Nucleic Acids Research, 2003, Vol. 31, No. 1 403–405
DOI: 10.1093/nar/gkg098

# The SBASE domain sequence library, release 10: domain architecture prediction

**Kristian Vlahoviček, Laszló Kaján, János Murvai[1], Zoltán Hegedűs[2] and Sándor Pongor***

ICGEB—International Center for Genetic Engineering and Biotechnology, Area Science Park, 34012 Trieste, Italy, [1]MSD, Inc., 131 Park Street NE, Vienna, VA 22180, USA and [2]Biological Research Center of Hungarian Academy Sciences, H-6726 Szeged, Temesvári krt 62, Hungary

Received October 2, 2002; Accepted October 3, 2002

## ABSTRACT

**SBASE (http://www.icgeb.trieste.it/sbase) is an on-line collection of protein domain sequences and related computational tools designed to facilitate detection of domain homologies based on simple database search. The 10th 'jubilee release' of the SBASE library of protein domain sequences contains 1 052 904 protein sequence segments annotated by structure, function, ligand-binding or cellular topology, clustered into over 6000 domain groups. Domain identification and functional prediction are based on a comparison of BLAST search outputs with a knowledge base of biologically significant similarities extracted from known domain groups. The knowledge base is generated automatically for each domain group from the comparison of within-group ('self') and out-of-group ('non-self') similarities. This is a memory-based approach wherein group-specific similarity functions are automatically learned from the database.**

## INTRODUCTION

SBASE is a collection of 1 052 904 protein domain sequences. Each SBASE domain record contains a sequence, assigned to one of 4340 functionally or structurally well-characterized groups (SBASE-A) and/or to one of the less well characterized 1863 groups (SBASE-B) described in terms of amino acid composition or cellular localization. All domains are cross-referenced back to their parent protein databases [SWISS-PROT+TrEMBL (1), PIR (2)] and to entries in other domain repositories, such as InterPro (3) or one of its constituent databases [largely Pfam (4), SMART (5) and PRINTS (6)]. Furthermore, SBASE records host taxonomy information and domain descriptions.

## DOMAIN ARCHITECTURE PREDICTION

From this year, SBASE provides users with an automated domain architecture prediction system, based on the so-called memory based learning approach of Stanfill and Waltz (7). The preprocessing step consists in comparing the SBASE library to itself, using BLAST in an 'all versus all' manner which results in a 'database similarity space' (8). Since all domains are previously annotated, we can distinguish two types of similarities for each group of domain records, i.e. the 'self-similarity' (i.e. the inter-group sequence similarity scores) and the 'non-self similarity' (the similarity scores between group members and non-member sequences in the neighborhood of the group). By extracting two similarity-space parameters from each group, namely the average number of significant similarities (NSD; 'average degree' in graph theory terms) and the average similarity score (AVS) separately for self- and non-self similarities, we can determine minimum requirements that a sequence segment has to satisfy in order to be considered part of the group. The partitioning of self- and non-self similarities can be as simple as determining the NSD and AVS thresholds (9) or, in case of groups with more 'noisy' backgrounds, it may include either probabilistic scoring (10) or neural networks (11). The classification procedure for new members is then as straightforward as running BLAST of new domain versus the database, determining NSD and AVS towards positively hit domains in each of the domain groups that occurs in the output and selecting the best-scoring group based on each group's partitioning parameters established in the previous step.

The primary advantage of the prediction system is its speed. Prediction run time is only marginally longer than the average BLAST run time for the same sequence against the database and has an prediction accuracy that compares very well to that of more time-consuming methods (9). Second, the principle is applicable to any kind of annotated sequence database. Finally, once the initial core set of domains are collected for each group, the prediction system can automatically add newly predicted samples to their respective groups and re-use them for prediction. This is especially valuable in large-scale annotation projects where high data throughput and automation are a priority.

