THE PIR INTEGRATED PROTEIN DATABASES AND DATA RETRIEVAL SYSTEM

H Huang[1], ZZ Hu[2], BE Suzek[3] and CH Wu[4]

[1-4] Department of Biochemistry and Molecular Biology, Georgetown University Medical Center, 3900 Reservoir Rd., NW, Box 571414, Washington, DC 20057-1414, USA
[1]Email: hh42@georgetown.edu
[2]Email: zh9@georgetown.edu
[3]Email: bes23@georgetown.edu
[4]Email: wuc@georgetown.edu

ABSTRACT

The Protein Information Resource (PIR) provides many databases and tools to support genomic and proteomic research. PIR is a member of UniProt—Universal Protein Resource—the central repository of protein sequence and function, which maintains UniProt Knowledgebase with extensively curated annotation, UniProt Reference databases to speed sequence searches, and UniProt Archive to reflect sequence history. PIR also provides PIRSF family classification system based on evolutionary relationships of full-length proteins, and iProClass integrated database of protein family, function, and structure. These databases are easily accessible from PIR web site using a centralized data retrieval system for information retrieval and knowledge discovery.

Keywords: Protein database, Database curation, Protein family, Data retrieval, Genomics, Proteomics.

1 INTRODUCTION

The high-throughput genome projects have resulted in a rapid accumulation of genome sequences for a large number of organisms. Meanwhile, scientists have begun to systematically tackle gene functions and other complex regulatory processes by studying organisms at the global scale of genomes, proteomes (Babnigg & Giometti, 2003), metabolomes (Bono, Nikaido, Kasukawa, Hayashizaki & Okazaki, 2003), interactomes (Walhout, Reboul, Shtanko, Bertin, Vaglio, Ge et al., 2002), and physiomes (Hunter & Bork, 2003). With accelerated accumulation of high-throughput genomic and proteomic data, computational approaches are increasingly important for deriving scientific knowledge and hypotheses. To fully explore these valuable data, advanced bioinformatics infrastructures must be developed for biological knowledge management. Associated with the enormous quantity and variety of data being produced is the growing number of molecular databases being generated and maintained (Galperin, 2004). One major challenge lies in the volume, complexity, and dynamic nature of the data, which are being collected and maintained in heterogeneous and distributed sources. New approaches need to be devised for data collection, maintenance, dissemination, query, and analysis.

As an integrated public bioinformatics resource to support genomic and proteomic research and scientific discovery, the Protein Information Resource (PIR) (Wu, Yeh, Huang, Arminski, Castro-Alvear, Chen et al., 2003a) provides many databases and analytical tools freely accessible to the scientific community for the past three decades. The major PIR scientific activities involve protein family classification, functional annotation, and data integration. This paper describes the PIR databases, the data organization and retrieval mechanisms, as well as our approach for data integration to facilitate knowledge discovery.

2 DATABASES

2.1 UniProt Protein databases

Protein sequence databases have become a crucial resource for molecular biologists, both as repositories for protein functional and structural data and as starting points for future experiments. Recently, PIR has joined forces with the European Bioinformatics Institute (EBI) and the Swiss Institute of Bioinformatics (SIB) to form the UniProt Consortium that aims to support biological research by maintaining a high quality database that serves as a stable, comprehensive, fully classified, richly and accurately annotated protein sequence knowledgebase, with extensive cross-references and querying interfaces freely accessible to the scientific community (Apweiler, Bairoch, Wu, Barker, Boeckmann, Ferro et al., 2004). Until recently, EBI and SIB together produced Swiss-Prot and TrEMBL (Boeckmann, Bairoch, Apweiler, Blatter, Estreicher, Gasteiger et al., 2003), while PIR produced the PIR-PSD (Wu et al., 2003a). These databases coexisted with differing protein sequence coverage and annotation priorities. The consortium decided to pool their resources, efforts, and expertise to create a central protein sequence and function resource, the UniProt databases.

The UniProt databases consist of three database layers (Figure 1), each optimized for different uses:

- *UniProt Archive (UniParc)* provides a stable, comprehensive sequence collection by storing the complete body of publicly available protein sequence data and reflecting their revision history. While most protein sequence data is derived from the translation of DDBJ/EMBL/GenBank sequences, a large amount of primary protein sequence data resulting from the direct sequencing of proteins is submitted directly to other sources, including Swiss-Prot, TrEMBL, and PIR-PSD; additional protein sequences are found in patent applications, PDB, International Protein Index (IPI), RefSeq, and a few other databases. UniParc represents each protein sequence once and only once, assigning it a unique UniParc identifier, and cross-references the accession numbers of the source databases. UniParc records carry no annotation, but this information can be found in the UniProt Knowledgebase.

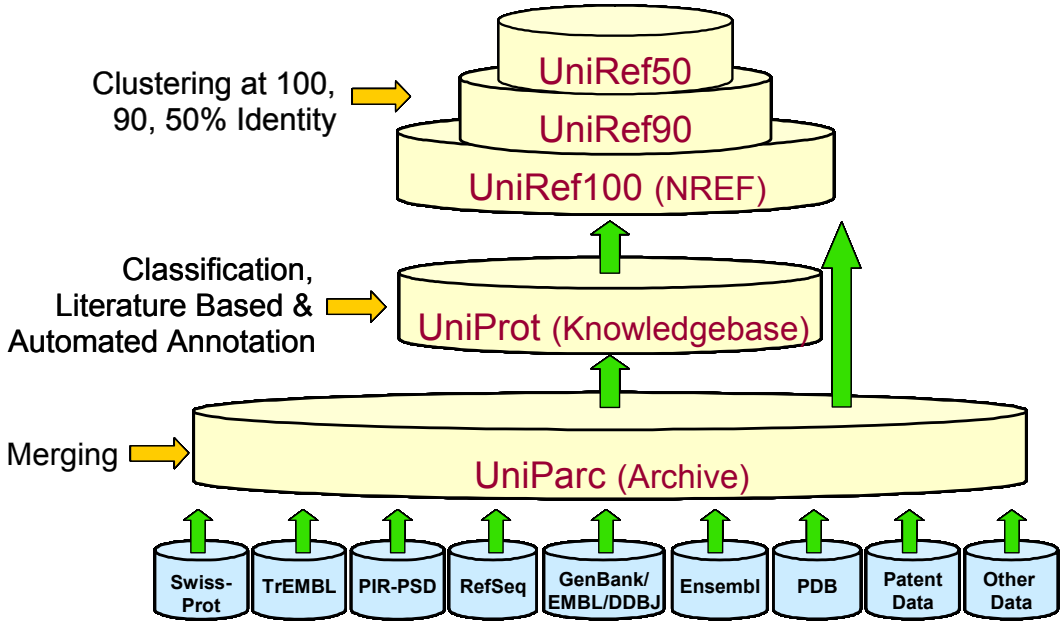


Figure 1. UniProt—Universal Protein Resource—the central repository of protein sequence and function, consisting of three database layers, UniProt Archive, UniProt Knowledgebase, and UniProt Reference databases

