1) SwePep precursors, 2) SwePep peptides, 3) SwePep predicted, and 4) mouse proteome. These sequence collections are available for download from www.swepep.org.

SwePep Precursor—The SwePep precursor sequence collection includes the sequences from the mouse peptide precursor proteins annotated in SwePep. Many precursor proteins, such as pro-opiomelanocortin, contain several known endogenous peptides (22) and a number of possible cleavage sites for endogenous peptides. Therefore this sequence collection should contain many of the endogenous peptides despite its moderate size of 123 protein sequences with a total number of 23,601 amino acid residues. Using unspecific cleavage and a maximum peptide length of 50 amino acid residues 4,406,615 peptides were derived from this sequence collection.

SwePep Peptides—The SwePep peptide sequence collection contains the sequences of the endogenous peptides annotated in SwePep from *Mus musculus*. It is constituted of 245 sequences and 6,776 amino acid residues. When using unspecific cleavage and a maximum peptide length of 50 amino acid residues this sequence collection generates 1,142,680 peptides.

SwePep Predicted—Endogenous neuropeptides are processed in many steps to become active peptides. Predominantly they are cleaved from their precursor at the C terminus of two basic amino acids, separated by 0, 2, 4, or 6 other residues, by endopeptidases such as prohormone convertase 1 (PC1/3)[1] and PC2 (13, 23). The basic residues at the C terminus are then removed by carboxypeptidase E (24). In the last step, the peptide may be modified. Important modifications on neuropeptides include C-terminal amidation and N-terminal acetylation (7).

By using the existing neuropeptide processing knowledge, possible peptide sequences were predicted from the mouse proteome (International Protein Index (IPI) mouse version 3.15, www.ebi.ac.uk/IPI/IPImouse.html) according to the following template: **(K/R)**X_m**(K/R)** $\downarrow X_k$**(K/R)**X_n**(K/R)** $\downarrow$ where m and n = 0, 2, 4, 6, X is any amino acid, and k = 3–50. Residues in bold signify amino acids that are not part of the final (detected) sequence. The C-terminal basic residues ($X_k \downarrow$ **(K/R)**X_n**(K/R)**) were removed, and the sequences X_k were stored in the SwePep predicted sequence collection.

It is possible to define digestion rules for the search engines so that the theoretical digest of the proteome is performed at dibasic sites on the fly, but the SwePep predicted sequence collection speeds up the search, and it can be curated to include special cases and peptides from more than one type of cleavage.

The SwePep predicted sequence collection was developed as a complement to the SwePep precursor and SwePep peptide sequence collections for identification of uncharacterized peptides and peptides from precursors not known to contain endogenous peptides. Peptides identified from the SwePep predicted sequence collection are likely to be biologically active because this collection only contains peptide sequences that have the specific cleavage pattern for neuropeptides. The SwePep predicted collection is constituted of precleaved sequences, and the searches are performed without any cleavage, *i.e.* the tandem mass spectra are directly matched against the sequences in the sequence collection. There are 3,413,034 predicted peptide sequences with 83,182,326 amino acid residues in this sequence collection. When using X! Tandem and its refinement function (25) this sequence collection generates 15,499,268 peptides with a maximum peptide length of 50 amino acid residues.

Mouse Proteome—To compare this new identification approach with the commonly used identification approach, a sequence collection constituted of the whole mouse proteome (IPI mouse version 3.15, www.ebi.ac.uk/IPI/IPImouse.html) was searched using unspecific cleavage. The sequence collection of the mouse proteome consists of 68,222 protein sequences with a total number of 27,668,712 amino acid residues. When using unspecific cleavage and a maximum peptide length of 50 amino acid residues 250,809,615 peptides are generated.

Search Engines

This study was performed using two different search engines, X! Tandem (26) and Mascot (27), for searching the four sequence collection s described above.

Search parameters were as follows. The SwePep peptides sequence collection, the SwePep precursor sequence collection, and mouse proteome sequence collection were searched using unspecific cleavage, and the precleaved SwePep predicted sequence collection was searched using no cleavage. The databases were searched using a peptide mass tolerance of ±2 Da and a fragment mass tolerance of ±0.7 Da. The first dataset was searched with a number of possible post-translational modifications (N-terminal acetylation, N-terminal pyroglutamic acid of glutamine, C-terminal amidation, deamidation of asparagine and glutamine, and oxidation of methionine). A full specification of search parameters is presented in the supplemental data. For X! Tandem the refinement function was used to allow unspecific cleavage of a precursor if one or more peptides have been identified from it (25).

[1] The abbreviation used is: PC, prohormone convertase.

