*D222–D226* *Nucleic Acids Research, 2015, Vol. 43, Database issue* *Published online 20 November 2014*
*doi: 10.1093/nar/gku1221*

# CDD: NCBI's conserved domain database

**Aron Marchler-Bauer*, Myra K. Derbyshire, Noreen R. Gonzales, Shennan Lu, Farideh Chitsaz, Lewis Y. Geer, Renata C. Geer, Jane He, Marc Gwadz, David I. Hurwitz, Christopher J. Lanczycki, Fu Lu, Gabriele H. Marchler, James S. Song, Narmada Thanki, Zhouxi Wang, Roxanne A. Yamashita, Dachuan Zhang, Chanjuan Zheng and Stephen H. Bryant**

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bldg. 38 A, Room 8N805, 8600 Rockville Pike, Bethesda, MD 20894, USA

Received November 06, 2014; Accepted November 08, 2014

## ABSTRACT

**NCBI's CDD, the Conserved Domain Database, enters its 15th year as a public resource for the annotation of proteins with the location of conserved domain footprints. Going forward, we strive to improve the coverage and consistency of domain annotation provided by CDD. We maintain a live search system as well as an archive of pre-computed domain annotation for sequences tracked in NCBI's Entrez protein database, which can be retrieved for single sequences or in bulk. We also maintain import procedures so that CDD contains domain models and domain definitions provided by several collections available in the public domain, as well as those produced by an in-house curation effort. The curation effort aims at increasing coverage and providing finer-grained classifications of common protein domains, for which a wealth of functional and structural data has become available. CDD curation generates alignment models of representative sequence fragments, which are in agreement with domain boundaries as observed in protein 3D structure, and which model the structurally conserved cores of domain families as well as annotate conserved features. CDD can be accessed at http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml.**

## INTRODUCTION AND STATISTICS ON CONSERVED DOMAIN DATABASE (CDD) COVERAGE

The amount of sequence data deposited into public repositories has made it impracticable to routinely run sequence similarity searches against single, generic and comprehensive sequence collections. A search set that grows exponentially in size does not fit well with the need for executing an ever increasing number of such searches—and quicker searches too, if possible. Curated collections of representative sequences offer a viable alternative, and so do profile database searches, as collections of profile models that represent evolutionarily conserved sequence fragments (or domains) are not expected to grow at an exorbitant pace. They also provide functional characterization that is based on the presence of signature sequence patterns and may serve as a starting point for functional annotation and classification.

Here, we briefly summarize recent updates to the CDD resource, with respect to its content and the functionality of search interfaces. The current live CDD version, v3.12, contains 46 675 protein- and protein domain-models, with content obtained from Pfam (1), SMART (2), the COGs collection (3), TIGRFAMs (4), the NCBI Protein Clusters collection (5) and NCBI's in-house data curation effort (6). CDD version v3.13 is being processed and likely to still be released in 2014, it will include the most recent release of TIGRFAMS, version 15.

Several large classifications for common and functionally diverse domain families have recently been updated or added to CDD, such as comprehensive hierarchies of models representing the catalytic domains of protein kinases (cd13968), type 2 periplasmic binding proteins (PBP2, cd00648), globins and globin-like domains (cd01067), the pleckstrin-homology domains (PH, cd00900), RNA recognition motif (RRM, cd00590), SH3 domains (cd00174), immunoglobulin domains (cd00096), LIM domains (cd08368), UBA_like domains (cd00194) or the thioredoxin superfamily (TRX, cd01659).

Table 1 below lists URLs for various entry points into CDD. CDD is part of NCBI's Entrez search and retrieval system and is cross-linked with other databases, such as Entrez/protein, Entrez/Gene, 3D-structure (MMDB), NCBI BioSystems, PubMed and PubChem. Domain and site annotation generated by CDD is visible in graphical views of protein sequences in Entrez. Currently, CDD annotates more than 120 million sequences in Entrez/protein, about 83% of the proteins excluding sequences from en-

*To whom correspondence should be addressed. Tel: +1 301 435 4919; Fax: +1 301 435 7793; Email: bauer@ncbi.nlm.nih.gov

Published by Oxford University Press on behalf of Nucleic Acids Research 2014.
This work is written by US Government employees and is in the public domain in the US.

