New Developments within the NCBI Conserved Domain Database:

Functional Annotation, Phylogenetics, and Structure Prediction

Eric W. Sayers, PhD
National Center for Biotechnology Information

ftp://ftp.ncbi.nih.gov/pub/sayers/Protein.2006

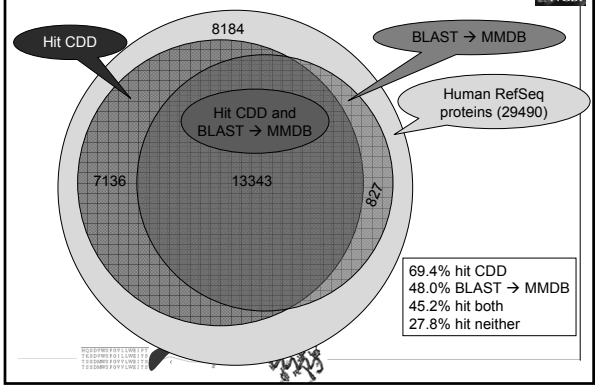
New Developments in CDD

- Overview of CDD
- Understanding PSSMs
- Processing and curating alignment data
- Improved tools for aligning PSSMs to protein sequences
- New tools for creating and refining PSSMs
  - Cn3D 4.2
  - CDTree

Overview of CDD

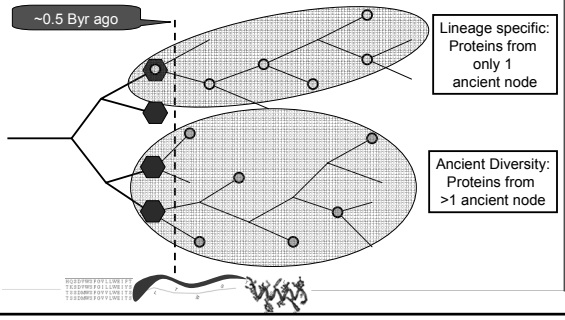
- CDD is a collection of protein domains and hierarchies of domains related by common descent
- CDD alignments are quantitatively represented by a PSSM (Position-Specific Score Matrix)
- Whenever possible, 3D structural data is used to define and refine alignment models
- CDD is supported by an active 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

Available Sequence Homology



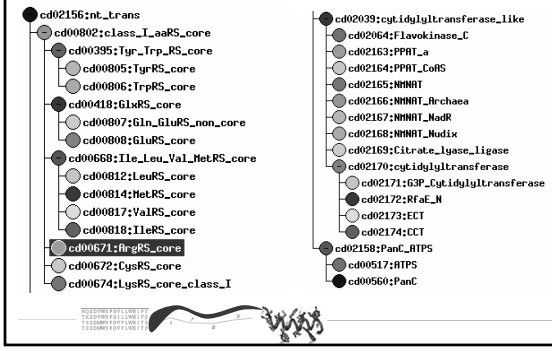
CDD: Units of Molecular Evolution

CDs represent domains that originated > ~0.5 billion years ago



CDD: Units of Molecular Evolution

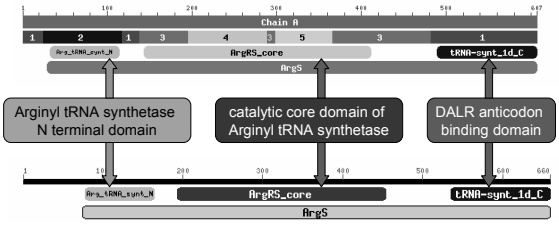
nucleotidyl transferase superfamily, containing amino acyl tRNA synthetases



CDD: Units of Molecular Structure and Function

1BS2: Yeast Arginyl-Trna Synthetase

Cn3D



arginyl-tRNA synthetase [Homo sapiens]

