*3610–3615* *Nucleic Acids Research, 1994, Vol. 22, No. 17*

# The SBASE protein domain library, release 3.0: a collection of annotated protein sequence segments

Sándor Pongor[1,2]*, Zsolt Hátsági[1,2], Kirill Degtyarenko[1§], Péter Fábián[1,2,], Vesna Skerl[1+], Hedvig Hegyi[2], János Murvai[2] and Valeria Bevilacqua[1]

[1]International Centre for Genetic Engineering and Biotechnology, Area Science Park, 34012 Trieste, Italy and [2]ABC Institute for Biochemistry and Protein Research, 2100 Gödöllö, Hungary

## ABSTRACT

**SBASE 3.0 is the third release of SBASE, a collection of annotated protein domain sequences. SBASE entries represent various structural, functional, ligand-binding and topogenic segments of proteins as defined by their publishing authors. SBASE can be used for establishing domain homologies using different database-search tools such as FASTA [Lipman and Pearson (1985) *Science*, 227, 1436–1441], and BLAST3 [Altschul and Lipman (1990) *Proc. Natl. Acad. Sci. USA*, 87, 5509–5513] which is especially useful in the case of loosely defined domain types for which efficient consensus patterns can not be established. The present release contains 41,749 entries provided with standardized names and cross-referenced to the major protein and nucleic acid databanks as well as to the PROSITE catalogue of protein sequence patterns. The entries are clustered into 2285 groups using the BLAST algorithm for computing similarity measures. SBASE 3.0 is freely available on request to the authors or by anonymous 'ftp' file transfer from <ftp.icgeb.trieste.it>. Individual records can be retrieved with the gopher server at <icgeb.trieste.it> and with a www-server at <http://www.icgeb.trieste.it>. Automated searching of SBASE by BLAST can be carried out with the electronic mail server <sbase@icgeb.trieste.it>. Another mail server <domain@hubi.abc.hu> assigns SBASE domain homologies on the basis of SWISS-PROT searches. A comparison of pertinent search strategies is presented.**

## INTRODUCTION

SBASE is a collection of protein domain sequences designed to facilitate the detection of distant similarities such as found between modules of multidomain proteins [1,2]. Typically, a multidomain protein can share a biologically significant sequence pattern with a number of different, functionally related proteins or protein domains, however the alignments may not be highly significant in the mathematical sense. SBASE can be considered as a conversion of the protein sequence database into a format that facilitates detection of such functional and structural similarities [3,4].

The current release 3.0 of SBASE contains over forty thousand annotated protein sequence segments consistently named by structure, function, biased composition, binding-specificity and/or similarity to other proteins. Searching this database with standard programs such as FASTA [5] or BLAST3 [6] yields information on the potential functions of the detected homology regions which may allow the detection of potential domain homologies and prediction of function.

The main developments with respect to the last release can be summarized as follows:

i) Release 3.0 contains 41749 sequence entries, 7231 (21%) more than release 2.0 (Table 1).

ii) Repetitive sequences are marked. In the mathematical sense, repetitive sequences have a low Kolmogorov string complexity measure [7]. The complexity of each domain sequence was calculated by John Wootton's seg program [8], and the fraction of low complexity region(s) in the sequence (a number between 0 and 1) is given in the CO line of those entries, in which the value is greater than an arbitrary threshold of 0.05. These values are meant as a warning to the experimenter to indicate that high similarity scores to low complexity sequences may be the result of chance rather than of biologically important similarity.

iii) The entries were subjected to cluster analysis using the BLAST algorithm. The BLAST score was used as a similarity measure, and entries with significant BLAST scores and not differing more than 50% in length were allowed to cluster together. The list of the resulting clusters is deposited into a separate database, SBASE-CLUSTERS, accessible through www-server (see below). Low-complexity entries were clustered separately and were given cluster numbers starting from 10000. SBASE-CLUSTERS now contains a total of 2287 clusters that each have at least 3 members. The cluster number(s) appear in the CL line. A further procedure was used to establish if entries were related also to other clusters, irrespective of the sequence length. An entry was considered related to a cluster if it had a

*To whom correspondence should be addressed

+Present address: Vinca Institute for Nuclear Sciences – Vinca, PO Box 522, Belgrade, Yugoslavia

