PROCLASS PROTEIN FAMILY DATABASE: NEW VERSION WITH MOTIF ALIGNMENTS

CATHY H. WU and SAILAJA SHIVAKUMAR
Department of Epidemiology/Biomathematics
University of Texas Health Center at Tyler, P. O. Box 2003
Tyler, TX 75710, USA

ProClass is a protein family database which organizes non-redundant sequence entries into families defined collectively by the ProSite patterns and PIR superfamilies. The database consists of about 100,000 entries, more than half of which are classified in about 3,000 families. The new version includes links to various protein family/domain and structural class databases and contains gapped motif alignments for all ProSite patterns. The motif sequences are retrieved from both SwissProt and PIR-international databases, including numerous new members detected by our GeneFIND family identification system. The motif collection represents a 50% increase from those catalogued in ProSite. The ProClass database can be used to maximize family information retrieval, help organize protein sequence databases, and support full-scale genomic annotation. The database and its query program are freely available for on-line record retrieval and direct file transfer from our WWW server at http://diana.uthct.edu/proclass.html.

1 Introduction

Effective large-scale genomic annotation involves the categorization of each sequence as identical to a known gene, related to a known gene or motif, or completely novel. The volume of the data poses new challenges to these search strategies, because the search time grows linearly with the number of database sequence entries, and the number of erroneous annotation increases with the large number of unrelated sequences. An important strategy to deal with this problem is to perform database searching against gene families. Family databases not only permit family-based searches, but also provide effective means to retrieve relevant information from vast amounts of data.

There are two major protein family databases: PIR-international [George et al., 1997], organized by superfamilies [Barker et al., 1996], and ProSite [Bairoch et al., 1997] which catalogues SwissProt [Bairoch and Apweiler, 1997] entries by using motif patterns. Other family or domain databases include BLOCKS [Henikoff et al., 1997], PRINTS [Attwood et al., 1997], ProDom [Sonnhammer and Kahn, 1994], Pfam [Sonnhammer et al., 1997], PIMA [Smith and Smith, 1992] and SBASE [Fabian et al., 1997]. The BLOCKS database (release 9.3, March 1997) has 932 protein groups compiled based on ProSite 13.0 and SwissProt 34.0. The PRINTS protein fingerprints database (release 16.0, June 1997) has 750 protein family entries, 350 of which are related to ProSite patterns. The PRINTS sequence entries are retrieved from the OWL non-redundant composite protein sequence database [Bleasby et al., 1994]. The ProDom protein domain database (release 34, April 1997) provides 18,086 domain families from non-fragmentary sequences in SwissProt 34.0. The Pfam database of protein domain families (release 2.0, March 1997) has 527 Pfam-A families which are constructed using hidden Markov models and 13,289 Pfam-B families which are clustered by the Domainer program used in ProDom. The SBASE protein domain library (release 5.0, October 1996) has a collection of 79,863 annotated protein sequence segments, clustered into more than 16,000 groups. Although these databases resolve sequences into families, none are designed to handle complex family relationship including hierachies and multiple membership, and address the database organization *per se*.

To provide a mechanism for organizing protein sequences and effectively annotating new sequences, we have developed a ProClass protein family database [Wu et al., 1996]. By combining global similarities and functional motifs in a single family organization scheme, ProClass provides a unique mechanism to reveal domain and family relationships and classify multi-domained proteins. This paper describes recent developments of the database, including the design and compilation of a new motif sub-database.

2 ProClass Database Design

ProClass is a second-generation, value-added database that organizes non-redundant sequence entries according to family relationships defined collectively by PIR superfamilies and ProSite patterns. The current release (2.0, October 1997) consists of 99,853 sequence entries retrieved from PIR-international (release 53.0, June 1997) and SwissProt (release 34.0, November 1996) databases, excluding unclassified sequence fragments or peptides of less than ten amino acids (Table 1).

