New Developments within the NCBI Conserved Domain Database: Functional Annotation, Phylogenetics, and Structure Prediction

Eric W. Sayers, PhD
National Center for Biotechnology Information


New Developments in CDD

• Content of CDD
• Understanding PSSMs
• Processing and curating alignment data
• Web tools for finding CDs in a protein sequence
• New tools for creating and refining PSSMs
  – Cn3D 4.2
  – CDTree

Overview of CDD

• CDD is a collection of protein domains and hierarchies of domains related by common descent
• CDD alignments are quantitatively represented by a PSSM (Position-Specific Score Matrix)
• Whenever possible, 3D structural data is used to define and refine alignment models
• CDD is supported by an active curation process that identifies new domains, builds domain hierarchies, and continually updates existing domains
• CDD is part of the NCBI Entrez system, and is linked to many NCBI resources
• CDD can be searched via the web, via a URL-API, or by standalone tools against custom local databases
What is a Domain?

• Domains in CDD represent
  – units of molecular evolution
  – units of molecular function
  – units of molecular 3D structure

• CDD records contain
  – a consensus sequence defining the length of the domain
  – a master sequence with 3D structure whenever possible
  – a multiple sequence alignment of portions of representative proteins that contain the domain
  – a PSSM encoding the alignment data

CDD Contains Ancient Domains

CDs represent domains that originated > ~0.5 billion years ago

CDD: Units of Molecular Evolution

CDs represent domains that originated > ~0.5 billion years ago

Lineage specific: Proteins from only 1 ancient node

Ancient Diversity: Proteins from >1 ancient node
Preferred Taxonomic Nodes

- Viruses – 35 nodes
  - Metazoa
  - Plants
  - Fungi
  - Other eukaryotes
  - Eubacteria
  - Archaea

- Metazoa – 12 nodes
- Plants – 5 nodes
- Fungi – 6 nodes
- Other eukaryotes – 11 nodes
- Eubacteria – 20 nodes
- Archaea – 13 nodes
Two Example Domains

cd02249 – ZZ zinc finger

46 residues long; found in over 950 proteins, mainly in eukaryotes

The ZZ zinc finger motif coordinates one or two zinc ions and most likely participates in ligand binding or molecular scaffolding.

cd00671 – core catalytic domain of arginyl tRNA synthetase

267 residues long; found in over 1600 proteins across all cellular organisms

The core domain is responsible for the ATP-dependent formation of the aminoacyl adenylate, which is then transferred to the 3' end of the tRNA.

CDD: Units of Molecular Evolution

Nucleotidyl transferase superfamily, containing amino acyl tRNA synthetases

CDD: Units of Molecular Function

Arginyl-tRNA synthetase [Homo sapiens]
The consensus sequence defines the positions of the PSSM, and thus the exact extent of the domain.

Understanding PSSMs

- PSSMs quantify the biological data in an alignment by combining the observed residue substitution frequencies at each position with general substitution frequencies observed in conserved proteins
- PSSMs are score matrices created by PSI-BLAST and searched by RPS-BLAST
- Scores are functions of logarithms of ratios of observed substitution frequencies to expected frequencies

\[
s \propto \ln \left( \frac{f_{\text{observed}}}{f_{\text{expected}}} \right)
\]

\[
s > 0 \quad \text{Positive for more likely substitutions}
\]

\[
s < 0 \quad \text{Negative for less likely substitutions}
\]

BLOSUM62: The Foundation

- True amino acids have low weights
- Rare amino acids have high weights
- BLOSUM scores are the same at every position
**PSI-BLAST**

Creating a PSSM:
1. Run BLASTp using BLOSUM62
2. Generate a PSSM from the resulting alignments
3. Run BLAST again using the new PSSM as the score matrix
4. Go to step 2

**Reading PSSMs**

- Displays a PSSM with scores, and sorts the matrix by scores for a given residue
- Displays a PSSM where the residues at each position are sorted by score
- Subset or highlight by residue name or category
- Shows positions on both the consensus and master sequences of the CD