Downloaded from https://academic.oup.com/nar/article/43/D1/D222/2439461 by guest on 21 August 2022

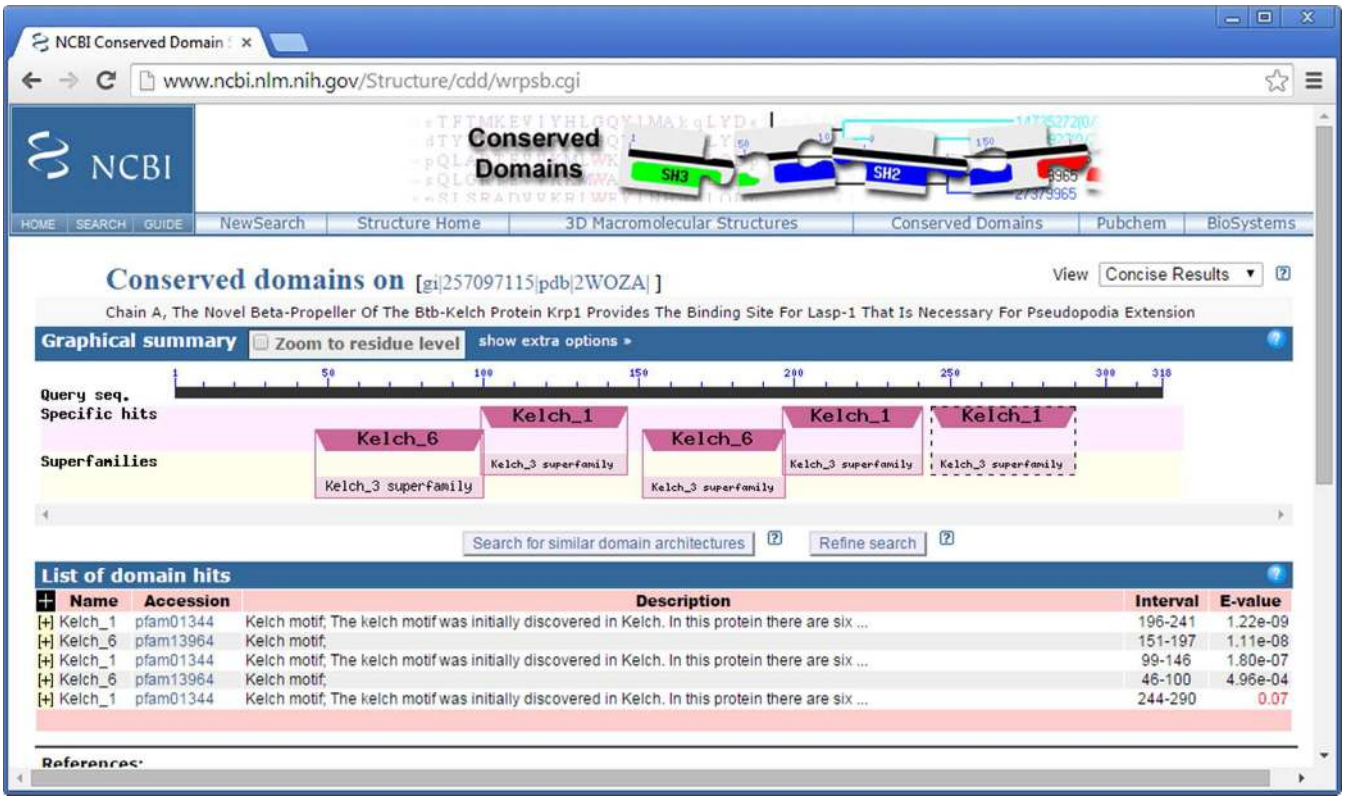


**Figure 1.** CD-Search reporting a 'rescued' domain annotation, which scores an *E*-value above the default reporting threshold of 0.01. The live search for the query sequence, derived from the PDB structure 2WOZ.

