240–243 *Nucleic Acids Research, 1997, Vol. 25, No. 1*

The SBASE protein domain library, release 5.0: a collection of annotated protein sequence segments

Péter Fábián[1,2], János Murvai[1], Zsolt Hátsági[1,2], Kristian Vlahovicek[2], Hedvig Hegyi[1] and Sándor Pongor[1,2,*]

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

Received October 15, 1996; Accepted October 21, 1996

ABSTRACT

SBASE 5.0 is the fifth release of SBASE, a collection of annotated protein domain sequences that represent various structural, functional, ligand-binding and topogenic segments of proteins. SBASE was designed to facilitate the detection of functional homologies and can be searched with standard database-search programs. The present release contains over 79863 entries provided with standardized names and is cross-referenced to all major sequence databases and sequence pattern collections. The information is assigned to individual domains rather than to entire protein sequences, thus SBASE contains substantially more cross-references and links than do the protein sequence databases. The entries are clustered into >16 000 groups in order to facilitate the detection of distant similarities. SBASE 5.0 is freely available by anonymous 'ftp' file transfer from <ftp.icgeb.trieste.it >. Automated searching of SBASE with BLAST can be carried out with the WWW-server <http://www.icgeb.trieste.it/sbase/ >. and with the electronic mail server <sbase@icgeb.trieste.it > which now also provides a graphic representation of the homologies. A related WWW-server <http://www.abc.hu/blast.html > and e-mail server <domain@hubi.abc.hu > predicts SBASE domain homologies on the basis of SWISS-PROT searches.

INTRODUCTION

Proteins often share biologically significant segments with a number of different, functionally unrelated proteins or protein domains, even though the sequence alignments may not be highly significant in the mathematical sense. SBASE is a collection of protein domain sequences designed to facilitate the detection of such distant similarities, typically found between multidomain proteins (1,2). SBASE can be considered a conversion of the protein sequence database into a form labelled by domain-names (3,4). However, this database contains more site-specific (i.e. domain-specific) information than the corresponding protein sequence databases, which facilitates the detection of functional and structural similarities.

The current release 5.0 of SBASE contains over seventy thousand annotated protein sequence segments consistently named by structure, function, biased composition, binding-specificity and/or similarity to other proteins. The format of the database is such that it can be searched with standard programs like FASTA (5) or BLASTP (6) and the information given allows for the prediction of function and the direct detection of potential domain homologies.

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

(i) Release 5.0 contains 79 863 sequence entries, 30% more than release 4.0 (Table 1).

(ii) The entries were clustered on the basis of the BLAST similarity scores as previously described (4). The list of all clusters with at least two members is deposited into a separate database, SBASE-CLUSTERS, which is now available through anonymous ftp as well as through the www-server. A total of 16 900 clusters were found.

(iii) An estimated 95% of the records are now provided with standard names. Short descriptions and literature reviews were prepared on some of the domain types that are not described in other collections. These are now available through the www-server, along with links to all the cross-referenced databases.

Table 1. Increase of data in SBASE release 5.0

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%)
4.0	15-June-95	61 137	3 281 782	(50 Mb)
		(+48%)	(+40%)	(+34%)
5.0	06-Oct-96	79 862	4 118 506	75 Mb
		(+30%)	(25%)	(50%)

* To whom correspondence should be addressed at: ICGEB, Area Science Park, Padriciano 99, 34012 Trieste, Italy. Tel: +39 40 3757300; Fax: +39 40 226555; Email: pongor@icgeb.trieste.it

```
ID   ANX1_COLLI-118-178 STANDARD;    PRT;
DT   06-OCT-96 (REL. 5, CREATED)
SN   ANNEXIN-REPEAT.
DE   ANNEXIN-REPEAT ANNEXIN I (LIPOCORTIN I) (CALPACTIN II) (CHROMOBINDIN 9)
DE   (P35) (PHOSPHOLIPASE A2 INHIBITORY PROTEIN).
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   PIR; A40153; LUPY1.
DR   PROSITE IN; PDOC00195; ANNEXIN.
DR   PRINTS11.0 IN;  ANNEXIN FM (ANNEXIN3).
DR   PRINTS11.0 IN;  ANNEXIN TYPE I (ANNEXINI4).
DR   PRODOM28 IN;  19 (ANNEXIN (LIPOCORTIN I) II) (CHROMOBINDIN (PLAC.
DR   BLOCKS9.0 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).
SQ   SEQUENCE 61 AA
     LRACMKGHGT DEDTLIEILA SRNNKEIREA CRYYKEVLKR DLTQDIISDT
     SGDFQKALVS L
//
```

Figure 1. Sample entries from the SBASE 5.0 protein domain library. An annexin repeat domain. 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.