- *UniProt Knowledgebase (UniProt)* is the central database of protein sequences with accurate, consistent, and rich sequence and functional annotation and the central access point for extensive curated protein information. The UniProt Knowledgebase contains two major components: a section containing manually-annotated records, based on information from the literature and curator-evaluated computational analysis (referred to as Swiss-Prot); and a section containing computationally-analyzed records awaiting manual annotation (referred to as TrEMBL). In addition, all suitable PIR-PSD entries and annotations not found in Swiss-Prot or TrEMBL are incorporated into the UniProt Knowledgebase. The Knowledgebase aims to describe in a single record all protein products derived from a certain gene from a certain species and to give not only the whole record an accession number but to assign to each protein form derived by alternative splicing, proteolytic cleavage, and post-translational modification isoform identifiers, which are accession numbers for the isoforms.

- *UniProt NREF (UniRef*) databases provide non-redundant reference collections of sequence data to speed searches at several resolutions (100, 90, 50% similarity) by combining closely related sequences across different species into a single record. UniRef100 is based on all UniProt Knowledgebase records, as well as UniParc records that represent sequences deemed over-represented in the Knowledgebase, DDBJ/EMBL/GenBank WGS CDS translations, Ensembl protein translations from various organisms, as well as IPI data. The production of UniRef100 begins with the clustering of all records by sequence identity. Identical sequences and subfragments are presented as a single UniRef100 entry, containing the accession numbers of all merged entries, the protein sequence, and a bibliography. UniRef90 and UniRef50 are built from UniRef100 to provide non-redundant sequence collections for faster homology searches. All records having > 90% or > 50% identity are merged together into a single UniRef90 or UniRef50 entry, respectively. The UniRef90 set is approximately 40% smaller than UniRef100, and the UniRef50 set is approximately 65% smaller than UniRef100.

PIR Evidence Attribution and Literature Mining, To promote annotation quality and database interoperability, UniProt employs controlled vocabulary for most annotations and adopts standard nomenclature whenever applicable. In the PIR-PSD database, feature annotation such as binding sites, catalytic sites, and modified sites are labeled with status tags "*experimental,*" "*predicted,*" "*absent*," or "*atypical*" to distinguish experimentally verified from computationally predicted data (Figure 2A). However, such "*experimental*" tag was not originally attributed with literature citations for the experimental evidence, even though the relevant citations are usually present in the Reference section of the PSD sequence report. To appropriately attribute bibliographic data to features with experimental evidence, we have been conducting a retrospective literature survey (Wu, Huang, Yeh & Barker, 2003b). The evidence-attributed PSD experimental feature data are being incorporated into the UniProt Knowledgebase. The literature survey involves both citation mapping (finding citations from the Reference section that describe the given experimental feature) and evidence tagging (tagging the sentences providing experimental evidence in an abstract and/or full-text article) (Figure 2B). Such curator-tagged texts can be used as training corpus to develop and benchmark Natural Language Processing-based methods for literature mining and annotation extraction (Hirschman, Park, Tsuji, Wong & Wu, 2002). PIR has developed a resource for protein literature mining—iProLINK (integrated Protein Literature, INformation and Knowledge), with special focuses on bibliography mapping, annotation extraction, protein named entity recognition, and protein ontology development (Hu, Mani, Hermoso, Liu, Wu, 2004). Annotation-tagged literature corpus, including several hundred abstracts and full-text articles tagged with experimentally validated post-translational modifications annotated in PIR-PSD, is directly accessible from iProLINK.

