W130–W132 *Nucleic Acids Research, 2005, Vol. 33, Web Server issue*
doi:10.1093/nar/gki427

DIAL: a web-based server for the automatic identification of structural domains in proteins

Ganesan Pugalenthi, Govindaraju Archunan[1] and Ramanathan Sowdhamini*

National Centre for Biological Sciences, Tata Institute of Fundamental Research, UAS-GKVK campus, Bellary Road, Bangalore 560 065, Karnataka, India and [1]Department of Animal Science, Bharathidasan University, Trichirapalli, Tamilnadu, 620 024, India

Received February 14, 2005; Revised and Accepted March 24, 2005

ABSTRACT

DIAL is a web server for the automatic identification of structural domains given the 3D coordinates of a protein. Delineation of the structural domains and their exact boundaries are the starting points for the better realization of distantly related members of the domain families, for the rational design of the experiments and for clearer understanding of the biological function. The current server can examine crystallographic multiple chains and provide structural domain solutions that can also describe domain swapping events. The server can be accessed from http://www.ncbs.res.in/~faculty/mini/DIAL/home.html. The Supplementary data can be accessed from http://www.ncbs.res.in/~faculty/mini/DIAL/supplement.html.

INTRODUCTION

Many proteins, especially those involved in the signal transduction, contain compact units or multiple domains performing a wide variety of functions (1,2). In some cases, the functional domains perform biological function sequentially—each involved in a series of steps in a biochemical reaction. In other instances, the functional domains work together in a manner that some domains decide the function and the efficiency of co-existing domains such that the catalytic domains are recruited in a particular pathway selectively. In few other instances, the structurally compact domains may be distinct in their functions at most times but can influence or organize the function of neighbouring domains at the interface through substantial conformational changes. Therefore, the compilation of biological information of the protein domains (3–11) is an useful step in different areas of the interface between computing and biology, e.g. protein sequence analysis, structure prediction, modelling, rational design of experiments (such as protein crystallization and site-directed mutagenesis or deletion experiments) and perception of biological function (such as signal transduction and allostery).

Several objective methods identify the protein structural domains starting from the atomic coordinates of proteins (12–18). DIAL is one such procedure that identifies the structural domains in proteins by clustering substructures on the basis of their spatial distances (17). This has been further improvised and compared (19) with other protein domain resources (see Supplementary data for some examples). Popular public domain resources often require careful manual examination (6) or consultation of several algorithms internally (20). There are also structural domain databases available over the public domain (12,18,19). However, in order that the structural domain boundaries can be identified for newer proteins and for addressing the overall structural domain architecture of proteins with multiple chains, we report the availability of DIAL web server. DIAL web server provides additional information such as the presence of secondary structures, conserved residues and functional motifs for the individual domains that are mapped both on sequence and structure.

DIAL SERVER

The non-hydrogen atomic coordinates of the protein form the input for DIAL server. Alternately, sequence of the query can be employed as an input to identify the nearest structural homologue for the examination of structural domains. The nearest structural homologue is identified by initiating a PSI-BLAST search against the Protein Data Bank (PDB) database at an *E*-value threshold of 10^{-3}. Segment of the PDB hit that matches 100% with the entire length of the query protein is considered for the domain delineation. Where the protein is reported as crystallographic or physiological multimers and the transformation matrix is provided, the server internally generates the multimer coordinates and offers structural domain architecture solutions for the entire quartenary arrangement. For instance, Figure S1 (in Supplementary data) shows

*To whom correspondence should be addressed. Tel: +91 80 23636421 Ext. 4240/1; Fax: +91 80 23636462; Email: mini@ncbs.res.in

the entire domain architecture of the three protein multimers where extensive interactions between the protomers are evident and the individual domains are composed of multiple chains; structures of these three protein examples further indicate domain swapping events [(21); for the structural domains see Supplementary data].

FEATURES OF DIAL SERVER

(i) Secondary structures and connecting loops are clustered using their structural distances and domains identified as described previously (19) for both single and multiple chains. Subsequent to our previous report (19), in our extension to address multiple chain entries, we are also currently considering the short segments (one or two residues long) as individual substructures since the interactions between the multiple chains often require small regions of interactions.

(ii) The best structural domain architectures are projected as convenient bar diagrams (as shown in Figure S1 in Supplementary data) and alternate domain definitions are also provided in a similar manner. Alternate domain definitions provide a structural hierarchy of locally compact units and also sometimes permit the user to recognize other structural domain solutions.

(iii) Nearest structural homologues or SCOP (6) entries are also reported.

(iv) Sequence and structural files can be downloaded for individual domains and viewed through RASMOL and CHIME (22) interfaces. Static images of the domain definitions are provided using MOLSCRIPT (23).

(v) Additional features such as secondary structural topology and conserved residues are provided. Sequences are structure-annotated using JOY (24). PSI-BLAST (25) is performed against the structural entries (26) to identify homologues, sequences aligned using CLUSTALW (27) and conserved residues identified using MOTIFS (28).

(vi) Functionally important residues, by PROSITE (29) definitions, are projected for individual domains. Domain interface residues are proposed by comparing the solvent accessibility (30) of individual domains in the free form and the entire protein context. Residues that undergo appreciable burial owing to adjacent structural domains are highlighted in the DIAL server as possible domain interface residues.

