

Chapter 3. Macromolecular Structure Databases

Eric Sayers and Steve Bryant

Created: October 9, 2002; Updated: August 13, 2003.

Summary

The resources provided by NCBI for studying the three-dimensional (3D) structures of proteins center around two databases: the Molecular Modeling Database (MMDB), which provides structural information about individual proteins; and the Conserved Domain Database (CDD), which provides a directory of sequence and structure alignments representing conserved functional domains within proteins (CDs). Together, these two databases allow scientists to retrieve and view structures, find structurally similar proteins to a protein of interest, and identify conserved functional sites.

To enable scientists to accomplish these tasks, NCBI has integrated MMDB and CDD into the Entrez retrieval system ([Chapter 15](#)). In addition, structures can be found by BLAST, because sequences derived from MMDB structures have been included in the BLAST databases ([Chapter 16](#)). Once a protein structure has been identified, the domains within the protein, as well as domain “neighbors” (i.e., those with similar structure) can be found. For novel data not yet included in Entrez, there are separate search services available.

Protein structures can be visualized using Cn3D, an interactive 3D graphic modeling tool. Details of the structure, such as ligand-binding sites, can be scrutinized and highlighted. Cn3D can also display multiple sequence alignments based on sequence and/or structural similarity among related sequences, 3D domains, or members of a CDD family. Cn3D images and alignments can be manipulated easily and exported to other applications for presentation or further analysis.

Overview

The Structure [homepage](#) (Figure 1) contains links to the more specialized pages for each of the main tools and databases, introduced below, as well as search facilities for the Molecular Modeling Database ([MMDB](#); Ref. 1).

[MMDB](#) is based on the structures within Protein Data Bank (PDB) and can be queried using the Entrez search engine, as well as via the more direct but less flexible Structure **Summary** search (see Figure 1). Once found, any structure of interest can be viewed using

The screenshot shows the NCBI Structure homepage. At the top, there's a navigation bar with links for PubMed, BLAST, OMIM, Taxonomy, and Entrez Structure. Below this is a search bar with 'Structure' selected in the dropdown and a 'Go' button. The main content area is divided into several sections: 'Whats New?' with links to MMDB, PDBBeast, Cn3D v4.0, VAST, and VAST Search; 'The NCBI Structure Group' with a description of MMDB and a 'try also:' section for Structure Summary search; 'Structure highlights' featuring Cn3D 4.1; and 'Read more about:' with links to MMDB, WWW-Entrez, and VAST. Resources for MMDB's FTP-site, NCBI Toolkit, MMDB-API, and Cn3D source code are also listed.

Figure 1. The Structure homepage. This page can be found by selecting the **Structure** link on the tool bar atop many NCBI Web pages. Two searches can be performed from this page, an **Entrez Structure** search or a Structure **Summary** search. Both query the MMDB database. The difference is that the **Entrez Structure** can take any text as a query (such as a PDB code, protein name, text word, author, or journal) and will result initially in a list of one or more document summaries, displayed within the Entrez environment (Chapter 15), whereas only a PDB code or MMDB ID number can be used for the Structure **Summary** search, resulting in direct display of the Structure Summary page for that record (Figure 2). Announcements about new features or updates can also be found on this page, as well as links to more specialized pages on the various Structure databases and tools.

Cn3D (2), a piece of software that can be freely downloaded for Mac, PC, and UNIX platforms.

Often used in conjunction with **Cn3D** is the Vector Alignment Search Tool (**VAST**; Refs. 3, 4). **VAST** is used to precompute “structure neighbors” or structures similar to each MMDB entry. For people that have a set of 3D coordinates for a protein not yet in MMDB, there is also a **VAST search service**. The output of the precomputed VAST searches is a list of structure records, each representing one of the **Non-Redundant PDB chain** sets (**nr-PDB**), which can also be downloaded. There are four clustered subsets of MMDB that compose nr-PDB, each consisting of clusters having a preset level of sequence similarity.

Figure 2. The Structure Summary page. The page consists of three parts: the header, the view bar, and the graphic display. The header contains basic identifying information about the record: a description of the protein (*Description:*), the author list (*Deposition:*), the species of origin (*Taxonomy:*), literature references (*Reference:*), the MMDB-ID (*MMDB:*), and the PDB code (*PDB:*). Several of these data serve as links to additional information. For example, the species name links to the Taxonomy browser, the literature references link to PubMed, and the PDB code links to the PDB Web site. The view bar allows the user to view the structure record either as a graphic with Cn3D or as a text record in either ASN.1, PDB (RasMol), or Mage formats. The latter can also be downloaded directly from this page. The graphic display contains a variety of information and links to related databases: (a) The Chain bar. Each chain of the molecule is displayed as a *dark bar* labeled with residue numbers. To the *left* of this bar is a **Protein** hyperlink that takes the user to a view of the protein record in Entrez Protein. The bar itself is also a hyperlink and displays the VAST neighbors of the chain. If a structure contains nucleotide sequences, they are displayed in the order contained in the PDB record. A **Nucleotide** hyperlink to their *left* takes the user to the appropriate record in Entrez Nucleotide. (b) The VAST (3D) Domain bar. The *colored bars* immediately below the Chain bar indicate the locations of structural domains found by the original MMDB processing of the protein. In many cases, such a domain contains unconnected sections of the protein sequence, and in such cases, discontinuous pieces making up the domain will have bars of the same color. To the *left* of the Domain bar is a 3D Domains hyperlink (*3d Domains*) that launches the 3D Domains browser in Entrez, where the user can find information about each constituent domain. Selecting a colored segment displays the VAST Structure Neighbors page for that domain. (c) The CD bar. Below the VAST Domain bar are *rounded, rectangular bars* representing conserved domains found by a CD-Search. The bars identify the best scoring hits; overlapping hits are shown only if the mutual overlap with hits having better scores is less than 50%. The *CDs* hyperlink to the *left* of the bar displays the CD records in Entrez Domains. Each of the colored bars is also a hyperlink that displays the corresponding CD Summary page configured to show the multiple alignment of the protein sequence with members of the selected CD.

The structures within MMDB are now being linked to the NCBI Taxonomy database ([Chapter 4](#)). Known as the [PDBeast](#) project, this effort makes it possible to find: (1) all MMDB structures from a particular organism; and (2) all structures within a node of the taxonomy tree (such as lizards or *Bacillus*), which launches the Taxonomy Browser showing the number of MMDB records in each node.

The second database within the **Structure** resources is the Conserved Domain Database (CDD; Ref. 5), originally based largely on [Pfam](#) and [SMART](#), collections of alignments that represent functional domains conserved across evolution. CDD now also contains the alignments of the NCBI COG database, the NCBI Library of Ancient Domains (LOAD) along with new curated alignments assembled at NCBI. CDD can be searched from the [CDD](#) page in several ways, including by a domain keyword search. Three tools have been developed to assist in analysis of CDD: (1) the [CD-Search](#), which uses a [BLAST](#)-based algorithm to search the position-specific scoring matrices (PSSM) of CDD alignments; (2) the CD-Browser, which provides a graphic display of domains of interest, along with the sequence alignment; and (3) the Conserved Domain Architecture Retrieval Tool ([CDART](#)), which searches for proteins with similar domain architectures.

All the above databases and tools are discussed in more detail in other parts of this Chapter, including tips on how to make the best use of them.

