Nucleic Acids Research, 2005, Vol. 33, Database issue ***D223–D225***
doi:10.1093/nar/gki112

The SBASE domain sequence resource, release 12: prediction of protein domain-architecture using support vector machines

Kristian Vlahoviček, László Kaján, Vilmos Ágoston[1] and Sándor Pongor*

ICGEB–International Center for Genetic Engineering and Biotechnology, Area Science Park, 34012 Trieste, Italy and [1]Bioinformatics Group, Biological Research Center of Hungarian Academy Sciences, H-6726 Szeged, Temesvári krt 62, Hungary

Received September 14, 2004; Revised and Accepted October 18, 2004

ABSTRACT

SBASE (http://www.icgeb.trieste.it/sbase) is an online resource designed to facilitate the detection of domain homologies based on sequence database search. The present release of the SBASE A library of protein domain sequences contains 972 397 protein sequence segments annotated by structure, function, ligand-binding or cellular topology, clustered into 8547 domain groups. SBASE B contains 169 916 domain sequences clustered into 2526 less well-characterized groups. Domain prediction is based on an evaluation of database search results in comparison with a 'similarity network' of inter-sequence similarity scores, using support vector machines trained on similarity search results of known domains.

INTRODUCTION

The SBASE project was initiated in order to develop a prediction scheme that can automatically recognize instances of known protein domains in the newly determined sequences, using similarity search on a reference domain sequence database (1–3). One of the project's main motivations has been to solve the prediction problem without the use of a manmade model or consensus description of domain sequence groups, in order to decrease maintenance costs while maintaining the generalization power of the prediction. The resulting system consists of a reference domain sequence database on one hand, and a related predictor program on the other, so the system's predicting power can be optimized by tuning both these components in concert.

SBASE 12.0 is a collection of 972, 397 protein domain sequences. Each SBASE domain record contains a sequence assigned to one of the 8 547 functionally or structurally well-characterized groups (SBASE A), or to one of the less well-characterized 2 018 groups described in terms of amino acid composition or cellular localization (SBASE B). All domains are cross-referenced back to their parent protein databases [Swiss-Prot + TrEMBL (4), PIR (5) and to entries in other domain repositories, such as INTERPRO (6), or its member databases such as Pfam (7), SMART (8) and PRINTS (9)].

Finding known domain types in new sequences includes two subtasks: (i) locating the potential domains—in the SBASE system this problem has been approached by analyzing the distribution of cumulative FASTA or BLAST similarity scores along the query sequence (10,11); and (ii) selecting/accepting the best candidate domains. This $task_{[lkl]}$ is a classification problem that was initially solved using significance values (11). In a subsequently developed analysis scheme, a database versus database comparison was used to create a similarity network in which the nodes are domain sequences and the (weighted) edges are similarity scores (12,13). In the resulting predictor algorithm each domain group was characterized by two variables: the average number of similarities above a selected threshold (NSD) and the average similarity score (AVS), which, in graph theory correspond to the terms 'degree' and 'average weight', respectively (12,13). Group-specific threshold values were calculated for both variables and the classification was based on a probabilistic score, which was calculated from the threshold values as well as measures derived from the distribution of the two characteristic parameters (NSD and AVS), for each domain group (14). Even though the system gave reassuring results in most groups (3,13), there were a number of persistent mispredictions that could not be eliminated by the optimization of threshold values.

In the present release of the system we introduce, in addition to the BLAST score and the degree (NSD), new variables, namely, (i) HSP length (alignment length) determined for the subject, (ii) score coverage, i.e. the similarity score divided by the self-similarity score of the subject (database entry);

*To whom correspondence should be addressed. Tel: +39 040 3757300; Fax: +39 040 226 555; Email: pongor@icgeb.org