Understanding PSSMs

- PSSMs quantify the biological data in an alignment by combining the observed residue substitution frequencies at each position with general substitution frequences 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

$$s \propto \ln \frac{f(\text{observed})}{f(\text{expected})}$$

f(observed) > f(expected) $s > 0$

f(observed) < f(expected) $s < 0$

P	C	Master	A	G	I	L	V	M	F	W	P	C	S	T	Y	N	Q	H	K	R	D	E
1	Y	7 - Y	-2	-4	2	-1	1	1	3	-1	0	-3	0	-2	6	0	-2	-1	-2	1	-3	-3
2	S	8 - T	0	-3	2	0	0	-1	1	-3	0	1	3	3	-2	-2	-1	-2	0	1	-2	0
3	C	9 - C	-1	-4	-2	-2	-2	-2	-3	-3	-4	11	-2	-2	-3	-4	-4	-4	-4	-5	-5	-5
4	D	10 - N	0	-2	-4	-4	-4	-3	-4	-5	-3	-4	-1	0	-3	4	-1	3	0	0	6	0
5	G	11 - E	0	4	1	-1	0	-2	-3	-3	-3	4	0	0	0	1	-2	3	-2	-3	-2	0
6	C	12 - C	-1	-4	-2	-2	-2	-2	-3	-3	-4	11	-2	-2	-3	-4	-4	-4	-4	-5	-5	-5

C only

D,N,H

C only

PSSM Viewer

Reading PSSMs

If a given position…	then the PSSM column will contain…
conserves a specific residue	one large, positive score (> ~6), with all other scores negative
conserves a residue type, or a set of residues	a few moderate, positive scores (~4-7), with all other scores negative
has minimal conservation	only small scores (< ~4)

PSSM Viewer

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

- 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

Analyzing PSSM Positions

NCBI

Residue Frequencies
cd02249 - Position 3

Residue	Frequency	Score
C	1.000	11

Residue Frequencies
cd02249 - Position 15

Residue	Frequency	Score
Y	0.286	6
F	0.286	5
W	0.179	7
V	0.071	0
I	0.071	1
M	0.071	1
L	0.036	0

Mutation Sites in F7

NCBI

homozygote: ~4% F7 activity

.0002 FACTOR VII DEFICIENCY [F7, CYS310PHE]

P	C	Master	A	G	I	L	V	M	F	W	P	C	S	T	Y	N	Q	H	K	R	D	E
155	C	222 - C	-2	-4	-3	-3	-1	-3	-4	-4	-3	11	-3	-3	-4	-4	-5	-5	-3	-3	-4	-5

heterozygote: 30-37% F7 activity

.0003 FACTOR VII DEFICIENCY [F7, CYS178TYR]

P	C	Master	A	G	I	L	V	M	F	W	P	C	S	T	Y	N	Q	H	K	R	D	E
27	C	77 - C	-1	-3	-3	-2	-2	-3	-2	0	-3	11	-3	-2	-4	-5	-5	-5	-5	-5	-5	-5

homozygote: asymptomatic

.0004 FACTOR VII MIE [F7, ARG247HIS]

P	C	Master	A	G	I	L	V	M	F	W	P	C	S	T	Y	N	Q	H	K	R	D	E
95	K	153 - K	-2	-4	-3	-3	-3	-3	-4	-2	-3	-5	-1	-1	-3	-1	2	1	3	4	2	2

CDD: Sources of Data

NCBI

CDD v2.08: *A database of Position Specific Score Matrices (PSSMs)*

Single Domains

Pfam	pfam01234 5252 (43%)	Sanger	Pfam-A seeds: HMM based models representing a wide variety of functional domains derived from SWISS-PROT
SMART	smart00123 575 (5%)	EMBL	HMM based models originally concentrating on eukaryotic signaling domains, now expanding
CD	**cd01234 2219 (18%)**	NCBI	**NCBI curated domains based on sequence and structural alignments**

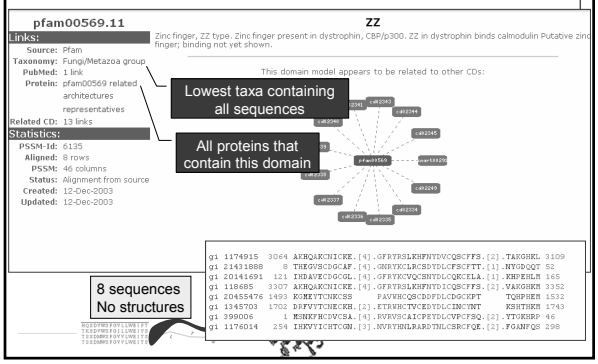
Protein Families

COG	COG0123 4101 (34%)	NCBI	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

