Challenges and Solutions in Proteomics

Hongzhan Huang[2,¶], Hem D. Shukla[1,¶], Cathy Wu[2] and Satya Saxena[1,*]

[1]Proteomics and Mass Spectrometry Unit, Research Resources Branch, National Institute on Aging, National Institutes of Health, Baltimore, MD, USA and [2]Department of Biochemistry and Molecular & Cellular Biology, Georgetown University Medical Center, Washington DC, USA

Abstract: The accelerated growth of proteomics data presents both opportunities and challenges. Large-scale proteomic profiling of biological samples such as cells, organelles or biological fluids has led to discovery of numerous key and novel proteins involved in many biological/disease processes including cancers, as well as to the identification of novel disease biomarkers and potential therapeutic targets. While proteomic data analysis has been greatly assisted by the many bioinformatics tools developed in recent years, a careful analysis of the major steps and flow of data in a typical high-throughput analysis reveals a few gaps that still need to be filled to fully realize the value of the data. To facilitate functional and pathway discovery for large-scale proteomic data, we have developed an integrated proteomic expression analysis system, iProXpress, which facilitates protein identification using a comprehensive sequence library and functional interpretation using integrated data. With its modular design, iProXpress complements and can be integrated with other software in a proteomic data analysis pipeline. This novel approach to complex biological questions involves the interrogation of multiple data sources, thereby facilitating hypothesis generation and knowledge discovery from the genomic-scale studies and fostering disease diagnosis and drug development.

Received on: November 23, 2006 - Revised on: December 10, 2006 - Accepted on: December 15, 2006

Key Words: Proteomic profiling, high-throughput analysis, biomarkers, bioinformatic tools, iProXpress, sequence library, pathway discovery, stage specific proteins.

INTRODUCTION

The human genome project has revolutionized the practice of biology and the future potential of medicine. The traditional one-gene-at-a-time approach, though effective in revealing detailed molecular functions of individual genes, does not provide a global view of gene function and temporal and spatial regulation for all genes at different physiological or pathological states or developmental stages. Researchers have begun to systematically tackle gene functions and complex regulatory processes by studying organisms on a global scale of genomes: transcriptomes (gene expression) [1], proteomes (protein expression) [2], metabolomes (metabolic networks) [3], and interactomes (protein-protein interactions) [4].

Proteomics aims to identify, characterize and quantify all proteins expressed in cells grown under a variety of conditions [5]. As most biological and disease processes are manifest at the protein level, proteomics has unique and significant advantages as an important complement to genomics and transcriptomics approaches. As a result, there is intense interest in applying proteomics to foster a better understanding of basic biology and disease processes, develop new biomarkers for diagnosis and early detection of disease, and accelerate drug development [6, 7]. Indeed, proteomics has been applied to the discovery of new diagnostic, prognostic and therapeutic targets for numerous diseases, including heart and cardiovascular disorders, cancer, infectious diseases, and diseases of the nervous system [8].

The accelerated growth of proteomics and other large-scale "omics" data presents both opportunities and challenges. The collective richness of the data allows researchers to ask complex biological questions and to gain new scientific insights, as illustrated in the integrated global profiling approach for studying the molecular basis of human cancer [9]. One major challenge, however, lies in the voluminous, complex, and dynamic data being maintained in heterogeneous and distributed sources. Advanced bioinformatics methods are needed for data integration and functional interpretation of large-scale 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 [10]. 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, such as the online tools listed at http://www.proteomecommons.org/tools.jsp. A number of database search programs (e.g., SEQUEST [11], Mascot [12] and X!Tandem [13]) are used to assign probable peptide sequences to MS/MS spectra and to infer protein identities. Utility programs such as PeptideProphet [14] and ProteinProphet [15] are designed to improve the accuracy of peptide and protein identification using statistical models, while DBParser [16] employs a parsimony principal to con-

*Address correspondence to this author at the Proteomics and Mass Spectrometry Unit, Research Resources Branch, National Institute on Aging, National Institutes of Health, 333 Cassell Dr., Baltimore, MD 21224, USA; Tel: (410) 558-8244; Fax: (410) 558-8249;
E-mail: ssaxena@grc.nia.nih.gov

[¶]These authors contributed equally.

