Laboratory for Computational Proteomics

www.FenyoLab.org

E-mail: Info@FenyoLab.org

Facebook: *NYUMC Computational Proteomics Laboratory*

Twitter: *@CompProteomics*

SwePep, a Database Designed for Endogenous Peptides and Mass Spectrometry*

Maria Fälth‡§, Karl Sköld‡§, Mathias Norrman‡, Marcus Svensson‡§, David Fenyö‖, and Per E. Andren‡§**

A new database, SwePep, specifically designed for endogenous peptides, has been constructed to significantly speed up the identification process from complex tissue samples utilizing mass spectrometry. In the identification process the experimental peptide masses are compared with the peptide masses stored in the database both with and without possible post-translational modifications. This intermediate identification step is fast and singles out peptides that are potential endogenous peptides and can later be confirmed with tandem mass spectrometry data. Successful applications of this methodology are presented. The SwePep database is a relational database developed using MySql and Java. The database contains 4180 annotated endogenous peptides from different tissues originating from 394 different species as well as 50 novel peptides from brain tissue identified in our laboratory. Information about the peptides, including mass, isoelectric point, sequence, and precursor protein, is also stored in the database. This new approach holds great potential for removing the bottleneck that occurs during the identification process in the field of peptidomics. The SwePep database is available to the public. ***Molecular & Cellular Proteomics 5:998–1005, 2006.***

Proteomic tools, including two-dimensional gel electrophoresis in combination with MS, are limited to the analysis of proteins >10 kDa, and therefore an important part of the proteome is generally ignored in proteomic studies. This part of the proteome consists of endogenous proteins and peptides that include well characterized families of neuropeptide transmitters, neuropeptide modulators, hormones, and fragments of functional proteins, some of which are essential in many biological processes (1, 2). The peptides exert potent biological actions in the respiratory, cardiovascular, endocrine, inflammatory, and nervous systems (1, 2).

The study of endogenously processed peptides has been termed "peptidomics" (3). Peptidomics complements molecular biological approaches in its ability to characterize the processing of functional gene products. It allows direct observation of changes in the amount of peptides and small proteins and their post-translational modifications. The main difficulties in the analysis of endogenous peptides are their rapid degradation during extraction and purification (4) and that their average tissue content is less than 0.1% of that of proteins (5). Endogenous peptides also often contain post-translational modifications (PTMs)[1] (*e.g.* acetylation, amidation, and phosphorylation), adding to the difficulty of deciphering the obtained mass spectra.

An important functional group of the peptidome is the endogenous peptides in the brain. The neuropeptides range in length from 3 to 100 amino residues and are up to 50 times larger than classical neurotransmitters (6). The neuroactive peptides are derived from the processing of secretory proteins that are formed in the cell body on polyribosomes attached to the cytoplasmic surface of the endoplasmic reticulum. They are then processed in the endoplasmic reticulum and moved to the Golgi apparatus for further processing. In the central nervous system, most neurons contain biologically active peptides together with classical neurotransmitters. Neuropeptides are implicated in the pathology of various neurological and psychiatric disorders such as depression, neurodegenerative diseases, and eating and sleeping disorders (2).

Despite their biological and physiological importance there is at the moment a lack of easily accessible information in the public databases regarding endogenous peptides, making it difficult to identify the endogenous peptides from complex samples. MS in combination with two-dimensional gel electrophoresis or LC has become the main tool in proteomics for the identification of peptides and proteins and typically generates large sets of data (7). By using a search engine, the data are compared with protein sequence collections such as UniProt Knowledgebase (8) or the non-redundant (nr) protein sequence collection from the National Center for Biotechnology Information (NCBI). These protein sequence databases also offer additional information, including brief functional descriptions (if available), an annotation of sequence features

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

Received, December 8, 2005, and in revised form, February 17, 2006

Published, MCP Papers in Press, February 26, 2006, DOI 10.1074/mcp.M500401-MCP200

[1] The abbreviations used are: PTM, post-translational modification; CLIP, corticotropin-lipotropin intermediary peptide; GPCR, G-protein-coupled receptor; LTQ, linear trap quadrupole; UniProt, Universal Protein Resource; XML, extensible markup language.

(*e.g.* modifications), secondary and tertiary structure predictions, key references, and links to other databases. Lately a number of databases have become more oriented against specific proteomic subareas (5, 9, 10). Although several of these databases are well organized and easy to use, they do not always fulfill all new demands. At present there is no searchable database specifically designed for identification of endogenous peptides.

In the present study we have developed a database for endogenous peptides and small proteins below 10 kDa. The database consists of biologically active peptides such as classical neuropeptides and hormones, potential biologically active peptides, and uncharacterized peptides. Several examples on improved neuropeptide identification utilizing SwePep and MS are demonstrated.

EXPERIMENTAL PROCEDURES

Software Architecture

SwePep is a Java (11) Enterprise Edition (J2EE) application implemented according to a multitier application model (12). 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 hypertext markup language (HTML), and the dynamic content was developed using Java ServerPages (JSP), which at runtime compiles to JavaServlets. This makes it possible for the web interface to communicate with the server side functions and the database. When a user sends a request to the server through the web interface the request is caught by a servlet. The servlet first validates the request and then processes the request. A request to SwePep often involves database queries, and they are managed by Enterprise Java Beans (EJB). After the request is processed the control servlet sends a response back to the web interface, and the result of the request is displayed to the user.