Content of the Molecular Modeling Database (MMDB)

Sources of Primary Data

To build MMDB (1), 3D structure data are retrieved from the PDB database (6) administered by the Research Collaboratory for Structural Bioinformatics ([RCSB](#)). In all cases, the structures in MMDB have been determined by experimental methods, primarily X-ray crystallography and Nuclear Magnetic Resonance ([NMR](#)) spectroscopy. Theoretical structure models are omitted. The data in each record are then checked for agreement between the atomic coordinates and the primary sequence, and the sequence data are then extracted from the coordinate set. The resulting agreement between sequence and structure allows the record to be linked efficiently into searches and alignment displays involving other NCBI databases.

The data are converted into [ASN.1](#) (7), which can be parsed easily and can also accept numerous annotations to the structure data. In contrast to a PDB record, a MMDB record in [ASN.1](#) contains all necessary bonding information in addition to sequence information, allowing consistent display of the 3D structure using [Cn3D](#). The annotations provided in the PDB record by the submitting authors are added, along with uniformly defined secondary structure and domain features. These features support structure-based similarity searches using [VAST](#). Finally, two coordinate subsets are added to the record: one containing only backbone atoms, and one representing a single-conformer model in cases where multiple conformations or structures were present in the PDB record. Both of these additions further simplify viewing both an individual structure and its alignments

with structure neighbors in Cn3D. When this process is complete, the record is assigned a unique Accession number, the MMDB-ID (Box 1), while also retaining the original four-character PDB code.

Box 1. Accession numbers.

MMDB records have several types of Accession numbers associated with them, representing the following data types:

- Each MMDB record has at least three Accession numbers: the PDB code of the corresponding PDB record (e.g., 1CYO, 1B8G); a unique MMDB-ID (e.g., 645, 12342); and a gi number for each protein chain. A new MMDB-ID is assigned whenever PDB updates either the sequence or coordinates of a structure record, even if the PDB code is retained.
- If an MMDB record contains more than one polypeptide or nucleotide chain, each chain in the MMDB record is assigned an Accession number in Entrez Protein or Nucleotide consisting of the PDB code followed by the letter designating that chain (e.g., 1B8GA, 3TATB, 1MUHB).
- Each 3D Domain identified in an MMDB record is assigned a unique integer identifier that is appended to the Accession number of the chain to which it belongs (e.g., 1B8G A 2). This new Accession number becomes its identifier in Entrez 3D Domains. New 3D Domain identifiers are assigned whenever a new MMDB-ID is assigned.
- For conserved domains, the Accession number is based on the source database:

Pfam:	pfam00049
SMART:	smart00078
LOAD:	LOAD Toprim
CD:	cd00101
COG:	COG5641

Annotation of 3D Domains

After initial processing, 3D domains are automatically identified within each MMDB record. 3D domains are annotations on individual MMDB structures that define the boundaries of compact substructures contained within them. In this way, they are similar to secondary structure annotations that define the boundaries of helical or β -strand substructures. Because proteins are often similar at the level of domains, VAST compares each 3D domain to every other one and to complete polypeptide chains. The results are stored in Entrez as a **Related 3D Domain** link.

To identify 3D domains within a polypeptide chain, MMDB's domain parser searches for one or more breakpoints in the structure. These breakpoints fall between major secondary

structure elements such that the ratio of intra- to interdomain contacts remains above a set threshold. The 3D domains identified in this way provide a means to both increase the sensitivity of structure neighbor calculations and also present 3D superpositions based on compact domains as well as on complete polypeptide chains. They are not intended to represent domains identified by comparative sequence and structure analysis, nor do they represent modules that recur in related proteins, although there is often good agreement between domain boundaries identified by these methods.

Links to Other NCBI Resources

After initially processing the PDB record, structure staff add a number of links and other information that further integrate the MMDB record with other NCBI resources. To begin, the sequence information extracted from the PDB record is entered into the Entrez Protein and/or Nucleotide databases as appropriate, providing a means to retrieve the structure information from sequence searches. As with all sequences in Entrez, precomputed BLAST searches are then performed on these sequences, linking them to other molecules of similar sequence. For proteins, these BLAST neighbors may be different than those determined by VAST; whereas VAST uses a conservative significance threshold, the structural similarities it detects often represent remote relationships not detectable by sequence comparison. The literature citations in the PDB record are linked to PubMed so that Entrez searches can allow access to the original descriptions of the structure determinations. Finally, semiautomatic processing of the “source” field of the PDB record provides links to the NCBI Taxonomy database. Although these links normally follow the genus and species information given, in some cases this information is either absent in the PDB record or refers only to how a sample was obtained. In these cases, the staff manually enters the appropriate taxonomy links.

The MMDB Record

The Structure Summary page for each MMDB record summarizes the database content for that record and serves as a starting point for analyzing the record using the NCBI structure tools (Figure 2).

VAST Structure Neighbors

Although VAST itself is not a database, the VAST results computed for each MMDB record are stored with this record and are summarized on a separate page for the whole polypeptide chain as well as for each 3D domain found in the protein (Figure 3). These pages can be accessed most easily by clicking on either the chain bar or the 3D Domain bar in the graphic display of the Structure Summary page (Figure 2).

nr-PDB

The non-redundant PDB database (nr-PDB) is a collection of four sets of sequence-dissimilar cluster PDB polypeptide chains assembled by NCBI Structure staff. The four sets differ only in their respective levels of non-redundancy. The staff assembles each set

NCBI

VAST
Structure Neighbors

PubMed BLAST Structure Taxonomy OMIM Help? Cn3d

Query: MMDB 20320, 1GW5 chain A domain 2
Description: Ap2 Clathrin Adaptor Core

View 3D Structure of All Atoms with Cn3D Display [Get Cn3D 4.1!](#)

View Alignment using Hypertext for Selected VAST neighbors

List subset NR, Blast_p < 10e-40 sorted by Aligned Length in Graphics

Find MMDB or PDB ids (separated by ","):

43 structure neighbors displayed.

1GW5 # 3d Dom. CDs

1 100 200 300 400 500 621 Ali_Res.

1 2 3 4 5 6 Rdapt.in_N

108K B 6 143

Figure 3. VAST Structure Neighbors page. The *top portion* of the page contains identifying information about the 3D Domain, along with three functional bars. (a) The View bar. This bar allows a user to view a selected alignment either as a graphic using Cn3D or as a sequence alignment in HTML, text, or mFASTA format. The user may select which chains to display in the alignment by checking the boxes that appear to the *left* of each neighbor in the *lower portion* of the page. (b) The nr-PDB bar. This bar allows a user to either display all matching records in MMDB or to limit the displayed domains to only representatives of the selected nr-PDB set. The user may also select how the matching domains are sorted in the display and whether the results are shown as graphics or as tabulated data. (c) The Find bar. This bar allows the user to find specific structural neighbors by entering their PDB or MMDB identifiers. (d) The *lower portion* of the page displays a graphical alignment of the various matching domains. The *upper three bars* show summary information about the query sequence: the *top bar* shows the maximum extent of alignment found on all the sequences displayed on the current page (users should note that the appearance of this bar, therefore, depends on which hits are displayed); the *middle bar* represents the query sequence itself that served as input for the VAST search; and the *lower bar* shows any matching CDs and is identical to the CD bar on the Structure Summary page. Listed below these three summary bars are the hits from the VAST search, sorted according to the selection in the nr-PDB bar. Aligned regions are shown in red, with gaps indicating unaligned regions. To the *left* of each domain accession is a check box that can be used to select any combination of domains to be displayed either on this page or using Cn3D. Moreover, each of the bars in the display is itself a link, and placing the mouse pointer over any bar reveals both the extent of the alignment by residue number and the data linked to the bar.

by comparing all the chains available from PDB with each other using the BLAST algorithm. The chains are then clustered into groups of similar sequence using a single-linkage clustering procedure. Chains within a sequence-similar group are automatically ranked according to the quality of their structural data. Details of the measures used to determine structure precision and completeness and the methodology of assembling the nr-PDB clusters can be found on the nr-PDB [Web page](#).