1389-2029/07 $50.00+.00

solidate redundant protein assignments. Several publicly available search algorithms were evaluated and benchmarked for sensitivity and specificity recently [17]. Furthermore, an entire pipeline has been developed for experiment annotation, database searching, peptide mining, and protein identification [18].

Once proteins are identified from the biological samples, they need to be analyzed for functional involvement in metabolic and signaling pathways and cellular functions and processes. Many programs have been developed for the biological interpretation of large lists of genes from genome-scale experiments, mostly for microarray gene expression data, with a few being extended to proteomics data. As the Gene Ontology (GO) [19] has become the common standard for genome annotation, most of these programs provide functional analysis in the context of GO (for examples, see http://www.geneontology.org/ GO.tools.microarray.shtml). A few examples include: (i) GoMiner [20], which presents gene lists in GO hierarchical views, (ii) MAPPFinder [21], integrating GO annotations with GenMAPP pathways, (iii) NetAffx [22], which renders a GO graph to display Affymetrix probe sets, (iv) DAVID [23], which includes additional information on Pfam domains [24] and KEGG pathways [25], and (v) Babelomics [26], which includes InterPro motifs [27], KEGG pathways, and Swiss-Prot keywords [28].

CHALLENGES IN PROTEOMICS RESEARCH

A major issue in proteomics and tandem mass spectroscopy protein identification is that the general purpose protein sequence databases leave out many alternative splice isoforms or include them only in the text comments. As a result, proteomic 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. The absence of real protein sequences in the sequence library may further lead to incorrect peptide and protein identification due to the presence of degenerate peptides corresponding to more than one protein in the sequence database.

Another common problem when dealing with a large list of proteins annotated in different places is the lack of standardization. Different protein IDs and names may be used for the same proteins if a different underlying target database is used for MS/MS protein search. Even different versions of the same protein database may result in different IDs if the database identifier is not stable. The lack of common protein identifiers and naming standards presents a challenge for integrating annotations from multiple heterogeneous sources for biological interpretation of proteomic data. Consequently, expression data analysis is often carried out in an ad hoc manner, resulting in a fragmented and inefficient use of rich annotations available in numerous information resources.

SOLUTIONS

Built upon the infrastructure developed by the investigative team at the Protein Information Resource [29], **iProXpress** facilitates protein identification using a comprehensive sequence library and functional interpretation using integrated data. iProXpress has been integrated with other programs in a proteomic data analysis pipeline that includes:

i) A comprehensive sequence library downloadable for use by proteomic search engines,

ii) Protein ID mapping web services,

iii) Protein annotation web services (returning a richly annotated “protein information matrix” from a protein list),

iv) Sequence analysis and protein curation services, and

v) A customized interactive web view for comparative profile analysis of multi-datasets/data types.

iProXpress System Overview

The iProXpress integrated protein expression analysis system is designed for function and pathway discovery from large-scale proteomic data, in a systems biology context, providing rich functional descriptions for individual proteins and detecting functional relationships among them. The system consists of three major software modules to support functionalities in protein mapping, functional annotation and expression profiling Fig. (**1**).

A) Protein Mapping

The protein mapping module is designed to map user-submitted data to corresponding UniProt entries for data analysis in the protein-centric framework. The major functionalities are summarized in Table **1**.

Input Data and Data Types

Currently, the accepted input data are protein IDs and, optionally, their associated peptide sequences, which may be generated from search programs such as SEQUEST or MASCOT. In the future, iProXpress will accept other user-supplied high-throughput data types.