Data Model

The SwePep database is implemented as a relational database (13) using an MySql database management system (11). 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 pI. The database is designed to minimize the data redundancy. Therefore some objects are split into two or more tables that are connected to each other, *e.g.* peptide and peptide type. This way the peptide sequence, mass, name, and pI are only stored once in the database even though the peptide occurs many times in different precursors (Fig. 1).

Information Collection

The information in SwePep is collected from three different sources: experimental data produced in our laboratory (4), peptide information from UniProt (version 49.0, released February 2006), and peer-reviewed publications. The data from UniProt will be updated every time a major release of UniProt is made. The rest of the data will be updated continuously. For all the peptides in the SwePep database, monoisotopic mass, average mass, and pI (14) have been calculated according to their amino acid sequence.

Currently SwePep consists of 4180 unique endogenous peptides, and many of these are post-translationally modified. So far, ~100 neuropeptides have been experimentally identified from brain tissue in our laboratory. The neuropeptides in SwePep have been derived from 1643 precursor proteins from 394 different species. All peptides have searchable descriptors such as mass (monoisotopic and average), modifications, precursor information, and organism affiliation. Because the experimental data contain peptides and proteins in the mass range up to 10 kDa, the SwePep database also contains 25,047 small proteins with sequence length less than or equal to 120 amino acids. This makes it possible to identify more of the contents in experimental samples. The current state of the number of peptides in the SwePep database is shown in Table I.

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

The SwePep database is also populated with novel peptides from brain tissue identified in our laboratory from different species. For this data set SwePep also contains information about the experimental conditions such as sample information (*i.e.* species and treatment), mass spectral raw data, and processed data.

Classification of Peptides in SwePep

To ensure that the information in the SwePep database is reliable, all peptides that are stored in SwePep are sorted into three different classes: (i) biologically active peptides, (ii) potential biologically active peptides, and (iii) uncharacterized peptides.

Biologically Active Peptides—This group of peptides contains the classical neuropeptides, such as substance P, neurotensin, enkephalins, and dynorphins, that are present in a neuron together with classical neurotransmitters. This group also contains peptides functioning as hormones, a class of peptides that are secreted into the blood stream to exert endocrine functions. All the neuropeptides and hormones in this group have known biological functions.

Potential Biologically Active Peptides—This group contains pharmacologically uncharacterized peptides (between 3 and 100 amino acids) that potentially are biologically active. They are identified in tissues or body fluids, which have been instantly proteolytically deactivated postmortem or postsampling, and have characteristics similar to the neuropeptides and hormones, *i.e.* they have specific convertase processing sites (15). Modifications such as amidation of the C terminus and N-terminal acetylation are regarded as important criteria because many bioactive peptides are amidated by conversion of a C-terminal glycine to a carboxamide.

Uncharacterized Peptides—Peptides that do not fulfill the criteria of the groups above belong to this group. Among others, this group consists of peptides from samples not rapidly proteolytically deactivated postsampling.

Sample Preparation and Mass Spectrometry Analysis

Rats (Sprague-Dawley) and mice (C57/BL6) were sacrificed as previously described (4) (Murimachi Kikai, Tokyo, Japan). The brain regions of interest were thereafter rapidly dissected out and stored at −80 °C. 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. The suspension was centrifuged at 20,000 × *g* for 30 min at 4 °C. The protein- and peptide-containing supernatant was transferred to a centrifugal filter (Microcon YM-10, Millipore, Bedford, MA) with a molecular mass limit of 10,000 Da and centrifuged at 14,000 × *g* for 45 min at 4 °C. Finally the peptide filtrate was immediately frozen and stored at −80 °C until analysis.

