 Nucleic Acids Research, 2000, Vol. 28, No. 1 273–276

ProClass protein family database

Hongzhan Huang, Chunlin Xiao and Cathy H. Wu*

Protein Information Resource, National Biomedical Research Foundation, 3900 Reservoir Road, NW, Washington, DC 20007, USA

Received October 1, 1999; Accepted October 7, 1999

ABSTRACT

ProClass is a protein family database that organizes non-redundant sequence entries into families defined collectively by PIR superfamilies and PROSITE patterns. By combining global similarities and functional motifs into a single classification scheme, ProClass helps to reveal domain and family relationships and classify multi-domain proteins. The database currently consists of >155 000 sequence entries retrieved from both PIR-International and SWISS-PROT databases. Approximately 92 000 or 60% of the ProClass entries are classified into ~6000 families, including a large number of new members detected by our GeneFIND family identification system. The ProClass motif collection contains ~72 000 motif sequences and >1300 multiple alignments for all PROSITE patterns, including >21 000 matches not listed in PROSITE and mostly detected from unique PIR sequences. To maximize family information retrieval, the database provides links to various protein family, domain, alignment and structural class databases. With its high classification rate and comprehensive family relationships, ProClass can be used to support full-scale genomic annotation. The database, now being implemented in an object-relational database management system, is available for online sequence search and record retrieval from our WWW server at http://pir.georgetown.edu/gfserver/proclass.html

INTRODUCTION

Molecular sequence data continue to grow at an accelerating pace due to the Human Genome Project and other large sequencing projects. Advanced databases are needed to facilitate the 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 a basic approach for large-scale genomic sequence annotation and database organization. Family classification requires two crucial elements: search tools that fully utilize the information embedded within families of homologous sequences, and databases organized according to family relationships. We have developed an integrated database and search system, consisting of the ProClass protein family database (1) and the GeneFIND (Gene Family Identification Network Design) search program (2).

ProClass is a secondary, value-added database that organizes non-redundant PIR-International (3) and SWISS-PROT (4) protein sequences according to family relationships defined collectively by PIR superfamilies (5) and PROSITE patterns (6). By combining global and motif sequence similarities into a single classification scheme, ProClass reveals domain and family relationships. To maximize family information retrieval, ProClass provides hypertext links to major family databases, including family and superfamily alignments [PIR-ALN (7) and ProtFam (8)], motif and domain databases [BLOCKS (9), PRINTS (10) and Pfam (11)], as well as structural class databases [SCOP (12), CATH (13) and HSSP (14)].

CONTENT OF CURRENT RELEASE

ProClass is updated in accordance with the releases of its underlying databases. The current release 5.0 (September 1999) is derived based on PIR release 61.05 (August 1999), SWISS-PROT release 38.0 (July 1999) and PROSITE release 16.0 (July 1999). It consists of 155 868 non-redundant sequence entries from PIR and SWISS-PROT databases, excluding unclassified sequence fragments or peptides of less than 10 amino acids (Table 1).

ProClass has three data subsets, ProClass_Family (PCFam) to define protein families, ProClass_Sequence (PCSeq) to describe sequence entries, and ProClass_Motif (PCMotif) to provide an up-to-date and comprehensive collection of motif sequences and alignments for all PROSITE patterns. The ProClass families are grouped into three categories: (i) PCFA for families defined by PROSITE patterns with or without PIR superfamilies; (ii) PCFB for families defined by PIR superfamilies without PROSITE patterns; and (iii) PCFC for the collection of unclassified entries (Table 1). A ProClass family has multiple subfamilies, each of which is defined by a unique PROSITE pattern and PIR superfamily combination.

Family members not classified by both PROSITE and PIR have been identified using our GeneFIND family identification system. Newly identified PIR superfamily members are incorporated into PIR releases and reflected in subsequent ProClass updates. New PROSITE members are directly shown in ProClass_Motif. As a result of the family cross-reference and GeneFIND search, ProClass shows a high classification rate of close to 60%. The number of classified entries is increased from ~41 000 in SWISS-PROT and ~74 000 in PIR to ~92 000 sequences in ProClass (Table 1).

