[Frontiers in Bioscience 12, 5071-5088, September 1, 2007]

Integration of bioinformatics resources for functional analysis of gene expression and proteomic data

Hongzhan Huang, Zhang-Zhi Hu, Cecilia N. Arighi, Cathy H. Wu

Protein Information Resource (PIR), Department of Biochemistry and Molecular and Cellular Biology, Georgetown University Medical Center, 3300 Whitehaven Street NW, Suite 1200, Washington, DC 20007, USA

TABLE OF CONTENTS

1. ABSTRACT

In the post-genome era, researchers are systematically tackling gene functions and complex regulatory processes by studying organisms on a global scale; however, a major challenge lies in the voluminous, complex, and dynamic data being maintained in heterogeneous sources, especially from proteomics experiments. Advanced computational methods are needed for integration, mining, comparative analysis, and functional interpretation of high-throughput proteomic data. In the first part of this review, we discuss aspects of data integration important for capturing all data relevant to functional analysis. We provide a list of databases commonly used in genomics and proteomics and explain strategies to connect the source data, with especial emphasis on our ID mapping service. Next, we describe iProClass, a central data infrastructure that supports both data integration and functional annotation of proteins, and give a brief introduction to the data search/retrieval and analysis tools currently available at our website (http://pir.georgetown.edu) that researchers can use for large-scale functional analysis. In the last part, we introduce iProXpress (integrated Protein eXpression), an integrated research and discovery platform for large-scale expression data analysis, and we show a prototype that has been useful for organelle proteome analysis.

2. INTRODUCTION

The traditional one-gene-at-a-time approach, though critical for revealing detailed molecular functions of individual genes, does not provide a global view of gene function or of the temporal and spatial regulation of all genes at different physiological states or developmental stages. In the post-genome era, researchers are beginning to systematically tackle gene functions and complex regulatory processes by studying organisms on a global scale: genomes, transcriptomes, proteomes, metabolomes (1), and interactomes (2).

Gene expression profiling using microarray technologies has been extensively used in the past decade. An enormous amount of data and knowledge have been obtained from studies ranging from yeast gene expression to human cancer biomarker identification. In addition to expression profiling at the transcript level, large-scale expression profiling at the protein level using mass spectrometry technology is now a widely used approach. Proteomics aims to identify, characterize, and quantify all proteins expressed in cells grown under a variety of conditions (3). As most physiological, developmental, and pathological processes are manifested at the protein level, proteomics has unique and significant advantages as an important complement to genomic and transcriptomic approaches. As a result, there is intense interest in applying proteomics to foster a better understanding of basic biology, as well as of disease processes (4, 5). Proteomics has been widely used for the analysis of complex biological systems (6). The use of quantitative mass spectrometry provides a powerful approach to the comprehensive analysis of macromolecular complexes (7).

The accelerated growth of proteomics and other large-scale "omics" data presents both opportunities and challenges. Numerous bioinformatics databases have been developed to organize and annotate the biological information for individual genes and proteins, and to facilitate sequence and functional analyses of genes and proteins. The collective richness of the data allows researchers to ask complex biological questions and gain new scientific insights, as illustrated in the integrated global profiling approach for studying the molecular basis of human cancer (8). One major challenge, however, lies in the voluminous, complex, and dynamic data being maintained in heterogeneous sources, particularly for proteomic data, due to the complexity of protein dynamics and the vast amount of information required for characterizing the proteome. Advanced computational methods are needed for integration, mining, comparative analysis, and functional interpretation of high-throughput proteomic data.

Due to its robustness, sensitivity and efficiency, tandem mass spectrometry (MS/MS) has become the method of choice for identification of proteins in high-throughput proteomics studies (9). This approach subjects protein mixtures to proteolytic digestion prior to liquid chromatography separation and MS/MS analysis of the resulting peptides. Many bioinformatics tools have been developed for the management and analysis of proteomics data (e.g., http://www.proteomecommons.org/tools.jsp). A number of database search programs (e.g., SEQUEST (10), Mascot (11) and X!Tandem (12)) are used to assign probable peptide sequences to MS/MS spectra and to infer protein identities. Tools such as PeptideProphet (13) and ProteinProphet (14) are designed to improve the accuracy of peptide and protein identification, while DBParser (15) consolidates redundant protein assignments. Several search algorithms have been benchmarked for sensitivity and specificity (16) and a pipeline developed for experiment annotation, database searching, peptide mining, and protein identification (17).

Once proteins are identified from the biological samples, they can be analyzed for functions and processes. Many programs have been developed for the biological interpretation of large lists of genes, mostly for microarray gene expression data, with a few being extended to proteomics data. As the Gene Ontology (GO) (18) has become the common standard for genome annotation, most programs provide functional analysis in the context of GO (http://www.geneontology.org/GO.tools.microarray.shtml). A few examples include: (i) GoMiner (19), which presents genes in GO hierarchical views, (ii) MAPPFinder (20), integrating GO with GenMAPP pathways, (iii) NetAffx (21), rendering GO graphs to display Affymetrix probe sets, (iv) DAVID (22), which includes additional information on Pfam domains (23) and KEGG pathways (24), and (v) Babelomics (25), which includes InterPro (26), KEGG, and Swiss-Prot keywords (27).

While bioinformatics tools have greatly assisted proteomic data analysis, a careful review of the major steps and flow of data in a typical high-throughput analysis reveals gaps that need to be addressed. There are unmet needs in the areas of protein identification and functional interpretation. One issue in proteomics and MS/MS protein identification is that the available general purpose protein sequence databases leave out many alternative splice isoforms. As a result, proteomics analysis may fail to identify bona fide protein products of alternative splice isoforms because the target sequence was not present in the database being searched. A second issue is the lack of standardization when dealing with a large list of proteins annotated in different places. Different protein IDs/names may be used for the same protein if a different underlying database is used for MS/MS protein search. Even different versions of the same database may result in different IDs if the database identifier is not stable. The lack of standards presents a challenge for integrating annotations from heterogeneous sources for biological interpretation of proteomic data. Consequently, expression analysis is often carried out in an *ad hoc* manner, with a fragmented and inefficient use of rich annotations available in numerous resources. In this article we will discuss the mapping and integration of heterogeneous molecular biology databases and the functional analysis of gene expression and proteomic data using an integrated expression analysis system (iProXpress). We then use organelle proteome analysis as a case study to illustrate the application of the iProXpress system.

3. DATA INTEGRATION

Molecular biology databases often have large volume and complex data structure. They are distributed through the Internet and are organized in a wide variety of formats. Each year, the Journal Nucleic Acids Research publishes an update for a list of several hundreds of molecular databases in its annual database issue (28). To effectively utilize this vast amount of data, it is essential to provide researchers with an integrated view of all data relevant to functional analysis. Using such integrated view, researchers can uncover important biological relationships among the large set of genes/proteins from their experiments for scientific discovery.

3.1. Data Sources

Databases commonly used in functional analysis of gene expression and proteomic data are listed in Table 1 under 13 categories: protein sequence, gene and genome, taxonomy, gene expression, protein peptide ID databases, protein expression, function and pathway, genetic variation and disease, ontology, interaction, modification, structure, and classification. Most of these molecular biology databases are uniquely structured, reflecting different underlying biological models. In order to use these data

Table 1. Molecular biology databases commonly used in functional analysis of gene expression and proteomic data

Database	Description	URL
Protein Sequence		
UniProtKB, UniRef100, UniParc	UniProt protein sequences	http://www.uniprot.org
RefSeq, GenPept, NR	NCBI protein sequences	http://www.ncbi.nlm.nih.gov
Gene and Genome		
GenBank/EMBL/DDBJ	DNA sequence databases	http://www.ncbi.nlm.nih.gov
UniGene	Non-redundant set of eukaryotic gene-oriented clusters of transcript sequences	http://www.ncbi.nlm.nih.gov
FlyBase	*Drosophila* sequences and genomic information	http://flybase.bio.indiana.edu
MGD	Mouse genome database	http://www.informatics.jax.org
SGD	*Saccharomyces* genome database	http://db.yeastgenome.org
WormBase	Data repository for *C.elegans* and *C.briggsae*	http://www.wormbase.org
TAIR	The *Arabidopsis* information resource	http://arabidopsis.org
TIGR	TIGR Microbial Database	http://www.tigr.org
Taxonomy		
NCBI Taxon	NCBI taxonomy database	http://www.ncbi.nlm.nih.gov
NEWT	UniProt taxonomy database	http://www.ebi.ac.uk/newt
Gene Expression		
GEO	gene expression profiles	http://www.ncbi.nlm.nih.gov
CleanEx	Expression reference database	http://www.cleanex.isb-sib.ch
SOURCE	Functional genomics resource for human, mouse and rat	http://source.stanford.edu
Proteomic Peptide ID Databases		
GPMDB	Global Proteome Machine Database	http://gpmdb.thegpm.org
PRIDE	PRoteomics IDEntifications database	http://www.ebi.ac.uk/pride
PeptideAtlas	Peptide database identified by MS experiments	http://www.peptideatlas.org
Protein Expression		
Swiss-2DPAGE	Annotated 2D gel electrophoresis database	http://www.expasy.org
PMG	2D gel data from Protein mapping group	http://proteomes.pex.anl.gov
Function and Pathway		
EC-IUBMB	Enzyme Nomenclature	http://www.chem.qmul.ac.uk
KEGG	Metabolic and regulatory pathways	http://www.genome.ad.jp
BioCyc	Microbial pathway/genome databases	http://biocyc.org
Genetic Variation and Disease		
OMIM	A catalog of human genetic and genomic disorders	http://www.ncbi.nlm.nih.gov
HapMap	Resource for human DNA sequence variation	http://www.hapmap.org
Ontology		
GO	Gene Ontology database	http://www.godatabase.org
Interaction		
IntAct	Protein–protein interaction data	http://www.ebi.ac.uk/intact
DIP	Database of interacting proteins	http://dip.doe-mbi.ucla.edu
Modification		
RESID	Post-translational protein modifications	http://srs.ebi.ac.uk/srsbin
Phosphosite	Database of phosphorylation sites	http://www.phosphosite.org
Structure		
PDB	Protein structure databank	http://www.rcsb.org
SCOP	Structural classification of proteins	http://scop.mrc-lmb.cam.ac.uk
CATH	Protein domain structures database	http://cathwww.biochem.ucl.ac.uk
MMDB	Database of 3D structures	http://www.ncbi.nlm.nih.gov/Structure/mmdb
PDBsum	Summaries and analyses of PDB structures	http://www.ebi.ac.uk/pdbsum/
Modbase	Annotated comparative protein structure models	http://modbase.compbio.ucsf.edu
Classification		
PIRSF	Family/superfamily classification of whole proteins	http://pir.georgetown.edu
UniRef50,90	UniProt non-redundant reference clusters	http://www.uniprot.org
PFam	Protein families of domains	http://www.sanger.ac.uk
InterPro	Integrated resource of protein families, domains and functional sites	http://www.ebi.ac.uk/interpro
PANTHER	Gene products organized by biological function	http://www.pantherdb.org/panther
COG	Clusters of orthologous groups of proteins	http://www.ncbi.nlm.nih.gov/COG
SMART	Resource for protein domain identification and the analysis of protein domain architectures	http://smart.embl-heidelberg.de/smart/
TIGRFAMs	TIGR protein families	http://cmr.tigr.org/tigr-scripts/CMR

effectively, one must understand the database schemas in each data source and their relationship. Data sources often contain overlapping or similar data elements, such as database identifiers, organism names, protein names and sequences, which are the keys to connecting the source data. However, there may be conflicting data definitions among data sources. Therefore, bioinformatics tools, such as ID mapping tools, are needed to uncover the relationship among databases and to map data from one database to another, regardless the names or descriptions that are given to corresponding objects and attributes in those databases.

