CD-Tree and Cn3D 4.2:

New Tools for Analyzing Protein Structures and Domains

Eric W. Sayers, PhD
National Center for Biotechnology Information

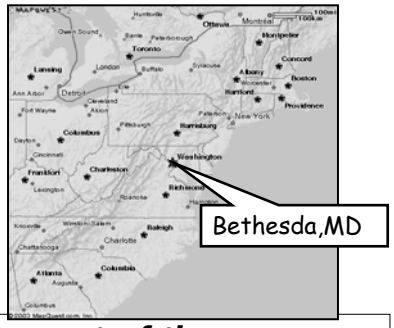
ftp://ftp.ncbi.nih.gov/pub/sayers/Biophysical.08

CD-Tree and Cn3D 4.2

- Overview of CDD
- Content of a CD
- CD Curation at NCBI
- Viewing CDs with CD-Tree
- Creating custom CDs with CD-Tree and Cn3D
- Searching with custom CDs

The National Center for Biotechnology Information

NCBI



Created in 1988 as a part of the National Library of Medicine at NIH

- Establish public databases
- Research in computational biology
- Develop software tools for sequence analysis
- Disseminate biomedical information

CD-Tree 3.0

NCBI

www.ncbi.nlm.nih.gov/

CD-Tree is…

A Viewing Tool for Conserved Domains (CDs)

- Sequences
- Phylogenetic trees
- Taxonomic trees
- 3D structure viewing in Cn3D 4.2

An Editing Tool for Conserved Domains (CDs)

- Automated interface to PSI-BLAST for updating CDs
- Creating new CDs or “child” CDs
- Remastering and sorting sequences
- Alignment editing in Cn3D 4.2

Cn3D 4.2

NCBI

What Cn3D Can Do…

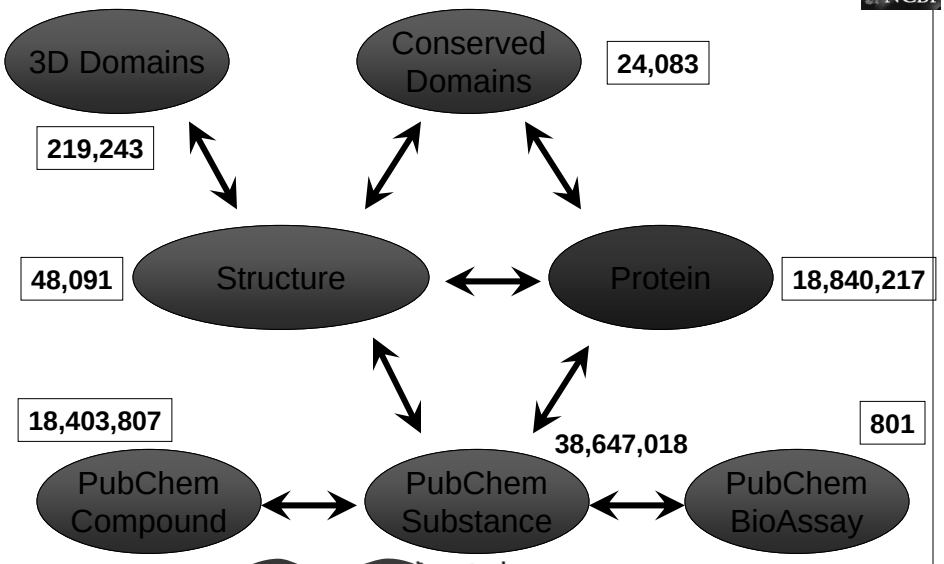
- Render, rotate, and annotate multiple structure models
- Create and edit multiple sequence alignments
- Import and align sequences and structures based on an existing alignment

What Cn3D *Can't* Do…

- Alter the structural coordinate data
- Create a theoretical structure or run MD simulations
- Read or write PDB files directly