Protein Sequence Library

iProXpress uses a comprehensive sequence library consisting of all known proteins and isoforms from the UniProt databases, namely UniRef100 and UniParc. UniRef100 is a non-redundant reference database maintained at PIR that provides complete non-redundant sequence coverage by combining identical sequences and sub-fragments into single entries. Currently containing about 3.1 million entries, UniRef100 consists of all published protein sequences from several database sources, including UniProtKB (the primary source with 2.8 million entries), and unique sequences from RefSeq, GenPept, IPI, Ensembl and PDB. The sequence space is further expanded by representing the variant sequences (such as splice variants and isoforms annotated in UniProtKB) as separate UniRef100 entries. This data list is extended on a regular basis to guarantee full coverage of sequence space. UniRef100 is the most comprehensive database of its kind due to the curation activities of the UniProt consortium. UniParc, currently containing over 5.5 million sequences, provides a complete sequence archive, including conceptual translations from gene models of newly sequenced genomes. Though many UniParc entries may still undergo sequence revisions, UniParc sequences not present in UniRef100 nevertheless are useful as a supplement to the sequence library for complete coverage.

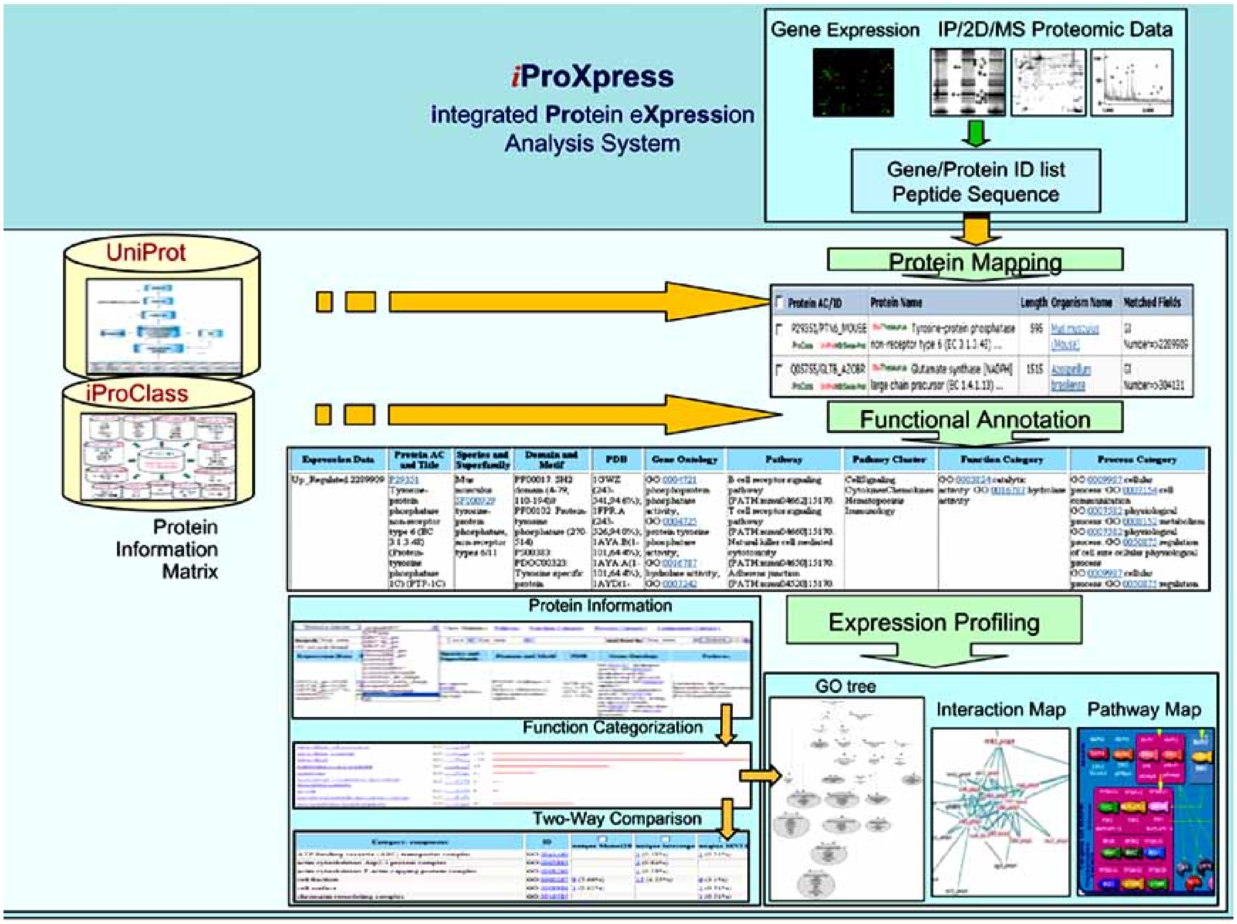


Fig. (1). iProXpress system design.

Table 1. Functionalities of Protein Mapping Module

Functionality	Current	Future
Input Data	Protein ID list, Peptide sequence	Protein or gene ID list, Peptide sequence
Data Type	Mass spectrometry (MS)	MS, 2D gel electrophoresis, immuno-precipitation (IP), gene expression
Sequence Library	UniProt (UniRef100 + unique UniParc)	UniProt (UniRef100 + unique UniParc), Genomic (SNP, EST, gene model), Protein identification DB (PRIDE, gpmDB)
Mapping Method	Protein ID, Peptide	Protein and gene ID, Peptide, Protein and gene name

This UniProt sequence library, consisting of UniRef100 and unique UniParc sequences, has several advantages over NCBI's nr database for use in the MS/MS database search engine because it will have (i) more sequences, (ii) stable identifiers for UniProtKB entries (NCBI's GI number many change from version to version), (iii) more extensive functional annotation, and (iv) less redundancy. While the sequence library covers proteins in source organisms from the entire taxonomic range, options will be provided for users to select sub-datasets based on taxonomy (e.g., human proteins only, or human and mouse proteins only).