Content of the Conserved Domain Database (CDD)

What Is a Conserved Domain (CD)?

CDs are recurring units in polypeptide chains (sequence and structure motifs), the extents of which can be determined by comparative analysis. Molecular evolution uses such domains as building blocks and these may be recombined in different arrangements to make different proteins with different functions. The CDD contains sequence alignments that define the features that are conserved within each domain family. Therefore, the CDD serves as a classification resource that groups proteins based on the presence of these predefined domains. CDD entries often name the domain family and describe the role of conserved residues in binding or catalysis. Conserved domains are displayed in MMDB Structure summaries and link to a sequence alignment showing other proteins in which the domain is conserved, which may provide clues on protein function.

Sources of Primary Data

The collections of domain alignments in the CDD are imported either from two databases outside of the NCBI, named Pfam (8) and SMART (9); from the NCBI COB database; from another NCBI collection named LOAD; and from a database curated by the CDD staff. The first task is to identify the underlying sequences in each collection and then link these sequences to the corresponding ones in Entrez. If the CDD staff cannot find the Accession numbers for the sequences in the records from the source databases, they locate appropriate sequences using BLAST. Particular attention is paid to any resulting match that is linked to a structure record in MMDB, and the staff substitute alignment rows with such sequences whenever possible. After the staff imports a collection, they then choose a sequence that best represents the family. Whenever possible, the staff chooses a representative that has a structure record in MMDB.

The Position-specific Score Matrix (PSSM)

Once imported and constructed, each domain alignment in CDD is used to calculate a model sequence, called a [consensus sequence](#), for each CD. The consensus sequence lists the most frequently found residue in each position in the alignment; however, for a sequence position to be included in the consensus sequence, it must be present in at least 50% of the aligned sequences. Aligned columns covered by the consensus sequence are then used to calculate a [PSSM](#), which memorizes the degree to which particular residues are conserved at each position in the sequence. Once calculated, the PSSM is stored with

the alignment and becomes part of the CDD. The [RPS-BLAST](#) tool locates CDs within a query sequence by searching against this database of PSSMs.

Reverse Position-specific BLAST (RPS-BLAST)


RPS-BLAST ([Chapter 16](#)) is a variant of the popular Position-specific Iterated BLAST (PSI-BLAST) program. PSI-BLAST finds sequences similar to the query and uses the resulting alignments to build a PSSM for the query. With this PSSM the database is scanned again to draw in more hits and further refine the scoring model. RPS-BLAST uses a query sequence to search a database of precalculated PSSMs and report significant hits in a single pass. The role of the PSSM has changed from “query” to “subject”; hence, the term “reverse” in RPS-BLAST. RPS-BLAST is the search tool used in the CD-Search service.

The CD Summary

Analogous to the Structure Summary page, the CD Summary page displays the available information about a given CD and offers various links for either viewing the CD alignment or initiating further searches ([Figure 4](#)). The CD Summary page can be retrieved by selecting the CD name on any page.

CD Records Curated at NCBI

In 2002, NCBI released the first group of curated CD records, a new and expanding set of annotated protein multiple sequence alignments and corresponding structure alignments. These new records have Accession numbers beginning with “cd” and have been added to the default CD-Search database. Most curated CD records are based on existing family descriptions from SMART and Pfam, but the alignments may have been revised extensively by quantitatively using three-dimensional structures and by re-examining the domain extent. In addition, CDD curators annotate conserved functional residues, ligands, and co-factors contained within the structures. They also record evidence for these sites as pointers to relevant literature or to three-dimensional structures exemplifying their properties. These annotations may be viewed using Cn3D and thus provide a direct way of visualizing functional properties of a protein domain in the context of its three-dimensional structure. (See [Box 3](#) and [Figure 7](#).)

 **Conserved Domain Database**

PubMed Nucleotide Protein Structure **CDD** Taxonomy Help?

CD: [pfam01602.6, Adaptin_N](#), Query added PSM-Id: 6518 Source: [Pfam\[US\]](#), [Pfam\[UK\]](#)

Description: Adaptin N terminal region. This family consists of the N terminal region of various alpha, beta and gamma subunits of the AP-1, AP-2 and AP-3 adaptor protein complexes. The adaptor protein (AP) complexes are involved in the formation of clathrin-coated pits and vesicles. The N-terminal region of the various adaptor proteins (APs) is constant by comparison to the C-terminal which is variable within members of the AP-2 family; and it has been proposed that this constant region interacts with another uniform component of the coated vesicles.

Taxa: [Eukaryota](#) **References:** [2 Pubmed Links](#)

Status: Alignment from source **Created:** 13-Jun-2002

Aligned: 35 rows **PSSM:** 535 columns **Representative:** Consensus

Proteins: [\[Click here for CDART summary of Proteins containing pfam01602\]](#)

View Alignment as width color at

Subset Rows

		10	20	30	40	50	60
consensus	1	..*... ...*... ...*... ...*... ...*...					
lgw5a(query)	28	aEIKRINKELANIRSKFKGdka-----ldGYSKKKYVCKLLFIFLLG----					EDISFL 44
gi_586420	37	EQEKRIQSEIVKIKQHFDAAkkkqgnhdrldGYSKKKYVAKLAYIYITSnttklNEILFG					96
gi_3372671	25	ERAVVRKECADIRALINEDd-----PHDRHRNLAKLMFIHMLG----					YPTHFG 69
gi_5902737	24	DERSLIQKESASIRTAFKDEd-----PFARHNNIAKLLYIHMLG----					YPAHFG 68
gi_12643299	23	QEREVIQKECAHIRASFRDgd-----PVHRHRQLAKLLYVHMLG----					YPAHFG 67
gi_12643391	22	EEREMIQKECAAIRSSFREEd-----NTYRCRNVAKLLYMHMLG----					YPAHFG 66
gi_3912968	27	AEVKRINKELANIRSKFKGdkt-----ldGYQKKKYVCKLLFIFLLG----					HDIDFG 74
gi_113339	28	AEIKRINKELANIRSKFKGdka-----ldGYSKKKYVCKLLFIFLLG----					HDIDFG 75
gi_15011827	27	AEIKRINKELANIRSKFKGdkt-----ldGYQKKKYVCKLLFIFLLG----					NDIDFG 74

Figure 4. CD summary page. The *top* of the page serves as a header and reports a variety of identifying information, including the name and description of the CD, other related CDs with links to their summary pages, as well as the source database, status, and creation date of the CD. A taxonomic node link (*Taxa:*) launches the Taxonomy Browser, whereas a Proteins link (*Proteins:*) uses CDART to show other proteins that contain the CD. *Below* the header is the interface for viewing the CD alignment, which can be done either graphically with Cn3D (if the CD contains a sequence with structural data) or in HTML, text, or mFASTA format. It is also possible to view a selected number of the top-listed sequences, sequences from the most diverse members, or sequences most similar to the query. In addition, users may now select sequences with the NCBI Taxonomy Common Tree tool. The *lower portion* of the page contains the alignment itself. Members with a structural record in MMDB are listed first, and the identifier of each sequence links to the corresponding record.

