Nucleic Acids Research, 1999, Vol. 27, No. 1 257–259

The SBASE protein domain library, release 6.0: a collection of annotated protein sequence segments

János Murvai[1], Kristian Vlahovicek[1], Endre Barta[2], Csaba Szepesvári[3], Cristina Acatrinei[1] and Sándor Pongor[1,2,*]

[1]International Centre for Genetic Engineering and Biotechnology, Area Science Park, 34012 Trieste, Italy, [2]ABC Institute for Biochemistry and Protein Research, 2100 Gödöllö, Hungary and [3]Research Group on Artificial Intelligence, József Attila Universisty, 6700 Szeged, Hungary

Received October 2, 1998; Accepted October 7, 1998

ABSTRACT

The sixth release of the SBASE protein domain library sequences contains 130 703 annotated and crossreferenced entries corresponding to structural, functional, ligand-binding and topogenic segments of proteins. The entries were grouped based on standard names (2312 groups) and futher classified on the basis of the BLAST similarity (2463 clusters). Automated searching with BLAST and a new sequence-plot representation of local domain similarities are available at the WWW-server http://www.icgeb.trieste.it/sbase . A mirror site is at http://sbase.abc.hu/sbase . The database is freely available by anonymous 'ftp' file transfer from ftp.icgeb.trieste.it

INTRODUCTION

Detection of domains in newly determined sequences is usually based on pattern collections that contain consensus representation domain types deduced from multiple alignments. Consensus descriptions come in different varieties such as regular expressions, sequence profiles, hidden Markov models, etc. Development of such a consesnsus description requires expertise and careful judgement hence pattern collections can hardly keep pace with the flow of new genome data. Another problem is the inevitable statistical bias of the consensus. Namely, atypical domains for which there are too few known examples, may not fit well with a consensus pattern developed with a numerous dataset of similar domains. Finally, there are domain types for which it is not easy to develop consensus representations because of weak similarity.

SBASE is a collection of protein domain sequences designed to facilitate detection domain homologies without the above problems (1,2). Here the method of domain recognition is database search rather than pattern search, so atypical and typical domains are equally well recognized. The underlying database, SBASE is preprocessed by BLAST similarity search (3) and the similarity groups (that can be best pictured as densely connected graphs) form the basis of domain recognition.

Table 1. Increase of data in SBASE 6.0

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-JUNE-95	61,137 (+46 %)	3,281,782 (+40%)	(50 Mb) (+34%)
5.0	06-OCT-96	79,862 (+30%)	4,118,506 (25%)	75 MB (+50%)
6.0	23-OCT-98	130,703 (63%)	10,457,771 (154%)	115 (53%)

The current release 6.0 of SBASE contains over 100 000 annotated protein sequence segments consistently named by structure, function, biased composition, binding-specificity and/or similarity to other proteins.

The main developments with respect to the previous release can be summarized as follows. (i) Release 6.0 contains 130 703 sequence entries, 63% more than release 5.0 (Table 1). (ii) All records are now provided with standard names and an effort was made to use domain names also used by other squence databases and pattern collections like Prosite (4) and PFAM (5). (iii) The entries were grouped based on standard names (2312 groups) and those with at least three entries (1039 groups) were futher classified on the basis of the BLAST similarity. A total of 2463 clusters with at least three members are deposited into a separate database, SBASE-CLUSTERS, which is now available through anonymous ftp as well as through links on the WWW-server (a description of the clustering procedure is given at the web-site). Within each standard name group the clusters are numbered, in such a way that clusters with more inter-member similarity have larger numbers. (iv) A new graphic output facility is added to the server whereby local domain similarity can be plotted along the sequence.

