*D112–D114* *Nucleic Acids Research, 2004, Vol. 32, Database issue*
*DOI: 10.1093/nar/gkh097*

# PIRSF: family classification system at the Protein Information Resource

**Cathy H. Wu*, Anastasia Nikolskaya, Hongzhan Huang, Lai-Su L. Yeh[1], Darren A. Natale, C. R. Vinayaka[1], Zhang-Zhi Hu[1], Raja Mazumder, Sandeep Kumar, Panagiotis Kourtesis[1], Robert S. Ledley[1], Baris E. Suzek, Leslie Arminski[1], Yongxing Chen[1], Jian Zhang[1], Jorge Louie Cardenas[1], Sehee Chung, Jorge Castro-Alvear[1], Georgi Dinkov[1] and Winona C. Barker[1]**

Department of Biochemistry and Molecular Biology, and [1]National Biomedical Research Foundation, Georgetown University Medical Center, 3900 Reservoir Road, NW, Box 571414, Washington, DC 20057-1414, USA

Received August 25, 2003; Accepted October 8, 2003

## ABSTRACT

**The Protein Information Resource (PIR) is an integrated public resource of protein informatics. To facilitate the sensible propagation and standardization of protein annotation and the systematic detection of annotation errors, PIR has extended its superfamily concept and developed the SuperFamily (PIRSF) classification system. Based on the evolutionary relationships of whole proteins, this classification system allows annotation of both specific biological and generic biochemical functions. The system adopts a network structure for protein classification from superfamily to subfamily levels. Protein family members are homologous (sharing common ancestry) and homeomorphic (sharing full-length sequence similarity with common domain architecture). The PIRSF database consists of two data sets, preliminary clusters and curated families. The curated families include family name, protein membership, parent–child relationship, domain architecture, and optional description and bibliography. PIRSF is accessible from the website at http://pir.georgetown.edu/pirsf/ for report retrieval and sequence classification. The report presents family annotation, membership statistics, cross-references to other databases, graphical display of domain architecture, and links to multiple sequence alignments and phylogenetic trees for curated families. PIRSF can be utilized to analyze phylogenetic profiles, to reveal functional convergence and divergence, and to identify interesting relationships between homeomorphic families, domains and structural classes.**

## INTRODUCTION

The Protein Information Resource (PIR) is an integrated public bioinformatics resource that supports genomic and proteomic research and scientific studies. For over three decades, PIR has provided many protein databases and analysis tools freely accessible to the scientific community, including the PIR-International Protein Sequence Database (PSD) of functionally annotated protein sequences, which grew out of the *Atlas of Protein Sequence and Structure* (1) edited by Margaret Dayhoff. PIR has recently joined forces with the European Bioinformatics Institute (EBI) and the Swiss Institute of Bioinformatics (SIB) to establish UniProt (the Universal Protein Knowledgebase) (2), the central resource of protein sequence and function, by unifying the database activities of PIR-PSD, Swiss-Prot and TrEMBL. In addition, we have implemented the new PIRSF (SuperFamily) classification system, which is described below. We have also enhanced iProClass (3), an integrated database of protein family, function and structure information with executive summaries and cross-references to over 50 molecular databases; maintained PIR-NREF (4), a non-redundant reference database; and improved the PIR website for scientific inquiry and system dissemination.

## PIRSF SYSTEM DEFINITION

The PIR superfamily/family concept (5), the original classification based on sequence similarity, has been used as a guiding principle to provide comprehensive and non-overlapping clustering of PIR protein sequences into a hierarchical order to reflect their evolutionary relationships (6). To facilitate the sensible propagation and standardization of protein annotation and the systematic detection of annotation errors as part of the UniProt project, PIR has extended its hierarchical superfamily concept and developed the PIRSF system, a 'network classification system based on the evolutionary relationships of whole proteins'. Classification based on whole proteins, rather than on the component domains,

*To whom correspondence should be addressed. Tel: +1 202 687 2121; Fax: +1 202 687 1662; Email: pirmail@georgetown.edu

 

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 primary level for curation is the homeomorphic family, which consists of proteins that are both homologous (evolved from a common ancestor as inferred by detectable sequence similarity) and homeomorphic (sharing full-length sequence similarity and a common domain architecture). Common domain architecture is indicated by the same type, number and order of core domains. Variation may exist for repeating domains and/or auxiliary domains, which are often mobile and may be easily lost, acquired or functionally replaced during evolution. Above the 'homeomorphic family' nodes in the network structure are parent superfamily nodes that connect distantly related families and orphan proteins based on common domains. They may be homeomorphic superfamilies, but are more likely to be domain superfamilies if the common domain regions do not extend over the entire full-length proteins. Below the homeomorphic family nodes are child subfamily nodes, which are homologous and homeomorphic clusters representing functional specialization and/or domain architecture variation within a family. The PIRSF system definition and working principles are detailed in the document, *A Proposal for the PIRSF Classification System*, available from the PIR website.

## PIRSF DATABASE CREATION AND CURATION

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 (% sequence identity, sequence length ratio and overlap length ratio) and cluster-based parameters (% matched members, distance to neighboring clusters and overall domain arrangement).

Systematic family curation is being conducted in a two-tier process to improve the quality of automated classification. Over 4500 families containing two or more members have been curated at the 'first-tier' for membership and domain architecture characteristic of the family. PIRSF has two membership types: regular members for proteins sharing end-to-end sequence similarity and associate members for proteins whose lengths deviate from the family length range, including incomplete sequences, alternate splice and initiator variants, and peptides derived from proteolytic processing. A subset of representative regular members is chosen as seed members for generating multiple sequence alignments, phylogenetic trees and Hidden Markov Models (HMMs) of the respective families. 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 (7). The incorporation of PIRSF families into InterPro and the implementation of a system to check the validity and integrity of existing families create additional means of ensuring accuracy and consistency in UniProt classification and annotation.