Table 2. Examples of domains in SBASE 5.0

Domain type	Number of records in SBASE 5.0	Domain type	Number of records in SBASE 5.0
Structural domains		**Homology domains**	
IG-like repeats	1542	**Ligand-binding domains**	
EGF-repeats	639	Calcium-binding	1108
Heptad-repeats	402	Zinc-fingers	2640
Sushi repeats	266	Other DNA-binding	388
FN3-repeats	405	RNA-binding	443
Ank-repeat	303	Lectin domains	411
Annexin-repeats	150	Homeobox	374
Kringle domain	110	HMG-box	126
TPR	158	Helix–turn–helix (HTH)	669
SH3	151	Helix–loop–helix (HLH)	128
SH2	144	Leucine-zipper	210
Domains with biased composition		**Cell topology domains**	
Ser-rich	486	Extracellular	3486
Gly-rich	385	Transmembrane	23856
Pro-rich	326	Cytosolic	3909
Cys-rich	170	Signal peptides	6632
Acidic	211	Transit to organelles	891
Basic	125	Nuclear localization signals	260
Hydrophilic	125		
Hydrophobic	96	**Miscellaneous repeats**	4310

DESCRIPTION OF THE DATA

Definition of protein domains

Domains included in SBASE are protein sequence segments with known structure and/or function (see refs 3,4 for details). The main entry classes are summarized in Table 2. The boundaries of the domains are either as previously defined in the original publications or determined by homology to domains with known boundaries.

Source and origin of data

SBASE data originate from three main sources: (i) from the SWISS-PROT protein sequence databank (11); (ii) from the Protein Sequence Database of the Protein Identification Resource (PIR) (12); and (iii) from the literature. The sequences are either translated from nucleotide sequence databases (13,14) or directly keyed in at the protein level. From a total of 79 863 records in SBASE 5.0, 58 389 (73%), 16 993 (21.2%) and 4154 (5%) are of eukaryotic, prokaryotic and viral origin, respectively. Domain sizes vary in length between 5 and 1000 amino acids. See Figure 1 for a sample entry.

Redundancy of sequences in SBASE 5.0 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 are included in several separate clusters in SBASE-CLUSTER.

Cross-references

SBASE 5.0 has cross-references to several protein and nucleic acid databanks, as well as to the PROSITE (15) PRINTS (7), PRODOM (8), and BLOCKS (9) databases (Table 3). In each record, the DR-lines contain the cross-reference data.

Record structure

The format of SBASE 5.0 follows that of the EMBL and SWISS-PROT databases and can be directly formatted under the GCG program package using (16). The field types used are listed in Table 4.

Table 3. Cross-references to other databases in SBASE

Database	Ref.	No of pointers in			
		SBASE 2.0	SBASE 3.0	SBASE 4.0	SBASE 5.0
EMBL	(13)	51 555	64 074	99 275	137 117
PIR	(12)	43 855	50 132	74 403	84 991
SWISS-PROT	(11)	34 518	41 749	61137	79 863
PRODOM	(8)	–	37 243	52 464	54 510
BLOCKS	(9)	–	12 483	17 245	26 930
PROSITE	(15)	6707	9307	16 029	26 384
PRINTS	(5)	–	8430	17 142	26 384
PDB	(17)	5438	1239	1109	3995
MIM	(18)	5149	6829	8570	11 161
FLYBASE	(19)	1354	1354	2321	2881
ECOGENE	(20)	1216	1300	2422	4442
HIV	(21)	58	51	92	92
REBASE	(22)	14	7	7	10

Citation

Users of SBASE and of the www/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 5.0″ via automated electronic mail server'.