DESCRIPTION OF THE DATA

Definition of protein domains

Domains included in SBASE are protein sequence segments with known structure and/or function. The main entry classes are summarized in Table 2. The boundaries of the domains are either

*To whom correspondence should be addressed at: ICGEB, Area Science Park, 34012 Trieste, Italy. Tel: +39 040 375 7300; Fax: +39 040 226 555; Email: pongor@icgeb.trieste.it

Table 2. Examples of domains in SBASE 6.0

Domain type	Number of records in SBASE 6.0	Domain type	Number of records in SBASE 5.0
STRUCTURAL DOMAINS		**HOMOLOGY DOMAINS**	
IG-like repeats	1942	**LIGAND-BINDING DOMAINS**	
EGF-repeats	1431	Calcium-binding	1214
Heptad-repeats	678	Zinc-fingers	2478
Sushi repeats	366	Other DNA-binding	6083
FN3-repeats	615	RNA-binding	585
Ank-repeat	486	Lectin domains	264
Annexin-repeats	198	Homeobox	538
Kringle domain	163	HMG-box	163
TPR	192	Helix-turn-helix (HTH)	806
SH3	198	Helix-loop-helix (HLH)	178
SH2	178	Leucine-zipper	260
Domains with biased composition		**CELL TOPOLOGY DOMAINS**	
Ser-rich	939	Extracellular	5429
Gly-rich	884	Transmembrane	32 474
Pro-rich	631	Cytosolic	6376
Cys-rich	272	Signal peptides	6632
Acidic	663	Transit to organelles	1149
Basic	487	Nuclear localization signals	538
Hydrophilic	136		
Hydrophobic	159	**MISCELLANEOUS REPEATS**	4591

Table 3. Cross-references to other databases in SBASE

DATABASE	Ref.	No of pointers in				
		SBASE 2.0	SBASE 3.0	SBASE 4.0	SBASE5.0	SBASE 6.0
EMBL	(13)	51,555	64,074	99,275	137,117	259,398
PIR International	(7)	43,855	50,132	74,403	84,991	116,657
SWISS-PROT	(6)	34,518	41,749	61137	79,863	130703
PRODOM	(9)	-	37,243	52,464	54,510	83,008
BLOCKS	(9)	-	12,483	17,245	26,930	64,220
PROSITE	(4)	6,707	9,307	16,029	26,384	54,246
PRINTS	(8)	-	8,430	17,142	26,384	77,587
PDB	(14)	5 ,438	1,239	1,109	3,995	7,123
MIM	(15)	5,149	6,829	8,570	11,161	17,554
FLYBASE	(16)	1,354	1,354	2,321	2,881	4,317
ECOGENE	(17)	1,216	1,300	2,422	4,442	5,583
HIV	(18)	58	51	92	92	1,769
REBASE	(19)	14	7	7	10	58

as previously defined in the original publications or determined by homology to domains with known boundaries. In this release, the boundaries used by PFAM (5) were adopted for a number of domain types.

Source and origin of data

SBASE data originate from three main sources: (i) from the SWISS-PROT protein sequence databank (6); (ii) from the Protein Sequence Database of the PIR International Protein sequence database (PIR) (7); and (iii) from the literature. From a total of 130 703 records in SBASE 6.0, 96 305 (73%), 27 089 (21%) and 6656 (5%) are of eukaryotic, prokaryotic and viral origin, respectively. Domain sizes vary in length between 5 and 1000 amino acids.

Redundancy of sequences in SBASE 6.0 is kept at a minimal level. In some cases, the domain definitions overlap.

Cross-references

SBASE 6.0 has cross-references to several protein and nucleic acid databanks, as well as to the PROSITE (4), PRINTS (8), PRODOM (9) and BLOCKS (10) databases (Table 3). In each record, the DR-lines contain the cross-reference data.

Record structure

The format of SBASE 6.0 (Fig. 1) follows that of the EMBL and SWISS-PROT databases and can be directly formatted under the GCG package The field types used are listed in Table 4. The

```
ID   ANX1_COLLI-126-178
DT   1-JUL-98 (REL. 7, CREATED)
SN   ANNEXINS.
DE   ANNEXINS ANNEXIN I (LIPOCORTIN I) (CALPACTIN II) (CHROMOBINDIN 9) (P35)
DE   (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 126-178
DR   EMBL; M22635; G213534; -.
DR   PIR; A40153; LUPY1.
DR   PROSITE_IN; PDOC00195; ANNEXIN.
DR   PRINTS16.0 NT;  ANNEXIN FM (ANNEXIN3)
DR   PRINTS16.0 NT;  ANNEXIN TYPE I (ANNEXINI4)
DR   PRODOM34 IN;  19 (ANNEXIN (LIPOCORTIN I) II) (CHROMOBINDIN (PLAC
DR   BLOCKS9.3 NT;  BL00223B Annexins repeat proteins domain proteins.
RA   HORSEMAN N.D.;
RL   MOL. ENDOCRINOL. 3:773-779(1989).
CL   ANNEXINS./8
SQ   SEQUENCE 53 AA
     GTDEDTLIEI LASRNNKEIR EACRYYKEVL KRDLTQDIIS DTSGDFQKAL
     VSL
//
```