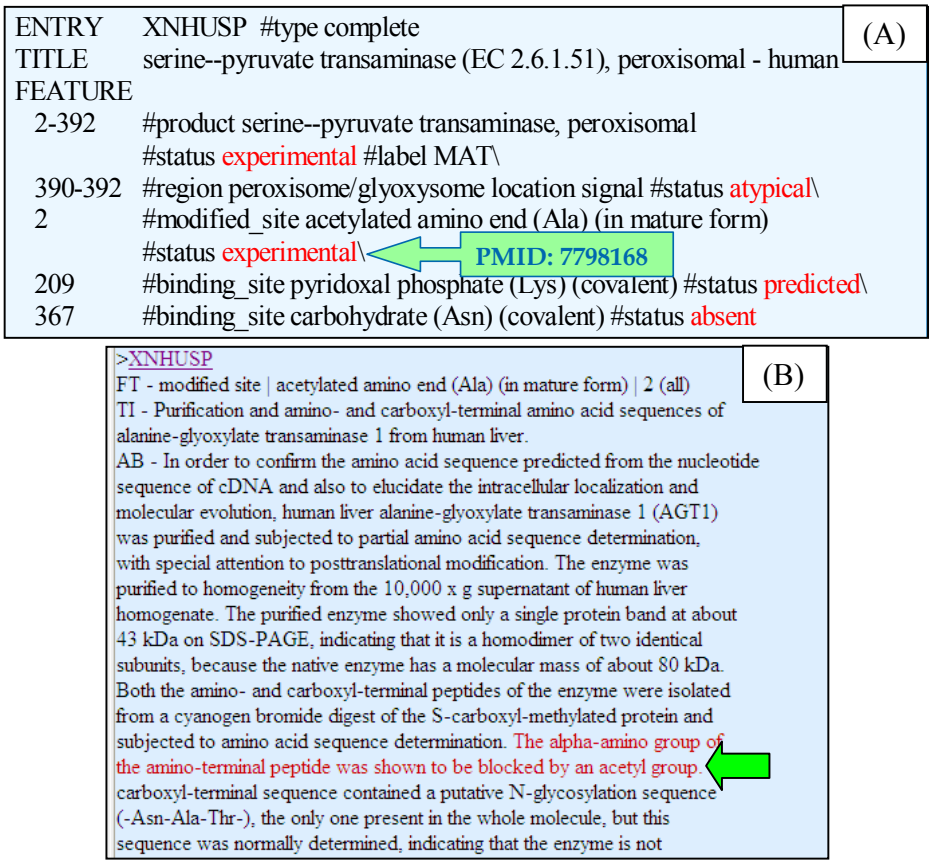


Figure 2. Evidence attribution of sequence features in PIR-PSD. (A) Controlled vocabulary for status tags and citation mapping with PubMed ID; (B) Annotation-tagged text where experimental evidence is indicated (see full report in http://pir.georgetown.edu/cgi-bin/retrosurvey.pl?id=XNHUSP)

2.2 PIRSF protein family classification system

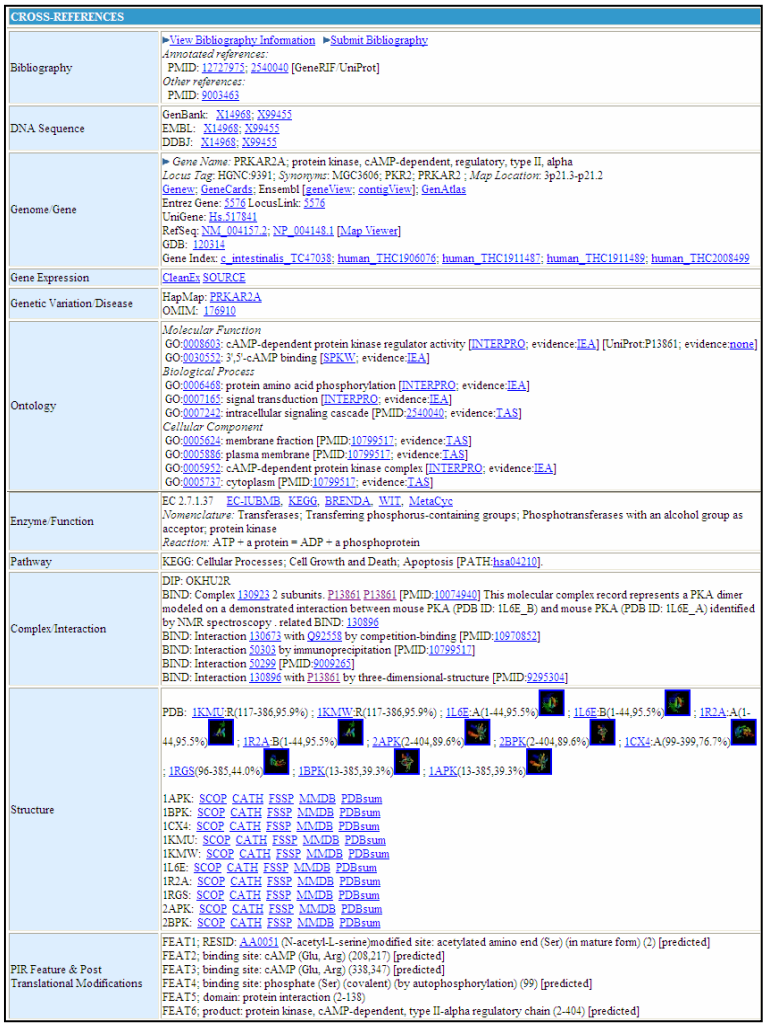
Protein family classification has been central to the annotation activities PIR since the pioneering work of Margaret Dayhoff. The original PIR superfamily concept (Dayhoff, 1976) was based on sequence similarity where protein family members are homologous (sharing common ancestry) and homeomorphic (sharing full-length sequence similarity with common domain architecture). It has been used as a guiding principle to provide comprehensive and non-overlapping clustering of protein sequences into a hierarchical structure to reflect their evolutionary relationships (Barker, Pfeiffer & George, 1996). To facilitate sensible propagation and standardization of protein annotation and systematic detection of annotation errors as part of the UniProt project, PIR has extended its hierarchical superfamily concept to the PIRSF system, a network classification system based on the evolutionary relationships of whole proteins (Wu, Nikolskaya, Huang, Yeh, Natale, Vinayaka et al., 2004a). Classification based on whole proteins, rather than on the component domains, allows annotation of both generic biochemical and specific biological functions. Furthermore, it permits the classification of proteins without well-defined domains. The network classification system accommodates a flexible number of levels that reflect varying degrees of sequence conservation. Such structure allows improved protein annotation, more accurate extraction of conserved functional residues, and classification of distantly related orphan proteins. The PIRSF system definition and working principles are detailed in the document, *A Proposal for the PIRSF Classification System*, available from the PIR web site (Protein Information Resource, 2003).

The PIRSF database consists of two data sets, preliminary clusters and curated families. Currently, about two-thirds of UniProt sequences are classified into over 32,000 preliminary clusters, including single-member clusters. The preliminary clusters are computationally defined using both pairwise-based parameters and cluster-based parameters. Systematic family curation is being conducted in a two-tier process to improve the quality of automated classification. Over 5000 families containing two or more members have been curated at the "first-tier" for membership and domain architecture characteristic of the family. The second-tier curation provides additional annotation, including family name, parent-child relationship, family description, and bibliography. Several hundred second-tier curated PIRSF families have been integrated into InterPro (Mulder, Apweiler, Attwood, Bairoch, Barrell, Bateman et al., 2003). Rule-based and classification-driven procedures are used to propagate information-rich annotations among similar sequences and to perform integrity checks based on PIR controlled vocabulary and thesaurus of synonyms or alternate names (Wu et al., 2003b).