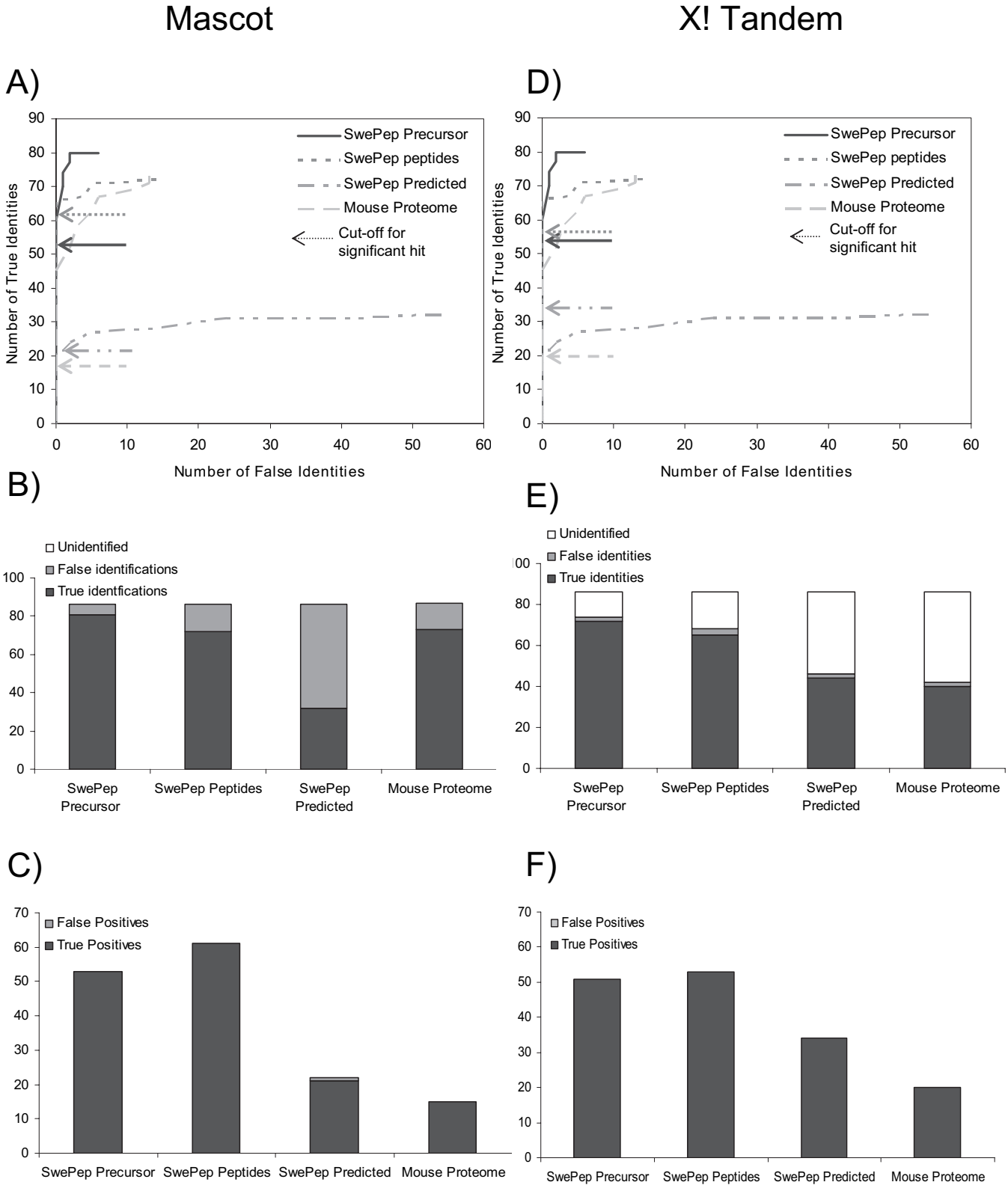


FIG. 1. *A* and *D* show the number of true and false identities of the 86 predetermined tandem mass spectra suggested by the search engines for the different databases. The *arrows* indicate the threshold for a significant hit suggested by the search engines for each database. *B* and *E* show the same thing but also the number of unidentified spectra. It should be noted that the number of false identities suggested by X! Tandem is much lower than for Mascot, but instead X! Tandem does not suggest any sequences for a number of spectra. *C* and *F* show the number of significantly (score over the threshold suggested by the search engines) identified peptides.

Mass Spectrometry Datasets

Two different MS datasets were used for searching the sequence collections. One set contained 86 tandem mass spectra with manually identified peptides in the mass range from 500 to 3500 Da and with charge states 1, 2, 3, or 4. All tandem mass spectra were manually evaluated, and the peptides were unambiguously identified. Because this dataset was manually composed of spectra with known identities it does not reflect a typical collection of tandem mass spectra from an LC-MS analysis of a peptidomic sample. Therefore, a second dataset was evaluated. This dataset was obtained by analyzing a peptidomic sample from mouse hypothalamus with nanoflow capillary LC-ESI-MS/MS and contained 2,867 tandem mass spectra.

Sample Preparation and Mass Spectrometry Analysis—The brain tissue was suspended in cold extraction solution (0.25% acetic acid) and homogenized by microtip sonication (Vibra cell 750, Sonics & Materials Inc., Newtown, CT) to a concentration of 0.2 mg of tissue/μl as described previously (4, 5). Briefly the suspension was centrifuged at 20,000 $\times$ *g* for 30 min at 4 °C. The protein- and peptide-containing supernatant was transferred to a centrifugal filter device (Microcon YM-10, Millipore, Bedford, MA) with a molecular mass limit of 10,000 Da and centrifuged at 14,000 $\times$ *g* for 45 min at 4 °C. Finally the peptide filtrate was frozen and stored at −80 °C until analysis.

Five microliters of peptide filtrate (equivalent to 1.0 mg of brain tissue) was desalted on a nano-precolumn (LC Packings, Amsterdam, The Netherlands) at 10 μl/min using a nano-LC system (Ettan MDLC, GE Healthcare). The filtrate was then separated using a fused silica capillary column (75-μm inner diameter, 15-cm length, NAN75-15-03-C18PM; LC Packings) by an isocratic flow of buffer A (0.25% acetic acid in water) for 35 min and eluted during a 60-min gradient from buffer A to B (35% acetonitrile in 0.25% acetic acid). The eluted peptides were analyzed by a linear trap quadrupole ion trap mass spectrometer (Thermo Electron, San Jose, CA). The spray voltage was 1.8 kV, the capillary temperature was 160 °C, and 35 units of collision energy were used to obtain fragment spectra. Four MS/MS spectra of the most intense peaks were obtained following each full-scan mass spectrum (Xcalibur 1.4 SR1). The dynamic exclusion feature was enabled to obtain MS/MS spectra on co-eluting peptides. Raw linear trap quadrupole data were converted to dta files by Xcalibur 1.4 SR1 and assembled by an in-house developed script to Mascot generic files.

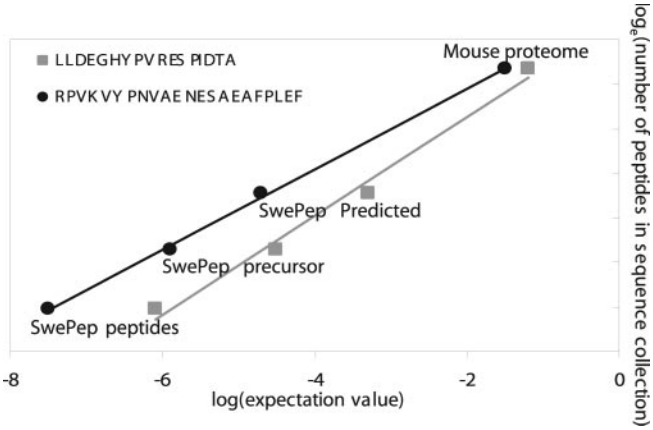


FIG. 2. **The logarithm of the expectation value as a function of the logarithm of the number of peptide sequences in the sequence collection for two of the peptides identified by X! Tandem.** To consider a hit as a significant hit the logarithm of the expectation value should be less than −2.

Verification of Search Results—An important step in the identification process is to verify the result from the search engines. One way to do this is to estimate the probability of false identifications (28, 29). This is often achieved by calculating an expectation value or by searching a decoy sequence collection, calculating the number of hits over a threshold, and dividing this number by the number of matches from the targeted sequence collection search (30, 31). A commonly used acceptance criteria for considering a protein to be identified is that at least two peptides have to be identified from that protein with scores over a calculated threshold. In contrast, each endogenously processed peptide may have a unique biological function, and it is therefore important to obtain sufficiently high quality data to trust the suggested identity. Also the thresholds selected for endogenous peptide identification have to be more stringent.