Figure 1. A sample entry from the SBASE 6.0 protein domain library. An annexin repeat domain. The underlined items are linked in the SBASE World Wide Web server so that the corresponding records can be viewed on the screen by 'clicking' on them.

Table 4. Types of comment lines in SBASE 6.0 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	**DaTe** 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	**S**tandard **N**ame/**CL**uster number
RA	**A**uthors of the literature **R**eference.
RL	**L**iterature **R**eference.
RM	**R**eference to **M**EDLINE/MEDLARS
SQ	**SeQuence**

clusters to which a sequence belongs are determined by (i) the standard name and (ii) the (optional) subclass number included in the CL field, e.g. ANNEXINS/8 (the CE field of previous releases is now abandoned).

DISTRIBUTION AND ACCESS

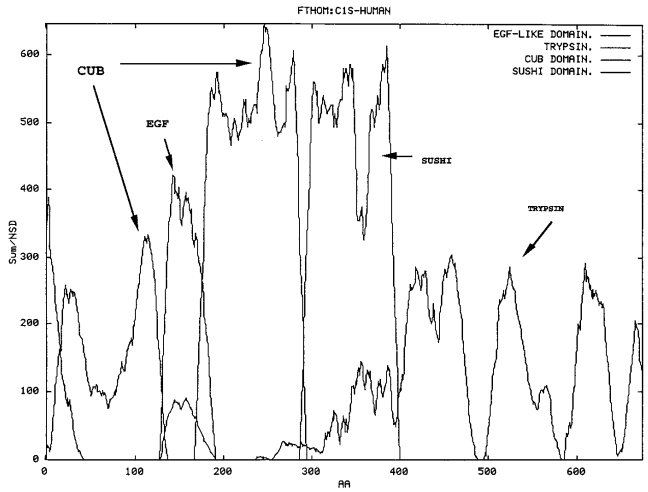
Distribution

SBASE 6.0 (23 October, 1998) 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.

Access by WWW: record retrieval and BLAST search

SBASE 6.0 and SBASE-CLUSTERS can be searched at the WWW-server http://base.icgeb.trieste.it/sbase and at the mirror site http://sbase.abc.hu/sbase . Record retrieval is with the SRS system. At present, cross-references to SBASE-CLUSTERS, EMBL, MEDLINE, MIM, PRINTS, PRODOM, PROSITE and SWISS-PROT can be directly accessed through the WWW-server. Prediction of domain homologies via BLAST searching is possible either by (i) running a search against SBASE, or (ii) running a search against SWISS-PROT and reprocessing the search output (11,12). In the output of the latter, local domain similarities are also graphically represented as a sequence-plot (Fig. 2).

A



B

Feature name	NSD	GN.	Sum.Score	Sum/NSD	Overlap Max.
ANNEXINS.	43	198	6112	142.1	213
EXTRACELLULAR.	3	4409	123	41.0	53
GLUTAMATE / LEUCINE / PHENYLALANINE / VALINE DEHYDROGENASES.	2	42	92	46.0	52
PYRIDOXAL-DEPENDENT DECARBOXYLASES CONSERVED DOMAIN.	2	42	88	44.0	44

Figure 2. (**A**) Graphic output of the domain similarity server (www.icgeb.trieste.it/sbase) in response to the query sequence C1S_HUMAN from SWISS-PROT. The known domain structure of this query is CUB-EGF-CUB-SUSHI-SUSHI-SPR (where S = signal, P = propeptide, SPR = serine protease). The ouput shows the plot of the BLAST similarities along with the SBASE standard names. Arrows have been added to help identification in black and white (original is in color). (**B**) Output of the domain homology WWW server (www.icgeb.trieste.it/sbase) in response to the annexin sequence shown in Figure 1 (detail). NSD: number of significant similarities found in the BLAST output; GN.: number of the given domain occuring in the database; Sum.Score: cumulative sum of BLAST scores belonging to a domain-name in the output; Overlap Max: maximum similarity score found (11). The server output contains alignments provided with annotations and a detailed explanation about evaluation (not shown).