By combining global similarities and functional motifs into a single family organization scheme, ProClass classifies multi-domained proteins, unveils domain and family relationships, and provides enriched family information. It has three sub-databases, ProClass_Family (PCFam) to define protein families, ProClass_ Sequence (PCSeq) to describe sequence entries, and ProClass_Motif (PCMotif) to collect motif alignments. The families are grouped into three categories: (1) PCFA for families defined by ProSite patterns with or without PIR superfamilies; (2) PCFB for families defined by PIR superfamilies without ProSite patterns; and (3) PCFC for entries not classified by either ProSite or PIR (Table 1). Subfamilies describe different domain or superfamily combinations within a ProClass family, thereby resolving the classification problem of multi-domained or multi-membership proteins. For example, ProClass family PCFA00175 (defined by

ProSite PS00197 pattern) has three sub-levels, 175A to group proteins containing PS00197 only, 175B for those containing both PS00197 and PS00198, and 175C for those with both PS00197 and PS00559 (Table 2).

Table 1: Summary of the ProClass database (Release 2.0, October 1997).

Total Number of Entries in ProClass = 99,853
SwissProt-PIR Redundant = 47,302; SwissProt Unique = 11,600; PIR Unique = 40,951
SwissProt Entries Classified with ProSite Patterns = 24,147 (41%)
PIR Entries Classified with Superfamilies = 43,569 (46%)
ProClass Classified Entries = 49,228 (49%) + 5,546 (6%) = 54,774 (55%)
ProClass Classified Entries in PCFA Families = 33,447 (33%)
ProClass Classified Entries in PCFB Families = 15,781 (16%)
ProClass Unclassified PCFC Entries = 50,625 (51%) - 5,546 (6%) = 45,079 (45%)
ProClass PCFC Entries Classified in PCMotif = 5,546
Number of PCFA Families (ProSite Patterns with/without Superfamilies) = 874
Number of PCFB Families (PIR Superfamilies without ProSite Patterns) = 2,278

Table 2: ProClass families can be used to (1) reveal domain relationships, (2) place sequence entries; and (3) define new patterns or superfamilies.

ProClass Family	ProSite Pattern Number & Name	PIR Superfamily Number & Name
(1) Domain relationships between related families and among multi-domained proteins		
PCFA00175Aa	PS00197 (2Fe2S_ferredoxin)	SFA00038 (ferredoxin [2Fe-2S])
PCFA00175Ba	PS00197|PS00198	SFA00119 (fumarate reductase iron-sulfer)
PCFA00175Ca	PS00197|PS00559 (molybdopterin)	SFA00083 (xanthine dehydrogenase)
PCFA00176Aa	PS00197|PS00198	SFA00119 (fumarate reductase iron-sulfer)
PCFA00176Ba	PS00198 (4Fe4S_ferredoxin)	SFA00039 (ferredoxin 2[4Fe-4S])
PCFA00176Bb	PS00198	SFA00092 (glycerol-3-p dehydrogenase C)
PCFA00176Bf	PS00198	SFA00212 (hydrogenase Fe large chain)
(2) Placement of sequence entries classified by ProSite or PIR only		
PCFA00058Aa	PS00059 (adh_zinc)	SFA00055 (alcohol dehydrogenase)
PCFA00058A#	PS00059	-
PCFA00058#a	-	SFA00055
(3) Definition of new ProSite patterns or PIR superfamilies		
PCFA00781	PS01019 (ADP-ribosylation)	-
PCFB00050	-	SFA00050 (phycocyanin)

Domain relationships which are otherwise difficult to identify systematically are revealed by the superfamily-motif cross-reference. There may be various combinations of domain structures between related protein families. This is illustrated by PCFA00175 and PCFA00176, where ferredoxin-related PIR superfamilies (SFA00038, SFA00039, and SFA00119) contain different combinations of PS00197 and PS00198 patterns (Table 2). The family cross-reference system groups a wide range of functionally related protein families that share the same motifs. An example is the PCFA00176Ba, 176Bb, and 176Bf sub-families, in which SFA00039, SFA00092, and SFA00212 PIR superfamilies all contain the PS00198 pattern (Table 2).

The cross-reference also permits efficient and correct placement of new sequence entries. A sequence entry can be placed into ProClass families if it is classified by either ProSite patterns or PIR superfamilies. This is illustrated by PCFA00058, which has 74 entries classified by both PS00059 and SFA00055, 37 by PS00059 only, and 39 by SFA00055 only (Table 2). As a result, the number of classified entries is increased from about 24,000 and 44,000 in SwissProt and PIR, to about 50,000 in ProClass, based on the cross-reference (Table 1). The motif-superfamily reference provides the means to locate potential candidates for new superfamily or motif definition. For example, new superfamily(ies) can be defined for the 34-membered PS01019 ProSite pattern, and patterns can be derived for the 101-membered SFA00050 PIR superfamily (Table 2). In fact, there are many large PCFB families, 225 of which have ten or more members. There are also many PCFA families without corresponding PIR superfamilies, 60 of which have at least ten members.