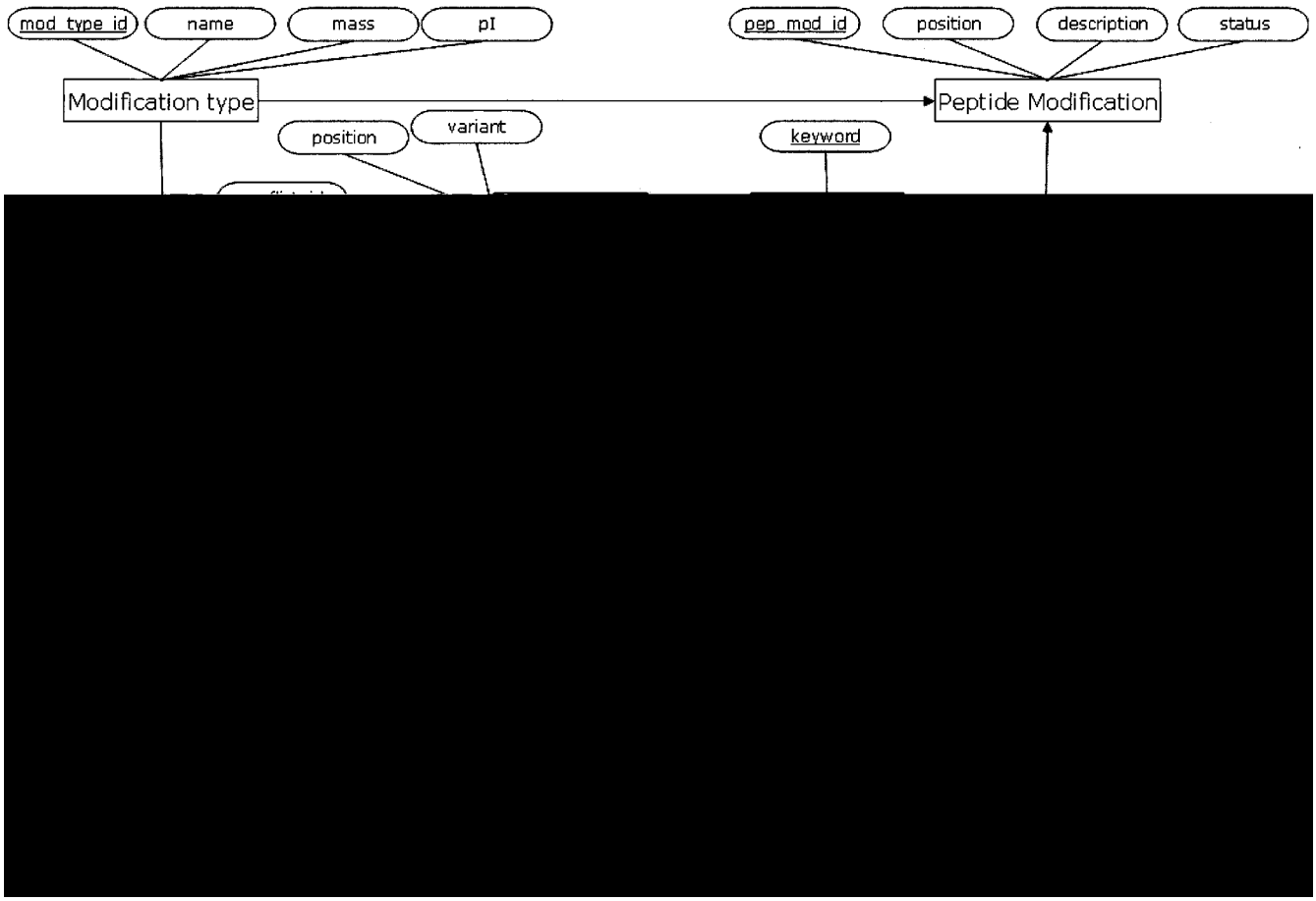

FIG. 1. **A peptide in SwePep has an identifier that is a reference to the peptide type table, which stores information about the peptide such as peptide sequence, name, pI, and mass.** The precursor table holds information about the name of the precursor, accession number associated with UniProt, amino acid sequence, and the length of the amino acid sequence. For each type of modification (mod) there is a row in the modification type table that holds information about the modification identification, name, mass, and pI. Modifications of a peptide or a precursor have to be connected to one and only one of the modification types in the modification type table.

TABLE I
Classification and number of peptides in SwePep

The SwePep database currently consists of 4180 endogenous peptides. 4136 of them have been found in UniProt Knowledgebase, and the rest have been identified in our laboratory. The peptides in SwePep are divided into three classes: biologically active peptides (peptides with known biological function), potential biologically active peptides (peptides from known peptide precursors with cleavage sites specific for endogenous peptides), and uncharacterized peptides (identified peptides that do not fulfill the criteria for the groups above).

	Biologically active peptides	Potential biologically active peptides	Uncharacterized Peptides
UniProt	4136	0	0
Experimentally identified peptides (in house)	37	28	16

The peptide extract was separated using on-line nanoflow reversed phase capillary liquid chromatography (Ettan MDLC, GE Healthcare, Uppsala, Sweden) and analyzed with ESI-MS using a Q-TOF (Waters) or Finnigan LTQ or LTQ-FT (Thermo Electron, San Jose, CA) mass spectrometer (4).

RESULTS AND DISCUSSION

Identification of endogenous peptides using MS data is time-consuming. The available software tools for identification are generally designed for proteins, which are cleaved to peptides by specific enzymes, such as trypsin, prior to MS analysis (7). However, the endogenous peptides will not contain many, if any, of such cleavage sites because they are processed at other specific sites by processing enzymes (proprotein convertases) that release the bioactive peptide from the precursor but also because of their small number of amino acid residues. This impairs the possibility of getting good peptide fragmentation data and a significant identification (16, 17). Furthermore another difficulty is that endogenous peptides often contain PTMs, which make them even more challenging to identify. The main purpose of the SwePep database is to speed up the identification process of endogenous peptides and to increase the number of identified peptides from complex tissue samples.

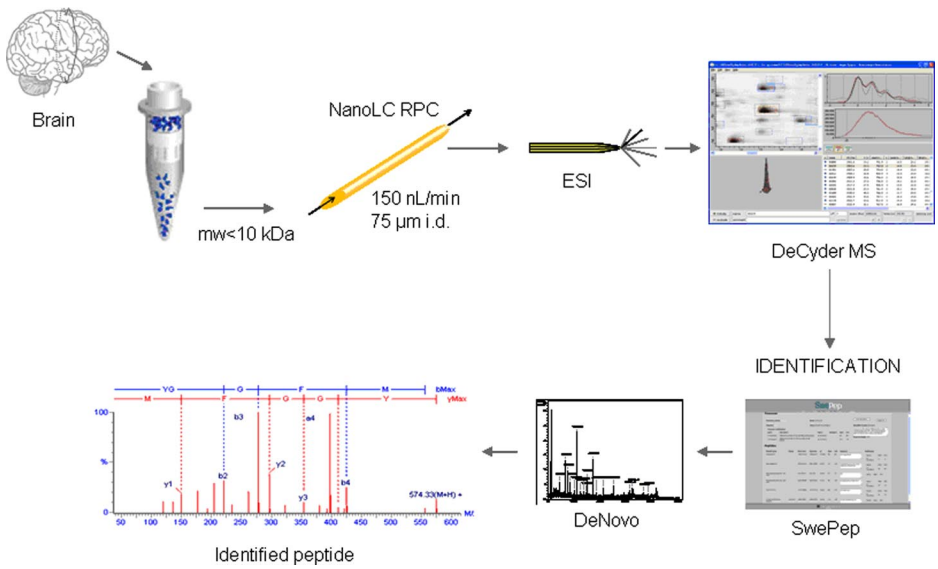

FIG. 2. **A schematic picture showing the workflow from sample to an identified peptide.** *RPC*, reverse phase chromatography; *i.d.*, inner diameter.