The false-positive rate for the first dataset in this study, containing known peptide identities, was calculated by dividing the number of false identified peptides with the number of true hits. For the second dataset the false-positive rate was estimated by searching the reversed sequence collections.

The threshold suggested by the search engine was used as the first verification step to evaluate the search result. All peptides with a score below the suggested threshold were manually verified or discarded. Secondly to increase the stringency the individual scores of the peptides were considered. After sorting out only the significant hits, the second best hit for each tandem mass spectrum was related to the best hit to confirm that the first (s_1) and second (s_2) best hits do not have a score that is too close to each other, *i.e.* $(s_1 - s_2) < 1$. If the scores for the first and second best hits were too close, manual inspection was used for determining the correct sequence.

RESULTS AND DISCUSSION

In the present report, we describe a methodological bioinformatics approach to detect and identify a large number of endogenous peptides. Many of the identified peptides represent previously uncharacterized and novel processed fragments of protein precursors. To identify more of the peptide

TABLE I
Summary of the advantages and disadvantages with the different sequence collections

	Advantage	Disadvantage
SwePep peptides	Contains all known peptides, fast search, many significant hits	Cannot identify novel peptides if they are not a part of an already known sequence
SwePep precursors	Contains more sequences than SwePep peptides but is still fast to search and gives a good search result	Only peptide precursors, cannot identify peptides from "novel" precursors
SwePep predicted	Contains all possible peptides with dibasic cleavage sites from the mouse proteome; good for identifying novel peptides	This should only be used for identifying novel peptides because it does not contain all already known peptides
Mouse proteome	Large sequence collection, possible to discover novel peptides	Gives low number of significant peptides identities, time-consuming to search with unspecific cleavage

TABLE II

Previously uncharacterized peptides from mouse brain tissue

All the peptides were identified using this new identification approach. The peptides in the table are identified from the sample in the second dataset or from similar samples. Mass is theoretical calculated in Da. See the supplemental data for a more detailed table and tandem mass spectra for all the peptides in the table. Residues in bold signify amino acids that are not part of the final (detected) sequence; * signifies a non-basic amino acid (not Lys or Arg).

UniProt accession no.	Precursor name	Sequence	Theoretical mass
			Da
O70176	Pituitary adenylate cyclase-activating polypeptide	**AR** ↓ GAGENLGGSAVDDPAPLT ↓ **KR**	1639.77
P01193	Corticotropin-lipotropin	**PR** ↓ KYVMGHF ↓ **R**R**	880.43
P01193	Corticotropin-lipotropin	**KR** ↓ ELEGERPLGLEQV ↓ **LE**	1467.76
P01193	Corticotropin-lipotropin	**EE** ↓ AVWGDGSPEPSPRE-amide ↓ **GKR**	1481.69
P01193	Corticotropin-lipotropin	**EA** ↓ VWGDGSPEPSPRE-amide ↓ **GKR**	1410.65
P12961	Neuroendocrine protein 7B2	**KK** ↓ acetyl-LLYEKMKGGQ ↓ **RR**	1207.63
P12961	Neuroendocrine protein 7B2	**KK** ↓ LLYEKMKGGQ ↓ **RR**	1165.62
P16014	Secretogranin-1	**KR** ↓ YPQSKWQEQE ↓ **KN**	1321.6
P16014	Secretogranin-1	**RPR** ↓ SEESQEREY ↓ **KR**	1155.47
P16014	Secretogranin-1	**R**R** ↓ DPADASGTRWASS ↓ **RE**	1319.57
P26339	Chromogranin A	**KR** ↓ LEGEDDPDRSMKLSF ↓ **R*R**	1737.79
P26339	Chromogranin A	**KR** ↓ LEGEDDPDRSM ↓ **K***R**	1262.51
P35455	Vasopressin neurophysin	**VQ** ↓ LAGTRESVDSAKPR ↓ **VY**	1485.79
P35455	Vasopressin neurophysin	**R**R** ↓ AREPSNATQLDGPA ↓ **R****R**	1425.68
P35455	Vasopressin neurophysin	**TR** ↓ ESVDSAKPRVY	1249.63
P41539	Protachykinin 1	**KR** ↓ DADSSVEKQVALLKALYGHGQIS ↓ **HKR**	2428.26
P47867	Secretogranin-3	**K****R** ↓ ELSAERPLNEQIAEAEADKI ↓ **KK**	2225.12
P47867	Secretogranin-3	**K****R** ↓ ELSAERPLNEQIAEAEAD ↓ **KIKK**	1983.94
P56388	Cocaine- and amphetamine-regulated transcript protein	**RA** ↓ pyro-Glu(Q)EDAELQPR ↓ **AL**	1067.49
Q03517	Secretogranin-2	**KR** ↓ IPVGSLKNEDTPNRQYLDEDM ↓ **LL**	2433.15
Q03517	Secretogranin-2	**KR** ↓ IPVGSLKNEDTPN ↓ **RQ**	1382.70
Q03517	Secretogranin-2	**KR** ↓ SGQLGLPDEEN ↓ **RR**	1157.52
Q03517	Secretogranin-2	**KR** ↓ TNEIVEEQYTPQSL ↓ **AT**	1649.78
Q03517	Secretogranin-2	**SVF** ↓ pyro-Glu(Q)ELGKLTGPSNQ ↓ **KR**[a]	1253.63
Q62361	Thyrotropin-releasing hormone	**KR** ↓ EEKEEDVEAEERGDLGEVGAWRPH ↓ **KR**	2765.25
Q9QXV0	Pro-SAAS	**RR** ↓ SVDQDLGPEVPPENVL-amide ↓ **GA**	1705.85
Q9QXV0	Pro-SAAS	**RA** ↓ WGSPRASDPPLAPDDDPDAPAAQLARAL ↓ **LR**	2869.36

[a] Novel modification; sequence identified previously by Che *et al.* (2).

content in mouse brain tissue, there are both experimental and bioinformatics requirements that have to be fulfilled. An automatic identification process is established for endogenous peptides that has high specificity and sensitivity. To accomplish this, the samples are searched against sequence collections specifically designed to better mimic the content of the samples. The sample preparation was performed to prevent protein degradation (4, 5) by fast inactivation of the enzymes in the tissue (5, 32) to maximize the endogenous peptide content of the sample. Because the sample predominantly contains endogenous peptides so should the sequence collections. In this study, the four different sequence collections were evaluated using two search engines, Mascot and X! Tandem. The in-house constituted sequence collections were evaluated by searching two different datasets against them.

