Nucleic Acids Research, 1999, Vol. 27, No. 1 229–232

SMART: identification and annotation of domains from signalling and extracellular protein sequences

Chris P. Ponting*, Jörg Schultz[1], Frank Milpetz[1] and Peer Bork[1]

University of Oxford, Fibrinolysis Research Unit, The Old Observatory, South Parks Road, Oxford OX1 3RH, UK and [1]European Molecular Biology Laboratory, Meyerhofstrasse 1, Heidelberg 69012, Germany

Received September 2, 1998; Accepted September 29, 1998

ABSTRACT

SMART is a simple modular architecture research tool and database that provides domain identification and annotation on the WWW (http://coot.embl-heidelberg.de/SMART). The tool compares query sequences with its databases of domain sequences and multiple alignments whilst concurrently identifying compositionally biased regions such as signal peptide, transmembrane and coiled coil segments. Annotated and unannotated regions of the sequence can be used as queries in searches of sequence databases. The SMART alignment collection represents more than 250 signalling and extracellular domains. Each alignment is curated to assign appropriate domain boundaries and to ensure its quality. In addition, each domain is annotated extensively with respect to cellular localisation, species distribution, functional class, tertiary structure and functionally important residues.

INTRODUCTION

Predicting the function of a protein from its sequence can be a laborious process that is fraught with pitfalls associated with sequence divergence, non-identical multidomain architectures and non-equivalent functions of homologues. In particular, detecting regulatory domains is of importance if the detailed cellular roles of multidomain proteins are to be predicted. Yet these domain homologues are mostly divergent in sequence and appear in numerous molecular contexts that may not be held in common even among homologues in diverse organisms. For example, families of protein kinase C isoenzymes are clearly apparent in yeast and vertebrates yet in no case are their arrangements of regulatory domains identical. To compound these difficulties, a family of domain homologues usually possesses a variety of distinct, albeit similar functions, that may be distinguished only by conservation of key active or binding site residue determinants.

We have addressed these problems by providing a Web-based tool and database (SMART, a simple modular architecture research tool) that allows the identification of divergent domain homologues in user-supplied sequences (1). Prediction of domain homologues obviates misannotations of sequences that arise when comparing pairs of multidomain protein sequences with contrasting domain architectures. This procedure also facilitates subsequent investigation of unannotated sequence regions thereby improving on the signal-to-noise ratio of the search (2). The initial SMART database release focused on signalling proteins. These contain a considerable variety of non-enzymatic regulatory domains (3), and mediate or regulate transduction of an extracellular signal towards the nucleus.

In 1998, SMART has been updated to allow the detection of extracellular domains (4) initially focusing on those that mediate cell–cell signalling events. Domains that function in prokaryotic and eukaryotic two component regulatory systems can now also be detected. Identified domains are annotated by providing links to (i) recent literature via ENTREZ (5), (ii) homologues with known three-dimensional structure via PDBsum (6), and (iii) the domain or motif collections of Pfam (7), BLOCKS (8) and PROSITE (9). Furthermore, each domain annotation has been extended to include information relating to (i) species range, (ii) subcellular localization, (iii) functional class and (iv) functional residues and secondary structure that are mapped onto alignments using consensus sequences. The output display has been simplified, and glossary and help pages are now provided. Here we describe the latest release of SMART (version 2.0; http://coot.embl-heidelberg.de/SMART) including changes made to methodologies and to the display of predictions.

DESCRIPTION OF THE DATA

Domains detectable by SMART are mostly represented by seed alignments. These are constructed by retaining only a single sequence from each branch of a phylogenetic tree that is shorter than a distance of 0.2 (1); this corresponds approximately to 80% identity (10). Seed alignments are now available to the academic community in MSF, CLUSTAL-W, PIR and FASTA (Pearson) formats. Alignments are constructed according to established procedures (reviewed in 2,11) and submission by others of similarly derived alignments of domains not yet represented by SMART is encouraged. Profiles (12) and profile/HMMs (13) are constructed from each alignment.