The motif database is designed to provide an up-to-date and comprehensive collection of motif sequences. It 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 memberships because it is keyed to the ProClass database containing both SwissProt and unique PIR sequences, whereas ProSite and BLOCKS are based on SwissProt only. A large number of new motif sequences not catalogued in ProSite are identified using our GeneFIND (Gene Family Identification Network Design) system, even at a stringent threshold condition (i.e., top 3% neural network hit, P(N) score of less than E-20 in BLAST search [Altschul et al., 1990], greater than 35% sequence identity in SSEARCH alignment [Smith and Waterman, 1981], and no more than two mismatches to the ProSite motif pattern) [Wu et al., 1997a]. Correspondingly, the memberships in PCMotif are grouped in four categories: PST for "T" (true positive without mismatch) patterns listed in ProSite; PSN for "N" patterns (false negative with mismatches) listed in ProSite, PCT for ProClass "T" patterns identified by GeneFIND, and PCN for ProClass "N" patterns identified by GeneFIND. Table 3 compares the membership data of PCMotif with those of other major protein domain/motif databases.

A total of about 15,000 PCT patterns are detected, one third of which are SwissProt entries not referenced by ProSite and two thirds are unique PIR entries. Two threshold conditions are used, high (with a "PCT" flag) and low ("PCt" flag).

The high threshold values are greater than 40% sequence identity extending more than 80% of query length, whereas the low threshold values are greater than 30% identity or BLAST scores of less than E-50 (but less than 40% identity at 80% overlap). Also detected are more than 1,000 PCN patterns, mostly "N1" (false negative with a single amino acid mismatch) or "N2" (with two mismatched amino acid residues) sequences. The PCN patterns are labeled with a "PCN" or "PCn" flag for satisfying high (40% identity at 80% overlap) or low (35% identity and E-50) conditions. Overall, the PCMotif database contains 45,080 "T" and 2,017 "N" patterns in 36,544 sequence entries (p.s. each sequence entry may contain multiple motif patterns). This motif collection represents a 50% increase from the 30,486 "T" and 964 "N" patterns of 24,147 sequence entries currently catalogued in ProSite (Table 3).

A large fraction of the PCT/PCN sequences in PCMotif are PCFC entries previously unclassified in both SwissProt and PIR. The GeneFIND identification of these motif sequences further increases the number of classified entries in ProClass by 5,546 sequences (6%), as shown in Table 1.

Table 3: Comparisons of protein domain/motif databases.

Database	Family Entries	Sequence Entries
ProSite 13.0	874	24,147
Blocks 9.3	932	19,138
Prints 16.0	750	24,844
Pfam 2.0 (Pfam-A)	527	28,170
ProClass 2.0 (PCMotif)	874	36,544

3 ProClass Database Format

Table 4 shows the ProClass database format using GATA-type zinc finger proteins as an example. The PCFam and PCSeq format has been previously described [Wu et al., 1996], except the new feature of hypertext links to various molecular databases. In addition to hypertext links (shown with underlines in Tables 4 and 5) to raw records of the SwissProt, PIR and ProSite databases, ProClass family and sequence entries now also link to PIR superfamily member list and sequence alignments, other family databases, and structural class databases (Table 4a-e). The PIR superfamily member list (PIR_SF) and their multiple sequence alignments (PIR_ALN) are available directly from the PIR. The MIPS [Mewes et al., 1997] further shows family alignments within superfamilies. Other linked domain or motif databases are: BLOCKS, PRINTS, ProDom and Pfam. Linked structural class databases are SCOP [Hubbard et al., 1997] and CATH (http://www.biochem.ucl.ac.uk/bsm/cath), from which one can view the PDB tertiary structure and classification based on protein folds, as well as the HSSP database of protein structure-sequence alignments [Schneider et al., 1997].

