D252–D255 *Nucleic Acids Research, 2005, Vol. 33, Database issue*
doi:10.1093/nar/gki087

GenDiS: Genomic Distribution of protein structural domain Superfamilies

Ganesan Pugalenthi, Anirban Bhaduri and Ramanathan Sowdhamini*

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

Received August 13, 2004; Revised and Accepted October 12, 2004

ABSTRACT

Several proteins that have substantially diverged during evolution retain similar three-dimensional structures and biological function inspite of poor sequence identity. The database on Genomic Distribution of protein structural domain Superfamilies (GenDiS) provides record for the distribution of 4001 protein domains organized as 1194 structural superfamilies across 18 997 genomes at various levels of hierarchy in taxonomy. GenDiS database provides a survey of protein domains enlisted in sequence databases employing a 3-fold sequence search approach. Lineage-specific literature is obtained from the taxonomy database for individual protein members to provide a platform for performing genomic and phyletic studies across organisms. The database documents residual properties and provides alignments for the various superfamily members in genomes, offering insights into the rational design of experiments and for the better understanding of a superfamily. GenDiS database can be accessed at http://www.ncbs.res.in/~faculty/mini/gendis/home.html.

INTRODUCTION

High-throughput large-scale sequencing efforts have illustrated the enormous diversity embedded within genomes owing to varied composition of the proteome. Fortunately, structural and sequence analyses suggest strong convergence, indicating that many proteins will share limited number of folds (1). Curation of protein structural entries in a hierarchy (2,3), compilation of sequence families (4,5) and superfamilies (6,7), establishing relationships between protein sequence and structural databases (8,9) and the analysis of genomic patterns (10,11) form representative approaches to understand the process of this strong convergence. Reliable association of unannotated protein sequences to pre-existing families of well-characterized structure and function allows the mapping of functionally important residues on sequence alignments that can provide important insights into functional mechanisms. However, similarity and inheritance of function among homologues related in the twilight zone have to be considered after careful validation (12).

Genomes are classified into taxons on the basis of morphology and genetic content under the taxonomy database (13). Classification of the organism at various taxonomic strata elaborates diversity among the organisms along with their proteomic content (14). Genome content and distribution of proteins provide better understanding of species phylogeny (15). Exploring the distribution of structural superfamilies across varied strata of taxons provides an addendum into our understanding of proteins and phylogeny of the organism. The database of Genomic Distribution of protein structural domain Superfamilies (GenDiS) aims to provide structural assignments to genes listed within the non-redundant protein sequence database at the superfamily level. Structural superfamily definitions are in correspondence with SCOP 1.63 (16) and PASS2 (17) databases. Search for homologues within the sequence databases have been performed using multiple approaches (see Methods). Assignments have been subsequently validated before inducting a member. Genomic lineage for every individual entry was obtained from the taxonomy database and corresponding taxon records were assigned. The database offers a platform for understanding and comparing the distribution of protein superfamilies across the different taxonomic strata.

METHODS

Searching for potential superfamily members in sequence databases

Potential members of the superfamilies have been searched using a 3-fold approach. Members of PASS2 database (17) have been queried in April 2003 release of non-redundant sequence database (13) employing PSI-BLAST (18) setting an expectation value of 10^{-3} for 20 iterations. The profile-to-sequence searches were complemented employing the HMMsearch tool of the HMMER suite (19). Hidden Markov

*To whom correspondence should be addressed. Tel: +91 80 3636421; Fax: +91 80 3636662; Email: mini@ncbs.res.in