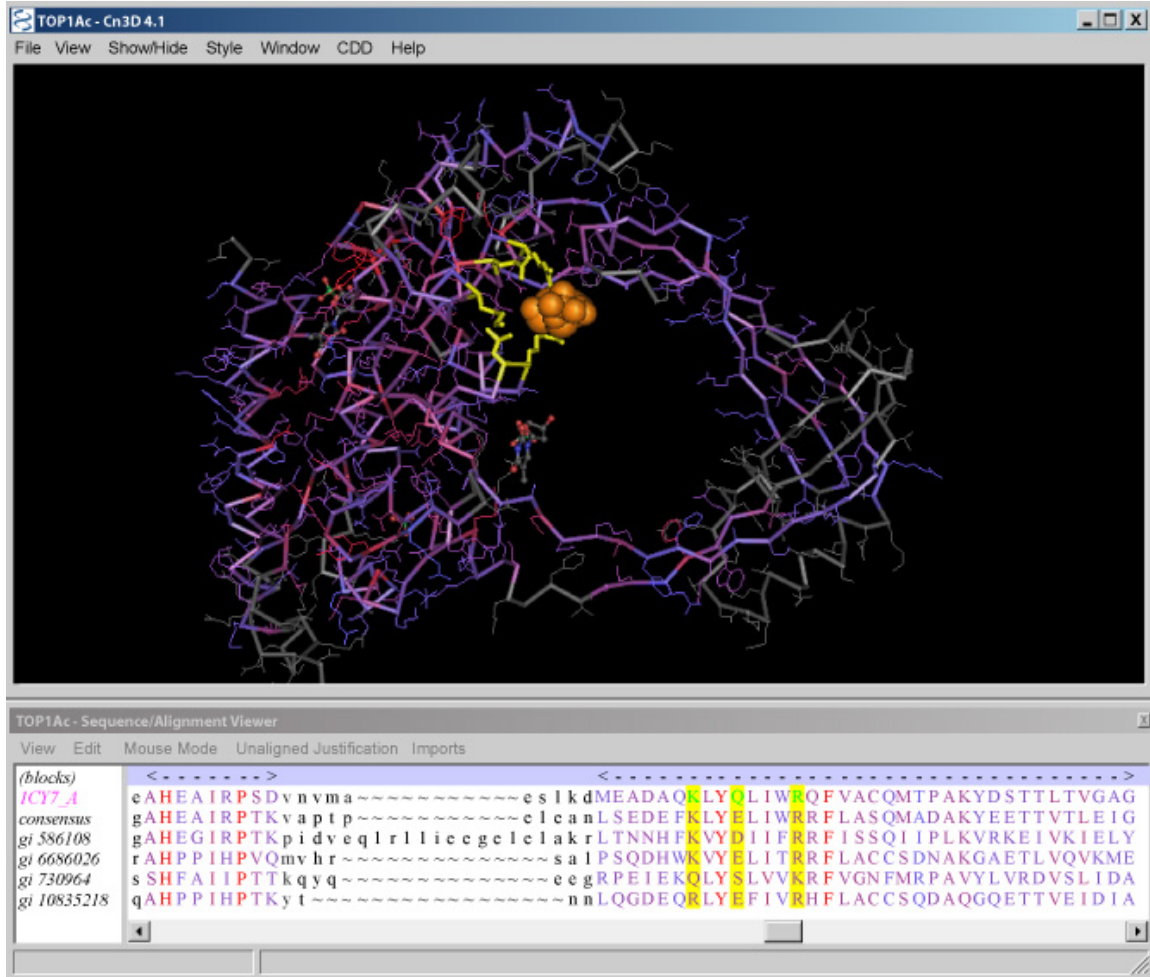


Figure 7. Sequence and structure views of the TOP1Ac conserved domain common to type III bacterial and eukaryotic DNA topoisomerases. The *upper window* displays the structure of the domain with the residues colored according to their sequence conservation, with *red* indicating high conservation and *blue* indicating low conservation. The nucleotide bound at site II is shown as an *orange* space-filling model, and the residues involved in this binding site are *yellow*. The *lower window* displays the sequence alignment for the domain with aligned residues shown as colored *capital letters*. Residues aligned to three of the binding site residues are highlighted in *yellow*. The sequence for NP_004609 (gi 10835218) occupies the *bottom row*.

Box 3. Example query: finding and viewing CDs in a protein.

Finding CDs in a Protein

Suppose that we are interested in topoisomerase enzymes and would like to find human topoisomerases that most closely resemble those found in eubacteria and thus may share a common ancestor. Further suppose that through a colleague, we are aware of a recent and particularly interesting crystal structure of a topoisomerase from *Escherichia coli* with PDB code 1I7D. How can we identify the conserved functional domains in this protein

Box 3 continues on next page...

Box 3 continued from previous page.

and then find human proteins with the same domains? From the Structure main page, we enter the PDB code 1I7D in the Structure **Summary** search box and quickly find the Structure Summary page for this record. We see that in this crystal structure, the protein is complexed with a single-stranded oligonucleotide. We also see that the protein has five 3D Domains. Two CDs align to the sequence as well, and they overlap with one another at the N-terminus of the protein in the region corresponding to the first 3D domain.

Analyzing CDs Found in a Protein

The Structure summary page displays only the CDs that give the best match to the protein sequence. To see all of the matching CDs, we can easily perform a full CD-Search. Select the **Protein** link to the left of the graphic to reveal the flatfile for the record. Then follow the **Domains** link in the **Link** menu on the right to view the results of the CD-Search. Select **Show Details** to see all CDs matching the query sequence. We find that nine CDs match this sequence, and that the statistics of each match are shown below the alignment graphic. The CD with the best hit is TopA from the COG database, and it is further clear that this domain consists of two smaller domains: TOPRIM (alignments from Pfam, SMART, and curated CD) and a topoisomerase domain (alignments from Pfam and curated CD). We can learn more about these CDs by studying the pairwise alignments at the bottom of the page and by studying their CD Summary pages, reached by selecting the links to their left.

Finding Other Proteins with Similar Domain Architecture

We now would like to find human proteins that have these same CDs. To perform a CDART search, simply select the **Show Domain Relatives** button. To limit these results to human proteins, we select the **Subset by Taxonomy** button. A taxonomic tree is then displayed, and we next check the box for **Mammal**, the lowest taxa including *Homo sapiens*. Selecting **Choose** then displays a Common Tree, and by clicking on the appropriate “scissor” icons, we can cut away all branches except the one leading to *H. sapiens*. We can execute this taxonomic restriction by selecting **Go back**, and we now find a much shorter list of CDART results. In the most similar group, we find two members, one of which is NP_004609. Selecting the **more>** link for this record shows the CD-Search results for this human protein. Interestingly, we find that the topoisomerase is very well conserved, whereas only a portion of the TOPRIM domain has been retained.