2.3 iProClass integrated protein database

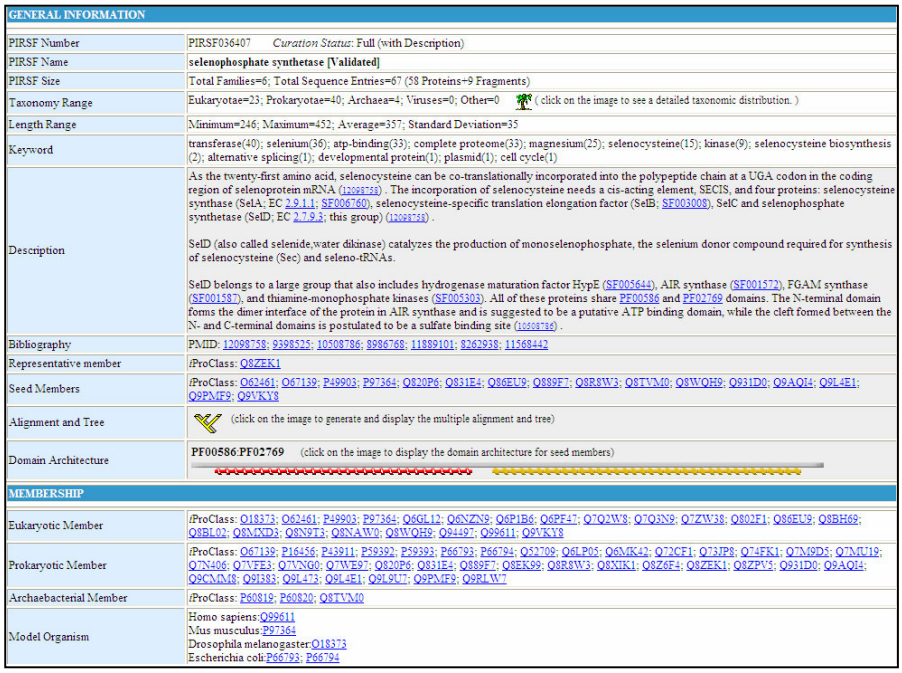
Originally designed to address data integration issues arising from the voluminous, heterogeneous, and distributed data (Wu, Xiao, Hou, Huang & Barker, 2001), the iProClass database (Wu, Huang, Nikolskaya, Hu & Barker, 2004b) provides comprehensive descriptions of proteins and serves as a framework for data integration in a distributed networking environment. The iProClass database contains comprehensive descriptions of all proteins with up-to-date information from many sources, thereby providing much richer annotation than can be found in any single database. Providing value-added views of the UniProt knowledgebase, iProClass includes information on protein family relationships at both global (superfamily/family) and local (domain, motif, site) levels, as well as structural and functional classifications and features of proteins. The database currently consists of more than 1.5 million UniProt sequence entries organized with 36,000 PIRSF families, 7600 domains, 1300 motifs, and 550,000 similarity clusters. iProClass provides rich links to over 90 molecular databases with source attribution, hypertext links, and related summary information extracted from the underlying sources, including databases of protein families (e.g., COG, InterPro), functions and pathways (e.g., KEGG, WIT), protein-protein interactions (e.g., DIP, BIND), post-translational modifications (e.g., RESID), structures and structural classifications (e.g., PDB, SCOP, CATH), genes and genomes (e.g., TIGR, GDB, OMIM), ontologies (e.g., GO), literature (PubMed), and taxonomy (NCBI Taxonomy).

Protein sequence (Figure 3) and PIRSF family (Figure 4) summary reports present extensive annotation information and include membership statistics and graphical display of domains and motifs. iProClass is implemented in Oracle 9i database management system, updated biweekly, and searchable by both sequence (BLAST search and peptide match) and text (unique identifiers and combinations of text strings).



CROSS-REFERENCES	
Bibliography	▶View Bibliography Information ▶Submit Bibliography *Annotated references:* PMID: 12727975; 2540040 [GeneRIF/UniProt] *Other references:* PMID: 9003463
DNA Sequence	GenBank: X14968; X99455 EMBL: X14968; X99455 DDBJ: X14968; X99455
Genome/Gene	▶ *Gene Name:* PRKAR2A; protein kinase, cAMP-dependent, regulatory, type II, alpha *Locus Tag:* HGNC:9391; *Synonyms:* MGC3606; PKR2; PRKAR2 ; *Map Location:* 3p21.3-p21.2 Genew; GeneCards; Ensembl [geneView; contigView]; GenAtlas Entrez Gene: 5576 LocusLink: 5576 UniGene: Hs.517841 RefSeq: NM_004157.2; NP_004148.1 [Map Viewer] GDB: 120314 Gene Index: c_intestinalis_TC47038; human_THC1906076; human_THC1911487; human_THC1911489; human_THC2008499
Gene Expression	CleanEx SOURCE
Genetic Variation/Disease	HapMap: PRKAR2A OMIM: 176910
Ontology	*Molecular Function* GO:0008603: cAMP-dependent protein kinase regulator activity [INTERPRO; evidence:IEA] [UniProt:P13861; evidence:none] GO:0030552: 3',5'-cAMP binding [SPKW; evidence:IEA] *Biological Process* GO:0006468: protein amino acid phosphorylation [INTERPRO; evidence:IEA] GO:0007165: signal transduction [INTERPRO; evidence:IEA] GO:0007242: intracellular signaling cascade [PMID:2540040; evidence:TAS] *Cellular Component* GO:0005624: membrane fraction [PMID:10799517; evidence:TAS] GO:0005886: plasma membrane [PMID:10799517; evidence:TAS] GO:0005952: cAMP-dependent protein kinase complex [INTERPRO; evidence:IEA] GO:0005737: cytoplasm [PMID:10799517; evidence:TAS]
Enzyme/Function	EC 2.7.1.37 EC-IUBMB, KEGG, BRENDA, WIT, MetaCyc *Nomenclature:* Transferases; Transferring phosphorus-containing groups; Phosphotransferases with an alcohol group as acceptor; protein kinase *Reaction:* ATP + a protein = ADP + a phosphoprotein
Pathway	KEGG: Cellular Processes; Cell Growth and Death; Apoptosis [PATH:hsa04210].
Complex/Interaction	DIP: OKHU2R BIND: Complex 130923 2 subunits. P13861 P13861 [PMID:10074940] This molecular complex record represents a PKA dimer modeled on a demonstrated interaction between mouse PKA (PDB ID: 1L6E_B) and mouse PKA (PDB ID: 1L6E_A) identified by NMR spectroscopy . related BIND: 130896 BIND: Interaction 130673 with Q92558 by competition-binding [PMID:10970852] BIND: Interaction 50303 by immunoprecipitation [PMID:10799517] BIND: Interaction 50299 [PMID:9009265] BIND: Interaction 130896 with P13861 by three-dimensional-structure [PMID:9295304]
Structure	PDB: 1KMU:R(117-386,95.9%) ; 1KMW:R(117-386,95.9%) ; 1L6E:A(1-44,95.5%) ; 1L6E:B(1-44,95.5%) ; 1R2A:A(1-44,95.5%) ; 1R2A:B(1-44,95.5%) ; 2APK(2-404,89.6%) ; 2BPK(2-404,89.6%) ; 1CX4:A(99-399,76.7%) ; 1RGS(96-385,44.0%) ; 1BPK(13-385,39.3%) ; 1APK(13-385,39.3%) 1APK: SCOP CATH FSSP MMDB PDBsum 1BPK: SCOP CATH FSSP MMDB PDBsum 1CX4: SCOP CATH FSSP MMDB PDBsum 1KMU: SCOP CATH FSSP MMDB PDBsum 1KMW: SCOP CATH FSSP MMDB PDBsum 1L6E: SCOP CATH FSSP MMDB PDBsum 1R2A: SCOP CATH FSSP MMDB PDBsum 1RGS: SCOP CATH FSSP MMDB PDBsum 2APK: SCOP CATH FSSP MMDB PDBsum 2BPK: SCOP CATH FSSP MMDB PDBsum
PIR Feature & Post Translational Modifications	FEAT1; RESID: AA0051 (N-acetyl-L-serine)modified site: acetylated amino end (Ser) (in mature form) (2) [predicted] FEAT2; binding site: cAMP (Glu, Arg) (208,217) [predicted] FEAT3; binding site: cAMP (Glu, Arg) (338,347) [predicted] FEAT4; binding site: phosphate (Ser) (covalent) (by autophosphorylation) (99) [predicted] FEAT5; domain: protein interaction (2-138) FEAT6; product: protein kinase, cAMP-dependent, type II-alpha regulatory chain (2-404) [predicted]

