Proteomics Data and Pathways/Ontology analysis

The resources below are sorted into three categories, 1. Proteomics data analysis, 2. Gene Ontology (GO)/pathways Analysis, and 3. Individual datasets with references. Note that to avoid misrepresentation, the information actually provided by the authors/websites is reiterated here without editing when possible, though not comprehensively, to assist in determining if the tools on the website will be useful. In addition, some of these resources are extensive in scope, and may include analysis tools and topics represented in a different Section.

Section 1. The first set is a collection of Proteomics oriented resources for learning about proteomics dataset analysis, analyzing and validating your proteomics data, uploading/downloading proteomics datasets, and much more. Find many standard prediction tools such as MS-Digest, MS-Product, MS-Isotope (Protein Prospector), searchable standard peptide MS/MS spectra databases such as SpectraST (available at Peptide Atlas), or a comprehensive set of protein analysis tools such as pairwise alignment (ClustalW), or BLAST to resolve isoform or polymorphism ambiguities that may arise from variation in accession information between databases, or to categorize proteins from new/unrepresented species.

Section 2. The second set are online database resources for ontology and pathway analysis. These resources can be used once you have a validated set of protein identifications, with unique peptides when necessary to resolve any questions about the precise family member or isoform identified in the data. These resources may center on gene designations rather than protein name: alternative spliceforms, isoforms, family member identities may need to be examined closely for accurate interpretation of the data.

HUPO, The Human Proteome Organization is relevant to all of the categories below. It is listed here first since it provides common ground for data analysis. The following is directly from the HUPO Proteomics Standards Intitiative:

The HUPO **Proteomics Standards Initiative** (PSI) defines community standards for data representation in proteomics to facilitate data comparison, exchange and verification.The main organizational unit of the Proteomics Standards Initiative is the work group. Currently, there are the following work groups:

- Molecular Interactions (MI)
- Mass Spectrometry (MS)
- Proteomics Informatics (PI)
- Protein Modifications (MOD)
- Protein Separation (PS)

The standard deliverables of each work group are

- Minimum Information Specification: For the given domain, this specifies the minimum information required for the useful reporting of experimental results in this domain.
- Formal exchange format for experimental results in the domain. This will usually be an XML format, capable of representing at least the Minimum Information, and normally significant additional detail.

Proteomics Data and Pathways/Ontology analysis

- Controlled vocabularies.
- Support for implementation of the standard in publicly available tools.

Section 1. Resources for analyzing and validating proteomics data

Protein Prospector

Proteomics tools for mining sequence databases in conjunction with Mass Spectrometry experiments.
http:prospector.ucsf.edu/
These programs were developed in the UCSF Mass Spectrometry Facility, which is directed by Dr. Alma Burlingame, Professor of Chemistry and Pharmaceutical Chemistry at UCSF and funded by the NIH National Institute for General Medical Sciences.

Peptide Atlas

http://www.peptideatlas.org/
PeptideAtlas is a multi-organism, publicly accessible compendium of peptides identified in a large set of tandem mass spectrometry proteomics experiments. Mass spectrometer output files are collected for human, mouse, yeast, and several other organisms, and searched using the latest search engines and protein sequences. All results of sequence and spectral library searching are subsequently processed through the Trans Proteomic Pipeline to derive a probability of correct identification for all results in a uniform manner to insure a high quality database, along with false discovery rates at the whole atlas level. Results may be queried and browsed at the PeptideAtlas web site. The raw data, search results, and full builds can also be downloaded for other uses.

PeptideAtlas is a product of the Seattle Proteome Center (SPC), a highly interactive, multi-disciplinary group aiming to enhance and develop innovative proteomic technologies and software tools and apply them to biological questions relevant to heart, lung, blood, and sleep.

This project has been funded in part with Federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, under contract No. N01-HV-28179.
PeptideAtlas is a module within SBEAMS, the Systems Biology Experiment Management System.