Viewing a CD Alignment with a 3D Structure

We now would like to view the alignment of the topoisomerase in the human protein to other members of this CD. On the CD-Search page, select the colored bar of this CD to see a CD-Browser window displaying the alignment. Because this is a curated CD record, we are able to view functional features of the protein domain on a structural template. The rightmost menu in the View Alignment bar shows the available features for this domain,

Box 3 continues on next page...

Box 3 continued from previous page.

whereas the topmost row in the alignment itself marks the residues involved in this feature with # symbols. The second row of the alignment is the consensus sequence of the CD record, whereas the third row contains the NP_004609 sequence, labeled “query”. At the bottom of the page, buttons allow Cn3D to be launched with various structural features highlighted. For example, if we are interested in nucleotide binding site II, Cn3D will launch with the view depicted in Figure 7, showing the bound nucleotide in orange. Additional Cn3D windows not shown in Figure 7 allow one to highlight the binding site residues yellow as shown, and these highlights also appear in the sequence window. In this figure, the NP_004609 sequence has been merged into the alignment (bottom row) using tools within Cn3D, and the result shows that this human protein closely conserves these important functional residues.

The Distinction between 3D Domains and CDs

The term “domain” refers in general to a distinct functional and/or structural unit of a protein. Each polypeptide chain in MMDB is analyzed for the presence of two classes of domains, and it is important for users to understand the difference between them. One class, called 3D Domains, is based solely on similar, compact substructures, whereas the second class, called Conserved Domains (CDs), is based solely on conserved sequence motifs. These two classifications often agree, because the compact substructures within a protein often correspond to domains joined by recombination in the evolutionary history of a protein. Note that CD links can be identified even when no 3D structures within a family are known. Moreover, 3D Domain links may also indicate relationships either to structures not included in CDD entries or to structures so distantly related that no significant similarity can be found by sequence comparisons.

Finding and Viewing Structures

For an example query on finding and viewing structures, see Box 2.

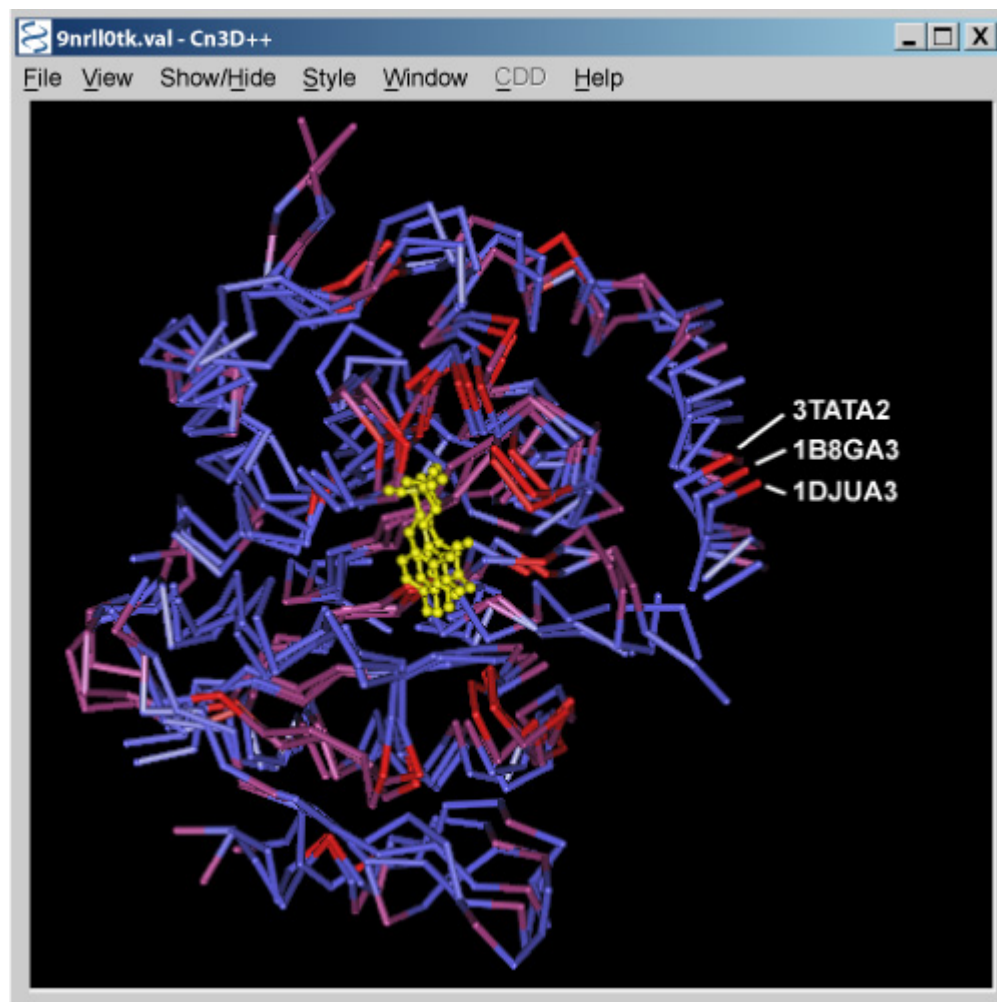


Figure 6. VAST structural alignment of 1B8GA3, 3TATA2, and 1DJUA3. The backbone atoms of the aligned residues of the three structures are shown colored according to their sequence conservation of each position in the alignment. Highly conserved positions are colored more *red*, whereas poorly conserved positions are colored more *blue*. The bound pyridoxal phosphate ligands are *yellow*.

Box 2. Example query: finding and viewing structural data of a protein.

Finding the Structure of a Protein

Suppose that we are interested in the biosynthesis of aminocyclopropanes and would like to find structural information on important active site residues in any available aminocyclopropane synthases. To begin, we would go to the **Structure** main page and enter “aminocyclopropane synthase” in the **Search** box. Pressing Enter displays a short list of structures, one of which is 1B8G, 1-aminocyclopropane-1-carboxylate synthase. Perhaps we would like to know the species from which this protein was derived. Selecting the **Taxonomy** link to the right shows that this protein was derived from *Malux x*

Box 2 continues on next page...

Box 2 continued from previous page.

domestica, or the common apple tree. Going back to the Entrez results page and selecting the PDB code (1B8G) opens the Structure Summary page for this record. The species is again displayed on this page, along with a link to the *Journal of Molecular Biology* article describing how the structure was determined. We immediately see from this page that this protein appears as a dimer in the structure, with each chain having three 3D domains, as identified by VAST. In addition, CD-Search has identified an “aminotran_1_2” CD in each chain. Now we are ready to view the 3D structure.

Viewing the 3D Structure

Once we have found the Structure Summary page, viewing the 3D structure is straightforward. To view the structure in Cn3D, we simply select the **View 3D Structure** button. The default view is to show helices in green, strands in brown, and loops in blue. This color scheme is also reflected in the Sequence/Alignment Viewer.

Locating an Active Site

Upon inspecting the structure, we immediately notice that a small molecule is bound to the protein, likely at the active site of the enzyme. How do we find out what that molecule is? One easy way is to return to the Structure Summary page and select the link to the PDB code, which takes us to the PDB Structure Explorer page for 1B8G. Quickly, we see that pyridoxal-5'-phosphate (PLP) is a HET group, or heterogen, in the structure. Our interest piqued, we would now like to know more about the structural domain containing the active site. Returning to Cn3D, we manipulate the structure so that PLP is easily visible and then use the mouse to double-click on any PLP atom. The molecule becomes selected and turns yellow. Now from the **Show/Hide** menu, we choose **Select by distance and Residues only** and enter 5 Angstroms for a search radius. Scanning the Sequence/Alignment Viewer, we see that seven residues are now highlighted: 117-119, 230, 268, 270, and 279. Glancing at the 3D Domain display in the Structure Summary page, we note that all of these residues lie in domain 3. We now focus our attention on this domain.

