W126–W129 *Nucleic Acids Research, 2005, Vol. 33, Web Server issue*
doi:10.1093/nar/gki474

PRODOC: a resource for the comparison of tethered protein domain architectures with in-built information on remotely related domain families

O. Krishnadev, N. Rekha, S. B. Pandit, S. Abhiman, S. Mohanty, L.S. Swapna, S. Gore[1] and N. Srinivasan*

Molecular Biophysics Unit and [1]Super Computer Education and Research Center, Indian Institute of Science, Bangalore 560 012, India

Received February 13, 2005; Revised March 23, 2005; Accepted April 15, 2005

ABSTRACT

PROtein Domain Organization and Comparison (PRODOC) comprises several programs that enable convenient comparison of proteins as a sequence of domains. The in-built dataset currently consists of ~698 000 proteins from 192 organisms with complete genomic data, and all the SWISSPROT proteins obtained from the Pfam database. All the entries in PRODOC are represented as a sequence of functional domains, assigned using hidden Markov models, instead of as a sequence of amino acids. On average 69% of the proteins in the proteomes and 49% of the residues are covered by functional domain assignments. Software tools allow the user to query the dataset with a sequence of domains and identify proteins with the same or a jumbled or circularly permuted arrangement of domains. As it is proposed that proteins with jumbled or the same domain sequences have similar functions, this search tool is useful in assigning the overall function of a multi-domain protein. Unique features of PRODOC include the generation of alignments between multi-domain proteins on the basis of the sequence of domains and in-built information on distantly related domain families forming superfamilies. It is also possible using PRODOC to identify domain sharing and gene fusion events across organisms. An exhaustive genome–genome comparison tool in PRODOC also enables the detection of successive domain sharing and domain fusion events across two organisms. The tool permits the identification of gene clusters involved in similar biological processes in two closely related organisms. The URL for PRODOC is http://hodgkin.mbu.iisc.ernet.in/~prodoc.

INTRODUCTION

Modular representation of gene products as sequences of functional domains, instead of as sequences of amino acids, is useful in understanding the molecular basis of the functions of multi-domain proteins (1–3). Knowledge of the functions of the individual domains of a multi-domain protein contributes to our understanding of the properties of the protein as a whole (4–6). Viewing multi-domain proteins as sequences of domains also enables the identification of gene fusion events, interacting proteins (7,8) and preferred domain associations (9–14), and the comparison of sequences of domains helps in obtaining clues about domain function. For example, in two multi-domain proteins with many common domains, alignment of a region of unknown function with a domain of known function raises the possibility of a distant relationship between the region of unknown function and the aligned domain.

Realizing the importance of viewing proteins as sequences of domains, many databases of protein domain families and sequences of protein domains have been developed, such as PRODOM (15), DOMO (16), BLOCKS (17), Pfam (18), SMART (19), InterPRO (20), PRINTS (21) and DART (22).

*To whom correspondence should be addressed. Tel: +91 80 2293 2837; Fax: +91 80 2360 0535; Email: ns@mbu.iisc.ernet.in
Present addresses:
S. B. Pandit, Center of Excellence in Bioinformatics, University at Buffalo, New York, USA
S. Abhiman, Centre for Genomics and Bioinformatics, Karolinska Institutet, Stockholm, Sweden
S. Gore, Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge, UK

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors

The entire compendium of proteins listed in SWISSPROT (23) is available in SWISSPFAM, wherein every SWISSPROT entry is represented as a sequence of domains. Domain assignments to various proteomes are also available in the form of databases (24–26).

Several software tools are available in PRODOC (PROtein Domain Organization and Comparison) to facilitate searching for a given sequence of domains in various genomes, identification of domain fusion events, recognition of gene products with identical or similar domain compositions and identification of proteins with a circularly permuted or jumbled arrangement of the order of domains. A tool for complete genome–genome comparison is also available in PRODOC. By considering two genomes at a time the program can identify series of gene products that exhibit domain sharing. This process enables the proposal of functional gene clusters in the two genomes. This is radically different from COG (27) as we consider the sharing of a domain family to be a criterion in identifying series of gene fusion events in a set of genes from two organisms. The database component of PRODOC is the sequence of functional domains of proteins encoded in a large number of organisms, as well as the entire set of proteins in the SWISSPROT database. The objective behind the generation of the PRODOC suite of programs is that it should provide a convenient platform to perform domain analysis at the genomic scale for the applications mentioned above.