§Permanent address: Institute of Biomedical Chemistry, Pogodinskaya 10, 119832 Moscow, Russia significant BLAST similarity score against all members of the cluster. The identification numbers of the related clusters appear in the CE line. Due to the nature of the clustering procedure used, overlapping domains of the same protein (e.g. EXTRACELLULAR and EGF-REPEAT of a receptor protein) may belong to several clusters at the same time.

iv) SBASE 3.0 is cross-referenced to the PRINTS 4.0 database of signatures [9], to the PRODOM 24 collection of homology domains [10] as well as to the BLOCKS 7.01 database of conserved sequence blocks [11]. The type of overlap is shown in the corresponding DR-lines (CT, NT = carboxy- or amino-terminal overlaps, respectively; CO = SBASE entry sequence contains a PRODOM, PRINTS or BLOCKS sequence, and IN = the opposite case).

v) The standardization of domain names proceeded further, now an estimated two third of the entries are provided with standard names assigned on the basis of the entry clusters in SBASE-CLUSTER.

vi) The electronic access to the data was substantially improved: two electronic mail servers and an on-line, cross-referenced hypertext server were installed (see below).

**Table 1.** Increase of data in SBASE release 3.0

| Release | Date | Records | Amino acids | Size [Mb] |
|---|---|---|---|---|
| 1.0 | 2-APR-92 | 27,221 | 1,551,445 | 17.2 |
| 2.0 | 13-FEB-93 | 34,518 | 1,922,524 | 24.9 |
| | | (+27%) | (+24%) | (+45%) |
| 3.0 | 28-MAY-94 | 41,749 | 2,339,538 | 37.3 |
| | | (+21%) | (+22%) | (+50%) |

**Table 2.** Examples of domains in SBASE 3.0

| Domain type | Number of records in SBASE 3.0 | Domain type | Number of records in SBASE 3.0 |
|---|---|---|---|
| STRUCTURAL DOMAINS | | HOMOLOGY DOMAINS | 3254 |
| IG-like repeats | 1231 | | |
| EGF-repeats | 352 | LIGAND-BINDING DOMAINS | |
| Heptad-repeats | 326 | Calcium-binding | 815 |
| Sushi repeats | 245 | Zinc-fingers | 1615 |
| FN3-repeats | 214 | Other DNA-binding | 1127 |
| Ank-repeat | 171 | RNA-binding | 162 |
| Annexin-repeats | 133 | Lectin domains | 84 |
| Kringle domain | 105 | Homeobox | 261 |
| TPR | 96 | HMG-box | 35 |
| SH3 | 80 | Helix-turn-helix (HTH) | 379 |
| SH2 | 63 | | |
| | | Helix-loop-helix (HLH) | 63 |
| Domains with biased composition | | Leucine-zipper | 127 |
| Ser-rich | 252 | | |
| Gly-rich | 249 | CELL TOPOLOGY DOMAINS | |
| Pro-rich | 190 | | |
| Cys-rich | 131 | Extracellular | 2057 |
| Acidic | 102 | Transmembrane | 9662 |
| Basic | 65 | Cytosolic | 2863 |
| Hydrophilic | 73 | Signal peptides | 4716 |
| Hydrophobic | 133 | Transit to organelles | 675 |
| | | Nuclear localization signals | 84 |

## DESCRIPTION OF THE DATA

### Definition of protein domains

Domains included in SBASE are protein sequence segments with known structure and/or function. The main classes of protein domains are defined as follows (Table 2): Structural domains are sequence segments with a known structure (like the protein modules [12] such as epidermal growth factor-like [EGF-REPEAT] and immunoglobulin-like [IG-LIKE] domains), biased composition (e.g. SER-RICH domains) as well as various sequence repeats. Homology domains are regions of homology to other proteins detected by the original authors. These homology regions are less well characterized than the 'established' structural domains but can be eventually used to define further domain types. Ligand-binding domains are sequence segments known to bind specific ligands (such as DNA, metals, sugars etc.). Cellular location domains are sequence segments known to be involved in protein targeting (signal peptides, nuclear localization signals, chloroplast transit peptides), as well as domains of transmembrane proteins (cytosolic, transmembrane, and extracellular domains). Examples of domain types are listed in Table 3.

