CDD: a database of conserved domain alignments with links to domain three-dimensional structure

Aron Marchler-Bauer*, Anna R. Panchenko, Benjamin A. Shoemaker, Paul A. Thiessen, Lewis Y. Geer and Stephen H. Bryant

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

Received August 15, 2001; Accepted October 1, 2001

ABSTRACT

The Conserved Domain Database (CDD) is a compilation of multiple sequence alignments representing protein domains conserved in molecular evolution. It has been populated with alignment data from the public collections Pfam and SMART, as well as with contributions from colleagues at NCBI. The current version of CDD (v.1.54) contains 3693 such models. CDD alignments are linked to protein sequence and structure data in Entrez. The molecular structure viewer Cn3D serves as a tool to interactively visualize alignments and three-dimensional structure, and to link three-dimensional residue coordinates to descriptions of evolutionary conservation. CDD can be accessed on the World Wide Web at http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml. Protein query sequences may be compared against databases of position-specific score matrices derived from alignments in CDD, using a service named CD-Search, which can be found at http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi. CD-Search runs reverse-position-specific BLAST (RPS-BLAST), a variant of the widely used PSI-BLAST algorithm. CD-Search is run by default for protein–protein queries submitted to NCBI's BLAST service at http://www.ncbi.nlm.nih.gov/BLAST.

INTRODUCTION

Protein domains may be thought of as proteins' structural and functional building blocks, dividing the primary and tertiary structure of a chain into distinct units. Domains are also mobile genetic units, rearranging in various combinations throughout the molecular evolution of proteins. To understand such processes, and the effect they have had on the present protein repertoire, proteins need to be analyzed not as full-length sequences but rather as collections of individual domains, each of which is important as a unit of molecular evolution.

Protein sequence comparison, as a tool to investigate patterns of conservation and divergence in molecular evolution, is more powerful when sequences are compared to models of protein families instead of other single sequences (1). This may be of particular importance if one wants to use sequence comparison for domain identification and annotation of new sequence data. Examples of such family models are protein profiles (2), hidden Markov models (3,4) and position-specific score matrices (5). The latter, for example, are constructed automatically from sets of pairwise alignments in an iterative database search procedure in PSI-BLAST, a popular addition to the BLAST family of programs (6). PSI-BLAST generates position-specific score matrices, anchored on query sequences, to serve as models of protein families. Explicit alignment models, though, are important for further analysis of divergent domain families and for the transfer of annotation.

Collections of domain alignment models thus are invaluable resources for the study of protein evolution and for large-scale annotation of genomic sequences. Examples of carefully compiled collections are Pfam (7) and SMART (8), which also come bundled with powerful, hidden Markov model-based search engines. Alignments from Pfam and SMART have been imported into the Conserved Domain Database (CDD), to serve a variety of purposes. CDD lets us tie explicit alignment models to a fast search system using BLAST heuristics, via position-specific score matrices computed from the alignments. A corresponding search service has been made available at http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi. Also, CDD makes possible the interactive visualization of both multiple alignment data and protein three-dimensional structure, using NCBI's Cn3D viewer (9). Such combined alignment/structure displays are available for both the browsing of CDD content, and for the display of database search results.

THE CONSERVED DOMAIN DATABASE

We have set up a protocol to import and present alignment data from external sources as well as from in-house collaborators. We attempt to identify the sequence fragments used by the alignments' authors, so that we can link to full-length sequences in Entrez (10). If accession codes supplied by the source databases cannot be identified, BLAST searches are run for the fragments in order to find identical or very similar sequences in NCBI's databases, requiring at least 90% sequence identity across the aligned fragment. Particular attention is paid to close matches with structure-linked sequences, and we substitute alignment rows with such sequences when possible. For substitution, we require a similarity threshold of at least 75% sequence identity in the aligned region and no more

*To whom correspondence should be addressed. Tel: +1 301 435 4919; Fax: +1 301 480 9241; Email: bauer@ncbi.nlm.nih.gov

than 5% of that region is allowed to be lost due to insertions and deletions.