In the future, the sequence library will be further expanded to improve the sensitivity of protein identification. New data sources to be integrated will include: (i) the genomic evidence for protein sequence variants, such as coding SNPs, ESTs and gene models, and (ii) protein identification databases such as gpmDB [30] and PRIDE [31] for comparison and cross-reference of spectra and protein identifications.

Protein Mapping Methods

For proteomic data interpretation, a high level of annotation, minimal level of redundancy, and high degree of data integration is critical. iProXpress maps protein lists and their associated peptide sequences to UniProt entries to facilitate functional and pathway analysis. Proteins with various database identifiers are mapped using the PIR ID mapping service, which maps protein and gene IDs from about 30 data sources to UniProt. All major NCBI identifiers, such as GI number and Entrez Gene and RefSeq IDs are included. To cross-validate the ID mapping, the peptide sequence of each mapped protein is matched against the cross-referenced UniProt sequence to confirm the correct assignment. For many-to-one mapping, where multiple IDs map to the same UniProt protein, as is often the case for GI numbers, this mapping removes redundancy effectively.

For proteins that cannot be mapped to UniProt entries through ID mapping, iProXpress searches their peptide sequences against the entire UniRef100 sequence set, or against a species-specific subset if appropriate. There are several mapping scenarios. In the case of one-to-one mapping, where the peptide matches exactly one UniProt protein, that distinct protein receives the assignment. In the case of 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. When all matched protein entries are in the same UniRef90 cluster, the peptides are mapped to the representative sequence from that cluster or to the member protein from the same species. Otherwise, if the proteins belong to different UniRef90 clusters; reliable protein assignments are made manually with retro-inspection of the original MS/MS protein identification results.

Proteins in the input list that are not mapped after ID and peptide mapping to UniRef100 are mapped to the unique UniParc sequence library. Since UniParc is a sequence archive with no functional annotation, the best-hit homologous protein in UniProtKB with at least 90% sequence identity is chosen as the sequence representative. All proteins mapped with sequence variations are flagged for manual validation.