Viewing Structure Neighbors of a 3D Domain

Given that this enzyme is a dimer, we arbitrarily choose domain 3 in chain A, the accession of which is thus 1B8GA3. By clicking on the 3D Domain bar at a point within domain 3, we are taken to the VAST Structure Neighbors page for this domain, where we find nearly 200 structure neighbors.

Restricting the Search by Taxonomy

Perhaps we would now like to identify some of the most evolutionarily distant structure neighbors of domain 1B8GA3 as a means of finding conserved residues that may be associated with its binding and/or catalytic function. One powerful way of doing this is to

Box 2 continues on next page...

Box 2 continued from previous page.

choose structure neighbors from phylogenetically distant organisms. We therefore need to combine our present search with a Taxonomy search. Given that 1B8G is derived from the superkingdom Eukaryota, we would like to find structure neighbors in other superkingdom taxa, such as Eubacteria and Archaea. Returning to the Structure Summary page, select the 3D Domains link in the graphic display to open the list of 3D Domains in Entrez. Finding 1B8GA3 in the list, selecting the **Related 3D Domains** link shows a list of all the structure neighbors of this domain. From this page, we select **Preview/Index**, which shows our recent queries. Suppose our set of related 3D Domains is #5. We then perform two searches:

1. #5 AND "Archaea"[Organism]
2. #5 AND "Eubacteria"[Organism]

Looking at the Archaea results, we find among them 1DJUA3, a domain from an aromatic aminotransferase from *Pyrococcus horikoshii*. Concerning the Eubacteria results, we find among the several hundred matching domains 3TATA2, a tyrosine aminotransferase from *Escherichia coli*.

Viewing a 3D Superposition of Active Sites

Returning to the VAST Structure Neighbors page for 1B8GA3, we want to select 1DJUA3 and 3TATA2 to display in a structural alignment. One way to do this is to enter these two Accession numbers in the **Find** box and press **Find**. We now see only these two neighbors, and we can select **View 3D Structure** to launch Cn3D.

Cn3D again displays the aligned residues in red, and we can highlight these further by selecting **Show aligned residues** from the **Show/Hide** menu. The excellent agreement between both the active site structures and the conformations of the bound ligands is readily apparent. Furthermore, by selecting **Style/Coloring Shortcuts/Sequence Conservation/Variety**, we can easily see that the most highly conserved residues are concentrated near the binding site (Figure 6).

Why Would I Want to Do This?

- To determine the overall shape and size of a protein
- To locate a residue of interest in the overall structure
- To locate residues in close proximity to a residue of interest
- To develop or test chemical hypotheses regarding an enzyme mechanism
- To locate or predict possible binding sites of a ligand
- To interpret mutation studies
- To find areas of positive or negative charge on the protein surface
- To locate particularly hydrophobic or hydrophilic regions of a protein

- To infer the 3D structure and related properties of a protein with unknown structure from the structure of a **homologous** protein
- To study evolutionary processes at the level of molecular structure
- To study the function of a protein
- To study the molecular basis of disease and design novel treatments

How to Begin

The first step to any structural analysis at NCBI is to find the structure records for the protein of interest or for proteins similar to it. One may search MMDB directly by entering search terms such as PDB code, protein name, author, or journal in the Entrez Structure **Search** box on the Structure [homepage](#). Alternative points of entry are shown below.

By using the full array of Entrez search tools, the resulting list of MMDB records can be honed, ideally, to a workable list from which a record can be selected. Users should note that multiple records may exist for a given protein, reflecting different experimental techniques, conditions, and the presence or absence of various ligands or metal ions. Records may also contain different fragments of the full-length molecule. In addition, many structures of mutant proteins are also available. The PDB record for a given structure generally contains some description of the experimental conditions under which the structure was determined, and this file can be accessed by selecting the PDB code link at the top of the Structure Summary page.

Alternative Points of Entry

Structure Summary pages can also be found from the following NCBI databases and tools:

- Select the Structure **links** to the right of any Entrez record found; records with Structure links can also be located by choosing **Structure links** from the **Display** pull-down menu.
- Select the **Related Sequences** link to the right of an Entrez record to find proteins related by sequence similarity and then select **Structure links** in the **Display** pull-down menu.
- Choose the PDB database from a **blastp** (protein-protein BLAST) search; only sequences with structure records will be retrieved by BLAST. The **Related Structures** link provides 3D views in Cn3D.
- Select the **3D Structures** button on any **BLink** report to show those BLAST hits for which structural data are available.
- From the results of any protein BLAST search, click on a red 'S' linkout to view the sequence alignment with a structure record.

Viewing 3D Structures

3D Domains

The 3D domains of a protein are displayed on the Structure Summary page. It is useful to know how many 3D domains a protein contains and whether they are continuous in sequence when viewing the full 3D structure of the molecule.

Secondary Structure

Knowing the secondary structure of a protein can also be a useful prelude to viewing the 3D structure of the molecule. The secondary structure can be viewed easily by first selecting the **Protein** link to the left of the desired chain in the graphic display. Finding oneself in Entrez Protein, selecting **Graphics** in the Display pull-down menu presents secondary structure diagrams for the molecule.

Full Protein Structures

Cn3D is a software package for displaying 3D structures of proteins. Once it has been [installed](#) and the Internet browser has been configured correctly, simply selecting the **View 3D Structure** button on a Structure Summary page launches the application. Once the structure is loaded, a user can manipulate and annotate it using an array of options as described in the [Cn3D Tutorial](#). By default, Cn3D colors the structure according to the secondary structure elements. However, another useful view is to color the protein by domain (see **Style** menu options), using the same color scheme as is shown in the graphic display on the Structure Summary page. These color changes also affect the residues displayed in the Sequence/Alignment Viewer, allowing the identification of domain or secondary structure elements in the primary sequence. In addition to Cn3D, users can also display 3D structures with RasMol or Mage. Structures can also be saved locally as an ASN.1, PDB, or Mage file (depending on the choice of structure viewer) for later display.

Finding and Viewing Structure Neighbors

For an example query on finding and viewing structure neighbors, see Box 2.

Why Would I Want to Do This?

- To determine structurally conserved regions in a protein family
- To locate the structural equivalent of a residue of interest in another related protein
- To gain insights into the allowable structural variability in a particular protein family
- To develop or test chemical hypotheses regarding an enzyme mechanism
- To predict possible binding sites of a ligand from the location of a binding site in a related protein
- To identify sites where conformational changes are concentrated
- To interpret mutation studies
- To find areas of conserved positive or negative charge on the protein surface

- To locate conserved hydrophobic or hydrophilic regions of a protein
- To identify evolutionary relationships across protein families
- To identify functionally equivalent proteins with little or no sequence conservation

How to Begin

The Vector Alignment Search Tool (VAST) is used to calculate similar structures on each protein contained in the MMDB. The graphic display on each Structure Summary page (Figure 2) links directly to the relevant VAST results for both whole proteins and 3D domains:

- The 3D Domains link transfers the user to Entrez 3D Domains, showing a list of the VAST neighbors.
- Selecting the chain bar displays the VAST Structure Neighbors page for the entire chain.
- Selecting a 3D Domain bar displays the VAST Structure Neighbors page for the selected domain.