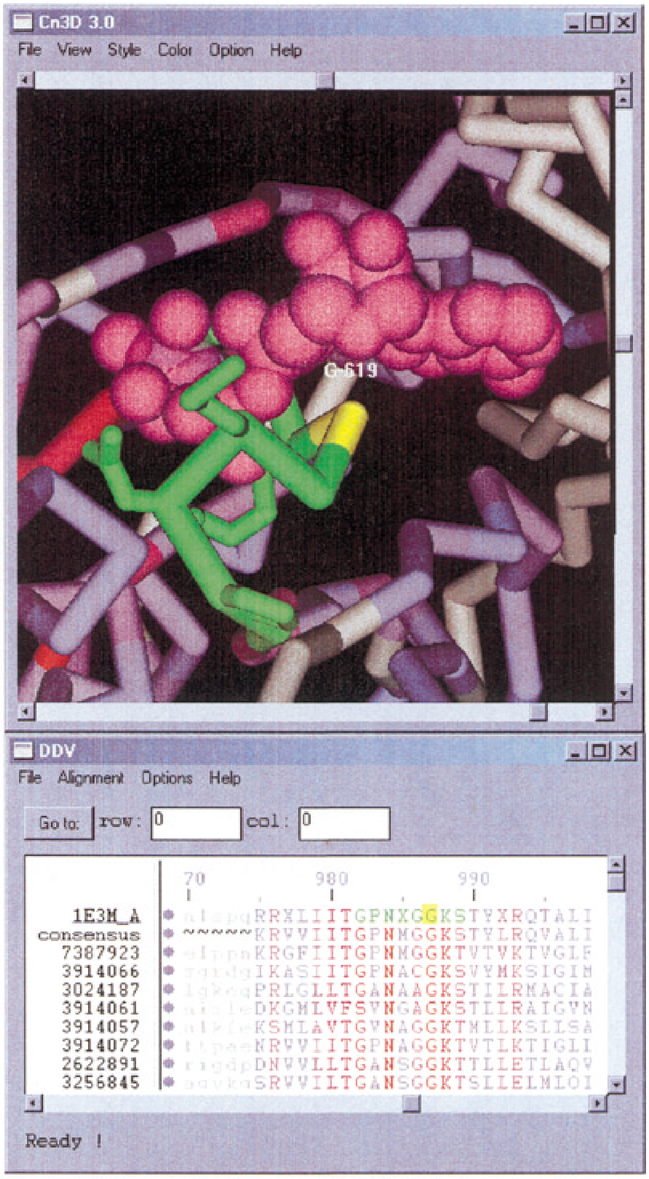
Upon import, multiple alignments are deconstructed into sets of pairwise alignments. A representative common to all the pairwise alignments is chosen as the sequence with the fewest deletions relative to other sequences, so that the loss of alignment information is minimal. In fact most of the alignments imported from Pfam or SMART have a very pronounced block structure, reducing the risk of losing information in this step. Structure-linked sequences are picked as representatives whenever available.

Imported domain alignments can be retrieved by accession or searched by keyword at http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml. The server generates summary pages from which several alignment visualization styles are available. If three-dimensional structure information is available, Cn3D 3.0, a molecular structure viewer distributed by NCBI, can be used to display integrated views of the domain's multiple alignment and its conservation patterns, as well as the three-dimensional structure of a representative member. This display allows interactive highlighting and feature annotation. Figure 1 shows an example of how this capability can be used to illustrate how genotypes may be linked to disease.

Domain alignments in CDD are used to calculate position-specific score matrices for database searching. For the representation of position-specific score matrix (PSSM) models, a consensus sequence is calculated for each conserved domain, reporting the most frequent residues in aligned columns. Although visible in alignment displays, the consensus sequence is not used directly in PSSM calculations. However, it determines the length of the PSSMs, as only columns with >50% aligned states are include in the consensus and PSSM calculation.

The search engine making use of CDD's collection of PSSM models is reverse-position-specific BLAST (RPS-BLAST), a variant of PSI-BLAST. It inverts the role of query and subject, comparing a single sequence against a database of PSSM models instead of searching a database of sequences with a single PSSM model. A web-based interface to RPS-BLAST, CD-Search, is available at http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi.

The ASN.1 data specification for domain multiple alignment data in CDD is available through the NCBI toolbox distribution at ftp://ncbi.nlm.nih.gov/toolbox, together with C program code that can be used to read, write and compute with CDD data in the context of the NCBI toolkit. The content of the CDD can be downloaded from NCBI's FTP site in machine-readable ASN.1 format, by following instructions on the CDD home page.


Cn3D 3.0
File View Style Color Option Help
G619
DDV
File Alignment Options Help
Go to:
row: 0
col: 0
Ready !

```
             70        980       990
             |         |         |
  1E3M_A  ● nlspqRRXLIITGPNXGGKSTYXRQTALI
consensus ● ~~~~~~KRVVIITGPNMGGKSTYLRQVALI
 7387923  ● mlppnKRGFIITGPNMGGKTVTVKTVGLF
 3914066  ● rqrdgIKASIITGPNACGKSVYMKSIGIM
 3024187  ● lgkeqPRLGLLTGANAAGKSTIIRMACIA
 3914061  ● nisleDKGMLVFSVNGAGKSTILRAIGVN
 3914057  ● nlkleKSMLAVTGVNAGGKTMLLKSLLSA
 3914072  ● ttpeeNRVVIITGPNAGGKTVTLKTIGLI
 2622891  ● rigdpDNVVLLTGANSGGKTTLLETLAQV
 3256845  ● sqvkqSRVVIITGANSGGKTSLLELMIQI
```

Figure 1. Cn3D 3.0 view showing a subset of aligned sequences from the CD sm (ATPase domain of DNA mismatch repair MUTS family). Residues corresponding to the P-loop motif around the ADP/Mg^{2+} binding site have been annotated in Cn3D and are highlighted in green: GLY614-x(4)-GLY619-LYS620-SER621 in 1E3M chain A (12). The ADP/Mg^{2+} complex is colored magenta. In the related human MSH2 protein, a somatic GLY→SER mutation in the position corresponding to GLY619 (highlighted in yellow) has been associated with type 1 familial non-polyposis colon cancer. From analyzing the image one may understand how a substitution of the side chain at this position interferes with ADP/Mg^{2+} binding and may therefore impede the DNA mismatch repair system.