**PSSM Viewer**

PSSM Viewing Modes

View by Residue
Matrix sorted by Cys

View by Position
Matrix showing only the positions where the consensus is Cys

Analyzing PSSM Positions

Residue Frequencies
- cd02249: Position 3

Residue Frequencies
- cd02249: Position 15

cd02249: Position 15
Analyzing PSSM Positions

Interpreting Mutations with PSSMs

Allelic Variants in F7
Locating the Mutation Sites

<table>
<thead>
<tr>
<th>NP: Pos 307; Master: Pos 153</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP: Pos 370; Master: Pos 222</td>
</tr>
</tbody>
</table>

Mutation Sites in the PSSM

<table>
<thead>
<tr>
<th>.002 Factor IV Deficiency FT: CYS175F</th>
</tr>
</thead>
<tbody>
<tr>
<td>.004 Factor IV Deficiency FT: ARG257F</td>
</tr>
</tbody>
</table>

The NCBI Scoremat

- ASN.1 representation of a single PSSM
- Two formats
  - SM(f): contains only substitution frequencies
  - SM(s): contains final substitution scores
- Produced by PSI-BLAST and Cn3D 4.2
- Available by ftp for all CDs
- Used to build RPS-BLAST databases
Inside a Scoremat

```

```

CDD: Sources of Data

**CDD v2.06**: A database of Position Specific Score Matrices (PSSMs)

<table>
<thead>
<tr>
<th>Single Domains</th>
<th>Protein Families</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pfam</strong></td>
<td><strong>NCBI</strong></td>
</tr>
<tr>
<td>pfam01234 5252 (45%)</td>
<td>COG 4101 (36%)</td>
</tr>
<tr>
<td>SMART</td>
<td></td>
</tr>
<tr>
<td>smad0123 575 (5%)</td>
<td></td>
</tr>
<tr>
<td>CD</td>
<td></td>
</tr>
<tr>
<td>cd01234 1602 (14%)</td>
<td></td>
</tr>
<tr>
<td>COG</td>
<td></td>
</tr>
<tr>
<td>COG0123</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Pfam-A seeds: HMM based models representing a wide variety of functional domains derived from SWISS-PROT

HMM based models originally concentrating on eukaryotic signaling domains, now expanding

BLAST based alignments derived from complete proteomes of prokaryotes

CDD: Initial Data Processing

1. Import alignments and identify proteins
   - Import seed alignment from Pfam, SMART, or COG
   - Associate each sequence with an NCBI protein GI
2. Add sequences with 3D structural data
   - If any sequence is sequence-similar to a sequence with 3D structural data, the original sequence is replaced with the one having 3D data
3. Create the consensus sequence
   - A column will be represented in the consensus if >50% of the aligned sequences have a residue in that column
   - The consensus residue for the column is that with the largest weighted frequency
4. Calculate the PSSM
Uncurated CD Summary

- pfas05660.11
- ZZ
- Lowest taxa containing all sequences
- All proteins that contain this domain
- 8 sequences
- No structures

CDD Curation

- Sequences that align to the PSSM and that have 3D structural data are added to a CD using PSI-BLAST and VAST
- A new alignment is built from representative proteins based on the structural features revealed by VAST and evidence in the literature
- Functional features reported in the literature are annotated on the alignment and structures
- New “child” alignments are generated from the original alignment based on phylogenetic analysis
- Existing alignments are updated with new sequences using PSI-BLAST
- Curation is performed by experts at NCBI using Cn3D and CDTree

VAST

Vector Alignment Search Tool

For each 3D domain, locate SSEs (secondary structure elements), and represent them as individual vectors.