The most distinguishing feature of PRODOC compared with similar resources for domain analysis is the use of the notion of remotely related domain families forming superfamilies. A superfamily is constituted by families which exhibit similarity in the functions and structures of protein domains (28). We have incorporated in PRODOC knowledge of such distantly related protein domain families in a superfamily with and without known three-dimensional structures (29,30). Thus it is possible to recognize those sequences of domains with one or more domains belonging to the same superfamily as those in the query. Such searches enable the user to study the evolution of the functions of multi-domain proteins. It has been suggested that homologous protein domains with extensive sequence divergence, forming protein domain superfamilies, are involved in novel domain combinations during gene fusion events while retaining the broad nature of the function (14). It is suggested that such variations in domain recruitment and high sequence divergence form turning points in otherwise similar biochemical pathways (14).

THE CONSTRUCTION AND ORGANIZATION OF PRODOC

The various tools and datasets present in PRODOC and the software's overall organization are shown in Figure 1, and these features are discussed below.

Domain assignments to genomes

The amino acid sequences of predicted gene products in the completely sequenced genomes of various organisms are available in public databases such as NCBI, ENSEMBL (31,32), FlyBase (33) and PlasmoDB (34). The hidden Markov models (HMMs) for protein domain families available in the PfamA dataset have been used in generating a database of HMMs for 7677 domain families available in Pfam (18) version 16. HMMER (35) enables the mapping of various domains along the amino acid sequence of the query. An *E*-value threshold of 10^{-2} is considered reasonable for the assignment of domains. In addition, it is ensured that the alignment is of considerable length (36). For the current and first major release of PRODOC, the domain assignments for the proteins from various genomes and for those proteins listed in SWISSPROT have been obtained from Pfam and SWISSPFAM, respectively. The domain assignments are confined to the regions showing a significant match with the HMMs of protein families, leaving a proportion of the gene products with no domain assignment. At the time of preparation of this article, the PRODOC database contained functional domain assignments for 192 completed proteomes (156 eubacterial, 19 archaeal and 17 eukaryotic proteomes) consisting of 697 976 proteins. Typically, 69% and 68% of the proteins of a proteome are covered by HMM-based domain assignments in prokaryotic and eukaryotic organisms, respectively. In every protein the domain assignments could be made for a substantial proportion, and on average 49% of the residues are covered by domain assignments. In the future the dataset will be updated periodically using HMMER2 running on locally available multi-processor systems.

