technical notes Journal of proteome research

Validation of Endogenous Peptide Identifications Using a Database of Tandem Mass Spectra

Maria Fälth,[†] Marcus Svensson,[†] Anna Nilsson,[†] Karl Sköld,[†] David Fenyö,[‡] and Per E. Andren*,[†]

Department of Pharmaceutical Biosciences, Medical Mass Spectrometry, Biomedical Centre, Box 583, Uppsala University, SE-75123 Uppsala, Sweden, and The Rockefeller University, 1230 York Avenue, New York, New York 10065

Received January 16, 2008

Abstract: The SwePep database is designed for endogenous peptides and mass spectrometry. It contains information about the peptides such as mass, p*I*, precursor protein and potential post-translational modifications. Here, we have improved and extended the SwePep database with tandem mass spectra, by adding a locally curated version of the global proteome machine database (GPMDB). In peptidomic experiment practice, many peptide sequences contain multiple tandem mass spectra with different quality. The new tandem mass spectra database in SwePep enables validation of low quality spectra using high quality tandem mass spectra. The validation is performed by comparing the fragmentation patterns of the two spectra using algorithms for calculating the correlation coefficient between the spectra. The present study is the first step in developing a tandem spectrum database for endogenous peptides that can be used for spectrum-to-spectrum identifications instead of peptide identifications using traditional protein sequence database searches.

Keywords: bioinformatics • neuropeptides • peptidomics • peptide identification • MS/MS database

Introduction

The recent advances in both mass spectrometers and assisting software have made the sequencing and identification of proteins, using enzymatically cleaved peptides, somewhat straightforward. However, the identification of endogenous peptides still remains a difficult and time-consuming task, because endogenous peptide precursors are often processed in several steps by different enzymes.[1] Some of these enzymes have unknown specificity, making it difficult to accurately predict the sequences of mature endogenous peptides. Therefore, when searching for endogenous peptides using tandem mass spectrometry data, 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, decreasing the sensitivity of identification, and peptides can only be identified when there is strong experimental evidence. In a typical peptidomics experiment, many hundreds of peptides are detected,[2] but only an order of magnitude less are identified confidently.

Many peptides are identified over and over again from different experiments by searching experimental spectra against sequence collections with search engines such as X! Tandem[3] and Mascot.[4] Since only a small amount of the generated tandem mass spectra in a typical peptidomics experiment are assigned to a sequence, the time and effort should be used to identifying those spectra instead of identifying the same peptides repeatedly. A possible solution for this problem is to collect tandem mass spectra and then use spectral library algorithms[5,6] to match experimental spectra to already identified spectra. This is a fast way to reidentify already identified peptides and the effort can instead be put into identifying sequences for the large part of the spectra that never get assigned to a sequence. To make these algorithms useful, it is necessary to have large collections of spectra. In fact, a growing number of tandem mass spectra from peptides are publicly available in databases.[7–9] However, these databases mainly contain proteolytic peptides produced by trypsin digestion and are not designed for endogenously processed peptides.

The SwePep (www.swepep.org) database was established to alleviate the identification problems for endogenous peptides.[10,11] Here, we report the extension of SwePep to include collision induced dissociation (CID) tandem mass spectra to allow for easier validation of identification results for endogenous peptides. The MS/MS database was created by adding a locally curated version of the global proteome machine database.[7] This is also a first step toward spectrum-to-spectrum identifications of endogenous peptides. In the future it is envisioned that the MS/MS database will be further extended with additional CID mass spectra and the complementary ETD/ECD[12–14] mass spectra.

The SwePep Database. SwePep is a database for endogenous peptides. The database contains information about the peptide precursors, p*I*, their post-translational modifications as well as references from the literature. To ensure that the information in the SwePep database is reliable, all peptides stored in SwePep are sorted into three different classes: (i) biologically active peptides, (ii) potentially biologically active peptides, and (iii) uncharacterized peptides. The group of *biologically active peptides* contains neuropeptides and hormones with previously

* To whom correspondence should be addressed: Dr. Per E. Andrén, Department of Pharmaceutical Biosciences, 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.
† Uppsala University.
‡ The Rockefeller University.