As an example, we take a set of MS data from an organelle proteome study on various stages of melanosomes from human melanoma cell lines (43) to illustrate the use of source data for data mapping, integration and interpretation. This data set is a list proteins represented by NCBI gi numbers and/or peptide sequences. To analyze the data, first we need to map the protein to UniProt Knowledgebase (UniProtKB) database entries for rich annotation, then to integrate all protein information for functional analysis.

3.2. Data mapping

Common data elements serve as keys for data mapping. In primary sequence (protein or gene) databases, only four elements, namely, database unique identifier (UID), name, source organism, and sequence, are commonly used to identify a gene or protein object in databases. *UID* is the ID and/or accession number assigned by an individual database to uniquely identify each sequence entry. Unlike the gene object, where there is a widely acceptable UID (i.e., GenBank/EMBL/DDBJ ID), there is no standard UID for a protein object. Various secondary databases use different UIDs to reference the same object from different protein sequence databases. *Protein name* is the word or phrase used to indicate a specific protein object in the scientific literature and biological databases. There is, however, a long-standing problem of nomenclature for proteins, where "profligate and undisciplined labeling is hampering communication." as discussed in *Nature* (29). Scientists may name a newly discovered or characterized protein based on its function, sequence features, gene name, cellular location, molecular weight, or other properties, as well as their combinations or abbreviations. The same protein is often named differently in different databases, and occasionally different proteins may share the same name.

To unambiguously identify a protein object, one also needs to know the *source organism* of the protein, as different organisms may share the same protein sequence. While there is a widely adopted standard for taxonomy (i.e., NCBI taxonomy), several problems are associated with the taxonomy of source organisms and their mapping among primary sequence databases. The problem may stem from non-specific names provided by biologists during direct submission of DNA sequence entries to GenBank. Often such names cannot be mapped to the taxonomy even at the species level (such as *Mus* sp.). To be useful, the source organism information should be specific, including the strain (or cultivar for plant) if possible, but many early submissions do not contain this information. There may also be a mapping problem when two databases describe the same protein object at different taxonomy levels, such as species vs. strain. The primary source of *protein sequences* is conceptual translation of GenBank DNA sequences. There are many possible sources of error. The translations may be made with an incorrect genetic code, in the wrong reading frame, with incorrect splice boundaries, or without corrections for RNA-editing and translational frame-shifting. In some complete bacterial genome reports where peptide sequences were available, early initiators had been chosen 20% of the time. Cases have been observed of late initiators having been chosen, revealed as carboxyl ends of incomplete homology domains at amino ends of translations. Other types of sequence variation may result from representations of alternate splice/initiator forms, precursor and mature forms, identical sequences from different loci, allelic variants, and post-translational modifications.

Data mapping typically involves the aforementioned four data elements. BioThesaurus (http://pir.georgetown.edu/pirwww/iprolink/biothesaurus.shtml) provides very detailed mappings of a comprehensive collection of protein and gene names to UniProtKB entries (30). Among the four data mappings, ID mapping is the most common type, and we will discuss this in detail in the following section.

ID mapping can be one of two types: 1) mapping among the biological objects, for an example, mapping between gi number and UniProtKB AC is a protein to protein mapping; 2) mapping from biological objects to their properties, such as mappings from gi number to Pathway or to GO ID. This type of mapping typically will produce many-to-one mapping. We will focus on the first type of mapping. ID mapping establishes the links between database identifiers, an important step for data integration. ID mapping is often directly used in gene expression and proteomic data analysis. Because of the data source heterogeneity, mapping between database identifiers can be very complex. There are three basic approaches for establishing the relationship between two IDs: the first and the most straightforward approach is to use the database cross-references from well curated databases, such as UniProtKB, which includes many cross-references in each protein entry, such as GenBank accession, KEGG pathway and GO ID. The second approach is to use other database identifiers as a bridge. For example, one can use GenBank accession to map from NCBI gi number to UniProtKB accession because both gi number and UniProtKB AC may cross-reference the same GenBank accession. Many mappings fall into this category. Finally, one can use computational approaches to establish the relationship between two database identifiers. For instance, one can compare the sequence identity between the entries from RefSeq and UniProtKB for the same organism to map the RefSeq accession and UniProtKB accession. Many methods, such as string match, BLAST (31) and FASTA (32), can be used for the sequence comparison. Usually, a 100% sequence identity is required for the mapping. In cases where this cannot be achieved, related entries with less than 100% identity will be mapped. The advantages of the computational approach are: 1) it does not require a third database; and 2) it can give useful mappings to related proteins, e.g., mapping uncharacterized protein entries to entries with valuable information, such as pathway information. The disadvantage of this approach is that it is usually computationally intense.

There are many ID mapping tools available on the web. AliasServer (http://cbi.labri.fr/outils/alias/) is a tool for identifier translation for about 35 species. MD Anderson GeneLink (http://bioinformatics.mdanderson.org/GeneLink.html) provides ID translation and search service for human IDs. MatchMiner (http://discover.nci.nih.gov/matchminer/index.jsp) provides ID translation service for mouse and human. Ariadne Genomics ID Mapping Service

Table 2. Database identifiers supported by PIR ID mapping service

Sequence ID	Other ID
ECOGENE ID	GO ID
EMBL ID	InterPro ID
Entrez Gene ID	Medline ID
FLYBASE ID	NCBI Taxon ID
GENEDB_SPOMBE ID	OMIM ID
GERMONLINE ID	PFAM ID
GI Number	PIRSF ID
GRAMENE ID	PRINTS ID
HIV ID	PRODOM ID
IPI ID	PROSITE ID
PDB ID	PubMed ID
PIR ID	SMART ID
REBASE ID	Taxon Group ID
Refseq Accession	TIGRFAMs ID
SGD ID	
TRANSFAC ID	
UniProtKB ID	
UniProtKB Accession	
WORMPEP ID	

(http://www.ariadnegenomics.com/services/idmap.html) is a tool for identifier translation, supporting 13 species and mapping between 15 different ID types, mainly for proteins and genes. NetAffx (http://www.affymetrix.com/products/software/specific/netaffx.affx) provides ID translation services for Affy probe set IDs. The caBIG GeneConnect project will provide ID mapping services for 10 different ID types supporting human, mouse and rat. At PIR, we provide an ID mapping service (http://pir.georgetown.edu/pirwww/search/idmapping.shtml) that maps between UniProtKB and more than 30 other data sources (Table 2) to support data interoperability among disparate data sources and to allow integration and querying of data from heterogeneous molecular biology databases. UniProt provides a mapping service to convert common gene IDs and protein IDs (such as NCBI gi number and Entrez Gene ID) to UniProtKB AC/ID and vice versa. Some mappings are inherited from cross-references within UniProtKB entries, some are based on the existing bridge between EMBL and GenBank entries, and others make use of cross-references obtained from the iProClass database. A subset of the latter (such as between UniProtKB and NCBI gi number) require matching based on sequence and taxonomy identity. Thus, it is possible to map between numerous databases using only a few sources for the mapping itself; these include UniProtKB, iProClass, RefSeq, GenBank, and nr.

For the melanosome data set mentioned earlier, the analysis first involved mapping of peptide sequences and protein lists (NCBI gi numbers from the nr database) derived from the MS data to UniProtKB entries. From a total of 2,298 gi numbers, 1,253 (55%) could be directly mapped to UniProtKB following ID mapping, while 1,506 (66%) mapped based on peptide sequences. When the results from both mappings were combined, 1,936 (84%) gi numbers were mapped to 1,438 UniProtKB sequences. The mapping revealed that the NCBI gi is an unstable database identifier, with many gi numbers changing from version to version or becoming obsolete. The result also indicates that the nr database is more redundant, with many gi numbers representing identical proteins that map to the same UniProtKB entries.

3.3. Data integration

Approaches for data integration can be divided into two major categories: 1) the data warehousing approach, and 2) the federated approach. The data warehouse approach put data sources into a centralized location with a global data schema and an indexing system for fast data retrieval. It requires reliable operation and maintenance, and stable underlying databases. On the other hand, the federated approach does not require a centralized database. It maintains a common data model and relies on a schema mapping to translate heterogeneous database schema into the target schema for integration. Therefore, it is modular, flexible and scalable.

Designed to address the data integration issue arising from voluminous, heterogeneous, and distributed data, iProClass (33, 34) uses data warehousing approach for fast data retrieval. It contains full descriptions of all known proteins with up-to-date information from many sources, thereby providing much richer annotation than can be found in any single database (35). The current version of the iProClass database provides value-added reports for about 4 million protein entries, including all entries in the UniProtKB (36) and unique NCBI (37) entries. It provides rich links and executive summaries from more than 90 databases of protein sequence, family, function, pathway, protein-protein interaction, post-translational modification, structure, genome, ontology, literature, and taxonomy. Source attribution and hypertext links facilitate exploration of additional information and examination of discrepant annotations from different sources. iProClass is implemented in the Oracle database management system. The underlying database schema and update procedures have been modified to interoperate with UniProtKB. iProClass also provides comprehensive views for more than 35,000 PIRSF protein families (38). PIRSF families are curated systematically based on literature review and integrative sequence and functional analysis, including sequence and structure similarity, domain architecture, functional association, genome context, and phyletic pattern. The results of classification and expert annotation are summarized in PIRSF family reports, with graphical viewers for taxonomic distribution, domain architecture, family hierarchy, and multiple alignment and phylogenetic tree (39).

iProClass, hosted on both the PIR and UniProt websites, now serves as the central data infrastructure at PIR/UniProt that supports both data integration and functional annotation of proteins. Coupled with the PIRSF classification, the data integration in iProClass reveals interesting relationships among protein sequence, structure and function, and facilitates functional analysis in a systems biology context. The integrative approach leads to novel prediction and functional inference for uncharacterized proteins, allows systematic detection of genome annotation errors, provides sensible propagation and standardization of protein annotation (40, 41), and assists comparative studies of protein function and evolution (39).

The database cross-references in iProClass underlie an ID mapping service (36) at PIR that maps gene and protein IDs from about 30 data sources to UniProtKB. The cross-references are also used to develop BioThesaurus (30), a web-based system for finding gene/protein synonyms for given proteins and for solving name ambiguities of proteins sharing common names. Currently covering more than 4 million UniProtKB proteins, BioThesaurus consists of over 5.7 million names collected from 23 biological databases, including gene and protein sequence databases and model organism databases.

For the 1438 melanosome proteins mapped to UniProtKB, comprehensive information from iProClass was retrieved and integrated into the data set, including protein sequence, family, function, pathway, protein-protein interaction, post-translational modification, structure, genome, ontology and literature, and taxonomy. The rich annotation for the protein s*et al*lowed to proceedwith functional profiling and functional analysis of the melanosome proteomes.

3.4. Data search and retrieval

Data integration provides a platform for researchers to efficiently query the databases. From the PIR website, the user has many options for data search and retrieval, including ID mapping (discussed previously), peptide search, batch retrieval, and text search. In this section, we focus on interactive data analysis using tools provided at PIR website. In an effort to develop an integrated expression analysis system for large scale transcriptomic and proteomic data, we initially built an prototype system, iProXpress (discussed in section 4), for analysis of large set of proteins, such as the melanosome MS data set.