Alignments usually represent the complete sequences of the entire family of detected homologues. However there are

*To whom correspondence should be addressed at present address: National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 38A, Bethesda, MD 20894, USA. Tel: +1 301 435 5916; Fax: +1 301 480 9241; Email: ponting@ncbi.nlm.nih.gov

exceptions: (i) for families where function clearly partitions with sequence similarity (as in the Ras family of small GTPases), an alignment of each subfamily (namely, Arf, Rab, Ran, Ras, Rho/Rac and Sar) has been constructed; (ii) families for which a single profile is insufficient to detect all known homologues are represented by two or more profiles; and (iii) families of homologues that cannot reliably be aligned throughout the domain are represented by alignments of their most conserved region(s) (i.e. 'motifs') only.

THE SMART WORLD WIDE WEB SERVER

SMART has been provided with a user interface (http://coot.embl-heidelberg.de/SMART/) that predicts the domain architecture of a user-supplied sequence via a graphical display. By default, sequences are searched against the collection of profiles using the WiseTools suite (12). True positive homologues are defined as sequences that score above a defined threshold (T_p), false negatives are detected as subsequent repeats that score above a secondary threshold (T_r) and pairs of repeats are detected where each of their scores is greater than $[(T_p+T_r)/2]$ (1). *E*-values (14) are provided but not used as thresholds since a wide variation in their values for different domain types has been observed [as implemented using the latest version of WiseTools (E.Birney, J.D.Thompson and T.J.Gibson, unpublished)].

Sequences are also searched for signal peptides, coiled coils, low complexity regions and transmembrane helices (15–18). SMART provides options to search sequences for SMART and Pfam (7) domain sets using the HMMER package (13). In addition, annotated and unannotated regions of the user's sequence may be subjected to WU-BLASTP analysis (WU-BLAST 2.0; W.Gish, unpublished) using a dedicated WU-Blast server (http://www.bork.embl-heidelberg.de/blast2 ; 19).

Several options have now been added to sequence searches. A user can specify that a search be made for either cytoplasmic or extracellular domains, and may additionally specify the query sequence using a sequence identifier or accession code. Due to high user demand at peak periods, searches are performed using a 4-processor machine and a queuing system is in place.

The SMART domain set: multiple alignments

The SMART alignment collection now contains multiple alignments of more than 100 cytoplasmic and more than 100 extracellular domains that are compared with each query sequence entered. SMART's domain coverage will increase in forthcoming years. Figure 1 shows the SMART-derived annotation of the Lambert-Eaton myasthenic syndrome antigen B (MYSB; SwissProt accession CCB2_HUMAN) which is the β2-subunit of the dihydropyridine-sensitive L-type calcium channel. The annotation procedure unexpectedly shows this protein to be a member of the membrane-associated guanylate kinase homologue (MAGUK) family (20), containing both an src homology 3 domain and a domain homologous to guanylate kinases. Annotation of the alignment of MYSB with guanylate kinase active enzyme sequences (Fig. 1) demonstrates that MYSB appears to have dispensed with its ATP-binding function and is likely to be inactive as an enzyme. It is emphasised that the annotation of binding or catalytic residues is essential if one is to infer the function of homologues from existing experimental data. From this example, it is clear that a functional prediction procedure that neglected annotation at the amino acid level would have wrongly predicted MYSB as an active guanylate kinase.

The SMART domain set: the 'Schnipsel' database

Query sequences are subjected to searches not only against HMM/profiles derived from multiple alignments, but also against a database of domain sequences using WU-BLAST. This 'Schnipsel' (a German word meaning 'snippet' or 'fragment') database contains the set of domain sequences identified by SMART from current protein sequence databases. This feature is an important asset to SMART since outliers of a family often cannot be detected by a profile, yet are detectable by pairwise similarity to one or more established members of a sequence family. This facility is similar to that employed by the SBASE protein domain library (21).