By classifying the peptides in SwePep into three different classes: (i) biologically active peptides, (ii) potential biologically active peptides, and (iii) uncharacterized peptides, it is possible to store peptides and protein fragments not proven to be biologically active. Peptides that belong to the group of potential or uncharacterized peptides are moved to the group of biologically active peptides if they demonstrate biological activity.

Previous studies suggest that the process of protein elimination and degradation should not be considered a random proteolysis yielding free amino acids subsequently utilized for various metabolic purposes. Instead it should be regarded as a complex process regulated by a system of tissue-specific enzymes and protein substrates. These peptides, complementary to the conventional regulatory systems, may be considered as another concept of a peptidergic regulatory system, giving rise to a large group of peptides, which are defined as tissue-specific peptide pool (18). For example, hemorphins are small peptides generated by enzymatic hydrolysis of hemoglobin or blood (19–21). Their physiological functions are discussed because they are found in a variety of mammalian tissues and fluids (22–27). Hemorphin peptides were previously found in brain tissue not proteolytically deactivated (28) but were not detected in tissue that had been proteolytically deactivated (4). However, hemorphins are claimed to have biological activity and produce constriction of coronary vessels and platelet aggregation (24) and to inhibit angiotensin-converting enzyme activity (29).

Searching for Peptide Identities Using SwePep—The whole peptide identification procedure, starting with experiment and ending with a list of identified peptides, is shown in Fig. 2. When using SwePep for peptide identification, the user typically starts by selecting a mass tolerance based on the mass accuracy of the mass spectrometer used in the peptide analysis. A file containing the experimental peptide masses are then matched against theoretical calculated masses in the database. The matching is performed both with and without annotated PTMs. It is also possible to add non-annotated modifications to the peptides to investigate other possible modifications.

The results from the database search are presented as a list containing peptide name, precursor name, species, peptide sequence, possible PTMs, monoisotopic mass, and mass difference between experimental and theoretical mass. A peptide identity from SwePep is a suggestion for an identity and needs to be confirmed by analysis of the corresponding tandem mass spectrum. Mostly there is only one suggested peptide identification for each experimental mass. This makes it easy to verify the results from SwePep. For the time being this confirmation has to be performed manually, but in the future tandem mass spectra will be stored in the database, and the confirmation will be implemented automatically. A database search in SwePep takes less than a minute, making it an efficient way of to identify known neuropeptides so that the effort can be put into the identification of novel peptides by *de novo* sequencing.

Example: Neuropeptide Identification in Hypothalamus—We have developed a new approach to study a large number of neuropeptides and used it for an investigation of the endogenous neuropeptide content of hypothalamic brain tissue samples from rat (4, 28). The MS data of the neuropeptide and small protein content in the hypothalamus were analyzed by an automated software program method for processing the results (DeCyder MS, GE Healthcare). The generated mass list consisting of deconvoluted mass data was matched against the SwePep database for neuropeptide matches. The absolute mass difference between the theoretical and experimental mass was selected not to exceed 0.2 Da for a match to be valid. All positive matches were recorded, and subsequent analysis was performed on experimental data pertaining to these matches to streamline the identification procedure. The final validation step was either manual inspection of tandem mass spectra, searching of sequence collections with tandem MS data for peptide identification, or a combination of both.

TABLE II

SwePep-matched rat hypothalamic neuropeptides

From a mass list with 400 experimentally observed peptide masses, the result from a search in SwePep contained 54 possible neuropeptides; 31 of those were positively confirmed by tandem mass spectrometry (in bold). MHC, melanin-concentrating hormone; CART, cocaine and amphetamine regulated transcript.