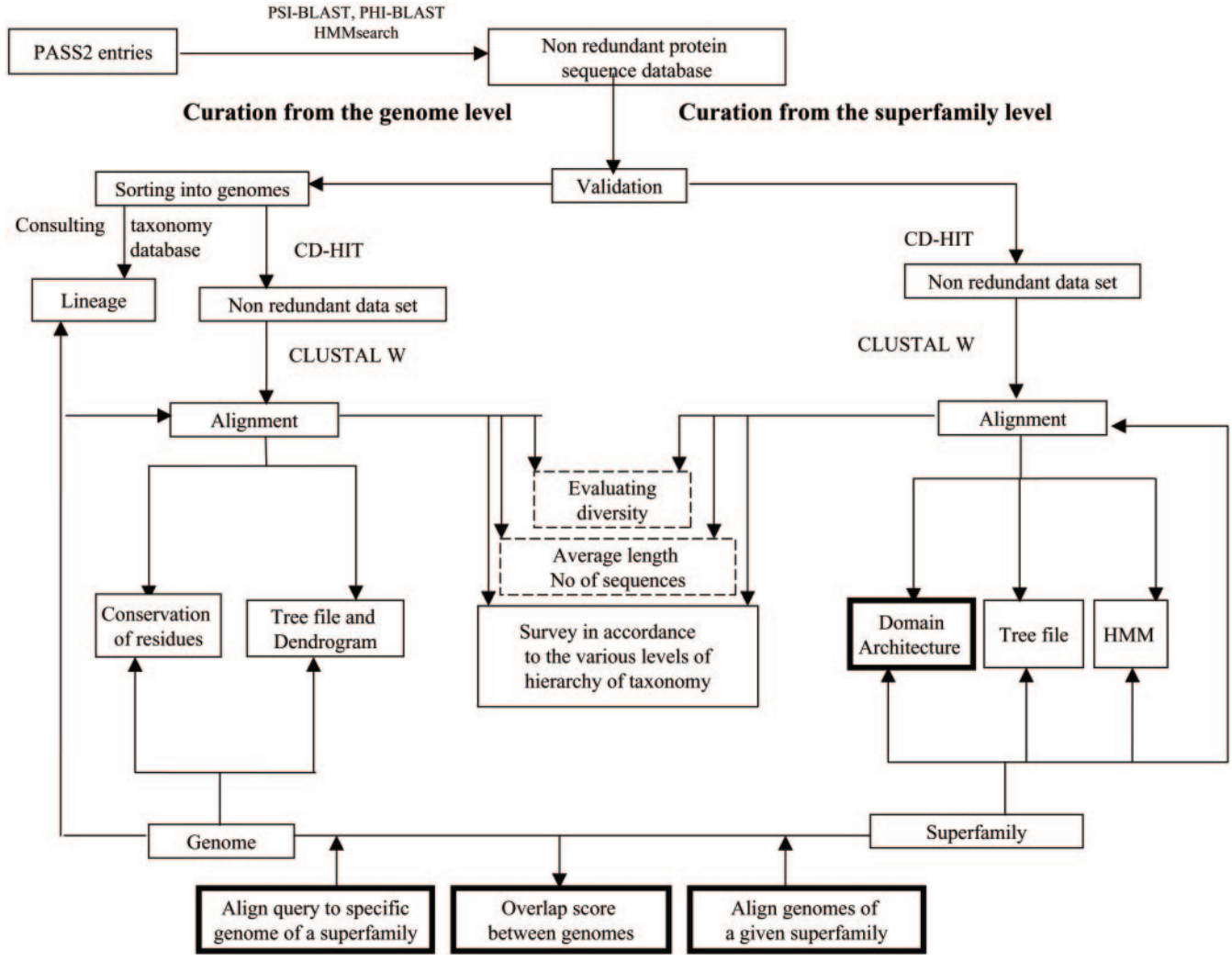

Figure 1. Flowchart representing the steps involved in the curation of the database and various features in GenDiS. Boxes marked in boldface represent the tools provided while the dotted boxes indicate the residual features evaluated for the protein members in GenDiS.

models (HMMs) were derived for domain superfamilies starting from structure-based sequence alignments of PASS2 members (17) with an expectation threshold of 0.1 during the searches. In addition, motif-constrained PHI-BLAST (20) searches were also carried out as reported previously (21,22) for a single iteration and an expectation value of 1.0. A composite set of domain assignments was obtained for individual superfamilies from these three approaches. The alignment lengths were compared with the query to ensure that it corresponds to the full length of PASS2 domains (23) (Figure 1). Redundant proteins were removed employing CD-HIT (24) at a stringent sequence identity cut-off of 100%. Domains assigned to a superfamily belonging to a genome were aligned using CLUSTALW (25). The alignments have been colour-coded by examining the conservation and similarity at the various positions.