10.1021/pr800036d CCC: $40.75

Published on Web 06/13/2008

described and documented biological functions. Peptides that are classified as *potentially biologically active peptides* are identified peptides, from samples that have been instantly proteolytically inactivated post mortem or post sampling,[2,15,16] with characteristics similar to the neuropeptides and hormones; that is, they have specific convertase processing sites[1,17,18] and/or modifications such as C-terminal amidation and N-terminal acetylation, that are common on bioactive peptides.[17] However, their potential biological function and activity require investigation. The last group, *uncharacterized peptides*, contain peptides which are confidently identified but do not fulfill the criteria of the groups above.

SwePep has been used in different ways for identifying endogenous peptides from complex tissue samples utilizing mass spectrometry. In the first study,[10] the masses obtained from experimental peptides were compared with the masses of annotated peptides in SwePep and then the identities were verified using tandem mass spectrometry. Performing large database searches using unspecific cleavage and allowing for a number of different post-translational modifications is time-consuming and the result of the search is often poor. However, when comparing the experimental peptide masses against the theoretical peptide masses using SwePep it was possible to add different modifications. This allowed for rapid selection of potentially modified peptides. The selected candidates could then be verified by tandem mass spectrometry. This procedure was very time-efficient compared to the standard approach since only a small number of tandem mass spectra need to be inspected.

In the second study,[11] the SwePep database was used to construct three targeted sequence collections that mimic the peptidomic samples: SwePep precursors, SwePep peptides, and SwePep predicted. The neuropeptide searches against these three sequence collections were compared with searches against the entire mouse proteome, which is commonly used to identify neuropeptides. Three times as many peptides were identified, with a false positive rate <1%, from these new sequence collections compared to the mouse proteome. The new sequence collections also made it possible to identify 27 previously uncharacterized peptides and potentially bioactive neuropeptides. These novel peptides were cleaved from the peptide precursors at sites that are characteristic for prohormone convertases,[1,17,18] and some of them have post-translational modifications that are characteristic for neuropeptides.[17]

In the present study, we have extended the SwePep database to include tandem mass spectra of identified endogenous peptides. Tandem mass spectra of 219 unique peptides that have been identified with a significant score have been added to the SwePep database and the peptides are linked to their corresponding tandem mass spectra. In total, there are 2700 tandem mass spectra identified using X! Tandem, from 389 unique peptides, and 219 of them have a score over the significance threshold suggested by the search engine. The tandem mass spectra with a below threshold score are only stored in the MS/MS database, but are possible to search for according to mass or sequence. The tandem mass spectra in SwePep can be used for validating peptide identification results and for designing targeted experiments to monitor selected peptides.