The PCMotif has three files, a data file for motif description and membership listing (PCMotif.dat), an alignment file for the motif sequences in ClustalW [Thompson et al., 1994] format (PCMotif.aln), and a sequence file for motif sequences in FASTA [Pearson and Lipman, 1988] format. The fields in the data records (Table 4g) are: PCM_AC and PCM_ID (the motif accession number and ID corresponding to the ProSite number and ID); PS_DE and PS_PA (the ProSite pattern description and definition); PROSITE (links to ProSite records); RELEASE (the ProClass release number and date); LENGTH (Conserve = the conserve length of the motif pattern excluding Xs, Maximum and Minimum = the length range of the pattern including Xs, the maximum and minimum lengths are different when indels occur in the motif region as indicated by X(m,n) in PS_PA); COUNT (the number of entries in each membership category and their number of motif occurrences in parentheses); PST, PSN, PCT, PCN (membership listing with links to ProClass sequence entries), PCF (link to the corresponding ProClass family entry) and PCMALN (link to the corresponding motif alignment record).

Each motif alignment record (Table 4f) starts with the record ID (i.e., PCM_AC linked to the corresponding motif data record), followed by sequence alignments generated by using the ClustalW program. Each sequence entry is annotated with three fields: PCS_ID (the ProClass sequence ID which is the concatenation of SwissProt and PIR ID); the membership category; and n (the beginning position of the motif). When there are multiple occurrences of the motif pattern, the sequences are concatenated with a ".." delimiter, with corresponding beginning positions concatenated with a "-" delimiter. Also displayed with each alignment record is a "conservation flag" generated from the alignment of all "T" patterns (i.e., PST and PCT). The flag uses "*" for complete conservation at the aligned position and "." for conservative substitution.

As illustrated, the GATA_ZN_FINGER family has 35 members in four different PIR superfamilies (Table 4a-c). All have a pair of the GATA-type zinc finger motif, except the two members of superfamily SFA02255. A total of 16 new PCT sequences are identified, including 5 SwissProt entries and 11 unique PIR entries (Table 4f-g). A second example is the ALDH_PNT_1 motif family. It includes two PCT and four PCN patterns (Table 5). Note that gaps are introduced in the alignment because the motif pattern is of variable lengths (i.e, represented by x(1,3) in the ProSite pattern). The alignment of PSN and PCN patterns are shown beneath the conservation flag.

Table 4: ProClass database format: GATA-type zinc finger proteins as an example.

```
//          a. Family Data Entry in PCFam.dat
PCF_AC      PCFA00300
PCF_DE      GATA-type zinc finger domain
PROSITE     PS00344; PDOC00300; GATA_ZN_FINGER
PIR_SFA     SFA02250; transcription factor GATA-1
PIR_SFA     SFA02251; transcription factor GATA-2
PIR_SFA     SFA02252; transcription factor GATA-4
PIR_SFA     SFA02255; nitrogen regulatory protein nit-2
BLOCKS      BL00344; GATA_ZN_FINGER
PRINTS      PR00619; transcription factor gata zinc finger signature
PFAM        PF00320; GATA family of transcription factors
COUNT       35
PCM_AC      PCM00344

//          b. Subfamily Data Entry in PCFam.dat
PCF_AC      PCFA00300Aa
PROSITE     PS00344; PDOC00300; GATA_ZN_FINGER
PIR_SFA     SFA02250; transcription factor GATA-1
BLOCKS      BL00344
PRINTS      PR00619
PFAM        PF00320
COUNT       7
PCS_ID      ELT1_CAEEL+A41267; GA1A_XENLA+A41602; GA1B_XENLA+B41602;
PCS_ID      GAT1_CHICK+A32993; GAT1_HUMAN+A34888; GAT1_MOUSE+S04655;
PCS_ID      GAT1_RAT+S48756;
PCM_AC      PCM00344

//          c. Family/Subfamily Summary Entries in PCFam.tb
PCFA00300Aa: PS00344: PDOC00300: SFA02250: -: 7
PCFA00300Ab: PS00344: PDOC00300: SFA02251: -: 7
PCFA00300Ac: PS00344: PDOC00300: SFA02252: -: 8
PCFA00300Ad: PS00344: PDOC00300: SFA02255: -: 2
PCFA00300A#: PS00344: PDOC00300: -: -: 6
PCFA00300#a: -: -: SFA02251: -: 2
PCFA00300#b: -: -: SFA02252: -: 2
PCFA00300#c: -: -: SFA02255: -: 1

//            d. Sequence Data Entry in PCSeq.dat
PCS_AC        PCS012829
PCS_ID        GAT1_CHICK+A32993
SP_DE         erythroid transcription factor (GATA-1) (eryf1) (nf-e1 DNA-binding
PIR_DE        transcription factor GATA-1 - chicken
SP_ENTRY      GAT1_CHICK; P17678; 304AA
PIR_ENTRY     PIR2; A32993; A32993; 304AA
PIR_SFA       SFA02250; transcription factor GATA-1
PIR_SFB       SFB03881; ALN02296; GATA-type zinc finger homology
PIR_ALN       ALN02296
PROSITE       PS00344; PDOC00300; GATA_ZN_FINGER; T
SP_SIMILAR    TO OTHER GATA-TYPE TRANSCRIPTION FACTORS.
```