3.4.1. Entry retrieval and batch retrieval

The PIR website provides a very simple way to retrieve protein entries by a single protein ID or one of many other sequence database identifiers. It also allows retrieval of protein entries using a batch of database identifiers. As discussed in the previous section, due to the diversity of databases and the lack of consistency in protein/gene names and/or identifiers in the literature, it can be difficult to retrieve multiple entries when protein and gene identifiers come from different sources. The batch retrieval tool (http://pir.georgetown.edu/pirwww/search/batch.shtml) overcomes this problem and provides high flexibility, allowing the retrieval of multiple entries from the iProClass database by selecting a specific identifier or a combination of them. The main sources of widely used identifiers are included: sequence databases (organism specific genome databases, NCBI and UniProt databases), function/feature databases (including EC-IUBMB, KEGG, RESID, and Gene Ontology), classification databases (PIRSF, PFam, COG, and PROSITE), organism databases (taxon group and taxon), and others (Entrez Gene, PDB, OMIM and PubMed). Batch retrieval of PIRSF families using a subset of these identifiers can be done as well.

3.4.2. Peptide search

Peptide sequences, such as those obtained by MS/MS, can be used as queries to search proteins containing exact matches to the peptide sequence from the UniProtKB or UniRef100 database. In the first case, the search can be performed on the whole set of proteins or on only those from taxon group or a specific organism, as in the example shown in Figure 1. Peptide Search may reveal protein sequence regions that are completely conserved in a certain group of organisms and that could be important for function.

3.4.3. Text search

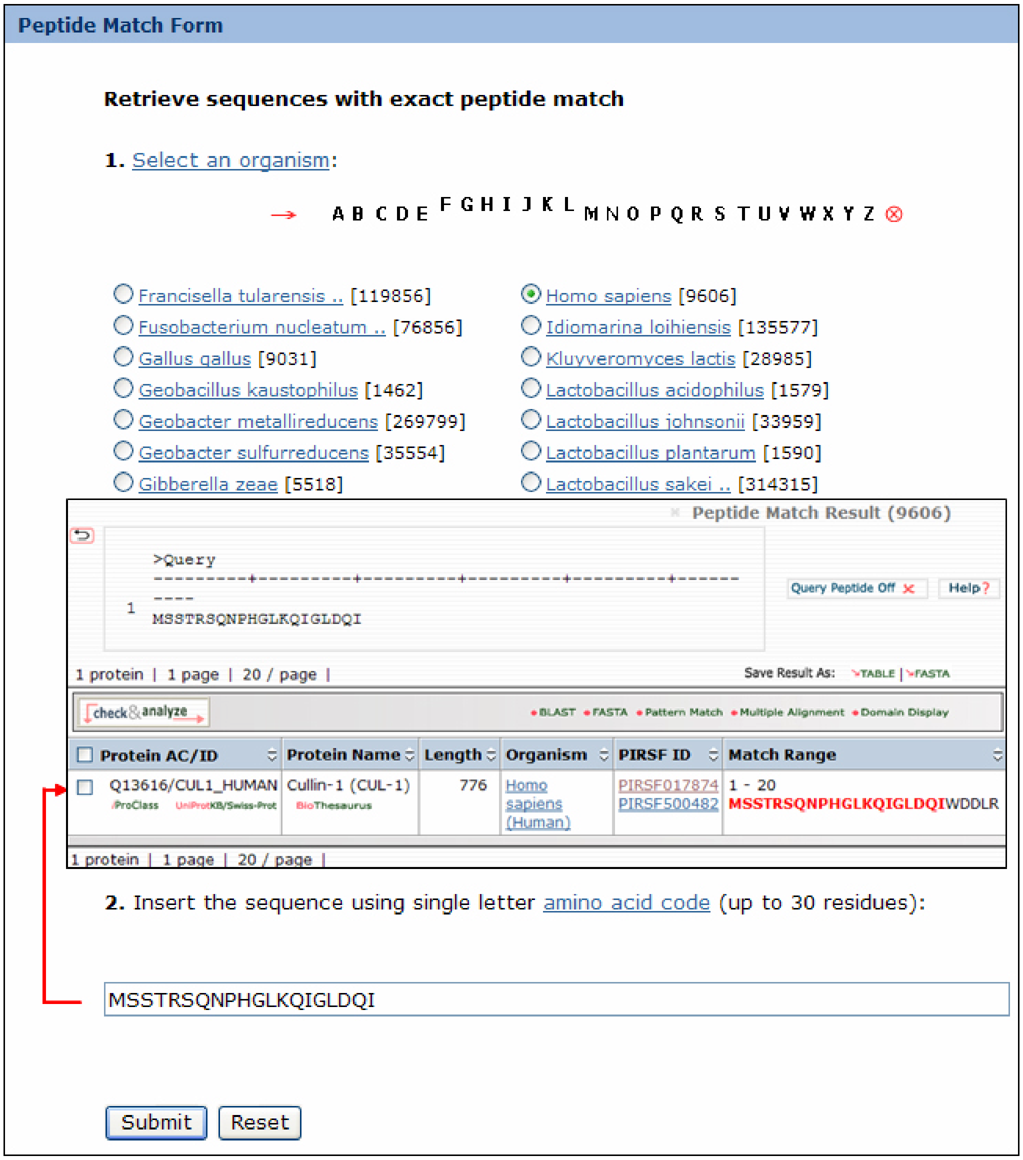
This is a widely used way to search the database, especially when a specific field needs to be searched. For proteomic/genomic data, common searches may include (but are not limited to): organism name, keywords such as "complete proteome" or a subcellular location, and pathway. Following the melanosome data set example, we could look into what proteins are annotated as melanosomal proteins in UniProtKB, using the corresponding controlled vocabulary for this subcellular location. Because there is a limit of 20,000 entries for retrieval, in some cases it may be necessary to download and parse the database. Figure 2A shows how to access the iProClass text search; Figure 2B shows the text search form. In this example the search has been restricted to Organism Name "homo sapiens" and KEGG pathway "TGF-beta" (Figure 2B). Figure 2C shows a partial section of the result page obtained after submitting the query displayed in Figure 2B.

Although all links and data can be found in the individual protein records, sometimes it is more convenient, for easy comparison, to have an overview of the information contained in selected fields. The Display Option box (Figure 2C, 1) gives the user options to customize the columns to be displayed. Any of the searchable fields can be selected, for example, it might be of interest to display the KEGG database metabolic or signaling pathway ID or name, to display OMIM ID to see entries that have an associated phenotype, or UniRef clusters to investigate homologous proteins (shown in this Figure 2C, 4).

3.4.4. Sequence similarity

The most widely used bioinformatics approach to assess function is by searching for homologous proteins. PIR provides four approaches for sequence similarity search and retrieval: 1) by real time BLAST or FASTA searches; 2) by the Related Sequence database; 3) by UniRef clusters; and 4) by the PIRSF database. Typically, people use BLAST or FASTA for sequence similarity search. This can be very time consuming due to the large size of the sequence database. In contrast, the other three databases provide several different ways for quick identification of sequence homologous sequences to the query sequence as discussed below.

Related Sequences (Figure 2C, 3): The Related Sequences column, which is present by default in the text search and batch retrieval result pages, serves at least two purposes: (1) to show proteins similar to the query, significantly faster than running BLAST in real time, and (2) to evidence tight protein clusters, since by expanding this column, the number of similar sequences at three different E-value cut-offs is shown (Figure 3). Figure 3 shows an example for two fungal proteins with unknown function. The number of related sequences for these entries



Peptide Match Form

Retrieve sequences with exact peptide match

1. Select an organism:

→ A B C D E F G H I J K L M N O P Q R S T U V W X Y Z ⊗

○ Francisella tularensis .. [119856]
○ Fusobacterium nucleatum .. [76856]
○ Gallus gallus [9031]
○ Geobacillus kaustophilus [1462]
○ Geobacter metallireducens [269799]
○ Geobacter sulfurreducens [35554]
○ Gibberella zeae [5518]

● Homo sapiens [9606]
○ Idiomarina loihiensis [135577]
○ Kluyveromyces lactis [28985]
○ Lactobacillus acidophilus [1579]
○ Lactobacillus johnsonii [33959]
○ Lactobacillus plantarum [1590]
○ Lactobacillus sakei .. [314315]

Peptide Match Result (9606)

```
   >Query
   ---------+---------+---------+---------+---------+------
   ----
1  MSSTRSQNPHGLKQIGLDQI
```

Query Peptide Off ✕ Help?

1 protein | 1 page | 20 / page |

Save Result As: TABLE | FASTA

check & analyze • BLAST • FASTA • Pattern Match • Multiple Alignment • Domain Display

□ Protein AC/ID	Protein Name	Length	Organism	PIRSF ID	Match Range
□ Q13616/CUL1_HUMAN ProClass UniProtKB/Swiss-Prot	Cullin-1 (CUL-1) BioThesaurus	776	Homo sapiens (Human)	PIRSF017874 PIRSF500482	1 - 20 MSSTRSQNPHGLKQIGLDQIWDDLR

1 protein | 1 page | 20 / page |

2. Insert the sequence using single letter amino acid code (up to 30 residues):

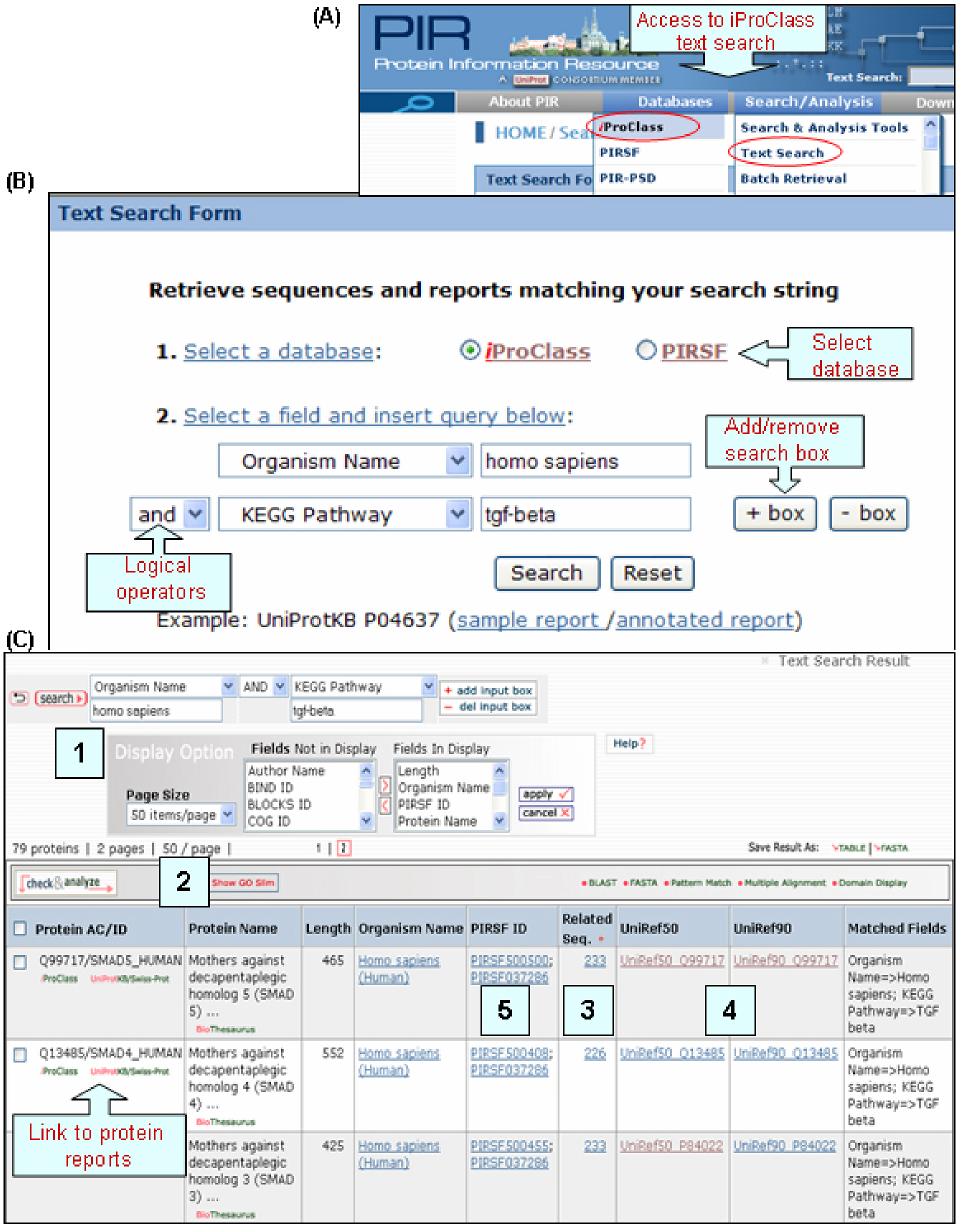
MSSTRSQNPHGLKQIGLDQI

Submit Reset

Figure 1. Organism specific Peptide Search. An example of the peptide search form, where *Homo sapiens* has been selected and the peptide sequence query entered in the blank box. The result page is shown next to the red arrow. (http://pir.georgetown.edu/pirwww/search/peptide.shtml).