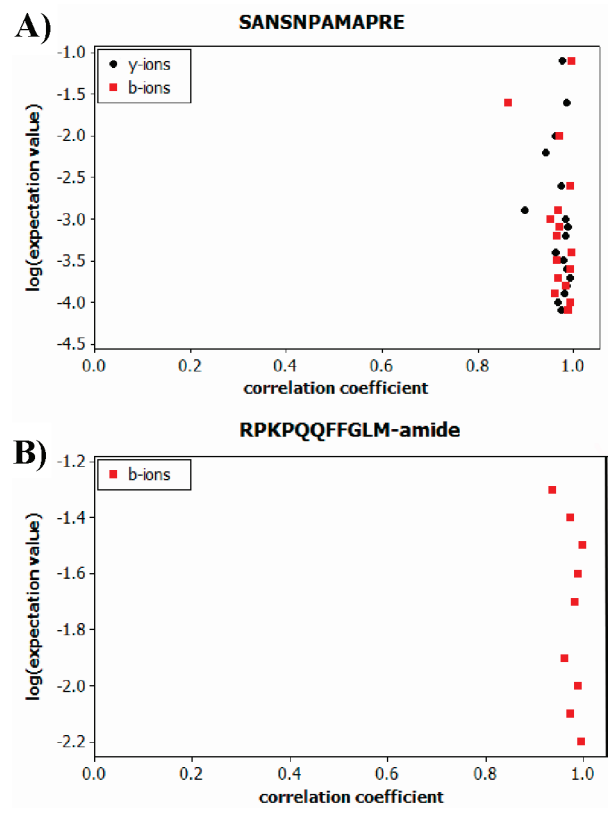

Figure 1. Correlation coefficients for the y-ions and b-ions series between high score tandem mass spectra and low score tandem mass spectra plotted against the X! Tandem score. (A) The correlation coefficients for the peptide SANSNPAMAPRE and (B) the peptide RPKPQQFFGLM-amide.

Material and Methods

Software Architecture. SwePep is a Java Enterprise Edition (J2EE) application. It consists of a dynamic Web interface, a relational database and a business tier, which uses the client-input from the Web interface to construct and execute queries to the database. The Web interface was developed using HTML and the dynamic content using Java ServerPages (JSP).

Data Model. The SwePep database is implemented as a relational database using MySql database management system. SwePep is specifically designed for endogenous peptides. Every peptide in the database is connected to the following information: name, sequence, precursor protein, position in precursor sequence, modifications, location, organisms, reference, mass and isoelectric point (p*I*).

Information Collection. The information in SwePep is collected from three different sources: experimental data,[2,11] peptide information from UniProt[19] (version 54.0, released July 2007), and peer reviewed publications. The database is updated continuously. For all the peptides in the SwePep database, monoisotopic mass, average mass, and p*I*[20] have been calculated according to their amino acid sequences.

Peptide and precursor protein data have been collected from UniProt by downloading the UniProt database in XML format. The XML file was searched for entries which had one or more annotated peptide. All entries with annotated peptides were saved into a new file which was used to automatically insert the entries into SwePep.

The SwePep database is also populated with uncharacterized peptides from brain tissue, identified in our laboratory from