The SMART domain annotations

For each segment of a query sequence that matches an annotated domain, (i) a multiple alignment can be accessed, (ii) a WU-Blast search can be initiated to identify the closest domain homologue, (iii) the alignment of the query sequence segment with domain consensus sequences (N.P.Brown *et al.*, unpublished) can be displayed, and (iv) an alignment against the domain profile is shown complete with secondary structure information from members of known three-dimensional structure. Furthermore, functional information ascribed to residues of family members are also mapped onto this alignment and the source of this information is hyperlinked. Caution should be shown, however, when transferring this type of information to the query (22).

Clicking on any domain icon ('bubble') on the SMART results page reveals the domain's annotation page from which alignments and results of additional BLAST procedures may be viewed. In addition the distribution of phyla that are known to contain proteins with this domain is given. Links are also provided to all protein sequences in a pseudo-non-redundant database with the particular domain type. This provides information such as the number (647, as of August 1998) of human protein kinase homologues represented in this database and, of these, the number (290) that contain additional annotated domains.

In version 2.0, annotation pages of predicted domains have been completely over-hauled. Manual annotations include (i) domain abbreviation, (ii) description line(s), (iii) selected key literature via hyperlinks to Entrez (5), (iv) the predicted cellular role of the protein (namely, chromatin-associated, metabolism, signalling, transport, translation, transcription, replication and interaction with the environment), (v) molecular features such as binding properties or catalytic activities, as well as (vi) links to PROSITE, BLOCKS and PFAM databases.

In addition, a number of automatic annotation tools have been added to retrieve quantitatively: (i) the distributions of organisms in which the domain has been identified; (ii) the predicted subcellular localisation of the homologues categorised as intracellular (subdivided into the nucleus, cytoplasm, ER-golgi, chloroplast and mitochondrion), extracellular, secreted, membrane-associated and transmembrane, using the protocol of Eisenhaber and Bork (23); and (iii) domain family members with known three-dimensional structures linked to a PDBviewer (http://www.biochem.ucl.ac.uk/bsm/pdbsum/ ; 6) and to Entrez-MMDB (http://www.ncbi.nlm.nih.gov/Entrez/structure.html ; 24).

