210–213 *Nucleic Acids Research, 1996, Vol. 24, No. 1*

The SBASE protein domain library, Release 4.0: a collection of annotated protein sequence segments

János Murvai[1], Andrei Gabrielian[2], Péter Fábián[1,2], Zsolt Hátsági[1,2], Kirill Degtyarenko[2,+], Hedvig Hegyi[1] and Sándor Pongor[1,2,*]

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

Received September 25, 1995; Revised and Accepted October 12, 1995

ABSTRACT

SBASE 4.0 is the fourth 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 tools, such as FASTA and BLAST3. The present release contains 61 137 entries provided with standardized names and cross-referenced to all major protein, nucleic acid and sequence pattern collections. The entries are clustered into 13 155 groups in order to facilitate detection of distant similarities. SBASE 4.0 is freely available 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 World Wide Web server at http://www.icgeb.trieste.it. Automated searching of SBASE with BLAST can be carried out with the electronic mail server sbase@icgeb.trieste.it, which now also provides a graphic representation of the homologies. A related mail server, domain@hubi.abc.hu, assigns SBASE domain homologies on the basis of SWISS-PROT searches.

INTRODUCTION

SBASE is a collection of protein domain sequences designed to facilitate the detection of distant similarities typically found between modules of multidomain proteins (1,2). A multidomain protein can share a biologically significant sequence pattern with a number of different, functionally related proteins or protein domains, even though the sequence 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 4.0 of SBASE contains over 60 000 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 BLAST3 (6), and the information given allows 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 4.0 contains 61 137 sequence entries, 48% more than release 3.0 (Table 1).

Table 1. Increase of data in SBASE Release 4.0

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 (+27%)	1 922 524 (+24%)	24.9 (+45%)
3.0	28 May 94	41 749 (+21%)	2 339 538 (+22%)	37.3 (+50%)
4.0	15 Jun 95	61 137 (+48 %)	3 281 782 (+40%)	50 (+34%)

(ii) The entries were clustered on the basis of BLAST similarity scores, as previously described (4). The list of all clusters having at least two members is deposited in a separate database, SBASE-CLUSTERS, which is now available through anonymous ftp, as well as through the World Wide Web (WWW) server. A total of 13 155 clusters were found. The definition of clusters is as previously described (4).

(iii) An estimated 90% of the records are now provided with standard names. In addition to domains types used in the previous releases (e.g. structural, functional, cellular topology and biased composition domains), standardized names were given to repeat units that have no known function, using the name of the parent protein or parent protein family followed by the word ‘REPEAT’. Short descriptions and literature reviews have been prepared on some of the domain types that are not described in other collections. These are now available through the WWW server.

(iv) The graphical interface of Sonnhammer and Durbin (7), capable of displaying domain homologies along the query sequence, was added to the WWW/email server sbase@icgeb.trieste.it.

*To whom correspondence should be addressed at: International Centre for Genetic Engineering and Biotechnology, Area Science Park, 34012 Trieste, Italy

+Present address: Department of Biochemistry and Molecular Biology, University of Leeds, Leeds LS2 9JT, UK

DESCRIPTION OF THE DATA

Definition of protein domains

Domains included in SBASE are protein sequence segments with known structure and/or function (for details see 3,4). The main entry clases are summarized in Table 2. As a rule domain boundaries were taken as indicated by the publishing authors.

Table 2. Examples of domains in SBASE 4.0

Domain type	Number of records in SBASE 4.0
Structural domains	
IG-like repeats	1546
EGF repeats	564
Heptad repeats	377
Sushi repeats	340
FN3 repeats	284
Ank repeat	256
Annexin repeats	145
Kringle domain	106
TPR	137
SH3	95
SH2	81
Domains with biased composition	75
Ser-rich	313
Gly-rich	297
Pro-rich	228
Cys-rich	162
Acidic	6
Basic	10
Hydrophilic	92
Hydrophobic	86
Homology domains	525
Ligand binding domains	
Calcium binding	924
Zinc fingers	2305
Other DNA binding	1537
RNA binding	335
Lectin domains	102
Homeobox	313
HMG-box	67
Helix–turn–helix (HTH)	531
Helix–loop–helix (HLH)	119
Leucine zipper	183
Cell topology domains	2967
fs20 extracellular	
Transmembrane	14 475
Cytosolic	3232
Signal peptides	5638
Transit to organelles	789
Nuclear localization signals	227
Miscellaneous repeats	3765