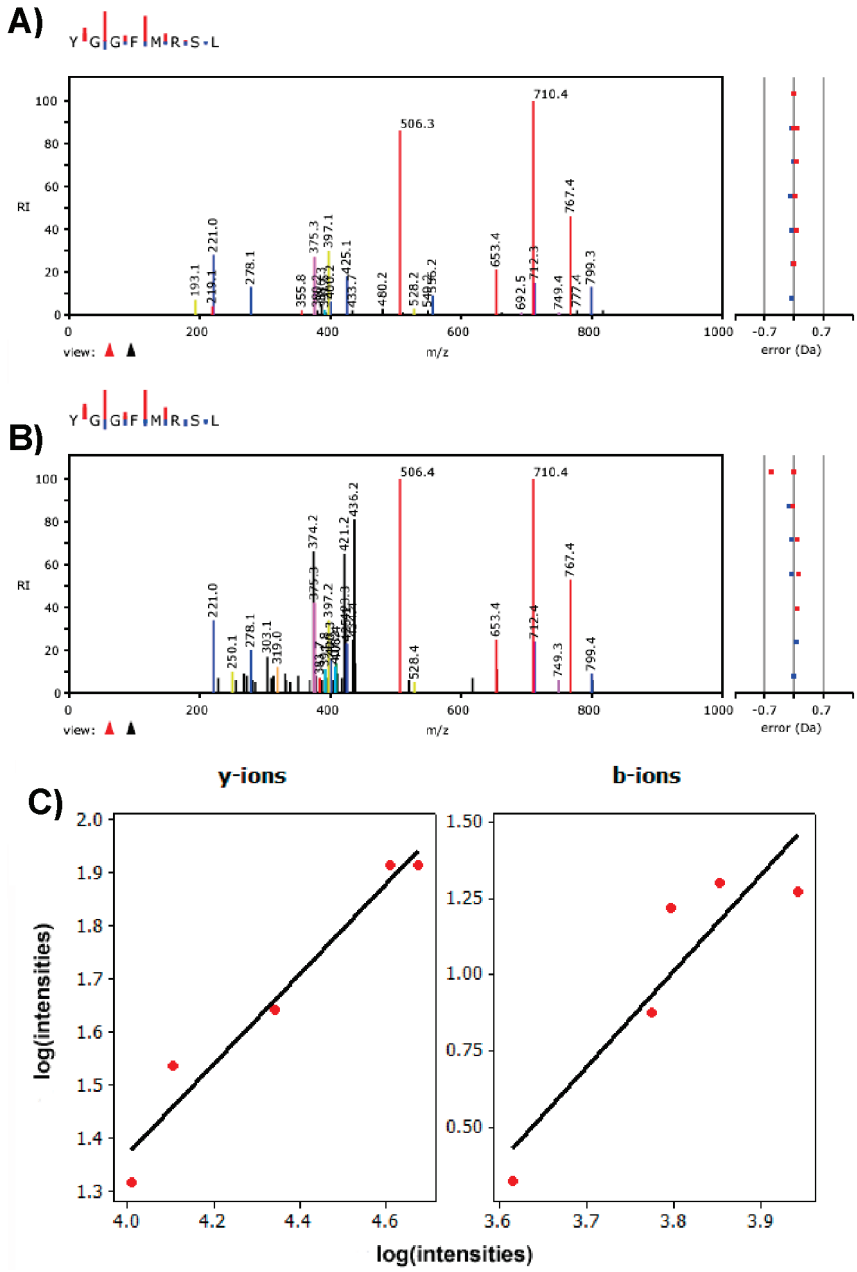

Figure 2. (A) Tandem mass spectrum of the peptide YGGFMRSL, derived from the Proenkephalin precursor, which was identified with high confidence, log(e) = −4.0. (B) Tandem mass spectrum that has been assigned with the same sequence, but the score of the peptide-spectrum match is log(e) = −1.4, which is below the threshold suggested by the search engine. (C) The correlation of the y-ion and the b-ion intensities of these two spectra. The Pearson correlation coefficients were 0.978 for the y-ions and 0.914 for the b-ions.

different species, mainly mouse. For this data set, SwePep contains information about the experimental conditions such as sample information (i.e., species, treatment) and tandem mass spectra.

Peptide Identification Using Tandem Mass Spectrometry. The Global Proteome Machine Database[7] (gpmDB) is used to visualize tandem mass spectra of endogenous peptides. The gpmDB is an open-source system developed for efficient storing and sharing proteomics data. The gpmDB was used to store search result from X! Tandem.[3]

The CID tandem mass spectra stored in SwePep were collected from analysis of different brain regions from the mouse. They were analyzed on an LTQ (Thermo Fisher Scientific) mass spectrometer by capillary liquid chromatography electrospray ionization tandem mass spectrometry (nanoLC-ESI MS/MS)[2,11,21] and identified by searching the spectra against SwePep Mouse precursors[11] using X! Tandem. All identifications are stored in the database, even if the scores are below the suggested threshold (log(e) > −2).

Statistical Analysis. Spectrum validation was performed by calculating the Pearson correlation coefficient[22] between two spectra. The correlation coefficient was calculated for the b- and y-ions series separately by taking the log intensities of the b- and y-ions that were detected in both spectra. A perl script for calculating the correlation between two tandem mass spectra is downloadable from the SwePep Web page.