Human IL-4
VAST: Refinement

- Aligned residues are red
- Cα atoms are added to the aligned SSEs
- Alignments are allowed to extend beyond SSE boundaries
- All atoms are added to the models, and the detailed backbone and sidechain positions are refined

VAST: Alignment of Sequence

- Aligned blocks represent structural core elements
- Aligned blocks have no internal gaps
- Aligned residues occupy the same position in space
- Aligned residues are shown in CAPITAL letters

Helix 1
Helix 2
Helix 3
Helix 4

The CDD Alignment Model

- Each sequence is aligned pairwise to the master sequence
- Aligned blocks represent secondary structure elements
- Aligned blocks have no internal gaps
- Aligned sequences have a residue at each column in the block
- Residues in the same column occupy the same position in space
### Curating CDs with VAST

<table>
<thead>
<tr>
<th>CD-ID</th>
<th>VAST</th>
<th>Cn3D</th>
</tr>
</thead>
<tbody>
<tr>
<td>smart00235</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cd00203</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### CD-Curation: Effect on model alignment accuracy

- **VAST**
- **RPS-BLAST before curation**
- **RPS-BLAST after curation**

### An Example: pfam00569

<table>
<thead>
<tr>
<th>Pfam-ID</th>
<th>Description</th>
<th>Accession</th>
<th>Score</th>
<th>$E$-Value</th>
<th>Identity</th>
<th>Coverage</th>
<th>Domain ID</th>
<th>Domain Name</th>
<th>Classification</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>pfam00569</td>
<td>ZZ</td>
<td>1TOT_A</td>
<td>1000</td>
<td>1e-10</td>
<td>90%</td>
<td>80%</td>
<td>1TOT_A</td>
<td>ZZ</td>
<td>Structural</td>
<td>Structural</td>
</tr>
</tbody>
</table>

- All proteins that contain this domain.
- 874 proteins
- 1 structure: 1TOT_A
VAST Neighbors of 1TOT

VAST finds 57 similar structures, but none align to the PSSM.

CDD Curation

- 1TOTA becomes the master sequence
- Pfam, literature and structure reveal two zinc binding sites

Tools for Curation

- Cn3D 4.2
  - Displays and renders structures and structural alignments
  - Imports and aligns sequences and structures manually or with BLAST and threading algorithms
  - Edits and analyzes CD records in conjunction with CDTree
  - Writes gapped FASTA and PSSMs
- CDTree
  - Reads/writes CD records and CD projects
  - Creates phylogenetic and taxonomic trees of CD sequences
  - Creates and edits CD hierarchies
  - Updates CDs using PSI-BLAST
  - Launches Cn3D sessions where alignments can be edited
  - Allows remastering of CD alignments

See demos at booth 521!
Curated CD: cd02249

CD Hierarchies

• Residues aligned in the parent must be aligned in the child
• The parent always contains the master sequence of the child

CD Family Values

Blocks in the child are often larger than those in the parent

Parent: cd00802, Catalytic core of class I amino acyl-tRNA synthetases

Child: cd00071, Catalytic core of arginyl tRNA synthetase
A Family of PSSMs

Parent: cd00802

Children are more opinionated than their parents!

Child: cd00671

Viewing Hierarchies on the Web

Displays the phylogenetic tree

Downloads the hierarchy into CDTree

CDD Page Links

Accession/Version: cd02249.2

Links:
- Original alignment
- PubMed abstracts describing this domain
- CDART

Attributes:
- Name: class_2_add2_core
- Source: Pfam
- Taxonomy: Eukaryota
- PubMed: 2 links
- Protein: cd02249 related architectures
- Related CDs: 13 links

HMM:
- PSSM: 16
- Number of aligned sequences: 30293
- Aligned: 26 rows
- Status: curated CD
- Created: 12 Oct 2003
- Updated: 29 Mar 2005
CDART: Conserved Domain Architecture Retrieval Tool

CDART - Architectures

CDART finds groups of proteins that share a set of CDs.

CDART - Subset Functions

CDART - Architecture Cluster

Retrieve sequences from Entrez

RPS-BLAST results
Searching CDD by Text

Most useful for fielded queries

- Use the quantitative power of the PSSM rather than arbitrary text annotations
- Annotates function and reveals sites of high sequence conservation
- Provides a 3D modeling template if the PSSM is curated

Limitations of Text Searches

- zinc finger → 104 records
- zz zinc finger → 14 records
- arginine tRNA → 0 records
- arginyl tRNA → 5 records

Searching with PSSMs
Searching for PSSMs

RPS-BLAST (CD-Search)

- What does it do?
  - Compares a protein sequence to a database of PSSMs
  - Alignment is scored by the matching PSSM