(iii) length coverage, i.e. the length of the aligned region (HSP length determined for the subject) divided by the subject length and (iv) length-to-score ratio, the length of the subject divided by the similarity score. The similarity of a query to a group is characterized by the average of these variables calculated from the BLAST alignments made between the query and the members of the group. The averaged variables are quite efficient in clustering domain sequences using BLAST searches. For instance, over 92% of the groups in the PFAM-SEED database (7) were completely separated from the non-group member neighbors by at least one of the variables (Figure 1). In order to get a robust separation of the sequence clusters we trained support vector machines (SVMs) with the linear kernel and the variables mentioned above, using the SVM utilities of the R package (www.r-project.org). Benchmark SVM training was based on comparing the non-redundant members of the PFAM-SEED 8.0 (128 780 sequences) to a set of their parent proteins (94 102 sequences) using a BLAST score cutoff of 40, and recording 'good' or 'bad' hits if 20% of an HSP overlapped respectively, with a domain of corresponding domain type or different domain type. The training took 10 h on a dual-processor (1400 MHz) AMD Opteron 240 machine. The predictive performance of the resulting system is shown by the fact that, if PFAM-SEED 8.0 is used as the reference database, over 60% of the groups had none or only one mistaken prediction (Table 1), and the average difference in the domain boundaries is <5 amino acids in 90% of the cases (Figure 2).

IMPROVEMENTS WITH RESPECT TO THE PREVIOUS RELEASE

(i) The examples included in the consolidated domain sequence collection SBASE A were filtered so as to discard conspicuously short or long domain examples. The consolidated sets were used to trainSVMs.

(ii) SBASE A group names have been renamed so as to match INTERPRO names. Domains, families and repeats are now classified into separate categories.