### Source and origin of data

SBASE data originate from three main sources: i) from the SWISS-PROT protein sequence databank [13]; ii) from the Protein Sequence Database of the Protein Identification Resource (PIR) [14]; and iii) from the literature. The sequences are either translated from nucleotide sequence databases [15, 16] or directly keyed in at the protein level. From a total of 41,749 records in SBASE 3.0, 33,342 (79.9%), 5,606 (13.4%) and 2,801 (6.7%) are of eukaryotic, prokaryotic and viral origin, respectively. Domain sizes vary in length between 5 and 1,000 amino acids. The boundaries of the domains are either as previously defined in the original publications or determined by homology to domains with defined boundaries.

**Table 3.** Examples of new standard domain names in SBASE 3.0

| SBASE standard name | Descriptive name (occurrence) |
|---|---|
| C1R/C1S-REPEAT | Repeat found in complement components C1r/C1s, bone morphogenetic protein 1, *Xenopus laevis* neuronal A5 antigen, sea urchin protein uEGF, *Drosophila* dorsal-ventral patterning tolloid protein |
| LRR | Leucine-rich repeat |
| PK | (catalytic domain similar to) protein serine/threonine kinases |
| PTK | (catalytic domain similar to) protein tyrosine kinases |
| PTPASE | (catalytic domain similar to) protein tyrosine phosphatases |
| SH2 | *src* homology 2 (SH2) domain (protein tyrosine kinases, protein tyrosine phosphatases, *ras* GTPase activating proteins, growth factor receptor-bound protein 2) |
| SH3 | *src* homology 3 (SH3) domain (protein tyrosine kinases, protein tyrosine phosphatases, *ras* GTPase activating proteins, growth factor receptor-bound protein 2, yeast actin-binding protein ABP1, myosin heavy chain, PI-3-kinase) |
| TPR | Tetratricopeptide repeat, 34 aa motif (yeast cell division cycle proteins CDC23, CDC16, heat shock STI1 protein, *Aspergillus* BimA protein, *Drosophila* crooked neck protein, human transformation-sensitive protein IEF) |
| VWFA-REPEAT | VWF-like A domain (von Willebrand factor, integrin a subunits, collagen VI, cartilage matrix protein) |
| N-REPEAT | N amino acids long repeat of unknown structure or function (e.g. 10-REPEAT) |

Redundancy of sequences in SBASE 3.0 is kept at a minimal level. In some cases, the domain definitions overlap, so the same sequence (e.g. EGF-REPEAT) can be present both as an independent entry and as part of another entry (e.g. EXTRACELLULAR domain of a receptor). For the same reason, entries can belong to separate clusters in SBASE-CLUSTER.

### Cross-references

SBASE 3.0 has cross-references to several protein and nucleic acid databanks, as well as to the PROSITE [17] PRINTS 4.0 [9], PRODOM 24 [10], and BLOCKS 7.01 [11] databases (Table 4). In each record, the DR-lines contain the cross-reference data.

### Record structure

The format of SBASE 3.0 follows that of the EMBL and SWISS-PROT databases and can be directly formatted under the GCG program package using [18]. Sample records are shown in Figure 1. The field types used are listed in Table 5.

### Citation

Users of SBASE and of the e-mail servers are asked to cite this article in their publications, e.g. in the following form: 'The sequence homologies were analyzed searching the SBASE protein domain sequence library release 3.0″ via automated electronic mail server'.

## DISTRIBUTION AND ACCESS

### Distribution

SBASE 3.0 (May 1994) is distributed by anonymous 'ftp' file transfer from <ftp.icgeb.trieste.it>. The complete database is 37.3 Mb, its compressed form is 4.3 Mb.

### Retrieval of records by gopher server

Individual entries are available through the gopher server [19] of ICGEB at <icgeb.trieste.it>. Entries can be retrieved by SBASE identifiers, standard names, description of the parent protein, organism and authors' names.