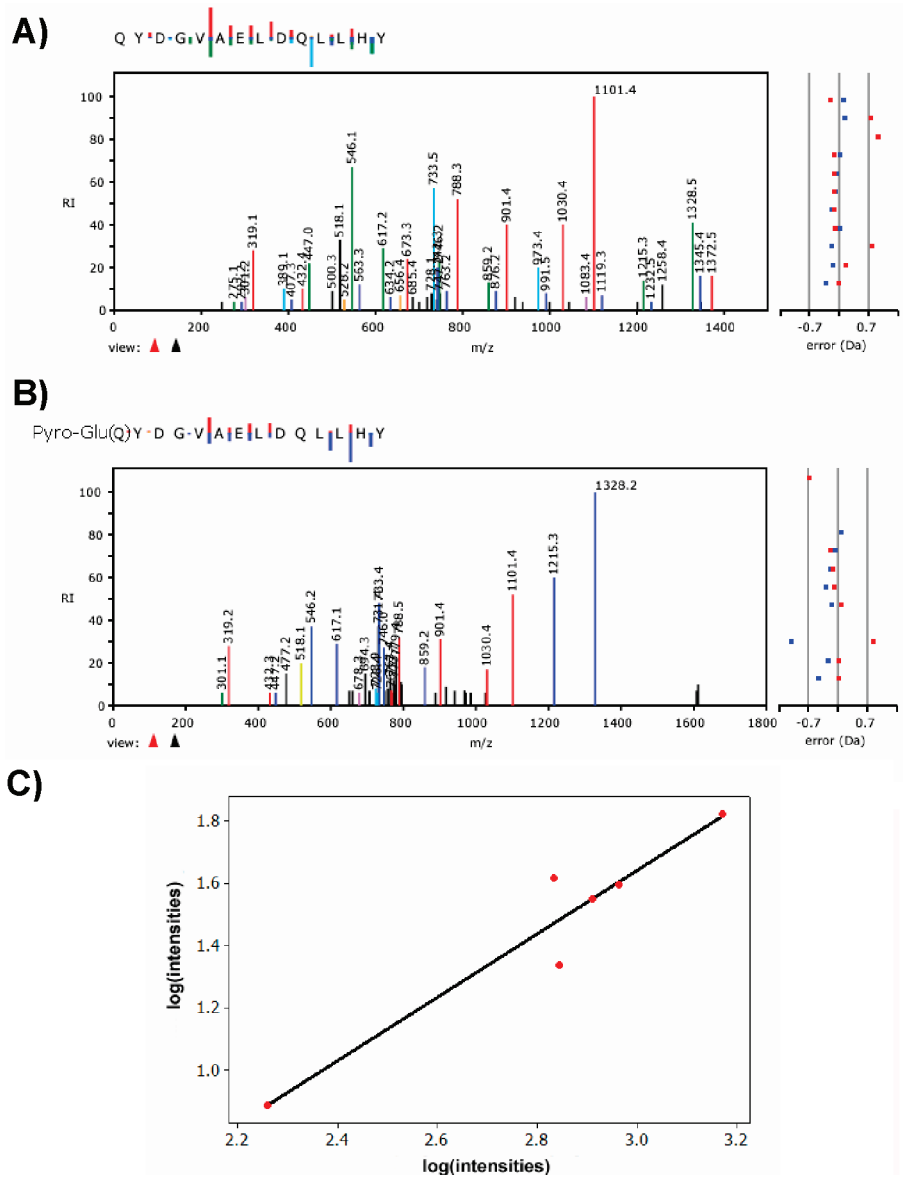

Figure 3. Two tandem mass spectra assigned to the peptide QYDGVAELDQLLHY (log(e) = −8.2) in (A) and Pyro-Glu(Q)YDGVAELDQLLHY (log(e) = −2.6) in (B), derived from the Secretogranin I precursor. (C) The intensity of the y-ion series of the modified and unmodified peptides plotted against each other and the Pearson correlation coefficient is calculated to be 0.959.

Results and Discussions

The tandem mass spectra stored in the SwePep database can be used as references when verifying poor fragmented spectra or spectra containing unassigned peaks. One possible way to verify a low score spectrum is by comparing the abundance of y-ions and b-ions for both the low score spectrum and a high score spectrum. Then the correlation coefficient between ion series of the spectra is calculated to establish the similarity between their fragmentation patterns. To investigate if the correlation coefficient is a good measurement of the similarity between spectra, the correlation coefficient was calculated for spectra assigned with the same sequence. This procedure was used for two different peptide sequences, SANSNPAMAPRE (SMS 28(1−12)) (Figure 1A) and RPKPQQFFGLM-amide (Substance P) (Figure 1B). All spectra were compared with the spectrum with the highest score, by calculating the correlation coefficient[22] between the log(intensities) of the y- and b-ion series. Since Substance P has two basic amino acids close to the N-terminal of the peptide, the b-ions are more abundant than the y-ions,[23] and therefore, only the b-ions are used for calculating the correlation between the spectra. The correlation between the spectra were high even when the score decreased, which suggests that the correlation coefficient is a good measure for the similarity of two spectra and that this procedure can be used for verification of the peptide sequence.