To demonstrate the ability of iProXpress to process large-scale proteomic datasets, we analyzed the proteome of a human embryonic carcinoma stem cell line, NTera2, using our robust offline multidimensional protein identification strategy together with automated linear ion trap mass spectrometry. iProXpress functionally classified and physiochemically characterized approximately 7000 proteins. This large scale proteomic profiling of the Ntera2 cell and its comparison with DNA microarray data contribute to a better understanding of gene regulation at a global level.

B. Functional Annotation

After the protein mapping, rich annotation is presented 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 only 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 proteins/sequence variants not in UniProt. Pre-computed sequence features include homologous proteins in KEGG, BioCarta and other curated pathway databases to populate pathway annotation, InterProScan for family, domain and motif identification, and Phobius for transmembrane helix and signal peptide prediction. Properties derived from homology-based inference are presented in the information matrix with evidence attribution.

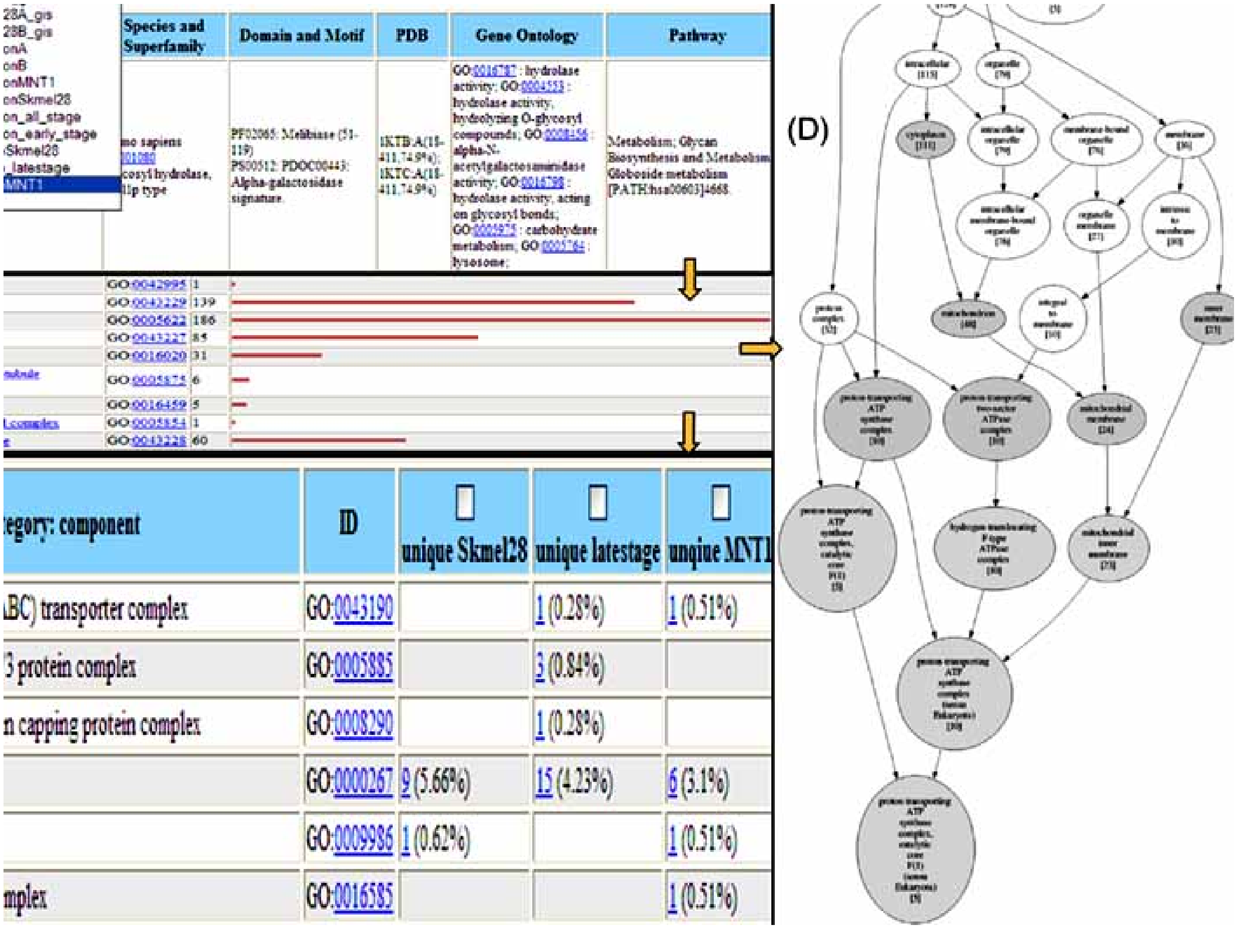
C. Expression Profiling

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 Fig. (**2**), 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 associations 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.

ANALYSIS OF PROTEOMIC DATA USING IPROXPRESS

The iProXpress system (http://pir.georgetown.edu/iproxpress/) has also been applied to the expression profile analysis for hCG-induced changes in MA-10 mouse Leydig tumor cells [32], organelle proteome analysis of various melanosome stages from human melanoma cell lines [33]*, and the comparative analyses of lysosome-related organelle (LRO) proteomes [34]*. Here we use the melanosome proteome analyses to illustrate the integrative approach for function



28A_gis
28B_gis
onA
onB
onMNT1
onSkmel28
on_all_stage
on_early_stage
Skmel28
_latestage
MNT1

Species and Superfamily	Domain and Motif	PDB	Gene Ontology	Pathway

mo sapiens
001080