Table 4: ProClass database format: GATA-type zinc finger proteins as an example (continued).

BLOCKS	BL00344
MIPS	A32993
PRINTS	GAT1_CHICK
PFAM	GAT1_CHICK
PRODOM	GAT1_CHICK; P17678; PS00344
SCOP	1GAT; 1GAU
CATH	1GAT; 1GAU
HSSP	1GAT; 1GAU
PCF_AC	PCFA00300Aa
PCM_AC	PCM00344

// e. Sequence Summary Entry in PCSeq.tb

PCS012829: GAT1_CHICK: P17678: A32993: PCFA00300Aa: PS00344: SFA02250: ALN02296: 304|304: erythroid transcription factor (GATA-1) (eryf1) (nf-e1 DNA-binding

// f. Motif Alignment Entry in PCMotif.aln (Partially Shown)

PCM_AC PCM00344

```
AREA_EMENI+S72883 PST 673     -------------------------..CTNCFTQTTPLWRRNPEGQPLCNAC
DA80_YEAST+S22781 PST 31      -------------------------..CQNCFTVKTPLWRRDEHGTVLCNAC
ELT1_CAEEL+A41267 PST 217-272 CVNCGVHNTPLWRRDGSGNYLCNAC..CVNCRTNTTTLWRRNGEGHPVCNAC
GA1A_XENLA+A41602 PST 178-232 CVNCGATVTPLWRRDMSGHYLCNAC..CSNCHTSTTTLWRRNASGDPVCNAC
GA1B_XENLA+B41602 PST 180-234 CVNCGATVTPLWRRDLSGHYLCNAC..CSNCHTSTTTLWRRNAGGDPVCNAC
GA5A_XENLA+I51419 PST 183-237 CVNCGAMSTPLWRRDGTGHYLCNAC..CTNCHTSTTTLWRRNSEGEPVCNAC
GA5B_XENLA+I51420 PST 184-238 CVNCGAMSTPLWRRDGTGHYLCNAC..CTNCHTSTTTLWRRNSEGEPVCNAC
GAT1_CHICK+A32993 PST 110-164 CVNCGATATPLWRRDGTGHYLCNAC..CSNCQTSTTTLWRRSPMGDPVCNAC
GAT1_HUMAN+A34888 PST 204-258 CVNCGATATPLWRRDRTGHYLCNAC..CTNCQTTTTTLWRRNASGDPVCNAC
GAT1_MOUSE+S04655 PST 204-258 CVNCGATATPLWRRDRTGHYLCNAC..CTNCQTTTTTLWRRNASGDPVCNAC
GAT1_RAT+S48756   PST 204-258 CVNCGATATPLWRRDRTGHYLCNAC..CTNCQTTTTTLWRRNASGDPVCNAC
GAT1_YEAST+S56233 PST 310     -------------------------..CSNCTTSTTPLWRKDPKGLPLCNAC
GAT2_CHICK+A36389 PST 281-335 CVNCGATATPLWRRDGTGHYLCNAC..CANCQTTTTTLWRRNANGDPVCNAC
GAT2_HUMAN+A40815 PST 295-349 CVNCGATATPLWRRDGTGHYLCNAC..CANCQTTTTTLWRRNANGDPVCNAC
GAT2_XENLA+C41602 PST 267-321 CVNCGATATPLWRRDGTGHYLCNAC..CANCQTSTTTLWRRNANGDPVCNAC
GAT3_CHICK+B36389 PST 264-318 CVNCGATSTPLWRRDGTGHYLCNAC..CANCQTTTTTLWRRNANGDPVCNAC
GAT3_HUMAN+A39794 PST 263-317 CVNCGATSTPLWRRDGTGHYLCNAC..CANCQTTTTTLWRRNANGDPVCNAC
GAT3_MOUSE+B39794 PST 263-317 CVNCGATSTPLWRRDGTGHYLCNAC..CANCQTTTTTLWRRNANGDPVCNAC
GAT3_XENLA+D41602 PST 256-310 CVNCGATSTPLWRRDGTGHYLCNAC..CANCQTTTTTLWRRNANGDPVCNAC
....
GZF3_YEAST+S53377 PST 131     -------------------------..CKNCLTSTTPLWRRDEHGAMLCNAC
NIT2_NEUCR+A34755 PST 743     -------------------------..CTNCFTQTTPLWRRNPDGQPLCNAC