- Why use it?
  - To annotate protein function
  - To identify important sequence motifs
  - To derive a structural model
  - To find homologs

<table>
<thead>
<tr>
<th>E</th>
<th>C</th>
<th>D</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-2</td>
<td>-2</td>
<td>-1</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
<td>10</td>
<td>-3</td>
</tr>
<tr>
<td>D</td>
<td>4</td>
<td>-5</td>
<td>-2</td>
</tr>
<tr>
<td>E</td>
<td>2</td>
<td>-4</td>
<td>2</td>
</tr>
<tr>
<td>F</td>
<td>-1</td>
<td>-3</td>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>E</th>
<th>C</th>
<th>D</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-2</td>
<td>-2</td>
<td>-1</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
<td>10</td>
<td>-3</td>
</tr>
<tr>
<td>D</td>
<td>4</td>
<td>-5</td>
<td>-2</td>
</tr>
<tr>
<td>E</td>
<td>2</td>
<td>-4</td>
<td>2</td>
</tr>
<tr>
<td>F</td>
<td>-1</td>
<td>-3</td>
<td>5</td>
</tr>
</tbody>
</table>

Structures in CDs

CDD v2.06

Single Domains

<table>
<thead>
<tr>
<th>Pfam</th>
<th>SMART</th>
<th>CD</th>
<th>COG</th>
</tr>
</thead>
<tbody>
<tr>
<td>pfam01234</td>
<td>smart0123</td>
<td>cd01234</td>
<td>COG0123</td>
</tr>
</tbody>
</table>

- 29.0% have structures (1523 of 5252)
- 60.3% have structures (347 of 575)
- 62.2% have structures directly (697 of 1602)
- 100% have structures in their parent (or self)
- 9.6% have structures (392 of 4101)

Protein Families

<table>
<thead>
<tr>
<th>Pfam</th>
<th>SMART</th>
<th>CD</th>
<th>COG</th>
</tr>
</thead>
<tbody>
<tr>
<td>pfam01234</td>
<td>smart0123</td>
<td>cd01234</td>
<td>COG0123</td>
</tr>
</tbody>
</table>

- 69.4% hit CDD
- 83.0% BLAST ➔ MMDB
- 45.2% hit both
- 27.8% hit neither

Available Sequence Homology

<table>
<thead>
<tr>
<th>Hit CDD</th>
<th>BLAST ➔ MMDB</th>
</tr>
</thead>
<tbody>
<tr>
<td>8184</td>
<td></td>
</tr>
</tbody>
</table>

Human RefSeq proteins (29490)

69.4% hit CDD
83.0% BLAST ➔ MMDB
45.2% hit both
27.8% hit neither
Precomputed RPS-BLAST

From Entrez Gene, follow the “Conserved Domains” link.

Provides a list of matching CDs in Entrez.

Provides full graphical and statistical output.

Precomputed RPS-BLAST

From Entrez Protein, look for the “Conserved Domains” link.

Web CD-Search Output

Click on a colored bar to align the query sequence to the CD.
### Viewing a Structural Model

![Graphic of a structural model](image)

### Web RPS-BLAST: Databases

<table>
<thead>
<tr>
<th>NCBI Conserved Domain Search</th>
<th>CDD v2.25</th>
</tr>
</thead>
<tbody>
<tr>
<td>Database</td>
<td>Records</td>
</tr>
<tr>
<td>CD</td>
<td>1682</td>
</tr>
<tr>
<td>Pfam</td>
<td>38252</td>
</tr>
<tr>
<td>SMART</td>
<td>375</td>
</tr>
<tr>
<td>COG</td>
<td>4151</td>
</tr>
<tr>
<td>KOG</td>
<td>0</td>
</tr>
</tbody>
</table>

CDD: All curated CDs + Pfam, SMART, and COG alignments that meet taxonomic requirements and are not redundant with other alignments

Entrez Conserved Domains is equivalent to the most recent CDD release

### Web RPS-BLAST: Options

- **E-value maximum:** Can be adjusted to detect weaker alignments, particularly for small domains
- **Low-complexity filter:** Can be turned off for queries containing low-complexity regions

