BIOINFORMATICS APPLICATIONS NOTE
Vol. 19 no. 3 2003, pages 429–430
DOI: 10.1093/bioinformatics/btg006

PDP: protein domain parser

Nickolai Alexandrov[1,*] and Ilya Shindyalov[2]

[1]Ceres Inc., 3007 Malibu Canyon Road, Malibu, CA 90265, USA and [2]San Diego Supercomputer Center, University of California, San Diego, MC 0537, 9500 Gilman Drive, La Jolla, CA 92093-0537, USA

Received on November 29, 2001; revised on September 27, 2002; accepted on October 6, 2002

ABSTRACT

Summary: We have developed a program for automatic identification of domains in protein three-dimensional structures. Performance of the program was assessed by three different benchmarks: (i) by comparison with the expert-curated SCOP database of structural domains; (ii) by comparison with a collection of manual domain assignments; and (iii) by comparison with a set of 55 proteins, frequently used as a benchmark for automatic domain assignment. In all these benchmarks PDP identified domains correctly in more than 80% of proteins.

Availability: http://123d.ncifcrf.gov/

Contact: nicka@ceres-inc.com; shindyal@sdsc.edu

The problem of parsing protein structures into domains has been approached before, (Crippen, 1978; Lesk and Rose, 1981; Holm and Sander, 1994; Islam *et al.*, 1995; Sowdhamini and Blundell, 1995; Jones *et al.*, 1998; Xu *et al.*, 2000), yet it cannot be considered as solved and be closed. This is because of the difficulty of this problem and also because of the ambiguity in the domain definition. Domain assignment is important for domain-based classification of proteins by three-dimensional structure (Murzin *et al.*, 1995; Holm and Sander, 1996; Orengo *et al.*, 1997) and by amino acid sequence (Bateman *et al.*, 2000; Apweiler *et al.*, 2001), for prediction of protein–protein interactions (Xenarios *et al.*, 2000), and for prediction of protein structure and function (Russel, 1999), We have developed and benchmarked a program PDP (Protein Domain Parser) for automatic domain assignment, which is currently used for preparation of the fold library for threading (Alexandrov *et al.*, 1996).

We define a domain as a set of continuous protein fragments. In the beginning, the whole protein chain is considered to be one domain consisting of one continuous fragment. At each recursive step, the program attempts to cut a domain into two domains by two ways: (1) by a single cut in all possible sites in the polypeptide chain, or (2) by a double cut in spatially close (distance between C$^{\alpha}$-atoms is less than 8 Å), but sequentially distant (more than 35 amino acids apart) sites of the polypeptide chain. After each attempt the number of contacts $nc(i, j)$ between two newly formed domains i and j is counted and normalized by the domain sizes: $nnc(i, j) = nc(i, j)/(|i|^{\alpha} \cdot |j|^{\alpha})$, where $|i|$ is the size of domain i, $\alpha = 0.43$. The initial assumption beyond this empirical formula is that the expected number of contacts between two domains depends on their surface areas, which is proportional to $n^{2/3}$ for a spherical domain of n amino acids. If the minimum of the normalized contacts is less than the threshold, the cut is implemented and the recursive step is repeated for the two new domains. The threshold is computed specifically for each given domain and is equal to one half of the average contact density for the whole domain. After all cuts are made, the contacts between all domains are checked again and domains with the large number of contacts (the normalized number of contacts is greater than two) are combined into one larger domain. At the last step PDP filters out all tiny domains (less than 30 amino acids).

The program was benchmarked with a set of 3548 non-redundant proteins for which domain annotation is available in SCOP (release 1.53). 84.2% of protein chains were cut into the same number of domains by both SCOP and PDP. A domain assignment was considered as correct, if (i) the number of domains in PDP and in SCOP assignments is equal and (ii) the agreement between these assignments is greater than 85%. Taking into account the accuracy of the domain assignment slightly reduces the fraction of the correct predictions to 83.9%. Most of the discrepancies belong to the category when SCOP considers the protein as one domain, but PDP splits it into two domains. Partially, this can be explained by those SCOP folds, where proteins are not cut into domains explicitly, yet are characterized as multi-domain proteins in the description of the fold. For example, perismatic binding protein-like I fold is typically a two-domain fold, however a representative PDB structure 2 lbp is a single entry in SCOP. Other discrepancies can be explained

[*]To whom correspondence should be addressed.