UniProt accession number	Peptide name	Peptide sequence	Annotated modification	Experimental mass	Mass difference
				Da	*Da*
O35314	**Secretogranin I precursor**	**^{585}SFAKAPHLDL^{594}**		**1097.688**	**0.102**
P01167	**Somatostatin-28-(1–12)**	**^{89}SANSNPAMAPRE^{100}**		**1243.659**	**0.098**
P01186	Arg-vasopressin	^{24}CYFQNCPRG^{32}		1086.624	0.186
P01186	**Vasopressin-neurophysin 2-copeptin precursor**	**^{151}VQLAGTQESVDSAKPRVY^{168}**		**1947.171**	**0.165**
P01194	Melanotropin γ	^{77}YVMGHFRWDRF^{87}	*C*-Amidation	1511.907	0.183
P01194	**Melanotropin α**	**^{124}SYSMEHFRWGKPV^{136}**	***C*-Amidation**	**1621.903**	**0.121**
P01194	**Melanotropin α**	**^{124}SYSMEHFRWGKPV^{136}**		**1622.814**	**0.048**
P01194	**CLIP**	**^{103}AEEETAGGDGRPEPSPRE^{120}**	***C*-Amidation**	**1881.996**	**0.151**
P01322	Insulin 1 A chain	^{90}GIVDQCCTSICSLYQLEN-YCN^{110}		2368.169	0.184
P04094	**Met-enkephalin**	**^{100}YGGFM^{104}**		**573.324**	**0.098**
P04094	**Met-enkephalin-Arg-Phe**	**^{263}YGGFMRF^{269}**		**876.488**	**0.092**
P04094	**Met-enkephalin-Arg-Gly-Leu**	**^{188}YGGFMRGL^{195}**		**899.504**	**0.072**
P04094	**Proenkephalin A precursor**	**^{198}SPQLEDEAKELQ^{209}**		**1385.772**	**0.104**
P04094	**Proenkephalin A precursor**	**^{198}SPQLEDEAKEL^{208}**		**1257.748**	**0.140**
P04094	**Proenkephalin A precursor**	**^{264}GGFMRF^{269}**		**713.422**	**0.090**
P04094	**Proenkephalin A precursor**	**^{219}VGRPEWWMDYQ^{229}**		**1465.805**	**0.160**
P06300	**Leu-enkephalin**	**^{166}YGGFL^{170}**		**555.299**	**0.030**
P06300	α-Neoendorphin	^{166}YGGFLRKYP^{174}		1099.696	0.115
P06767	**Neurokinin A**	**^{98}HKTDSFVGLM^{107}**	***C*-Amidation**	**1132.678**	**0.108**
P06767	**Substance P**	**^{58}RPKPQQFFGLM^{68}**	***C*-Amidation**	**1346.828**	**0.100**
P06767	**C-terminal flanking peptide**	**^{111}ALNSVAYERSAMQNYE^{126}**		**1845.010**	**0.173**
P07490	**Gonadoliberin I**	**^{24}QHWSYGLRPG^{33}**	**Pyrrolidone carboxyl acid *C*-amidation**	**1181.702**	**0.129**
P08435	Neurokinin B	^{82}DMHDFFVGLM^{91}		1210.671	0.156
P10354	**WE-14**	**^{361}WSRMDQLAKELTAE^{374}**		**1676.999**	**0.180**
P10354	**Chromogranin A precursor**	**^{395}AYGFRDPGPQL^{405}**		**1219.721**	**0.123**
P10362	Secretoneurin	^{184}TNEIVEEQYTPQSLATLESV-FQELGKLTGPSNQ^{216}		3649.899	0.099
P10683	Galanin	^{33}GWTLNSAGYLLGPHAIDN-HRSFSDKHGLT^{61}	*C*-Amidation	3162.765	0.190
P13432	SMR1-related undecapeptide	^{23}VRGPRRQHNPR^{33}		1371.797	0.026
P13589	**Pituitary adenylate cyclase-activating polypeptide precursor**	**^{111}GMGENLAAAAVDDRAPLT^{128}**		**1771.030**	**0.173**
P13668	**Stathmin**	**^{2}ASSDIQVKELEKRASGQAF^{20}**	Acetylation	**2105.264**	**0.189**
P14200	**Neuropeptide-glutamic acid-isoleucine**	**^{131}EIGDEENSAKFPI^{143}**	***C*-Amidation**	**1446.855**	**0.156**

TABLE II—*continued*

UniProt accession number	Peptide name	Peptide sequence	Annotated modification	Experimental mass	Mass difference
				Da	*Da*
P14200	**Pro-MCH precursor**	**^{131}EIGDEENSAKFPIG^{144}**		**1504.853**	**0.149**
P20068	Tail peptide	^{165}ASYYY^{169}		665.371	0.102
P20068	Neurotensin	^{150}QLYENKPRRPYIL^{162}		1688.949	0.013
P20068	**Neurotensin**	**^{150}QLYENKPRRPYIL^{162}**	**Pyroglutamic acid**	**1671.945**	**0.036**
P20156	**VGF protein precursor**	**^{491}PPEPVPPPRAAPAPTHV^{507}**		**1729.067**	**0.136**
P23436	**Cerebellin**	**^{1}SGSAKVAFSAIRSTNH^{16}**		**1631.942**	**0.104**
P27682	C-terminal peptide	^{198}SVPHFSEEEKEPE^{210}		1542.802	0.118
P27682	C-terminal peptide	^{198}SVPHFSEEEKEPE^{210}	Phosphorylation	1622.814	0.164
P28841	**Neuroendocrine convertase 2 precursor**	**^{94}IKMALQQEGFD^{104}**		**1278.764**	**0.137**
P49192	**CART protein precursor**	**^{82}IPIYE^{86}**		**633.377**	**0.040**
P60042	Somatostatin-14	^{103}AGCKNFFWKTFTSC^{116}	Disulfide bond	1636.876	0.160
P60042	Somatostatin-14	^{103}AGCKNFFWKTFTSC^{116}		1638.915	0.183
P81278	Prolactin-releasing peptide PrRP20	^{33}TPDINPAWYTGRGIRPVGRF^{52}	*C*-Amidation	2271.374	0.171
P98087	Cerebellin 2	^{88}SGSAKVAFSATRSTNH^{103}		1619.922	0.121
Q62923	Nociceptin	^{135}FGGFTGARKSARKLANQ^{151}		1808.094	0.114
Q62923	Neuropeptide 2	^{154}FSEFMRQYLVLSMQSSQ^{170}		2080.138	0.162
Q8BFS3	Relaxin 3 A chain	^{117}DVLAGLSSSCCEWGCSKSQ-ISSLC^{140}	Disulfide bond	2460.213	0.170
Q8BFS3	Relaxin 3 B chain	^{24}RPAPYGVKLCGREFIRAVIFTCG-GSRW^{50}		3038.772	0.187
Q9QXU9	**Big PEN**	**^{245}LENSSPQAPARRLLPP^{260}**		**1745.069**	**0.111**
Q9QXU9	**Little SAAS**	**^{42}SLSAASAPLAETSTPLRL^{59}**		**1784.130**	**0.162**
Q9QZQ4	Urotensin-2	^{110}QHGTAPECFWKYCI^{123}		1681.893	0.155
Q9R0R3	Apelin-13	^{65}QRPRLSHKGPMPF^{77}	Pyrrolidone carboxyl acid	1532.852	0.048

From hypothalamic mouse brain tissue DeCyder MS detected ~400 specific peptide masses. SwePep suggested 54 neuropeptide candidates, and of these, 31 neuropeptides were verified by tandem mass spectrometry (Table II).