is very distinct. Q7SBU9 belongs to a large group of proteins since the number of related sequences reached the limit (299 sequences) even at the lowest E-value. Inspecting the related sequence link reveals that this hypothetical protein is highly similar to a large group of proteins that has glucosidase activity. In contrast, Q756W7 has very few related sequences, suggesting that this protein belongs to a very tight group. In this case, looking into the similar sequences reveals that this protein is similar to a specific RNA polymerase I-specific transcription initiation factor found in fungi (although it should be noted that a tight cluster should not be translated into lineage specificity). In conclusion, related sequences might help to quickly investigate similar sequences and to detect tight sequence clusters.

UniRef clusters: The UniRef databases provide clustered sets of sequences from UniProtKB (including splice variants and isoforms) and selected UniParc records in order to obtain complete coverage of sequence space at several resolutions while hiding redundant sequences (but not their descriptions) from view. The UniRef100 database combines identical sequences and sub-fragments with 11 or more residues (from any organism) into a single UniRef entry, displaying the sequence of a representative protein. UniRef90 and UniRef50 are built by clustering UniRef100 sequences that have at least 90% or 50% sequence identity, respectively, to the representative sequence. UniRef90 and UniRef50 yield a database size reduction of approximately 40% and 65%, respectively, providing for significantly faster sequence searches. In addition to speeding searches



Text Search Form

Retrieve sequences and reports matching your search string

1. Select a database: ● iProClass ○ PIRSF

2. Select a field and insert query below:

	Organism Name	homo sapiens
and	KEGG Pathway	tgf-beta

+ box | - box

Search | Reset

Example: UniProtKB P04637 (sample report / annotated report)

Text Search Result

search | Organism Name: homo sapiens | AND | KEGG Pathway: tgf-beta | + add input box | − del input box

Display Option — Page Size: 50 items/page — Fields Not in Display: Author Name, BIND ID, BLOCKS ID, COG ID — Fields In Display: Length, Organism Name, PIRSF ID, Protein Name — apply ✓ | cancel ✗ — Help ?

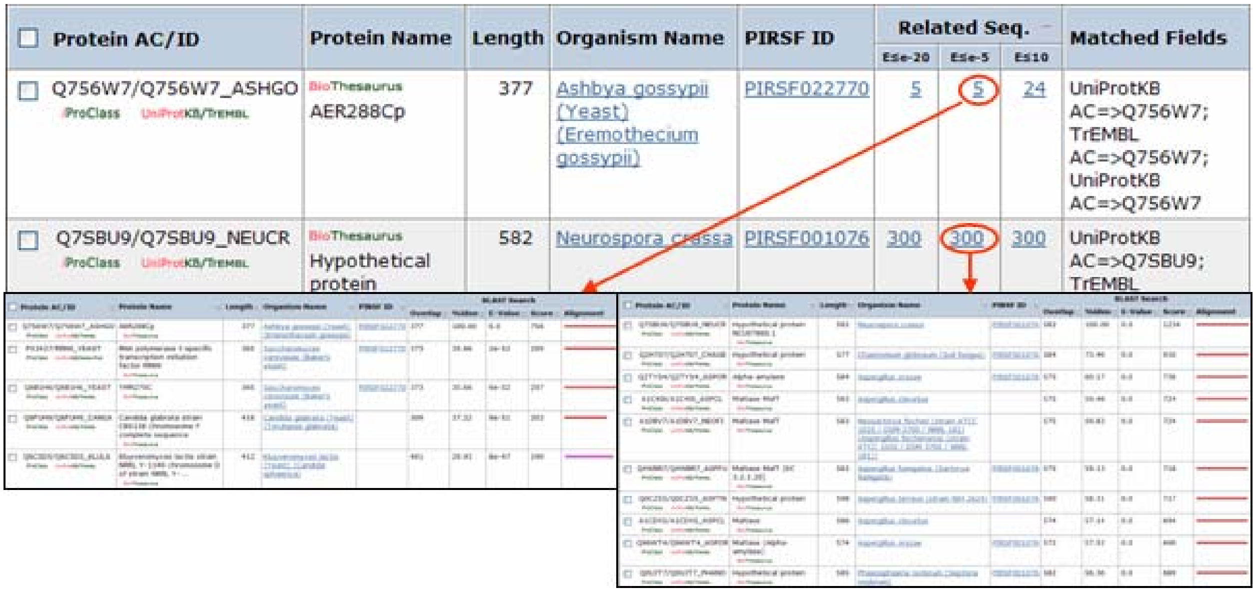
79 proteins | 2 pages | 50 / page | 1 | 2 — Save Result As: TABLE | FASTA

check & analyze | Show GO Slim | BLAST | FASTA | Pattern Match | Multiple Alignment | Domain Display

□ Protein AC/ID	Protein Name	Length	Organism Name	PIRSF ID	Related Seq.	UniRef50	UniRef90	Matched Fields
□ Q99717/SMAD5_HUMAN iProClass UniProtKB/Swiss-Prot	Mothers against decapentaplegic homolog 5 (SMAD 5) ... BioThesaurus	465	Homo sapiens (Human)	PIRSF500500; PIRSF037286	233	UniRef50_Q99717	UniRef90_Q99717	Organism Name=>Homo sapiens; KEGG Pathway=>TGF beta
□ Q13485/SMAD4_HUMAN iProClass UniProtKB/Swiss-Prot	Mothers against decapentaplegic homolog 4 (SMAD 4) ... BioThesaurus	552	Homo sapiens (Human)	PIRSF500408; PIRSF037286	226	UniRef50_Q13485	UniRef90_Q13485	Organism Name=>Homo sapiens; KEGG Pathway=>TGF beta
	Mothers against decapentaplegic homolog 3 (SMAD 3) ... BioThesaurus	425	Homo sapiens (Human)	PIRSF500455; PIRSF037286	233	UniRef50_P84022	UniRef90_P84022	Organism Name=>Homo sapiens; KEGG Pathway=>TGF beta

Figure 2. iProClass text search. (A) Ways to access iProClass from the PIR Web site (http://pir.georgetown.edu). (B) text search form including a query for human proteins that are annotated in the TGF-beta pathway in the KEGG database. (C) Example of the result page obtained after performing a text search for human proteins with TGF-beta pathway annotation in KEGG database. On top, the search terms are shown, which can be modified to either refine results, or to do a new search. Below is the Display Option (**1**) button that allows customizing the table columns. There are two saving options: as Table to save the information displayed in the table, or as FASTA to save sequences in this format. Most of the analysis tools are listed in the gray bar and all numbered items are described in the text.

and being more comprehensive, UniRef clusters can be useful for functional analysis. The UniRef cluster columns (Figure 2C, 4) can be added using the Display Option box (Figure 2C, 1). Some information about a protein with unknown function may be inferred if any of the clusters to which it belongs contains characterized members. The more similar it is to a characterized protein, the more likely the proteins are to be orthologous and, therefore, to have the same or very similar function. As an example, UniProtKB entry Q2TAS2 corresponding to a frog protein annotated as hypothetical, belongs to UniRef clusters: UniRef100_Q2TAS2 that contains only this entry, UniRef90_Q9HAU4, Smad ubiquitination regulatory factor 2 related cluster, that includes a characterized member from human known as Smurf2, and the UniRef50_Q9HCE7 Smad ubiquitination regulatory factor 1 related cluster, which contains not only Smurf2 but also its paralog Smurf1. In human, Smurf1 and Smurf2 differ in which



□ Protein AC/ID	Protein Name	Length	Organism Name	PIRSF ID	Related Seq.			Matched Fields
					Ese-20	Ese-5	Es10	
□ Q756W7/Q756W7_ASHGO ProClass UniProtKB/TrEMBL	BioThesaurus AER288Cp	377	Ashbya gossypii (Yeast) (Eremothecium gossypii)	PIRSF022770	5	5	24	UniProtKB AC=>Q756W7; TrEMBL AC=>Q756W7; UniProtKB AC=>Q756W7
□ Q7SBU9/Q7SBU9_NEUCR ProClass UniProtKB/TrEMBL	BioThesaurus Hypothetical protein	582	Neurospora crassa	PIRSF001076	300	300	300	UniProtKB AC=>Q7SBU9; TrEMBL

Figure 3. View of expanded Related Sequences column. The number of sequences obtained by pre-computed BLAST is shown at three different E-value cut-offs: e^{-20}, e^{-5}, e^{10} . The maximum number is 299 sequences. Below is the related sequences table for the cut-off value e^{-5}. In the case of Q7SBU9, only a partial view of the result is displayed.

Smad partners they bind, and that may determine whether they regulate the TGF or the BMP signaling pathway.

PIRSF database: This is the most reliable source for functional assessment. The PIRSF protein classification system is a network with multiple levels of sequence diversity, from superfamilies to subfamilies, that reflects the evolutionary relationship of full-length proteins and domains. PIRSF is based on whole proteins rather than on the component domains; therefore, it allows annotation of generic biochemical and specific biological functions, as well as classification of proteins without well-defined domains (38). Automatically generated protein clusters are manually curated for membership, domain architecture, annotation of sequence features, and specific biological functions and biochemical activities, when possible.

PIRSF is one of the default columns in the result pages for text and sequence related searches (Figure 2C, 5). If an entry is a member of a PIRSF, then the PIRSF ID(s) will be displayed in this column. Proteins with the same PIRSF ID belong to the same family, as is the case for the entries shown in Figure 2C that belong to the family PIRSF037286. However, some of the entries in this family belong to different subfamilies, usually reflecting functional specialization. Selection of a particular PIRSF ID links to the family report that contains information about membership, taxonomic distribution, PIRSF hierarchy, multiple alignment and domain architecture. Even if the protein of interest is not assigned to a PIRSF, if one of the best BLAST hits is a protein that belongs to a curated PIRSF, one may be able to infer some functional properties of the protein.