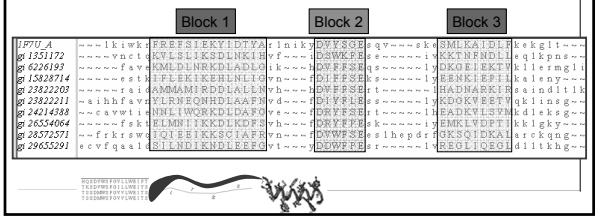


CDD Curation

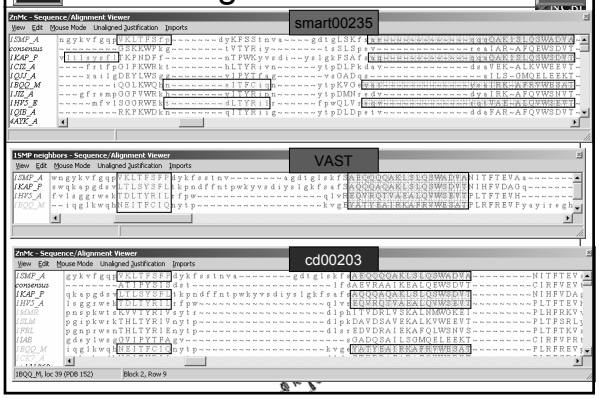
- Sequences that align to the PSSM and that have 3D structural data are added to a CD using PSI-BLAST and VAST
- A new alignment is built from representative proteins based on the structural features revealed by VAST and evidence in the literature
- Functional features reported in the literature are 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

The CDD Alignment Model

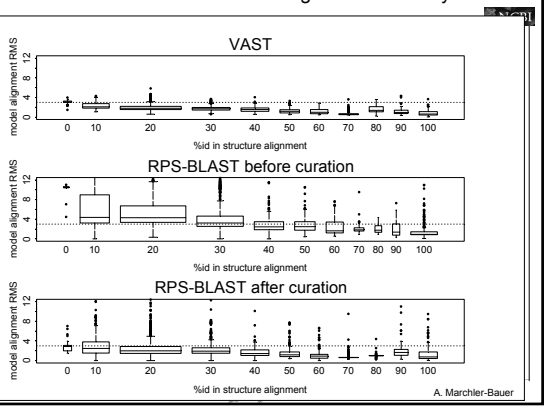
- Each sequence is aligned pairwise to the master sequence
- Aligned blocks represent secondary structure elements
- Aligned blocks have no internal gaps
- Aligned sequences have a residue at each column in the block
- Residues in the same column occupy the same position in space



Curating CDs with VAST



CD-Curation: Effect on model alignment accuracy



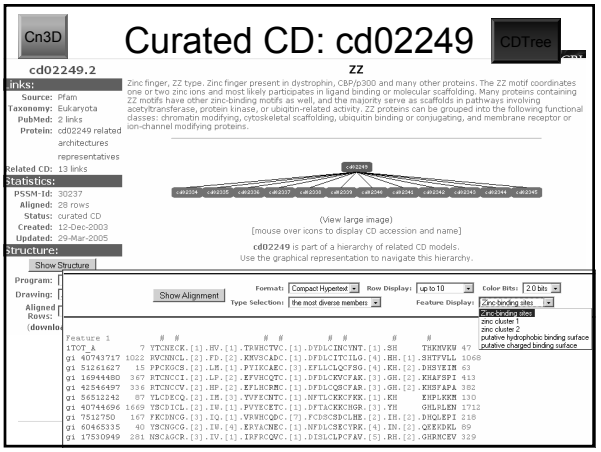
Tools for Curation

See demos at booth 521!