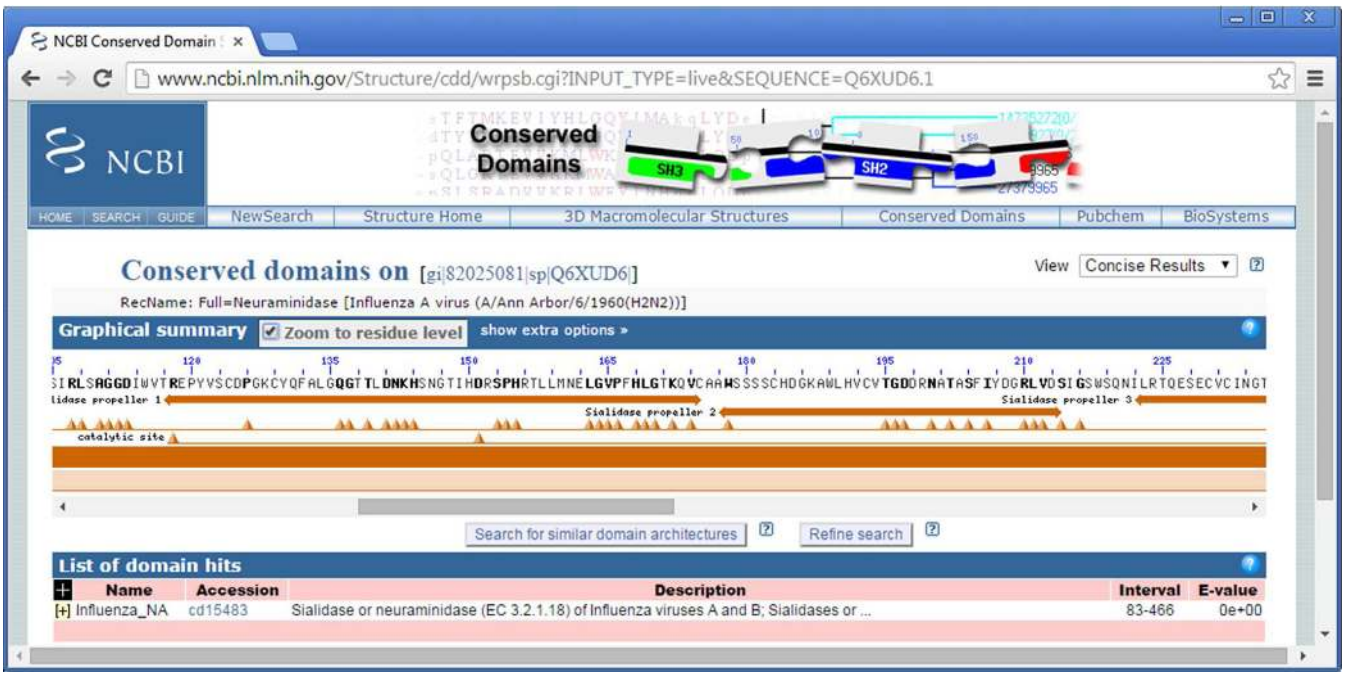


**Figure 2.** CD-Search results for SwissProt Q6XUD6, zoomed in to 'residue level' display so that the precise locations of domain boundaries and functional sites become apparent. Query sequence residues highlighted in bold print have been identified as part of a functional site (such as the 'catalytic site' mapping to R118 and D151, plus other residues not shown in this example). Structural motifs are shown as double-headed arrows.