Taxonomic annotation of the superfamily members and alignments

Non-redundant sequences, maintained in the NCBI, form a composite resource of several genome databases. GenDiS records the source organism of the assigned proteins and a detailed taxonomic lineage of the species in correspondence with the taxonomy database (13). Taxonomic classifications at the phyla, class, order, family, genus and species levels have been recorded against individual entries. Proteins belonging to similar taxons are clustered together and further sub-grouped at the superfamily level (Figure 1).

TOOLS AND SERVICES AT THE GenDiS SERVER

GenDiS database can be navigated through a user-friendly search engine to obtain relevant information on taxonomic and superfamily distribution. The database has been linked to taxonomy and other protein databases. GenDiS server provides several useful tools for performing genome and cross-genome analysis.

Information about superfamily members

The presence of superfamily members at the different taxonomic levels is summarized. Domains of the various superfamilies before and following the validation (pruned set) are

downloadable. Domain architecture was identified for validated members of GenDiS employing IMPALA (26) against PASS2 profiles of structural domains. Average domain length, sequence diversity within genomes and at the superfamily level are listed. HMMs can be obtained for the various superfamilies.

Genome and taxonomic information

The full list of the diverse superfamilies residing at the various taxonomic hierarchies can be retrieved from the database. Information about the occurrences of the various descending taxons within a particular hierarchy level of taxonomy is provided. Completely sequenced genomes have been separately listed and can be browsed through the complete genome list. The number of superfamilies and homologous sequences present in the various genomes can be obtained. Alignments of the members of particular superfamilies within genomes and conserved regions of the alignment are provided. For multimembered superfamilies, diversity score evaluated by the Makowski and Soares (27) method and the phylogenetic tree obtained on the basis of protein dissimilarity are presented. Domain architectures can also be retrieved at the phyla, class, order and genus levels at the taxonomical hierarchy.

Overlap score within genomes

Distinction among organisms results from the composite proteome encoded by the genome. Comprehensive structural domain assignments at the proteome level provide opportunities to study the distribution of the common and unique superfamilies among the completely sequenced genomes. The overlap score for a pair of completed genomes along with the listing of common and unique superfamilies demonstrates similarity among the organisms at a more holistic level.

Alignments of desired query to superfamilies

Options are provided for aligning query sequences to superfamily members within a genome or by performing genome-wide alignments for specific superfamilies. The alignments are performed employing CLUSTALW (25).

Assigning structural domain architectures

Domain architectural assignments of unannotated sequences elucidate the combination of structural domains embedded within the polypeptide aiding its detailed characterization (28). Structural domains can be assigned to a query sequence by probing against sequence profiles of PASS2 members employing IMPALA (26).

CONCLUSION

GenDiS is a compendium of sequence domains of evolutionarily related proteins grouped at the superfamily level in direct correspondence with SCOP (16) and PASS2 (17) databases. Furthermore, it is possible to obtain links between structural hierarchy and taxonomic levels at GenDis. Availability of alignments for sequence domains in the various genomes over the World Wide Web facilitates the study and design of experiments on specific superfamilies. The database creates a framework for a systematic survey and analysis of various structural superfamilies. The database may be accessed and downloaded across the World Wide Web (http://caps.ncbs.res.in/gendis/download.html).

