SwePep – A database optimally designed for storing and mining of neuropeptide MS data

Maria Falth[1], Karl Skold[1], Marcus Svensson[1], Mathias Norrman[1], Anna Nilsson[1], David Fenyo[2], Per E. Andren[1],
[1]Laboratory for Biological and Medical Mass Spectrometry, Uppsala University, Uppsala, Sweden,
[2]The Rockefeller University, New York, NY

UPPSALA UNIVERSITET

OVERVIEW

Purpose

Construction and validation of a database consisting of annotated neuropeptides in order to provide valuable help in the identification process of endogenous neuropeptides derived from mass spectrometry analyzes.

Methods

A relational database was created using MySql and consists of annotated neuropeptides and small proteins. Validation of the database was performed by comparing experimental peptide mass data to the theoretical peptide masses in the database.

Results

Using SwePep 36 neuropeptides were identified from one LC-MS analysis. Masses automatically chosen for MS/MS by the mass spectrometry software verified the identity 15 of the peptides (100% success rate).

INTRODUCTION

Identification of peptides and proteins in big datasets generated by mass spectrometry analyses has become one of the major tasks in proteomics. By using a tool-box of different search engines the data is compared against protein sequence collections such as SwissProt and NCBI. Lately a number of databases have become more oriented against specific proteomic sub areas. Although some databases e.g. SwissProt are well organized and easy to use, they do not always fulfil all new demands. Today there is no searchable database specifically designed for identification of endogenous neuropeptides. Therefore a unique neuropeptide database was developed.

For the time being, SwePep consists of 1600 neuropeptide precursor proteins, 3600 neuropeptides, and 25000 small proteins (mw<10kDa) derived from 4000 different organisms. All peptides have the searchable descriptors such as: mass (monoisotopic and average), pI value, tissue origin, sub cellular location, modifications, precursor information and organism affiliation. To make SwePep flexible and user-friendly, the user interface has multiple options for choosing different representations of the output data.

METHODS

Database construction

The database management system used for SwePep is MySql. A simplified database schema for SwePep is shown in Figure 1. Server side functions are developed using Java enterprise edition and the client side by using JSP and HTML. SwePep consists of in-house produced data, re-organized SwissProt data and data collected from literature.

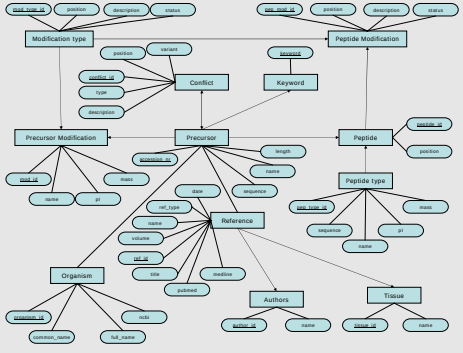

Figure 1. *A simplified database schema for the SwePep database*

Sample preparation and LC-MS

Hypothalamus from rat was dissected out and stored at -80 C. The tissue was suspended in cold extraction solution (0.25% acetic acid), homogenized by sonication (Sonics & Materials) and filtered through a 10 kDa mw cut-off membrane (Millipore). The peptide extract was analyzed using on-line nanoscale reversed phase capillary liquid chromatography (GE Healthcare) and ESI-LTQ (Thermo Electron).

Data analysis

The neuropeptide and small protein content in hypothalamus from rat was analyzed by DeCyder MS (GE Healthcare). The generated mass list consisting of deconvoluted mass data was matched to the previously stored information in the SwePep database for neuropeptide hits. The identification step utilizes peptide mass, and is verified by embedded MS/MS information if available. All experimental information is stored in the database, including MS/MS, to facilitate future identification processes.

Possible neuropeptide modifications were considered in the matching procedure. The absolute mass difference between the theoretical and experimental mass was selected not to exceed 0.2 Da in order 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 MS/MS spectra, searching of sequence collections with MS/MS data for peptide identification or a combination of both.

RESULTS

Data stored in the SwePep database has been used to identify neuropeptides in hypothalamus of rat brain. DeCyder MS detected approximately 1200 specific peptide masses. The generated mass list were matched to SwePep for identification with a mass accuracy of 0.2 Da. SwePep suggested 36 neuropeptide hits from the 1200 potential peptides. 15 neuropeptides were also verified by MS/MS (Table 1).

Figure 2. *Neuropeptide and small protein content in different brain regions from rat in the mass range up to 10 kDa were investigated using NanoLC ESI MS. The MS output data was analyzed by DeCyder MS (GE Healthcare) and was used for peptide identification by matching experimental data to the SwePep database.*