Downloaded from https://academic.oup.com/nar/article/43/D1/D222/2439461 by guest on 21 August 2022

**Table 1.** URLs and other resources associated with the CDD project

| Task | URL | Description |
|---|---|---|
| ..find conserved domains in a protein query sequence | http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi | CD-Search interface to the RPS-BLAST algorithm and CDD search databases |
| ..look up conserved domains for a protein with known sequence identifier | http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi | CD-Search interface to the CDART database, or use 'Identify Conserved Domains' link on Entrez/protein pages |
| ..find or look up conserved domains for many protein queries | http://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi | BATCH CD-Search interface to the RPS-BLAST algorithm and the CDART database |
| ..find a conserved domain model by name or arbitrary text term | http://www.ncbi.nlm.nih.gov/cdd | Entrez interface to CDD |
| ..learn more about CDD | http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml | CDD project home page |
| ..retrieve and analyze domain architectures | http://www.ncbi.nlm.nih.gov/Structure/lexington/lexington.cgi | CDART domain architecture viewer |
| ..download CDD data | ftp://ftp.ncbi.nih.gov/pub/mmdb/cdd | CDD FTP site, see README file for content |
| ..download and install data viewers CDTree and Cn3D | http://www.ncbi.nlm.nih.gov/Structure/cdtree/cdtree.shtml | Domain hierarchy editor/viewer and protein structure/alignment viewer |
| ..download and install capabilities for local sequence searches | ftp://ftp.ncbi.nlm.nih.gov/toolbox executables can be obtained from: http://www.ncbi.nlm.nih.gov/BLAST/download.shtml | RPS-BLAST stand-alone tool for searching databases of profile models, part of the NCBI toolkit distribution |

vironmental sampling. CDD annotates 98% of structure-derived protein sequences in Entrez that are over 30 residues long. An active curation program is examining the remaining and newly deposited 3D structures and is adding models for novel domains to the collection. Currently, CDD tracks 123 domain superfamilies that are solely represented by models curated as part of the NCBI effort.

CDD curators annotate functional sites on NCBI-curated models, such as active sites and binding sites, which are mapped onto protein (query) sequences. Currently, a total of 21 202 site annotations have been created, on 8642 out of 10 789 domain models. These facilitate the identification of specific amino acids involved in a protein's function and therefore specific points at which allelic variation might directly affect that function.

## REVIEW OF SUPERFAMILY CLUSTERING

RPS-BLAST, as used for CDD data production and live RPS-BLAST searches, has recently been updated with the option to use composition-corrected scoring (7). Masking of compositionally biased regions on query sequences is no longer necessary in order to avoid an abundance of false-positive hits. In turn, conserved domain models representing compositionally biased domain or protein families can now be utilized more productively in protein annotation. The default reporting *E*-value threshold (0.01) for CD-Search (8) and CDD data production was not changed, as the overall coverage obtained with the CDD annotation resource was not affected by the change in scoring. However, details of the annotation for numerous individual protein sequences have changed.

These changes in the pre-computed domain annotation that is stored in the CDART (9) database have affected the results of computational clustering of protein domain models (with overlapping coverage) into protein domain 'superfamilies'. Domain superfamily clusters are created so that domain architectures can be evaluated more efficiently and consistently, as CDD is a redundant resource, which collects domain models from various sources, and homologous sequence fragments from two different proteins that share one and the same architecture might get annotated by different models. If these different models belong to the same CDART superfamilies, the proteins will be sorted into the same domain architecture.

To avoid false clustering of unrelated domain family models, formation of novel larger clusters was disabled during the production of CDD version v3.11, released in early 2014. For the current version v3.12, CDD curators manually reviewed 649 newly formed clusters, those that merged previous smaller clusters or domain models that had not been included in clusters at all. Clusters that bridged domain models not related by common descent were broken up into their respective smaller units, if such a determination could be made based on available 3D structure, functional annotation, and comparison with other public resources of similar scope and coverage (1). CDD curators used the Cytoscape (10) program to visualize clusters of domains with overlapping sequence, as network displays indicating the presence and amount of overlapping sequence annotation between dozens or hundreds of domain models in a cluster were invaluable in spotting and investigating questionable relationships. More than 700 individual domain model pairs were blocked from forming clusters as a result of this review, largely avoiding false clusters that would confound domain architecture analysis available via the CDART resource (9), as the latter depends on the assignment of superfamily-level domain footprints. Manual review of newly formed family clusters will be continued as long as necessary and while an updated clustering procedure is being developed.