The fact that only 54 of the 400 peptides detected by DeCyder MS could be identified by SwePep clearly demonstrates the challenge of identifying endogenous peptides. There are only 195 endogenous peptides from the mouse in the database, and many of these originate from other tissues than the brain. This indicates that a large portion of our detected peptides from hypothalamus are novel and not annotated and therefore do not exist in SwePep. Furthermore of these 195 peptides, 73 have annotated disulfide bonds, which impair the possibility to identify these peptides because the disulfide bonds may inhibit fragmentation in tandem MS.

Post-translational Modifications in SwePep—It is an important task to characterize all modifications for understanding of the biological function and the regulations of the peptides. Unfortunately it is both time-consuming and difficult to fully characterize peptides and proteins with respect to their modifications. Important modifications include acetylation, amidation, phosphorylation, and sulfation (2), and ~300 different modifications have been reported for proteins (30). For example, 50–90% of eukaryotic proteins synthesized in the cytoplasm are isolated with their N termini acetylated (31), including the opioid neuropeptide dynorphin that is acetylated after it has been cleaved from its larger precursor (32). It is also estimated that about 30% of mammalian proteins are phosphorylated (33).

Furthermore disulfide bonds are frequent modifications among peptides. Because of the small size of the peptides, disulfide bonds provide the necessary constraints for the peptides to have a well defined three-dimensional structure. This adds another level of complexity because many disulfide-linked peptides remain intact in tandem MS as mentioned above (34). The fact that endogenous peptides often are modified is also reflected in SwePep where the majority of the peptides are modified, *e.g.* 122 of the 195 peptides found in mouse have annotated modification, and 58 of the 122 have more than one annotated modification. By having information about modifications and thereby taking into account possible changes in the molecular mass, identification of modified peptides is easier.

In the example above analyzing the hypothalamic brain tissue we could identify a number of neuropeptides with different PTMs using SwePep. Several of the identified neuropeptides, such as corticotropin-lipotropin intermediary peptide (CLIP) and substance P, had C-terminal amidation. N-terminally acetylated stathmin was identified as well as gonadoliberin I with both a pyrrolidone carboxyl acid and

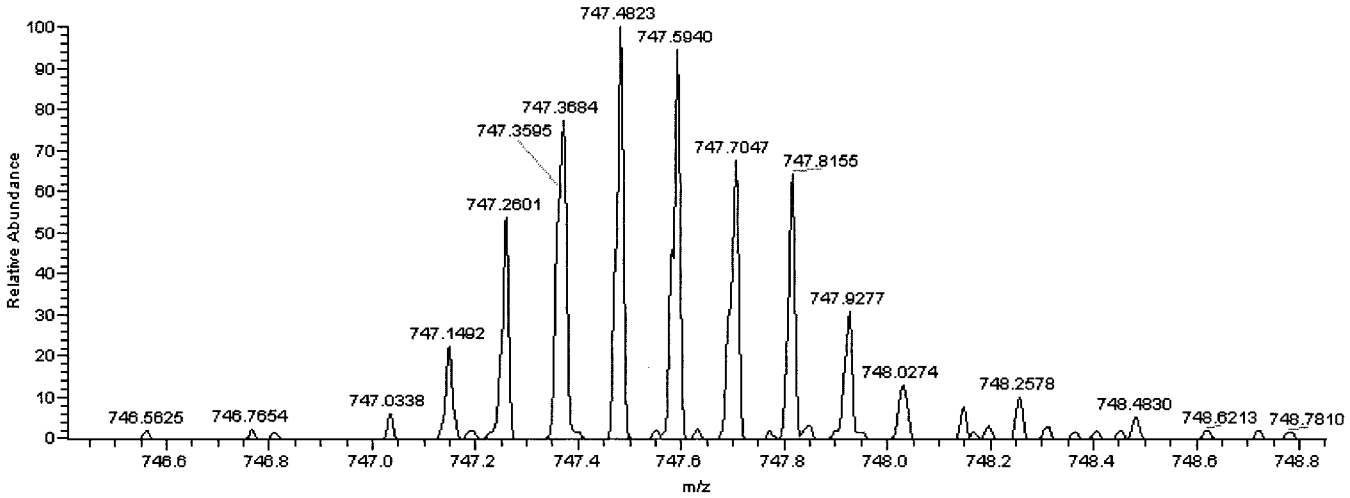

FIG. 3. **FT-ICR mass spectrum of PEP-19 showing the ion envelope at monoisotopic *m/z* 747.0338 (charge state, +9).**

C-terminal amidation. Additionally a phosphorylated (at Ser^{14}) and non-phosphorylated form of CLIP was also identified. Searching the SwePep database for peptides matching the experimental peptide masses 2505.01 and 2585.23 Da with a mass accuracy of 0.2 Da generated one matching peptide for each of the two masses. The suggested identities were Arg-CLIP and the phosphorylated species of Arg-CLIP. The identities were confirmed by tandem mass spectrometry. Some of these neuropeptides would have been difficult to identify without the suggested identity from SwePep.

