D188–D191 *Nucleic Acids Research, 2005, Vol. 33, Database issue*
doi:10.1093/nar/gki096

ADDA: a domain database with global coverage of the protein universe

Andreas Heger[1,*], Christopher Andrew Wilton[1], Ashwin Sivakumar[1] and Liisa Holm[1,2]

[1]Institute of Biotechnology and [2]Department of Genetics, University of Helsinki, 00014 Helsinki, Finland

Received August 16, 2004; Revised and Accepted October 13, 2004

ABSTRACT

We used the Automatic Domain Decomposition Algorithm (ADDA) to generate a database of protein domain families with complete coverage of all protein sequences. Sequences are split into domains and domains are grouped into protein domain families in a completely automated process. The current database contains domains for more than 1.5 million sequences in more than 40 000 domain families. In particular, there are 3828 novel domain families that do not overlap with the curated domain databases Pfam, SCOP and InterPro. The data are freely available for downloading and querying via a web interface (http://ekhidna.biocenter.helsinki.fi:9801/sqgraph/pairsdb).

INTRODUCTION

The past and present periods of large-scale genome sequencing have brought an enormous wealth of protein sequences that makes managing, navigating and mining the data an area of research in its own right.

Protein evolution suggests domains as a convenient unit of classification. Mutations, insertions and deletions create sequence diversity inside a domain family. On a higher level, recombination events combine domains in different architectures to give the single- or multi-domain proteins we observe.

Various invaluable tools exist to reduce the diversity of proteins into a reduced set of protein domain families. Semi-automated methods such as Pfam (1), PROSITE (2) or SMART (3) extrapolate the information gained from known members of protein domain families by matching sequences to libraries of hidden Markov models (HMMs), profiles or patterns. Integrative projects such as InterPro (4) combine various primary sources to yield a summary view on protein sequences. Fully automated methods such as ProDom (5) or DOMO (6) apply algorithms to achieve a classification based on first principles.

Previously, we have introduced the Automatic Domain Decomposition Algorithm (ADDA) (7), a method for clustering very large sets of protein sequences. ADDA first splits sequences into domains and then organizes these domains into protein domain families. The classification is constructed in an entirely automated fashion from first principles and thus is not biased by human curation, but only limited by the applied algorithms.

We have applied ADDA to the set of all known protein sequences that are available in the major public databases. Using all sequences for clustering has the advantage of drawing the boundaries between protein domain families in a globally consistent manner. This is in contrast to scanning methods such as Pfam and SMART, where novel or hypothetical sequences are scanned against a library of HMMs or profiles. Matches at the borderline of significance can be due to a newly discovered remote relative or to spurious similarity. In such events, domain families have to be assigned case by case.

Here, we describe a database with a web interface that allows scientists to download and browse the results. The web interface lets a scientist explore the context of a protein sequence in the protein universe: its immediate neighbours as determined by pre-computed sequence similarity searches, and its remote homologues as determined by its domain composition. Alternatively, a scientist can browse the domain families to hunt for domain families of interest.

DATABASE CONTENTS

The database contains more than 1.5 million sequences from UniProt/Swiss-Prot, UniProt/TrEMBL (8), Ensembl (9), NCBI genomes (10) and other sources of protein sequences. The clustering yields 2.7 million domains, which are grouped into 123 000 families. Of these, 40 000 families have more than five members. The database is built in an entirely automated fashion and is updated regularly.

DOMAIN AND DOMAIN FAMILY DEFINITION

The domain and domain family definitions result from an automated clustering procedure applied to the set of all protein

*To whom correspondence should be addressed. Tel: +358 9 191 59115; Fax: +358 9 191 59366; Email: andreas.heger@helsinki.fi
Correspondence may be addressed to Liisa Holm. Email: liisa.holm@helsinki.fi

sequences in the major public databases. The process starts by removing sequences with >40% sequence identity to any other sequence from the input set (11). The remaining representative sequences are then aligned in an all-versus-all manner using the BLAST (12) program.

Representative sequences are split into domains by the ADDA algorithm. Domains are defined so that a minimum number of alignments are intersected by domain boundaries and these alignments cover domains as much as possible.

After splitting protein sequences, the resultant domains are grouped into families of related sequences by a single-linkage clustering algorithm. Domains joined by alignments are grouped into a family if their domain boundaries are consistent. In addition, domains are compared using sensitive profile–profile comparison. The merging process stops when the sequence similarity between domains drops below a threshold.

Finally, domain boundaries are mapped from the representative sequences onto all sequences in the database.

QUALITY CONTROL

Quality control monitors two aspects of the clustering by comparing ADDA domains to curated databases of domain families like Pfam and SCOP (13). The correspondence of domain boundaries is checked by computing the relative overlap between ADDA and reference domains. ADDA tends to split conservatively; ADDA domains are, on average, larger than reference domains.