The majority of protein domain models in CDD are "singletons" that do not form clusters with other domain models. As a result, they are the sole member of their superfamily. Superfamilies that contain more than one model are indexed in the Entrez/CDD database, and their content can be visualized as conserved domain summary pages or via link data in Entrez that associate each superfamily cluster with its constituent models. Superfamily cluster models are given the accession prefix 'cl'.

Downloaded from https://academic.oup.com/nar/article/43/D1/D222/2439461 by guest on 21 August 2022

## IMPROVING DOMAIN ANNOTATION VIA COMMON DOMAIN ARCHITECTURES

Conserved domain models utilized in CDD's annotation procedures may not yield significant scores with all sequence fragments that are true members of the respective family. One reason is that domains may be relatively short and sequence-diverse, giving rise to position-specific scoring matrices that fail to separate all true positive family members from false positives.

Recently, we have added an option to the live CD-Search interface that lets users run searches at an implicitly higher level of sensitivity and post-processes the results, so that domain hits falling just below the default borderline of significance may still be reported if they match a domain architecture context provided by other domain hits in the same query. When a user selects the corresponding search option, the live search is run at a higher *E*-value threshold (1.0 for the default reporting *E*-value threshold of 0.01), and additional domain hits are collected. CD-Search will then evaluate alternative domain architectures that can be formed including all the original hits and one or more of the additional domain hits. Alternative domain architectures are ranked by frequency. If the most frequent alternative architecture is found in NCBI's non-redundant protein database (NR) at least 20 times (this current threshold is subject to change as the size of the NR database increases), the additional domain hits contributing to that architecture are reported, using a novel display style that highlights their tentative status.

If additional domain hits that score above the default reporting *E*-value threshold are tandem repeats of a neighboring domain that was detected at the default *E*-value threshold, they are reported irrespective of the domain architecture's frequency in the NR database. Figure 1 shows the result of such a live search, for a query sequence derived from the PDB structure 2WOZ (11), as tracked by MMDB (12). In this example, an additional hit to a domain model is reported which is already found several times in a tandem repeat arrangement.

We are evaluating an additional option to also suppress borderline hits that score *E*-values slightly below the default reporting threshold but give rise to unique or very unusual domain architectures, as they may be false positives and should not be reported. Domain architectures based on additional 'rescued' and relatively fewer 'suppressed' domain hits will be recorded as the pre-computed domain architecture in the near future and are intended to improve the overall consistency of domain footprint annotation. Currently, the detection and display of 'rescued' domain annotation is limited to live CD-Searches.

## ANNOTATION OF STRUCTURAL MOTIFS

CDD curators record the location of functional motifs on protein domain models, so that these motifs can be mapped onto protein sequences and facilitate the interpretation of sequence conservation and variation, for example. Site annotations provided by CDD include a large number of active sites, chemical binding and protein-protein interaction sites, and complement, to some extent, experimentally derived or computationally generated site annotations tied to individual protein records, such as found in the SwissProt data set (13), for example. We have now added 'structural motifs' to the list of motifs or sites that may be recorded and mapped. Structural motifs are not necessary functional, but provide more detailed annotation on query sequences. They will include short structural repeats, such as beta-propellers, coiled coils and transmembrane segments, as well as short functional motifs, such as DNA-binding zinc fingers, for example. Upcoming versions of CDD will contain a novel type of model, called a 'structural domain', with the accession prefix 'sd'. Structural domain models are being assembled solely for the purpose of providing structural motif annotation, but structural motif annotation can also be found on regular conserved domain models with the accession prefix 'cd'. Figure 2 gives an example of how such structural motif annotation delineates the extent of beta-propellers on a query sequence from an Influenza virus. Figure 2 also displays a novel feature of the CD-Search interface, the ability to zoom the graphical displays so that individual query sequence residues become visible and let the user map domain extents and the location of conserved sites more precisely. Individual residues that are parts of functional sites (but not structural motifs) are highlighted in bold font.