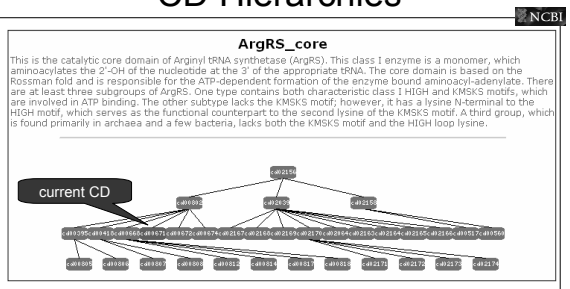
ftp://ftp.ncbi.nih.gov/pub/mmdb/cdtree

- **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

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: cd00671, 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

Coming soon

Downloads the hierarchy into CDTree

Searching with PSSMs

- Uses 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

PSI-BLAST

Attractor = PSSM

PSSM

RPS-BLAST

Attractor = sequence

PSSM PSSM PSSM PSSM PSSM PSSM

Precomputed RPS-BLAST

From **Entrez Protein**, look for the "Conserved Domains" link

1: XP_444859 Reports BLink, Conserved Domains, Links
hypothetical protein CAGL0A02189g [Candida glabrata CBS138]
gi|50284863|ref|XP_444859.1|[50284863]

Query sequence: [gi|50284863|ref|XP_444859.1|]
hypothetical protein CAGL0A02189g [Candida glabrata CBS138]

Concise Result Full Result Show Search Information

1 125 250 375 500 695

Arg_tRNA_synt_N ArgRS_core tRNA-synt_1d_C
ArgS

Descriptions

Title	Pssmid	Multi-Dom	E-value
[+]cd00671, ArgRS_core, This is the catalytic core domain of Arginyl tRNA synthetase (Arg...	28809	No	4e-71
[+]pfam05746, tRNA-synt_1d_C, DALR anticodon binding domain. This all alpha helical domai...	45641	No	4e-25
[+]pfam03485, Arg_tRNA_synt_N, Arginyl tRNA synthetase N terminal domain. This domain is ...	43411	No	3e-14
[+]COG0018, ArgS, Arginyl-tRNA synthetase [Translation, ribosomal structure and biogenesis].	30368	Yes	6e-157

Search for similar domain architectures

Web CD-Search Output

Click on a colored bar to align the query sequence to the CD

Query sequence: [gi|50284863|ref|XP_444859.1|]
hypothetical protein CAGL0A02189g [Candida glabrata CBS138]

Concise Result Full Result Show Search Information

1 125 250 375 500 695

Arg_tRNA_synt_N ArgRS_core tRNA-synt_1d_C
class_I_aaRS_core
ArgS
tRNA-synt_1d

curated CDs child parent

Descriptions

Title	Pssmid	Multi-Dom	E-value
[+]cd00671, ArgRS_core, This is the catalytic core domain of Arginyl tRNA synthetase (Arg...	28809	No	4e-71
[+]pfam05746, tRNA-synt_1d_C, DALR anticodon binding domain. This all alpha helical domai...	45641	No	4e-25
[+]pfam03485, Arg_tRNA_synt_N, Arginyl tRNA synthetase N terminal domain. This domain is ...	43411	No	3e-14
[+]cd00802, class_I_aaRS_core, Class I amino acyl-tRNA synthetase (aaRS) catalytic core d...	28812	No	0.002
[+]COG0018, ArgS, Arginyl-tRNA synthetase [Translation, ribosomal structure and biogenesis].	30368	Yes	6e-157
[+]pfam00750, tRNA-synt_1d, tRNA synthetases class I (R). Other tRNA synthetase sub-famil...	40830	Yes	6e-87

Search for similar domain architectures

Standalone RPS-BLAST

NCBI

rpsblast

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

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

mFASTA

- 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!