FUTURE DEVELOPMENTS

DART, the domain architecture retrieval tool, is an application making use of CDD data. It runs a variant of CD-search for protein query sequences and compares the inferred domain architecture of the query with pre-calculated domain architectures of database proteins, showing a list of neighbors with matching sets of domains. DART can be accessed at http://www.ncbi.nlm.nih.gov/Structure/lexington/lexington.cgi?cmd=rps.

The DART service covers a non-redundant subset of Entrez's protein database only. In an ongoing effort, CDD will be made a fully integrated part of NCBI's Entrez system, making use of Entrez's powerful search and query refinement engine. As a prerequisite, all proteins in Entrez will be neighbored to CDD, thus adding domain annotations to all protein sequences, annotations which will be refreshed periodically, as both Entrez-protein and CDD will continue to grow. Conserved domain models in Entrez will not only be linked to proteins, but also to three-dimensional structure data, literature, nodes in the taxonomic classification and to other conserved domain models, using suitable definitions for conserved domain neighbor relationships.

We have started to curate multiple alignments in CDD. Our goal is the reconciliation of sequence alignment data with quantitative information from protein three-dimensional structure and structure comparison, resulting in sets of pairwise alignments between each family member and a structure-linked representative that could be used to instantiate initial three-dimensional models for family members. One of the prerequisites is, for example, that structures inferred from alignments do not violate basic geometric principles. We also plan to validate CDD alignments with sequence-structure threading methods (11) as a means to detect errors in the alignments, outliers requiring manual curation and contamination with sequences from outside the family. Setting up a curation pipeline will allow us to periodically update conserved domain alignments with new family members, as sequence databases continue to grow. Carefully curated alignments will also serve as a means to link functional annotation to conserved residues, and to make this annotation accessible in visualization services. A set of curated conserved domains will be available later this year.

ACKNOWLEDGEMENTS

We are grateful to the authors of Pfam and SMART, for creating an invaluable resource. We thank Chris Ponting, Alex Bateman, Eugene Koonin and David Lipman for discussions and helpful suggestions, L. Aravind for providing sequence alignment data, Tom Madden and Sergei Shavirin for making RPS-BLAST available, Richard Copley for providing access to SMART data, and Naomi Ariel for help with the initial analysis of imported alignments.

REFERENCES

1. Park,J., Karplus,K., Barrett,C., Hughey,R., Haussler,D., Hubbard,T. and Chothia,C. (1998) Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.*, **284**, 1201–1210.
2. Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. USA*, **84**, 4355–4358.
3. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
4. Karplus,K., Barrett,C. and Hughey,R. (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, **14**, 846–856.
5. Henikoff,S. and Henikoff,J.G. (1997) Embedding strategies for effective use of information from multiple sequence alignments. *Protein Sci.*, **6**, 698–705.
6. Altschul,S.F., Madden,T.L., Schäffer,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.
7. Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Howe,K.L. and Sonnhammer,E.L.L. (2000) The Pfam proteins family database. *Nucleic Acids Res.*, **28**, 263–266. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 276–280.
8. Schultz,J., Copley,R.R., Doerks,T., Ponting,C.P. and Bork,P. (2000) SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.*, **28**, 231–234. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 242–244.
9. Wang,Y., Geer,L.Y., Chappey,C., Kans,J.A. and Bryant,S.H. (2000) Cn3D: sequence and structure views from Entrez. *Trends Biochem. Sci.*, **25**, 300–302.
10. Wheeler,D.L., Church,D.M., Lash,A.E., Leipe,D.D., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Tatusova,T.A., Wagner,L. and Rapp,B.A. (2001) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **29**, 11–16. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 13–16.
11. Panchenko,A.R., Marchler-Bauer,A. and Bryant,S.H. (1999) Threading with explicit models for evolutionary conservation of structure and sequence. *Proteins*, **37** (Suppl. 3), 133–140.
12. Lamers,M.H., Perrakis,A., Enzlin,J.H., Winterwerp,H.H., de Wind,N. and Sixma,T.K. (2000) The crystal structure of DNA mismatch repair protein MutS binding to a G × T mismatch. *Nature*, **407**, 711–717.