*To whom correspondence should be addressed. Email: pongor@icgeb.trieste.it

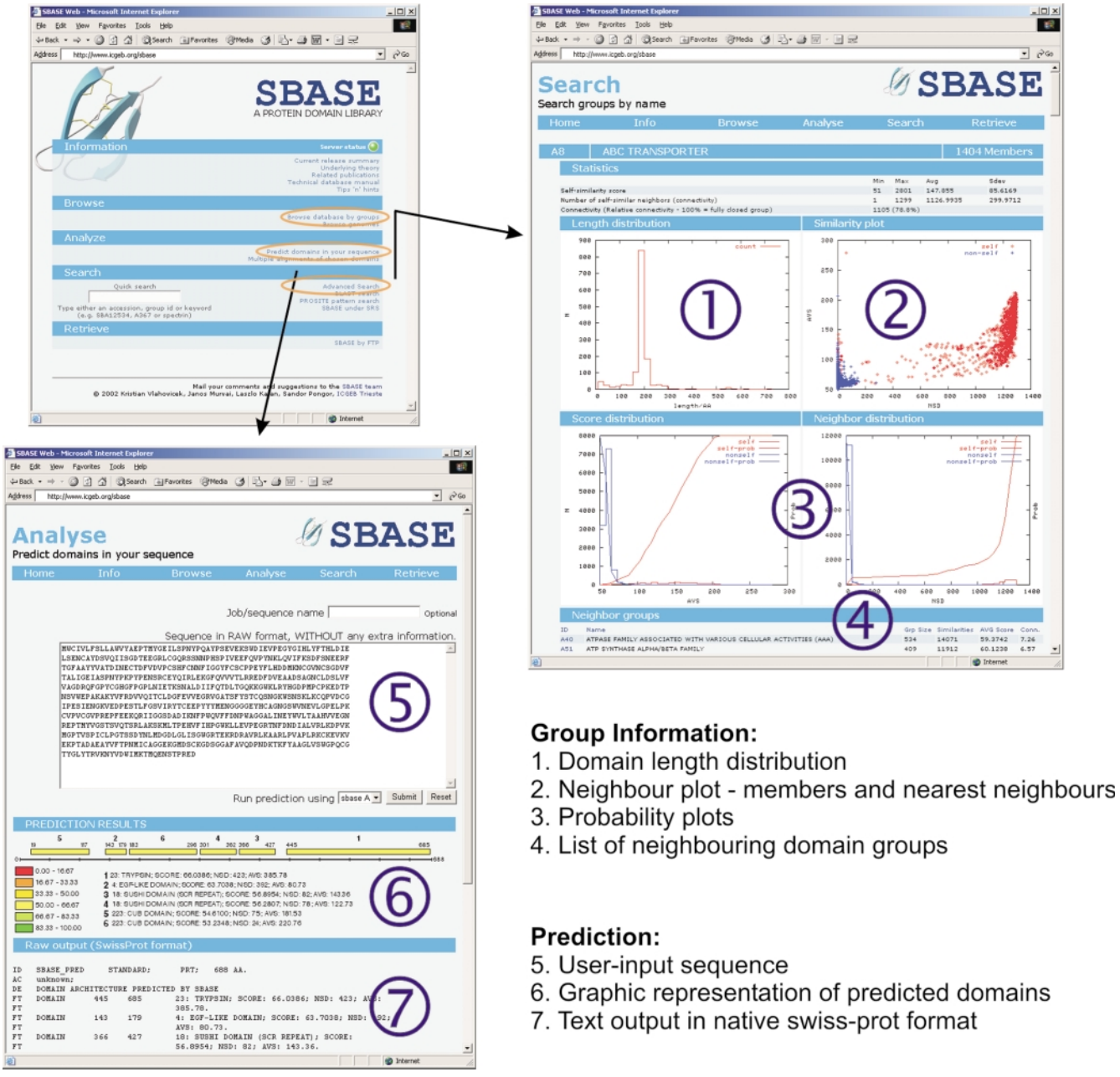


**Figure 1.** The SBASE domain library www interface; http://www.icgeb.org/sbase. The home page (top left) provides links for browsing the database groups and retrieving domain group information (top right) used in domain architecture prediction (bottom left). Each group contains graphical information on domain length distribution (1) and the AVS (average similarity) versus NSD (number of similar domains) plots (2) for group members (red) and nearest group neighbors (blue) as well as distribution histograms and posterior probability curves for AVS (3; left plot) and NSD (3; right plot) within a group. Each group also contains a cross-reference list of groups containing neighboring domains (4). Domain architecture prediction involves a simple procedure of user pasting a sequence for prediction (5) and pushing the submit button. Prediction results [the submitted sequence in this case was human complement component C1s (SPROT:C1S_HUMAN)] reveal the positions and lengths of predicted domains both graphically (6), where domains are ranked and coloured by prediction reliability score (scale from red to green, red being the least reliable), and as a raw swissprot file, with annotations (7). Domain architecture sketch is 'clickable' and links each prediction to the corresponding domain group information page. The sketch can be easily transferred, via copy-paste, to users' publication.

