Nucleic Acids Research, 2006, Vol. 34, Database issue ***D507–D510***
doi:10.1093/nar/gkj122

LMPD: LIPID MAPS proteome database

Dawn Cotter[1], Andreia Maer[1], Chittibabu Guda[2], Brian Saunders[1] and Shankar Subramaniam[1,3,*]

[1]San Diego Supercomputer Center, University of California, 9500 Gilman Drive, La Jolla, CA 92093-0505, USA, [2]Gen*NY*sis Center for Excellence in Cancer Genomics and Department of Epidemiology and Biostatistics, University at Albany, State University of New York, One Discovery drive, Rensselaer, NY 12144-3456, USA and [3]Departments of Bioengineering, Chemistry and Biochemistry, University of California, 9500 Gilman Drive, La Jolla, CA 92093-0612, USA

Received August 15, 2005; Revised and Accepted October 20, 2005

ABSTRACT

The LIPID MAPS Proteome Database (LMPD) is an object-relational database of lipid-associated protein sequences and annotations. The initial release contains 2959 records, representing human and mouse proteins involved in lipid metabolism. UniProt IDs were obtained based on keyword search of KEGG and GO databases, and this LMPD protein list was then enhanced with annotations from UniProt, EntrezGene, ENZYME, GO, KEGG and other public resources. We also assigned associations with general lipid categories, based on GO and KEGG annotations. Users may search LMPD by database ID or keyword, and filter by species and/or lipid class associations; from the search results, one can then access a compilation of data relevant to each protein of interest, cross-linked to external databases. The LIPID MAPS Proteome Database (LMPD) is publicly available from the LIPID MAPS Consortium website (http://www.lipidmaps.org/). The direct URL is http://www.lipidmaps.org/data/proteome/index.cgi.

INTRODUCTION

Lipids play central roles in energy storage, cell membrane structure, cellular communication and regulation of biological processes such as inflammatory response, neuronal signal transmission and carbohydrate metabolism. They are furthermore known to be involved in many disease states, including Alzheimer's, asthma, cancer, malaria and rheumatoid arthritis.

The LIPID Metabolites and Pathways Strategy (LIPID MAPS) Consortium represents a multi-institutional effort to develop a detailed understanding of lipid structure and function. As part of this effort, we will develop 'parts lists' of lipid metabolites and assemble these into metabolic networks. These networks will then provide an infrastructure for subsequent modeling using quantitative data from LIPID MAPS experiments. LMPD embodies the protein, gene and pathway parts lists for these networks.

Existing lipid databases include LIPIDAT (1) (http://www.lipidat.chemistry.ohio-state.edu/), LIPID BANK for Web (http://www.lipidbank.jp/) and LIPIDBASE (http://www.lipidbase.jp/category.html). They are primarily organized around the properties and structures of lipids, while LMPD is focused on proteins and genes associated with all lipids. The Arabidopsis Lipid Gene Database focuses on *Arabidopsis thaliana* and includes genes and proteins that are involved in acyl-lipid metabolism (2).

BIOINFORMATICS

Identifying lipid-associated proteins

A new classification system has recently been developed for lipids (3). This classification system organizes lipids into eight main classes, with two levels of subclasses. The top-level lipid categories are fatty acyls, glycerolipids, glycerophospholipids, sphingolipids, sterol lipids, prenol lipids, saccharolipids and polyketides. A list of lipid-related GO (Gene Ontology, http://www.geneontology.org) (4) terms and KEGG (Kyoto Encyclopedia of Genes and Genomes, http://www.genome.ad.jp/kegg) (5) pathways was compiled, using lipid-specific keywords, such as trivial names of classes, subclasses and individual lipid compounds. The UniProt (6) proteins annotated with those GO and KEGG terms were then collected. The GO terms identify lipid-related enzymatic activity and metabolic processes; KEGG terms identify lipid-related pathways. The proteins are associated with one or multiple lipid classes based on these GO/KEGG annotations. By this process we have identified ∼1600 human proteins and ∼1300 mouse proteins in UniProt. About 2500 out of the 2900 proteins are associated

*To whom correspondence should be addressed at 9500 Gilman Drive, MC-0505, San Diego, CA 92093-0505, USA. Tel: +1 858 822 0986; Fax: +1 858 822 3752; Email: shankar@sdsc.edu