## ACKNOWLEDGEMENTS

We thank Paul Thiessen, Lianyi Han and the NCBI Information Engineering Branch for assistance with software development. We are grateful to the authors of Pfam, SMART, COGs, TIGRFAMs and NCBI's Protein Clusters database for providing access to their resources and data, and the users of CDD for their acknowledgements and invaluable feedback.

## FUNDING

Intramural Research Program of the National Institutes of Health, National Library of Medicine. Funding open access charge: Intramural Research Program of the National Institutes of Health, National Library of Medicine.
*Conflict of interest statement.* None declared.

## REFERENCES

1. Finn,R.D., Bateman,A., Clements,J., Coggill,P.C., Eberhardt,R.Y., Eddy,S.R., Heger,A., Hetherington,K., Holm,L., Mistry,J. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.*, **42**, D222–D230.
2. Letunic,I., Doerks,T. and Bork,P. (2014) SMART: recent updates, new developments, and status in 2015. *Nucleic Acids Res.*, doi:10.1093/nar/gku949.
3. Tatusov,R.L., Natale,D.A., Garkavtsev,I.V., Tatusova,T.A., Shankavaram,U.T., Rao,B.S., Kiryutin,B., Galperin,M.Y., Fedorova,N.D. and Koonin,E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**, 22–28.
4. Haft,D.H., Selengut,J.D., Richter,A.R., Harkins,D., Basu,M.K. and Beck,E. (2013) TIGRFAMs and Genome Properties in 2013. *Nucleic Acids Res.*, **41**, D387–D395.
5. Klimke,W., Agarwala,R., Badretdin,A., Chetvernin,S., Ciufo,S., Fedorov,B., Kiryutin,B., O'Neill,K., Resch,W., Resenchuk,S. *et al.* (2009) The National Center for Biotechnology Information's Protein Clusters Database. *Nucleic Acids Res.*, **37**, D216–D223.

Downloaded from https://academic.oup.com/nar/article/43/D1/D222/2439461 by guest on 21 August 2022

6. Marchler-Bauer,A., Anderson,J.B., DeWeese-Scott,C., Fedorova,N.D., Geer,L.Y., He,S., Hurwitz,D.I., Jackson,J.D., Jacobs,A.R., Lanczycki,C.J. *et al.* (2003) CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Res.*, **31**, D383–D387.
7. Schäffer,A.A., Aravind,L., Madden,T.L., Shavirin,S., Spouge,J.L., Wolf,Y.I., Koonin,E.V. and Altschul,S.F. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.*, **29**, 2994–3005.
8. Marchler-Bauer,A. and Bryant,S.H. (2005) CD-Search: protein domain annotations on the fly. *Nucleic Acids Res.*, **32**, W327–W331.
9. Geer,L.Y., Domrachev,M., Lipman,D.J. and Bryant,S.H. (2002) CDART: protein homology by domain architecture. *Genome Res.*, **12**, 1619–1623.
10. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
11. Gray,C.H., McGarry,L.C., Spence,H.J., Riboldi-Tunnicliffe,A. and Ozanne,B.W. (2009) Novel beta-propeller of the BTB-Kelch protein Krp1 provides a binding site for Lasp-1 that is necessary for pseudopodial extension. *J. Biol. Chem.*, **284**, 30489–30507.
12. Madej,T., Lanczycki,C.J., Zhang,D., Thiessen,P.A., Geer,R.C., Marchler-Bauer,A. and Bryant,S.H. (2014) MMDB and VAST+: tracking structural similarities between macromolecular complexes. *Nucleic Acids Res.*, **42**, D297–D303.
13. Derbyshire,M.K., Lanczycki,C.J., Bryant,S.H. and Marchler-Bauer,A. (2012) Annotation of functional sites with the Conserved Domain Database. *Database*, doi:10.1093/database/bar058.

Downloaded from https://academic.oup.com/nar/article/43/D1/D222/2439461 by guest on 21 August 2022