Source and origin of data

SBASE data originate from three main sources: (i) from the SWISS-PROT protein sequence databank (8); (ii) from the Protein Sequence Database of the Protein Identification Resource (PIR) (9); (iii) from the literature. The sequences are either translated from nucleotide sequence databases (10,11) or directly keyed in at the protein level. From a total of 61 137 records in SBASE 4.0, 47 765 (78.1%), 9538 (15.6%) and 3834 (6.3%) are of eukaryotic, prokaryotic and viral origin respectively. Domain sizes vary in length between five and 1000 amino acids. The boundaries of the domains are either as previously defined in the original publications or determined by homology to domains with known boundaries.

Redundancy of sequences in SBASE 4.0 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 4.0 has cross-references to several protein and nucleic acid data banks, as well as to the PRINTS (12), PRODOM (13), BLOCKS (14) and PROSITE (15) databases (Table 3). In each record the DR lines contain the cross-reference data.

Table 3. Cross-references to other databases in SBASE

Database	Ref.	No of pointers in		
		SBASE 2.0	SBASE 3.0	SBASE 4.0
EMBL 37	10	51 555	64 074	99 275
PIR 38	9	43 855	50 132	74 403
SWISS-PROT 28	8	34 518	41 749	61 137
PRODOM 28	13		37 243	52 464
BLOCKS 8.0	14		12 483	17 245
PROSITE 12	15	6707	9307	16 029
PRINTS 9.0	12		8430	17 142
PDB	20	5438	1239	1109
MIM	21	5149	6829	8570
TFD	22	1572	1554	
FLYBASE	23	1354	1354	2321
ECOGENE	24	1216	1300	2422
SWISS-2DPAGE	25	360	182	
HIV	26	58	51	92
REBASE	27	14	7	7

Record structure

The format of SBASE 4.0 follows that of the EMBL and SWISS-PROT databases and can be directly formatted under the GCG program package (16). A sample record is shown in Figure 1. The field types used are listed in Table 4.

```
ID   ANX1_COLLI-118-178 STANDARD;      PRT;
DT   21-SEP-95 (REL. 4, 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; NEOGNAT
OC   COLUMBIFORMES.
DR   SWISS-PROT; ANX1_COLLI; P14950; AA 118-178
DR   EMBL; M22635; CLCAL.
DR   PROSITE11; IN_PDOC00195; ANNEXIN.
DR   PRINTS9.0 IN;  ANNEXIN FM (ANNEXIN3).
DR   PRINTS9.0 IN;  ANNEXIN TYPE I (ANNEXINI4).
DR   PRODOM28 IN;  28 (ANNEXIN (LIPOCORTIN I) II) (CHROMOBINDIN (PLAC.
DR   PRODOM28 CT;  28 (ANNEXIN (LIPOCORTIN I) II) (CHROMOBINDIN (PLAC.
DR   BLOCKS8.0 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).
RM   89330493 [MEDLINE , MEDLARS]
SQ   SEQUENCE    61 AA;
     LRACMKGHGT DEDTLIEILA SRNNKEIREA CRYYKEVLKR DLTQDIISDT SGDFQKALVS
     L
//
```

Figure 1. Sample entries from the SBASE 4.0 protein domain library. An annexin repeat domain. The underlined items are linked in the SBASE WWW server and so the corresponding records can be viewed on screen by 'clicking' on them.

Table 4. Types of comment lines in SBASE 4.0 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 have 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	Literature Reference Authors
RL	Reference Literature
RM	Reference to MEDLINE/MEDLARS
SQ	SeQuence

Citation

Users of SBASE and of the email servers are asked to cite this article in their publications, for example in the following form 'The sequence homologies were analysed by searching the SBASE Protein Domain Sequence Library release 4.0 via automated email server'.

DISTRIBUTION AND ACCESS

Distribution

SBASE 4.0 (21 September 1995) is distributed by anonymous ftp file transfer from ftp.icgeb.trieste.it. The complete database is 50 Mb, in compressed form 6.3 Mb.