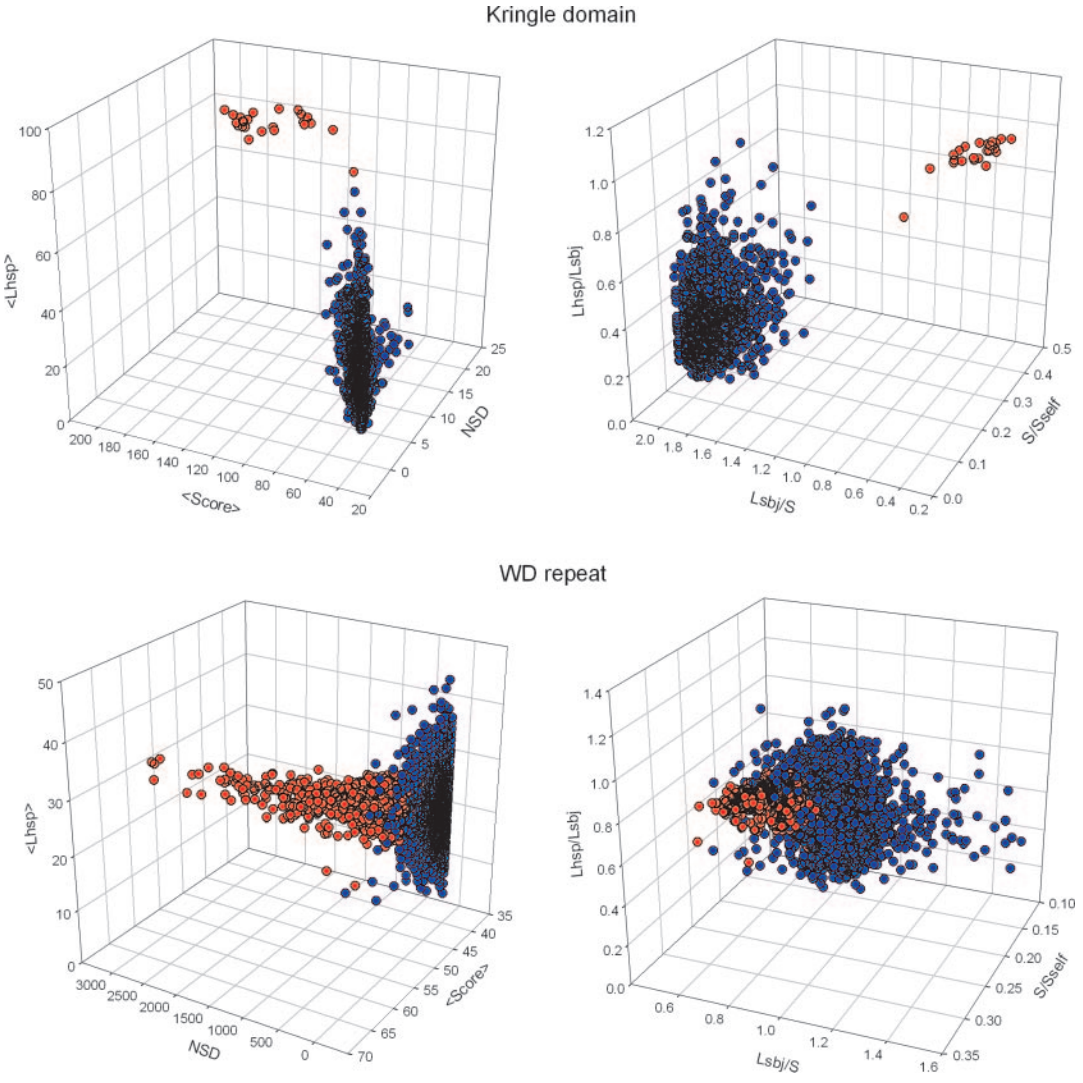

Figure 1. Separation of domain group members from neighbors in three dimensions. The kringle group is one of the perfectly separated groups, WD repeat is one of the critical cases (L_{hsp} = length of HSP, L_{sbj} = length of subject (database entry), S/S_{self} = score coverage; see text for explanations).

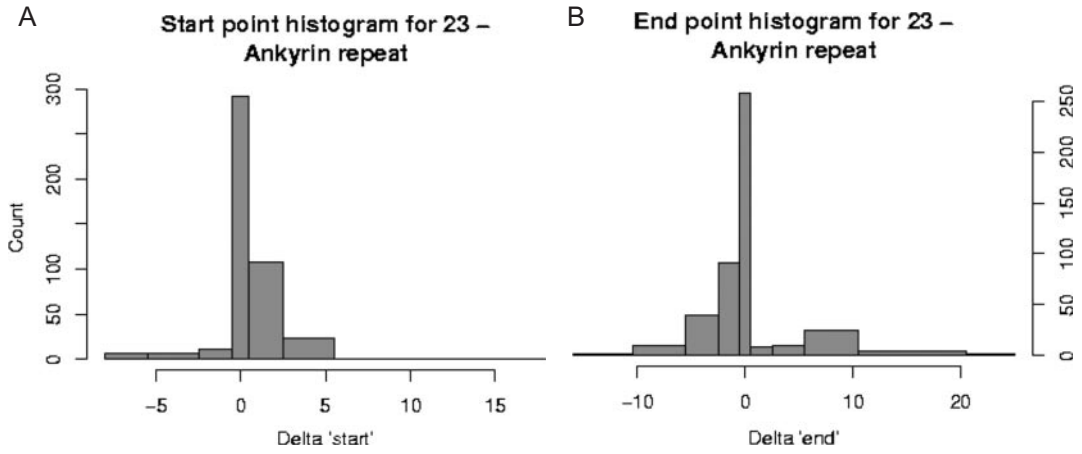

Figure 2. Domain boundary prediction statistics available at the SBASE homepage.

Table 1. SVM benchmark figures[a]

Domain group	No. of sequences		Match	Mismatch	Unpredicted
	Learning set	Test set			
Kringle domain	24	9	9	0	0
Fibronectin type III domain	108	352	328	2	22
WD repeat	1924	673	542	12	125
EGF-like domain	87	290	262	13	15
Protein kinase domain	67	545	505	5	35
Annexin repeat	181	34	32	1	2
Sushi domain	80	119	103	0	3
Trypsin family	83	128	110	0	1
Globin family	79	59	57	1	0
ABC transporter domain	63	564	563	1	0
Ank repeat	1195	736	535	5	196
Total	128 780	60 457	56 891	238	3328

[a]The learning set consisted of the parent protein sequences of domains in PFAM-SEED 8.0. The test included parent proteins with annotated domains not included in PFAM-SEED.