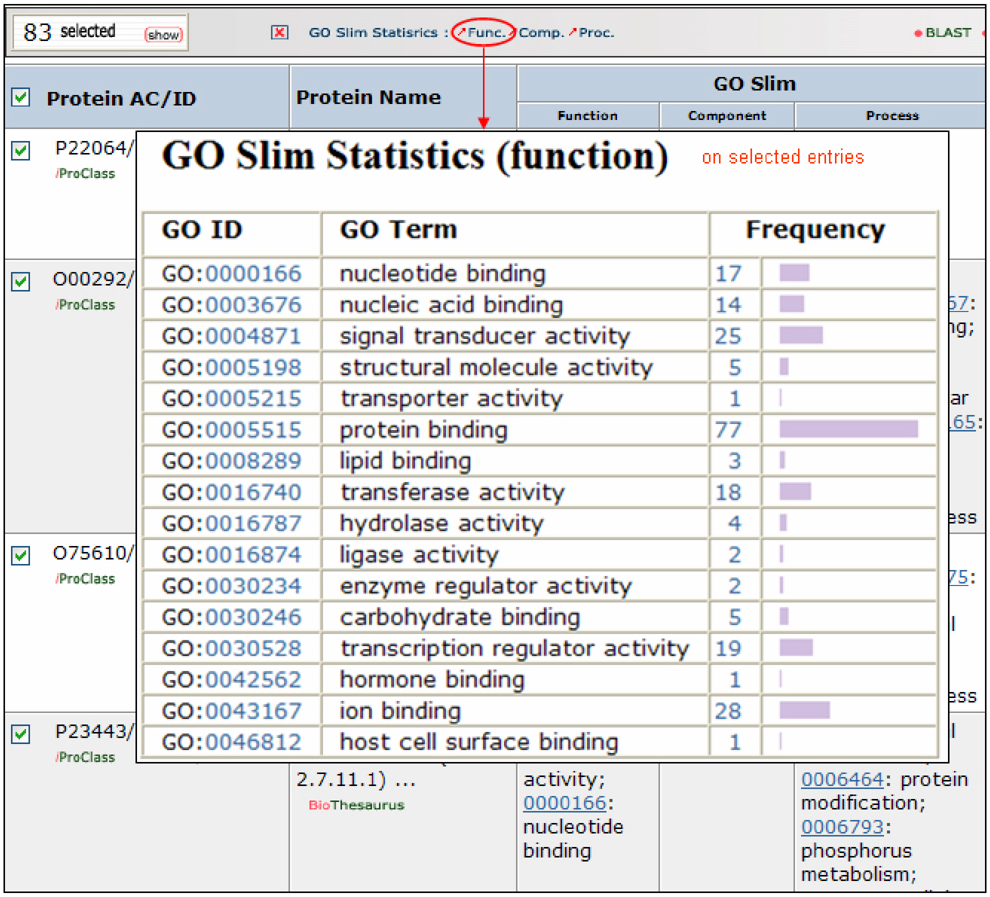
3.5. Data Analysis

At the PIR website, sequence analysis tools are integrated into the data search and retrieval result pages to assist functional analysis of the retrieved data set. These tools, which are located in the gray bar below the Display Option box, include: FASTA and BLAST for sequence similarity searches, multiple alignment with tree display, domain display, and pattern match. Of special interest for functional analysis is the GO Slim tool as described below.

GO Slim is a light version of the Gene Ontology, containing a subset of the terms residing at the high level (node) of GO term hierarchy. The GO terms in GO Slim give a broad overview of the ontology content without the detail of the specific fine grained terms. GO slims are particularly useful for summarizing the results based on the GO annotation of a genome or proteome when broad classification of gene product function is required. From the results page, one can view the GO slim terms by selecting the “Show GO Slim” button (Figure 2C, 2, and Figure 4) in the analysis tool bar. One can then view statistics for the individual ontologies (Molecular *Function*, Cellular *Component*, and Biological *Process*) by checking entries of interest and selecting the ontology to show (Figure 4). This will lead to the functional profiling of a set of proteins.

4. INTEGRATED EXPRESSION ANALYSIS SYSTEM

As discussed in previous sections, analyzing large expression data such as the melanosome MS proteomic data requires at least three steps: 1) protein mapping, 2) functional annotation and 3) functional profiling. To provide an integrated research and discovery platform for large-scale gene expression and proteomic data analysis, we have developed methods and prototype software tools specifically designed to address the limitations and gaps in the current methods and systems for protein mapping and functional annotation of proteomic



GO Slim Statistics (function) on selected entries

GO ID	GO Term	Frequency
GO:0000166	nucleotide binding	17
GO:0003676	nucleic acid binding	14
GO:0004871	signal transducer activity	25
GO:0005198	structural molecule activity	5
GO:0005215	transporter activity	1
GO:0005515	protein binding	77
GO:0008289	lipid binding	3
GO:0016740	transferase activity	18
GO:0016787	hydrolase activity	4
GO:0016874	ligase activity	2
GO:0030234	enzyme regulator activity	2
GO:0030246	carbohydrate binding	5
GO:0030528	transcription regulator activity	19
GO:0042562	hormone binding	1
GO:0043167	ion binding	28
GO:0046812	host cell surface binding	1

Figure 4. The GO Slim functionality. Three columns corresponding to the GO components molecular function, cellular component and biological process are displayed when the GO Slim button is selected. The available GO annotation for a subset of terms is shown with links to AMIGO. In addition, the statistics for a particular ontology can easily be viewed by selecting entries and then GO statistics.

data. A prototype expression analysis system, integrated Protein eXpression (iProXpress), was recently developed at PIR and has been applied to several studies (42) (43) (44). A public iProXpress website (http://pir.georgetown.edu/iproxpress) is accessible for browsing the published proteomics data sets.

4.1. System design

An overview of the iProXpress system design is shown in Figure 5. The system contains several components, including a data warehouse composed of the UniProtKB and iProClass databases, and analysis tools for protein mapping, functional annotation and expression profiling. Sequence homology analysis tools are included in the protein mapping tools. System integration by iProXpress also supports iterative functional analysis for algorithm enhancement.

4.2. Major functionalities

The major functionalities provided by the iProXpress system include the mapping of gene/protein sequences with different types of IDs from gene expression and proteomic data to UniProtKB protein entries, and the functional annotation and profiling of the mapped proteins for functional analysis. This integration of bioinformatics tools and databases from a large number of resources supports functional annotation and function/pathway profiling of proteomic data in a systems biology context.

4.2.1. Protein Mapping

As rich annotation, minimal redundancy, and a high degree of data integration are critical for gene expression and proteomic data interpretation, the protein mapping tool is designed to map these data to corresponding UniProtKB entries to facilitate functional analysis. The accepted input data will be protein IDs and their associated peptide sequences when available, which may be generated from search programs such as SEQUEST, MASCOT or X!Tandem. To integrate with other high-throughput data types, the iProXpress system will also accept gene expression data such as lists of gene probes from cDNA microarray experiments.

Protein lists and peptide sequences are mapped to UniProtKB entries based on ID and peptide mapping, respectively. The PIR ID mapping service maps protein/gene IDs from about 30 data sources (see data mapping section) to UniProtKB. To cross-validate the ID

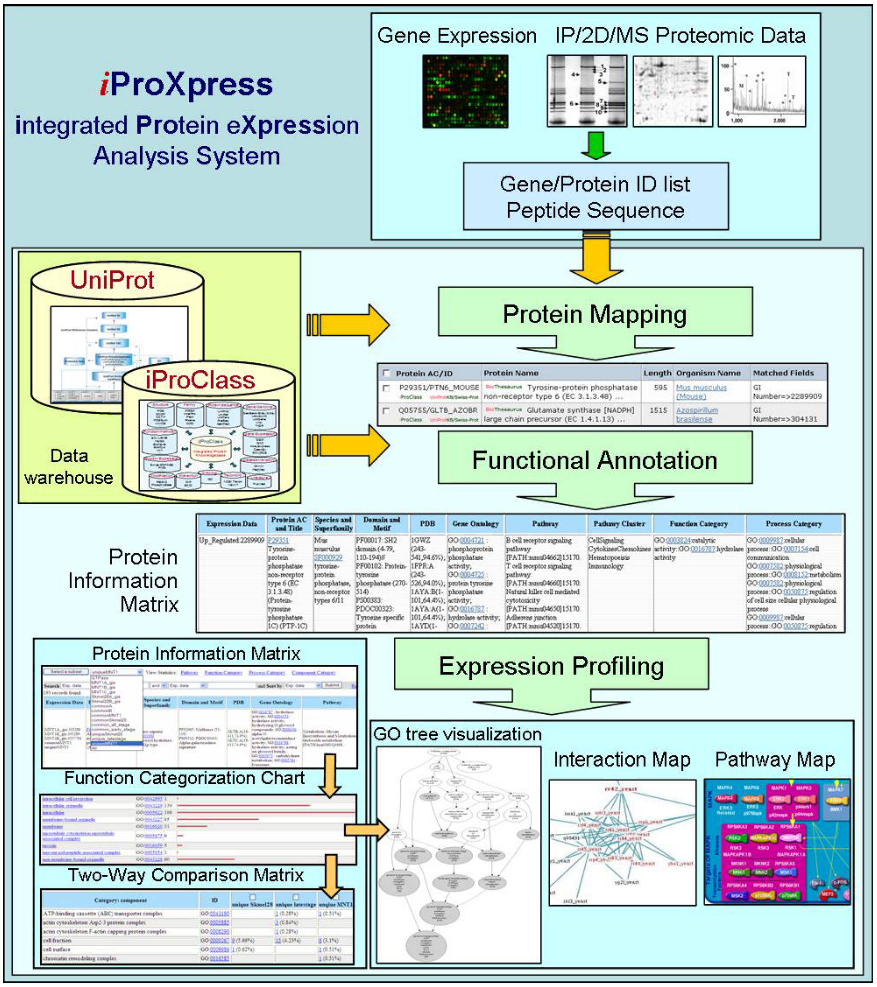


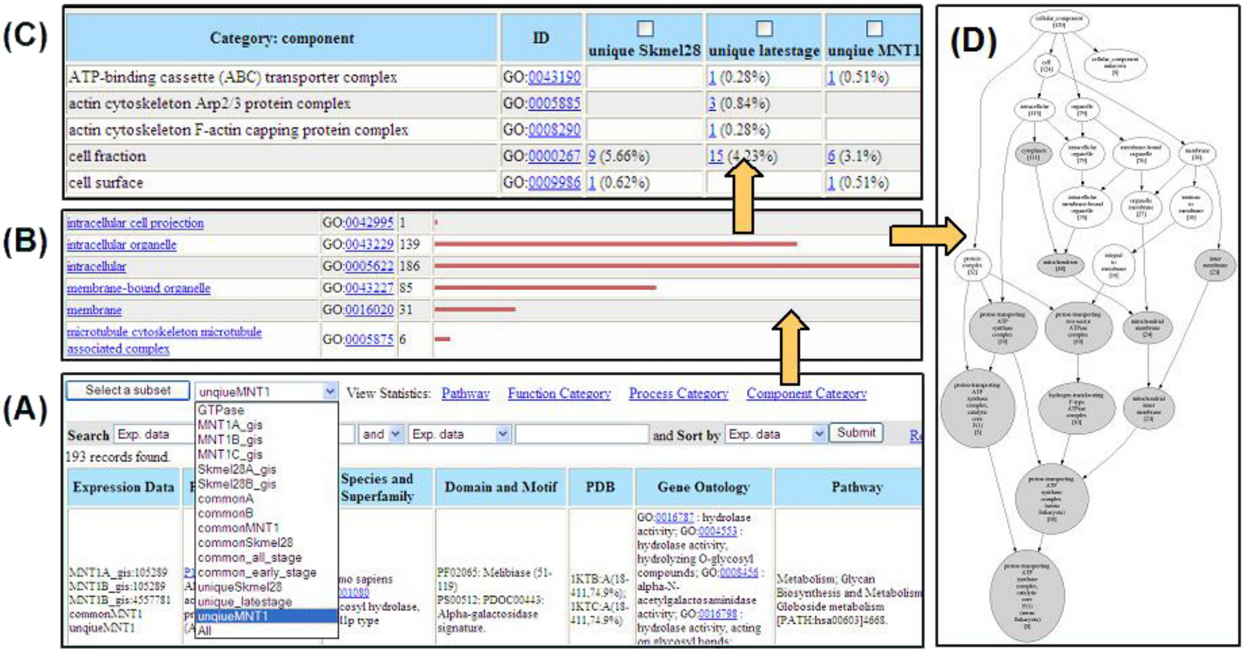
Figure 5. Overview of the iProXpress system design.