Retrieval of records by gopher server

Individual entries are available through the gopher server 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 email server

SBASE 4.0 can be searched by the BLAST program using the email server sbase@icgeb.trieste.it. An example of a search request sent by email is presented in Figure 2A. The results of the search appear as best potential domain homologies ranked according to BLAST score (Fig. 2C). A related email server, domain@hubi.abc.hu, was created in order to assign SBASE domain homologies on the basis of BLAST searches performed on the SWISS-PROT database (17). Users can obtain all the necessary information by sending an email to sbase@icgeb.trieste.it or to domain@hubi.abc.hu with the word HELP in the body of the message.

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, PRINTS8.0, PRODOM28, PROSITE12 and SWISS-PROT29 (see underlined items in Fig. 1) can be directly accessed through the WWW server.

A

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

B

Sequences producing High-scoring Segment Pairs:		High Score	Smallest Poisson Probability 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 name	FT freq	score sum
DOMAIN ANNEXIN-REPEAT.	32	4089
DOMAIN PHYCOBILIN-LIKE.	1	46
PEPTIDE C-TERMINAL EXTENSION PEPTIDE (CTEP). (POTENTIAL).	1	40

Figure 2. Automated email servers based on the SBASE domain library. (**A**) Input file format (see the respective help files for detailed instructions). (**B**) Output of the SBASE email server (sbase@icgeb.trieste.it) in response to the annexin repeat shown in Figure 1 (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 (17). Both server outputs contain alignments provided with annotations and detailed explanation about evaluation (not shown).

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 Sequence 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 Engng*, **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 Sonnhammer,E.L.L. and Durbin,R. (1994). *Comput. Appl. Biosci.*, **10**, 301–307.
8 Bairoch,A. and Boeckmann,B. (1994) *Nucleic Acids Res.*, **22**, 3578–3580.
9 George,D.G., Barker,W.C., Mewes,H.-W., Pfeiffer,F. and Tsugita,A. (1994) *Nucleic Acids Res.*, **22**, 3569–3573.
10 Emmert,D.B., Stoehr,P.J., Stoesser,G., and Cameron,G.N. (1994) *Nucleic Acids Res.*, **22**, 3445–3449.
11 Benson,D., Boguski,M., Lipman,D.J. and Ostell,J. (1994) *Nucleic Acids Res.*, **22**, 3441–3444.
12 Attwood,T.K., Beck,M.E., Bleasby,A.J. and Parry-Smith,D.J. (1994) *Nucleic Acids Res.*, **22**, 3590–3596
13 Sonnhammer,E.L.L. and Kahn,D. (1994) *Protein Sci.*, **3**, 482–492.
14 Henikoff,S. and Henikoff,J.G. (1994) *Genomics*, **19**, 97–107.
15 Bairoch,A. and Bucher,P. (1994) *Nucleic Acids Res.*, **22**, 3583–3589.
16 Devereux,J., Haeberli,P. and Smithies,O. (1984) *Nucleic Acids Res.*, **12**, 387–395.
17 Hegyi,H. and Pongor,S. (1992) *Comput. Appl. Biosci.*, **9**, 371–372.
18 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.
19 Pearson,P., Francomano,C., Foster,P., Bocchini,C., Li,P. and McKusick,V.M. (1994) *Nucleic Acids Res.*, **22**, 3470–3473.
20 Ghosh,D. (1993) *Nucleic Acids Res.*, **21**, 3117–3118.
21 FlyBase Consortium, (1994) *Nucleic Acids Res.*, **22**, 3456–3458.
22 Rudd,K.E., Bouffard,G. and Miller,G. (1992) In Davies,K.E. and Tilghmann,S.M. (eds), *Genome Analysis.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 1–38.
23 Appel,R.D., Sanchez,J.-C., Bairoch,A., Golaz,O., Ravier,F., Pasquali,C., Hughes,G.J. and Hochstrasser,D.F. (1994) *Nucleic Acids Res.*, **22**, 3581–3582.
24 Myers,F. (1990) Human Retrovirus and AIDS Database, Los Alamos National Laboratory, Los Alamos, NM.
25 Roberts,R.J. and Macelis,D. (1994) *Nucleic Acids Res.*, **22**, 3628–3639.