Improvements with respect to the previous release:

i. Completed migration to the relational database management system (RDBMS). All data are now being stored in a MySQL database, which enables more complex data querying methods, consistent data retrieval and allows for permanent domain group accession numbers that in turn makes the domain library easier to interface with other available sequence annotation projects (e.g. InterPro). The RDBMS will also decrease data maintenance times and will permit more frequent database updates in the future.

ii. A new web based interface (Fig. 1). The interface provides full access to the domain library in terms of browsing and/or searching for a particular domain, as well as an interactive domain architecture prediction system that graphically presents output results in form of a domain sketch that can be readily included in users' publications. The interface provides advanced database querying methods that include profile searches on domain sequences, taxonomy information searches and a 'genome browser'—a list of domains found in genomes completed so far.

iii. Improved prediction system. Domain architecture of user-submitted protein sequences is predicted using a three-step algorithm, described in (9–11). In order to increase prediction efficacy and decrease response time, the prediction system and the domain library are distributed over a cluster of 15 computers.

iv. Inclusion of TrEMBL domains. The tenth release of SBASE includes annotated domains from TrEMBL, thus greatly increasing the database size to one million domains (an increase of 300% with respect to the previous release). Such an increase resulted in improved prediction accuracy, largely due to consolidation of domain groups that previously had few samples.

## FUTURE DIRECTIONS

The SBASE development efforts will concentrate on further data consistency improvements, in terms of stricter domain length checks and removal of highly redundant domain samples. Tuning the web interface and including new functionality will be done according to valuable suggestions from the user community.

## DISTRIBUTION AND ACCESS

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

## ACKNOWLEDGEMENTS

The SBASE is established 1990 and is maintained collaboratively by ICGEB, Trieste and the Biological Research Center of the Hungarian Academy of Sciences, Szeged, Hungary. K.V. wishes to thank Boris Lenhard (Center for Genomics and Bioinformatics, Sweden) for useful comments during the SBASE web interface development. S.P. is a recipient of an Albert Szent-Györgyi Fellowship Award from the Hungarian Ministry of Education.

## REFERENCES

1. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
2. Wu,C.H., Huang,H., Arminski,L., Castro-Alvear,J., Chen,Y., Hu,Z.Z., Ledley,R.S., Lewis,K.C., Mewes,H.W., Orcutt,B.C. *et al.* (2002) The Protein Information Resource: an integrated public resource of functional annotation of proteins. *Nucleic Acids Res.*, **30**, 35–37.
3. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P., Bucher,P. *et al.* (2002) InterPro: an integrated documentation resource for protein families, domains and functional sites. *Brief. Bioinform.*, **3**, 225–235.
4. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.
5. Letunic,I., Goodstadt,L., Dickens,N.J., Doerks,T., Schultz,J., Mott,R., Ciccarelli,F., Copley,R.R., Ponting,C.P. and Bork,P. (2002) Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Res.*, **30**, 242–244.
6. Attwood,T.K., Blythe,M.J., Flower,D.R., Gaulton,A., Mabey,J.E., Maudling,N., McGregor,L., Mitchell,A.L., Moulton,G., Paine,K. *et al.* (2002) PRINTS and PRINTS-S shed light on protein ancestry. *Nucleic Acids Res.*, **30**, 239–241.
7. Stanfill,C. and Waltz,D. (1986) Toward memory-based reasoning. *Commun. ACM*, **29**, 1213–1228.
8. Johnson,M.A. and Maggiora,G.M. (1990) *Concepts and Applications of Molecular Similarity Analysis*, 1st Edn. John Wiley and Sons Inc., New York, Chichester.
9. Murvai,J., Vlahovicek,K. and Pongor,S. (2001) Towards a memory-based interpretation of proteome data. In Pifat-Mrzljak,G. (ed.), *Supramolecular Structure and Function 7*. Kluwer Academic Publishers, Dordrecht, pp. 155–166.
10. Murvai,J., Vlahovicek,K. and Pongor,S. (2000) A simple probabilistic scoring method for protein domain identification. *Bioinformatics*, **16**, 1155–1156.
11. 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.