Update on human genome completion and annotations: Protein information resource

Cathy Wu[1] *and Daniel W. Nebert*[2]*

[1]Director of PIR, Department of Biochemistry & Molecular Biology, Georgetown University Medical Center, Washington, DC, USA
[2]Department of Environmental Health and Center for Environmental Genetics (CEG), University of Cincinnati Medical Center, Cincinnati, OH 45267–0056, USA
**Correspondence to*: Tel: +1 513 558 4347; Fax: +1 513 558 3562; E-mail: dan.nebert@uc.edu

Date received (in revised form): 11th January 2004

Abstract

The Protein Information Resource (PIR) recently joined the European Bioinformatics Institute (EBI) and Swiss Institute of Bioinformatics (SIB) to establish UniProt—the Universal Protein Resource—which now unifies the PIR, Swiss-Prot and TrEMBL databases. The PIRSF (SuperFamily) classification system is central to the PIR/UniProt functional annotation of proteins, by providing classifications of whole proteins into a network structure to reflect their evolutionary relationships. Data integration and associative studies of protein family, function and structure are supported by the iProClass database, which offers value-added descriptions of all UniProt proteins with highly informative links to more than 50 other databases. The PIR system allows consistent, rich and accurate protein annotation for all investigators.

Keywords: *protein web sites, protein family, functional annotation*

Introduction

The high-throughput genome projects have resulted in a rapid accumulation of genome sequences for a large number of organisms. Meanwhile, researchers have begun to tackle gene functions and other complex regulatory processes by studying organisms at the global scales for various levels of biological organisation. To fully exploit the value of the data, bioinformatics infrastructures are urgently needed to identify proteins encoded by these genomes and understand how these proteins function in making up a living cell.

As a public bioinformatics database, the Protein Information Resource (PIR) is located at the Georgetown University Medical Center (Washington, DC). PIR (http://pir.georgetown.edu) provides an advanced framework for comparative analysis and functional annotation of proteins. PIR recently joined the European Bioinformatics Institute (EBI) and Swiss Institute of Bioinformatics (SIB) to establish UniProt[1] (http://www.uniprot.org), the world's most comprehensive catalogue of information on proteins. It is a central repository of protein sequence and function—created by joining the information contained in PIR-PSD, Swiss-Prot and TrEMBL. To facilitate consistent and accurate protein annotation, PIR has extended its superfamily concept and developed the PIRSF classification system.[2] This framework is supported by the iProClass integrated database of protein family, function and structure.[3] iProClass offers value-added descriptions of all UniProt proteins—with highly informative links to more than 50 other databases of protein family, function, pathway, interaction, modification, structure, genome, ontology, literature and taxonomy (Figure 1).

PIR, then and now

For more than three decades, PIR has provided many protein databases and analysis tools, freely accessible to the scientific community, including Protein Sequence Database (PSD), the first international protein database, which grew out of the *Atlas of Protein Sequence and Structure*, edited by Margaret Dayhoff [1965–1978], a pioneer in molecular evolution research. As a core resource, the PIR environment is widely used by researchers to develop other bioinformatics infrastructures and algorithms and to enable basic and applied scientific research.

The current version (January 2004) consists of more than 1,232,000 (non-redundant PIR-PSD, SwissProt and TrEMBL) proteins—organised into more than 36,290 PIR superfamilies,

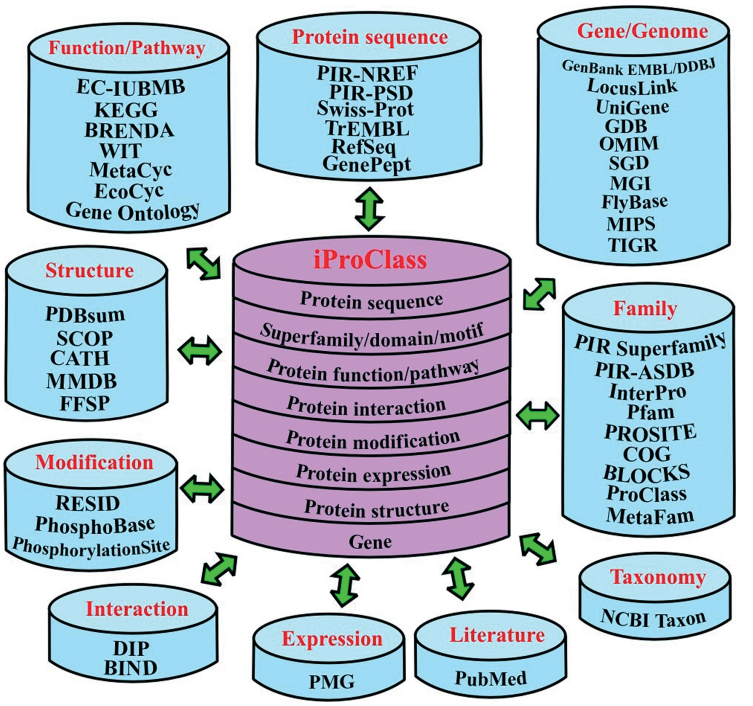

Figure 1. Diagram of the interrelated links of the iProClass database. Comprehensive protein and superfamily views exist in two types of summary reports. The protein sequence report covers information on family, structure, function, gene, genetics, disease, ontology, taxonomy, literature, with cross-references to relevant molecular databases and executive summary lines, as well as graphical display of domain and motif regions. The superfamily report provides PIR superfamily membership information with length, taxonomy and keyword statistics, complete member listing separated by major kingdoms, family relationships at the whole protein and domain and motif levels with direct mapping to other classifications, structure and function cross-references, and domain and motif graphical display.

145,340 families, 5,720 Pfam and PIR homology domains, 1,300 PROSITE/ProClass motifs, 280 RESID post-translational modification sites, 550,000 FASTA similarity clusters and links to more than 50 molecular biology databases. iProClass cross-references include databases for protein families (eg COG, InterPro), functions and pathways (eg KEGG, WIT), protein–protein interactions (eg DIP), structures and structural classifications (eg PDB, SCOP, CATH, PDBSum), genes and genomes (eg *TIGR*, *OMIM*), ontologies (eg gene ontology), literature (NCBI PubMed) and taxonomy (NCBI taxonomy).

Coupling protein classification and data integration allows associative studies of protein family, function and structure.[3] Domain-based or structural classification-based searches allow identification of protein families sharing domains or structural-fold classes. Functional convergence (unrelated proteins with the same activity) and functional divergence are revealed by the relationships between the enzyme classification and protein family classification. With the underlying taxonomic information in hand, protein families that occur in given lineages can be identified. Combining phylogenetic-pattern and biochemical-pathway information for protein families allows identification of alternative pathways to the same end product in different taxonomic groups, which may suggest potential drug targets. The systematic approach for protein-family curation, using integrative data, leads to novel predictions and functional inference for uncharacterised 'hypothetical' proteins, and to detection and correction of genome annotation errors (a few examples are listed in Table 1). Such studies may serve as a basis for further analysis of protein functional evolution and its relationship to the co-evolution of metabolic pathways, cellular networks and organisms.

Organisational levels of protein groups

PIR has three organisation levels of protein groups—namely, protein sequence entry, homeomorphic superfamily/family/subfamily and domain superfamily.

Protein sequence entries

A UniProt protein sequence entry represents the unprocessed normal product of a gene (or, sometimes, of very

Table 1. Protein family classification and integrative associative analysis for functional annotation.*

A. Functional inference of uncharacterised hypothetical proteins	
SF034452	TIM-barrel signal transduction protein
SF004961	metal-dependent hydrolase
SF005928	Nucleotidyltransferase
SF005933	ATPase with chaperone activity and inactive LON protease domain
SF005211	alpha/beta hydrolase
SF014673	lipid carrier protein
SF005019	[Ni,Fe]-hydrogenase-3-type complex, membrane protein EhaA
B. Correction, or improvement, of genome annotations	
SF025624	ligand-binding protein with an ACT domain
SF005003	inactive homologue of metal-dependent protease
SF000378	glycyl radical cofactor protein YfiD
SF000876	chemotaxis response regulator methylesterase CheB
SF000881	thioesterase, type II
SF002845	bifunctional tetrapyrrole methylase and MazG NTPase
C. Enhanced understanding of structure, function and evolutionary relationships	
SF005965	chorismate mutase, AroH class
SF001501	chorismate mutase, AroQ class, prokaryotic type

*PIRSF protein family reports detail supporting evidence for both experimentally validated and computationally predicted annotations

closely related genes) from a single species. (A number of Swiss-Prot entries still contain identical sequences from different species, which will be un-merged in future releases.) The mature protein chain and its modifications are detailed in the feature table. To the extent that is practical, UniProt aims to have one entry for each genetic locus that encodes protein. When the sequence variation is more extensive than can be conveniently represented within the entry, however, additional entries may be constructed for splice variants, allelic variants and strain variants. The source data from which entries are constructed include entries that represent a single sequence report, either published or deposited in a databank. Often, such reports will need to be 'merged' with other reports representing the same protein sequence. The UniProt staff attempts to identify these cases and perform the required merges.

Protein families

For purposes of standardising annotation, database entries are organised into families of closely related sequences. These generally represent proteins with the same function in various organisms. The taxonomic distribution within a family will depend on how well conserved are the structure and function of the protein. As a general guideline, sequences having more than 50 per cent sequence identity are usually similar in structure and function, and the major sequence features are unambiguously aligned by commonly used multiple sequence alignment programs. Therefore, 50 per cent sequence identity is used by the database staff for the provisional clustering of proteins into families. This threshold is appropriate in many cases; however, some families may be repartitioned into more convenient clusters after PIR review.

Homeomorphic superfamilies/families/subfamilies

The PIR superfamily concept,[4] 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.[5] To facilitate sensible propagation and standardisation of protein annotation and

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.

Proteins are considered 'homeomorphic' if they share full-length sequence similarity and a common domain architecture, as indicated by the same type, number and order of defined domains. Length deviation may occur for alternative-splice and alternate-initiator variants, sequence fragments and peptides derived from proteolytic processing. Variation of the domain architecture may exist for repeating domains and/or auxiliary domains, which are often mobile and may easily be lost, acquired or functionally replaced during evolution. Classification based on whole proteins, rather than on the component domains, allows annotation of both generic biochemical and specific biological functions.

The network structure accommodates a flexible number of levels that reflect varying degrees of sequence conservation (superfamily, family and subfamily). The threshold values of sequence similarity may vary at each level, depending on the evolutionary rate in each group of proteins (ie the taxonomic distribution within a protein group will depend on how well conserved are the structure and function of the protein). The network structure allows improved protein annotation, more accurate extraction of conserved functional residues and classification of distantly related orphan proteins. Homeomorphic families and subfamilies—generally representing proteins with the same function in various organisms—are suitable for propagating standardised protein names, position-specific features (such as functional sites) and keywords. Distantly related homeomorphic families and orphan proteins, sharing a common domain architecture, may form a homeomorphic superfamily. It is assumed that, although in most cases this has not been investigated in detail, the molecules in a homeomorphic superfamily share a common evolutionary history because of the acquisition of their constituent domains. Thus, it should be valid to construct an evolutionary tree from the members of a homeomorphic superfamily. If two groups of proteins with the same architecture are shown to have come to that structure independently (convergent evolution), they are appropriately separated into two homeomorphic superfamilies.

Domain superfamilies

Many types of domains have been found in diverse proteins. In common use, for example, the term 'protein kinase superfamily' refers to the collection of all proteins that contain a protein kinase-like domain. PIR calls such a group a 'domain superfamily'. Any given protein sequence will be assigned to only one homeomorphic superfamily, but it may contain sequence segments belonging to several domain superfamilies.[5]

Recent directions for additional protein analyses and databases

With the new surge in interest in the fields of subcellular and intracellular signal transduction circuitry and 'systems biology',[6] confirmed protein–protein interactions are being registered at the Human Protein Reference Database (HPRD; http://www.hprd.org).[7] Another bioinformatics database being developed is the Secreted Protein Discovery Initiative (SPDI), which has begun to identify novel and transmembrane proteins.[8] A Bayesian-networks approach for predicting protein–protein interactions, genome-wide, in yeast[9] is available at: http://genecensus.org/intint. A protein-interaction map for *Drosophila melanogaster* has very recently been developed,[10] as a starting point of a systems-biology modelling for multicellular organisms, including humans.

OrthoMCL provides a scalable method for constructing orthologous groups across multiple eukaryotic taxa, using a Markov Cluster algorithm when applied to two genomes, bu can be extended to cluster-orthologue analysis across multiple species (http://www.cbil.upenn.edu/gene-family). Analysis of clusters incorporating *Plasmodium falciparum* genes, for example, identifies numerous enzymes that were incompletely annotated in first-pass annotation of that parasite genome.[11] Finally, the evolutionary divergence of large enzyme protein families, based on complexities of their substrates, can be compared by a profile Hidden Markov Model method; the method was recently used to classify 47 glycosyltransferase families in the CAZy database into four superfamilies.[12]

Acknowledgments

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 (C.W.). The writing of this article was funded, in part, by NIH grant P30 ES06096 (D.W.N.). We very much appreciate the graphics assistance by Dr Marian Miller.

References

1. Apweiler, R., Bairoch, A., Wu, C.H., *et al.* (2004), 'UniProt: Universal protein knowledgebase', *Nucleic Acids Res.* Vol. 32, D115–D119.
2. Wu, C.H., Nikolskaya, A., Huang, H., *et al.* (2004), 'PIRSF family classification system at the Protein Information Resource', *Nucleic Acids Res.* Vol. 32, D112–D114.
3. Wu, C.H., Huang, H., Nikolskaya, A., *et al.* (2004), 'The iProClass integrated database for protein functional analysis', *Comput. Biol. Chem.* Vol. 28, in press.
4. Dayhoff, M.O. (1976), 'The origin and evolution of protein superfamilies', *Fed. Proc.* Vol. 35, 2132–2138.
5. Barker, W.C., Pfeiffer, F. and George, D.G. (1996), 'Superfamily classification in the PIR—International Protein Sequence Database', *Meth. Enzymol.* Vol. 266, 59–71.
6. Ehrenberg, M., Elf, J., Aurell, E., *et al.* (2003), 'Systems biology is taking off', *Genome Res.* Vol. 13, 2377–2380.

7. Peri, S., Navarro, J.D., Amanchy, R., *et al.* (2003), 'Development of human protein reference database as an initial platform for approaching systems biology in humans', *Genome Res.* Vol. 13, 2363–2371.
8. Clark, H.F., Gurney, A.L., Abaya, E., *et al.* (2003), 'The secreted protein discovery initiative (SPDI), a large-scale effort to identify novel human secreted and transmembrane proteins: A bioinformatics assessment', *Genome Res.* Vol. 13, 2265–2270.
9. Jansen, R., Yu, H., Greenbaum, D., *et al.* (2003), 'A Bayesian networks approach for predicting protein–protein interactions from genomic data', *Science* Vol. 302, 449–453.
10. Giot, L., Bader, J.S., Brouwer, C., *et al.* (2003), 'A protein interaction map of *Drosophila melanogaster*', *Science* Vol. 302, 1727–1736.
11. Li, L., Stoeckert, Jr., C.J. and Roos, D.D. (2003), 'OrthoMCL: Identification of ortholog groups for eukaryotic genomes', *Genome Res.* Vol. 13, 2178–2189.
12. Kikuchi, N., Kwon, Y.-D., Gotoh, M., *et al.* (2003), 'Comparison of glycosyltransferase families using the profile Hidden Markov model', *Biochem. Biophys. Res. Commun.* Vol. 310, 574–579.