URB1_USTMA+S27473 PST 338-482 CSNCGVTSTPLWRRAPDGSTICNAC..CTNCQTTTTPLWRRDEDGNNICNAC
GAT4_MOUSE+       PST 216-270 CVNCGAMSTPLWRRDGTGHYLCNAC..CANCQTTTTTLWRRNAEGEPVCNAC
ELT2_CAEEL+A56953 PCt 237     -------------------------..CSNCNGTNTTLWRRNAEGDPVCNAC
GAF2_SCHPO+       PCt 172     -------------------------..CQNCATTNTPLWRRDESGNPICNAC
GATB_BOMMO+       PCt 322     -------------------------..CTNCQTTATSLWRRNVQGETVCNAC
PNR_DROME+        PCt 169-226 CVNCGAISTPLWRRDGTGHYLCNAC..CTNCGTRTTTLWRRNNDGEPVCNAC
SRP_DROME+S40382  PCt 319     -------------------------..CSNCHTTHTSLWRRNPAGEPVCNAC
+A41782           PCT 289-343 CVNCGATATPLWRRDGTGHYLCNAC..CANCQTTTTTLWRRNANGDPVCNAC
+A48099           PCT 215-269 CVNCGAMSTPLWRRDGTGHYLCNAC..CANCQTTTTTLWRRNAEGEPVCNAC
+A53741           PCt 322     -------------------------..CTNCQTTATSLWRRNVQGETVCNAC
+A57601           PCt 261-321 CVNCGATSTPLWRRDGTGHYLCNAC..CANCKTTTTTLWRRNASGEPVCNAC
+B48099           PCT 5-59    CVNCGATATPLWRRDGTGHYLCNAC..CANCQTTTTTLWRRNANGDPVCNAC
+I57561           PCT 214-268 CVNCGAMSTPLWRRDGTGHYLCNAC..CANCQTTTTTLWRRNAEGEPVCNAC
+JC6170           PCt 94-238  CSNCGTKSTPLWRRSPTGAMICNAC..CQNCGTTVTPLWRRDEQGHPICNAC
+S51493           PCT 662     -------------------------..CTNCFTQTTPLWRRNPEGQPLCNAC
+S53811           PCt 141     -------------------------..CTNCQTTATSLWRRNVQGETVCNAC
+S53812           PCt 55      -------------------------..CTNCQTTATSLWRRNVQGETVCNAC
+S70168           PCT 673     -------------------------..CTNCFTQTTPLWRRNPEGQPLCNAC
                                                         * **    * ***.   *  .****
```

Table 4: ProClass database format: GATA-type zinc finger proteins as an example (continued).

```
//              g. Motif Data Entry in PCMotif.dat (Partially Shown)
PCM_AC          PCM00344
PCM_ID          GATA_ZN_FINGER; MOTIF.
PS_DE           GATA-type zinc finger domain
PS_PA           C-x-N-C-x(4)-T-x-L-W-R-[RK]-x(3)-G-x(3)-C-N-A-C.
PROSITE         PS00344; PDOC00300
RELEASE         PROCLASS 2.0 (October 1997)
LENGTH          Conserve = 13aa; Maximum = 25aa; Minimum = 25aa;
COUNT           PST= 30 (54); PSN= 0;
COUNT           PCT= 16 (23); PCN= 0;
PST             .......
PCT             ELT2_CAEEL+A56953; GAF2_SCHPO+; ... PNR_DROME+ (2); ...
PCF             PCFA00300
PCMALN          PCM00344
```