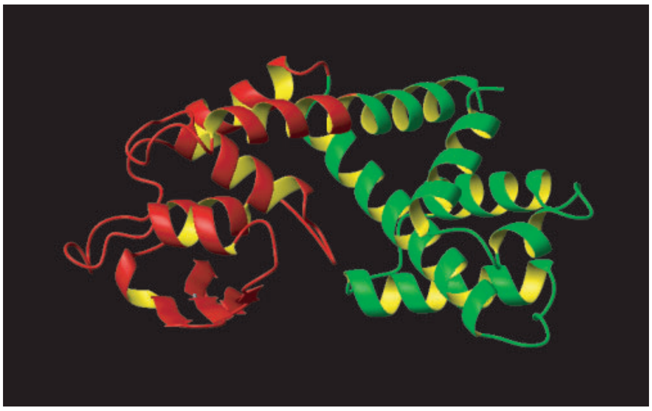

Fig. 1. PDP domain assignment in a catalytic core of eukaryotic DNA topoisomerase I from Vaccina virus (PDB ID 1a41). The boundaries of the domains are: domain 1 (red) 99–211; domain 2 consists of two fragments: 81–98 and 212–310. 50 Ca-atoms of the second domain can be aligned with bacteriocin AS-48 from Enterococcus Faecalis (PDB ID 1e68), resulting in rmsd of 2.45 Å. Such recurrence of the domain topology indicates that the domain assignment has a biological meaning. Figure prepared with the program MOLMOL (Koradi *et al.*, 1996).

by different principles of domain definition. Figure 1 shows an example of such discrepancy: a catalytic core of eukaryotic DNA topoisomerase I from Vaccina virus (PDB ID 1a41) is presented as one-domain protein in SCOP, where as PDP cuts it into two domains. Topology, similar to the second (green) domain occurs also in a short protein Bacteriocin AS-48 from Enterococcus Faecalis (PDB ID 1e68), which can be seen as an additional argument in favor of the biological meaning of this domain assignment. Top SCOP folds contributing to the discrepancies are: PLP-dependent transferase (fold representative 1bkgB), alpha-alpha superhelix (3bct), TIM beta/alpha-barrel (1aq0B), Periplasmic binding protein-like II (1jetA).

We also benchmarked the program using a collection of non-redundant proteins with manual domain assignment from the protein domain server http://www.bmm.icnet.uk/∼domains/. PDP assigned domains correctly for 81.4% of proteins. On the frequently used benchmark of 55 proteins (Jones *et al.*, 1998; Xu *et al.*, 2000), PDP assignment was correct in 47 proteins (85%), which is better than achieved by others (Jones *et al.*, 1998; Xu *et al.*, 2000). These 55 proteins were not used for optimization of the threshold values.

REFERENCES

Alexandrov,N.N., Nussinov,R. *et al.*, (1996) Fast protein fold recognition via sequence to structure alignment and contact capacity potentials. *Pac. Symp. Biocomput.*, 53–72.

Apweiler,R., Attwood,T.K. *et al.*, (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, **29**, 37–40.

Bateman,A., Birney,E. *et al.*, (2000) The Pfam protein families database. *Nucleic Acids Res.*, **28**, 263–266.

Crippen,G.M. (1978) The tree structural organization of proteins. *J. Mol. Biol.*, **126**, 315–332.

Holm,L. and Sander,C. (1994) Parser for protein folding units. *Proteins*, **19**, 256–268.

Holm,L. and Sander,C. (1996) Mapping the protein universe. *Science*, **273**, 595–603.

Islam,S.A., Luo,J. *et al.*, (1995) Identification and analysis of domains in proteins. *Protein Eng.*, **8**, 513–525.

Jones,S., Stewart,M. *et al.*, (1998) Domain assignment for protein structures using a consensus approach: characterization and analysis. *Protein Sci.*, **7**, 233–242.

Koradi,R., Billeter,M. *et al.*, (1996) MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.*, **14**, 29–32, 51–55.

Lesk,A.M. and Rose,G.D. (1981) Folding units in globular proteins. *Proc. Natl Acad. Sci. USA*, **78**, 4304–4308.

Murzin,A.G., Brenner,S.E. *et al.*, (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–640.

Orengo,C.A., Michie,A.D. *et al.*, (1997) CATH—a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.

Russel,R. (1999) A guide to structure prediction, http://www.bmm.icnet.uk/people/rob/CCP11BBS/.

Sowdhamini,R. and Blundell,T.L. (1995) An automatic method involving cluster analysis of secondary structures for the identification of domains in proteins. *Protein Sci.*, **4**, 506–520.

Xenarios,I., Rice,D.W. *et al.*, (2000) DIP: the database of interacting proteins. *Nucleic Acids Res.*, **28**, 289–291.

Xu,Y., Xu,D. *et al.*, (2000) Protein domain decomposition using a graph-theoretic approach. *Bioinformatics*, **16**, 1091–1104.