Manually Identified Brain Peptides (Dataset 1)—Dataset 1 constituted 86 tandem mass spectra. Fig. 1 shows the search result obtained by searching Dataset 1 against the different sequences collections. The highest numbers of identified

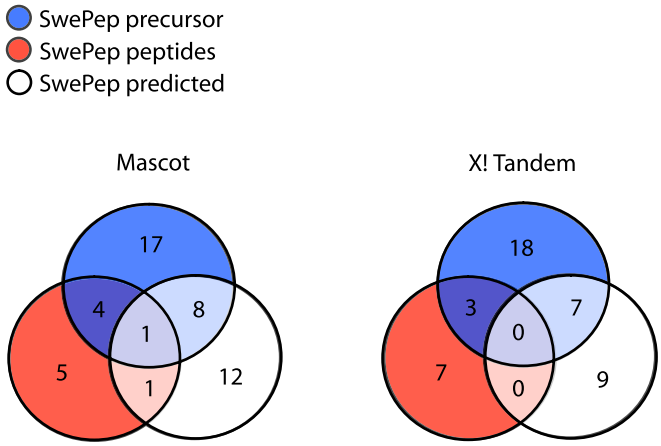


FIG. 3. **The number of uncharacterized peptides in a tissue sample of mouse hypothalamus identified from the different sequence collection using Mascot and X! Tandem.** Many of the peptides are identified from more than one sequence collection; this shows once again that the sequence collection mimics the samples well. Some of the peptides are only identified from the predicted sequence collection, showing the importance of using it as well.

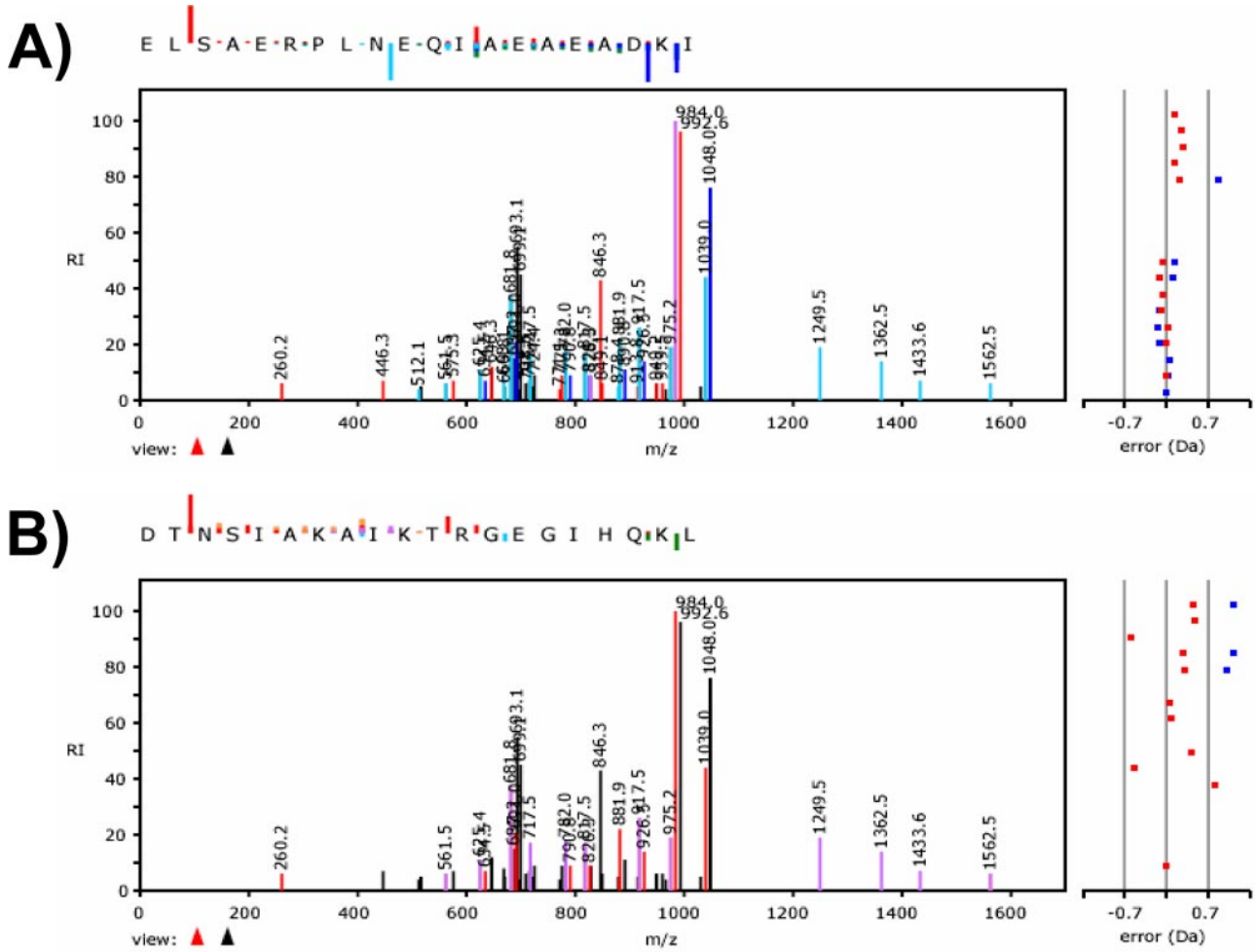
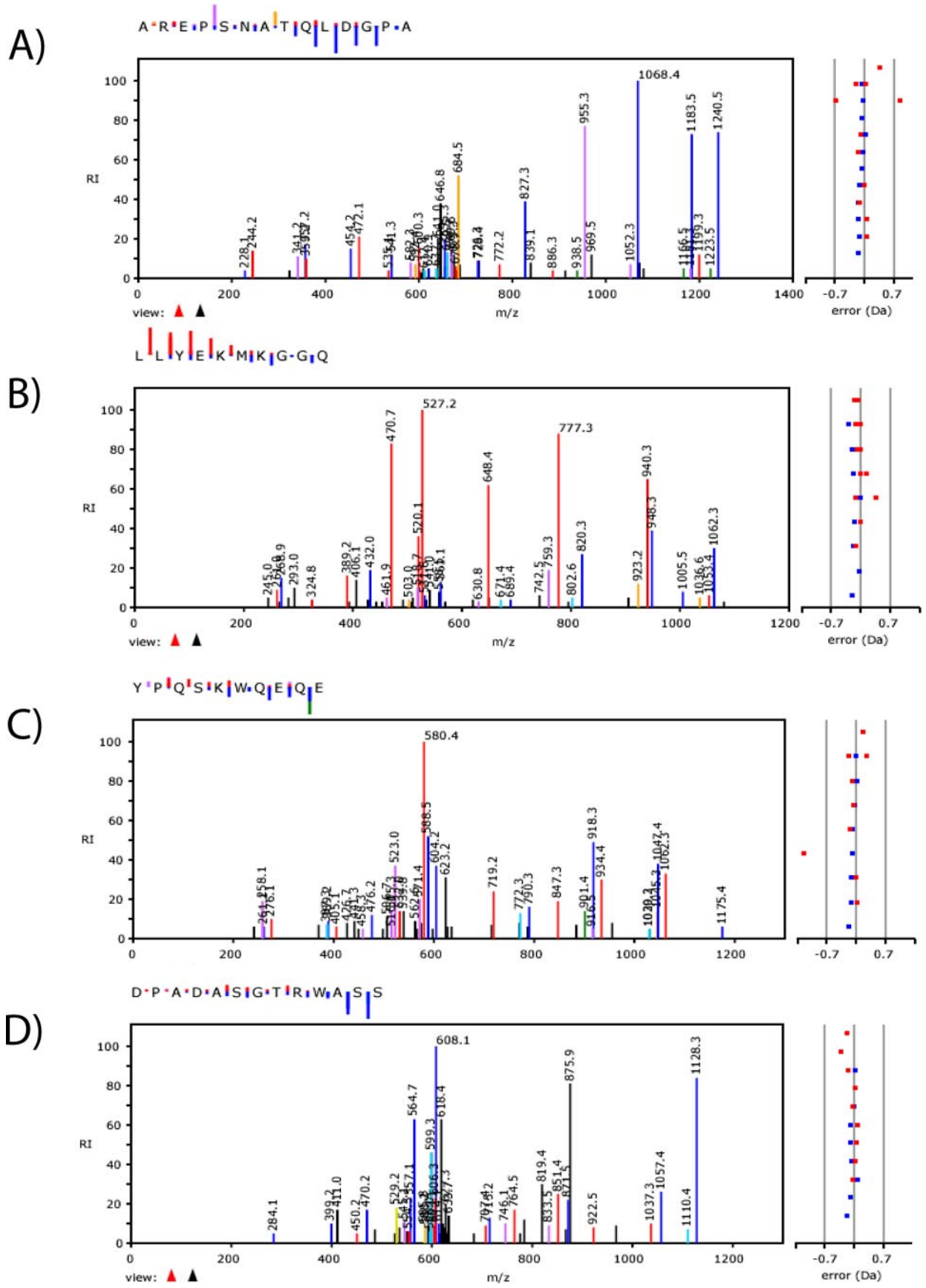