### BLAST search by electronic mail server

SBASE 3.0 can be searched by the BLAST program using the e-mail server sbase@icgeb.trieste.it. A query sequence in a simple format (Figure 2A) can be sent to this address, and the results (Figure 2B) are returned by electronic mail. Based on the FTHOM algorithm [30], a related e-mail server was created in order to assign SBASE domain homologies on the basis of BLAST searches performed on the SWISS-PROT database. The results of this search appear as best potential domain homologies ranked according to BLAST score (Figure 2C).

**Table 4.** Cross-references to other databases in SBASE

| DATABASE | Ref. | No. of pointers in SBASE 2.0 | No. of pointers in SBASE 3.0 |
|---|---|---|---|
| EMBL | [15] | 51,555 | 64,074 |
| PIR | [14] | 43,855 | 50,132 |
| SWISS-PROT | [13] | 34,518 | 41,749 |
| PRODOM24 | [10] | – | 37,243 |
| BLOCKS7.01 | [11] | – | 12,483 |
| PROSITE | [17] | 6,707 | 9,307 |
| PRINTS4.0 | [9] | – | 8,430 |
| PDB | [22] | 5 ,438 | 1,239 |
| MIM | [23] | 5,149 | 6,829 |
| TFD | [24] | 1, 572 | 1,554 |
| FLYBASE | [25] | 1,354 | 1,354 |
| ECOGENE | [26] | 1,216 | 1,300 |
| SWISS-2DPAGE | [27] | 360 | 182 |
| HIV | [28] | 58 | 51 |
| REBASE | [29] | 14 | 7 |

### Access by www-server

All the above services can be accessed on-line also using the www-server [20] at <http://www.icgeb.trieste.it>. At present, cross-references to SBASE-CLUSTERS, EMBL, MEDLINE, MIM, PRINTS4.0, PRODOM24, PROSITE11, and SWISS-PROT28 (see underlined items in Figure 1) can be directly accessed through the www-server. At present, SBASE-CLUSTERS is available only through the www-server.

## DISCUSSION: A COMPARISON WITH OTHER DATABASES

Detection of protein domain homologies can be carried out by a number of methods and search strategies; many of these are based on specific databases such as BLOCKS [11], PROSITE [17], PRINTS [9], PRODOM [10], PLSEARCH [21]. The crucial difference between these databases is how the domain homology groups are defined and how their (consensus) structure is represented.

PROSITE, BLOCKS, PRINTS and PLSEARCH are databases of consensus representations for domain groups. PROSITE is based on heuristically defined domains which are predominantly represented as short regular expressions (signatures). BLOCKS is essentially based on the PROSITE domains but uses an extended representation of short conserved regions (blocks) for the sequences. PRINTS is also based on heuristically defined domains but uses yet another representation of regular expressions (fingerprints) that are more robust than the short PROSITE signatures. PLSEARCH uses an automated clustering procedure for the identification of homology groups which are then represented as consensus sequences. All of these databases are

**Table 5.** Types of comment lines in SBASE 3.0 records

| Line identifier | Content of the comment line |
|---|---|
| ID | Unique record IDentifier. If the SWISS-PROT name is available, it is followed by the starting and the ending positions of the domain (*e.g.* A20_HUMAN-286–317). Since release 2.0, we started to store, in the rest of the ID-line, a short domain description for the sake of easier interpretation of database search data. |
| DT | DaTe of entry. |
| SN | Standard Name. |
| DP | Definition of the Parent protein. |
| DE | DEfinition of the domain (same as SN + DP in short). |
| OS | Source Organism Species name. |
| OC | Organism Classification (taxonomy line). |
| DR | Database Reference (cross-reference). |
| CO | Low COmplexity. |
| CL | CLUSTER CLuster number |
| CE | DISTCLUST Related clusters (Distclust) |
| RA | Authors of the literature Reference. |
| RL | Literature Reference. |
| RM | Reference to MEDLINE/MEDLARS |
| SQ | SeQuence |