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

Table 1. Summary of the ProClass database (Release 5.0, September 1999)

Total number of entries in ProClass = 155 868
PIR & SWISS-PROT redundant = 59 468; PIR unique = 75 984; SWISS-PROT unique = 20 416
PIR entries classified with superfamilies = 73 975 (53%)
SWISS-PROT entries classified with PROSITE patterns = 41 576 (51%)
ProClass classified entries = 88 053 (57%) + 3732 (2%) = 91 785 (59%)
ProClass classified entries in PCFA families = 58 925 (38%)
ProClass classified entries in PCFB families = 29 128 (19%)
ProClass PCFC entries classified in PCMotif = 3732 (2%)
Unclassified ProClass PCFC entries = 67 815 (43%) – 3732 (2%) = 64 083 (41%)
Number of PCFA families/subfamilies (PROSITE patterns with/without superfamilies) = 1021/6684
Number of PCFB families/subfamilies (PIR superfamilies without PROSITE patterns) = 5150/5190
Number of PROSITE sequence entries/patterns = 41 576/50 472
Number of ProClass PCMotif sequence entries/patterns = 53 855/71 960
Number of PCMotif 'T' (true positive) patterns = 64 615
Number of PCMotif 'N' (false negative with mismatches) patterns = 7345
Number of PCMotif 'PCT' patterns not listed in PROSITE = 16 298
Number of PCMotif 'PCN' patterns not listed in PROSITE = 5190

Table 2. ProClass database files

Data file	Description	Format link[a]	FTP link[b]
ProClass_Sequence (PCSeq)			
PCSeq.dat	Detail sequence report	#pcs.tb	~/PCSeq.dat.Z
PCSeq.tb	Summary sequence information	#pcs.tb	–
PCSeq.fasta	Sequence in FASTA format	–	~/PCSeq.fasta.Z
ProClass_Family (PCFam)			
PCFam.dat	Detail family/subfamily report with membership	#pcf.dat	~/PCFam.dat.Z
PCFam.tb	Summary family/subfamily information	#pcf.tb	–
ProClass_Motif (PCMotif)			
PCMotif.dat	Detail motif report with membership	#pcm.dat	~/PCMotif.dat.Z
PCMotif.aln	Motif sequence alignments in ClustalW	#pcm.aln	~/PCMotif.aln.Z
PCMotif.fasta	Motif sequence in FASTA format	#pcf.seq	~/PCMotif.fasta.Z

[a]Format home: http://pir.georgetown.edu/gfserver/pc_format.html
[b]FTP home: ftp://nbrfa.georgetown.edu/pir/databases/proclass/

The motif collection currently includes all PROSITE patterns (i.e. motifs of PCFA families), and can be regarded as a supplement to PROSITE and BLOCKS. PCMotif has more complete membership because it is keyed to the ProClass database containing additional unique PIR sequences (~76 000 in the present release), whereas PROSITE and BLOCKS are based on SWISS-PROT only. The PROSITE members identified by GeneFIND are flagged as PCT (for 'T' true positive patterns) or PCN (for 'N' false negative patterns with mismatches) entries to distinguish from the PST/PSN (for 'T' and 'N' patterns) listed in PROSITE. Together, PCMotif has ~65 000 'T' and >7000 'N' motif patterns, including >16 000 PCT and >5000 PCN patterns (Table 1).

The ProClass database has grown by >15% since the last major ProClass release (4.0, November 1998) in terms of both numbers of sequence entries and classified entries. The current release has ~26 000 more PCSeq sequence entries, ~12 000 more classified entries, and ~9700 more PCMotif patterns.

DATABASE FORMAT

The ProClass database contains the following distribution files as summarized in Table 2: PCSeq.dat (sequence data), PCSeq.tb (sequence summary), PCSeq.fasta (protein sequence), PCFam.dat (family/subfamily data), PCFam.tb (family/subfamily summary), PCMotif.dat (motif data), PCMotif.aln (motif

alignment) and PCMotif.seq (motif sequence). The formats of the individual database files are shown with multiple examples on our Web site at http://pir.georgetown.edu/gfserver/pc_format (Table 2).