FIG. 4. **Correlation of tandem mass spectra with the fragmentation for the sequence ELSAERPLNEQIAEAEADKI (*A*) and the sequence acetyl-DTNSIAKAIKTRGEGIHQKL+2 deamidation(NQ) (*B*).** The error for the fragment ions shows a consistent pattern for the first sequence, whereas for the second sequence pattern of the fragment the ion error is inconsistent. This implies that the first sequence is most likely to be the correct one.

peptides were obtained when searching the SwePep peptide sequence collection. The least number of peptides were identified searching the mouse proteome. This is due to the fact that the cutoff scores for a significant identification increase with the size of the sequence collection because the possibility that the hit is random increases (33). For example, the cutoff score for a significant identification in Mascot increases from 31 to 70 when going from the smallest to the largest sequence collections (5% false-positive rate). Fig. 2 shows the logarithm of the expectation value as a function of the logarithm of the number of peptide sequences in the sequence collection for a few peptides identified by X! Tandem. This means that the number of false negatives will increase as the size of the sequence collection increases. For Dataset 1, the only false identity was suggested by Mascot when the SwePep predicted sequence collection was searched. When searching these more targeted sequences collections, 3 times as many peptides were identified compared with searching the mouse proteome. The search of the mouse proteome did not contribute any identities that were not identified when searching the more targeted sequence collections. Another drawback with searching large sequence collections using unspecific cleavage is that it is time-consuming, especially if the search includes a number of different post-translational modifications.

Not all of the tandem mass spectra in Dataset 1 were positively identified in this study. This might depend on the fact that many of the algorithms for identifying peptides and proteins are designed for tryptic peptides and that the fragmentation pattern of endogenous peptides differs from the fragmentation pattern of peptides digested with trypsin. Tabb *et al.* (34) studied the fragmentation of peptides generated by proteinase K digestion to investigate how the peak intensity for b- and y-ions depends on the position of the basic residue. The study demonstrated that the position of basic amino acids strongly influences the fragmentation pattern of the



peptide. When the basic residue was positioned near the N terminus of the peptides, the b-ion series was more prominent than the y-ion series, and when the basic residue was positioned close to the C terminus of the peptide, the y-ion series dominated. Peptides derived from trypsin digestion generally have a single basic residue, arginine or lysine, at the C terminus of the peptide, and therefore a high intensity y-ion series is often observed (unless the peptides contain missed cleavage sites and have more than one basic residue).

The sequence collections all generated somewhat different results in the present study. Table I summarizes the advantages and disadvantages for the different sequence collections. Each sequence collection, except the one of the mouse proteome, generated sequence collection-specific peptide identities; although they existed in the other sequence collections, they were not significantly identified. For more information about Dataset 1 see Table 1 in the supplemental data.

Hypothalamic Brain Sample (Dataset 2)—Dataset 2 contained tandem mass spectra from a tissue sample from mouse hypothalamus, which was analyzed using nano-LC-ESI-MS/MS. The aim of this study was to investigate the possibility to identify uncharacterized peptides using the more targeted sequence collections.

When using Dataset 2 to search against the three different collections of sequences 85 peptides were identified. Some of the peptides were well characterized neuropeptides such as melanotropin α, little SAAS, WE-14, and Met-enkephalin-Arg-Phe. Others were fragments of characterized neuropeptides as well as previously uncharacterized peptides. A total of 27 uncharacterized peptides were identified (Table II). See the supplemental data for more detailed information and tandem mass spectra for the peptides. Many of these peptides are processed from known peptide precursors at sites that correspond to the cleavage sites of the proprotein convertases PC1/3 and PC2. Fig. 3 shows in which sequence collection the peptides were identified. There are overlaps between the sequence collections; however, some of the peptides were only identified in one of the sequence collections, showing the importance of using all of them. Some of these previously uncharacterized peptides have the potential to be novel biologically active neuropeptides. From a single peptidomics experiment it is not possible to determine whether the observed peptides are endogenous or degradation products. However, in a time course study where the peptide levels are measured as a function of postmortem time to the first order approximation, the level of endogenous peptides will decrease and the level of degradation products will increase as the postmortem time increases, *i.e.* the dynamics of the changes in peptide level can be used to give an indication whether an observed peptide is endogenous or a degradation product.