cosyl hydrolase, ilp 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; GO:0005975 : carbohydrate metabolism; GO:0005764 : lysosome;	Metabolism; Glycan Biosynthesis and Metabolism Globoside metabolism [PATH:hsa00603]4668.

	GO:0042995	1
	GO:0043229	139
	GO:0005622	186
	GO:0043227	85
	GO:0016020	31
nubule	GO:0005875	6
	GO:0016459	5
l complex	GO:0005854	1
e	GO:0043228	60

tegory: component	ID	□ unique Skmel28	□ unique latestage	□ unqiue MNT1
BC) transporter complex	GO:0043190		1 (0.28%)	1 (0.51%)
3 protein complex	GO:0005885		3 (0.84%)	
n capping protein complex	GO:0008290		1 (0.28%)	
	GO:0000267	9 (5.66%)	15 (4.23%)	6 (3.1%)
	GO:0009986	1 (0.62%)		1 (0.51%)
mplex	GO:0016585			1 (0.51%)

(D)
intracellular [115]
organelle [79]
cytoplasm [111]
intracellular organelle [79]
membrane-bound organelle [76]
membrane [36]
intracellular membrane-bound organelle [76]
organelle membrane [27]
intrinsic to membrane [10]
protein complex [52]
mitochondrion [48]
integral to membrane [10]
inner membrane [23]
proton-transporting ATP synthase complex [10]
proton-transporting two-sector ATPase complex [10]
mitochondrial membrane [24]
proton-transporting ATP synthase complex, catalytic core F(1) [5]
hydrogen-translocating F-type ATPase complex [10]
mitochondrial inner membrane [23]
proton-transporting ATP synthase complex (sensu Eukaryota) [10]
proton-transporting ATP synthase complex, catalytic core F(1) (sensu Eukaryota) [5]

Fig. (2). Functional profiling: (A) protein information matrix, (B) functional categorization chart, (C) cross-comparison matrix, (D) graphical GO hierarchy.

and pathway exploration and knowledge discovery using iProXpress.

Melanosomes are membrane-bound organelles 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 bioinformatics analysis of the melanosome proteomic data using the iProXpress system first involved mapping peptide sequences from the MS data and protein lists (NCBI gi numbers from the nr database) to UniProtKB entries. From a total of 2,298 gi numbers, 1,253 (55%) were mapped to UniProtKB following ID mapping, and 1,506 (66%) mapped based on peptide sequences. When the results from both mappings were combined, 1,936 (84%) gi numbers were mapped to 1438 UniProtKB sequences. The mapping reveals 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.