Figure 3. iProClass protein sequence report with rich cross references (see full report in http://pir.georgetown.edu/cgi-bin/ipcEntry?id=P13861)



GENERAL INFORMATION	
PIRSF Number	PIRSF036407 *Curation Status*: Full (with Description)
PIRSF Name	**selenophosphate synthetase [Validated]**
PIRSF Size	Total Families=6; Total Sequence Entries=67 (58 Proteins+9 Fragments)
Taxonomy Range	Eukaryotae=23; Prokaryotae=40; Archaea=4; Viruses=0; Other=0 (click on the image to see a detailed taxonomic distribution.)
Length Range	Minimum=246; Maximum=452; Average=357; Standard Deviation=35
Keyword	transferase(40); selenium(36); atp-binding(33); complete proteome(33); magnesium(25); selenocysteine(15); kinase(9); selenocysteine biosynthesis (2); alternative splicing(1); developmental protein(1); plasmid(1); cell cycle(1)
Description	As the twenty-first amino acid, selenocysteine can be co-translationally incorporated into the polypeptide chain at a UGA codon in the coding region of selenoprotein mRNA (12098758) . The incorporation of selenocysteine needs a cis-acting element, SECIS, and four proteins: selenocysteine synthase (SelA; EC 2.9.1.1; SF006760), selenocysteine-specific translation elongation factor (SelB; SF003009), SelC and selenophosphate synthetase (SelD; EC 2.7.9.3; this group) (12098758) . SelD (also called selenide,water dikinase) catalyzes the production of monoselenophosphate, the selenium donor compound required for synthesis of selenocysteine (Sec) and seleno-tRNAs. SelD belongs to a large group that also includes hydrogenase maturation factor HypE (SF005644), AIR synthase (SF001572), FGAM synthase (SF001587), and thiamine-monophosphate kinases (SF005303). All of these proteins share PF00586 and PF02769 domains. The N-terminal domain forms the dimer interface of the protein in AIR synthase and is suggested to be a putative ATP binding domain, while the cleft formed between the N- and C-terminal domains is postulated to be a sulfate binding site (16608786) .
Bibliography	PMID: 12098758; 9398525; 10508786; 8986768; 11889101; 8262938; 11568442
Representative member	*i*ProClass: Q8ZEK1
Seed Members	*i*ProClass: O62461; O67139; P49903; P97364; Q820P6; Q831E4; Q86EU9; Q889F7; Q8R8W3; Q8TVM0; Q8WQH9; Q931D0; Q9AQI4; Q9L4E1; Q9PMF9; Q9VKY8
Alignment and Tree	(click on the image to generate and display the multiple alignment and tree)
Domain Architecture	**PF00586:PF02769** (click on the image to display the domain architecture for seed members)
MEMBERSHIP	
Eukaryotic Member	*i*ProClass: O18373; O62461; P49903; P97364; Q6GL12; Q6NZN9; Q6P1B6; Q6PF47; Q7Q2W8; Q7Q3N9; Q7ZW38; Q802F1; Q86EU9; Q8BH69; Q8BL02; Q8MXD3; Q8N9T3; Q8NAW0; Q8WQH9; Q94497; Q99611; Q9VKY8
Prokaryotic Member	*i*ProClass: O67139; P16456; P43911; P59392; P59393; P66793; P66794; Q52709; Q6LP05; Q6MK42; Q72CF1; Q73JP8; Q74FK1; Q7M9D5; Q7MU19; Q7N406; Q7VFE3; Q7VNG0; Q7WE97; Q820P6; Q831E4; Q889F7; Q8EK99; Q8R8W3; Q8XIK1; Q8Z6F4; Q8ZEK1; Q8ZPV5; Q931D0; Q9AQI4; Q9CMM8; Q9I383; Q9L473; Q9L4E1; Q9L9U7; Q9PMF9; Q9RLW7
Archaebacterial Member	*i*ProClass: P60818; P60820; Q8TVM0
Model Organism	Homo sapiens:Q99611 Mus musculus:P97364 Drosophila melanogaster:O18373 Escherichia coli:P66793; P66794

Figure 4. PIRSF family report with family and membership information (see full report in http://pir.georgetown.edu/cgi-bin/ipcSF?id=PIRSF036407)

The data integration in iProClass, coupled with the PIRSF classification, allows us to identify interesting relationships between protein sequence, structure, and function (Wu et al., 2004b). For example, Pfam domain-based searches can identify all PIRSFs sharing one or more pfam domains (Bateman, Coin, Durbin, Finn, Hollich, Griffiths-Jones et al., 2004). Likewise, SCOP (Andreeva, Howorth, Brenner, Hubbard, Chothia & Murzin, 2004) structural classification-based searches can identify PIRSFs in common SCOP superfamily class. In combination with the underlying taxonomic information, one can retrieve PIRSFs that occur only in given lineages or share common phyletic patterns. Functional convergence and divergence can be revealed by the many-to-one and one-to-many relationships between the enzyme classification (EC number) and PIRSF classification. Such knowledge is fundamental to the understanding of protein evolution, structure, and function, and crucial to functional genomic and proteomic research.

3 DATA RETRIEVAL SYSTEM

3.1 Data storage

The bioinformatics infrastructure at the PIR is devised to support the management and retrieval of large amounts of data. We have considered and addressed several data storage issues.

- *The file sizes are too large to fit one hard disk or volume.* One major data collection underlying our databases is the Related Sequence database, consisting of pre-computed BLAST neighbors of UniProt sequences. The database is compiled from the all-against-all BLAST search results of all 1.5 million sequences, and requires more 100 gigabytes of storage space. To store such big files, one can use hardware-based solutions, such as network attached storage (NAS) or storage area networks (SAN). Alternatively, one can use software-based solutions, such as distributing data to different hard disks and developing

indexing schemes accordingly. We use the latter approach since the hardware solution becomes expensive with increasing storage need and it is easier to manage smaller files distributed to different hard disks.

- *Some old 32-bit applications or software development libraries limit the maximum file size to 2 or 4 GB*. We divide the large data files into at most 2GB chunks to maintain the compatibility issues stemming from 32-bit development.
- *There is always a need for more space, but not all data are used frequently.* At PIR we prioritize files according to their usage. While heavily used files are stored in local disks, the rest are moved to disks mounted from Linux-based network file system (NFS) servers. This approach not only saves local disk space but also provides scalability and cost-effectiveness.

3.2 Data indexing

To support efficient data retrieval, PIR uses a text indexing and an entry indexing systems, in addition to the indexing of the Oracle 9i database management system.

The text indexing system is used to index about 60 text and unique identifier fields extracted from the underlying protein databases (e.g., UniProt ID, protein name, Superfamily Description, MedLine abstract, and author name). The system is developed using the Callable Personal Librarian (CPL) (America Online, 2003), a C-callable application programmer's interface (API). Since CPL originally was not developed for the bioinformatics domain, some default configurations and working mechanisms have been modified as detailed below.

- By default, the CPL tokenizer ignores characters other than numbers and letters when indexing text. However, special characters such as period (.) and hyphen (-) are often part of the word or identifier (e.g. "EC 1.9.3.1" or "5-phosphate") in protein databases. CPL tokenizer rules are modified to recognize such characters. Such rule change does not cause many tokenization discrepancies because the indexed fields do not contain many sentences that end with full-stop.
- CPL, unlike SQL, does not have a documented "NOT NULL" operator. Hence, to accommodate queries checking for the presence or absence of a certain field in a record, a hidden field that keeps a list of fields containing non-null values in that record is created.
- CPL is capable of substring searches using the wildcard character (*). For instance, one can use the query "*acetyl*" to search for "deacetylase." Unfortunately, the performance of CPL in processing such queries is poor. To work around this problem, PIR indexes all possible substrings of letter-based words. Since not all substrings are meaningful and the indexes take up large storage space, one future improvement would be creating a dictionary of meaningful substrings for every bioinformatics term and indexing accordingly.

To provide fast retrieval of entries in plain text format from our databases, an entry indexing system using a C language-based utility "uniprot_index" is developed. The indexing utility generates a mapping of protein entry identifiers to the file offsets where corresponding entries reside in the data file.

3.3 Data retrieval

The PIR Data Retrieval System (Figure 5) allows an easy access to all PIR databases. The retrieval system consists of three components, corresponding to the three indexing systems described above.

- *Text Search System:* This CPL-based search system uses the text indexing system to support exact text searches, substring searches (e.g., "phosp" instead of "phosphorylation"), and range searches (e.g., "10 <" sequence length).
- *Entry Retrieval System:* This system uses a C language retrieval utility called "uniport_get" for entry retrieval based on indexes generated by the entry indexing system. It supports direct record retrieval using entry identifiers. For instance, for a given UniProt identifier, one can retrieve the protein report or the result of BLAST searches from Related Sequence database. On an AlphaServer 4100 computer running Tru64 V5.1, the average entry retrieval time for "nget" varies between 10 and 80 ms depending on the entry size.
- *Perl Database Interface (DBI) and Oracle Database Driver (DBD):* For databases stored in the Oracle 9i database management system, the queries are processed via the Perl DBI module (Perl DBI, 2003) and the Oracle DBD driver (Bunce, 2003).