mapping results, the peptide sequence of each mapped protein is matched against the cross-referenced UniProtKB sequence to confirm the assignment. For many-to-one mappings, as is often the case for gi numbers, the mapping effectively removes redundancy. For proteins not mapped through ID mapping, their peptide sequences are matched against the UniRef100 library. In one-to-one mapping, where the peptide matches exactly one UniProt protein, that protein is given the assignment. In one-to-many mapping, where the peptide sequence matches to more than one UniProt entry, sequence variations are identified by UniRef90 clusters in which members share at least 90% sequence identity to the representative sequence. If the proteins belong to different UniRef90 clusters, manual validation with retro-inspection of the original MS/MS protein identification results is required to make reliable assignment. The proteins not mapped after ID and peptide mapping to UniRef100 are mapped to the unique UniParc sequence library. Finally, the remaining proteins not mapped by the above steps are mapped by sequence similarity.

Gene lists are mapped to UniProtKB based on gene/protein IDs, sequences, or gene/protein names. Genes with common identifiers such as GenBank, UniGene or Entrez Gene are mapped based on the PIR ID mapping service, which continues to add cross-references. For genes with no ID match, the mapping uses sequence comparison, or uses name matching if the sequence is not available. Computer-assisted name mapping is provided using the PIR BioThesaurus, currently with nearly 5 million protein/gene names. When corresponding IDs are assigned to both genes and proteins, the system also links gene expression and proteomic data for comparative analysis.

4.2.2. Functional Annotation

After the protein mapping, rich annotation can be fully described in a protein information matrix based on sequence analysis and integration of information from the iProClass database. iProClass also includes pre-computed sequence analysis results (e.g. BLAST related sequences) to support reliable annotation transfer from well-curated homologs to poorly characterized proteins, which is useful because an estimated 40-50% of proteins from complete genomes are “hypothetical”, and a small fraction of proteins have experimentally validated annotations. We pre-compute and regularly update sequence features of functional significance for UniProt proteins, and make the sequence analysis tools available for online analysis of


(C)
(B)
(A)
(D)
Select a subset
unqiueMNT1
GTPase
MNT1A_gis
MNT1B_gis
MNT1C_gis
Skmel28A_gis
Skmel28B_gis
commonA
commonB
commonMNT1
commonSkmel28
common_all_stage
common_early_stage
uniqueSkmel28
unique_latestage
unqiueMNT1
All
View Statistics: Pathway Function Category Process Category Component Category
Search Exp. data and Exp. data and Sort by Exp. data Submit
cellular_component
cell
cellular_component unknown
intracellular
organelle
cytoplasm
intracellular organelle
membrane-bound organelle
membrane
intracellular membrane-bound organelle
organelle membrane
intrinsic to membrane
protein complex
mitochondrion
integral to membrane
inner membrane
proton-transporting ATP synthase complex
proton-transporting two-sector ATPase complex
mitochondrial membrane
proton-transporting ATP synthase complex, catalytic core F(1)
hydrogen-translocating F-type ATPase complex
mitochondrial inner membrane
proton-transporting ATP synthase complex (sensu Eukaryota)
proton-transporting ATP synthase complex, catalytic core F(1) (sensu Eukaryota)

Category: component	ID	□ unique Skmel28	□ unique latestage	□ unqiue MNT1
ATP-binding cassette (ABC) transporter complex	GO:0043190		1 (0.28%)	1 (0.51%)
actin cytoskeleton Arp2/3 protein complex	GO:0005885		3 (0.84%)	
actin cytoskeleton F-actin capping protein complex	GO:0008290		1 (0.28%)	
cell fraction	GO:0000267	9 (5.66%)	15 (4.23%)	6 (3.1%)
cell surface	GO:0009986	1 (0.62%)		1 (0.51%)

intracellular cell projection	GO:0042995	1	
intracellular organelle	GO:0043229	139	
intracellular	GO:0005622	186	
membrane-bound organelle	GO:0043227	85	
membrane	GO:0016020	31	
microtubule cytoskeleton microtubule associated complex	GO:0005875	6	

193 records found.

Expression Data		Species and Superfamily	Domain and Motif	PDB	Gene Ontology	Pathway

MNT1A_gis:105289
MNT1B_gis:105289
MNT1B_gis:4557781
commonMNT1
unqiueMNT1
		mo sapiens
		001080
		cosyl hydrolase,
		llp type
			PF02065: Melibiase (51-119)
			PS00512: PDOC00443: Alpha-galactosidase signature.
				1KTB:A(18-411,74.9%);
				1KTC:A(18-411,74.9%)
					GO:0016787 : hydrolase activity; GO:0004553 : hydrolase activity, hydrolyzing O-glycosyl compounds; GO:0008456 : alpha-N-acetylgalactosaminidase activity; GO:0016798 : hydrolase activity, acting on glycosyl bonds;	Metabolism; Glycan Biosynthesis and Metabolism Globoside metabolism [PATH:hsa00603]4668.

Figure 6. Functional profiling analysis: (A) protein information matrix, (B) functional categorization chart, (C) cross-comparison matrix, (D) graphical GO hierarchy. Abbreviations: GO: Gene Ontology, iProXpress: integrated Protein eXpression, LRO: lysosome-related organelle, MS/MS: tandem mass spectrometry, PIR: Protein Information Resource, UID: database unique identifier, UniProtKB: UniProt Knowledgebase.

proteins/sequence variations not in UniProt. Sequence features pre-computed include homologous proteins in KEGG, BioCarta and other curated pathway databases to populate pathway annotation, InterProScan (45) for family, domain and motif identification, and Phobius for transmembrane helix and signal peptide prediction (46). Properties derived from homology-based inference will be presented in the information matrix with evidence attribution.

4.2.3. Functional profiling

Functional profiling analysis aims at discovering the functional significance of expressed proteins, the plausible functions and pathways, and the hidden relationships and interconnecting components of proteins, such as proteins sharing common functions, pathways, or cellular networks. As shown in Figure 6 (the example mockups), the extensive annotation in the protein information matrix (A) allows functional categorization and detailed analysis of expressed proteins in a given dataset, as well as cross-comparison of co-expressed or differentially-expressed proteins from multiple datasets. For functional categorization, proteins are grouped based on annotations such as GO terms, and KEGG and BioCarta pathways, and then correlated with sequence similarity to identify relationships among individual proteins or protein groups. The functional categorization chart (B) displays the frequency (number of occurrences) of proteins in each functional category. Categorization and sorting of proteins based on functions, pathways, and/or other attributes in the information matrix generate various protein clusters, from which interesting unique or common proteins in different datasets can be identified in combination with manual examination. The cross-comparison matrix (C) shows the comparative distribution of functional categories in multiple datasets.

To correlate functional association of expressed proteins in different samples, the relative enrichment of a given functional category in each sample will be calculated to identify all samples that contain a statistically significant proportion of proteins that are associated with the given category. Likewise, the system will point to groups of proteins that show a statistically significant correlation with certain pathways or functions, thus enabling characterization of biological pathways. Evidence on differential protein expression, protein interactions, pathway membership and other attributes is combined to provide the evidence for pathway and network participation. This allows relative ranking of the proteins involved in the biological response to identify the critical nodes in the response pathway and hidden relationships.

4.3. Analysis of proteomic data

The prototype version of iProXpress system has been applied to the expression profile analysis for hCG-induced changes in MA-10 mouse Leydig tumor cells (42), organelle proteome analysis of various melanosome stages from human melanoma cell lines (43), and the comparative analyses of lysosome-related organelle (LRO) proteomes (44). Here we use the organelle proteome analyses to illustrate the integrative approach for function and pathway exploration and knowledge discovery using iProXpress.

4.3.1. Melanosome proteomes

Melanosomes are membrane-bound organelles

Table 3. Stage-related melanosome proteins (partial list)

	UniProtKB	Gene Name	Protein Name	Notes[1]
Stage I	P36955	PEDF	Pigment epithelium-derived factor precursor	*Validated*
	P51148	RAB5C	Ras-related protein Rab-5C (RAB5L) (L1880)	
	P05556	ITGB1	Integrin beta-1	
	Q9UMX9	Matp	Membrane-associated transporter protein (SLC45A2)	*Homolog*
	O14880	MGST3	Microsomal glutathione S-transferase 3	*Proposed*
	Q14254	FLOT2	Flotillin-2 (Epidermal surface antigen)	
Stage II	Q9UMX9	Matp	Membrane-associated transporter protein (SLC45A2)	*Homolog*
	Q04656	ATP7A	Copper-transporting ATPase 1	
	Q9P0L0	VAPA	Vesicle-associated membrane protein-associated protein A	*Proposed*
	P53992	SEC24C	Protein transport protein Sec24C	
	O95782	AP2A	Adapter-related protein complex 2 alpha- 1 subunit	
Stage IV	O95670	ATP6G2	Vacuolar ATP synthase subunit G 2	*Validated*
	Q71RS6	SLC24A	Sodium/potassium/calcium exchanger 5 precursor	
	P57729	Rab-38	Ras-related protein Rab-38	*Homolog*
	P51159	RAB27A	Ras-related protein Rab-27A (Rab-27)	
	Q9Y4I1	Myo5a	Myosin-5A (Myosin Va)	
	Q99698	LYST	Lysosomal trafficking regulator	
	Q16643	DBN1	Drebrin	*Proposed*
	P59998	ARPC4	Actin-related protein 2/3 complex subunit 4	
	P18206	VCL	Vinculin (Metavinculin)	
	P63000	RAC1	Ras-related C3 botulinum toxin substrate 1 (p21-Rac1)	
	P51148	RAB5C	Ras-related protein Rab-5C (RAB5L)	

[1] All proteins listed are identified in the melanosome proteomes. *Validated* –shown to be localized in melanosomes by immunostaining; *Homolog* – homologous to known mouse coat color genes; *Proposed* – proposed as protein of functional interest for validation.

specialized in the production and distribution of melanin pigment and are conserved in structure from primitive organisms to mammals. Dysfunctions in pigmentation and melanosome biogenesis are associated with a wide variety of inherited genetic disorders and pigmentary diseases, including oculocutaneous albinism and Hermansky–Pudlak syndrome. Melanosome-specific proteins also provide important markers for malignant melanoma. In mammals, melanosomes mature from undifferentiated vesicles (stage I) to an elongated form with internal fibrils (stage II). In the presence of tyrosinase and other enzymes, melanin is synthesized and deposited on the internal fibrils (stage III) and can become uniformly dense (stage IV) in heavily-pigmented melanocytes. As melanosomes mature, they are gradually transported to the peripheries of the melanocytes in which they form, and in human skin, they are transferred to neighboring keratinocytes. A detailed understanding of how melanosomes mature and move within and between cells requires a comprehensive knowledge of the proteins comprising them. A combination of immunoblotting, immunofluorescence microscopy, and bioinformatics analysis was used to characterize the protein profiles of melanosomes at various biogenic stages.