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 (Table **2**). 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 through manual examination. The stage-related proteins provide direct evidence of protein sorting and trafficking to this organelle and provide information about melanosome 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. Using the functional information matrices, melanosome stage-specific proteins were proposed, 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. 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 **2** selectively lists three groups of proteins that are functionally important at each stage of melanosome differentiation: 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 for 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).

Table 2. Stage-Related Melanosome Proteins (Partial List)

Stages	UniProtKB	Gene Name	Protein Name	Notes*
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)	

*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.

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. Thus, the proteomic analysis of melanosomes using iProXpress illustrates that bioinformatics characterization of melanosome proteomes facilitates a better understanding of the biogenesis and function of melanosomes.

SUMMARY

Advances in proteomic technology in combination with state-of-the-art bioinformatics tools have greatly facilitated proteomic data analysis and pathway discovery. iProXpress, a unique bioinformatics tool designed to facilitate functional annotation and pathway discovery from large scale proteomic data, will help researchers to answer complex biological questions by searching multiple data sources. The integration of iProXpress with high-throughput genomics and proteomics technology will expedite the prediction of pathway and functional relationships, thereby supporting hypothesis generation and knowledge discovery from genome wide studies.

ACKNOWLEDGEMENT

This research was supported in part by the Intramural Research Program of the NIH, National Institute on Aging and the NIH grants U01-HG02712 and HHSN2662004000-61C to the Protein Information Resource (PIR) bioinformatics resources.

REFERENCES

[1] Brentani, R.R., Carraro, D.M., Verjovski-Almeida, S., Reis, E. M., Neves, E.J., de Souza, S.J., Carvalho, A.F., Brentani, H., Reis, L. F. Gene expression arrays in cancer research: methods and applications. *Crit. Rev. Oncol. Hematol.* **2005**, *54*: 95-105.

[2] Anderson, N.L., Anderson, N.G. Gene expression arrays in cancer research: methods and applications. *Crit. Rev. Oncol. Hematol.* **2005**, *54*: 95-105.

[3] Bono, H., Nikaido, I., Kasukawa, T., Hayashizaki, Y., Okazaki, Y. Comprehensive analysis of the mouse metabolome based on the transcriptome. *Genome Res.*, **2003**, *13*: 1345-9.

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

[5] Pandey, A., Mann, M. Proteomics to study genes and genomes. *Nature* **2000**, *405*: 837-846.

[6] Srivastava, S., Srivastava, R.G. Proteomics in the Forefront of Cancer Biomarker Discovery. *J. Proteome Res.* **2005**, *4*: 1098-1103.

[7] Hanash, S. Disease proteomics. *Nature* **2003**, *422*: 226-232.

[8] Burkhard, P.R., Rodrigo, N., May, D., Sztajzel, R., Sanchez, J.C., Hochstrasser, D.F., Schiffer, E., Reverdin, A., Lacroix, J.S. Assessing cerebrospinal fluid rhinorrhea: a two-dimensional electrophoresis. *Electrophoresis* **2001**, *22*: 1826-1833.

[9] Hanash, S. Integrated global profiling of cancer. *Nat. Rev. Cancer* **2004**, *4*: 638-644.

[10] Aebersold, R., Goodlett, D.R. Mass spectrometry in proteomics. *Chem. Rev.* **2001**, *101*: 269-295.

[11] Eng, J.K., McCormack, A.L., Yates, J.R.I. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **1994**, *5*: 976-989.

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

[13] Fenyö, D., Beavis, R.C. A Method for Assessing the Statistical Significance of Mass Spectrometry-Based Protein Identifications Using General Scoring Schemes. *Anal. Chem.* **2003**, *75*: 768-774.

[14] Keller, A., Nesvizhskii, A.I., Kolker, E., Aebersold, R. An explanation of the Peptide Prophet algorithm developed. *Anal. Chem.* **2002**, *74*: 5383 –5392.