Another strength of the SwePep database was demonstrated when one peptide was identified with high confidence tandem mass spectrum in one experiment, while in another experiment, the peptide was identified with a below threshold score (Figure 2A,B). By comparing the fragmentation patterns of the two spectra, it was clear that they were very similar except for the noise peaks in the second spectrum which lowered its score because of their relatively high abundance. Using the present method, the correlations coefficients[22] were 0.98 for the y-ions and 0.91 for the b-ions, inferring a high similarity between the two spectra (Figure 2C). One could verify the second identifica-

tion because the fragmentation pattern is similar between the two spectra. The major difference between the two spectra is that the second spectrum contains some unidentified peaks. By looking at the sum of the intensities of the y- and b-ions in the tandem mass spectra (180 516 for the MS/MS with the high score and 329 for the MS/MS with the lower score), it becomes clear that the noise in the second spectrum was due to the overall low intensity in the spectrum. Without having the high quality spectrum at hand, it would not be possible to verify the noise contaminated spectrum using the standard method for peptide identification.

Many of the peptide sequences are assigned to multiple spectra in the SwePep database, which makes it possible to study the fragmentation pattern for a specific peptide. Some of the peptides are identified in more than one charge state and both with and without post-translational modifications. This information could be used for comparing fragmentation patterns of peptides.

The peptide QYDGVAELDQLLHY derived from the Secretogranin I precursor is often identified both with (Figure 3A) and without (Figure 3B) a pyro-glutamine acid at the N-terminal. The modified peptide was identified with a lower score than the unmodified peptide and needed to be validated. One possible way to validate the identity of the modified peptide is by comparing the intensities of the y-ion series of the modified and unmodified peptide. By plotting the log(intensities) of both y-ion series against each other and calculating the correlation coefficient, it was possible to determine if the y-ion series are the same or not. The correlation coefficient was 0.96 which indicated that the y-ion series have a strong correlation and it was most likely that the identity of the modified peptide was correct.

Conclusions

In the present study, CID tandem mass spectra of endogenous peptides have been added to the SwePep database. These spectra can be used for validation of other experimentally derived spectra or for studying fragmentation patterns of peptides without specific enzymatic cleavage sites. This is also the first step in developing a tandem spectrum database for endogenous peptides that can be used for spectrum-to-spectrum identifications instead of peptide identifications using database searches.

Acknowledgment. This study was sponsored by the Swedish Research Council (VR), Grant No. 2004-3417, 621-2007-4686, 521-2007-3017, the K&A Wallenberg Foundation, and the Karolinska Institutet Centre for Medical Innovations, Research Program in Medical Bioinformatics.