The correspondence of domain families is measured by matching each ADDA family to the best matching reference family and counting the relative frequency of other reference families in the ADDA family (selectivity) and the relative frequency of reference domains assigned to different ADDA families (sensitivity). On average, an ADDA domain family unifies 93% of the members of a Pfam family while containing only 5% contamination.

Charts and summary statistics are available on the web server.

ANNOTATING DOMAIN FAMILIES

In multi-domain proteins, domains of different protein families co-occur. Based on the observed architecture of protein sequences, domain families can be divided into two groups: mobile modules and associated families. Mobile modules are promiscuous and recombine with several other domain families. Associated families either always occur in single-domain proteins or are always associated with the same domain family. In the present release, there are 9252 mobile modules and 49 455 associated families (Figure 1A). While the latter tend to be specific to a single kingdom (Archaea, Bacteria and Eukaryota), mobile modules have a larger taxonomic range (Figure 1B).

A domain family is declared to be structurally covered if one of its domains can be mapped onto a structure in the PDB database of protein structures (14). For each ADDA domain, we register the sequence overlap with domains from curated domain databases (Pfam, SCOP and InterPro). An ADDA domain family is classified as novel if <5% of its domains overlap with domains from the curated domain databases.

ADDA contains 3828 novel mobile modules that are not known to curated domain databases and for which there is no structural information available (Figure 1C). Novel domain families tend to have fewer members (<200) than well-known domain families. The number of novel domains in associated families is even larger comprising 40 505 domain families.

DATABASE ACCESS AND INTERFACE

The complete protein domain classification can be downloaded from http://ekhidna.biocenter.helsinki.fi:8080/downloads/adda. A web interface is available for browsing and querying at http://ekhidna.biocenter.helsinki.fi:9801/sqgraph/pairsdb.

The interface allows the user to query for protein sequences by identifier, accession number or sequence similarity. The sequence view shows the decomposition of a protein sequence into domains. Links allow the user to browse similar sequences in the direct neighbourhood of a query sequence (multiple alignments pre-computed using BLAST and PSI-BLAST) and to switch to the domain families to get all related sequences beyond the immediate neighbourhood.

Domain families of interest can be retrieved by custom queries. Attributes available for querying are the size of the family, its taxonomic spread, its structural coverage, the number of associated domains (querying for mobile modules), the overlap with other domain classifications (querying for novel domain families) and others. The domain family view includes a summary overview over the protein family and links to other domain classifications. The sequences of all domains in a domain family can be downloaded for local analysis. If structures are known in the family, the domain boundaries are mapped onto the structures for visualization with RASMOL (15).

In the browsing section, the user finds links to precompiled domain sets of interest, e.g. all exclusively eukaryotic mobile domains or a list of domain families without known structures (structural genomics targets). In addition, a genome browser allows access to all or a selection of domain families occurring in a genome.

The web server contains outgoing links to external databases for sequences and domains. For example, ADDA is linked to by the Dali domain database (16) and vice versa.

EXAMPLE

Links and attributes can be queried in numerous ways. One application of the database interface is to hunt for novel domain families. For example, typing 'sapiens' into the genome browser lists 28 791 domain families in human protein sequences. The result set can be restricted to exclude domain families that have domains from Archaea or Bacteria and/or are not novel and not a mobile module giving 7933 domain families. Modifying the query to include only domains with more than 20 members produces 108 novel domain families that occur in human protein sequences, are specific to eukaryotes and have at least 20 members.