ExPASy Bioinformatics Resource Portal-Proteomics

http://www.expasy.org/proteomics/

In June 2011 the SIB ExPASy Bioformatics Resources Portal was launched by the SIB Swiss Institute of Bioinformatics. In particular, ExPASy has been designed, developed and maintained by the SIB Web Team in co-operation with several other SIB groups and ExPASy users. ExPASy is an extensible and integrative portal accessing many scientific resources, databases and software tools in different areas of life sciences. Refer to Features for more details. The portal enhances the original ExPASy server, previously known as "Expert Protein Analysis System"

Proteomics Data and Pathways/Ontology analysis

Originally, ExPASy was created in August 1993: it was one of the first Web servers for biological sciences. Since that date it has undergone constant modifications and improvements. Please contact the ExPASy Helpdesk if you have comments or encounter problems when using it.

In case of scientific publication, the SIB Swiss Institute of Bioinformatics should be mentioned. You can cite the following publication:

Artimo P, Jonnalagedda M, Arnold K, Baratin D, Csardi G, de Castro E, Duvaud S, Flegel V, Fortier A, Gasteiger E, Grosdidier A, Hernandez C, Ioannidis V, Kuznetsov D, Liechti R, Moretti S, Mostaguir K, Redaschi N, Rossier G, Xenarios I, and Stockinger H. ExPASy: SIB bioinformatics resource portal, Nucleic Acids Res, 40(W1):W597-W603, 2012.

UCSD-Center for Computational Mass Spectrometry
Serving the Biomedical and Bioinformatics Research Community
http://proteomics.ucsd.edu/

Who we are
The Center for Computational Mass Spectrometry (CCMS) is the National
Biomedical Technology Research Resource (BTRR) funded by the National Institute of General Medical Sciences of the National Institutes of Health. This Resource has been established to serve the biomedical research community by developing and integrating new computational proteomics technologies for collaborative studies, disseminating the new software, and training scientists in its use.

What CCMS Does
Housed in the Department of Computer Science and Engineering at UCSD, CCMS develops and supports a powerful suite of proteomic data analysis tools as well as operates a large computing infrastructure available to the biomedical community and mass spectrometry researchers.

Section 2. Resources for Gene Ontology (GO) and pathways analysis

These resources generally assume that you've accurately identified isoforms before you check pathways and ontology. The key to being certain about which isoform or family member you have is at least one statistically significant unique peptide which unambiguously identifies the isoform/spliceform.

http://geneontology.org/

The Gene Ontology Project
The Gene Ontology (GO) project is a major bioinformatics initiative to develop a computational representation of our evolving knowledge of how genes encode biological functions at the molecular, cellular and tissue system levels. Biological systems are so complex that we need to rely on computers to represent this knowledge. The project has developed formal ontologies that represent over 40,000 biological concepts, and are constantly being revised to reflect new discoveries. To date, these concepts

have been used to "annotate" gene functions based on experiments reported in over 100,000 peer-reviewed scientific papers.

The Gene Ontology project provides a controlled vocabulary of terms for describing gene product characteristics and gene product annotation data from GO Consortium members, as well as tools to access and process these data. Read more about the Gene Ontology.

The Gene Ontology Consortium (GOC) includes groups from around the world that collaborate closely on the development of the Gene Ontology and creation of gene function annotations.
The GOC is co-directed by (in alphabetical order):

Judith Blake, Jackson Laboratory (mouse gene annotation, ontology development)
J. Michael Cherry, Stanford University (yeast gene annotation, data production processes)
Suzanna Lewis, Lawrence Berkeley National Laboratory (GOC software development, ontology development)
Paul Sternberg, Caltech (C. elegans gene annotation, Common Annotation Framework development)
Paul Thomas, University of Southern California (phylogenetic annotation, ontology development)

Read more about the GO Consortium member groups.
The GO Council consists of the heads of major participating groups (in alphabetical order):