Tool for the comparison of proteins with linear and shuffled domain order

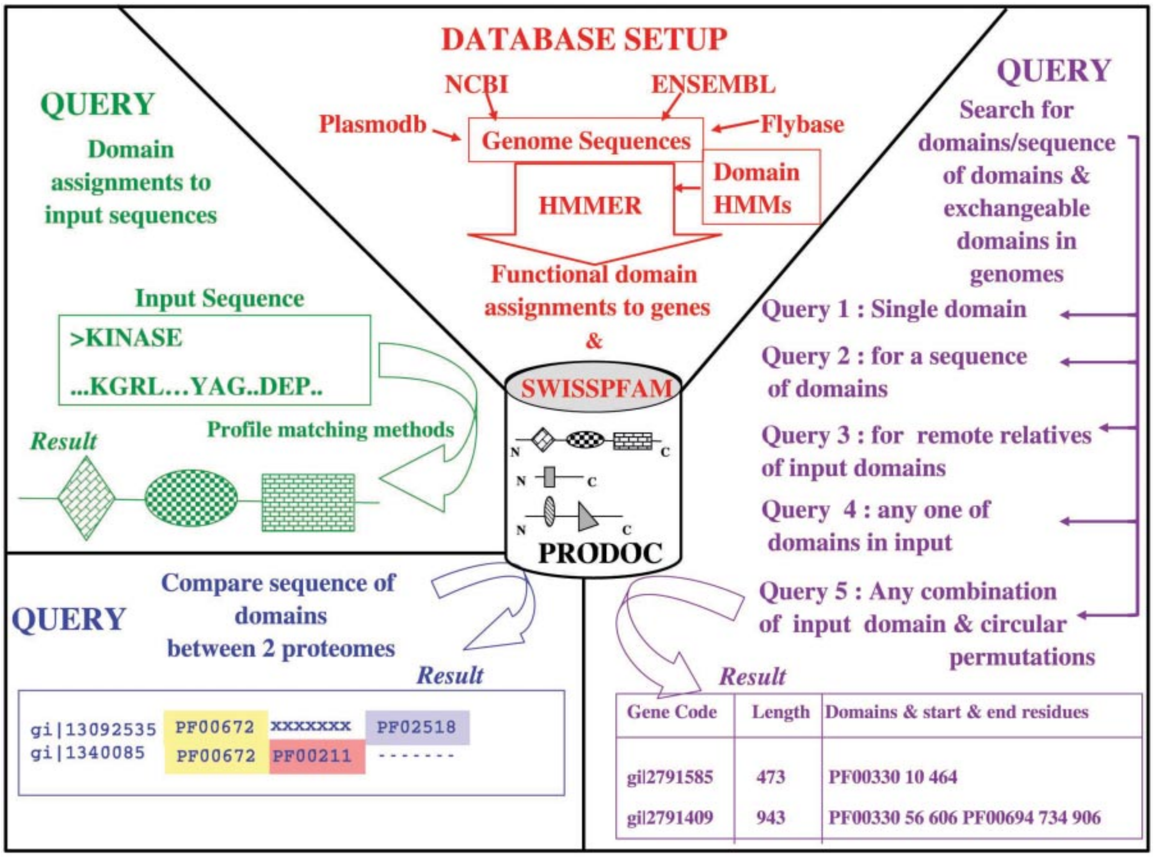
One of the tools available in PRODOC allows the user to query the datasets for occurrences of a sequence of domains. It has been observed that in many similar multi-domain proteins, the order of occurrence of domains in the primary structure is different. Such cases cannot be easily detected by simple amino acid sequence search methods, but a tool has been built in PRODOC to search for such cases. The user is allowed to input a number of domains as a query. Following this step, a search is made in the dataset of interest to identify all the multi-domain proteins with a different or cyclically permuted order of domains compared with the query protein. It is known that the overall functions of two proteins related by jumbled domain architectures are often similar (37).

Tool for the comparison of the sequences of domain families considering superfamily relationships

When comparing the sequences of the domains of two multi-domain proteins, it is possible that some of the domains in one protein are distantly related to domain(s) in the other protein (superfamilies). We have formed a dataset of distantly related Pfam domain families by relating Pfam families with proteins of known three-dimensional structure and by identifying new potential sequence superfamilies (29,30). This information is used in the domain architecture search tool to result in the identification of distantly related multi-domain proteins with one or more domains related by a superfamily connection (14).

Clues to the functions of domain-unassigned regions

Tools in PRODOC can also aid remote homology detection and function annotation based on alignment of a domain with a region with no domain assignment. For example, in two



```
gi|13092535 PF00672 xxxxxxx PF02518
gi|1340085  PF00672 PF00211 -------
```

Gene Code	Length	Domains & start & end residues
gi|2791585	473	PF00330 10 464
gi|2791409	943	PF00330 56 606 PF00694 734 906

Figure 1. The organization of PRODOC and the utilities offered.

multi-domain proteins with many common domains, alignment of a region of unknown function with a domain of known function raises the possibility of a distant relationship between the region of unknown function and the aligned domain. Thus PRODOC can be helpful in suggesting new possibilities for the functional annotation of domain-unassigned regions.

Tools for the identification of domain sharing, gene fusion and functional clusters

Putative gene fusion events across two organisms can be identified using PRODOC. This can be illustrated as follows. Let us consider that two different gene products in organism A encode for domain families P and Q, respectively. If, in a closely related organism B, a protein with domain families P and Q fused as a single gene product can be identified, this forms a potential gene fusion event and the possibility of functional interaction between the two gene products in organism A is raised (7,8,11,38,39).