Alternative Points of Entry

- From any Entrez search, select **Related 3D Domains** to the right of any record found to view the Vast Structure Neighbors page.

Viewing a 2D Alignment of Structure Neighbors

A graphic 2D HTML alignment of VAST neighbors can be viewed as follows:

- On the lower portion of the VAST Structure Neighbors page (Figure 3), select the desired neighbors to view by checking the boxes to their left.
- On the **View/Save** bar, configure the pull-down menus to the right of the **View Alignment** button.
- Select **View Alignment**.

Viewing a 3D Alignment of Structure Neighbors

Alignments of VAST structure neighbors can be viewed as a 3D image using Cn3D.

- On the lower portion of the VAST Structure Neighbors page (Figure 3), select the desired neighbors to view by checking the boxes to their left.
- On the **View/Save** bar, configure the pull-down menus to the right of the **View 3D Structure** button.
- Select **View 3D Structure**.

Cn3D automatically launches and displays the aligned structures. Each displayed chain has a unique color; however, the portions of the structures involved in the alignment are shown in red. These same colors are also reflected in the Sequence/Alignment Viewer. Among the many viewing options provided by Cn3D, of particular use is the **Show/Hide**

menu that allows only the aligned residues to be viewed, only the aligned domains, or all residues of each chain.

Finding and Viewing Conserved Domains

For an example query on finding and viewing conserved domains, see Box 3.

Why Would I Want to Do This?

- To locate functional domains within a protein
- To predict the function of a protein whose function is unknown
- To establish evolutionary relationships across protein families
- To interpret mutation studies
- To predict the structure of a protein of unknown structure

How to Begin

Following the Domains link for any protein in Entrez, one can find the conserved domains within that protein. The [CD-Search](#) (or Protein BLAST, with CD-Search option selected) can be used to find conserved domains (CDs) within a protein. Either the Accession number, gi number, or the [FASTA](#) sequence can be used as a query.

Alternative Points of Entry

Information on the CDs contained within a protein can also be found from these databases and tools:

- From any Entrez search: select the **Domains** link to the right of a displayed record.
- From the Structure Summary page of a MMDB record: this page displays the CDs within each protein chain immediately below the 3D Domain bar in the graphic display. Selecting the **CDs** link shows the CD-Search results page.
- From an Entrez Domains search: choose **Domains** from the Entrez **Search** pull-down menu and enter a search term to retrieve a list of CDs. Clicking on any resulting CD displays the CD Summary page. To find the location of this CD in an aligned protein, select the CD link following a protein name in the bottom portion of this page.
- From the CDD page: locate CDs by entering text terms into the search box and proceed as for an Entrez CD search.
- From a BLink report: select the **CDD-Search** button to display the CD-Search results page.
- From the BLAST main page: follow the RPS-BLAST link to load the CD-Search page.

Viewing Conserved Domains

Results from a CD search are displayed as colored bars underneath a sequence ruler. Moving the mouse over these bars reveals the identity of each domain; domains are also listed in a format similar to BLAST summary output ([Chapter 16](#)). Pairwise alignments between the matched region of the target protein and the representative sequence of each domain are shown below the bar. Red letters indicate residues identical to those in the representative sequence, whereas blue letters indicate residues with a positive BLOSUM62 score in the BLAST alignment.

Viewing Multiple Alignments of a Query Protein with Members of a Conserved Domain

These can be displayed by clicking a CD bar within a MMDB Structure Summary page or from a hyperlinked CD name on a CD-Search results page.

Viewing CD Alignments in the Context of 3D Structure

If members of a CD have MMDB records, one of these records can be viewed as a 3D image along with the sequence alignment using Cn3D (launched by selecting the pink dot on a CD-Search results page). As in other alignment views, colored capital letters indicate aligned residues, allowing the sequence of the protein sequence of interest to be mapped onto the available 3D structure.

Finding and Viewing Proteins with Similar Domain Architectures

For an example query on finding and viewing proteins with similar domain architectures, see [Box 3](#).

Why Would I Want to Do This?

- To locate related functional domains in other protein families
- To gain insights into how a given CD is situated within a protein relative to other CDs
- To explore functional links between different CDs
- To predict the function of a protein whose function is unknown
- To establish evolutionary relationships across protein families

How to Begin

Following the **Domain Relatives** link for any protein in Entrez, one can find other proteins with similar domain architecture. The Conserved Domain Architecture Retrieval Tool ([CDART](#)) can take an Accession number or the FASTA sequence as a query to find out the domain architecture of a protein sequence and list other proteins with related domain architectures.

Alternative Points of Entry

- From a CD-Search results page, click **Show Domain Relatives**
- From a CD-Summary page, click the **Proteins** link
- From an Entrez Domains search, click the **Proteins** link in the Links menu

Results of a CDART Search

These are described in Figure 5. The protein “hits”, which have similar domain architectures to the query sequence, can be further refined by taxonomic group, in which the results can be limited to selected nodes of the taxonomic tree. Furthermore, search results may be limited to those that contain only particular conserved domains.

Links Between Structure and Other Resources

Integration with Other NCBI Resources

As illustrated in the sections above, there are numerous connections between the Structure resources and other databases and tools available at the NCBI. What follows is a listing of major tools that support connections.

Entrez

Because Entrez is an integrated database system ([Chapter 15](#)), the links attached to each structure give immediate access to PubMed, Protein, Nucleotide, 3D Domain, CDD, or Taxonomy records.

BLAST

Although the BLAST service is designed to find matches based solely on sequence, the sequences of Structure records are included in the BLAST databases, and by selecting the PDB search database, BLAST searches only the protein sequences provided by MMDB records. A new **Related Structure** link provides 3D views for sequences with structure data identified in a BLAST search.

BLink

The BLink report represents a precomputed list of similar proteins for many proteins (see, for example, links from Entrez Gene records; [Chapter 19](#)). The **3D Structures** option on any BLink report shows the BLAST hits that have 3D structure data in MMDB, whereas the **CDD-Search** button displays the CD-Search results page for the query protein.

Microbial Genomes

A particularly useful interface with the structural databases is provided on the [Microbial Genomes page](#) (10). To the left of the list of genomes are several hyperlinks, two of which offer users direct access to structural information. The red **[D]** link displays a listing of every protein in the genome, each with a link to a BLink page showing the results of a