Table 5: A ProClass motif example with gaps in the alignment: ALADH_PNT_1 Motif.

```
//          a. Motif Data Record
PCM_AC      PCM00836
PCM_ID      ALADH_PNT_1; MOTIF.
PS_DE       Alanine dehydrogenase and pyridine nucleotide transhydrogenase signature
PS_PA       G-[LIVM]-P-x-E-x(3)-N-E-x(1,3)-R-V-A-x-[ST]-P-x-[GST]-V-x(2)-L-x-[KRH]-x-G.
PROSITE     PS00836; PDOC00654
RELEASE     PROCLASS 2.0 (October 1997)
LENGTH      Conserve = 16aa; Maximum = 27aa; Minimum = 29aa;
COUNT       PST= 5; PSN= 0;
COUNT       PCT= 2; PCN= 4;
PST         DHA_BACSH+A34261; DHA_BACST+B34261; DHA_MYCTU+A43830;
PST         PNTA_ECOLI+DEECXA; NNTM_BOVIN+DEBOXM;
PCT         +G02257; +S54876;
PCN         DHA_BACSU+A49337; PNTA_HAEIN+E64119; +S74638; +S77433;
PCMALN      PCM00836

//          b.  Motif Alignment Record
PCM_AC      PCM00836
DHA_BACSH+A34261        PST    4     GIPKEIKNNE--NRVAMTPAGVVSLTHAG
DHA_BACST+B34261        PST    4     GIPKEIKNNE--NRVAITPAGVMTLVKAG
DHA_MYCTU+A43830        PST    4     GIPTETKNNEFQFRVAITPAGVAELTRRG
NNTM_BOVIN+DEBOXM       PST    60    GVPKEIFQNE--KRVALSPAGVQALVKQG
PNTA_ECOLI+DEECXA       PST    4     GIPRERLTNE--TRVAATPKTVEQLLKLG
+G02257                 PCT    60    GVPKEIFQNE--KRVALSPAGVQNLVKQG
+S54876                 PCT    60    GVPKEIFQNE--KRVALSPAGVQALVKQG
                                     *.* *   **   *** .*  *  * . *
DHA_BACSU+A49337        PCN1   4     GVPKEIKNNE--NRVALTPGGVSQLISNG
PNTA_HAEIN+E64119       PCN1   4     GVPRELLENE--SRVAATPKTVQQILKLG
+S74638                 PCn3   4     GVPKEIKDQE--FRVGLTPSSVRALLSQG
+S77433                 PCN2   23    GVPRESFDQE--CRVAMTPDTAQKLQKLG
```

4 ProClass System Distribution

A WWW on-line server has been set up for the distribution of our system [Wu et al., 1997b]. The ProClass database is accessible for family information retrieval using various search keys at http://diana.uthct.edu/proclass.html. Free copies of the ProClass database and ProQuery program can be obtained via anonymous FTP to ftp://diana.uthct.edu/pub/ProClass/. The ProQuery program can be installed on UNIX machines for ProClass database record retrieval in batch-mode or via WWW interface. The GeneFIND system is available for on-line family identification of query sequences at http://diana.uthct.edu/genefind.html.

5 Conclusion

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 ProSite patterns and PIR superfamilies) and sequences (all non-redundant SwissProt and PIR sequences). Consisting of approximately 55,000 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 our GeneFIND family identification system. The motif collection provides a useful supplement to ProSite and BLOCKS. In addition to the 50% increase in membership, the motif alignments contain gaps for variable-length motif patterns (vs. ungapped blocks), and all occurrences of motif patterns within a sequence are shown. Furthermore, to allow hypertext navigation, ProClass entries are linked to other family/domain and structural class databases in addition to the raw PIR, SwissProt and ProSite records.

The ProClass database can be used to support full-scale genomic annotation effort in several aspects. The database constitutes ideal data sets that can be used for database search against individual protein families, and the collection of motif sequences is directly searchable for motif detection using database search and alignment tools. Although only references ProSite motifs at the present, the motif database will be extended to PRINTS motifs and PIR domains (in collaboration with PIR). The ProClass database can be used to compile training sets for family-based search tools including hidden Markov models [Krogh et al., 1994; Eddy et al. 1995], profiles [Gribskov et al., 1989], and neural networks [Wu, 1996], whose search sensitivity would be improved by a more completely classified database. The enriched protein family information assists membership confirmation and sequence annotation, as being used in our GeneFIND system. Finally, the motif alignments, which embed effective conservative substitution information for all known protein families, can be used to compile alternative scoring matrices or amino acid substitution groups. Such prior information is known to be crucial for improving the accuracy of various database search and alignment algorithms.