The determined melanosome proteomes contain ~1,500 proteins combined from all stages of melanosomes, with ~600 in any given stage. Protein information matrices were generated for corresponding UniProtKB entries of identified melanosome proteins, summarizing salient features retrieved from the underlying PIR data warehouse or inferred based on sequence homology. Iterative categorization and sorting of proteins were carried out to generate various protein clusters, from which interesting unique or common proteins at different stages of melanosome biogenesis were identified in combination with manual examination. The stage-related proteins provide direct evidence of protein sorting and trafficking to this organelle and provide information about their biogenesis as lysosome-related organelles. Approximately 100 proteins shared by melanosomes from pigmented and non-pigmented melanocytes at all stages define the essential melanosome proteome. These common proteins are considered constituent or resident proteins throughout melanosome biogenesis. Melanosome stage-specific proteins were proposed using the functional information matrices, some of which have been subsequently validated for their melanosomal localization, including PEDF (pigment-epithelium derived factor) and SLC24A5 (sodium/potassium/calcium exchanger 5, NCKX5).

Based on the functional profiling, a more detailed melanosomal biogenic pathway has been proposed that will facilitate understanding of the dynamic process of melanosome biogenesis, including the contribution of elements and complex membrane protein traffic input from several other organelles (43). Besides proteins previously known as melanosome-specific proteins (e.g., Pmel17, TYR, Tyrp1), this study provided a comprehensive list of proteins comprising this dynamic organelle. Table 3 selectively lists three groups of proteins that are functionally important at each stage of melanosomes: 1) newly identified and validated in this study (e.g., PEDF and SLC24A5); 2) human homologs of mouse color genes identified in this study (e.g., Atp7a and MyoVa); 3) proposed stage-related proteins newly identified (e.g., Sec24 and Vinculin); 4) proteins known as melanosome proteins from previous studies (e.g., Pmel17 and TYR). Many proteins detected in stage IV melanosomes are molecular motor- and cytoskeleton-related proteins (not listed), which may be necessary for directing fully pigmented melanosomes towards the cell periphery and their eventual transfer to keratinocytes. Some proteins are found in all stages and are also common to other organelles, e.g., LAMP1 in lysosome. While it is obvious that multiple sources of cellular components contribute to the biogenesis of melanosomes, proteins more abundant in specific stages may define unique functions in that stage (e.g., the ion transporters VATPase and SLC24A5).

Therefore, it is possible to deduce a set of signature proteins for melanosomes that will consist of previously known melanosome-specific proteins, the proposed melanosome stage-specific proteins, and other constituent proteins commonly found in several other organelles. This study illustrates that bioinformatics characterization of melanosome proteomes facilitates a better understanding of the biogenesis and function of melanosomes.

4.3.2. Comparative organelle proteome analysis

Due to a better understanding of complex pathways and interactions at the molecular level, organelles are no longer considered fixed entities, but rather are dynamic structures interacting with each other and remodeling themselves in response to various stimuli. Accordingly, it is unlikely that a discrete proteome can be assigned to any of the subcellular compartments. The same organelle in different tissues or cell types may have different profiles (47, 48). Many proteins may be associated with more than one organelle or subcellular component, and temporal and spatial regulation of organellar proteins is common (49). Due to the dynamic nature of organellar proteins, complete and accurate cataloging of protein subcellular localizations is challenging. Despite the technical challenges and the biological reality, large-scale MS proteomic profiling, coupled with separation techniques, represents the best current technology and has led to the characterization of a number of organelle proteomes, including those of mitochondria, the plasma membrane, the cytosol, the nucleus, and even subnuclear structures, such as the nucleolus.

A systematic bioinformatics analysis of proteome profiles of lysosome-related organelles (LROs), a family of organelles that includes lysosomes, platelet dense bodies, and melanosomes, was recently conducted using the iProXpress system to provide functional insights for LRO biogenesis and functions (44). Proteins found in only one type of LRO and those found in a group of LROs are profiled, and large families of proteins, such as Rab family proteins, are examined for their distribution among all the LROs. The comparative organelle proteome analysis provides some interesting concepts regarding the biogenesis, interactions, and functions of LROs. Proteins detected in only one type of LRO are likely to contribute to the specific function of that organelle, while those shared by one or more LROs suggest common functions among them. The promiscuous localization of the majority of proteins in LROs also reflects their common origins as well as their transient and dynamic natures.

In previous studies, lysosomes, melanosomes, and platelet dense bodies had been identified as LROs based on common defects seen in various diseases, such as Hermansky-Pudlak syndrome and Chediak-Higashi syndrome, where their various functions were significantly affected (50). More recently, proteomic analyses have revealed other members of the LRO family (e.g., neuromelanin granules, exosomes, and synaptosomes) which in retrospect is quite reasonable, based on their phenotypes and functions. This comparative LRO proteome study underscores the common biogenesis of those organelles and their interactions.

The compiled catalogs of LRO proteomes also serve as “reference data” for the scientific community to query and browse for answering specific questions, such as “which proteins are most often seen in both melanosomes and neuromelanin granules?” As organelle proteomic research continues, this reference data set can be updated, thus providing a valuable resource for the LRO research.

5. FUTURE DIRECTIONS

While expression profiling based on the comprehensive protein information matrix provided by iProXpress allows functional views of the microarray and proteomic data, methods such as various classification schemes that provide quantitative assessment and global behavior of the expression data can be used to enhance the ability of iProXpress in studying differences among physiological or disease states or developmental stages. Current classification methods can be roughly divided into two categories, clustering and machine leaning. However, for clustering methods determining the number of clusters is challenging, and widely-accepted measures for performance evaluation are lacking. Although machine learning methods process gene/protein features jointly, group behaviors and interactions of genes/proteins are not taken into consideration. A novel ensemble dependence model has recently been developed and applied to the expression data analysis, which effectively captures the global behavior of genes and proteins from gene expression and proteomic data (51). The model has promising performance for the microarray and proteomic data classification and can be used to build a dependence network (51).

A microarray gene expression data set for gastric cancer was analyzed using a network modeling method (52). Seven potential biomarkers for the cancer were identified, six with significantly increased expression levels and one with decreased expression. The seven genes have been shown to be biologically relevant in gastric and other cancers. The six up-regulated genes include extracellular matrix components and those that mediate cell-matrix interactions, which tend to be more highly expressed in tumors of the diffuse histological type. This is consistent with greater propensity of the group of tumors for invasive growth, often provoking a dense fibrous reaction, and a reflection of reciprocal interactions between tumor and stromal cells that play important roles in tumor biology. Proteins encoded by three of the six biomarker genes (SPARC, COL3A1, and THY1) are known extracellular matrix components. Of special interest is that both SPARC and COL3A1 are concurrently observed in several studies as valuable biomarkers for gastric cancers (as connected core nodes in the network). The network modeling approach has provided a novel and consistent mathematic model to define potential cancer biomarkers, which imply functional associations or interactions that are important for the underlying cancer biology. Therefore, the network modeling method will be incorporated into the iProXpress system for functional and pathway discovery from gene expression and proteomic data in a broad range of biological systems.

The prototype iProXpress system will be further developed into a pipelined expression analysis tool, which will be made available to the research community through the iProXpress website. The web interface will be interactive and will allow data input (e.g., batch IDs) and output (e.g., protein information matrix), and support browsing, sorting, and categorization of proteins based on individual or combined attributes in the protein matrix to identify hidden relationships among different functional or pathway categories or correlation of expression profiles to certain salient properties.

6. CONCLUSION

To effectively utilize the vast amounts of data generated from proteomics experiments, it is essential to provide researchers with an integrated view of all data relevant to functional analysis. From such an integrated view, researchers can infer important biological relationships for scientific discovery. The challenge lies in the diversity and large number of existing databases that have to be handled. Especially, the lack of standardization of protein/gene ID and names makes it very difficult to map to the same entity among different databases.

For this reason, we believe that a key aspect of data integration is achieving a reliable mapping system to link these sources. As we discussed in the ID mapping section, there are two types of ID mappings: mapping among the biological objects, and mapping from biological objects to their properties, which usually produces many-to-one mapping. The PIR ID mapping service allows integration and querying of data from heterogeneous molecular biology databases. We have successfully developed iProClass, which combines both data warehouse and hypertext navigation methods for integrating data, providing a comprehensive picture of protein properties that may lead to novel prediction and functional inference for previously uncharacterized "hypothetical" proteins and protein groups. In addition, we provide a user-friendly interface to search, retrieve, and analyze large amounts of data, and to assist in its functional analysis. For example, we adopted GO Slim, a subset of the GO terms containing the more general nodes, which allows inspection of the general GO annotation on a set of selected proteins. We also provide easy ways to investigate homologous proteins: from pre-computed BLAST results and UniRef clusters that can give some preliminary idea of homologous proteins to the PIRSF database, whose manually curated families contain reliable information for functional inference.

In our continuing effort towards facilitating this task further, we have developed an integrated research and discovery platform for large-scale gene expression and proteomic data analysis, called iProXpress. The prototype system has been very useful for organelle-related proteomic studies, and the complete system will be made widely available to the research community. iProXpress consists of three major components: (i) the PIR data warehouse with integrated protein information, (ii) analytical tools for sequence analysis and functional annotation, and (iii) graphical user interface for categorization and visualization of expression data. The major functionalities include gene/peptide to protein mapping, protein information matrix, and protein data analysis. Through iterative categorization and sorting of proteins in the information matrix, users can correlate expression/interaction patterns to protein properties for pathway and network discovery.

7. ACKNOWLEDGEMENT

The PIR bioinformatics resource is supported in part by grants U01-HG02712 and HHSN266200400061C from the National Institutes of Health. We would like to thank Dr. Winona Barker for reviewing the manuscript and providing useful comments.