**A**

```
ID      ANX1_COLLI-118-178 STANDARD; PRT;
DT      26-NOV-93 (REL. 3, CREATED)
SN      ANNEXIN-REPEAT
DE      ANNEXIN-REPEAT ANNEXIN I (LIPOCORTIN I)
        (CALPACTIN II) (CHROMOBI
DP      ANNEXIN I (LIPOCORTIN I) (CALPACTIN II)
        (CHROMOBINDIN 9) (P35)
DP      (PHOSPHOLIPASE A2 INHIBITORY PROTEIN).
OS      COLUMBA LIVIA (DOMESTIC PIGEON).
OC      EUKARYOTA; METAZOA; CHORDATA; VERTEBRATA;
        TETRAPODA; AVES; NEOGNATHAE;
OC      COLUMBIFORMES.
DR      SWISS-PROT; ANX1_COLLI; P14950; AA 118-178
DR      EMBL; M22635; CLCAL.
DR      PROSITE; PS00223; ANNEXIN.
DR      PRINTS4.0 IN; ANNEXIN FM (ANNEXIN3).
DR      PRINTS4.0 IN; ANNEXIN TYPE I (ANNEXINI4).
DR      PRODOM24 IN; 28 (ANNEXIN (LIPOCORTIN I) II)
        (CHROMOBINDIN (PLAC.
DR      PRODOM24 CT; 28 (ANNEXIN (LIPOCORTIN I) II)
        (CHROMOBINDIN (PLAC.
DR      BLOCKS7.01 NT; BL00223B ANNEXINS REPEAT
        PROTEINS DOMAIN PROTEINS.
CL      CLUSTER 1469
CE      DISTCLUST 1484; 1468; 1470; 10432.
RA      HORSEMAN N.D.;
RL      MOL. ENDOCRINOL. 3:773-779(1989).
RM      89330493 [MEDLINE, MEDLARS]
SQ      SEQUENCE 61 AA;
        LRACMKGHGT DEDTLIEILA SRNNKEIREA
        CRYYKEVLKR DLTQDIISDT SGDFQKALVS
        L
//
```

**B**

```
ID      PGCA_HUMAN-2163-2200 STANDARD; PRT;
DT      26-NOV-93 (REL. 3, CREATED)
SN      EGF-REPEAT
DE      EGF-REPEAT CARTILAGE-SPECIFIC PROTEOGLYCAN
        CORE PROTEIN P
DP      CARTILAGE-SPECIFIC PROTEOGLYCAN CORE PROTEIN
        PRECURSOR (CSPCP)
DP      (AGGRECAN) (CHONDROITIN SULFATE PROTEOGLYCAN
        CORE PROTEIN 1).
OS      HOMO SAPIENS (HUMAN).
OC      EUKARYOTA; METAZOA; CHORDATA; VERTEBRATA;
        TETRAPODA; MAMMALIA;
OC      EUTHERIA; PRIMATES.
DR      SWISS-PROT; PGCA_HUMAN; P16112; AA 2163-2200
DR      EMBL; M55172; HSAGPRO.
DR      EMBL; J05062; HSCSPCP.
DR      EMBL; X17406; HSCSP.
DR      PIR; S08042; S08042.
DR      MIM; 155760; TENTH EDITION.
DR      PROSITE; PS00022; EGF.
DR      PRODOM24 NT; 4888 ((AGGRECAN). (CSPCP) PRECURSOR
        CORE PROTEOGLY.
DR      PRODOM24 IN; 12273 (P16112 CARTILAGE-SPECIFIC
        PROTEOGLYCAN CORE.
DR      BLOCKS7.01 IN; BL00022 EGF-LIKE DOMAIN PROTEINS
        CYSTEINE PATTERN.
CL      CLUSTER 1169
CE      DISTCLUST 90; 1653; 3355; 196; 1176; 1166.
RA      DOEGE K.J., SASAKI M., KIMURA T., YAMADA Y.;
RL      J. BIOL. CHEM. 266:894-902(1991).
RM      91093289 [MEDLINE , MEDLARS]
SQ      SEQUENCE 38 AA;
        APARSCAEEP CGAGTCKETE GHVICLCPPG YTGEHCNI
//
```