Accurate Mass Identification of PEP-19 Using ESI FT-ICR MS—In a proteomic study of an animal model of Parkinson disease, we observed a decreased level of a 6.7-kDa peptide in mouse striatum using nano-LC ESI Q-TOF MS (35). Subsequent accurate mass data of the protein were acquired using nano-LC ESI LTQ-FT, and the MS data were compared with the SwePep. Because the mass accuracy of the LTQ-FT mass spectrometer is specified to less than 2 ppm by the manufacturer using external calibration (36), all possible peptide matches in the database were ensured by limiting the search to 10 ppm. Two matches corresponding to the molecular mass of the peptide were retrieved from the search, *i.e.* acetylated PEP-19 (mass, 6714.2604 Da) from mouse/rat and small venom protein 1 precursor (mass, 6714.2433 Da) from parasitoid wasp. The mass was calculated from the most intense charge state at *m/z* 747.0338 (Fig. 3). The suggested identity of the protein was also confirmed to be acetylated PEP-19 by tandem MS.

Potential Biologically Active Peptides—Traditionally the discovery of several novel peptides has been achieved by searching for ligands to the G-protein-coupled receptors (GPCRs). For example, the first neuropeptide GPCR ligand to be discovered was orphaninFQ/nociceptin, which is a ligand for an opioid-like receptor (37, 38). It is interesting to note that there exist about 550 GPCR genes in the human genome and that neuropeptides are ligands for about 20% of them. The classical transmitters constitute up to 55% of the GPCR ligands, and 25% have no known ligand (6).

Recently we were able to identify a number of novel endogenous peptides from rat hypothalamus. Moreover post-translational modifications of some of these novel peptides were also identified. These novel peptides from rat hypothalamus have been added to SwePep. We also have identified and added an additional 30 novel peptides from various regions in the mouse and rat brain to SwePep. The identities of these peptides will be published separately. Our technology, which includes instant deactivation of processing enzymes in the brain and highly sensitive MS analysis (4), may contribute to additional identification of novel biologically active neuropeptides, which will be added to the SwePep database.

Concluding Remarks—We have developed a novel database for endogenous peptides, SwePep, that contain approximately 4200 endogenous peptides, hormones, potential neuropeptides, and uncharacterized peptides from 394 different species to facilitate and improve endogenous peptide identification utilizing MS. A light version of the SwePep database is accessible through the internet, www.swepep.org. The website will grow continuously. It is possible to search for peptides according to mass, name, organism affiliation, UniProt accession number, or a combination of them. The result of the search contain detailed information about the peptide such as precursor name, precursor sequence, peptide name, mass, sequence, peptide function, and references.

* This study was sponsored by Swedish Research Council (VR) Grants 621-2004-3417 and 521-2002-6116, an institutional grant from the Swedish Foundation for International Cooperation in Research and Higher Education (STINT), the K&A Wallenberg Foundation, and the Karolinska Institutet Centre for Medical Innovations, Research Programme 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.