The PCSeq.tb and PCSeq.dat provide condensed information and detail reports for sequence entries with fields such as unique sequence identifiers of ProClass (PCS_AC), PIR (PIR_ID) and SWISS-PROT (SP_ID), family identifiers of ProClass (PCF_AC, PCM#), PIR (SFA#, DA/SA/FA#), and PROSITE (PS_AC), as well as attributes like sequence title, length and source organism. The PCSeq.fasta provides ProClass sequences in the FASTA format (15) for use in database sequence similarity searching. The PCFam.tb and PCFam.dat provide summary information and detail report for family entries with family identifiers of ProClass (PCF_AC), PIR (SFA#, DA/SA/FA#) and PROSITE (PS_AC, PS_DOC#), and count of family members. Links to other databases are only shown in the PCSeq.dat and PCFam.dat data files.

The PCMotif.dat reports motif description (pattern regular expression, description and length range) and membership (PCS_ID in the format of SP_ID+PIR_ID) which is separately listed and counted in the four categories (PST, PCT, PSN and PCN). Each PCMotif.aln alignment record contains sequence alignments generated by using the ClustalW program (16), with annotations including PCS_ID, membership category, beginning position of the motif and the motif, as well as 'conservation flags' separately generated from the alignment of all 'T' patterns and from both 'T' and 'N' patterns. Redundant motif sequences are represented only once in the alignment. All motif sequences (including multiple occurrences in one protein sequence) are collected in PCMotif.seq in FASTA format, with the header containing membership category and beginning position of the motif.

The ProClass is being implemented in the Oracle object-relational database management system to assist database query and management. Underlying base tables and views are built conforming to the data model.

DATABASE ACCESS AND USAGE

A WWW online server (now at http://pir.georgetown.edu/gfserver) has been set up for the distribution of the ProClass database and GeneFIND family search system (17). The ProClass database can be accessed in three different modes.

Keyword search

Individual ProClass records or lists can be retrieved based on a keyword search at http://pir.georgetown.edu/gfserver/proclass.html . As shown in Table 3 with examples and hypertext links, the database can be searched using sequence identifiers of ProClass, PIR or SWISS-PROT; using family identifiers of ProClass family or motif, PIR superfamily, or PROSITE group; or using family keyword. The search results are returned to users as HTML documents. Standard relational search options will be available.

Sequence similarity search and family classification

ProClass records and family information can also be retrieved based on the result of a GeneFIND search at http://pir.georgetown.edu/gfserver/genefind.html . The program searches the user-supplied query sequence against all ProClass sequences and returns family classification results with links to ProClass. The report displays the top matching PIR superfamilies, PROSITE patterns and unclassified sequences, with summary lines of corresponding PCSeq.tb and PCFam.tb entries, as well as full-length pairwise alignments and multiple motif alignments.

Table 3. ProClass database search options

Identifier	Example & Link[a]
Text search with sequence identifiers	
PIR entry ID (PIR_ID)	CCRZ (#2)
SWISS-PROT ID (SP_ID)	CYC_ORYSA (#1)
ProClass sequence ID (PCS_AC)	PCS009869 (#3)
Text search with family identifiers or keywords	
PIR superfamily number (SFA#)	SFA00001 (#5)
PROSITE number (PS_AC)	PS00190 (#4)
ProClass family ID (PCF_AC)	PCFA00169 (#6)
ProClass motif ID (PCM_AC)	PCM00190 (#7)
Family keyword	cytochrome c (#8)
Sequence search with sequence identifiers or query sequences	
Sequence identifier	JC4383 (gf_demo.html)
Query sequence	– (genefind.html)

[a]Text search page: http://pir.georgetown.edu/gfserver/pc_demo.html and sequence search page: http://pir.georgetown.edu/gfserver/genefind.html

File transfer

The ProClass distribution files in ASCII form are available from the anonymous FTP server at ftp://nbrfa.georgetown.edu/pir/databases/proclass/ . ProClass object/relational base tables can be obtained by sending a request to wuc@nbrf.georgetown.edu

CONCLUSIONS

The major objectives of the ProClass protein family database are to maximize family information retrieval and help organize existing protein sequence databases. As a family information resource, ProClass has a comprehensive collection of families (i.e., all PIR superfamilies and PROSITE patterns) and sequences (all non-redundant PIR and SWISS-PROT sequences). Consisting of ~92 000 classified entries, it has one of the highest classification rates among all major family or domain databases, attributable to the motif-superfamily cross-reference scheme and the GeneFIND classification.

