NeuroPedia: neuropeptide database and spectral library

Yoon Kim¹, Steven Bark², Vivian Hook²,³,⁴,⁵ and Nuno Bandeira²,⁶,*

¹Department of Electrical and Computer Engineering, ²Skaggs School of Pharmacy and Pharmaceutical Sciences, ³Department of Neurosciences, ⁴Department of Pharmacology, ⁵Department of Medicine and ⁶Department of Computer Science and Engineering, University of California, San Diego, La Jolla, California 92093-0744, USA

Received on May 7, 2011; revised on July 11, 2011; accepted on July 25, 2011

1 INTRODUCTION

Neuropeptides are peptide neurotransmitters and hormones that mediate cell-cell communication for regulation of physiological functions and biological processes. Understanding the role and regulation of neuropeptide forms in health, disease and drug treatments requires the ability to globally analyze neuropeptide expression in an unbiased form. Mass spectrometry (MS) based neuropeptidomics is highly suited for untargeted, global neuropeptide studies (Bora et al., 2008; Fricker et al., 2007; Hook et al. 2010; Li et al. 2008; Svensson et al., 2007). However, the unique characteristics of neuropeptides (e.g. short/long sequences or non-tryptic) presents difficulties for identification from tandem mass spectrometry (MS/MS) spectra. Many neuropeptides are non-tryptic, meaning they lack of tryptic residues to facilitate peptide ionization/fragmentation. Therefore, while some neuropeptides can be identified with current database search tools, the resulting data has not been structured to facilitate further analysis by tandem mass spectrometry (MS/MS), the main technology for high-throughput neuropeptide identification. Many neuropeptides are difficult to identify when searching MS/MS spectra against large protein databases because of their atypical lengths (e.g. shorter/longer than common tryptic peptides) and lack of tryptic residues to facilitate peptide ionization/fragmentation.

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

Availability: http://proteomics.ucsd.edu/Software/NeuroPedia.html

Supplementary information: Supplementary materials are available at Bioinformatics online.

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

NeuroPedia is a specialized neuropeptide database and spectral library that is directly searchable using mass spectrometry data. In addition to the expected improvement in sensitivity from searching against a small targeted sequence database, the neuropeptide spectral libraries further improve identification efficiency, sensitivity and reliability by considering all spectral features, including actual fragment intensities, neutral losses from fragments and various uncommon or even unknown fragments to determine the best matches. NeuroPedia spectral libraries are compatible with the publicly available spectral library search tool M-SPLIT (Wang et al., 2010) and can be easily converted to other spectral library formats. To further facilitate visual evaluation of neuropeptide MS/MS spectra, NeuroPedia provides annotated spectrum images for every library spectrum and further separates spectral libraries by species, digestion enzyme and instrument type (see Supplementary Table).

2 METHODS

NeuroPedia aggregates neuropeptide data from in-house mass spectrometry experiments with data from multiple reference websites and public spectral library. Using python HTML parsing and manual typing, we gathered neuropeptide names, gene families, gene names and their protein names.
from the neuropeptide repository, Handbook of Biologically Active Peptides (Kastin et al., 2006), and UniProt (Jain et al., 2009). In addition, we also obtained neuropeptide sequences, their start and end positions on the precursor protein, species, RefSeq ID and UniProt ID from UniProt. We collected NCBI taxonomy ID and gene reference ID from NCBI. Using cluster searching at 50% identity in UniProt, we expanded the catalog of species from human into chimpanzee, mouse, rat, bovine, rhesus macaque and California sea hare. We further analyzed the collected neuropeptide sequences to classify sequence similarities between neuropeptides into three match types: (a) identical if the sequences are exactly the same, (b) overlapping if the prefix of one sequence exactly matches the suffix of the other sequence for at least k characters, where k is half the length of longest sequence and (c) homolog if overlapping as in (b) but allowing up to two amino acid substitutions. Neuropeptide spectra were collected from NIST spectral libraries (Stein et al., 2009), and in-house datasets (Braund et al., 2011; Gupta et al., 2010).

All collected MS/MS spectra were searched against the NeuroPedia sequences database using InsPect (Tanner et al., 2005) at http://proteomics.ucsd.edu with search parameters: Instrument (ESI-Ion-TRAP or QToF), Cysteine protecting group (Carbamidomethylation +57), Protease (Trypsin, None), 2Da Parent mass tolerance, 0.5Da Ion tolerance, no post-translational modifications and including common contaminants (digestion enzymes and Human Keratins). V8 digested runs were searched as above but with the protease parameter set to ‘None’.

3 RESULTS

The NeuroPedia spectral library contains a total of 3401 identified spectra in 10 MGF files as described in Supplementary Table. In addition to providing libraries for all identified spectra, NeuroPedia also contains libraries of manually validated high/low-quality spectra for unique combinations of peptide sequence and precursor charge states. The NeuroPedia sequence database contains 847 neuropeptides from human, chimpanzee, mouse, rat, cow, California sea hare, rhesus macaque and medicinal leech. Using InsPect or any other database search tool, new MS/MS data can be searched against this sequence database. As shown in Total/UniProt in Supplementary Table, searching against NeuroPedia identifies many more spectra than the UniProt database. Out of all possible 340725 pairs of neuropeptides sequences (without considering species), there are 531 pairs with identical sequences (type a), 5020 pairs with overlapping sequences (type b), and 9185 pairs with homolog sequences (type c). A detailed description of the format of the databases is provided in Supplementary Materials.

4 CONCLUSIONS

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

NeuroPedia can be accessed at http://proteomics.ucsd.edu/Software/NeuroPedia.html.

Funding: This work was partially supported by the National Institutes of Health grant 1-P41-RR024851 from the National Center for Research Resources (to N.B.) and National Institutes of Health grants 5K01DA23065 (to S.B.) and R01 NS24553, R01 DA04271, R01 MH077305, and P01 HL58120 (to V.H.).

Conflict of Interest: none declared.

REFERENCES

Jain,E. et al. (2009) Infrastructure for the life sciences: design and implementation of the UniProt website. BMC Bioinformatics, 10, 136.