Nucleic Acids Research, 2006, Vol. 34, Database issue ***D285–D286***
doi:10.1093/nar/gkj125

iMOTdb—a comprehensive collection of spatially interacting motifs in proteins

Ganesan Pugalenthi, Anirban Bhaduri and R. Sowdhamini*

National Centre for Biological Sciences, Tata Institute of Fundamental Research, UAS-GKVK Campus, Bellary Road, Bangalore 560 065, India

Received August 15, 2005; Accepted October 21, 2005

ABSTRACT

Realization of conserved residues that represent a protein family is crucial for clearer understanding of biological function as well as for the better recognition of additional members in sequence databases. Functionally important residues are recognized well due to their high degree of conservation in closely related sequences and are annotated in functional motif databases. Structural motifs are central to the integrity of the fold and require careful analysis for their identification. We report the availability of a database of spatially interacting motifs in single protein structures as well as those among distantly related protein structures that belong to a superfamily. Spatial interactions amongst conserved motifs are automatically measured using sequence similarity scores and distance calculations. Interactions between pairs of conserved motifs are described in the form of pseudoenergies. iMOTdb database provides information for 854 488 motifs corresponding to 60 849 protein structural domains and 22 648 protein structural entries.

INTRODUCTION

The central dogma in protein folding problem is how proteins arrive at their unique three-dimensional fold spontaneously. Anfinsens' hypothesis has stated that the entire information about the tertiary structure of a protein is contained in its amino acid sequence (1). Proteins are largely tolerant to mutations and a large amount of information in homologous protein families reveals that mutations are more likely in structurally variable regions (2–8). Structurally invariant regions point to solvent-buried residues that undergo permitted amino acid exchanges. We had earlier identified such structurally invariant residues amongst superfamilies where proteins are distantly related but retain similar biological functions (9,10). The structurally invariant residues undergo permitted amino acid mutations where the amino acids exchanged still retain similar chemical groups.

Functionally important residues can be recognized from mutagenesis experiments or simply from their high sequence and structural conservation among protein families and superfamilies. Information on such functional residues can be obtained from popular motif databases (11). However, conserved residues crucial for the structural integrity are hard to recognize since they undergo permitted amino acid exchanges. We had earlier employed conserved residues that are spatially interacting with other motifs in the fold to recognize additional putative members of a protein family (12) and developed a web server for the automatic identification of spatially interacting conserved residues (13). There have been similar attempts by other groups on the visualization of conserved regions on protein structures (14). In this paper, we report the availability of a database containing spatially proximate conserved motifs where iMOT has been applied to whole database of protein structural superfamilies (7,10) and all structural entries in the Protein Structural Databank (15).

CONTENTS OF THE DATABASE

This database provides interacting motifs for 60 849 protein structural domain superfamilies derived from SCOP database 1.67 release (7). All the 1731 problematic entries in the SCOP database could not be considered for our database owing to spurious values in the calculations or lack of spatial interactions of conserved residues or lack of homologues or entries with only C^{α} coordinates. For each structural member in the superfamily that has been considered in SCOP database, homologous sequences are individually identified. Alignment positions are provided an average similarity score after consulting amino acid exchange matrix (16). Contiguous residues with an average similarity score of more than 50

*To whom correspondence should be addressed. Tel: +91 80 23636421; Fax: +91 80 23636462; Email: mini@ncbs.res.in
Present address:
Anirban Bhaduri, Information and Mathematical Sciences, Genome Institute of Singapore, Genome, 60 Biopolis Street, Singapore

© The Author 2006. Published by Oxford University Press. All rights reserved.

The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article for non-commercial purposes provided that: the original authorship is properly and fully attributed; the Journal and Oxford University Press are attributed as the original place of publication with the correct citation details given; if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated. For commercial re-use, please contact journals.permissions@oxfordjournals.org

are treated as conserved residues or motifs. These motifs are mapped on to the structural superfamily member to examine their spatial proximity with each other. Pairs of conserved residues are further examined by calculating psuedoenergies that describe the strength of interactions (13). Spatially interacting motifs are mapped on to the alignment of the superfamily to further recognize spatially interacting motifs that are conserved throughout the superfamily [for details, please see the help web pages and (12,13)]. Interacting motifs are provided for all the 22 648 protein structures submitted in PDB database (May 2005 release). iMOTdb pertains to 854 488 motifs of 60 849 protein structural domains corresponding to SCOP 1.67 database (7).

FEATURES OF THE DATABASE

- Spatially interacting motifs identified in protein structures are mapped and colour-coded on sequence alignment as well as on structure [using MOLSCRIPT (17) and CHIME (MDL Information Systems, Inc.)].
- The extent of spatial interaction between all possible pairs of motifs is provided as a symmetric matrix where the values are described as pseudoenergies (13). Pseudoenergies are classified, by benchmarking on known structural motifs, as strong (better than −125), medium (between −125 and −50) and weak (worse than −50) and colour-coded accordingly.
- Structural information about individual motifs is provided that includes the presence of motifs in secondary structures, solvent accessibility patterns and positional variations amongst superfamily members (reflected as root mean square deviations).
- This database provides the user with an option to search genome databases using selected interacting motifs as in SCANMOT server (18) and using PHIBLAST (19).
- Hyperlinks to other online resources, such as PROSITE (11), CKAAPsDB (20), PRINTS and (21) eMOTIFS (22), are provided so that direct comparison of motif definitions and peptide signatures (23) may be possible.

APPLICATIONS

Spatially interacting motifs can be critical for structure and/or function. They are useful in searching for distant homologues and establishing remote homologies among the largely unassigned sequences in genome databases. Availability of information on structural motifs in large number of protein structures should be useful as starting points to perform detailed analysis, for the rational design of experiments in protein folding, site-directed mutagenesis and to understand mechanism of action and conformational changes in proteins. iMOTdb database can be accessed from http://caps.ncbs.res.in/imotdb/.

ACKNOWLEDGEMENTS

R.S. is a Senior Research Fellow of the Wellcome Trust, UK. We also thank NCBS (TIFR) for financial and infrastructural support. The stay of G.P. was supported by Wellcome Trust, UK. Funding to pay the Open Access publication charges for this article was provided by Wellcome Trust, UK.

Conflict of interest statement. None declared.

REFERENCES

1. Anfinsen,C.B. (1973) Principles that govern the folding of protein chains. *Science*, **181**, 223–230.
2. Rossmann,M.G. and Argos,P. (1977) The taxonomy of protein structure. *J. Mol. Biol*, **109**, 99–129.
3. Richardson,J.S. (1981) The anatomy and taxonomy of protein structure. *Adv. Prot. Chem*, **34**, 167–339.
4. Chothia,C. (1984) Principles that determine the structure of proteins. *Annu. Rev. Biochem*, **53**, 537–572.
5. Holm,L. and Sander,C. (1998) Touring protein fold space with Dali/FSSP. *Nucleic Acids Res*, **26**, 316–319.
6. Overington,J., Johnson,M.S., Sali,A. and Blundell,T.L. (1990) Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. *Proc. Biol. Sci*.,, **B241**, 132–145.
7. 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.
8. 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.
9. Chakrabarti,S., Venkataramanan,K. and Sowdhamini,R. (2003) SMoS: a database of structural motifs of superfamilies. *Protein Eng*, **16**, 791–793.
10. Bhaduri,A., Pugalenthi,G. and Sowdhamini,R. (2004) PASS2: an automated database of protein alignments organised as structural superfamilies. *BMC Bioinformatics*, **5**, 35.
11. Falquet,L., Pagni,M., Bucher,P., Hulo,N., Sigrist,C.J., Hofmann,K. and Bairoch,A. (2002) *The PROSITE database, its status in 2002*. *Nucleic Acids Res*, **30**, 235–238.
12. Bhaduri,A., Ravishankar,R. and Sowdhamini,R. (2004) Conserved spatially interacting motifs of protein superfamilies: application to fold recognition and function annotation of genome data. *Proteins*, **54**, 657–670.
13. 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.
14. Bennett,S.P., Nevill-Manning,C.G. and Brutlag,D.L. (2003) 3MOTIF: visualizing conserved protein sequence motifs in the protein structure database. *Bioinformatics*, **19**, 541–542.
15. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res*, **28**, 235–242.
16. Johnson,M.S., Overington,J.P. and Blundell,T.L. (1993) Alignment and searching for common protein folds using a data bank of structural templates. *J. Mol. Biol*, **231**, 735–752.
17. Kraulis,P.J. (1991) MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr*., **24**, 946–950.
18. Chakrabarti,S., Anand,A.P., Bhardwaj,N., Pugalenthi,G. and Sowdhamini,R. (2005) SCANMOT: searching for similar sequences using a simultaneous scan of multiple sequence motifs. *Nucleic Acid Res*, **33**, W274–W276.
19. Zhang,Z., Schaffer,A.A., Miller,W., Madden,T.L., Lipman,D.J., Koonin,E.V. and Altschul,S.F. (1998) Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Res*, **26**, 3986–3990.
20. Li,W.W., Reddy,B.V., Tate,J.G., Shindyalov,I.N. and Bourne,P.E. (2002) CKAAPs DB: a Conserved Key Amino Acid Positions DataBase. *Nucleic Acids Res*, **30**, 409–411.
21. Attwood,T.K. (2002) The PRINTS database: a resource for identification of protein families. *Brief. Bioinform*, **3**, 252–263.
22. Huang,J.Y. and Brutlag,D.L. (2001) The EMOTIF database. *Nucleic Acids Res*, **29**, 202–204.
23. Prakash,T., Khandelwal,M., Dasgupta,D., Dash,D. and Brahmachari,S.K. (2004) CoPS: Comprehensive Peptide Signature database. *Bioinformatics*, **20**, 2886–2888.