Using the NCBI Peptidome Repository

Douglas J. Slotta
National Center For Biotechnology Information, National Library of Medicine, National Institutes of Health, 45 Center Drive, Bethesda, MD 20894

http://www.ncbi.nlm.nih.gov/peptidome/

Submission Requirements

The submission process requires the deposit of four types of data:

- *Metadata*: Information about the study, sample(s), instrument, and data processing.
- *Spectra*: Raw data in an open format, such as mzXML or mzML.
- *Search output*: (optional) Search engine output files, such as Mascot, OMSSA, pepXML, or mzIdentML, to annotate the results.
- *Results*: The list of identified proteins, peptides, and spectra.

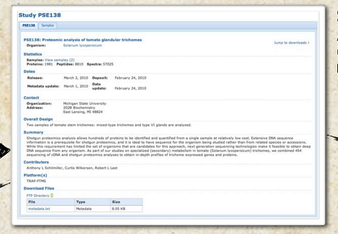
ProteomExchange

A consortium of proteomics repositories formed to facilitate the transfer of data between repositories, while negating the need for a client to submit to multiple sites. Members also include (but are not limited to) Pride, PeptideAtlas, and Tranche.

Peptidome

A public repository for the results and supporting data for tandem mass spectrometry-based Proteomics experiments.

43 studies



Studies

A collection of related *Samples*. A *Study* usually contains the data that one would place into a publishable paper.

310 samples

Why should data be submitted to Peptidome?

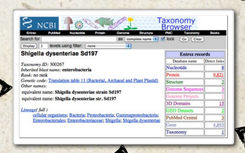
- Provides a location for a publication's supplementary data.
- *Studies* and *Samples* are issued stable accessions for ease of referral.
- All submitted data may be held private until published.
- Human curators will facilitate the submission process.
- Integrates experimental results with the rest of NCBI's datasets.
- Allows broader dissemination of results beyond the field of mass spectrometry-based proteomics.
- Ensures compliance with the data sharing policies of NIH and other funding agencies.
- Serves as a primary data source for researchers who study peptide identification or protein inference, for those who wish to build spectral libraries, or for other activities that will improve experimental analysis for everyone.
- It is free!

Entrez

Entrez is the integrated, text-based search and retrieval system used at NCBI for the major databases, including PubMed, Nucleotide and Protein Sequences, Protein Structures, Complete Genomes, Taxonomy, and others.

172 public studies and samples

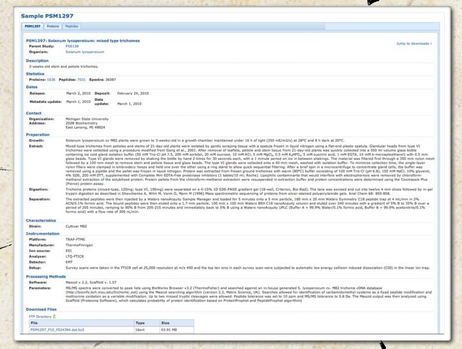
590,344 taxa



Taxonomy

The NCBI taxonomy database contains the names of all organisms that are represented in the genetic databases with at least one nucleotide or protein sequence.

20 taxa

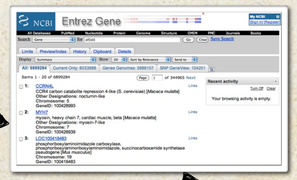


Samples

A container of all the *proteins*, *peptides*, and *spectra* from a single biological source. It may contain data from one or more instrument runs (see *Fractions*).

68,992,278 genes

32,862,759 proteins



Gene

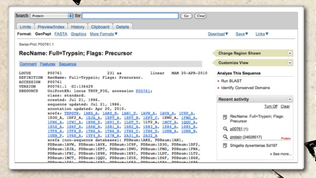
Entrez Gene is a searchable database of genes, from RefSeq genomes, and defined by sequence and/or located in the NCBI Map Viewer.

251 taxa

60,436 proteins

Retrieving the Data

Data from Peptidome may be downloaded for all public *Samples* and *Studies* via FTP. The data are available in *PeptidomeXML* format, as well as the original source data and any supplementary files. For larger transfers, the high-performance Aspera file transfer client may also be used.

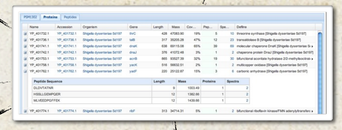


Protein (Entrez)

The protein entries in the Entrez search and retrieval system have been compiled from a variety of sources, including SwissProt, PIR, PRF, PDB, and translations from annotated coding regions in GenBank and RefSeq.

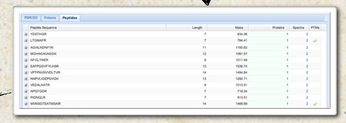
29,187 genes

267,166 unique peptides



Proteins

A tabular list of all proteins identified in a single *Sample*, including the number of peptides and spectra used for each identification. Links are included to the Entrez Gene and Protein databases, and results may be visualized with the NCBI Sequence Viewer.



Peptides

A tabular list of all peptides identified in a single *Sample*, including the number of spectra used for the identification. A link to all proteins (within the same Sample) is also provided.

46,098 proteins

46,098 proteins

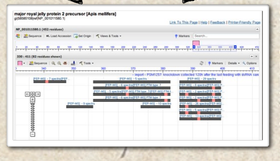
46,098 proteins

4,338,544 peptide identifications

4,338,544 peptide identifications

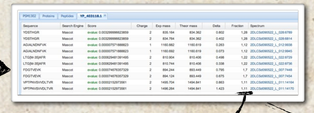
How big is Peptidome?

43	Studies
310	Samples
60,436	Proteins
267,166	Peptides
3,692,671	Identified MS^2 Spectra
52,099,517	MS Spectra
48,148,925	MS^2 Spectra
21,499	MS^3 Spectra
3.3 GB	MSSQL Database size
424 GB	Spectra Files (msHDF5)
559 GB	Supplemental files



Sequence Viewer

The NCBI Sequence Viewer has been developed to be the graphical display for the Nucleotide and Protein databases. The Graphical Sequence Viewer is the web version of NCBI's Genome Workbench. They are based on the same code libraries and developed by the same team.



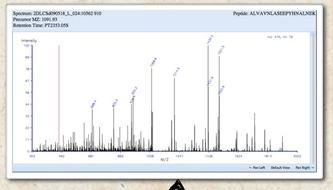
Peptide Identifications

A tabular list of information showing how each spectrum was identified as a given peptide. Contains the quality scores, post-translational modifications, experimental and theoretical masses, etc.

Fraction

This contains all of the *spectra* from a single instrument run. The source is typically a spot on a 2D gel or a band from a gel column. There may be one or more fractions per sample.

3,692,671 spectra



Spectra

All spectra are stored in msHDF5, which is a binary transformation and mapping of a set of mzXML and mzML files Into HDF5. This allows a drastic decrease in storage requirements while still allowing nearly instantaneous retrieval of individual spectra.

For example, a given data.mzXML file is:

- 845MB uncompressed
- 342MB bzip2 compressed
- 266MB msHDF5

Acknowledgements

This work would not be possible without the development efforts of Dmitry Rudnev, Dennis Troup, Rolf Muertter, Alexandra Soboleva, Maxim Tomashevsky, and Oluwabukunmi Ayanbule. Ji Li and Patti Sherman have performed nearly all of the data wrangling. Advice and support was provided by Ron Edgar, Tanya Barrett, Lewis Geer, Sechi Salvatore, and Sandy Markey. This research was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine.