*i*ProClass: an integrated, comprehensive and annotated protein classification database

Cathy H. Wu*, Chunlin Xiao, Zhenglin Hou, Hongzhan Huang and Winona C. Barker

Protein Information Resource, National Biomedical Research Foundation, Georgetown University Medical Center, 3900 Reservoir Road, NW Washington, DC 20007-2195, USA

Received September 5, 2000; Revised and Accepted October 27, 2000

ABSTRACT

The *i*ProClass database is an integrated resource that provides comprehensive family relationships and structural and functional features of proteins, with rich links to various databases. It is extended from ProClass, a protein family database that integrates PIR superfamilies and PROSITE motifs. The *i*ProClass currently consists of more than 200 000 non-redundant PIR and SWISS-PROT proteins organized with more than 28 000 superfamilies, 2600 domains, 1300 motifs, 280 post-translational modification sites and links to more than 30 databases of protein families, structures, functions, genes, genomes, literature and taxonomy. Protein and family summary reports provide rich annotations, including membership information with length, taxonomy and keyword statistics, full family relationships, comprehensive enzyme and PDB cross-references and graphical feature display. The database facilitates classification-driven annotation for protein sequence databases and complete genomes, and supports structural and functional genomic research. The *i*ProClass is implemented in Oracle 8i object-relational system and available for sequence search and report retrieval at http://pir.georgetown.edu/iproclass/.

INTRODUCTION

In this post-genomic era, advanced databases are essential to facilitate retrieval of relevant information from the voluminous data and to provide insight into protein structure and function. Protein family classification is now well recognized as an effective approach for large-scale genomic sequence annotation. Moreover, it provides an important mechanism for database organization and integration of protein sequence, structure and function. There has been a proliferation of protein databases and a variety of classification schemes to organize the data. Major protein family organizations include hierarchical families of proteins, such as the superfamilies (1) in the PIR-International Protein Sequence Database (PIR-PSD) (2); families of protein domains, such as those in Pfam (3); sequence motifs or conserved regions, as in PROSITE (4); and structural classes, as in SCOP (5). InterPro (http://www.ebi.ac.uk/interpro/) has taken a further step, integrating PROSITE, PRINTS (6), Pfam and ProDom (7) protein signature databases. MetaFam (8) is more comprehensive, assembling about 10 public domain classification databases into a 'superset' using set theory and providing a distinctive graphical interface. Still, none of these databases integrates sequence and family annotations with structure and function classifications, which would be valuable not only for data mining and information retrieval, but also as an integral part of family identification algorithms. Moreover, none of these currently include the PIR superfamily and MIPS family classifications (1), which are unique in being based on end-to-end sequence comparisons and include over 182 000 sequences.

The *i*ProClass is an integrated classification database devised as a central resource of annotated protein family information with comprehensive family relationships and structural and functional features of proteins. Its design is extended from ProClass (9,10), the first integrated protein family database, which organizes proteins based on PIR superfamilies and PROSITE motifs. The objectives of the *i*ProClass database are to support knowledge discovery by easy retrieval of family information, database management by full-scale family assignment and complete database organization, and genomic sequence annotation by classification-driven annotation of protein sequences.

*i*PROCLASS OVERVIEW AND CURRENT CONTENTS

The vast information in *i*ProClass is organized into multiple data sets (Fig. 1). The ClassSeQuence (CSQ) component describes protein sequence entries, the ClassSuperFamily (CSF), ClassDoMain (CDM) and ClassMoTif (CMT) define family relationships at the superfamily, domain and motif levels, and the ClassFuNction (CFN) and ClassSTructure (CST) describes protein functional (activity/enzyme) and structural properties and relationships.

The database has three major features: integration, comprehensiveness and annotation. It integrates protein sequence families with functional and structural classes. It also integrates family classification at the whole protein, domain and motif/site levels, supported by PIR/MIPS superfamilies/families, PIR homology domains and Pfam domains, ProClass/PROSITE motifs and PIR binding, active and modification sites. Since the introduction of the concept three decades ago, the PIR superfamily classification is still the only inclusive

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

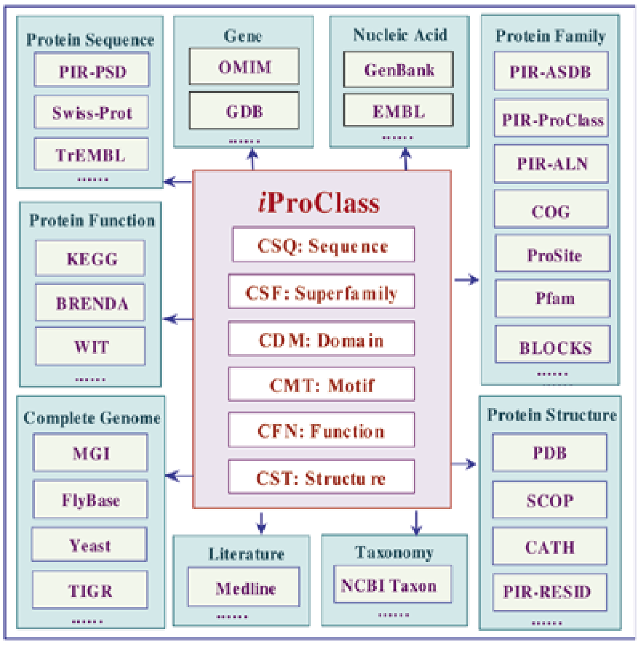

Figure 1. *i*ProClass database overview.

scheme that provides a unique hierarchical ordering of proteins to reflect their evolutionary origins and relationships.

The *i*ProClass is comprehensive, containing data derived from and links to several PIR databases, including PIR-PSD, ProClass, PIR-ALN alignment database (11), RESID database of post-translational modifications (12) and PIR-ASDB of precompiled FASTA similarity results, as well as numerous external databases. The current version (β release, September 30, 2000) is based on PIR-PSD release 66.00 (09/00), ProClass 6.0 (08/00), PIR-ALN 25.03 (08/00), RESID 22.01 (07/00), SWISS-PROT 39.0 (05/00) and TrEMBL 14.0 (06/00) (13), Pfam 5.4 (06/00), BLOCKS 12.0 (06/00) (14), PRINTS 27.0 (04/00) (6), PROSITE 14.0 (07/99), PDB (07/00) (15) and COG (01/00) (16).

The database presently consists of more than 200 000 non-redundant sequences, derived from PIR and SWISS-PROT, and more than 29 000 PIR superfamilies, correlated with 100 000 MIPS families, 380 PIR homology domains, 2200 Pfam domains, 1300 ProClass/PROSITE motifs, 280 post-translational modification sites. Also included are 3400 PDB identity links (100% identity), 40 000 PDB similarity links (30–99% sequence identity) and 30 000 enzyme (EC) links. Complete PIR/MIPS superfamily/family, domain and motif alignments are provided in MIPS-ProtFam, PIR-ALN and ProClass, respectively. The *i*ProClass provides cross-references and links to more than 30 databases of protein sequence (PSD, SWISS-PROT, TrEMBL), family and alignment (PIR-ASDB, ProClass, Pfam, PROSITE, PRINTS, BLOCKS, COG, MetaFam, ProtFam, PIR-ALN), protein enzyme/pathway (KEGG, BRENDA, WIT, EcoCyc), protein structure and structural class (PDB, SCOP, CATH, RESID), gene and genome (GenBank, EMBL, DDBJ, TIGR, UWGP, SGD, Flybase, MGI, GDB, OMIM), literature (MEDLINE) and taxonomy (NCBI Taxonomy).

Annotated summary reports and lists have been compiled for all *i*ProClass protein sequence entries and PIR superfamilies. Homology domain, ProClass motif, protein function and structure reports will be included in future releases, as will sequences unique to TrEMBL. Each sequence summary report has sections on general information, database cross-references, family assignments/relationships, and functional and structural information. Family reports contain additional summaries on length, taxonomy and keyword statistics, and a membership section that lists all sequence entries separated by major kingdoms, and denotes those from model organisms or with validated experimental status. Each protein sequence report also includes a graphical feature display that delineates regions of domains, motifs, sites and known structure chains, whichever is applicable.

Table 1. *i*ProClass database search options with examples

Identifier	Example	Demo/example link[a]
A. Report retrieval with unique identifiers		~/RPTdemo.html
Protein entry ID (PIR or SWISS-PROT ID)	A31997	~/RPTPex.html
Superfamily ID	SF000130	~/RPTFex.html
B. Protein list retrieval with text search		~/LSTPdemo.html
Enzyme classification (EC number)	2.7.1.25	~/LSTPex1.html
Protein title and species	‘bifunctional’ and ‘*Escherichia coli*’	~/LSTPex2.html
C. Family list retrieval with text search		~/LSTFdemo.html
Superfamily name	zinc finger	~/LSTFex1.html
Keyword or Pfam ID	‘GMP biosynthesis’ or ‘PF00478’	~/LSTFex2.html
D. List retrieval with sequence search		~/BLASTdemo.html
Sequence identifier (PIR or SWISS-PROT ID)	>F71428	~/BLASTex.html
Query sequence (FASTA format)	MDQTVSENLIQVK …	~/BLASTex.html

[a] *i*ProClass search home: http://pir.georgetown.edu/iproclass.

DATABASE ACCESS AND USAGE

The *i*ProClass is implemented in the Oracle 8i object-relational database management system to support database query and management. It is freely accessible from our web site at http://pir.georgetown.edu/iproclass and searchable using different modes (Table 1). Direct report retrieval is based on report unique identifiers such as PIR protein ID or superfamily number. Matching lists of summary reports are retrievable by sequence search or text search. Sequence search, based on BLAST search (17) of user-supplied query sequence against all *i*ProClass protein sequences, returns lists of best-matched families and all sequences above a given threshold. Text search provides list retrieval by using combinations of text string (and substring) searches, including protein title, superfamily or domain name, EC number, keyword, species and other database unique identifiers (such as GenBank accession and protein ID, MEDLINE ID, PDB, SWISS-PROT, TrEMBL, Pfam, PROSITE and COG).

The recognition of sequence similarity at the whole protein (superfamily) level, together with a full view of family (superfamily–domain–motif) relationships, is essential for accurate genomic sequence annotation. The integration in *i*ProClass allows it to present relationships not available in individual databases alone and to contain more comprehensive information than any other single information resource. The database supports both sequence-based and annotation-based searches. Comparative studies between or among the various family relationships will be facilitated. Knowledge of these relationships is crucial to our understanding of protein evolution, structure and function, and important for functional and structural genomic research.

ACKNOWLEDGEMENTS

This study is supported in part by grant DBI-9974855 from National Science Foundation, and grant P41 LM05798 from National Library of Medicine, NIH.

REFERENCES

1. Barker,W.C., Pfeiffer,F. and George,D. (1996) Superfamily classification in PIR-international protein sequence database. *Methods Enzymol.*, **266**, 59–71.
2. Barker,W.C., Garavelli,J.S., Hou,Z., Huang,H., Ledley,R.S., McGarvey,P.B., Mewes,H.-W., Orcutt,B.C., Pfeiffer,F., Tsugita,A, Vinayaka,C.R., Xiao,C., Yeh,L.-S.L and Wu,C. (2001) Protein Information Resource: a community resource for expert annotation of protein data. *Nucleic Acids Res.*, **29**, 29–32.
3. Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Howe,K.L. and Sonnhammer,E.L.L. (2000) The Pfam protein families database. *Nucleic Acids Res.*, **28**, 263–266.
4. Hofmann,K., Bucher,P., Falquet,L. and Bairoch,A. (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res.*, **27**, 215–219.
5. Lo Conte,L., Ailey,B., Hubbard,T.J.P., Brenner,S.E., Murzin,A.G. and Chothia,C. (2000) SCOP: a structural classification of proteins database. *Nucleic Acids Res.*, **27**, 254–256.
6. Attwood,T.K., Croning,M.D.R., Flower,D.R., Lewis,A.P., Mabey,J.E., Scordisw,P., Selley,J.N. and Wright,W. (2000) PRINTS-S: the database formerly known as PRINTS. *Nucleic Acids Res.*, **28**, 225–227.
7. Copet,F., Servant,F., Gouzy,J. and Kahn,D. (2000) ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res.*, **28**, 267–269.
8. Silverstein,K.A.T., Shoop,E., Johnson,J.E., Kilian,A., Freeman,J.L., Kunau,T.M., Awad,I.A., Mayer,M. and Retzel,E.F. (2001) The MetaFam Server: a comprehensive protein family resource. *Nucleic Acids Res.*, **29**, 49–51.
9. Wu,C., Zhao,S. and Chen,H.L. (1996) A protein class database organized with ProSite protein groups and PIR superfamilies. *J. Comp. Biol.*, **3**, 547–562.
10. Huang,H., Xiao,C. and Wu,C.H. (2000) ProClass protein family database. *Nucleic Acids Res.*, **28**, 273–276.
11. Srinivasarao,G.Y., Yeh,L.-S., Marzec,C.R., Orcutt,B.C. and Barker,W.C. (1999) PIR- ALN: A database of protein sequence alignments, *Bioinformatics*, **15**, 382–390.
12. Garavelli,J.S. (2000) The RESID database of protein structure modifications: 2000 update. *Nucleic Acids Res.*, **28**, 209–211. Updated article in this issue: *Nucleic Acids Res.* (2001), **29**, 199–201.
13. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
14. Henikoff,J.G., Greene,E.A., Pietrokovski,S. and Henikoff,S. (2000) Increased coverage of protein families with the Blocks database servers. *Nucleic Acids Res.*, **28**, 228–230.
15. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242. Updated article in this issue: *Nucleic Acids Res.* (2001), **29**, 214–218.
16. Tatusov,R.L., Galperin,M.Y., Natale,D.A. and Koonin,E.V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.*, **28**, 33–36. Updated article in this issue: *Nucleic Acids Res.* (2001), **29**, 22–28.
17. 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.