260–262 Nucleic Acids Research, 2000, Vol. 28, No. 1

The SBASE protein domain library, release 7.0: a collection of annotated protein sequence segments

János Murvai[1], Kristian Vlahovicek[1], Endre Barta[2], Bruno Cataletto[1] and Sándor Pongor[1,2,*]

[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

Received September 29, 1999; Revised and Accepted October 15, 1999

ABSTRACT

SBASE 7.0 is the seventh release of the SBASE protein domain library sequences that contains 237 937 annotated structural, functional, ligand-binding and topogenic segments of proteins, cross-referenced to all major sequence databases and sequence pattern collections. The entries are clustered into over 1811 groups and are provided with two WWW-based search facilities for on-line use. SBASE 7.0 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 servers http://www.icgeb.trieste.it/sbase/ and http://sbase.abc.hu/sbase/

INTRODUCTION

Prediction of domains is usually based on pattern collections that contain consensus representations of domain types deduced from multiple alignments. Consensus representation of sequences (such as consensus sequences, regular expressions, sequence profiles, hidden Markov models, etc.) requires human 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 consensus representations. Namely, reliable multiple alignments require a good number of domain examples, and, as a consequence, atypical domains for which there are too few known examples, may be difficult to recognize. Finally, there are domain types for which it is not easy to develop consensus representations because of the weak similarity.

SBASE is a collection of protein domain sequences designed to facilitate detection of domain homologies without the above problems (1,2). The method of domain recognition is database search rather than pattern search, so atypical and typical domains are equally well recognized (3). The central concept is the 'similarity group', i.e. a group of domain sequences that have BLAST similarity to each other. One can distinguish tight and loose groups depending on how many significant similarity connections exist, on average, between the members. Briefly, a new sequence is considered member of a given group if its similarity parameters (3) are above the threshold levels automatically established for that group, and if it has no sequential overlap with any other domain group. Validated domain groups i.e. the 1550 groups that satisfy these criteria are deposited in SBASE-A; these are the well-known structural and functional domain types. SBASE-B contains a 261 groups that are either (i) less well characterized than the groups of SBASE-A, or (ii) are defined by composition (e.g. glycine-rich), cellular location (e.g. transmembrane, etc.). These groups are sometimes defined in an overlapping manner, e.g. an extracellular domain (SBASE-B) may contain an EGF-module (SBASE-A).

The current release 7.0 of SBASE contains over 230 000 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 [release 6] can be summarized as follows:
(i) Release 7.0 contains 237 937 sequence entries, 82% more than release 6.0 (Table 1).

Table 1. Increase of data in SBASE release 7.0

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-JUN-95	61,137	3,281,782	50
		(+48%)	(+40%)	(+34%)
5.0	06-OCT-96	79,862	4,118,506	75
		(+30%)	(+25%)	(+50%)
6.0	23-OCT-98	130,703	10,457,771	115
		(+63%)	(+154%)	(+53%)
7.0	23-OCT-99	237,937	18,964,500	221
		(+82%)	(+81%)	(+92%)

(ii) The entries were grouped by standard names and further classified on the basis of the BLAST similarity scores. The list of all clusters having at least two members is deposited into a separate database, SBASE-CLUSTERS, which is now available through anonymous ftp as well as through links on the WWW-server. A total of 1811 domain groups were found, of which 1550 validated groups (1936 clusters) are in SBASE-A and 261 groups (382 clusters) are in SBASE-B. The clusters are identified by the standard name and by the (optional) subclass number included in the SC field. The CL and CE fields of previous releases are now abandoned.

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

*To whom correspondence should be addressed at: International Centre for Genetic Engineering and Biotechnology, Area Science Park, 34012 Trieste, Italy. Tel: +39 040 3757 300; Fax: +39 040 226 555; Email: pongor@icgeb.trieste.it

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 such as given in the PROT-FAM (4) and in the PFAM databases (5).