![Options menu](image)

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Display</td>
<td>E-value, X-score, Structural identity, Oi, or Segment as a TFASTA format</td>
</tr>
<tr>
<td>Output format</td>
<td>TFASTA, FASTA, or EIALPHA format</td>
</tr>
</tbody>
</table>

Entrez protein CD: Shares in protein
### Standalone RPS-BLAST

**rpsblast**

```
```

**rpsblast** `-i` **infile** `-d` **database** `-o` **outfile**

- Why run RPS-BLAST on your own machine?
  - ultimate flexibility
  - customized databases
  - batch processing
  - parsing output
  - speed
  - it's now much easier than it might sound!

### Making Custom Databases

**formatrpsdb**

```
```

`formatrpsdb` `-i` **infile** `-n` **dbname** `-T` `-f` **9.82** `-S` **100**

**database**

**name**

**scaling**

**parameters**

- `cd02249.smp`
- `cd02335.smp`
- `cd02336.smp`
- `cd02337.smp`
- `mypssm1.smp`
- `mypssm2.smp`
- `testcd.smp`

- text file containing scoremat file names

- Standard CDD database - 11530 PSSMs required to format: ~2.8 GB
  - final size: ~620 MB

### Accessing CDD Data Files

**Archive Collections**

- `acd.tar.gz` - *.acd files – full CD records
  - read by Cn3D and CDTree
  - ~4.2 GB uncompressed
- `cdd.tar.gz` - *.smp files – NCBI scoremats (PSSMs)
  - read by formatrpsdb
  - ~2.2 GB uncompressed
- `fasta.tar.gz` - *.FASTA files – gapped FASTA for each CD
  - read by third-party alignment software
  - ~250 MB uncompressed

---

Generating Custom Scoremats

1. **FASTA** → **PSI-BLAST** → **SM**
   - Start with a sequence or a scoremat, collect matching sequences from a database, then generate an alignment from these sequences.

2. **FASTA** → **Cn3D 4.2** → **SM**
   - Start with unaligned sequences, a VAST alignment or a CD and build a block alignment based on evidence from 3D structures.

3. **gFASTA** → **FA2CD’** → **SM**
   - Convert an existing alignment in gapped FASTA format to a scoremat.

---

PSI-BLAST: Making a PSSM

1. **blastpgp**
   - blastpgp -i infile -d database -o outfile -j 3 -u 1 -C pssm1.smp -J T
     - Sets checkpoint output to ASCII scoremat.
     - File name for ASCII scoremat.
     - Number of PSI-BLAST iterations.

   **database** — standard BLASTp database generated by formatdb.

---

PSI-BLAST: Searching with a PSSM

1. **blastpgp**
   - blastpgp -d database -o outfile -j 3 -u 1 -C pssm2.smp -J T -q 1 -R pssm1.smp

   **database** — standard BLASTp database generated by formatdb.
Cn3D 4.2

- Launch from a CD, a VAST alignment, or a structure
- Import additional sequences and structures
- Align sequences to the master using...
  - BLAST/PSSM: BLASTp scored by the current PSSM in Cn3D
  - BLOCK Aligner: Algorithm using blocks as "words"
  - Threader: Use contact potentials and the PSSM to align
  - Manually
- Output gapped FASTA, a CD, or a scoremat

Modeling the C-terminus

'FA2CD'

- command-line conversion utility
- Input: gFASTA or a CD
- Output: CD, SM(f), SM(s)
CDTree: Refining PSSMs

- Import CDs or CD hierarchies
- Collect new sequences using PSI-BLAST
- Remaster the alignment (particular if new structures are found)
- Refine the alignment using Cn3D
- Analyze alignment using phylogenetic and taxonomic trees
- Create or extend a CD hierarchy based on the tree analysis
- Validate new child CDs using cross hits
- Output CDs or CD hierarchies

The CDD Data Tree

Come see us!

- Booth 521 – Try out CDTree and Cn3D 4.2
- Tuesday Poster 2221, Board B569 – CDD
- Wednesday Poster 2789, Board B456 - CDTree