PRODOC also facilitates searches to identify several domain fusion events successively. For example, it is possible that, in organism B, the gene product with domain families P and Q is also tethered to another domain family, R. In such a situation one can search in organism A for a gene product containing the domain family R. If such a gene product can be found in organism A and domain family S is tethered to R, a further search can be made in organism B for a protein with domain S, and so on. Such a repetitive search for successive domain fusion events across two organisms will eventually result in two sets of genes from two organisms with several domains shared between the sets. Such sets of gene products can be considered functional clusters of proteins involved in similar series of events in similar biological pathways across the two organisms.

Using PRODOC, the user can easily compare domain organization between two genomes of interest. Pairs of proteins that contain at least a common domain are displayed as output. This information can be harnessed to derive cases of gene fusion and functional gene clusters.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We thank Ms V. S. Gowri and Dr S. Balaji for their timely update of a database of structure-based alignments of homologous proteins. We thank the anonymous referees of an earlier version of this article and others who assessed the PRODOC site for useful inputs. This research is supported by the award of a Senior Fellowship in Biomedical Sciences to N.S. by the Wellcome Trust, UK, as well as by the computational genomics project supported by the Department of Biotechnology, Government of India. N.R. and S.B.P. are supported by fellowships from the Council of Scientific and Industrial Research, Government of India. The Open Access publication charges for this article were waived by Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES

1. Copley,R.R., Doerks,T., Letunic,I. and Bork,P. (2002) Protein domain analysis in the era of complete genomes. *FEBS Lett.*, **513**, 129–134.
2. Copley,R.R., Goodstadt,L. and Ponting,C. (2003) Eukaryotic domain evolution inferred from genome comparisons. *Curr. Opin. Genet. Dev.*, **13**, 623–628.
3. Copley,R.R., Ponting,C.P., Schultz,J. and Bork,P. (2002) Sequence analysis of multidomain proteins: past perspectives and future directions. *Adv. Protein Chem.*, **61**, 75–98.
4. Ponting,C.P. and Russell,R.R. (2002) The natural history of protein domains. *Annu. Rev. Biophys Biomol. Struct.*, **31**, 45–71.
5. Vogel,C., Bashton,M., Kerrison,N.D., Chothia,C. and Teichmann,S.A. (2004) Structure, function and evolution of multidomain proteins. *Curr. Opin. Struct. Biol.*, **14**, 208–216.
6. Pawson,T. and Nash,P. (2003) Assembly of cell regulatory systems through protein interaction domains. *Science*, **300**, 445–452.
7. Enright,A.J., Iliopoulos,I., Kyrpides,N.C. and Ouzounis,C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.
8. Marcotte,E.M., Pellegrini,M., Ng,H.L., Rice,D.W., Yeates,T.O. and Eisenberg,D. (1999) Detecting protein function and protein–protein interactions from genome sequences. *Science*, **285**, 751–753.
9. Ye,Y. and Godzik,A. (2004) Comparative analysis of protein domain organization. *Genome Res.*, **14**, 343–353.
10. Snel,B., Bork,P. and Huynen,M.A. (2002) The identification of functional modules from the genomic association of genes. *Proc. Natl Acad. Sci. USA*, **99**, 5890–5895.
11. Tsoka,S. and Ouzounis,C.A. (2000) Prediction of protein interactions: metabolic enzymes are frequently involved in gene fusion. *Nature Genet.*, **26**, 141–142.
12. Apic,G., Huber,W. and Teichmann,S.A. (2003) Multi-domain protein families and domain pairs: comparison with known structures and a random model of domain recombination. *J. Struct. Funct. Genomics*, **4**, 67–78.
13. Apic,G., Gough,J. and Teichmann,S.A. (2001) Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J. Mol. Biol.*, **310**, 311–325.
14. Rekha,N., Suman,M.M., Cyndhavi,N., Krupa,A. and Srinivasan,N. (2005) Interaction interfaces of protein domains are not topologically equivalent across families within superfamilies: implications for metabolic and signalling pathways. *Proteins*, **58**, 339–353.
15. Sonnhammer,E.L. and Kahn,D. (1994) Modular arrangement of proteins as inferred from analysis of homology. *Protein Sci.*, **3**, 482–492.
16. Gracy,J. and Argos,P. (1998) DOMO: a new database of aligned protein domains. *Trends Biochem. Sci.*, **23**, 495–497.
17. Henikoff,J.G., Henikoff,S. and Pietrokovski,S. (1999) New features of the Blocks Database servers. *Nucleic Acids Res.*, **27**, 226–228.
18. Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Finn,R.D. and Sonnhammer,E.L. (1999) Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucleic Acids Res.*, **27**, 260–262.
19. Letunic,I., Copley,R.R., Schmidt,S., Ciccarelli,F.D., Doerks,T., Schultz,J., Ponting,C.P. and Bork,P. (2004) SMART 4.0: towards genomic data integration. *Nucleic Acids Res.*, **32**, 142–144.
20. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al.* (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.*, **31**, 315–318.
21. Attwood,T.K., Flower,D.R., Lewis,A.P., Mabey,J.E., Morgan,S.R., Scordis,P., Selley,J.N. and Wright,W. (1999) PRINTS prepares for the new millennium. *Nucleic Acids Res.*, **27**, 220–225.
22. Geer,L.Y., Domrachev,M., Lipman,D.J. and Bryant,S.H. (2002) CDART: protein homology by domain architecture. *Genome Res.*, **12**, 1619–1623.
23. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
24. Buchan,D.W., Rison,S.C., Bray,J.E., Lee,D., Pearl,F., Thornton,J.M. and Orengo,C.A. (2003) Gene3D: structural assignments for the biologist and bioinformaticist alike. *Nucleic Acids Res.*, **31**, 469–473.
25. Kersey,P.J., Morris,L., Hermjakob,H. and Apweiler,R. (2003) Integr8: enhanced inter-operability of European molecular biology databases. *Methods Inf. Med.*, **42**, 154–160.
26. Madera,M., Vogel,C., Kummerfeld,S.K., Chothia,C. and Gough,J. (2004) The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Res.*, **32**, 235–239.
27. Tatusov,R.L., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Kiryutin,B., Koonin,E.V., Krylov,D.M., Mazumder,R., Mekhedov,S.L., Nikolskaya,A.N. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
28. 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.
29. Pandit,S.B., Gosar,D., Abhiman,S., Sujatha,S., Dixit,S.S., Mhatre,N.S., Sowdhamini,R. and Srinivasan,N. (2002) SUPFAM—a database of potential protein superfamily relationships derived by comparing sequence-based and structure-based families: implications for structural genomics and function annotation in genomes. *Nucleic Acids Res.*, **30**, 289–293.
30. Pandit,S.B., Bhadra,R., Gowri,V.S., Balaji,S., Anand,B. and Srinivasan,N. (2004) SUPFAM: a database of sequence superfamilies of protein domains. *BMC Bioinformatics*, **5**, 28.
31. Birney,E., Andrews,D., Bevan,P., Caccamo,M., Cameron,G., Chen,Y., Clarke,L., Coates,G., Cox,T., Cuff,J. *et al.* (2004) Ensembl 2004. *Nucleic Acids Res.*, **32**, 468–470.
32. Hubbard,T., Barker,D., Birney,E., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.
33. FlyBase Consortium (2003) The FlyBase database of the *Drosophila* genome projects and community literature, *Nucleic Acids Res.*, **31**, 172–175.
34. Bahl,A., Brunk,B., Crabtree,J., Fraunholz,M.J., Gajria,B., Grant,G.R., Ginsburg,H., Gupta,D., Kissinger,J.C., Labo,P. *et al.* (2003) PlasmoDB: the *Plasmodium* genome resource. A database integrating experimental and computational data. *Nucleic Acids Res.*, **31**, 212–215.
35. Hofmann,K. (2000) Sensitive protein comparisons with profiles and hidden Markov models. *Brief. Bioinform.*, **1**, 167–178.
36. Madera,M. and Gough,J.A. (2002) Comparison of profile hidden Markov model procedures for remote homology detection. *Nucleic Acids Res.*, **30**, 4321–4328.
37. Bashton,M. and Chothia,C. (2002) The geometry of domain combination in proteins. *J. Mol. Biol.*, **315**, 927–939.
38. Enright,A.J. and Ouzounis,C.A. (2001) Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions. *Genome Biol.*, **2**, RESEARCH0034.
39. Yanai,I., Derti,A. and DeLisi,C. (2001) Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes. *Proc. Natl Acad. Sci. USA*, **98**, 7940–7945.