Acknowledgments

This study is supported in part by grant R29 LM05524 from National Library of Medicine.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410.

Attwood, T. K., Beck, M. E., Bleasby, A. J., *et al.* (1997) Novel developments with the PRINTS protein fingerprint database. *Nuc. Acids Res.* **25**, 212-216.

Bairoch, A. and Apweiler, R. (1997) The SWISS-PROT protein sequence data bank and its supplement TrEMBL. *Nuc. Acids Res.* **25**, 31-36.

Bairoch, A., Bucher, P. and Hofmann, K. (1997) The PROSITE database, its status in 1997. *Nuc. Acids Res.* **25**, 217-221.

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

Bleasby A. J., Akrigg D. and Attwood T. K. (1994) OWL - a non-redundant composite protein sequence database. *Nuc. Acids Res.* **22**, 3574-3577.

Eddy, S. R., Mitchison, G and Durbin, R. (1995) Maximum Discrimination hidden Markov models of sequence consensus. *J. Comp. Biol.* **2**, 9-23.

Fabian, P., Murvai, J., Hatsagi, Z., Vlahovicek, K., Hegyi, H. and Pongor, S. (1997) The SBASE protein domain library, release 5.0: a collection of annotated protein sequence segments. *Nuc. Acids Res.* **25**, 240-243.

George, D.G., Dodson, R.J., Garavelli, J.S., *et al.* (1997) The protein information resource (PIR) and PIR-international protein sequence database. *Nuc. Acids Res.* **25**, 24-27.

Gribskov, M., Luthy, R. and Eisenberg, D. (1989) Profile analysis. *Methods Enzymol.*, **183**, 146-159.

Henikoff, J.G., Peitrokovski, S. and Henikoff, S. (1997) Recent enhancements to the Blocks database servers. *Nuc. Acids Res.* **25**, 222-225.

Hubbard, T. J. P., Murzin, A. G., Brenner, S. E. and Chothia, C. (1997) SCOP: a structural classification of proteins database. *Nucleic Acids Res* **25,** 236-239.

Krogh, A., Brown, M., Mian, I. S., Sjolander, K. and Haussler, D. (1994) Hidden markov models in computational biology: applications to protein modeling. *J. Mol. Biol.* **235**, 1501-1531.

Mewes, H.W., Albermann, K., Heumann, K., Liebl, S. and Pfeiffer, F. (1997) MIPS: a database for protein sequences, homology data and yeast genome information. *Nuc. Acids Res.* **25,** 28-30.

Pearson, W. R. and Lipman, D. J. (1988) Improved tools for biological sequence comparisons. *Proc. Nat. Acad. Sci.* USA **85**, 2444-2448.

Schneider, R., Daruvar, A. and Sander, C. (1997) The HSSP database of protein structure-sequence alignment. *Nuc. Acids Res.*, **25**, 226-230.

Smith, R. F. and Smith, T. F. (1992) Pattern-induced multi-sequence alignment (PIMA) algorithm employing secondary structure-dependent gap penalties for use in comparative protein modelling. *Prot. Eng.* **5**, 35-41.

Smith, T.F., and Waterman, M.S. (1981) Comparison of bio-sequences. *Adv. Appl. Math.* **2**, 482-489.

Sonnhammer, E.L.L. and Kahn, D. (1994) Modular arrangement of proteins as inferred from analysis of homology. *Prot. Sci.* **3**, 482-492.

Sonnhammer, E. L., Eddy, S. R. and Durbin, R. (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* **28**, 405-420.

Thompson, J. D., Higgins, D. G. and 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. *Nuc. Acids Res.*, **22**, 4673-4680.

Wu, C. H. (1996) Gene classification artificial neural system. *Methods Enzymol.*, **266**, 71-88.

Wu, C.H., 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.

Wu, C. H., Shivakumar, S. and Barker, W. C. (1997a). Protein family identification and information retrieval using a ProClass database and a motif neural design. *Math. Model. & Sci. Comp.* **8**, (in press).

Wu, C.H., Shivakumar, S., Shivakumar, C. and Chen, S.C. (1997b) GeneFIND web server for protein family identification and information retrieval. *Comp. Applic. Biosci.* **13**, (in press).