Associating different proteins with structurally similar and evolutionarily related proteins enhance our functional understanding of a protein superfamily. Complete taxonomic information corresponding to individual sequences in GenDiS database provides a platform for performing cross-genomic or phyletic analysis at various levels of hierarchy in taxonomy. A World Wide Web interface would provide an understanding of the various sequence relatives across the various genomes, their conservation and sequence diversity enhancing our comprehension corresponding to the protein superfamily or an organism.

ACKNOWLEDGEMENTS

R.S. is a Senior Research Fellow of the Wellcome Trust (UK). G.P. is also supported by the Wellcome Trust. We also thank NCBS (TIFR) for infrastructural support.

REFERENCES

1. Chothia,C. (1992) Proteins. One thousand families for the molecular biologist. *Nature*, **357**, 543–544.
2. 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.
3. 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.
4. Sonnhammer,E.L., Eddy,S.R. and Durbin,R. (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, **28**, 405–420.
5. Schultz,J., Milpetz,F., Bork,P. and Ponting,C.P. (1998) SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl Acad. Sci. USA*, **95**, 5857–5864.
6. Buchan,D.W., Shepherd,A.J., Lee,D., Pearl,F.M., Rison,S.C., Thornton,J.M. and Orengo,C.A. (2002) Gene3D: structural assignment for whole genes and genomes using the CATH domain structure database. *Genome Res.*, **12**, 503–514.
7. 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**, D235–D239.
8. Sander,C. and Schneider,R. (1993) The HSSP database of protein structure–sequence alignments. *Nucleic Acids Res.*, **21**, 3105–3109.
9. 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.
10. Apic,G., Gough,J. and Teichmann,S.A. (2001) Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J. Mol. Biol.*, **310**, 311–325.
11. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
12. Todd,A., Orengo,C. and Thornton,J. (2001) Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.*, **307**, 1113–1143.
13. Wheeler,D.L., Church,D.M., Edgar,R., Federhen,S., Helmberg,W., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E., Suzek,T.O., Tatusova,T.A. and Wagner,L. (2004) Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res.*, **32**, D35–D40.

14. Lin,J., Qian,J., Greenbaum,D., Bertone,P., Das,R., Echols,N., Senes,A., Stenger,B. and Gerstein,M. (2002) GeneCensus: genome comparisons in terms of metabolic pathway activity and protein family sharing. *Nucleic Acids Res.*, **30**, 4574–4582.
15. Snel,B., Bork,P. and Huynen,M.A. (1999) Genome phylogeny based on gene content. *Nature Genet.*, **21**, 108–110.
16. Andreeva,A., Howorth,D., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.*, **32**, D226–D229.
17. Bhaduri,A., Pugalenthi,G. and Sowdhamini,R. (2004) PASS2: an automated database of protein alignments organised as structural superfamilies. *BMC Bioinformatics*, **5**, 35.
18. 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.
19. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
20. 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.
21. 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.
22. 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.
23. McLysaght,A., Baldi,P.F. and Gaut,B.S. (2003) Extensive gene gain associated with adaptive evolution of poxviruses. *Proc. Natl Acad. Sci. USA*, **100**, 15655–15660.
24. Li,W., Jaroszewski,L. and Godzik,A. (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, **17**, 282–283.
25. Chenna,R., Sugawara,H., Koike,T., Lopez,R., Gibson,T.J., Higgins,D.G. and Thompson,J.D. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.*, **31**, 3497–3500.
26. Schaffer,A.A., Wolf,Y.I., Ponting,C.P., Koonin,E.V., Aravind,L. and Altschul,S.F. (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics*, **15**, 1000–1011.
27. Makowski,L. and Soares,A. (2003) Estimating the diversity of peptide populations from limited sequence data. *Bioinformatics*, **19**, 483–489.
28. Vogel,C., Berzuini,C., Bashton,M., Gough,J. and Teichmann,S.A. (2004) Supra-domains: evolutionary units larger than single protein domains. *J. Mol. Biol.*, **336**, 809–823.