A Versatile Structural Domain Analysis Server Using Profile Weight Matrices

Steffen Schmidt,[†,§] Peer Bork,[†,‡] and Thomas Dandekar*[,†,‡,§,#]

EMBL, Postfach 102209, D-69012 Heidelberg, Germany, Max Delbrueck Centre for Molecular Medicine, Robert-Roessle-Strasse 10, 13092 Berlin-Buch, Germany, Parasitology, University of Heidelberg, Im Neuenheimer Feld 324, 69120 Heidelberg, Germany, and Institute for Molecular Medicine, Breisacherstrasse 66, 79106 Freiburg, Germany

Received June 8, 2001

The WEB tool "AnDom" assigns to a given protein sequence all experimentally determined structural domains contained within it, including multidomain and large proteins. The server uses profile specific matrices from custom generated multiple sequence alignments of all known SCOP domains (SCOP version 1.50). Prediction time is short allowing numerous applications for structural genomics including investigation of complex eucaryotic protein families. The WWW server is at http://www.bork.embl-heidelberg.de/AnDom, and profiles can be downloaded at ftp.bork.embl-heidelberg.de/pub/users/ schmidt/AnDom.

INTRODUCTION

A number of approaches (e.g. SMART;[1] PFAM;[2] COGs;[3] conserved domain database[4] http://www.ncbi.nlm.nih.gov) use sequence information to analyze protein domains and/or function. All of them have different advantages and limitations. In practice it is important to retest function and domain assignment by several independent methods to minimize false positive and false negative domain assignments.[5] Here we offer for such evaluations a sequence-structure mapping tool. It assigns to a protein all domains of known three-dimensional structure. Specific domains identified in this way mediate functions such as nucleotide or cofactor binding and are known in atomic detail. It uses a specifically generated database, calculating structural domain family profiles and clustering according to the SCOP classification.[6] Found similarities are given as unbiased as possible for each different part of the sequence, and the user can decide which significance values to accept (ranges can be set from e^{-30} to 10.0).

Limitations are according to the sensitivity and coverage of the profiles (see below), significance level selected, and the SCOP classification itself.

PROCEDURES

Query. A query is posted by simply pasting the sequence into the query window (accepted formats: Raw, FASTA). Run time scales only lineary with protein sequence length (5 s/ 100 amino acids using a 4 × 550 MHz PIII Xeon with 2GB RAM). The output obtained allows rapid assignment of the different structural domains contained just by visual inspection of the different color coded SCOP domains and is explained in detail in Figure 1.

Database Generation. Known SCOP domains (version 1.50) were collected from the ASTRAL[7] compendium of structural domains selecting a cutoff for respective sequence identities of less than 40%. Each SCOP domain sequence was augmented by iterative sequence alignment searches against the nonredundant database using PSI-BLAST,[8] allowing 10 iterations. Low complexity regions were filtered out using the program SEG.[9] Multiple alignments derived included the full family of structurally related sequences using a cutoff of $e = 10^{-3}$ before matrix profiles were calculated.

Matrix Profiles. These were calculated for each of the specifically generated structural (via SCOP) related sequence alignments. We used the IMPALA package to obtain position specific score matrices (PSSMs) from the PSI-BLAST outputs. As described[10] the program MAKEMAT converts byte-encoded frequency ratios to integral ASCII score matrices and the program COPYMAT converts the matrix files from our database into one large byte-encoded integer matrix, but the local alignment is improved by using the rigorous Smith-Waterman algorithm.

RESULTS AND DISCUSSION

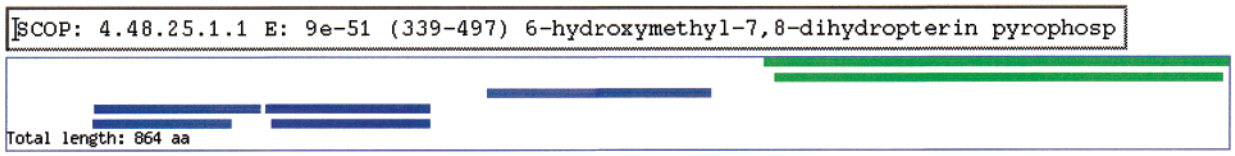
Efficiency. The SCOP domain assignments made by our WEB server were tested on SWISS-PROT (release 38.0).[11] Seventy-seven percent of the entries (16 742 from 21 553 entries) with E.C. classification are rapidly and efficiently assigned (expected value below 10^{-3} to avoid PSSM divergence[12]) a structural domain. From these entries, 11 492 (69%) were assigned completely (less than 100 amino acids unassigned) and nonoverlapping (less than 10 amino acids overlap) to different SCOP domains. One thousand one hundred ninety entries (7%) had overlapping SCOP domains, and 4060 were not completely assigned (24%; these had at least 100 amino acids unassigned). The default parameters for the server and database allow an efficient coverage yet unambiguous assignment of structural domains from sequence.