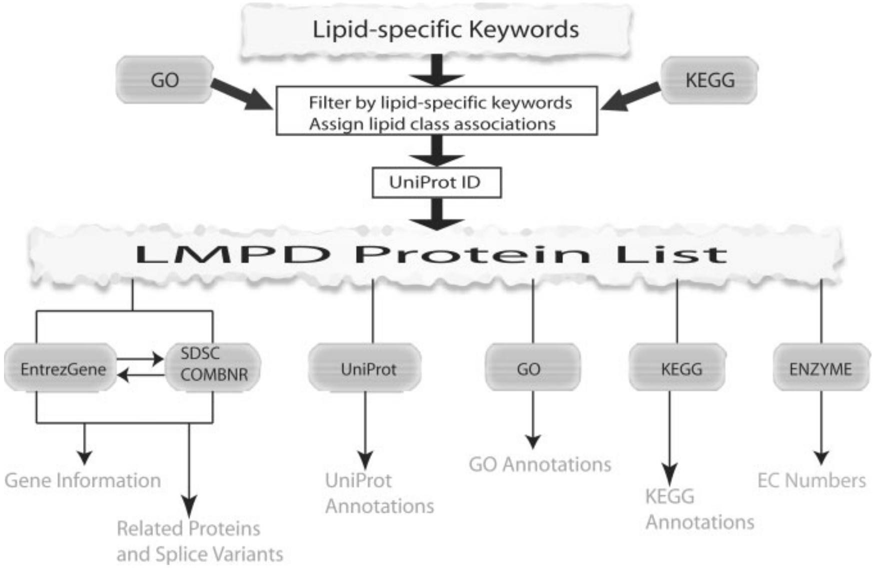

Figure 1. Overview of Bioinformatics Process.

with at least one of the eight main lipid classes. An overview of the bioinformatics process is illustrated in Figure 1.

Protein and gene annotation

Protein annotations were primarily obtained from UniProt (7) and include UniProt accession, UniProt entry name, domain information and Swiss-Prot (8) comments such as function, catalytic activity, subcellular location and similarity. Accession numbers were used to link to various public databases and collect available information for each protein. From NCBI EntrezGene (8) we collected gene information, such as Gene ID, alternate names/synonyms and symbols, chromosomal mapping and cross-references to other databases.

The 2959 proteins comprising the LMPD protein list correspond to ∼2300 unique genes. We also gathered all the GO and KEGG annotations, not just the lipid-specific ones, using Gene IDs and UniProt accessions. Other protein records with sequences that are (i) identical with the ones in our list, (ii) splice variants of those or (iii) related (from the same gene/locus) were gathered using EntrezGene and an in-house generated non-redundant protein sequence database compiled from the most well-known public protein sequence databases such as Swiss-Prot, Trembl and GenBank (9).

Enzyme information

EC numbers were obtained from KEGG ENZYME and UniProt ENZYME. A single protein may be associated with multiple EC numbers and multiple proteins may be associated with the same EC number. EC numbers were then used to obtain information such as enzyme name and synonyms, reaction, substrate(s) and product(s), from ExPASy ENZYME (10) (Enzyme nomenclature database) and KEGG ENZYME (11) database.

Pathways

Proteins/genes that have associated KEGG annotations or EC numbers are hyperlinked/mapped to KEGG metabolic pathways. Future work will include manual mapping of those proteins to more detailed/specific lipid metabolic pathway maps such as SphinGOMAP (http://www.sphingomap.com) (12) and to signaling pathways.

DATABASE IMPLEMENTATION AND USER INTERFACES

LMPD is implemented as an object-relational database, using Oracle9i Enterprise Edition Release 9.2.0.2.0, running on a Sun Fire 880. Perl scripts and Oracle SQL*Loader were used to parse and load flat-file data into Oracle database tables. The LMPD graphical user interface (GUI) is based on Perl, and is served by the Apache 1.3.26 web server, running on a Sun Ultra-80. Both Sun machines are running Solaris 9.

An entity-relationship diagram is available as Supplementary Data at http://www.lipidmaps.org/data/proteome/supplementarymaterial/NAR/2005/ER_diagram.gif.

Query forms

The default query form allows the user to browse the protein list, with an option to browse by associated lipid category.

The 'Advanced' query form provides options for conducting a more focused search, including options to search by database ID or keyword and to filter by species and/or lipid class association. Database ID fields searched include UniProt accession, UniProt entry name, gene symbols, GenBank GI, EC number, GO ID and KEGG pathway ID. Keyword search fields include Uniprot description and Swiss-Prot comments.

Results summary

The results summary page presents a sortable list of proteins matching the query criteria, along with selected summary information, including LMPD_ID, accession, protein name, protein symbol and associated lipid categories. From the summary page, the user may display complete LMPD annotations for each protein.

Record details

For each record selected from the results summary, all LMPD data relevant to that protein are displayed, with external database IDs linked to their respective resources.

Annotations are organized by category: Record Overview, Gene/GO/KEGG Information, UniProt Annotations, and Related Proteins. The record overview contains LMPD_ID, species, description, gene symbols, lipid categories, EC number, molecular weight, sequence length and protein sequence. Gene information includes Entrez Gene ID, chromosome, map location, primary name, primary symbol and alternate names and symbols; Gene Ontology (GO) IDs and descriptions, and KEGG pathway IDs and descriptions. UniProt annotations include primary accession number, entry name and comments such as catalytic activity, enzyme regulation, function and similarity. For related proteins and splice variants, we display source database, database ID, sequence length, and title.

DISCUSSION

The initial release of LMPD establishes a framework for creating a lipid-associated protein list, collecting relevant annotations, databasing this information and providing a user interface. This initial release includes data collected from mouse and human taxonomies.