**C**

```
CL      #1169 CONTAINS:
AGRI_CHICK-1489-1521 EGF-REPEAT 3 AGRIN PRECURSOR.
AGRI_RAT-1441-1476 EGF-REPEAT 2 AGRIN PRECURSOR.
AGRI_RAT-1480-1515 EGF-REPEAT 3 AGRIN PRECURSOR.
CO6_HUMAN-522-556 LDLR-REPEAT, TYPE B COMPLEMENT C6 PRECURSOR.
```

```
CO9_HUMAN-509-543 LDLR-REPEAT, TYPE B COMPLEMENT C9 PRECURSOR
CRB_DROME-1207-1244 EGF-REPEAT 20 CRUMBS PROTEIN PRECURSOR (95F).
CRB_DROME-1759-1796 EGF-REPEAT 23 CRUMBS PROTEIN PRECURSOR (95F).
CRB_DROME-1874-1914 EGF-REPEAT 26 CRUMBS PROTEIN PRECURSOR (95F).
CRB_DROME-2030-2071 EGF-REPEAT 30 CRUMBS PROTEIN PRECURSOR (95F).
CRB_DROME-388-426 EGF-REPEAT 4 CRUMBS PROTEIN PRECURSOR (95F).
CRB_DROME-543-581 EGF-REPEAT 8 CRUMBS PROTEIN PRECURSOR (95F).
CRB_DROME-611-647 EGF-REPEAT 10 CRUMBS PROTEIN PRECURSOR (95F).
CRB_DROME-980-1022 EGF-REPEAT 19 CRUMBS PROTEIN PRECURSOR (95F).
CRIO_HUMAN-82-106 EGF-REPEAT EPIDERMAL GROWTH FACTOR-LIKE CRIPTO
FA12_HUMAN-188-209 EGF-REPEAT, TYPE A COAGULATION FACTOR XII
...
UROM_HUMAN-32-63   EGF-REPEAT   DOMAIN   UROMODULIN
PRECURSOR (TAMM-
UROM_HUMAN-69-106   EGF-REPEAT   DOMAIN,   UROMODULIN
PRECURSOR
UROM_RAT-34-65   EGF-REPEAT   DOMAIN   UROMODULIN
PRECURSOR (TAMM-HORSFALL
UROT_HUMAN-86-119 EGF-REPEAT, TYPE A TISSUE PLASMINOGEN
ACTIVATOR
UROT_MOUSE-83-116 EGF-REPEAT, TYPE A TISSUE PLASMINOGEN
ACTIVATOR
```

**Figure 1.** Sample entries from the SBASE 3.0 protein domain library. **A.** An annexin repeat domain; **B.** An epidermal growth factor-like domain (EGF-REPEAT) from aggrecan. The underlined items are linked in the SBASE World Wide Web server so the corresponding records can be viewed on the screen by 'clicking' on them. **C.** A record of SBASE-CLUSTERS containing EGF-repeats (partially shown). The entry shown in B is a member of this cluster.

**A**

```
MATRIX PAM120
EXPECT PARAMETER 25
SCORE PARAMETER 35
ALIGNMENTS YES
ANNOTATIONS NO
OUTPUT_PAGES 10
-SORT_BY_PVALUE
BEGIN
> mysequence
LRNGVDINTCNQNGLNGLHLASKEGHVKMVVELL....
```

**B**

```
                                                                                      Smallest
                                                                                      Poisson
                                                                         High         Probability
Sequences producing high-scoring segment pairs:                          Score        P(N)         N

ANX1_COLLI-118-178     ANNEXIN-REPEAT    ANNEXIN    I (LIPOCORTIN...     317          1.0e-43      1
ANX1_CAVCU-123-183     ANNEXIN-REPEAT    ANNEXIN    I (LIPOCORTIN...     221          1.8e-29      1
ANX1_HUMAN-122-182     ANNEXIN-REPEAT    ANNEXIN    I (LIPOCORTIN...     221          1.8e-29      1
ANX1_RAT-122-182       ANNEXIN-REPEAT    ANNEXIN    I (LIPOCORTIN...     213          2.8e-28      1
ANX1_MOUSE-122-182     ANNEXIN-REPEAT    ANNEXIN    I (LIPOCORTIN...     212          4.0e-28      1
ANX2_HUMAN-113-173     ANNEXIN-REPEAT    ANNEXIN    II (LIPOCORTI...     188          1.5e-24      1
ANX2_MOUSE-113-173     ANNEXIN-REPEAT    ANNEXIN    II (LIPOCORTI...     188          1.5e-24      1
ANX2_BOVIN-113-173     ANNEXIN-REPEAT    ANNEXIN    II (LIPOCORTI...     185          4.0e-24      1
ANX2_CHICK-113-173     ANNEXIN-REPEAT    ANNEXIN    II (LIPOCORTI...     180          2.2e-23      1
ANX2_XENLA-114-174     ANNEXIN-REPEAT    ANNEXIN    II TYPE II (L...     170          6.8e-22      1
ANXB_XENLA-114-174     ANNEXIN-REPEAT    ANNEXIN    II TYPE I (LI...     170          6.8e-22      1
ANX3_HUMAN-99-159      ANNEXIN-REPEAT    ANNEXIN    III (LIPOCORT...     163          7.5e-21      1
ANX3_RAT-100-160       ANNEXIN-REPEAT    ANNEXIN    III (LIPOCORT...     163          7.5e-21      1
```