## PIRSF SYSTEM ACCESS

The PIRSF database is accessible from the PIR website at http://pir.georgetown.edu/pirsf/ for report retrieval and online sequence search and classification. The family report can be retrieved directly based on the PIRSF unique identifier (see http://pir.georgetown.edu/cgi-bin/pirsf?id=PIRSF001500 as an example). The reports present membership information with length, taxonomy and keyword statistics, members listed according to major kingdoms, family relationships at the whole protein, domain and motif levels with direct mapping to other classification schemes such as InterPro and SCOP (8), structure and function cross-references, and graphical display of domain and motif architecture. For curated PIRSF families, the report also includes additional family annotation and links to dynamically generated multiple sequence alignments and phylogenetic trees. The value-added protein data and cross-references in the PIRSF report are derived from the iProClass database (4). In addition to direct report retrieval, PIRSF is searchable by text strings. The text searches return PIRSF entries listed in summary lines with information on family name, membership summary, length range, domain and motif, with hypertext links to full reports. More than 20 PIRSF fields are searchable, including database unique identifiers (e.g. Pfam ID, EC number and PDB ID) and annotations (e.g. family name, keywords and length).

Protein sequence search and family classification is supported by HMM (9) and BLAST-based (10) methods. The HMM-based searches classify query sequences into curated PIRSF families based on a combination of full-length and domain HMM matches and length constraint. The algorithm, PIRSF-Scan, has been incorporated into the InterProScan program. The BLAST search of a query sequence against all UniProt sequences in the PIRSF database returns a list of best-matched families (preliminary clusters and curated families) and all protein sequences above a given threshold. The HMM and BLAST search summaries are hypertext linked to full PIRSF reports as well as to detailed search results.

## PIRSF SYSTEM APPLICATIONS

The PIRSF system provides a systematic approach for standardized and rich protein annotation (11), especially for position-specific features, protein names and keywords. Position-specific feature rules for annotating and propagating functional sites, active sites and binding sites are being developed based on manually curated multiple sequence alignments and HMMs of homeomorphic families and subfamilies, starting with those that contain at least one known 3D structure with experimentally verified site information. PIRSF also helps to detect and correct genome annotation errors, many of which have been propagated throughout molecular databases. Since many proteins are multifunctional, the assignment of a single function, which is still common in genome projects, results in incomplete or incorrect information. Numerous annotation errors have

resulted from identifications based only on local domain similarities and subsequently propagated based on transitivity (11). PIRSF classification, which considers both full-length similarity and domain architecture, discriminates between single- and multi-domain proteins where functional differences are associated with the presence or absence of one or more domains.

The data integration in PIRSF allows the identification of interesting relationships between different classification schemes. For example, Pfam-based searches can identify all PIRSFs sharing one or more Pfam domains (12). Likewise, CATH (13) or SCOP-based searches can identify PIRSFs in common CATH homology levels or SCOP superfamily levels. In combination with the underlying taxonomic information, one can retrieve PIRSFs that occur only in given lineages or share common phyletic/phylogenetic profiles. Functional convergence (non-orthologous gene displacement) and functional divergence can be revealed by the many-to-one and one-to-many relationships between the enzyme classification (EC number) and PIRSF classification. Knowledge of such relationships is fundamental to the understanding of protein evolution, structure and function, and crucial to functional genomic and proteomic research.

## ACKNOWLEDGEMENTS

The 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.

## REFERENCES

1. Dayhoff,M.O. (1965–1978) *Atlas of Protein Sequence and Structure*. 5 vols, 3 Supplements. National Biomedical Research Foundation, Washington, DC.
2. Apweiler,R., Bairoch,A., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.*, **32**, D115–D119.
3. Wu,C.H., Yeh,L.-S. L., Huang,H., Arminski,L., Castro-Alvear,J., Chen,Y., Hu,Z., Kourtesis,P., Ledley,R.S., Suzek,B.E. *et al.* (2003) The Protein Information Resource. *Nucleic Acids Res.*, **31**, 345–347.
4. Huang,H., Barker,W.C., Chen,Y. and Wu,C.H. (2003) iProClass: an integrated database of protein family classification, function and structure information. *Nucleic Acids Res.*, **31**, 390–392.
5. Dayhoff,M.O. (1976) The origin and evolution of protein superfamilies. *Fed. Proc.*, **35**, 2132–2138.
6. Barker,W.C., Pfeiffer,F. and George,D.G. (1996) Superfamily classification in PIR—International Protein Sequence Database. *Methods Enzymol.*, **266**, 59–71.
7. Mulder,N., Apweiler,R., Attwood,T., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al.* (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.*, **31**, 315–318.
8. Lo Conte,L., Brenner,S.E., Hubbard,T.J.P, Chothia,C. and Murzin,A.G. (2002) SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.*, **30**, 264–267.
9. Eddy,S.R., Mitchison,G. and Durbin,R. (1995) Maximum Discrimination Hidden Markov Models of sequence consensus. *J. Comp. Biol.*, **2**, 9–23.
10. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
11. Wu,C.H., Huang,H., Yeh,L.-S. and Barker,W.C. (2003) Protein family classification and functional annotation. *Comp. Biol. Chem.*, **27**, 37–47.
12. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E. (2002) The Pfam Protein Families Database. *Nucleic Acids Res.*, **30**, 276–280.
13. Pearl,F.M., Bennett,C.F., Bray,J.E., Harrison,A.P., Martin,N., Shepherd,A., Sillitoe,I., Thornton,J. and Orengo,C.A. (2003) The CATH database: an extended protein family resource for structural and functional genomics. *Nucleic Acids Res.*, **31**, 452–455.