The server is implemented and ready to use at http://www.bork.embl-heidelberg.de/AnDom. Newly generated for the server were the following: The preparation of SCOP structural domain related sequence families from iterative

* Corresponding author phone: 06221-387-466; fax: 06221-387-306; e-mail: dandekar@embl-heidelberg.de.
† EMBL.
‡ Max Delbrueck Centre for Molecular Medicine.
§ University of Heidelberg.
# Institute for Molecular Medicine.

10.1021/ci010374r CCC: $22.00
Published on Web 01/18/2002

a)



b)

Sequences producing significant alignments:			Score (bits)	E Value
PDB	SCOP			
d1aj2__	3.1.20.1.1	Dihydropteroate synthetase {Escherichia c...	265	5e-72
d1ad1a_	3.1.20.1.2	Dihydropteroate synthetase {Staphylococcu...	263	1e-71
d1hka__	4.48.25.1.1	6-hydroxymethyl-7,8-dihydropterin pyroph...	194	9e-51
d1b9la_	4.80.1.3.2	7,8-dihydroneopterin triphosphate epimera...	100	2e-22
d1dhn__	4.80.1.3.1	7,8-dihidroneopterin aldolase {Staphyloco...	99	3e-22

c)

```
> d1hka__ 4.48.25.1.1 6-hydroxymethyl-7,8-dihydropterin pyrophosphokinase, HPPK
          Length = 158

 Score =  194 bits (490), Expect = 9e-51
 Identities = 52/160 (32%), Positives = 52/160 (32%), Gaps = 8/160 (5%)

Query: 339 RAFLAFGSNIGDRFKHIQMALQLLSREKTVKLRNISSIFESEPMYFKDQTPFMNGCVEVE 398

Sbjct: 2   VAYIAIGSNLASPLEQVNAALKALGDIPESHILTVSSFYRTPPLGPQDQPDYLNAAVALE 61

Query: 399 TLLTPSELLKLCKKIEYEELQRVKHFDNGPRTIDLDIVMFLNSAGEDIIVNEPDLNIPHP 458

Sbjct: 62  TSLAPEELLNHTQRIELQQGRVRKAERWGPRTLDLDIMLF-----GNEVINTERLTVPHY 116

Query: 459 RMLERTFVLEPLCELISPVHLHPVTAEPIVDHLKQ-LYDK 497

Sbjct: 117 DMKNRGFMLWPLFE-IAPELVFP-DGEMLRQILHTRAFDK 154
```

Figure 1. Analysis example: FAS_YEAST (predicted to be a folic acid synthesis protein in yeast). (a) Significant similarities and their lengths to structural domains found (move mouse over for display), indicating SCOP class color codes. Color shades distinguish different folds (shades of black, red, green, blue, pink, yellow, turquise for SCOP 1.* to SCOP 7.*). (b) SCOP domains found including PDB code, SCOP identification number, and short description of the domain. If several SCOP hits are found for the same part of the sequence, all are displayed. In this case this allows easy assignment and identification of three different structural domains with catalytic activity each participating in the structure for this sequence. Additional catalytic domains are also rapidly identified, e.g. in viral polymerase sequences a helicase activity in addition to the polymerase domain. (c) Individual alignments of SCOP hits are given in addition. The server seeks an optimal gapped local alignment of the query sequence Q against each PSSM M. Each PSSM uses a representative sequence as a place holder only to display pairwise alignments in BLAST format.

sequence searches, the calculation of corresponding profile matrices, the rapid query interface, script annotation of input and output, the parser, assembly, and graphical presentation of results as well as the WEB output surface. All are fully available (the profiles used via anonymous ftp at ftp-.bork.embl-heidelberg.de/pub/users/schmidt/AnDom). All the settings as well as the database are completely adjustable and customizable to user specified needs.

Server Specific Features. Several independent and different methods have been advocated to maximize structure prediction accuracy particularly in genome projects.[13] Compared to other available software for domain assignment our tool does not rely on ProDom,[14] Pfam, or SMART sequence families (CDD server, http://www.ncbi.nlm.nih.gov[4]). Sequence family information from any of these databases is valuable and important, but regions of known three-dimensional structure become only apparent applying further analysis and tools.

In contrast, our tool recognizes specifically protein structure domains. An example output is demonstrated and explained in Figure 1. To maximize structural information and recognition of structural homologues we augmented and completed all known SCOP folds by iterative sequence alignments in a custom generated database. Any known structural domains the query sequence is homologous to are rapidly identified and indicated.

Our server is also independent from threading approaches with contact potentials to thread sequences on three-dimensional structures such as the 123D+ server (http://www-lmmb.ncifcif.gov/~nicka/123D.html[15]). Furthermore, no comparative modeling steps[16] are required or involved. Program specific scoring functions affect threading predictions in different and partly unpredictable ways depending on the structure examined, so that in fact even specific programs for cross-comparison of threading predictions have been developed.[17]

The value of direct structural information for genome annotation has also been demonstrated by DiGennaro et al. in a recent study.[18] However, whereas their approach considers only a three-dimensional description of functional sites (so-called "FFF", fuzzy functional forms), we consider and indicate by our server complete and well described structural domains.