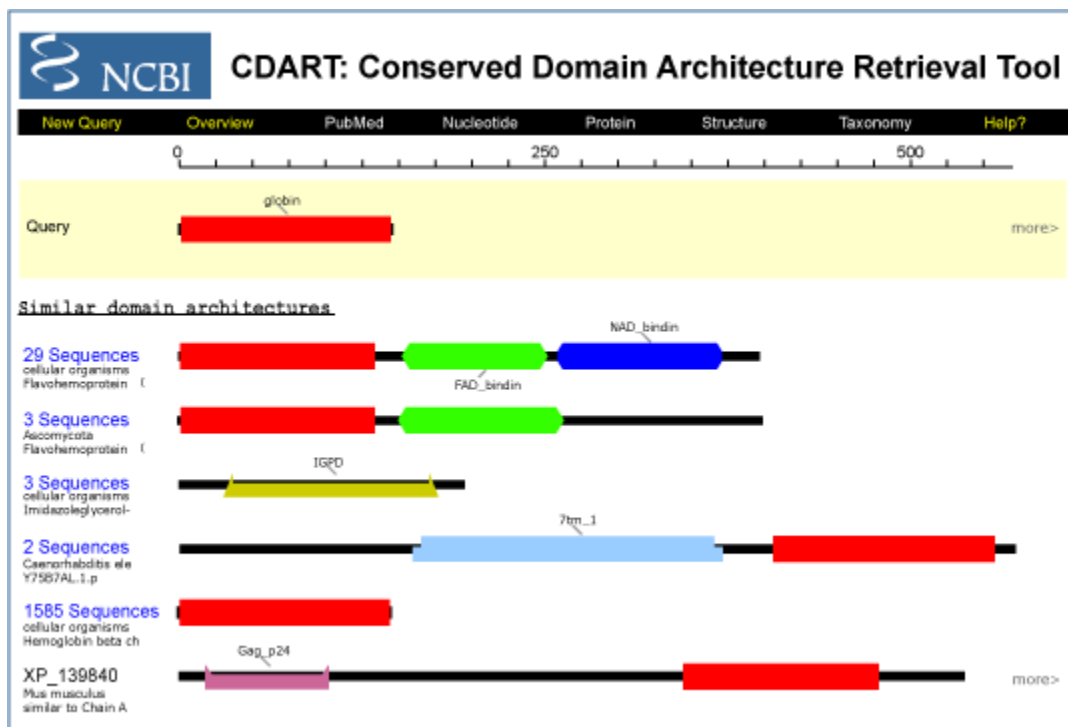


Figure 5. A CDART results page. At the *top* of the CDART results page in a *yellow box*, the query sequence CDs are represented as “beads on a string”. Each CD had a unique color and shape and is labeled both in the display itself and in a legend located at the *bottom* of the page. The shapes representing CDs are hyperlinked to the corresponding CD summary page. The matching proteins to the query are listed *below* the yellow box, ranked according to the number of non-redundant hits to the domains in the query sequence. Each match is either a single protein, in which case its Accession number is shown, or is a cluster of very similar proteins, in which case the number of members in the cluster is shown. Cluster members can be displayed by selecting the logo to the *left* of its diagram. Selecting any protein Accession number displays the flatfile for that protein. To the *right* of any drawing for a single protein (either on the main results page or after expanding a protein cluster) is a **more>** link, which displays the CD-Search results page for the selected protein so that the sequence alignment, e.g., of a CDART hit with a CD contained in the original protein of interest, can be examined.

BLAST pdb search for that protein. The [S] link displays a similar protein list for the selected genome, but now with a listing of the conserved domains found in each protein by a CD-Search.

Links to Non-NCBI Resources

The Protein Data Bank (PDB)

As stated elsewhere, all records in the MMDB are obtained originally from the Protein Data Bank (PDB) (6). Links to the original PDB records are located on the Structure Summary page of each MMDB record. Updates of the MMDB with new PDB records occur once a month.

Pfam and SMART

The CDD staff imports CD collections from both the Pfam and SMART databases. Links to the original records in these databases are located on the appropriate CD Summary page. Both Pfam and SMART are updated several times per year in roughly bimonthly intervals, and the CDD staff update CDD accordingly.

Saving Output from Database Searches

Exporting Graphics Files from Cn3D

Structures displayed in Cn3D can be exported as a Portable Network Graphics (PNG) file from within Cn3D (the Export PNG command in the **File** menu). The structure file itself, in the orientation currently being viewed, can also be saved for later launching in Cn3D.

Saving Individual MMDB Records

Individual MMDB records can be saved/downloaded to a local computer directly from the Structure Summary page for that record. **Save File** in the **View** bar downloads the file in a choice of three formats: ASN.1 (select **Cn3D**); PDB (select **RasMol**); or Mage (select **Mage**).

Saving VAST Alignments

Alignments of VAST neighbors can be saved/downloaded from the VAST Structure Neighbors page of any MMDB record. By selecting options in the **View Alignment** pull-down menu, the alignment data can be saved, formatted as HTML, text, or mFASTA, and then saved.

FTP

MMDB

Users can download the NCBI Structure databases from the NCBI FTP site: <ftp://ftp.ncbi.nih.gov/mmdb>. A Readme file contains descriptions of the contents and information about recent updates. Within the mmdb directory are four subdirectories that contain the following data:

- mmdbdata: the current MMDB database (NOTE: these files can not be read directly by Cn3D).
- vastdata: the current set of VAST neighbor annotations to MMDB records
- nrtable: the current non-redundant PDB database
- pdbeast: table listing the taxonomic classification of MMDB records

CDD

CDD data can be downloaded from <ftp://ftp.ncbi.nih.gov/pub/mmdb/cdd>. A Readme file contains descriptions of the data archives. Users can download the PSSMs for each CD record, the sequence alignments in mFASTA format, or a text file containing the accessions and descriptions of all CD records.

Frequently Asked Questions

- Cn3D
- VAST searches
- CDD

References

1. Wang Y, Anderson JB, Chen J, Geer LY, He S, Hurwitz DI, Liebert CA, Madej T, Marchler GH, Marchler-Bauer A, Panchenko AR, Shoemaker BA, Song JS, Thiessen PA, Yamashita RA, Bryant SH. MMDB: Entrez's 3D-structure database. *Nucleic Acids Res.* 2002;30:249–252. PubMed PMID: 11752307.
2. Wang Y, Geer LY, Chappay C, Kans JA, Bryant SH. Cn3D: sequence and structure views for Entrez. *Trends Biochem Sci.* 2000;25:300–302. PubMed PMID: 10838572.
3. Madej T, Gibrat J-F, Bryant SH. Threading a database of protein cores. *Proteins.* 1995;23:356–369. PubMed PMID: 8710828.
4. Gibrat J-F, Madej T, Bryant SH. Surprising similarities in structure comparison. *Curr Opin Struct Biol.* 1996;6:377–385. PubMed PMID: 8804824.
5. Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY, Bryant SH. CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res.* 2002;30:281–283. PubMed PMID: 11752315.
6. Westbrook J, Feng Z, Jain S, Bhat TN, Thanki N, Ravichandran V, Gilliland GL, Bluhm W, Weissig H, Greer DS, Bourne PE, Berman HM. The Protein Data Bank: unifying the archive. *Nucleic Acids Res.* 2002;30:245–248. PubMed PMID: 11752306.
7. Ohkawa H, Ostell J, Bryant S. MMDB: an ASN.1 specification for macromolecular structure. *Proc Int Conf Intell Syst Mol Biol.* 1995;3:259–267. PubMed PMID: 7584445.
8. Bateman A, Birney E, Cerruti L, Durbin R, Ewlinger L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer ELL. The Pfam proteins family database. *Nucleic Acids Res.* 2002;30:276–280. PubMed PMID: 11752314.
9. Letunic I, Goodstadt L, Dickens NJ, Doerks T, Schultz J, Mott R, Ciccarelli F, Copley RR, Ponting CP, Bork P. SMART: a Web-based tool for the study of genetically mobile domains. Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Res.* 2002;30:242–244. PubMed PMID: 11752305.
10. Wang Y, Bryant S, Tatusov R, Tatusova T. Links from genome proteins to known 3D structures. *Genome Res.* 2000;10:1643–1647. PubMed PMID: 11042161.