References

(1) Steiner, D. F. The proprotein convertases. *Curr. Opin. Chem. Biol.* **1998**, *2*, 31–39.
(2) Svensson, M.; Skold, K.; Svenningsson, P.; Andren, P. E. Peptidomics-based discovery of novel neuropeptides. *J. Proteome Res.* **2003**, *2*, 213–219.
(3) Craig, R.; Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20*, 1466–1467.
(4) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20*, 3551–3567.
(5) Craig, R.; Cortens, J. C.; Fenyo, D.; Beavis, R. C. Using annotated peptide mass spectrum libraries for protein identification. *J. Proteome Res.* **2006**, *5*, 1843–1849.
(6) Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; King, N.; Stein, S. E.; Aebersold, R. Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **2007**, *7*, 655–667.
(7) Craig, R.; Cortens, J. P.; Beavis, R. C. Open source system for analyzing, validating, and storing protein identification data. *J. Proteome Res.* **2004**, *3*, 1234–1242.
(8) Desiere, F.; Deutsch, E. W.; King, N. L.; Nesvizhskii, A. I.; Mallick, P.; Eng, J.; Chen, S.; Eddes, J.; Loevenich, S. N.; Aebersold, R. The PeptideAtlas project. *Nucleic Acids Res.* **2006**, *34*, D655–658.
(9) Falth, M.; Savitski, M. M.; Nielsen, M. L.; Kjeldsen, F.; Andren, P. E.; Zubarev, R. A. SwedCAD, a database of annotated high-mass accuracy MS/MS spectra of tryptic peptides. *J. Proteome Res.* **2007**, *6*, 4063–4067.
(10) Falth, M.; Skold, K.; Norrman, M.; Svensson, M.; Fenyo, D.; Andren, P. E. SwePep, a database designed for endogenous peptides and mass spectrometry. *Mol. Cell. Proteomics* **2006**, *5*, 998–1005.
(11) Falth, M.; Skold, K.; Svensson, M.; Nilsson, A.; Fenyo, D.; Andren, P. E. Neuropeptidomics strategies for specific and sensitive identification of endogenous peptides. *Mol. Cell. Proteomics* **2007**, *6*, 1188–1197.
(12) Savitski, M. M.; Kjeldsen, F.; Nielsen, M. L.; Zubarev, R. A. Complementary sequence preferences of electron-capture dissociation and vibrational excitation in fragmentation of polypeptide polycations. *Angew. Chem., Int. Ed. Engl.* **2006**, *45*, 5301–5303.
(13) Syka, J. E.; Coon, J. J.; Schroeder, M. J.; Shabanowitz, J.; Hunt, D. F. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 9528–9533.
(14) Zubarev, R. A.; Kelleher, N. L.; McLafferty, F. W. Electron capture dissociation of multiply charged protein cations. A nonergodic process. *J. Am. Chem. Soc.* **1998**, *120*, 3265–3266.
(15) Skold, K.; Svensson, M.; Kaplan, A.; Bjorkesten, L.; Astrom, J.; Andren, P. E. A neuroproteomic approach to targeting neuropeptides in the brain. *Proteomics* **2002**, *2*, 447–454.
(16) Skold, K.; Svensson, M.; Norrman, M.; Sjogren, B.; Svenningsson, P.; Andren, P. E. The significance of biochemical and molecular sample integrity in brain proteomics and peptidomics: Stathmin 2−20 and peptides as sample quality indicators. *Proteomics* **2007**, *7*, 4445–4456.
(17) Fricker, L. D. Neuropeptide-processing enzymes: applications for drug discovery. *AAPS J.* **2005**, *7*, E449–455.
(18) Zhou, A.; Webb, G.; Zhu, X.; Steiner, D. F. Proteolytic processing in the secretory pathway. *J. Biol. Chem.* **1999**, *274*, 20745–20748.
(19) UniProt Consortium. The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **2007**, *35*, D193−197.
(20) Bjellqvist, B.; Hughes, G. J.; Pasquali, C.; Paquet, N.; Ravier, F.; Sanchez, J. C.; Frutiger, S.; Hochstrasser, D. The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences. *Electrophoresis* **1993**, *14*, 1023–1031.
(21) Svensson, M.; Sköld, K.; Nilsson, A.; Fälth, M.; Nydahl, K.; Svenningsson, P.; Andrén, P. Neuropeptidomics: MS Applied to the Discovery of Novel Peptides from the Brain. *Anal. Chem.* **2007**, *79*, 14–21.
(22) Pearson, K. Mathematical contributions to the theory of evolution: III. Regression, heredity, and panmixia. *Phil. Trans. R. Soc. London* **1896**, *187*, 253–318.
(23) Tabb, D. L.; Huang, Y.; Wysocki, V. H.; Yates, J. R. 3rd, Influence of basic residue content on fragment ion peak intensities in low-energy collision-induced dissociation spectra of peptides. *Anal. Chem.* **2004**, *76*, 1243–1248.

PR800036D