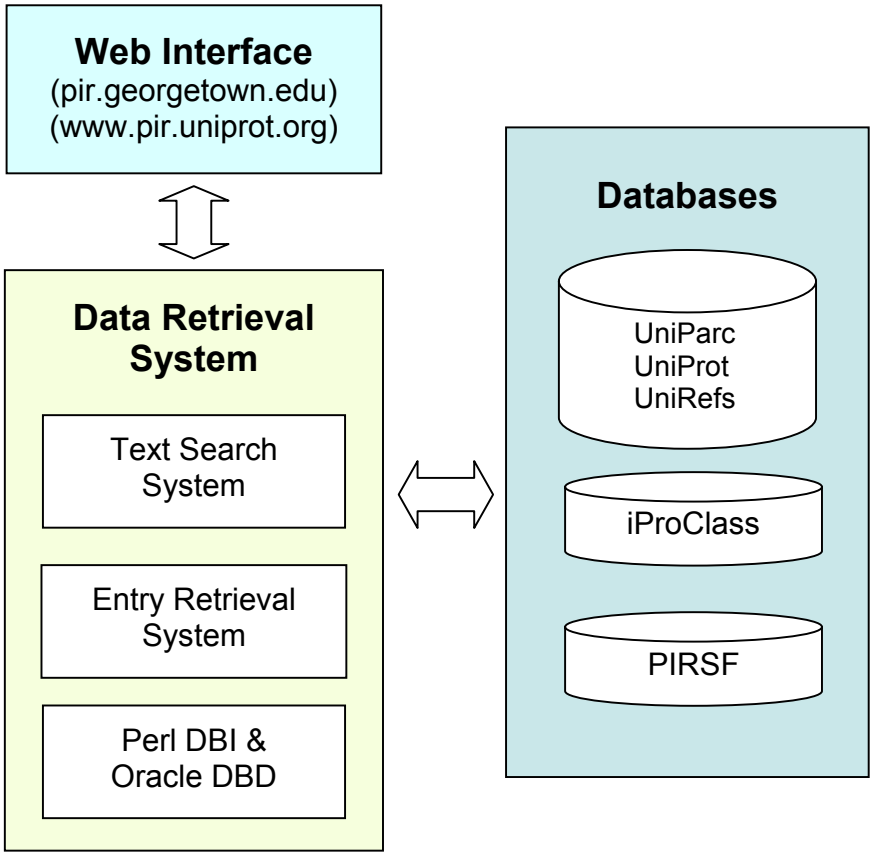


Figure 5. PIR data retrieval system for easy access to all PIR databases

4 PIR WEB INTERFACE

With integrated databases and centralized data retrieval system, PIR allows users to answer complex biological questions that may typically involve querying multiple sources and serves as a primary resource for exploration of proteins information.

The PIR web site (Protein Information Resource, 2004a) connects data mining and sequence analysis tools to underlying databases for information retrieval and knowledge discovery, with functionalities for interactive queries, combinations of sequence and annotation text searches, and sorting and visual exploration of search results. The databases are accessible by text search for entry and list retrieval, as well as BLAST search and peptide match. Direct entry report retrieval is based on sequence unique identifiers of all underlying databases. Species-based browsing and searching are supported for about 150 organisms, including over 120 complete genomes. Basic and advanced text searches return protein entries listed in summary lines with information on protein IDs, matched fields, protein name, taxonomy, superfamily, domain, and motif, with hypertext links to the full entry report and to cross-referenced databases. More than 50 fields are searchable, including about 30 database unique identifiers (e.g., PDB ID, EC number, PubMed ID, and KEGG pathway number) and a wide range of annotation texts (e.g., protein name, organism name, sequence feature, and paper title) (Figure 6), as listed in http://pir.georgetown.edu/cgi-bin/searchExpl (Protein Information Resource, 2004b). The BLAST search and peptide search likewise return lists of matched entries with summary lines that also contain search statistics and matched sequence region. Protein entries returned from text and sequence searches can be selected for further analysis, including BLAST (Altschul, Madden, Schaffer, Zhang, Zhang, Miller et al., 1997) and FASTA search (Person & Lipman, 1988), pattern match, ClustalW (Thompson, Higgins & Gibson, 1994) multiple sequence alignments, and graphical display of superfamily, domain and motif relationships. The related sequences in FASTA clusters are retrievable based on sequence unique

identifiers where neighbors are listed with annotation information and graphical display of matched sequence region. Similar PIR web design and data retrieval mechanisms have been adopted in the UniProt web interface (UniProt, 2004).

The UniProt databases and iProClass are updated biweekly and made immediately available from the PIR web site for searching and browsing, as well as from the FTP site for free downloading. UniProt, UniRefs and iProClass are distributed XML formats with associated DTD (Document Type Definition) file. The sequence files of the databases are distributed in FASTA format.

	Search Field	Example					
1	Author Name	Huberman		29	PC Motif ID	PCM00487	UID
2	BIND ID	10658	UID	30	PDB ID	1B3O:A	UID
3	BLOCKS ID	IPB000644	UID	31	PIR Accession	I52303	UID
4	COG ID	COG1009	UID	32	PIR HD ID	HD00040	UID
5	Common Name	man		33	PIR ID	A31997	UID
6	EC Number	1.1.1.205	UID	34	PRINTS ID	PR01434	UID
7	Enzyme Name	Oxidoreductases		35	Paper Title	dehydrogenase gene	
8	FLY ID	FBgn0002940	UID	36	Pfam ID	PF00478	UID
9	Family#	FAM0008667	UID	37	Pfam Name	IMP dehydrogenase	
10	Feature	active site:		38	Prosite Motif Name	GMP reductase signature	
11	GDB ID	128086	UID	39	NREF Entry Name	IMP dehydrogenase	
12	GO ID	0006177	UID	40	Protein Name	IMP dehydrogenase	
13	GO Term	GMP biosynthesis		41	PubMed ID	7999076	UID
14	GenBank/DDBJ/EMBL ID	J04208	UID	42	RESID ID	AA0005	UID
15	Gene Name	IMPDH2		43	Refseq Accession	NP_213588	UID
16	Genpept Accession	AAH15567.1	UID	44	Refseq ID	g15606211	UID
17	Genpept ID	g15990412	UID	45	SGD ID	S0001047	UID
18	Journal Name	Biochem Pharmacol		46	Superfamily Name	IMP dehydrogenase	
19	KEGG Pathway	Metabolism		47	Superfamily#	SF000130	UID
20	KEGG Pathway ID	hsa00230	UID	48	SwissProt Accession	P12268	UID
21	Keyword	NAD		49	SwissProt ID	IMD2_HUMAN	UID
22	LOCUS ID	3615	UID	50	TIGR ID	PFB0865w	UID
23	MGI ID	87986	UID	51	Taxon Group	Euk/Animal	
24	NCBI GI Number	g15606211	UID	52	Taxon Group ID	2759	UID
25	NCBI Taxon ID	9606	UID	53	TrEMBL Accession	O67027	UID
26	NREF ID	NF00078343	UID	54	TrEMBL ID	O67027	UID
27	OMMIM ID	146691	UID	55	UWGP ID	b3640	UID
28	Organism Name	Homo sapiens		56	iProclass ID	A31997+IMD2_HUMAN	UID

Figure 6. Text search fields in PIR data retrieval system, including unique identifiers (UID) of various databases and annotation texts (e.g. protein name, keyword)

5 ACKNOWLEDGMENTS

The work at PIR is supported by grant U01-HG02712 from the National Institutes of Health and grants DBI-0138188 and ITR-0205470 from the National Science Foundation. The computing resources are partially supported by an Academic Excellence Grant (AEG) grant from Sun Microsystems and a Shared University Research (SUR) grant from IBM.