8. REFERENCES

1. H. Bono, I. Nikaido, T. Kasukawa, Y. Hayashizaki, Y. Okazaki: Comprehensive analysis of the mouse metabolome based on the transcriptome. *Genome Res* 13, 1345-1349 (2003)

2. A. J. Walhout, J. Reboul, O. Shtanko, N. Bertin, P. Vaglio, H. Ge, H. Lee, L. Doucette-Stamm, K. C. Gunsalus, A. J. Schetter, D. G. Morton, K. J. Kemphues, V. Reinke, S. K. Kim, F. Piano, M. Vidal: Integrating interactome, phenome, and transcriptome mapping data for the C. elegans germline. *Curr Biol* 12, 1952-1958 (2002)

3. A. Pandey, M. Mann: Proteomics to study genes and genomes. *Nature* 405, 837-846 (2000)

4. S. Hanash: Disease proteomics. *Nature* 422, 226-232 (2003)

5. N. L. Anderson, A. D. Matheson, S. Steiner: Proteomics: applications in basic and applied biology. *Curr Opin Biotechnol* 11, 408-412 (2000)

6. C. C. Wu, M. MacCoss: Shotgun proteomics: Tools for the analysis of complex biological systems. *Curr Opin Mol Ther* 4, 242-250 (2002)

7. J. A. Ranish, E. C. Yi, D. M. Leslie, S. O. Purvine, D. R. Goodlett, J. Eng, R. Aebersold: The study of macromolecular complexes by quantitative proteomics. *Nat Genet* 33, 349-355 (2003)

8. S. Hanash: Integrated global profiling of cancer. *Nat Rev Cancer* 4, 638-644 (2004)

9. R. Aebersold, D. R. Goodlett: Mass spectrometry in proteomics. *Chem Rev* 101, 269-295 (2001)

10. J. M. A. Eng, J. R. Yates: An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* 5, 976-989 (1994)

11. D. N. Perkins, D. J. Pappin, D. M. Creasy, J. S. Cottrell: Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20, 3551-3567 (1999)

12. D. Fenyo, R. C. Beavis: A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal Chem* 75, 768-774 (2003)

13. A. Keller A. I., Nesvizhskii, E. Kolker, R. Aebersold: Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* 74, 5383-5392 (2002)

14. A. I. Nesvizhskii, A. Keller, E. Kolker, R. Aebersold: A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* 75, 4646-4658 (2003)

15. X. Yang, V. Dondeti, R. Dezube, D. M. Maynard, L. Y. Geer, J. Epstein, X. Chen, S. P. Markey, J. A. Kowalak: DBParser: web-based software for shotgun proteomic data analyses. *J Proteome Res* 3, 1002-1008 (2004)

16. E. A. Kapp, F. Schutz, L. M. Connolly, J. A. Chakel, J. E. Meza, C. A. Miller, D. Fenyo, J. K. Eng, J. N. Adkins, G. S. Omenn, R. J. Simpson: An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis. *Proteomics* 5, 3475-3490 (2005)

17. A. Rauch, M. Bellew, J. Eng, M. Fitzgibbon, T. Holzman, P. Hussey, M. Igra, B. Maclean, C. W. Lin, A. Detter, R. Fang, V. Faca, P. Gafken, H. Zhang, J. Whitaker, D. States, S. Hanash, A. Paulovich, M. W. McIntosh: Computational Proteomics Analysis System (CPAS): an extensible, open-source analytic system for evaluating and publishing proteomic data and high throughput biological experiments. *J Proteome Res* 5, 112-121 (2006)

18. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, G. Sherlock: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25, 25-29 (2000)

19. B. R. Zeeberg, W. Feng, G. Wang, M. D. Wang, A. T. Fojo, M. Sunshine, S. Narasimhan, D. W. Kane, W. C. Reinhold, S. Lababidi, K. J. Bussey, J. Riss, J. C. Barrett, J. N. Weinstein: GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol* 4, R28 (2003)

20. S. W. Doniger, N. Salomonis, K. D. Dahlquist, K. Vranizan, S. C. Lawlor, B. R. Conklin: MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol* 4, R7 (2003)

21. J. Cheng, S. Sun, A. Tracy, E. Hubbell, J. Morris, V. Valmeekam, A. Kimbrough, M. S. Cline, G. Liu, R. Shigeta, D. Kulp, M. A. Siani-Rose: NetAffx Gene Ontology Mining Tool: a visual approach for microarray data analysis. *Bioinformatics* 20, 1462-1463 (2004)

22. G. Dennis, Jr., B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane, R. A. Lempicki: DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* 4, P3 (2003)

23. A. Bateman, L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. Sonnhammer, D. J. Studholme, C. Yeats, S. R. Eddy: The Pfam protein families database. *Nucleic Acids Res* 32, D138-141 (2004)

24. M. Kanehisa, S. Goto, S. Kawashima, A. Nakaya: The KEGG databases at GenomeNet. *Nucleic Acids Res* 30, 42-46 (2002)

25. F. Al-Shahrour, P. Minguez, J. M. Vaquerizas, L. Conde, J. Dopazo: BABELOMICS: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments. *Nucleic Acids Res* 33, W460-464 (2005)

26. N. J. Mulder, R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bradley, P. Bork, P. Bucher, L. Cerutti, R. Copley, E. Courcelle, U. Das, R. Durbin, W. Fleischmann, J. Gough, D. Haft, N. Harte, N. Hulo, D. Kahn, A. Kanapin, M. Krestyaninova, D. Lonsdale, R. Lopez, I. Letunic, M. Madera, J. Maslen, J. McDowall, A. Mitchell, A. N. Nikolskaya, S. Orchard, M. Pagni, C. P. Ponting, E. Quevillon, J. Selengut, C. J. Sigrist, V. Silventoinen, D. J. Studholme, R. Vaughan, C. H. Wu: InterPro, progress and status in 2005. *Nucleic Acids Res* 33, D201-205 (2005)

27. B. Boeckmann, A. Bairoch, R. Apweiler, M. C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, M. Schneider: The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* 31, 365-370 (2003)

28. M. Y. Galperin: The Molecular Biology Database Collection: 2006 update. *Nucleic Acids Res* 34, D3-5 (2006)

29. Opinion: Obstacles in Nomenclature. Nature 389, 1 (1997)

30. H. F. Liu, Z. Z. Hu, J. Zhang, C. H. Wu: BioThesaurus: a web-based thesaurus of protein and gene names. *Bioinformatics* 22, 103-105 (2006)

31. S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25, 3389-3402 (1997)

32. W. R. Pearson, D. J. Lipman: Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci USA* 85, 2444-2448 (1988)

33. H. Huang, W. C. Barker, Y. Chen, C. H. Wu: iProClass: an integrated database of protein family, function and structure information. *Nucleic Acids Res* 31, 390-392 (2003)

34. C. H. Wu, C. Xiao, Z. Hou, H. Huang, W. C. Barker: iProClass: An integrated and comprehensive protein classification database. *Nucleic Acids Res* 29, 52-54 (2001)

35. C. H. Wu, H. Huang, A. Nikolskaya, Z. Hu, W. C. Barker: The iProClass integrated database for protein functional analysis. *Comput Biol Chem* 28, 87-96 (2004)

36. C. H. Wu, R. Apweiler, A. Bairoch, D. A. Natale, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, R. Mazumder, C. O'Donovan, N. Redaschi, B. Suzek: The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res* 34, D187-191 (2006)

37. D. L. Wheeler, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, L. Y. Geer, W. Helmberg, Y. Kapustin, D. L. Kenton, O. Khovayko, D. J. Lipman, T. L. Madden, D. R. Maglott, J. Ostell, K. D. Pruitt, G. D. Schuler, L. M. Schriml, E. Sequeira, S. T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, T. O. Suzek, R. Tatusov, T. A. Tatusova, L. Wagner, E. Yaschenko: Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 34, D173-180 (2006)

38. C. H. Wu, A. Nikolskaya, H. Huang, L-S. Yeh, D. Natale, C. R. Vinayaka, Z. Hu, R. Mazumder, S. Kumar, P. Kourtesis, R. S. Ledley, B. E. Suzek, L. Arminski, Y. Chen, J. Zhang, J. L. Cardenas, S. Chung, J. Castro-Alvear, G. Dinkov, W. C. Barker: PIRSF family classification system at the Protein Information Resource. *Nucleic Acids Res* 32, D112-114 (2004)

39. A. N. Nikolskaya, C. Arighi, H. Huang, W. C. Barker, C. H. Wu: PIRSF family classification system for protein functional and evolutionary analysis. *Evolutionary Bioinformatics Online* 2, 209-221 (2006).

40. H. Huang, A. N. Nikolskaya, C. R. Vinayaka, S. Chung, J. Zhang, C. H. Wu: Family classification and integrative associative analysis for protein functional annotation. In: *Trends in Bioinformatics Research*, 33-57. Ed: Peter V. Yan *Nova Science Publishers, Inc* (2005)

41. C. R. Vinayaka, C. H. Wu, D. Natale. Large-scale, classification-driven, rule-based functional annotation of proteins. In: Online Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics. Part 4. Bioinformatics 4.3. Protein Function and Annotation. Ed: Shankar Subramaniam, *John Wiley & Sons, Ltd* (2005)

42. W. Li, H. Amri, H. Huang, C. Wu, V. Papadopoulos: Gene and protein profiling of the response of MA-10 Leydig tumor cells to human chorionic gonadotropin. *J Androl* 25, 900-13 (2004)

43. A. Chi, J. C. Valencia, Z. Z. Hu, H. Watabe, H. Yamaguchi, N. Mangini, H. Huang, V. A. Canfield, K. Cheng, J. Shabanowitz, V. J. Hearing, C. Wu, E. Appella, D. F. Hunt: A proteomics and bioinformatics approach to define the biogenesis and function of melanosomes. *J Proteome Res* 5, 3135-44 (2007)

44. Z. Z. Hu, J. C. Valencia, H. Huang, A. Chi, J. Shabanowitz, V. J. Hearing, E. Appella and C. H. Wu: Comparative Bioinformatics Analyses and Profiling of Lysosome-Related Organelle Proteomes. *Int J Mass Spectrom* 259, 147-160 (2007)

45. E. M. Zdobnov, R. Apweiler: InterProScan--an integration platform for the signaturerecognition methods in InterPro. *Bioinformatics* 17, 847-848 (2001)

46. L. Kall, A. Krogh, E. L. Sonnhammer: A combined transmembrane topology and signal peptide prediction method. *J Mol Biol* 338, 1027-1036 (2004)

47. T. Kislinger, B. Cox, A. Kannan, C. Chung, P. Hu, A. Ignatchenko, M. S. Scott, A. O. Gramolini, Q. Morris, M. T. Hallett, J. Rossant, T. R. Hughes, B. Frey, A. Emili: Global survey of organ and organelle protein expression in mouse: combined proteomic and transcriptomic profiling. *Cell* 125, 173-186 (2006)

48. F. Forner, L. J. Foster, S. Campanaro, G. Valle, M. Mann: Quantitative proteomic comparison of rat mitochondria from muscle, heart, and liver. *Mol Cell Proteomics* 5, 608-619 (2006)

49. L. J. Foster, C. L. de Hoog, Y. Zhang, Y. Zhang, X. Xie, V. K. Mootha, M. Mann: A mammalian organelle map by protein correlation profiling. *Cell* 125, 187-199 (2006)

50. M. Huizing, Y. Anikster, W. A. Gahl: Hermansky-Pudlak syndrome and Chediak-Higashi syndrome: disorders of vesicle formation and trafficking. *Thromb Haemost* 86, 233-245 (2001)

51. P. Qiu, Z. J. Wang, K. J. Liu: Ensemble dependence model for classification and prediction of cancer and normal gene expression data. *Bioinformatics* 21, 3114-3121 (2005)

52. P. Qiu, Z. J. Wang, K. J. R. Liu, Z. Z. Hu, C. H. Wu: Dependence Network Modeling for Biomarker Identification. *Bioinformatics* 23,198-206 (2007)

Key Words: Bioinformatics, Proteomics, Data Integration, Database, Dependence Network Modeling, Functional Annotation, Functional Profiling, Gene Ontology, Id Mapping, Lysosome-Related Organelle, Mass Spectrometry, Melanosome, Microarray, Organelle Proteome, Pathway, Protein Mapping, Review

Send correspondence to: Dr. Cathy H. Wu, Department of Biochemistry and Molecular and Cellular Biology, Georgetown University Medical Center, 3300 Whitehaven Street, NW, Suite 1200, Washington, DC 20007, USA, Tel: 202-687-1039, Fax: 202-680-0057, E-mail: wuc@georgetown.edu

http://www.bioscience.org/current/vol12.htm