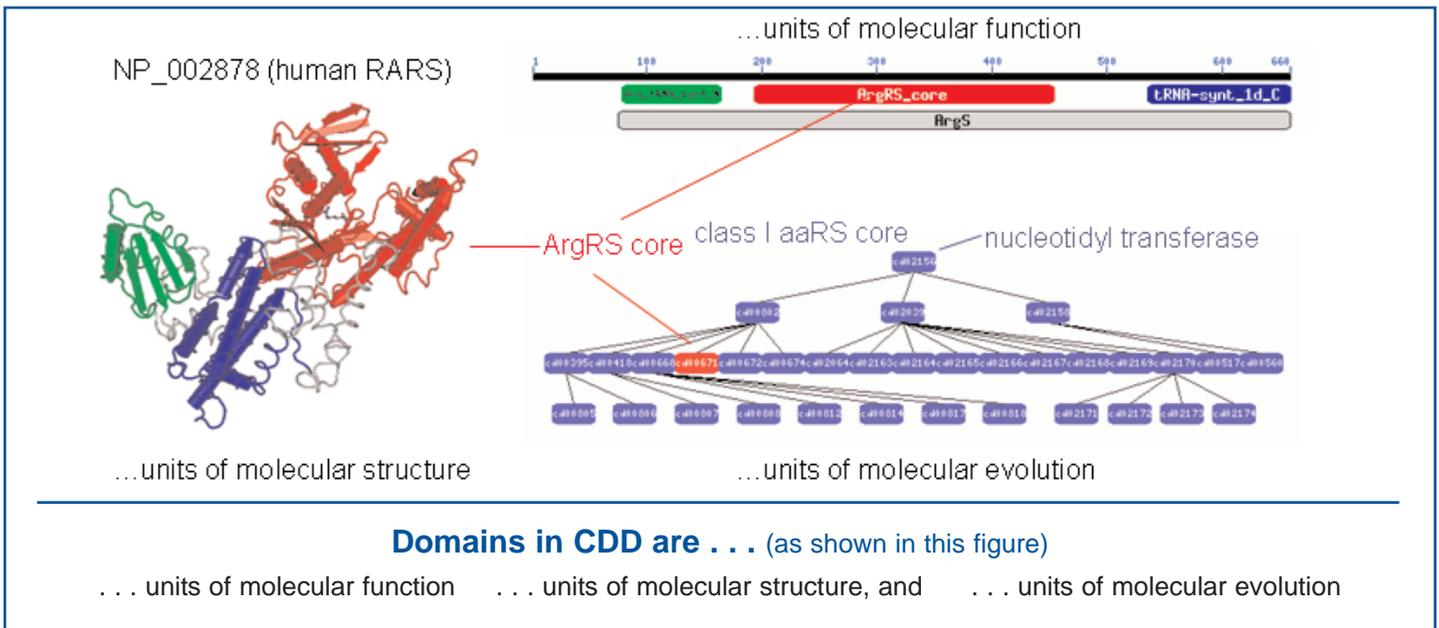




NCBI Conserved Domain Database (CDD)

National Center for Biotechnology Information ■ National Library of Medicine ■ National Institutes of Health ■ Department of Health and Human Services

The NCBI Conserved Domain Database (CDD) is a protein domain annotation resource, and includes a collection of ancient (> ~0.5 Byr old) protein domains and hierarchies of domains related by common descent. Each CD record consists of a protein multiple sequence alignment, a consensus sequence, and a PSSM (position-specific score matrix) that quantitatively represents the information in the alignment. Whenever possible, 3D structural information is used to define and refine the alignment models. CDD is supported by an active curation effort that identifies new domains, builds domain hierarchies, and continually updates existing records with new sequences. CDD is part of the NCBI Entrez system, and thus is extensively linked with other NCBI data. CDs can be found by direct text searching in Entrez, by Entrez links from any protein sequence, or by RPS-BLAST. Each of the CD PSSMs is now represented in a "scoremat" format, and these scoremats can be used to build custom domain databases for RPS-BLAST searches.



Database Content — CDD v2.11

Database	Accession	No. of records	Source URL
<i>Curated data</i>			
NCBI CD	cd01234	2908 domains	www.ncbi.nlm.nih.gov/Structure/cdd/cdd/shtml
<i>Non-curated data</i>			
Pfam v22.0	pfam01234	9318 domains	www.sanger.ac.uk/Software/Pfam/
SMART v4.0	smart0123	663 domains	smart.embl-heidelberg.de/
COG	cog0123	4873 domains	www.ncbi.nlm.nih.gov/COG/

Elements of a CD Record

Sequence Alignment

Each aligned sequence is matched to a record in Entrez Protein, and if 3D structural data are available for a sequence, then the chosen protein sequence will be one from PDB. If sequences with structural data exist, one of them will be chosen as the master sequence of the alignment.

Consensus sequence

Each position in the consensus sequence contains the residue with the highest weighted frequency in that column of the alignment. For a column to be included in the consensus sequence, at least 50% of the sequences must have an aligned residue in that column.

PSSM

For each position in the consensus, frequency ratios (expected/observed) are calculated for each amino acid, and these frequencies are converted to scores in a PSSM. The PSSM thus has the same number of columns as the consensus and precisely defines the extent of the domain.