**C**

```
Number of entries:55
Query: ANX1_COLLI-118-178                          Feature   FT      score
                                                      name   freq    sum
                                    DOMAIN ANNEXIN-REPEAT.   32      4089
                                   DOMAIN PHYCOBILIN-LIKE.    1        46
           PEPTIDE C-TERMINAL EXTENSION PEPTIDE (CTEP).
                                              (POTENTIAL).    1        40
```

**Figure 2.** Automated electronic mail servers based on the SBASE domain library. **A.** Input file format (see the respective help files for detailed instructions). **B.** Output of the sbase e-mail server (sbase@icgeb.trieste.it) in response to an annexin repeat shown in Figure 1A (detail); **C.** Output of the FTHOM domain homology server (domain@hubi.abc.hu) in response to the same query sequence (detail). Fr: Frequency of domain-name found in the BLAST output; Score sum: sum of BLAST scores belonging to a domain-name in the output [30]. Both server outputs contain alignments provided with annotations and detailed explanation about evaluation (not shown).

based on *a priori* knowledge of the domain consensus structure and require specific software for the analysis of a new sequence.

SBASE is a collection of domain sequences rather than of consensus structures. The domains are heuristically defined and many domain types are included for which consensus structures are not yet available. The added information in SBASE are the standard names, the domain-clusters and the cross-references to various consensus structure databases, which greatly facilitates the interpretation of similarities with standard programs such as BLAST or FASTA. PRODOM is the second domain database published so far which is, however, based on a very different principle: PRODOM domains are determined by sequence similarity clustering rather than by structure/function definitions. This procedure has the advantage of automation and can possibly lead to the identification of new domain types. On the other hand, PRODOM domains lack systematic names and do not necessarily coincide with accepted domain boundaries. For example, the cartilage specific proteoglycan core protein aggrecan (PGCA_HUMAN in SWISS-PROT) has, among others, an EGF-REPEAT at position 2163–2200. This is a 'regular' EGF sequence that contains the PROSITE and BLOCKS consensus sequences. In SBASE-CLUSTER, the EGF-REPEAT in question is clustered together with 44 other EGF-REPEATs (Figure 1C). In contrast, PRODOM24 does not cluster this domain together with other epidermal growth factor-like domains. Instead, an internal part of the sequence (2167–2190) appears as an unnamed unique domain of PRODOM. The advantage of SBASE and PRODOM is that they do not require a previously established consensus representation or a specific software for domain identification - searches can simply be carried out by programs like BLAST, FASTA and patterns, if any, extracted directly from the search results [31].

Future work in SBASE will concentrate on further standardization of domain names in the next releases so as to allow simple retrieval of more domain types. An automated procedure, based on standard domain names, is being developed for the identification of domain homologies by the FTHOM algorithm [31] and will be made available through the electronic mail server <domain@hubi.abc.hu>. Furthermore, the clustering procedure will be used to automatically identify potential homologues to SBASE entries that will be incorporated in SBASE as domain candidates. The next full release 4.0 will appear in April 1995.

## ACKNOWLEDGEMENTS

SBASE was established in 1990 and is maintained collaboratively by the International Center for Genetic Engineering and Biotechnology, Trieste, Italy and the ABC Institute for Biochemistry and Protein Research, Gödöllö, Hungary.