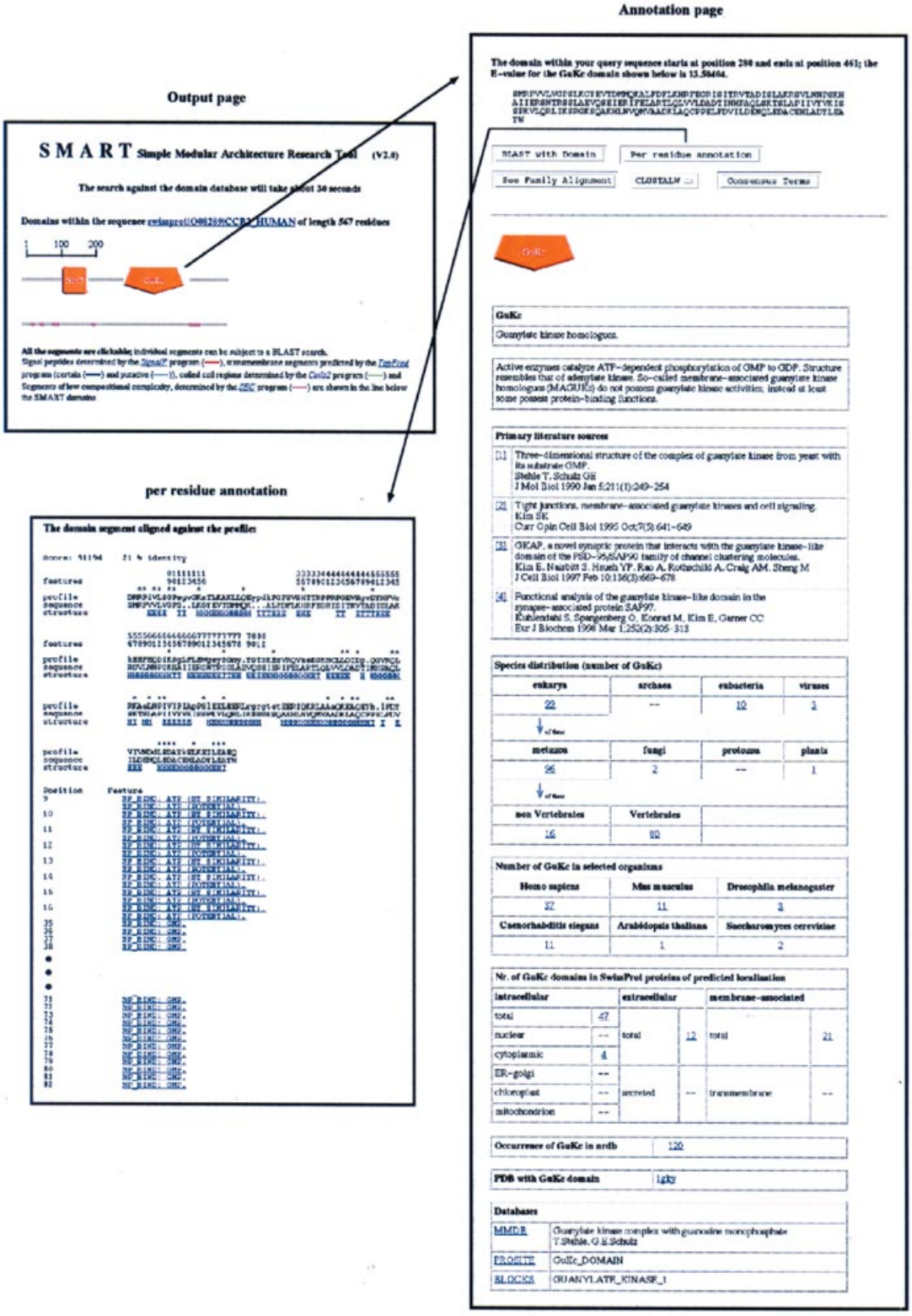

Figure 1. An example of the use of the SMART WWW server. An output page (upper right) shows that a query sequence (SwissProt code CCB2_HUMAN) is predicted to contain single src homology 3 (SH3) and guanylate kinase homologous (GuKc) domains. These domains are represented using 'bubbles'. The annotation pages for the GuKc domain (right and lower left) provide a description of the domain family, literature links and numbers detected in different phyla from a pseudo-non-redundant database. The 'per residue' annotation page (lower left) maps SwissProt annotation features onto an alignment of the CCB2_HUMAN GuKc domain sequence.

FUTURE DIRECTIONS

The major goals of this project are two-fold: (i) the provision of a Web-based annotation tool that is able to predict the function and structure of protein sequences, and (ii) the use of this tool to infer the evolution of mobile (non-catalytic) domains and their functional networks in diverse organisms (25). Thus, SMART's domain coverage is lower than those of Prosite (9), BLOCKS (8), PRINTS (26) and Pfam (7) and users are encouraged to use the Web-based servers of SMART, Prosite (http://expasy.hcuge.ch/sprot/prosite.html), BLOCKS (http://www.blocks.fhcrc.org), PRINTS (http://www.biochem.ucl.ac.uk/bsm/dbbrowser/), Pfam (http://genome.wustl.edu/Pfam/ or http://www.sanger.ac.uk/Software/Pfam3/) and ProfileScan (http://ulrec3.unil.ch/software/PFSCAN_form.html) in parallel since there is considerable complementarity among these services. Users of the SMART research tool and database are encouraged to cite this article in resulting publications.