The false-positive rate was calculated using the reversed database approach and was below 1% for all the sequence collections except for the SwePep peptide collection where the false-positive rate was 2% when using Mascot and 4% using X! Tandem. When the sequence collections were evaluated using Dataset 1, it was shown that the mouse proteome did not contribute any significantly identified peptides other than the peptides identified from the other sequence collections. Because the searches against the mouse proteome are time-consuming and the high thresholds yield few significantly identified peptides, the sequence collection containing the mouse proteome is instead used for verification of the uncharacterized peptides. Their tandem mass spectra are searched against the mouse proteome to verify that they are the primary hit even when the spectra are searched against a larger sequence collection.

Some of the uncharacterized peptides had to be manually validated because they did not fulfill the above stated criteria for the identification process of endogenous peptides. One of the peptides was not suggested as a primary hit when verifying the search result by searching the mouse proteome using Mascot. The peptide suggested from the searches of the more targeted sequences collections was pyro-Glu(Q)E-DAELQPR. When validating the search result by searching the mouse proteome using Mascot the primary hit for this tandem mass spectrum was KQPASQAIPQdeamide-amide (data not shown). This peptide sequence suggested by Mascot is likely to be incorrect. First the peptide sequence suggested by Mascot has a basic residue, lysine, at the N terminus. This would imply that the b-ion series should be the most prominent, but when examining the tandem mass spectra the most intense peaks are assigned as y-ions. Secondly the tandem mass spectra of the sequence pyro-Glu(Q)EDAELQPR showed poor fragmentation. Because this peptide is singly charged, the C-terminal residue is an arginine, and the sequence contains an aspartic acid, it is likely that there will be "charge remote fragmentation" (35). It will generate a tandem mass spectrum that will have the most intense peak between aspartic acid and alanine. The most abundant ion series should be the y-ion series because the basic residue is positioned at the C terminus. Taken together, the peptide sequence suggested by Mascot using the mouse proteome sequence collection is most likely a false positive.

This explanation is in line with the work published by Kapp *et al.* (36) where the effect of proton mobility on patterns in peptide fragmentation was investigated. It was noted that peptides without a mobile proton often receive low scores from search engines. It was also suggested that the use of an

Fig. 5. **Tandem mass spectra for potential active peptides.** *A*, **R**R** ↓ AREPSNATQLDGPA ↓ **R****R** derived from vasopressin-neurophysin 2-copeptin precursor (P35455). *B*, **KK** ↓ acetyl-LLYEKMKGGQ ↓ **RR** from 7B2 (P12961). *C* and *D*, **KR** ↓ YPQSKWQEQE ↓ **KN** (*C*) and **R**R** ↓ DPADASGTRWASS ↓ **RE** (*D*), both processed from secretogranin-1 (P16014).

additional proton mobility-based scoring would compensate for this effect. An automatic implementation of such proton mobility scoring would be of great value for the identification of endogenous peptides.

Another example of when manual inspection of the Mascot search result is needed was when the difference between the scores for the primary and secondary hits was too close (*i.e.* <1). The primary hit corresponded to the sequence acetyl-DTNSIAKAIKTRGEGIHQKL+2 deamidation as the most likely match for the tandem mass spectrum. The secondary hit corresponded to the sequence ELSAERPLNEQIAEAEADKI. The manual inspection of the expected fragmentation patterns showed that the secondary hit was presumably the right one (Fig. 4). The sequence ELSAERPLNEQIAEAEADKI has two basic residues; according to the study performed by Tabb *et al.* (34) arginine is the dominant residue, and it is the position of arginine that decides which type of ions that will be most abundant. In this case the b-ions should be the most intense peaks, and indeed 16 of the most 21 intense peaks in the tandem mass spectrum could be assigned as b-ions.

Potential Bioactive Endogenous Peptides—Some of the uncharacterized peptides discovered in this study are likely to have biological functions. One example is the peptide **R**R** ↓ AREPSNATQLDGPA ↓ **R****R** where ***** signifies a non-basic amino acid (not Lys or Arg) (Fig. 5*A*) from vasopressin-neurophysin 2-copeptin precursor (P35455; all accession numbers cited are from UniProt). The peptide is processed from the precursor protein at sites that are specific for neuropeptides (20, 24, 37), and it is the N-terminal part of the known biologically active peptide copeptin. The precursor also contains the bioactive peptide Arg-vasopressin; many precursor proteins contain more than one bioactive peptide (38). Another example is the peptide **KK** ↓ acetyl-LLYEKMKGGQ ↓ **RR** (Fig. 5*B*) from 7B2 (P12961). 7B2 was first discovered in 1982 (39, 40); it is a bifunctional polypeptide that is highly conserved in multiple species (41–46). The N-terminal peptide of the polypeptide binds to PC2 and is essential for its activation; in contrast the C-terminal peptide functions as its inhibitor (47, 48). The peptide identified in this study is the C-terminal part of the N-terminal peptide; it has both characteristic cleavage sites and has been identified both with and without an N-terminal acetylation. Two peptides, **KR** ↓ YPQSKWQEQE ↓ **KN** (Fig. 5*C*) and **R**R** ↓ DPADASGTRWASS ↓ **RE** (Fig. 5*D*), from secretogranin-1 (P16014) were also identified in this study. Both of the peptides have characteristic cleavage sites at the N terminus but not on the C terminus. Many peptides derived from secretogranin-1 have been discovered in mouse and rat brain tissue during the last couple of years (1, 2, 5, 32, 49, 50).

In conclusion, a new approach for identification of endogenous peptides has been developed. This new approach uses sequence collections created from the SwePep database. These sequence collections mimic the peptidomic samples better than nonspecific digestion of the entire mouse proteome and yield more and higher confidence peptide identifications. This study showed that 3 times as many peptides were significantly identified from the sequence collections created from the SwePep database and the predicted sequence collection together than from the mouse proteome sequence collection. This new approach was also successfully applied to identifying uncharacterized potentially novel bioactive peptides. The described methodology will benefit from using high mass accuracy instruments like orbitrap (51–53), FTMS (54), and time of flight (55) and will benefit greatly from the use of complementary fragmentation techniques such as electron capture dissociation (56, 57) and electron transfer dissociation (58).