Table 4. Types of comment lines in SBASE 5.0 records

Line identifier	Content of the comment line
ID	Unique record **ID**entifier. 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	**D**a**T**e of entry.
SN	**S**tandard **N**ame.
DP	**D**efinition of the **P**arent protein.
DE	**DE**finition of the domain (same as SN + DP in short).
OS	Source **O**rganism **S**pecies name.
OC	**O**rganism **C**lassification (taxonomy line).
DR	**D**atabase **R**eference (cross-reference).
CO	Low **CO**mplexity.
CL	CLUSTER **CL**uster number
CE	DISTCLUST Related clusters (Distclust)
RA	**A**uthors of the literature **R**eference.
RL	**L**iterature **R**eference.
RM	**R**eference to **M**EDLINE/MEDLARS
SQ	**S**e**Q**uence

A

```
QUERY=UROT_MOUSE  1|================================================|559
>UROT_DESRO-225-477                    ---------------- ----  SPR, SALIVARY PLASMIN
>UROT_MOUSE-213-294                 -------                  KRINGLE, TISSUE PLASM
>UROT_MOUSE-124-205              -------                     KRINGLE, TISSUE PLASM
>UROK_MOUSE-180-433                    ---------------- ----  SPR, UROKINASE-TYPE P
>UROT_RAT-213-294                   -------                  KRINGLE, TISSUE PLASM
>UROT_RAT-124-205                -------                     KRINGLE, TISSUE PLASM
>UROK_RAT-179-432                      ---------------- ----  SPR, UROKINASE-TYPE P
>UROT_HUMAN-127-208              -------                     KRINGLE, TISSUE PLASM
>UROK_PIG-190-442                      ---------------- ----  SPR, UROKINASE-TYPE P
>UROK_HUMAN-179-431                    ---------------- -----  SPR, UROKINASE-TYPE P
>UROK_CHICK-173-434                    -- - ---------- ----  SPR, UROKINASE-TYPE P
>UROT_DESRO-128-209              -------                     KRINGLE, SALIVARY PLA
>UROT_HUMAN-215-296                 -------                  KRINGLE, TISSUE PLASM
>UROT_MOUSE-38-82       ---                                  FN1-REPEAT, TISSUE PL
>UROT_RAT-38-82         ---                                  FN1-REPEAT, TISSUE PL
>FA12_HUMAN-373-615                       ------ ------ - --  SPR, COAGULATION FACT
>UROK_PAPCY-178-433                    ---------------------  SPR, UROKINASE-TYPE P
>FA11_HUMAN-384-618                     -------       -----  SPR, COAGULATION FACT
>UROT_HUMAN-41-85       ---                                  FN1-REPEAT, TISSUE PL
>UROK_PIG-72-153              -----                         KRINGLE, UROKINASE-TY
>UROK_MOUSE-71-152            -----                         KRINGLE, UROKINASE-TY
>UROT_MOUSE-83-116         ---                              EGF-REPEAT, TISSUE PL
>UROT_RAT-83-116           ---                              EGF-REPEAT, TISSUE PL
>UROT_HUMAN-86-119         ---                              EGF-REPEAT, TISSUE PL
>UROK_RAT-70-151                    ------                  KRINGLE, UROKINASE-TY
>UROK_PAPCY-69-150            -----                         KRINGLE, UROKINASE-TY
>UROK_HUMAN-70-151                  ------                  KRINGLE, UROKINASE-TY
>UROT_DESRO-80-120         ---                              EGF-REPEAT, SALIVARY
>UROK_CHICK-79-158              ---                         KRINGLE, UROKINASE-TY
>UROK_CHICK-79-158                   --                     KRINGLE, UROKINASE-TY
```

B

```
Number of entries:82

Query: annexin-repeat

                Feature      FT    score
                 name       freq    sum

          ANNEXIN-REPEAT     35     5063
           39 KD PROTEIN      3       72
             BP74-REPEAT      2       47
              ANK-REPEAT      1       42
                      PK      2       40
             EXTRACELLULAR    1       32
DOMAIN RODLIKE TAIL           1       26
   DNA-BINDING; HTH           1       16
             ATP-BINDING      1       10
```

Figure 2. Automated electronic search servers based on the SBASE domain library. (**A**) Output of the sbase www/e-mail server (www.icgeb.trieste.it/sbase and sbase@icgeb.trieste.it) in response to a tissue plasminogen activator sequence query (detail). The known domain structure of this query is S-P-FN1-EGF-KRINGLE-KRINGLE-SPR, where S = signal peptide, P = propeptide, FN1 = Fibronectin type 1 repeat, SPR = Serine protease. The weak homologies to signal peptide and propeptide appear at the bottom of the list, omitted from this picture. Also, the domain description lines are truncated in this figure. The server uses the 'HSPCRUNCH' program by Sonnhammer and Durbin (24); (**B**) Output of the FTHOM domain homology server (www.abc.hu/blast.html and 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 (23). Both server outputs contain alignments provided with annotations and a detailed explanation about evaluation (not shown).