Table 2. Examples of domains in SBASE 7.0

Domain type	Number of records in SBASE 7.0	Domain type	Number of records in SBASE 7.0
STRUCTURAL DOMAINS		**HOMOLOGY DOMAINS**	
IG-like repeats	3143	**LIGAND-BINDING DOMAINS**	
EGF-repeats	1419	Calcium-binding	2820
Heptad-repeats	1237	Zinc-fingers	5827
Sushi repeats	557	DNA-binding	8039
FN3-repeats	955	RNA-binding	783
Ank-repeat	579	Lectin domains	426
Annexin-repeats	262	Homeobox	971
Kringle domain	172	HMG-box	641
TPR	64	Helix-turn-helix (HTH)	1289
SH3	281	Helix-loop-helix (HLH)	335
SH2	235	Leucine-zipper	513
Domains with biased composition		**CELL TOPOLOGY DOMAINS**	
Ser-rich	1426	Extracellular	7850
Gly-rich	1206	Transmembrane	69836
Pro-rich	1151	Cytosolic	9100
Cys-rich	329	Signal peptides	17039
Acidic	115	Transit to organelles	2447
Basic	442	Nuclear localization signals	229
Hydrophilic	136		
Hydrophobic	393	**MISCELLANEOUS REPEATS**	1406

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 Protein Identification Resource (PIR International) (7); and (iii) from the literature. The sequences are either translated from nucleotide sequence databases (8,9) or directly keyed in at the protein level. From a total of 237 937 records in SBASE 7.0, 136 367 (57%), 53 307 (22.4%) and 38 083 (20.6%) are of eukaryotic, prokaryotic and viral origin, respectively. Domain sizes vary in length between 5 and 1000 amino acids.

Cross-references

SBASE 7.0 has cross-references to several protein and nucleic acid databanks, as well as to the PROSITE (10) PRINTS (11), ProDom (12), BLOCKS (13) and PFAM (5) domain databases, the Protein Structure Data Bank (14) and the database of human Mendelian inheritance (15) (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	SBASE5.0	SBASE6.0	SBASE7.0
EMBL (9)	51,555	64,074	99,275	137,117	259,398	435,109
PIR-Int'l (7)	43,855	50,132	74,403	84,991	116,657	180,741
SWISSPROT (6)	34,518	41,749	61137	79,863	130,703	192,668
PRODOM (12)	-	37,243	52,464	54,510	83,003	252,900
BLOCKS (13)	-	12,483	17,245	26,930	64,220	54,821
PROSITE (10)	6,707	9,307	16,029	26,384	54,246	83,501
PRINTS (11)	-	8,430	17,142	26,384	77,587	42,623
PDB (14)	5 ,438	1,239	1,109	3,995	7,123	10,872
MIM (15)	5,149	6,829	8,570	11,161	17,554	26,255
FLYBASE (16)	1,354	1,354	2,321	2,881	4,317	5,704
ECOGENE (17)	1,216	1,300	2,422	4,442	5,583	8,816
HIV (18)	58	51	92	92	1,769	3,467
REBASE (19)	14	7	7	10	58	84
PFAM (5)	0	0	0	0	0	118804

Record structure

The format of SBASE 7.0 follows that of the EMBL and SWISS-PROT databases and can be directly formatted under the GCG program package using (16).

DISTRIBUTION AND ACCESS

Distribution

SBASE 7.0 (6 October, 1999) is distributed by anonymous 'ftp' file transfer from ftp.icgeb.trieste.it . The complete database (including the records and list of clusters), is 221 Mb, its compressed form is 16.3 Mb.

BLAST search by WWW-server

SBASE 7.0 can be searched by the BLAST program using the WWW-server http://www.icgeb.trieste.it/sbase . A related server was created in order to assign SBASE domain homologies on the basis of BLAST searches performed on the SWISS-PROT database and on the PIR International databases (7). This service (available at http://www.abc.hu/blast.html and at domain@abc.hu) returns the best potential domain homologies ranked according to BLAST score.

Access by WWW-server

Record retrieval and the above services can be accessed also using the WWW-server at http://www.icgeb.trieste.it/sbase . At present, cross-references to SBASE-CLUSTERS, EMBL, MEDLINE, MIM, PRINTS, ProDom, PROSITE and SWISS-PROT can be directly accessed through the WWW-server.