** To whom correspondence should be addressed. Tel.: 46-18-471-7206; Fax: 46-18-471-4422; E-mail: per.andren@bmms.uu.se.

REFERENCES

1. 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
2. Hokfelt, T., Broberger, C., Xu, Z. Q., Sergeyev, V., Ubink, R., and Diez, M. (2000) Neuropeptides—an overview. *Neuropharmacology* **39,** 1337–1356
3. Schulz-Knappe, P., Zucht, H. D., Heine, G., Jurgens, M., Hess, R., and Schrader, M. (2001) Peptidomics: the comprehensive analysis of peptides in complex biological mixtures. *Comb. Chem. High Throughput Screen.* **4,** 207–217
4. Svensson, M., Skold, K., Svenningsson, P., and Andren, P. E. (2003) Peptidomics-based discovery of novel neuropeptides. *J. Proteome Res.* **2,** 213–219
5. Minamino, N., Tanaka, J., Kuwahara, H., Kihara, T., Satomi, Y., Matsubae, M., and Takao, T. (2003) Determination of endogenous peptides in the porcine brain: possible construction of peptidome, a fact database for endogenous peptides. *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.* **792,** 33–48
6. Hokfelt, T., Bartfai, T., and Bloom, F. (2003) Neuropeptides: opportunities for drug discovery. *Lancet Neurol.* **2,** 463–472
7. Fenyo, D. (2000) Identifying the proteome: software tools. *Curr. Opin. Biotechnol.* **11,** 391–395
8. Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N., and Yeh, L. S. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **33,** D154–D159
9. Lu, P., Szafron, D., Greiner, R., Wishart, D. S., Fyshe, A., Pearcy, B., Poulin, B., Eisner, R., Ngo, D., and Lamb, N. (2005) PA-GOSUB: a searchable database of model organism protein sequences with their predicted Gene Ontology molecular function and subcellular localization. *Nucleic Acids Res.* **33,** D147–D153
10. Xenarios, I., Salwinski, L., Duan, X. J., Higney, P., Kim, S. M., and Eisenberg, D. (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* **30,** 303–305
11. (2001) The Human Genome Issue. *Nature* **409,** 745–953
12. Barish, G. (2002) *Building Scalable and High-Performance Java Web Applications Using J2EE Technology*, Addison-Wesley, Boston
13. Silberschatz, A., Korth, H., and Sudarshan, S. (2002) *Database System Concepts*, 4th Ed., McGraw-Hill, New York
14. Bjellqvist, B., Hughes, G. J., Pasquali, C., Paquet, N., Ravier, F., Sanchez, J. C., Frutiger, S., and Hochstrasser, D. (1993) The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences. *Electrophoresis* **14,** 1023–1031
15. Steiner, D. F. (1998) The proprotein convertases. *Curr. Opin. Chem. Biol.* **2,** 31–39
16. 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
17. 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
18. Karelin, A. A., Blishchenko, E., and Ivanov, V. T. (1999) Fragments of functional proteins: role in endocrine regulation. *Neurochem. Res.* **24,** 1117–1124
19. Ivanov, V. T., Karelin, A. A., Philippova, M. M., Nazimov, I. V., and Pletnev, V. Z. (1997) Hemoglobin as a source of endogenous bioactive peptides: the concept of tissue-specific peptide pool. *Biopolymers* **43,** 171–188
20. Piot, J. M., Zhao, Q., Guillochon, D., Ricart, G., and Thomas, D. (1992) Isolation and characterization of two opioid peptides from a bovine hemoglobin peptic hydrolysate. *Biochem. Biophys. Res. Commun.* **189,** 101–110
21. Brantl, V., Gramsch, C., Lottspeich, F., Mertz, R., Jaeger, K. H., and Herz, A. (1986) Novel opioid peptides derived from hemoglobin: hemorphins. *Eur. J. Pharmacol.* **125,** 309–310
22. Nishimura, K., and Hazato, T. (1993) Isolation and identification of an endogenous inhibitor of enkephalin-degrading enzymes from bovine spinal cord. *Biochem. Biophys. Res. Commun.* **194,** 713–719
23. Glamsta, E. L., Marklund, A., Hellman, U., Wernstedt, C., Terenius, L., and Nyberg, F. (1991) Isolation and characterization of a hemoglobin-derived opioid peptide from the human pituitary gland. *Regul. Pept.* **34,** 169–179
24. Barkhudaryan, N., Oberthuer, W., Lottspeich, F., and Galoyan, A. (1992) Structure of hypothalamic coronaro-constrictory peptide factors. *Neurochem. Res.* **17,** 1217–1221
25. Glamsta, E. L., Morkrid, L., Lantz, I., and Nyberg, F. (1993) Concomitant increase in blood plasma levels of immunoreactive hemorphin-7 and β-endorphin following long distance running. *Regul. Pept.* **49,** 9–18
26. Moisan, S., Harvey, N., Beaudry, G., Forzani, P., Burhop, K. E., Drapeau, G., and Rioux, F. (1998) Structural requirements and mechanism of the pressor activity of Leu-Val-Val-hemorphin-7, a fragment of hemoglobin β-chain in rats. *Peptides* **19,** 119–131
27. Moisan, S., Drapeau, G., Burhop, K. E., and Rioux, F. (1998) Mechanism of the acute pressor effect and bradycardia elicited by diaspirin crosslinked hemoglobin in anesthetized rats. *Can J. Physiol. Pharmacol.* **76,** 434–442
28. 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
29. Lantz, I., Glamsta, E. L., Talback, L., and Nyberg, F. (1991) Hemorphins derived from hemoglobin have an inhibitory action on angiotensin converting enzyme activity. *FEBS Lett.* **287,** 39–41
30. Jensen, O. N. (2004) Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. *Curr. Opin. Chem. Biol.* **8,** 33–41
31. Polevoda, B., and Sherman, F. (2003) N-terminal acetyltransferases and sequence requirements for N-terminal acetylation of eukaryotic proteins. *J. Mol. Biol.* **325,** 595–622
32. Robinson, P., Toney, K., James, S., and Bennett, H. P. (1995) Mass spectrometric and biological characterization of guinea-pig corticotrophin. *Regul. Pept.* **56,** 89–97
33. 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
34. Gorman, J. J., Wallis, T. P., and Pitt, J. J. (2002) Protein disulfide bond determination by mass spectrometry. *Mass Spectrom. Rev.* **21,** 183–216
35. Skold, K., Svensson, M., Nilsson, A., Zhang, X., Nydahl, K., Caprioli, R. M., Svenningsson, P., and Andren, P. E. (2006) Decreased striatal levels of PEP-19 following MPTP lesion in the mouse. *J. Proteome Res.* **5,** 262–269
36. Metelmann-Strupat, W., Strupat, K., Peterman, S., and Muenster, H. (2004) Accurate mass measurements using the Finnigan LTQ FT. Application Note 30045, Thermo Electron Corp., Waltham, MA
37. Meunier, J. C., Mollereau, C., Toll, L., Suaudeau, C., Moisand, C., Alvinerie, P., Butour, J. L., Guillemot, J. C., Ferrara, P., Monsarrat, B., Mazarguil, H., Vassart, G., Parmentier, M., and Costentin, J. (1995) Isolation and structure of the endogenous agonist of opioid receptor-like ORL1 receptor. *Nature* **377,** 532–535
38. Reinscheid, R. K., Nothacker, H. P., Bourson, A., Ardati, A., Henningsen, R. A., Bunzow, J. R., Grandy, D. K., Langen, H., Monsma, F. J., Jr., Civelli, O. (1995) Orphanin FQ: a neuropeptide that activates an opioid-like G protein-coupled receptor. *Science* **270,** 792–794