Citation

Users of SBASE and of the WWW/Email servers are asked to cite this article in their publications.

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. The authors wish to thank the support of EMBnet, the European Molecular Biology Network. The Protein Structure and Function Group is supported by EMBnet in the framework of EU grant ERB-BIO4-CT96-0030. Work at ABC was supported by ICGEB collaborative research grant no CRP/HUN9603.

REFERENCES

1 Pongor,S., Skerl,V., Cserzo,M., Hatsagi,Z., Simon,G. and Bevilacqua,V. (1993) *Protein Engng.*, **6**, 391–395.
2 Fabian,P., Murvai,J., Hatsagi,Z., Vlahovicek,K., Hegyi,H. and Pongor,S. (1997) *Nucleic Acids Res.*, **25**, 240–243.
3 Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res.*, **25**, 3389–3402.
4 Bairoch,A., Bucher,P. and Hofmann,K. (1996) *Nucleic Acids Res.*, **24**, 189–196.
5 Sonnhammer,E.L., Eddy,S.R., Birney,E., Bateman,A. and Durbin,R. (1998) *Nucleic Acids Res.*, **26**, 320–322.
6 Bairoch,A. and Apweiler,R. (1998) *Nucleic Acids Res.*, **26**, 38–42.
7 Barker,W.C., Garavelli,J.S., Haft,D.H., Hunt,L.T., Marzec,C.R., Orcutt,B.C., Srinivasarao,G.Y., Yeh,L.S.L., Ledley,R.S., Mewes,H.W., Pfeiffer,F. and Tsugita,A. (1998) *Nucleic Acids Res.*, **26**, 27–32.
8 Attwood,T.K., Beck,M.E., Flower,D.R., Scordis,P. and Selley,J.N. (1998) *Nucleic Acids Res.*, **26**, 304–308.
9 Corpet,F., Gouzy,J. and Kahn,D. (1998) *Nucleic Acids Res.*, **26**, 323–326.
10 Henikoff,S., Pietrokovski,S. and Henikoff,J.G. (1998) *Nucleic Acids Res.*, **26**, 309–312.
11 Murvai,J., Vlahovicek,K., Barta,E., PFeiffer,F., Hegyi,H. and Pongor,S. (1998) *Bioinformatics*, in press.
12 Hegyi,H. and Pongor,S. (1993) *Comput. Applic. Biosci.*, **9**, 371–372.
13 Stoesser,G., Moseley,M.A., Sleep,J., McGowran,M., Garcia-Pastor,M. and Sterk,P. (1998) *Nucleic Acids Res.*, **26**, 8–15.
14 Bernstein,F.C., Koetzle,T.F., Williams,G.J., Meyer,E.E.,Jr, Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) *J. Mol. Biol.*, **112**, 535–542.
15 Pearson,P., Francomano,C., Foster,P., Bocchini,C., Li,P. and McKusick,V. (1994) *Nucleic Acids Res.*, **22**, 3470–3473.
16 Flybase Consortium (1998) *Nucleic Acids Res.*, **26**, 85–88.
17 Rudd,K.E., Bouffard,G. and Miller,G. (1992) In Davies,K.E. and Tilghman,S.M. (eds), *Genome Analysis*. Cold Spring Harbor Laboratory Press, New York, pp. 1–38.
18 Myers,F. (1990) *Human Retrovirus and Aids Database*. Los Alamos National Laboratory, Los Alamos, NM, USA.
19 Roberts,R.J. and Macelis,D. (1998) *Nucleic Acids Res.*, **26**, 338–350.