* This study was supported by Swedish Research Council (VR) Grant 11565, 2004-3417; the Swedish Foundation for International Cooperation in Research and Higher Education (STINT) institutional grant; the K&A Wallenberg Foundation; and the Karolinska Institutet Centre for Medical Innovations, Research Program in Medical Bioinformatics. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

S The on-line version of this article (available at http://www.mcponline.org) contains supplemental material.

‖ To whom correspondence should be addressed: Laboratory for Biological and Medical Mass Spectrometry, Uppsala University, Box 583 Biomedical Centre, SE-75123 Uppsala, Sweden. Tel.: 46-18-471-7206; Fax: 46-18-471-4422; E-mail: per.andren@bmms.uu.se.

REFERENCES

1. Svensson, M., Sköld, K., Nilsson, A., Fälth, M., Nydahl, K., Svenningsson, P., and Andrén, P. (2007) Neuropeptidomics: MS applied to the discovery of novel peptides from the brain. *Anal. Chem.* **79,** 14–21
2. Che, F. Y., Lim, J., Pan, H., Biswas, R., and Fricker, L. D. (2005) Quantitative neuropeptidomics of microwave-irradiated mouse brain and pituitary. *Mol. Cell. Proteomics* **4,** 1391–1405
3. Clynen, E., Baggerman, G., Veelaert, D., Cerstiaens, A., Van der Horst, D., Harthoorn, L., Derua, R., Waelkens, E., De Loof, A., and Schoofs, L. (2001) Peptidomics of the pars intercerebralis-corpus cardiacum complex of the migratory locust, Locusta migratoria. *Eur. J. Biochem.* **268,** 1929–1939
4. Skold, K., Svensson, M., Kaplan, A., Bjorkesten, L., Astrom, J., and Andren, P. E. (2002) A neuroproteomic approach to targeting neuropeptides in the brain. *Proteomics* **2,** 447–454
5. Svensson, M., Skold, K., Svenningsson, P., and Andren, P. E. (2003) Peptidomics-based discovery of novel neuropeptides. *J. Proteome Res.* **2,** 213–219
6. Verhaert, P., Uttenweiler-Joseph, S., de Vries, M., Loboda, A., Ens, W., and Standing, K. G. (2001) Matrix-assisted laser desorption/ionization quadrupole time-of-flight mass spectrometry: an elegant tool for peptidomics. *Proteomics* **1,** 118–131
7. Hokfelt, T., Broberger, C., Xu, Z. Q., Sergeyev, V., Ubink, R., and Diez, M. (2000) Neuropeptides—an overview. *Neuropharmacology* **39,** 1337–1356
8. Hokfelt, T., Millhorn, D., Seroogy, K., Tsuruo, Y., Ceccatelli, S., Lindh, B., Meister, B., Melander, T., Schalling, M., Bartfai, T., and Terenius, L. (1987) Coexistence of peptides with classical neurotransmitters. *Experientia* **43,** 768–780
9. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., and Whitehouse, C. M. (1989) Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246,** 64–71
10. Yamashita, M., and Fenn, J. B. (1984) Electrospray ion source. Another variation on the free-jet theme. *J. Phys. Chem.* **88,** 4451–4459
11. Karas, M., and Hillenkamp, F. (1988) Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal. Chem.* **60,**