Citation

Users of SBASE and of the WWW 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 7.0' via automated electronic mail (WWW) server'.

ACKNOWLEDGEMENTS

This work was supported in part by EMBnet, the European Molecular Biology Network in the framework of EU grant ERBBIO4-CT96-0030. SBASE was established in 1990 and is maintained collaboratively by the International Center for Genetic Engineering and Biotechnology, Trieste, Italy and the Agricultural Biotechnology Center, Gödöllö, Hungary. The help of Suzanne Kerbavcic with the manuscript is gratefully acknowledged.

REFERENCES

1. Pongor,S., Skerl,V., Cserzo,M., Hatsagi,Z., Simon,G. and Bevilacqua,V. (1993) *Protein Eng.*, **6**, 391–395.
2. Murvai,J., Vlahovicek,K., Barta,E., Szepesvári,C., Acatrinei,C. and Pongor,S. (1999) *Nucleic Acids Res.*, **27**, 257–259.
3. Murvai,J., Vlahovicek,K., Barta,E., Parthasarathy,S., Hegyi,H., Pfeiffer,F. and Pongor,S. (1999) *Bioinformatics*, **15**, 343–344.
4. Mewes,H.W., Heumann,K., Kaps,A., Mayer,K., Pfeiffer,F., Stocker,S. and Frishman,D. (1999) *Nucleic Acids Res.*, **27**, 44–48. Updated article in this issue: *Nucleic Acids Res*. (2000), **28**, 37–40.
5. Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Finn,R.D. and Sonnhammer,E.L. (1999) *Nucleic Acids Res.*, **27**, 260–262. Updated article in this issue: *Nucleic Acids Res*. (2000), **28**, 263–266.
6. Bairoch,A. and Apweiler,R. (1999) *Nucleic Acids Res.*, **27**, 49–54. Updated article in this issue: *Nucleic Acids Res*. (2000), **28**, 45–48.
7. Barker,W.C., Garavelli,J.S., McGarvey,P.B., Marzec,C.R., Orcutt,B.C., Srinivasarao,G.Y., Yeh,L.S., Ledley,R.S., Mewes,H.W., Pfeiffer,F., Tsugita,A. and Wu,C. (1999) *Nucleic Acids Res.*, **27**, 39–43.

8. Benson,D.A., Boguski,M.S., Lipman,D.J., Ostell,J., Ouellette,B.F., Rapp,B.A. and Wheeler,D.L. (1999) *Nucleic Acids Res.*, **27**, 12–17. Updated article in this issue: *Nucleic Acids Res*. (2000), **28**, 15–18.
9. Stoesser,G., Tuli,M.A., Lopez,R. and Sterk,P. (1999) *Nucleic Acids Res.*, **27**, 18–24. Updated article in this issue: *Nucleic Acids Res*. (2000), **28**, 19–23.
10. Hofmann,K., Bucher,P., Falquet,L. and Bairoch,A. (1999) *Nucleic Acids Res.*, **27**, 215–219.
11. Attwood,T.K., Flower,D.R., Lewis,A.P., Mabey,J.E., Morgan,S.R., Scordis,P., Selley,J.N. and Wright,W. (1999) *Nucleic Acids Res.*, **27**, 220–225. Updated article in this issue: *Nucleic Acids Res*. (2000), **28**, 225–227.
12. Corpet,F., Gouzy,J. and Kahn,D. (1999) *Nucleic Acids Res.*, **27**, 263–267. Updated article in this issue: *Nucleic Acids Res*. (2000), **28**, 267–269.
13. Henikoff,J.G., Henikoff,S. and Pietrokovski,S. (1999) *Nucleic Acids Res.*, **27**, 226–228. Updated article in this issue: *Nucleic Acids Res*. (2000), **28**, 228–230.
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 (1999) *Nucleic Acids Res.*, **27**, 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, Cold Spring Harbor, NY, pp. 1–38.
18. Roberts,R.J. and Macelis,D. (1999) *Nucleic Acids Res.*, **27**, 312–313. Updated article in this issue: *Nucleic Acids Res*. (2000), **28**, 306–307.