To support full-scale genomic annotation efforts, the ProClass database can be used for direct searching against gene families and for motif detection. Weekly updates in accordance with PIR releases will soon become available on the Web server. A collection of motif patterns being derived from PIR-ALN and ProtFam alignment databases will be incorporated into future ProClass updates.

ACKNOWLEDGEMENT

This study is supported in part by grant LM05524 from the National Library of Medicine.

REFERENCES

1. Wu,C.H., Zhao,S. and Chen,H.L. (1996) *J. Comp. Biol.*, **3**, 547–562.
2. Wu,C.H., Huang,H. and McLarty,J. (1999) *Int. J. Artif. Intell. Tools*, **8**, in press.
3. Barker,W.C., Garavelli,J.S., Huang,H., McGarvey,P.B., Orcutt,B.C., Srinivasarao,G.Y., Xiao,X., Yeh,L.-S.L., Ledley,R.S., Janda,J.F., Pfeiffer,F., Mewes,H.-W., Tsugita,A. and Wu,C.H. (2000) *Nucleic Acids Res.*, **28**, 41–44 (this issue).
4. Bairoch,A. and Apweiler,R. (2000) *Nucleic Acids Res.*, **28**, 45–48 (this issue).
5. Barker,W.C., Pfeiffer,F. and George,D. (1996) *Methods Enzymol.*, **266**, 59–71.
6. Hofmann,K., Bucher,P., Falquet,L. and Bairoch,A. (1999) *Nucleic Acids Res.*, **27**, 215–219.
7. Srinivasarao,G.Y., Yeh,L.-S.L., Marzec,C.R., Orcutt,B.C. and Barker,W.C. (1999) *Bioinformatics*, **15**, 382–390.
8. Mewes,H.W., Hani,J., Pfeiffer,F. and Frishman,D. (1998) *Nucleic Acids Res.*, **26**, 33–37. Updated article in this issue: *Nucleic Acids Res.* (2000), **28**, 37–40.
9. Henikoff,J.G., Henikoff,S. and Pietrokovski,S. (1999) *Nucleic Acids Res.*, **27**, 226–228. Updated article in this issue: *Nucleic Acids Res.* (2000), **28**, 228–230.
10. Attwood,T.K., Flower,D.R., Lewis,A.P., Mabey,J.E., Morgan,S.R., Scordis,P., Selley,J.N. and Wright,W. (1999) *Nucleic Acids Res.*, **27**, 220–225. Updated article in this issue: *Nucleic Acids Res.* (2000), **28**, 225–227.
11. Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Finn,R.D. and Sonnhammer,E.L.L. (1999) *Nucleic Acids Res.*, **27**, 260–262. Updated article in this issue: *Nucleic Acids Res.* (2000), **28**, 263–266.
12. Hubbard,T.J.P., Ailey,B., Brenner,S.E., Murzin,A.G. and Chothia,C. (1999) *Nucleic Acids Res.*, **27**, 254–256. Updated article in this issue: *Nucleic Acids Res.* (2000), **28**, 257–259.
13. Orengo,C.A., Pearl,F.M.G., Bray,J.E., Todd,A.E., Martin,A.C., Lo Conte,L. and Thornton,J.M. (1999) *Nucleic Acids Res.*, **27**, 275–279. Updated article in this issue: *Nucleic Acids Res.* (2000), **28**, 277–282.
14. Dodge,C., Schneider,R. and Sander,C. (1998) *Nucleic Acids Res.*, **26**, 313–315.
15. Pearson,W.R. and Lipman,D.J. (1988) *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
16. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) *Nucleic Acids Res.*, **22**, 4673–4680.
17. Wu,C.H., Shivakumar,S., Shivakumar,C.V. and Chen,S. (1998) *Bioinformatics*, **14**, 223–224.