2299–2301
12. Olsen, J. V., Ong, S. E., and Mann, M. (2004) Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Mol. Cell. Proteomics* **3,** 608–614
13. Steiner, D. F. (1998) The proprotein convertases. *Curr. Opin. Chem. Biol.* **2,** 31–39
14. Jensen, O. N. (2004) Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. *Curr. Opin. Chem. Biol.* **8,** 33–41
15. Jensen, O. N. (2006) Interpreting the protein language using proteomics. *Nat. Rev.* **7,** 391–403
16. Larsen, M. R., Trelle, M. B., Thingholm, T. E., and Jensen, O. N. (2006) Analysis of posttranslational modifications of proteins by tandem mass spectrometry. *BioTechniques* **40,** 790–798
17. Mann, M., and Jensen, O. N. (2003) Proteomic analysis of post-translational modifications. *Nat. Biotechnol.* **21,** 255–261
18. Mann, M., Ong, S. E., Gronborg, M., Steen, H., Jensen, O. N., and Pandey, A. (2002) Analysis of protein phosphorylation using mass spectrometry: deciphering the phosphoproteome. *Trends Biotechnol.* **20,** 261–268
19. Eipper, B. A., Milgram, S. L., Husten, E. J., Yun, H. Y., and Mains, R. E. (1993) Peptidylglycine α-amidating monooxygenase: a multifunctional protein with catalytic, processing, and routing domains. *Protein Sci.* **2,** 489–497
20. Fricker, L. D. (2005) Neuropeptide-processing enzymes: applications for drug discovery. *AAPS J.* **7,** E449–E455
21. Falth, M., Skold, K., Norrman, M., Svensson, M., Fenyo, D., and Andren, P. E. (2006) SwePep, a database designed for endogenous peptides and mass spectrometry. *Mol. Cell. Proteomics* **5,** 998–1005
22. Wilkinson, C. W. (2006) Roles of acetylation and other post-translational modifications in melanocortin function and interactions with endorphins. *Peptides* **27,** 453–471
23. Seidah, N. G., and Chretien, M. (1997) Eukaryotic protein processing: endoproteolysis of precursor proteins. *Curr. Opin. Biotechnol.* **8,** 602–607
24. Zhou, A., Webb, G., Zhu, X., and Steiner, D. F. (1999) Proteolytic processing in the secretory pathway. *J. Biol. Chem.* **274,** 20745–20748
25. Craig, R., and Beavis, R. C. (2003) A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Commun. Mass Spectrom.* **17,** 2310–2316
26. Craig, R., and Beavis, R. C. (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* (*Oxf.*) **20,** 1466–1467
27. Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20,** 3551–3567
28. Eriksson, J., Chait, B. T., and Fenyo, D. (2000) A statistical basis for testing the significance of mass spectrometric protein identification results. *Anal. Chem.* **72,** 999–1005
29. Fenyo, D., and Beavis, R. C. (2003) A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal. Chem.* **75,** 768–774
30. Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J., and Gygi, S. P. (2003) Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J. Proteome Res.* **2,** 43–50
31. Moore, R. E., Young, M. K., and Lee, T. D. (2002) Qscore: an algorithm for evaluating SEQUEST database search results. *J. Am. Soc. Mass Spectrom.* **13,** 378–386
32. Dowell, J. A., Heyden, W. V., and Li, L. (2006) Rat neuropeptidomics by LC-MS/MS and MALDI-FTMS: Enhanced dissection and extraction techniques coupled with 2D RP-RP HPLC. *J. Proteome Res.* **5,** 3368–3375
33. Eriksson, J., and Fenyo, D. (2004) The statistical significance of protein identification results as a function of the number of protein sequences searched. *J. Proteome Res.* **3,** 979–982
34. Tabb, D. L., Huang, Y., Wysocki, V. H., and Yates, J. R., III (2004) Influence of basic residue content on fragment ion peak intensities in low-energy collision-induced dissociation spectra of peptides. *Anal. Chem.* **76,** 1243–1248
35. Wysocki, V. H., Resing, K. A., Zhang, Q., and Cheng, G. (2005) Mass spectrometry of peptides and proteins. *Methods* (*San Diego*) **35,** 211–222
36. Kapp, E. A., Schutz, F., Reid, G. E., Eddes, J. S., Moritz, R. L., O'Hair, R. A., Speed, T. P., and Simpson, R. J. (2003) Mining a tandem mass spectrometry database to determine the trends and global factors influencing peptide fragmentation. *Anal. Chem.* **75,** 6251–6264
37. Seidah, N. G., and Prat, A. (2002) Precursor convertases in the secretory pathway, cytosol and extracellular milieu. *Essays Biochem.* **38,** 79–94
38. Dixon, J. E., Haun, R. S., Minth, C. D., and Nichols, R. (1987) Neuropeptide gene families, in *Neuropeptides and Their Peptidases* (Turner, A. J., ed) pp. 9–30, Weinheim, New York
39. Hsi, K. L., Seidah, N. G., De Serres, G., and Chretien, M. (1982) Isolation and NH2-terminal sequence of a novel porcine anterior pituitary polypeptide. Homology to proinsulin, secretin and Rous sarcoma virus transforming protein TVFV60. *FEBS Lett.* **147,** 261–266
40. Seidah, N. G., Hsi, K. L., De Serres, G., Rochemont, J., Hamelin, J., Antakly, T., Cantin, M., and Chretien, M. (1983) Isolation and NH_2-terminal sequence of a highly conserved human and porcine pituitary protein belonging to a new superfamily. Immunocytochemical localization in pars distalis and pars nervosa of the pituitary and in the supraoptic nucleus of the hypothalamus. *Arch. Biochem. Biophys.* **225,** 525–534
41. Lindberg, I., Tu, B., Muller, L., and Dickerson, I. M. (1998) Cloning and functional analysis of C. elegans 7B2. *DNA Cell Biol.* **17,** 727–734
42. Martens, G. J. (1988) Cloning and sequence analysis of human pituitary cDNA encoding the novel polypeptide 7B2. *FEBS Lett.* **234,** 160–164
43. Martens, G. J., Bussemakers, M. J., Ayoubi, T. A., and Jenks, B. G. (1989) The novel pituitary polypeptide 7B2 is a highly-conserved protein coexpressed with proopiomelanocortin. *Eur. J. Biochem.* **181,** 75–79
44. Mbikay, M., Grant, S. G., Sirois, F., Tadros, H., Skowronski, J., Lazure, C., Seidah, N. G., Hanahan, D., and Chretien, M. (1989) cDNA sequence of neuroendocrine protein 7B2 expressed in beta cell tumors of transgenic mice. *Int. J. Pept. Protein Res.* **33,** 39–45
45. Spijker, S., Smit, A. B., Martens, G. J., and Geraerts, W. P. (1997) Identification of a molluscan homologue of the neuroendocrine polypeptide 7B2. *J. Biol. Chem.* **272,** 4116–4120
46. Waldbieser, G. C., Aimi, J., and Dixon, J. E. (1991) Cloning and characterization of the rat complementary deoxyribonucleic acid and gene encoding the neuroendocrine peptide 7B2. *Endocrinology* **128,** 3228–3236
47. Fugere, M., and Day, R. (2002) Inhibitors of the subtilase-like pro-protein convertases (SPCs). *Curr. Pharm. Des.* **8,** 549–562
48. Martens, G. J., Braks, J. A., Eib, D. W., Zhou, Y., and Lindberg, I. (1994) The neuroendocrine polypeptide 7B2 is an endogenous inhibitor of prohormone convertase PC2. *Proc. Natl. Acad. Sci. U. S. A.* **91,** 5784–5787
49. Lim, J., Berezniuk, I., Che, F. Y., Parikh, R., Biswas, R., Pan, H., and Fricker, L. D. (2006) Altered neuropeptide processing in prefrontal cortex of Cpe (fat/fat) mice: implications for neuropeptide discovery. *J. Neurochem.* **96,** 1169–1181
50. Parkin, M. C., Wei, H., O'Callaghan, J. P., and Kennedy, R. T. (2005) Sample-dependent effects on the neuropeptidome detected in rat brain tissue preparations by capillary liquid chromatography with tandem mass spectrometry. *Anal. Chem.* **77,** 6331–6338
51. Hardman, M., and Makarov, A. A. (2003) Interfacing the orbitrap mass analyzer to an electrospray ion source. *Anal. Chem.* **75,** 1699–1705
52. Makarov, A. (2000) Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Anal. Chem.* **72,** 1156–1162
53. Olsen, J. V., de Godoy, L. M., Li, G., Macek, B., Mortensen, P., Pesch, R., Makarov, A., Lange, O., Horning, S., and Mann, M. (2005) Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. *Mol. Cell. Proteomics* **4,** 2010–2021
54. Marshall, A. G., Hendrickson, C. L., and Shi, S. D. (2002) Scaling MS plateaus with high-resolution FT-ICRMS. *Anal. Chem.* **74,** 252A–259A
55. Chernushevich, I. V., Loboda, A. V., and Thomson, B. A. (2001) An introduction to quadrupole-time-of-flight mass spectrometry. *J. Mass Spectrom.* **36,** 849–865
56. Nielsen, M. L., Savitski, M. M., and Zubarev, R. A. (2005) Improving protein identification using complementary fragmentation techniques in Fourier transform mass spectrometry. *Mol. Cell. Proteomics* **4,** 835–845
57. Zubarev, R. A., Kelleher, N. L., and McLafferty, F. W. (1998) Electron capture dissociation of multiply charged protein cations. A nonergodic process. *J. Am. Chem. Soc.* **120,** 3265–3266
58. Syka, J. E., Coon, J. J., Schroeder, M. J., Shabanowitz, J., and Hunt, D. F. (2004) Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 9528–9533