The NCBI Scoremat

New

NCBI

ASN.1 representation of a PSSM in two formats

SM(f) only substitution frequencies

SM(s) final substitution scores

used by blastpgp

```
intermediateData {
    freqRatios {
        { 0, 10, 0 },
        { 299632125379044, 10, -16 },
        { 0, 10, 0 },
        { 558686711792638, 10, -17 },
        { 139896854109911, 10, -16 },
        { 193133860706877, 10, -16 },
        { 109099606656283, 10, -15 },
        { 171697465661704, 10, -16 },
        { 117137722228048, 10, -16 },
        { 102752285827591, 10, -15 },
        { 223075360092639, 10, -16 },
        { 482871602722152, 10, -16 },
        { 307401569741692, 10, -16 },
        { 399867114754777, 10, -16 },
        { 478659137545398, 10, -16 },
        { 156973370804235, 10, -16 },
        { 778137132249873, 10, -16 },
        { 635768329974439, 10, -16 },
        { 247884702502916, 10, -16 },
        { 100884034314323, 10, -15 },
        { 767071894370943, 10, -17 },
        { 0, 10, 0 },
        { 194351660103265, 10, -15 },
        { 0, 10, 0 },
        { 0, 10, 0 },
        { 0, 10, 0 },
```

Pos 1

```
-
A
B
C
D
E
F
G
H
I
K
L
M
N
P
Q
R
S
T
V
W
X
Y
Z
U
*
```

```
finalData {
    scores {
        -32768,
        -304,
        -32768,
        -393,
        -427,
        -375,
        330,
        -463,
        -200,
        220,
        -301,
        -198,
        100,
        -37,
        -27,
        -317,
        132,
        -36,
        -272,
        143,
        -175,
        -101,
        572,
        -32768,
        -32768,
        -404,
```

produced by Cn3D 4.2

available via ftp

Scores scaled by 100

Making Custom Databases

NCBI

SM → **formatrpsdb** → rpsblast database

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

```
formatrpsdb -i infile -n dbname -o 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 - 12147 PSSMs
required to format: ~2.8 GB
final size: ~620 MB

Generating Custom Scorematrs

Use any scoremat as input for formatrpsdb

1 FASTA / SM(f) → **PSI-BLAST** → SM(f)

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

Coming soon

3 gFASTA → **'FA2CD'** → SM

Convert an existing alignment in gapped FASTA format to a scoremat

1 Using Scorematrs with PSI-BLAST

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

Creating a new scoremat

```
blastpgp -i infile -d database -o outfile -j 3
         -u 1 -C pssm1.smp -J T
```

Creates a new PSSM from the alignment produced by BLASTp

Starting PSI-BLAST from an existing scoremat

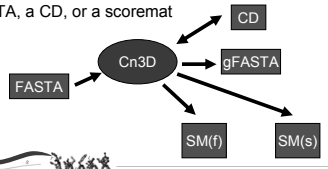
```
blastpgp -d database -o outfile -j 3
         -u 1 -C pssm2.smp -J T -q 1 -R pssm1.smp
```

Finds all proteins that align to the input PSSM

2 Cn3D 4.2

Cn3D

- 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

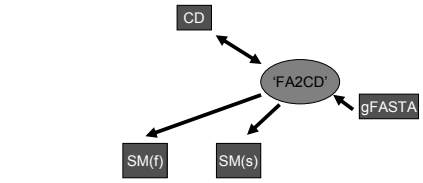


3

‘FA2CD’

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

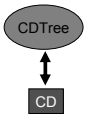
Coming soon



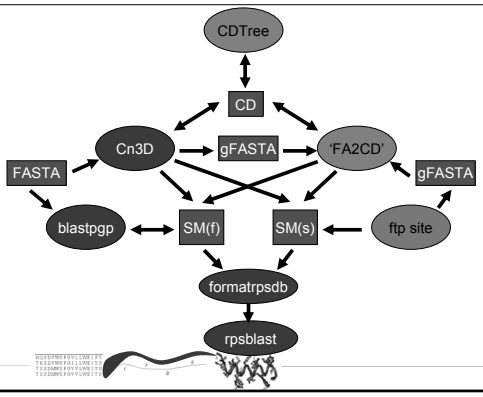
CDTree

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 207 – Try out CDTree and Cn3D 4.2