Neglecting the specific domain architecture, PSSMs have previously been shown to be useful for assignment of known pdb protein files, notably in *M. genitalium* with a cutoff of protein size till 800 amino acids[19] (http://www.bmm.icnet.uk/servers/3dpssm/) or to compare protein fold distributions.[20]

The server package presented here offers now a handy solution for a protein of any length (e.g. in contrast to ref 19) with rapid detection of each individual SCOP domain and the specific domain architecture contained including

multidomain and large proteins (a further specific advantage). It is well suited for structural genomics investigations including detailed analysis of eucaryotic protein families and a concise, user-friendly output. It is easily adjusted to user specific databases and threshold choices e.g. user specific inhouse databases and/or highly specific, hand curated alignments.

ACKNOWLEDGMENT

We thank Richard Copley for suggestions and DFG for support (SFB 544/B2 and BO-1099/5-2).

REFERENCES AND NOTES

(1) Schultz, J.; Copley, R. R.; Doerks, T.; Ponting, C. P.; Bork, P. SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.* **2000**, *28*, 231−234.
(2) Bateman, A.; Birney, E.; Durbin, R.; Eddy, S. R.; Howe, K. L.; Sonnhammer, E. L. The Pfam Protein Families Database. *Nucleic Acids Res.* **2000**, *28*, 263−266.
(3) Tatusov, R. L.; Natale, D. A.; Garkavtsev, I. V.; Tatusova, T. A.; Shankavaram, U. T.; Rao, B. S.; Kiryutin, B.; Galperin, M. Y.; Fedorova, N. D.; Koonin, E. V. The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* **2001**, *29*, 22−28.
(4) Wheeler, D. L.; Church, D. M.; Lash, A. E.; Leipe, D. D.; Madden, T. L.; Pontius, J. U.; Schuler, G. D.; Schriml, L. M.; Tatusova, T. A.; Wagner, L.; Rapp, B. A. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2001**, *29*, 11−16.
(5) Bork, P.; Dandekar, T.; Diaz-Lazcoz, Y.; Eisenhaber, F.; Huynen, M.; Yuan, Y. Predicting function: from genes to genomes and back. *J. Mol. Biol.* **1998**, *283*, 707−725.
(6) Lo Conte, L.; Alley, B.; Hubbard, T. J.; Brenner, S. E.; Murzin, A. G.; Chothia, C. SCOP: A structural classification of proteins database. *Nucleic Acids Res.* **2000**, *28*, 257−259.
(7) Brenner, S. E.; Koehl, P.; Levitt, M. The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res.* **2000**, *28*, 254−256.
(8) Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI−BLAST, a new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389−3402.
(9) Wootton, J. C. Nonglobular domains in protein sequences: automated segmentation using complexity measures. *Comput. Chem.* **1994**, *18*, 269−285.
(10) Schäffer, A. A.; Wolf, Y. I.; Ponting, C. P.; Koonin, E. V.; Aravind, L.; Altschul, S. F. IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics* **1999**, *15*, 1000−1011.
(11) Bairoch, A.; Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **2000**, *28*, 45−48.
(12) Muller, A.; MacCallum, R. M.; Sternberg, M. J. Benchmarking PSI−BLAST in genome annotation. *J. Mol. Biol.* **1999**, *293*, 1257−1271.
(13) Baker, D.; Sali, A.; Protein structure prediction and structural genomics. *Science* **2001**, *294*, 93−96.
(14) Gouzy, J.; Corpet, F.; Kahn, D. Whole genome protein domain analysis using a new method for domain clustering. *Comput. Chem.* **1999**, *23*, 333−340.
(15) Alexandrov, N. N.; Nussinov, R.; Zimmer, R. R. Fast protein fold recognition via sequence to structure alignment and contact capacity potentials. In *Pacific symposium on Biocomputing 96*; Hunter, L., Klein, T. E., Eds.; World Scientific Publishing Co.; Singapore, pp 53−72.
(16) Kolinski, A.; Betancourt, M. R.; Kihara, D.; Rotkiewicz, P.; Skolnick J. Generalized comparative modeling (GENECOMP): A combination of sequence comparison, threading, and lattice modeling for protein structure prediction and refinement. *Proteins* **2001**, *44*, 133−149.
(17) Douguet, D.; Labesse, G. Easier threading through web-based comparisons and cross-validations. *Bioinformatics* **2001**, *17*, 752−753.
(18) Di Gennaro, J. A.; Siew, N.; Hoffman, B. T.; Zhang, L.; Skolnick, J.; Neilson, L. I.; Fetrow, J. S. Enhanced functional annotation of protein sequences via the use of structural descriptors. *J. Struct. Biol.* **2001**, *134*, 232−245.
(19) Kelley, L. A.; MacCallum, R. M.; Sternberg, M. J. Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.* **2000**, *299*, 499−520.
(20) Wolf, Y. I.; Brenner, S. E.; Bash, P. A.; Koonin, E. V. Distribution of protein folds in the three superkingdoms of life. *Genome Res.* **1999**, *9*, 17−26.

CI010374R