DISTRIBUTION AND ACCESS

Distribution

SBASE 5.0 (October 6, 1996) is distributed by anonymous 'ftp' file transfer from <ftp.icgeb.trieste.it >. The complete database (including the records and list of clusters), is 75 Mb, its compressed form is 8.3 Mb.

BLAST search by electronic mail server

SBASE 5.0 can be searched by the BLAST program using the WWW-server <http://base.icgeb.trieste.it/sbase/ > and the e-mail server <sbase@icgeb.trieste.it >. A related server was created in order to assign SBASE domain homologies on the basis of BLAST searches performed on the SWISS-PROT database (23). This service (available at <http://www.abc.hu/blast.html > and at <domain@hubi.abc.hu >) returns the best potential domain homologies ranked according to BLAST score (Fig. 2B).

Access by www-server

All the above services can be accessed on-line also using the www-server at <http://www.icgeb.trieste.it >. At present, cross-references to SBASE-CLUSTERS, EMBL, MEDLINE, MIM, PRINTS, PRODOM, PROSITE, and SWISS-PROT can be directly accessed through the www-server.

ACKNOWLEDGEMENTS

SBASE was established in 1990 and is maintained collaboratively by the International Centre for Genetic Engineering and Biotechnology, Trieste, Italy and the ABC Institute for Biochemistry and Protein Research, Gödöllö, Hungary. The help of Suzanne Kerbavcic with the manuscript is gratefully acknowledged.

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 Biochem.*, **16**, 13–17.
3 Pongor, S., Skerl, V., Cserzö, M. and Hátsági, Z., Simon, G. and Bevilacqua, V. (1993) *Protein Eng.*, **6**, 391–395.
4 Pongor, S., Hátsági, Z., Degtyarenko, K., Fábián, P., Skerl, V.,Hegyi, H., Murvai, J. and Bevilacqua, V. (1994) *Nucleic Acids Res.*, **22**, 3610–3615.
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 Attwood, T.K., Beck, M.E., Bleasby, A.J., Degtyarenko, K., and Parry-Smith D.J. (1994) *Nucleic Acids Res.*, **24**, 182–188
8 Sonnhammer, E.L.L. and Kahn, D. (1994) *Protein Sci.*, **3**, 482–492.
9 Pietrokovski, S., Henikoff, J.G.. and Henikoff, S. (1994) *Nucleic Acids. Res.* **24**, 197–200.
10 Bork, P. (1992) *Curr. Opin. Struct. Biol.*, **2**, 413–421.
11 Bairoch, A. and Appweiler, R. (1996) *Nucleic Acids Res*., **24**, 21–25
12 George, D.G., Barker, W.C., Mewes, H.-W., Pfeiffer, F. and Tsugita, A. (1996) *Nucleic Acids Res*., **24**, 17–20.
13 Rodriguez-Thomé, P. Stoehr, P.J., Cameron, G.N. and Flores, T.P. (1996) *Nucleic Acids Res*., **24**, 6–12
14 Benson, D., Boguski, M., Lipman, D.J. and Ostell, J. (1996) *Nucleic Acids Res*., **24**, 1–5.
15 Bairoch, A., Bucher, P. and Hoffmann, K.(1996) *Nucleic Acids Res*., **24**, 189–196.
16 Devereux, J., Haeberli, P., and Smithies, O. (1984) *Nucleic Acids Res.* **12**, 387–395.
17 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.
18 Pearson, P., Francomano, C., Foster, P., Bocchini, C., Li, P., and McKusick, V.M. (1994) *Nucleic Acids Res*., **22**, 3470–3473.
19 FlyBase Consortium, (1996) *Nucleic Acids Res*., **24**, 53–56.
20 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.
21 Myers, F. (1990) Human Retrovirus and AIDS Database, Los Alamos National Laboratory, NM.
22 Roberts, R.J. and Macelis, D. (1996) *Nucleic Acids Res*., **24**, 223–235.
23 Hegyi, H. and Pongor, S. (1992) *Comput. Applic. Biosci*., **9**, 371–372.
24 Sonnhammer, E. L. L and Durbin, R. (1994). *Comput. Applic. Biosci*., **10**, 301–307.