Finding CDs

Text Searching in Entrez

To find . . .

serine kinases
Curated domains

Domains containing only proteins from archaea

use this query . . .

serine kinase
cdd[database]
archaea[orgn]

Linking in Entrez

To find . . .

CDs in a protein sequence
CDs encoded by a gene

CDs that occur with a given CD in a single protein

use this link . . .

Conserved Domain link in Protein
Conserved Domains link in Gene
Co-occurring Domain

RPS-BLAST

Web: www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi

Standalone: ftp://ftp.ncbi.nlm.nih.gov/blast/

CDD Software Tools

CDTree — analyzes sequence alignments, constructs phylogenetic trees, and constructs CD hierarchies; reads/writes CDs

www.ncbi.nlm.nih.gov/Structure/cdtree/cdtree.shtml

Cn3D 4.1 — renders and aligns structures; creates and edits sequence alignments; reads/writes CDs; reads mFASTA, writes gapped FASTA; v4.2 will write PSSMs as NCBI scoremats

www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml

formatrpsdb — command-line utility (part of the BLAST package) that converts scoremats into rpsblast databases

ftp.ncbi.nih.gov/blast/executables/LATEST/

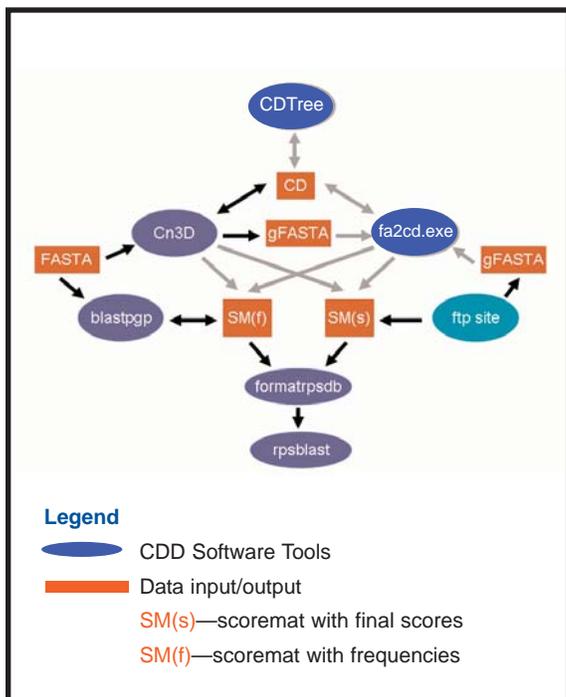
rpsblast — command-line program that searches query protein sequences against a database of PSSMs produced by formatrpsdb (part of the BLAST package)

ftp.ncbi.nih.gov/blast/executables/LATEST/

blastpgp — (PSI-BLAST) command-line program that creates PSSMs iteratively from proteins that are sequence-similar to a query protein; writes frequency scoremats only; can initiate a search with a frequency scoremat

ftp.ncbi.nih.gov/blast/executables/LATEST/

fa2cd.exe — a command-line utility that converts gapped FASTA into a CD file readable by CDTree and Cn3D.



CDD Summary and Annotation Files Available by FTP

ftp.ncbi.nih.gov/pub/mmdb/cdd

Full CD — *.acd

These files contain all data for each CD record. These files can be read and written by both Cn3D and CDTree. — [acd.tar.gz](http://ftp.ncbi.nih.gov/pub/mmdb/cdd/acd.tar.gz)

CD scoremats (SM(s)) — *.smp

These files are ASN.1 representations of the PSSMs containing only final scores (not frequencies), and can be used as input to formatrpsdb. They cannot be used to initiate a PSI-BLAST search. — [cdd.tar.gz](http://ftp.ncbi.nih.gov/pub/mmdb/cdd/smp.tar.gz)

FASTA alignments — *.FASTA

Each of these files contains the entire CD alignment, including the consensus sequence, in gapped FASTA format. These files are suitable for importing CD alignments into sequence analysis software. — [fasta.tar.gz](http://ftp.ncbi.nih.gov/pub/mmdb/cdd/FASTA.tar.gz)

CD Summaries

This file contains the PSSM-Id, CD accession, title, description, and PSSM length (number of columns) of each CD. — [cddid.tbl.gz](http://ftp.ncbi.nih.gov/pub/mmdb/cdd/cddid.tbl.gz)

Curated Annotations

This file is an index of all annotations present on NCBI curated CDs. — [cddannot.dat.gz](http://ftp.ncbi.nih.gov/pub/mmdb/cdd/cddannot.dat.gz)

Master Sequences

This file contains the master sequence of each CD in FASTA format. — [cddmasters.fa.gz](http://ftp.ncbi.nih.gov/pub/mmdb/cdd/cddmasters.fa.gz)

CDD Versions

This file contains accessions, PSSM-IDs, names, and create dates of all current and previous versions of CDD records.

Web Tools Related to CDD

CDD Home

www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml

CDART

www.ncbi.nlm.nih.gov/Structure/lexington/lexington.cgi

Web RPS-BLAST

www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi

PSSM Viewer

www.ncbi.nlm.nih.gov/Class/Structure/pssm/pssm_viewer.cgi

NCBI Handbook

www.ncbi.nlm.nih.gov/books/bv.fcgi?call=bv.View..ShowSection&rid=handbook.section.110

CDD Help

www.ncbi.nlm.nih.gov/Structure/cdd/cdd_help.shtml