(iii) The SVMs were included into the predictor algorithm. Evaluation of a protein sequence query typically takes 10 s including 5 s of the BLAST run.

DISTRIBUTION AND ACCESS

The SBASE domain library browser and domain architecture prediction system are accessible through the web-interface at http://www.icgeb.org/sbase.

ACKNOWLEDGEMENTS

SBASE was established in 1990 and is maintained collaboratively by ICGEB, Trieste and the Biological Research Center of the Hungarian Academy of Sciences, (Szeged, Hungary). This work was supported by the grant BIO-00001/2002 as well as by an Albert Szent-Györgyi Award to S.P. (Hungarian Ministry of Education)

REFERENCES

1. Simon,G., Paladini,R., Tisminetzky,S., Cserzö,M., Hátsági,Z., Tossi,A. and Pongor,S. (1992) Improved detection of homology in distantly related proteins: similarity of adducin with actin-binding proteins. *Protein Seq. Data Anal.*, **5**, 39–42.
2. Pongor,S., Skerl,V., Cserzo,M., Hatsagi,Z., Simon,G. and Bevilacqua,V. (1993) The SBASE domain library: a collection of annotated protein segments. *Protein Eng.*, **6**, 391–395.
3. Vlahovicek,K., Kajan,L., Murvai,J., Hegedus,Z. and Pongor,S. (2003) The SBASE domain sequence library, release 10: domain architecture prediction. *Nucleic Acids Res.*, **31**, 403–405.
4. Apweiler,R., Bairoch,A., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.*, **32**, D115–D119.
5. Wu,C.H., Nikolskaya,A., Huang,H., Yeh,L.S., Natale,D.A., Vinayaka,C.R., Hu,Z.Z., Mazumder,R., Kumar,S., Kourtesis,P. *et al.* (2004) PIRSF: family classification system at the Protein Information Resource. *Nucleic Acids Res.*, **32**, D112–D114.
6. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al.* (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.*, **31**, 315–318.
7. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
8. Letunic,I., Copley,R.R., Schmidt,S., Ciccarelli,F.D., Doerks,T., Schultz,J., Ponting,C.P. and Bork,P. (2004) SMART 4.0: towards genomic data integration. *Nucleic Acids Res.*, **32**, D142–D144.
9. Attwood,T.K., Bradley,P., Flower,D.R., Gaulton,A., Maudling,N., Mitchell,A.L., Moulton,G., Nordle,A., Paine,K., Taylor,P. *et al.* (2003) PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res.*, **31**, 400–402.
10. Hegyi,H. and Pongor,S. (1993) Predicting potential domain homologies from FASTA search results. *Comput. Appl. Biosci.*, **9**, 371–372.
11. Murvai,J., Vlahovicek,K., Barta,E., Pfeiffer,F., Hegyi,H. and Pongor,S. (1998) The Domain-server: direct prediction of protein domain-homologies from BLAST search. *Bioinformatics*, **4**, 343–344.
12. Murvai,J., Vlahovicek,K. and Pongor,S. (2001) A memory-based approach to protein sequence similarity searching. In Pifat-Mrzljak,G. (ed.), *Supramolecular Structure and Function 7*. Kluwer Academic Publishers, Dordrecht, pp. 155–166.
13. Murvai,J., Vlahovicek,K., Szepesvári,C. and Pongor,S. (2001) Prediction of protein functional domains from sequences using artificial neural networks. *Genome Res.*, **11**, 1410–1417.
14. Murvai,J., Vlahovicek,K. and Pongor,S. (2000) A simple probabilistic scoring method for protein domain identification. *Bioinformatics*, **16**, 1155–1156.