6 REFERENCES

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res. 25*(17), 3389-3402.

America Online (2003) Callable Personal Librarian (CPL) (version 6.5). Retrieved August 6, 2003 from http://web.archive.org/web/20030806033023/http://www.pls.com/index.html

Andreeva, A., Howorth, D., Brenner, S.E., Hubbard, T.J., Chothia, C., & Murzin, A.G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res*. *32* (Database issue), D226-229.

Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N., & Yeh, L.-S. (2004) UniProt: Universal Protein Knowledgebase. *Nucleic Acids Res*. *32* (Database issue), D115-119.

Babnigg, G. & Giometti, C.S. (2003) ProteomeWeb: a web-based interface for the display and interrogation of proteomes. *Proteomics 3*(5), 584-600.

Barker, W.C., Pfeiffer, F. & George, D.G. (1996) Superfamily classification in PIR-intemational protein sequence database. *Methods Enzymol*. *266*, 59-71.

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

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. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*. *31*(1), 365-370.

Bono, H., Nikaido, I., Kasukawa, T., Hayashizaki, Y., & Okazaki, Y. (2003) Comprehensive analysis of the mouse metabolome based on the transcriptome. *Genome Res*. *13*(6B), 1345-1349.

Bunce, T. (2003) DBD-Oracle-1.14. Retrieved January 20, 2004 from the CPAN Seach website: http://search.cpan.org/~timb/DBD-Oracle-1.14/

Dayhoff, M.O. (1976) The origin and evolution of protein superfamilies. *Fed. Proc*. *35*(10), 2132-2138.

Galperin, M.Y. (2004) The Molecular Biology Database Collection: 2004 update. *Nucleic Acids Res*. *32* (Database issue), D3-22.

Hirschman, L., Park, J.C., Tsuji, J., Wong, L., & Wu, C.H. (2002) Accomplishments and challenges in literature data mining for biology. *Bioinformatics 18*(12), 1553-1561.

Hu, Z.Z., Mani, I., Hermoso, V., Liu, H., & Wu, C.H. (2004) iProLINK: an integrated protein resource for literature mining. *Comput. Biol. Chem*. *28*, 409-416.

Hunter, P.J. & Borg, T.K. (2003) Integration from proteins to organs: the Physiome Project. *Nat. Rev. Mol. Cell Biol. 4*(3), 237-243.

Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Barrell, D., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P., Bucher, P., Copley, R.R., Courcelle, E., Das, U., Durbin, R., Falquet, L., Fleischmann, W., Griffiths-Jones, S., Haft, D., Harte, N., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lopez, R., Letunic, I., Lonsdale, D., Silventoinen, V., Orchard, S.E., Pagni, M., Peyruc, D., Ponting, C.P., Selengut, J.D., Servant, F., Sigrist, C.J., Vaughan, R. & Zdobnov, E.M. (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res. 31*(1), 315-318.

Pearson, W.R., & Lipman, D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA 85*(8), 2444-2448.

Perl DBI (2003) Homepage of Perl DBI. Available from http://dbi.perl.org/

Protein Information Resource (2003) A Proposal for the PIRSF Classification System. Retrieved January 20, 2004 from the PIR web site: http://pir.georgetown.edu/pirsf/PIRSF.pdf

Protein Information Resource (2004a) Homepage of Protein Information Resource. Available from http://pir.georgetown.edu

Protein Information Resource (2004b) Example for PIR text search. Retrieved January 20, 2004 from the PIR web site: http://pir.georgetown.edu/cgi-bin/searchExpl

Thompson, J.D., Higgins, D.G., & Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*. *22*(22), 4673-4680.

UniProt (2004) Homepage of UniProt. Available from http://www.pir.uniprot.org

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. (2002) Integrating interactome, phenome, and transcriptome mapping data for the C. elegans germline. *Curr. Biol*. *12*(22), 1952-1958.

Wu, C.H., Xiao, C., Hou, Z., Huang, H., & Barker, W.C. (2001) iProClass: an integrated, comprehensive and annotated protein classification database. *Nucleic Acids Res*. *29*(1), 52-54.

Wu, C.H., Yeh, L.-S., Huang, H., Arminski, L., Castro-Alvear, J., Chen, Y., Hu, Z., Kourtesis, P., Ledley, R.S., Suzek, B.E., Vinayaka, C.R., Zhang, J., & Barker WC. (2003a) The Protein Information Resource. *Nucleic Acids Res*. *31*(1), 345-347.

Wu, C.H., Huang, H., Yeh, L.-S., & Barker, W.C. (2003b) Protein family classification and functional annotation. *Comput. Biol. Chem*. *27*(1), 37-47.

Wu, C.H., Nikolskaya, A., Huang, H., Yeh, L.-S., Natale, D., Vinayaka, C.R., Hu, Z., Mazumder, R., Kumar, S., Kourtesis, P., Ledley, R.S., Suzek, B.E., Arminski, L., Chen, Y., Zhang, J., Cardenas, J.L., Chung, S., Castro-Alvear, J., Dinkov, G., & Barker WC. (2004a) PIRSF family classification system at the Protein Information Resource. *Nucleic Acids Res*. *32* (Database issue), D112-D114.

Wu, C.H., Huang, H., Nikolskaya, A., Hu, Z., & Barker, W.C. (2004b) The iProClass integrated database for protein functional analysis. *Comput. Biol. Chem*. *28*(1), 87-96.