Swissprot ac-number	Peptide name	Peptide sequence	Annotated modification	Exprimental mass	Mass difference
P01186	Arg-vasopressin	24-CYFQNCPRG-32		1086.623627	0.186107
P01194	Met-enkephalin	205-YGGFM-209		573.3241474	0.0984474
P01194	Melanotropin alpha	124-SYSMEHFRWGKPV-136		1622.814427	0.048067
P01194	Melanotropin alpha	124-SYSMEHFRWGKPV-136	C-Amidation	1621.903148	0.120808
P01194	Melanotropin gamma	77-YVMGHFRWDRF-87	C-Amidation	1511.90742	0.18299
P01322	Insulin 1 A chain	90-GIVDQCCTSICSLYQLENYCN-110		2368.168877	0.184497
P04094	Met-enkephalin-Arg-Gly-Leu	188-YGGFMRGL-195		899.5038684	0.0715284
P04094	Met-enkephalin-Arg-Phe	263-YGGFMRF-269		876.4877246	0.0924946
P04094	Met-enkephalin	100-YGGFM-104		573.3241474	0.0984474
P04094	Leu-enkephalin	232-YGGFL-236		555.2993982	0.0301182
P06300	Leu-enkephalin	166-YGGFL-170		555.2993982	0.0301182
P06300	Alpha-neoendorphin	166-YGGFLRKYP-174		1099.6961	0.11466
P06767	C-terminal flanking peptide	111-ALNSVAYERSAMQNYE-126		1845.009567	0.173267
P06767	Substance P	58-RPKPQQFFGLM-68	C-Amidation	1346.827927	0.099807
P06767	Neurokinin A	98-HKTDSFVGLM-107	C-Amidation	1132.678003	0.108113
P07490	Gonadoliberin I	24-QHWSYGLRPG-33	Pyrrolidone carboxyl acid, C-Amidation	1181.701541	0.128541
P08435	Neurokinin B	82-DMHDFFVGLM-91		1210.670673	0.155603
P10354	WE-14	361-WSRMDQLAKELTAE-374		1676.999294	0.180124
P10362	Secretoneurin	184-TNEIVEEQYTPQSLATLESVFQELGKLTGPSNQ-216		3649.898526	0.098666
P10683	Galanin	33-GWTLNSAGYLLGPHAIDNHRSFSDKHGLT-61	C-Amidation	3162.765108	0.190418
P13432	SMR1-related undecapeptide	23-VRGPRRQHNPR-33		1371.796932	0.026122
P14200	Neuropeptide-glutamic acid-isoleucine	131-EIGDEENSAKFPI-143	C-Amidation	1446.85488	0.15585
P20068	Tail peptide	165-ASYYY-169		665.3712777	0.1015877
P20068	Neurotensin	150-QLYENKPRRPYIL-162		1688.94896	0.01276
P27682	C-terminal peptide	198-SVPHFSEEEKEPE-210	Phosphorylation	1622.814427	0.164327
P27682	C-terminal peptide	198-SVPHFSEEEKEPE-210		1542.801839	0.118069
P60042	Somatostatin-14	103-AGCKNFFWKTFTSC-116		1638.91481	0.18252
P60042	Somatostatin-14	103-AGCKNFFWKTFTSC-116	Disulphide bond	1636.876182	0.159542
P81278	Prolactin-releasing peptide PrRP20	33-TPDINPAWYTGRGIRPVGRF-52	C-Amidation	2271.373894	0.171454
P98087	Cerebellin 2	88-SGSAKVAFSATRSTNH-103		1619.922462	0.120902
Q62923	Neuropeptide 2	154-FSEFMRQYLVLSMQSSQ-170		2080.138235	0.162455
Q62923	Nociceptin	135-FGGFTGARKSARKLANQ-151		1808.094283	0.113773
Q8BFS3	Relaxin 3 B chain	24-RPAPYGVKLCGREFIRAVIFTCGGSRW-50		3038.771956	0.187276
Q8BFS3	Relaxin 3 A chain	117-DVLAGLSSSCCEWGCSKSQISSLC-140	Disulphide bond	2460.213362	0.170422
Q9QZQ4	Urotensin-2	110-QHGTAPECFWKYCI-123		1681.893213	0.155123
Q9R0R3	Apelin-13	65-QRPRLSHKGPMPF-77	Pyrrolidone carboxyl acid	1532.851824	0.048414

Confirmed by MS/MS spectra

Table 1. *From 1407 exprimental peptide masses detected by DeCyder MS, we could match them to 36 annotatade neuropeptides in SwePep. 13 of these were then validated by MS/MS spectra.*

CONCLUSIONS

A neuropeptide database has been developed, available at www.swepep.com (Figure 3). SwePep has been used to identify annotated neuropeptides from MS data in a fast and reliable way. To facilitate future identification processes, all experimental information is stored in the database, i.e. SwePep is under constant growth.

The neuropeptide content of hypothalamus from rat was analyzed using LC ESI-MS. The mass spectrometry raw data was processed by DeCyder MS and the generated output was matched to SwePep. With the mass accuracy of 0.2 Da, 36 neuropeptides were identified. So far, 15 neuropeptide identities were confirmed by MS/MS.

FUTURE PLANS

- Populate the database with additional novel neuropeptides
- Store MS/MS spectra in the SwePep database for automatic validation
- Curate the database by incorporating literature references to all neuropeptides

Figure 3. *Screenshots of SwePep, available at www.swepep.com*

ACKNOWLEDGEMENTS

This work was supported by the Swedish Research Council, Medicine #11565, Swedish Foundation for International Cooperation in Research and Higher Education (STINT), the Knowledge Foundation, Karolinska Institute Programme in Medical Bioinformatics, and K&A Wallenberg Foundation.