CONCLUSIONS

Owing to the structural genomics initiative (31) and the recent high-throughput computational structure prediction of gene products, there will be increasing numbers of proteins whose structures are available and their biological function waiting to be determined. The availability of domain boundaries would be an useful starting point for such analyses. In general, DIAL domain definitions and boundaries compare very well with crystallographers' definition and other objective identification methods such as 3Dee (18) with a mean overlap score of 93 and 97%, respectively [(19); also for details see Table S1 in Supplementary data]. In small number of cases, e.g. PDB code 1bia in Table S1 (in Supplementary data), the domain definitions from pure distance-based methods such as DIAL cannot be compared with the functional domains defined by other resources. The accurate delineation of structural domains is often a non-trivial problem and requires expert opinion where the domains may be heavily interacting or discontinuous in sequence or involve multiple chains. The availability of a web server for the understanding of structural domain architecture of protein structures should be useful for the study of newer proteins. DIAL server is especially suited for the study of structural domain architecture of multiple chain systems that must give rise to a biologically more meaningful picture of structural domain organization.

ACKNOWLEDGEMENTS

R.S. is a Senior Research Fellow of the Wellcome Trust, UK. The Wellcome Trust supports G.P. Authors also thank NCBS (TIFR) for financial and infrastructural support. The Open Access publication charges for this article were waived by Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES

1. Ilsley,J.L., Sudol,M. and Winder,S.J. (2002) The WW domain: linking cell signalling to the membrane cytoskeleton. *Cell. Signal.*, **14**, 183–189.
2. Chamnongpol,S. and Li,X. (2004) SH3 domain protein-binding arrays. *Methods Mol. Biol.*, **264**, 183–189.
3. Heger,A. and Holm,L. (2003) Exhaustive enumeration of protein domain families. *J. Mol. Biol.*, **328**, 749–767.
4. Russell,R.B. (1994) Domain insertion. *Protein Eng.*, **7**, 1407–1410.
5. Finn,R.D., Marshall,M. and Bateman,A. (2005) iPfam: visualization of protein–protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics*, **21**, 410–412.
6. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
7. Marchler-Bauer,A., Anderson,J.B., Cherukuri,P.F., DeWeese-Scott,C., Geer,L.Y., Gwadz,M., He,S., Hurwitz,D.I., Jackson,J.D., Ke,Z. *et al.* (2005) CDD: a conserved domain database for protein classification. *Nucleic Acids Res.*, **33**, D192–D196.
8. 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.
9. Liu,J. and Rost,B. (2004) CHOP: parsing proteins into structural domains. *Nucleic Acids Res.*, **32**, W569–W571.
10. Dengler,U., Siddiqui,A.S. and Barton,G.J. (2001) Protein structural domains: analysis of the 3Dee domains database. *Proteins*, **42**, 332–344.
11. Pugalenthi,G., Bhaduri,A. and Sowdhamini,R. (2005) GenDiS: genomic distribution of protein structural domain superfamilies. *Nucleic Acids Res.*, **33**, D252–D255.
12. Alexandrov,N. and Shindyalov,I. (2003) PDP: Protein Domain Parser. *Bioinformatics*, **19**, 429–430.
13. Taylor,W.R. (1999) Protein structural domain identification. *Protein Eng.*, **12**, 203–216.
14. Schulz,G.E. (1977) Structural rules for globular proteins. *Angew. Chem. Int. Ed. Engl.*, **16**, 23–32.
15. Swindells,M.B. (1995) A procedure for detecting structural domains in proteins. *Protein Sci.*, **4**, 103–112.
16. Xu,Y., Xu,D. and Gabow,H.N. (2000) Protein domain decomposition using a graph–theoretic approach. *Bioinformatics*, **16**, 1091–1104.
17. 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.
18. Siddiqui,A.S., Dengler,U. and Barton,G.J. (2001) 3Dee: a database of protein structural domains. *Bioinformatics*, **17**, 200–201.
19. Vinayagam,A., Shi,J., Pugalenthi,G., Meenakshi,B., Blundell,T.L. and Sowdhamini,R. (2003) DDBASE2.0: updated domain database with

improved identification of structural domains. *Bioinformatics*, **19**, 1760–1764.
20. Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) CATH—a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.
21. Liu,Y. and Eisenberg,D. (2002) 3D domain swapping: as domains continue to swap. *Protein Sci*., **11**, 1285–1299.
22. Sayle,R.A. and Milner-White,E.J. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem. Sci*., **20**, 374.
23. Kraulis,P.J. (1991) MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr*., **24**, 946–950.
24. Mizuguchi,K., Deane,C.M., Blundell,T.L., Johnson,M.S. and Overington,J.P. (1998) JOY: protein sequence-structure representation and analysis. *Bioinformatics*, **14**, 617–623.
25. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.
26. Bourne,P.E., Addess,K.J., Bluhm,W.F., Chen,L., Deshpande,N., Feng,Z., Fleri,W., Green,R., Merino-Ott,J.C., Townsend-Merino,W. *et al*. (2004) The distribution and query systems of the RCSB Protein Data Bank. *Nucleic Acids Res*., **32**, D223–D225.
27. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*., **22**, 4673–4680.
28. Bhaduri,A., Pugalenthi,G., Gupta,N. and Sowdhamini,R. (2004) iMOT: an interactive package for the selection of spatially interacting motifs. *Nucleic Acids Res*., **32**, W602–W605.
29. Sigrist,C.J., Cerutti,L., Hulo,N., Gattiker,A., Falquet,L., Pagni,M., Bairoch,A. and Bucher,P. (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief. Bioinformatics*, **3**, 265–274.
30. Lee,B. and Richards,F.M. (1971) The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol*., **55**, 379–400.
31. Chance,M.R., Fiser,A., Sali,A., Pieper,U., Eswar,N., Xu,G., Fajardo,J.E., Radhakannan,T. and Marinkovic,N. (2004) High-throughput computational and experimental techniques in structural genomics. *Genome Res*., **14**, 2145–2154.