## REFERENCES

1. Barker, W.C., Hunt, L.T. and George, D.G. (1988) *Protein Seq. Data Anal.*, 1, 363–373.
2. Baron, M., Norman, D.G. and Campbell, I.D. (1991) *Trends in Biochemistry*, 16, 13–17.
3. Pongor, S., Skerl, V., Cserzö, M., H ts gi, Z., Simon, G. and Bevilacqua, V. (1993) *Protein Engineering*, 6, 391–395.
4. Pongor, S., Skerl, V., Cserzö, M.,H ts gi, Z., Simon, G. and Bevilacqua, V. (1993) *Nucleic Acids. Res*, 21, 3111–3115
5. Lipman, D.J. and Pearson, W.R. (1985) *Science*, 227, 1436–1441.
6. Altschul, S.F. and Lipman, D.J. (1990) *Proc. Natl. Acad. Sci. USA*, 87, 5509–5513.
7. Wootton, J.C. and Federhen, S. (1993) *Computers Chem.* 17, 149–163
8. Wootton, J.C. (1994) *Computers Chem.* 18 (in press)
9. Attwood, T.K. and Beck, M.E. (1994) *Protein Engineering*, 7, in press.
10. Sonnhammer, E.L.L. and Kahn, D. (1994) *Protein Science* 3, 482–492.
11. Henikoff, S. and Henikoff, J.G. (1991) *Nucleic Acids Res.*, 19, 6565–6572.
12. Bork, P. (1992) *Curr. Opin. Struct. Biol.*, 2, 413–421.
13. Bairoch, A. and Boeckmann, B. (1993) *Nucleic Acids Res.*, 21, 3093–3096.
14. Barker, W.C., George, D.G., Mewes, H.-W., Pfeiffer, F. and Tsugita, A. (1993) *Nucleic Acids Res.*, 21, 3089–3092.
15. Rice, C.M., Fuchs, R., Higgins, D.G., Stoehr, P.J. and Cameron, G.N. (1993) *Nucleic Acids Res.*, 21, 2967–2971.
16. Benson, D., Lipman, D.J. and Ostell, J. (1993) *Nucleic Acids Res.*, 21, 2963–2965.
17. Bairoch, A. (1993) *Nucleic Acids Res.*, 21, 3097–3103.
18. Devereux, J., Haeberli, P., and Smithies, O. (1984) *Nucleic Acids Res.* 12, 387–395.
19. Alberti, R., Anklesaria, F., Lindner, P., McCahill, M. and Torrey, D. (1991–1992) The Internet Gopher Protocol: a distributed document search and retrieval protocol. University of Minnesota Microcomputer and Workstation Networks Center.
20. The Wide Area Information Server Project (1991) Thinking Machines Corporation, Cambridge, MA.
21. Smith, R.F. and Smith, T.F. (1990) *Proc. Natl. Acad. Sci. USA*, 87, 118–122.
22. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) *J. Mol. Biol.*, 112, 535–542.
23. McKusick, V.M. (1990) Mendelian Inheritance in Man. John H Hopkins University Press, Baltimore, MD.
24. Ghosh, D. (1993) *Nucleic Acids Res.*, 21, 3117–3118.
25. FlyBase Consortium, Biological Laboratories, Harvard University, Cambridge, MA.
26. Rudd, K.E., Bouffard, G. and Miller, G. (1992) *In* 'Genome analysis', K.E. Davies and S.M.Tilghmann (eds.), pp. 1–38 Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
27. Appel, R.D., Sanchez, J.-C., Bairoch, A., Golaz, O., Miu, M., Pasquali, C., Reynaldo Vargas, J., Hughes, G.J. and Hochstrasser, D.F. (1993) *Electrophoresis* 14, 1232–1238.
28. Myers, F. (1990) Human Retrovirus and AIDS Database, Los Alamos National Laboratory, NM.
29. Roberts, R.J. and Macelis, D. (1993) *Nucleic Acids Res.*, 21, 3125–3137.
30. Hegyi, H. and Pongor, S. (1992) *CABIOS*, 9, 371–372
31. Polner, G., Skerl, V. and Pongor, S. (1993) *Protein Seq. Data Anal.*, 5, 409–413.