We have chosen to gather lipid-associated proteins and associated them with lipid classes based on GO and KEGG annotations because their data are high quality, annotated and curated with the help of experts and literature. Because an automatic keyword search has some inherent pitfalls such as having false positives, we will continuously try to refine our keywords list in order to minimize the number of false positives. In the next major release (planned for summer of 2006), we will improve and refine our keyword list to capture additional proteins that we may have missed. Expanding the knowledge-based keyword search from GO and KEGG annotations to protein and gene names and synonyms, and to enzyme names, substrates and products names will also greatly reduce the number of unassigned proteins. We will also include in the next release additional annotations, such as those from ENZYME and OMIM (Online Mendelian Inheritance in Man) [McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD), 2000. world wide web URL: http://www.ncbi.nlm.nih.gov/omim/]. In subsequent releases we will add annotations from Ensembl (13), Homologene (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=homologene) and Unigene (14).

We will also add other species, such as *Escherichia coli* and *Saccromyces cerevisiae* in subsequent releases. We intend to have two major releases a year, but we will be updating the annotation of the existing records every couple of months.

Future work will also include integration of LMPD with the LIPID MAPS Lipid Structure database. This process will involve expert help from the LIPID MAPS Consortium and will be done mostly manually. For the LMPD proteins that have EC numbers, a semi-automatic part of it will involve using the enzyme annotation from KEGG database, finding the substrates and products which are lipids and then mapping those lipids to the LIPID MAPS Lipid structure database based on name and structure matching if available. The expert knowledge and manual work will also help with the lipid class and sub-class assignment of proteins. This mapping of proteins to lipid classes and subclasses will permit users to search for proteins and then access data from corresponding lipids, and vice versa.

We ultimately aim to develop lipid interaction networks that will integrate lipid metabolic pathways and signaling networks, and tools for exploring these networks. These networks and tools, coupled with LIPID MAPS experimental data, may provide insight into the biological processes underlying lipid-involved disease processes and lead to the identification of potential drug targets.

ACKNOWLEDGEMENTS

This work was supported by National Institutes of Health (NIH) National Institute of General Medical Sciences (NIGMS) Glue Grant NIH/NIGMS Grant 1 U54 GM69338. Funding to pay the Open Access publication charges for this article was provided by National Institutes of Health (NIH), National Institute of General Medical Sciences (NIGMS) Glue, Grant NIH/NIGMS Grant 1 U54GM69338.

Conflict of interest statement. None declared.

REFERENCES

1. Caffrey,M. and Hogan,J. (1992) LIPIDAT: a database of lipid phase transition temperatures and enthalpy changes. DMPC data subset analysis. *Chem. Phys. Lipids*, **61**, 1–109.
2. Beisson,F., Koo,A.K.K., Ruuska,S., Schwender,J., Pollard,M., Thelen,J., Paddock,T., Salas,J., Savage,L. and Milcamps,A.*et.al*. (2003) *Arabidopsis thaliana* genes involved in acyl lipid metabolism. a 2003 census of the candidates, a study of the distribution of expressed sequence tags in organs, and a web-based database. *Plant Physiol*., **132**, 681–697.
3. Fahy,E., Subramaniam,S., Brown,H.A., Glass,C.K., Merrill,A.H.Jr, Murphy, R.C., Raetz,C.R., Russell,D.W., Seyama,Y. and Shaw,W.*et. al*. (2005) A comprehensive classification system for lipids. *J. Lipid Res*., **46**, 839–862.
4. Harris,M.A., Clark,J., Ireland,A., Lomax,J., .Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B. and Mungall,C., *et. al*. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*., **32**, D258–D261.
5. Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*., **28**, 27–30.
6. Apweiler,R., Bairoch,A., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R. and Magrane,M.*et. al*. (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res*., **32**, D115–D119.
7. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C. and Phan,I.*et. al*. (2003) The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*., **31**, 365–370.
8. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*., **33**, D54–D58.
9. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2003) GenBank. *Nucleic Acids Res*., **31**, 23–27.
10. Bairoch,A. (2000) The ENZYME database in 2000. *Nucleic Acids Res*., **28**, 304–305.
11. Goto,S., Nishioka,T. and Kanehisa,M. (2000) LIGAND: chemical database of enzyme reactions. *Nucleic Acids Res*., **28**, 380–382.

12. Merril,A.H.,Jr (2005) SphinGOMAP—a web-based biosynthetic pathway map of sphingolipids and glycosphingolipids. *Glycobiology*, **15**(6), 15G.
13. Hubbard,T., Andrews,D., Caccamo,M., Cameron,G., Chen,Y., Clamp,M., Clarke,L., Coates,G., Cox,T. and Cunningham,F.*et. al*. (2005) Ensembl 2005. *Nucleic Acids Res*., **33**, D447–D453.
14. Wheeler,D.L., Church,D.M., Federhen,S., Lash,A.E., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E. and Tatusova,T.A.*et. al*. (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Res*., **31**, 28–33.