ACKNOWLEDGEMENTS

The authors thank each of their colleagues in the Bork group for valuable discussions and contributions to this work, in particular F. Eisenhaber for helping in linking predicted localisation information and Y. P. Yuan for system administration support. C.P.P. is a Wellcome Trust Research Fellow and is a member of the Oxford Centre for Molecular Sciences. J.S., F.M. and P.B. are supported by the DFG and by the EC (grant 01KW9602/6) as well as by the BMBF grants MEDSEQ and TARGID.

REFERENCES

1 Schultz,J., Milpetz,F., Bork,P. and Ponting,C.P. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 5857–5864.
2 Bork,P. and Koonin,E.V. (1998) *Nature Genet.*, **18**, 313–318.
3 Bork,P., Schultz,J. and Ponting,C.P. (1997) *Trends Biochem. Sci.*, **22**, 296–298.
4 Bork,P., Downing,A.K., Kieffer,B. and Campbell,I.D. (1996) *Q. Rev. Biophys.*, **29**, 119–167.
5 McEntyre,J. (1998) *Trends Genet.*, **14**, 39–40.
6 Laskowski,R.A., Hutchinson,E.G., Mitchie,A.D., Wallace,A.C., Jones,M.L. and Thornton,J.M. (1997) *Trends Biochem. Sci.*, **22**, 488–490.
7 Sonnhammer,E.L.L., Eddy,S.R., Birney,E., Bateman,A. and Durbin,R. (1998) *Nucleic Acids Res.*, **26**, 320–322.
8 Henikoff,S., Pietrokovski,S. and Henikoff,J.G. (1998) *Nucleic Acids Res.*, **26**, 309–312.
9 Bairoch,A., Bucher,P. and Hofmann,K. (1997) *Nucleic Acids Res.*, **25**, 217–221.
10 Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) *Nucleic Acids Res.*, **22**, 4673–4680.
11 Bork,P. and Gibson,T.J. (1996) *Methods Enzymol.*, **266**, 162–184.
12 Birney,E., Thompson,J.D. and Gibson,T.J. (1996) *Nucleic Acids Res.*, **24**, 2730–2739.
13 Eddy,S.R., Mitchison,G. and Durbin,R.J. (1995) *J. Comput. Biol.*, **2**, 9–23.
14 Altschul,S.F. and Gish,W. (1996) *Methods Enzymol.*, **266**, 460–480.
15 Nielsen,H., Engelbrecht,J., Brunak,S. and von Heijne,G. (1997) *Protein Engng.*, **10**, 1–6.
16 Lupas,A., Van Dyke,M. and Stock,J. (1991) *Science*, **252**, 1162–1164.
17 Wootton,J.C. and Federhen,S. (1996) *Methods Enzymol.*, **266**, 554–573.
18 von Heijne,G. (1992) *J. Mol. Biol.*, **225**, 487–494.
19 Yuan,Y., Eulenstein,O., Vingron,M. and Bork,P. (1998) *Bioinformatics*, **14**, 285–289.
20 Gomperts,S.N. (1996) *Cell*, **84**, 659–662.
21 Fabian,P., Murvai,J., Hatsagi,Z., Vlahovicek,K., Hegyi,H. and Pongor,S. (1997) *Nucleic Acids Res.*, **25**, 240–243. http://www2.icgeb.trieste.it/~sbasesrv/
22 Doerks,T., Bairoch,A. and Bork,P. (1998) *Trends Genet.*, **14**, 248–250.
23 Eisenhaber,F. and Bork,P. (1998) *Trends Cell Biol.*, **8**, 169–170.
24 Ohkawa,H., Ostell,J. and Bryant,S. (1995) *ISMB*, **3**, 259–267.
25 Bork,P. (1992) *Curr. Opin. Struct. Biol.*, **2**, 413–421.
26 Attwood,T.K., Beck,M.E., Flower,D.R., Scordis,P. and Selley,J.N. (1998) *Nucleic Acids Res.*, **26**, 304–308.