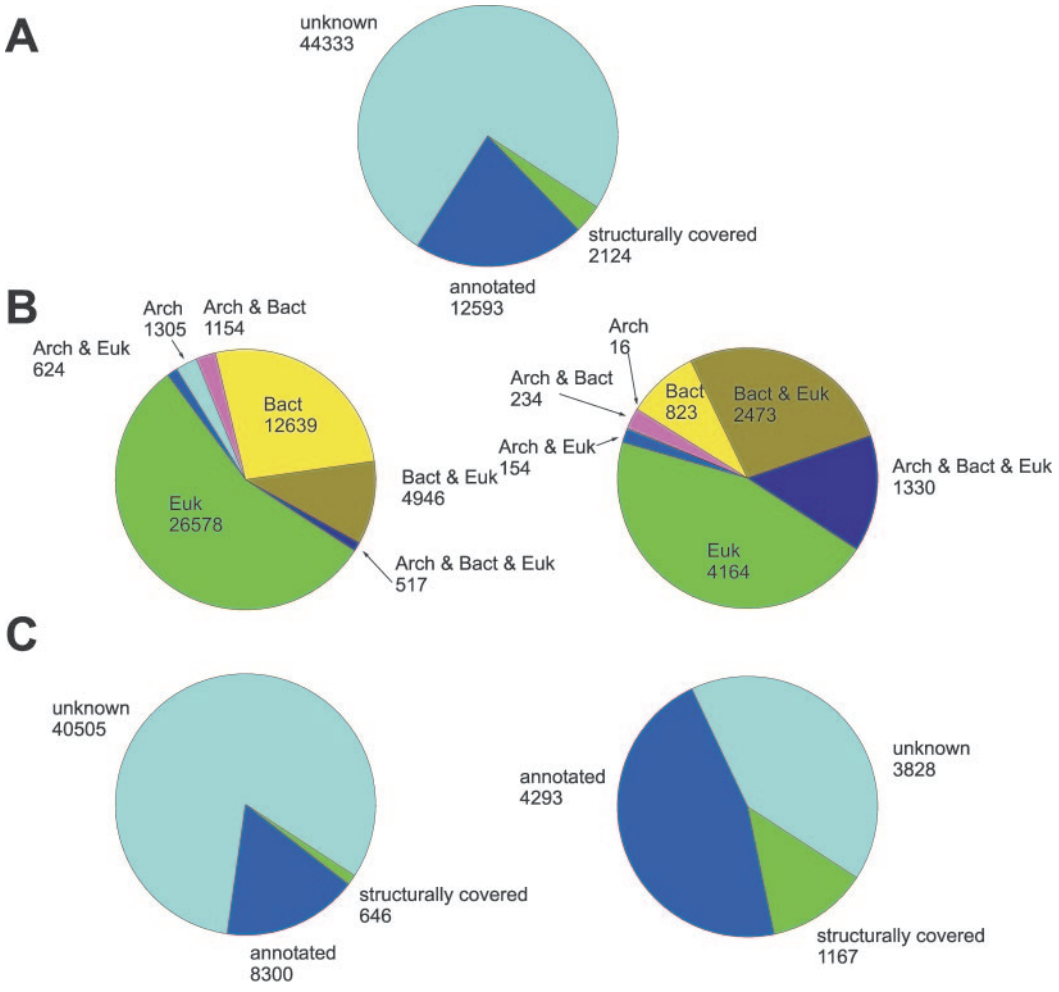

Figure 1. Overview of domain families in ADDA. The number of families is given in the last row of each category label. (**A**) Mobile modules, domain families that co-occur with a variety of different domain families, constitute only a fraction of all domain families. Many domains only occur in single-domain proteins or are always associated with the same domain family (associated families). The majority of domain families contain only a single representative sequence on the 40% similarity level (singletons). (**B**) Taxonomic distribution of domain families over the three superkingdoms (Archaea, Bacteria and Eukaryota). Left: only associated domain families excluding singletons. Right: only mobile modules. Mobile modules tend to be more widely distributed than associated domains. (**C**) Annotation of domain families. Left: only associated domain families excluding singletons. Right: only mobile modules. Novel domain families do not overlap with domain families from Pfam, SCOP and InterPro. Mobile modules are well known to curated domain databases, but there are many novel domain families left to be explored.

WORK IN PROGRESS

Our aim is to push the functional annotation of proteins as far as possible using only automated methods. Defining domain families is the first step. We are currently testing methods to split domain families into groups of orthologous proteins and automated methods to define functionally important residues in a family.

REFERENCES

1. 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.
2. Hulo,N., Sigrist,C.J.A., Le Saux,V., Langendijk-Genevaux,P.S., Bordoli,L., Gattiker,A., De Castro,E., Bucher,P. and Bairoch,A. (2004) Recent improvements to the PROSITE database. *Nucleic Acids Res.*, **32**, D134–D137.
3. 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**, D142–D144.
4. 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.
5. Servant,F., Bru,C., Carrere,S., Courcelle,E., Gouzy,J., Peyruc,D. and Kahn,D. (2002) ProDom: automated clustering of homologous domains. *Brief Bioinformatics*, **3**, 246–251.
6. Gracy,J. and Argos,P. (1998) DOMO: a new database of aligned protein domains. *Trends Biochem. Sci.*, **23**, 495–497.
7. Heger,A. and Holm,L. (2003) Exhaustive enumeration of protein domain families. *J. Mol. Biol.*, **328**, 749–767.
8. Apweiler,R., Bairoch,A., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.*, **32**, 115–119.
9. Birney,E., Andrews,T.D., Bevan,P., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cuff,J., Curwen,V., Cutts,T. *et al.* (2004) An overview of Ensembl. *Genome Res.*, **5**, 925–928.
10. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2004) GenBank: update. *Nucleic Acids Res.*, **32**, D23–D26.
11. Park,J., Holm,L., Heger,A. and Chothia,C. (2000) RSDB: representative protein sequence databases have high information content. *Bioinformatics*, **16**, 458–464.

12. 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.
13. Andreeva,A., Howorth,D., Brenner,S.E., Hubbard,T.J.P., Chothia,C. and Murzin,A.G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res*., **32**, D226–D229.
14. 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.
15. Sayle,R.S. and Milner-White,E.J. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem. Sci*., **20**, 374l.
16. Dietmann,S., Park,J., Notredame,C., Heger,A., Lappe,M. and Holm,L. (2001) A fully automatic evolutionary classification of protein folds: dali domain dictionary version 3. *Nucleic Acids Res*., **29**, D55–D57.