Alex Bateman, UniProt, European Bioinformatics Institute (gene annotation)
Nick Brown, FlyBase, University of Cambridge (Drosophila gene annotation)
Rex Chisholm, DictyBase, Northwestern University (Dictyostelium gene annotation)
James Hu, Texas A&M (gene annotation)
Eva Huala, The Arabidopsis Information Resource, Carnegie Institution for Science (Arabidopsis gene annotation)
Claire O'Donovan, UniProt, European Bioinformatics Institute (gene annotation)
Helen Parkinson, Ontologies Team lead, European Bioinformatics Institute (ontology development)
Monte Westerfield, University of Oregon (zebrafish gene annotation)

GOPubMed
http://www.gopubmed.org/web/gopubmed/
GoPubMed® allows users to find information significantly faster and guarantees completeness of search results. The fundamental difference between GoPubMed's® semantic search technology and traditional search engines such as PubMed or Google is the use of background knowledge. Semantic algorithms connect text – abstracts from the MEDLINE database – to background knowledge in the form of semantic networks of concept categories, also called ontologies or knowledge base. This is done by meaning and not by keywords only. So results are meaningfully structured and intelligent semantic navigation becomes possible. The concept categories come from the Gene Ontology (GO), the Medical Subject Headings (MeSH), the Universal Protein Resource (UniProt), Authors, Locations, Journals, and Publication Dates. In GoPubMed® the user does the ranking. Examples below clearly demonstrate this semantic power.

Proteomics Data and Pathways/Ontology analysis

Kyoto Encyclopedia of Genes and Genomes
http://www.genome.jp/kegg/

KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from genomic and molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies. It is a computer representation of the biological system, consisting of molecular building blocks of genes and proteins (genomic information) and chemical substances (chemical information) that are integrated with the knowledge on molecular wiring diagrams of interaction, reaction and relation networks (systems information). It also contains disease and drug information (health information) as perturbations to the biological system.

The KEGG database has been in development by Kanehisa Laboratories since 1995, and is now a prominent reference knowledge base for integration and interpretation of large-scale molecular data sets generated by genome sequencing and other high-throughput experimental technologies.

KEGG is an integrated database resource consisting of the seventeen main databases shown below. They are broadly categorized into systems information, genomic information, chemical information and health information, which are distinguished by color coding of web pages.

DAVID (Database for Annotation, Visualization and Integrated Discovery)
http://david.abcc.ncifcrf.gov/
"DAVID bioinformatics resources consists of an integrated biological knowledgebase and analytic tools aimed at systematically extracting biological meaning from large gene/protein lists."
Reference: Nature Protocols 4, - 44 - 57 (2009), Da Wei Huang, Brad T Sherman & Richard A Lempicki

The Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.7 is an update to the sixth version of our original web-accessible programs. DAVID now provides a comprehensive set of functional annotation tools for investigators to understand biological meaning behind large list of genes. For any given gene list, DAVID tools are able to:

Identify enriched biological themes, particularly GO terms
Discover enriched functional-related gene groups
Cluster redundant annotation terms
Visualize genes on BioCarta & KEGG pathway maps
Display related many-genes-to-many-terms on 2-D view.
Search for other functionally related genes not in the list
List interacting proteins
Explore gene names in batch
Link gene-disease associations
Highlight protein functional domains and motifs
Redirect to related literatures
Convert gene identifiers from one type to another

Proteomics Data and Pathways/Ontology analysis

Pathway Commons

http://www.pathwaycommons.org/

Pathway Commons is a collection of publicly available pathway information from multiple organisms. It provides researchers with convenient access to a comprehensive collection of biological pathways from multiple sources represented in a common language for gene and metabolic pathway analysis. Access is via a web portal for query and download. Database providers can share their pathway data via a common repository and avoid duplication of effort and reduce software development costs. Bioinformatics software developers can increase efficiency by sharing pathway analysis software components. Pathways can include biochemical reactions, complex assembly, transport and catalysis events, physical interactions involving proteins, DNA, RNA, small molecules and complexes, gene regulation events and genetic interactions involving genes.