[15] Nesvizhskii, A.I., Keller, A., Kolker, E., Abersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **2003**, *75*: 4646–4658.

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

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

[18] Rauch, A., Bellew, M., Eng, J., Fitzgibbon, M., Holzman, T., Hussey, P., Igra, M., Maclean, B., Lin, C.W., Detter, A., Fang, R., Faca, V., Gafken, P., Zhang, H., Whiteaker, J., States, D., Hanash, S., Paulovich, A., McIntosh, M.W. 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.* **2006**, *5*: 112-121.

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

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

[21] Doniger, S.W., Salomonis, N., Dahlquist, K.D., Vranizan, K., Lawlo, S.C., Conklin, B.R. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biology* **2003**, *4*: R7.

[22] Cheng, J., Sun, S., Tracy, A., Hubbell, E., Morris, J., Valmeekam, V., Kimbrough, A., Cline, M.S., Liu, G., Shigeta, R., Kulp, D., Siani-Rose, M.A. NetAffx gene ontology mining tool: a visual approach for microarray data analysis. *Bioinformatics* **2004**, *20*: 1462–1463.

[23] Dennis, G., Sherman, B.T., Hosack, D. A., Yang, J., Gao, W., Lane H.C., Lempicki, R.A. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.* **2003**, *4*: R60.

[24] Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Jones, S.G., Khanna, A., Marshall, M., Moxon, S., Erik, L., Sonnhammer, L., Studholme, D.J., Yeats, C., Eddy, S.R. The Pfam protein families' database. *Nucleic Acids Res.* **2004**, *32*: D138-D141.

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

[26] Al-Shahrour, F., Minguez, P., Vaquerizas, J.M., Conde, L., Dopaz, J. BABELOMICS: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments. *Nucleic Acids Res.* **2005**, *33*: W460-W464.

[27] Mulder, N.J., Apweiler R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bradley, P., Bork, P., Bucher, P., Cerutti, L., Copley, R., Courcelle, E., Das, U., Durbin, R., Fleischmann, W., Gough, J., Haft, D., Harte, N., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lonsdale, D., Lopez, R., Letunic, I., Madera, M., Maslen, J., McDowall, J., Mitchell, A., Nikolskaya, A.N., Orchard, S., Pagni., M., Ponting, C.P., Quevillon, E., Selengut, J., Sigrist, C.J.A., Silventoinen, V., Studholme, D.J., Vaughan, R., Wu, C.H. InterPro, progress and status in 2005. *Nucleic Acids Res.* **2005**, *33*: D201-D205.

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

[29] Wu, C.H., Yeh, A.S.L., Huang, H., Arminski, L., Castro-Alvean, J., Chen, Y., Hu, Z., Kourtesis, P., Ledley, R. S., Suzek, B.E., Vinaya-

ka, C.R., Zhang, J., Barker, C.W. The Protein Information Resource. *Nucleic Acids Res.* **2003**, *31*: 345-347.

[30] Craig, R., Cortens, J.P., Beavis, R.C. Open source identification data.system for analyzing, validating, and storing protein. *J. Proteome Res.* **2004**, *3*: 1234-1242.

[31] Martens, L., Vandekerckhove, J., Gevaert, K. DBToolkit: processing protein databases for peptide-centric proteomics. *Bioinformatics* **2005**, *21*: 3584-3585.

[32] Li, W., Amri, H., Huang, H., Wu, C., Papadopoulos, V. Gene and protein profiling of the response of MA-10 Leydig Tumor Cells to Human Chorionic Gonadotropin. *J. Andrology* **2004**, *25*: 900-913.

[33] Chi, A., Valencia, J.C., Hu, Z.Z., Watabe, H., Yamaguchi, H., Mangini, N., Huang, H., Canfield, V.A., Cheng, K., Shabanowitz, J., Hearing, V.J., Wu, C., Appella, E., Hunt, D.F. A proteomics and bioinformatics approach to define the biogenesis and function of melanosomes. *J. Prot. Res.* **2006**, *5*: 3135-3144.

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