Structural Informatics in Entrez

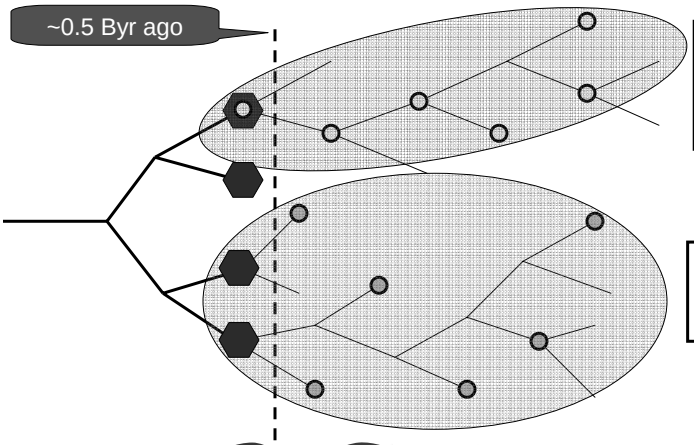
NCBI



CDD: Units of Molecular Evolution

NCBI

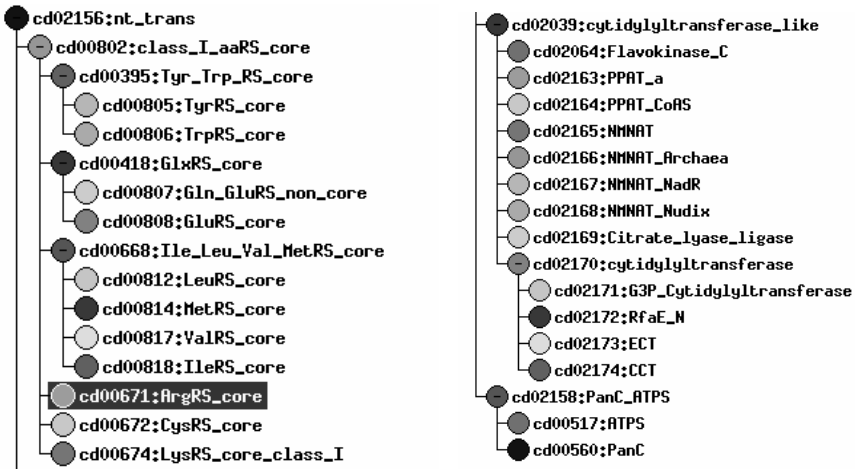
CDs represent domains that originated > ~0.5 billion years ago



CDD: Units of Molecular Evolution

NCBI

nucleotidyl transferase superfamily, containing amino acyl tRNA synthetases

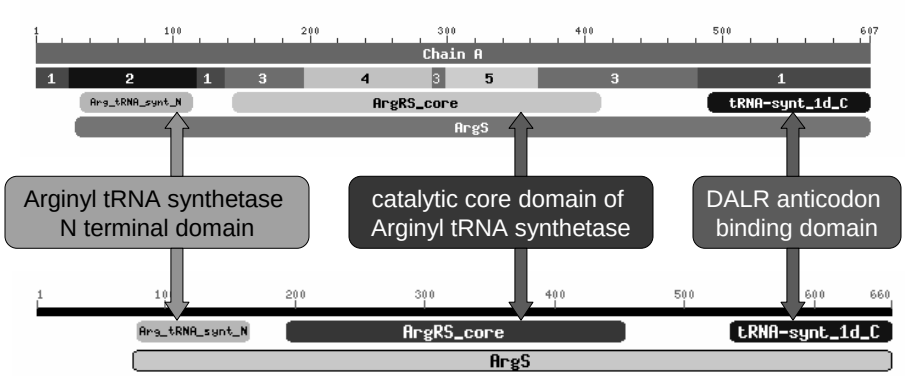


CDD: Units of Molecular Structure and Function

NCBI

1BS2: Yeast Arginyl-Trna Synthetase

Cn3D



arginyl-tRNA synthetase [Homo sapiens]

Entrez Conserved Domains (CDD)

NCBI

CDD v2.12: *A database of Position Specific Score Matrices (PSSMs)*

Single Domains

Pfam	pfam01234 8957 (38%)	Sanger	Pfam-A seeds (v21): HMM based models representing functional domains derived from SWISS-PROT
SMART	smart00123 663 (3%)	EMBL	(v4): HMM based models originally concentrating on eukaryotic signaling domains, now expanding
CD	**cd01234** **3078 (13%)**	NCBI	**NCBI curated domains based on sequence and structural alignments**

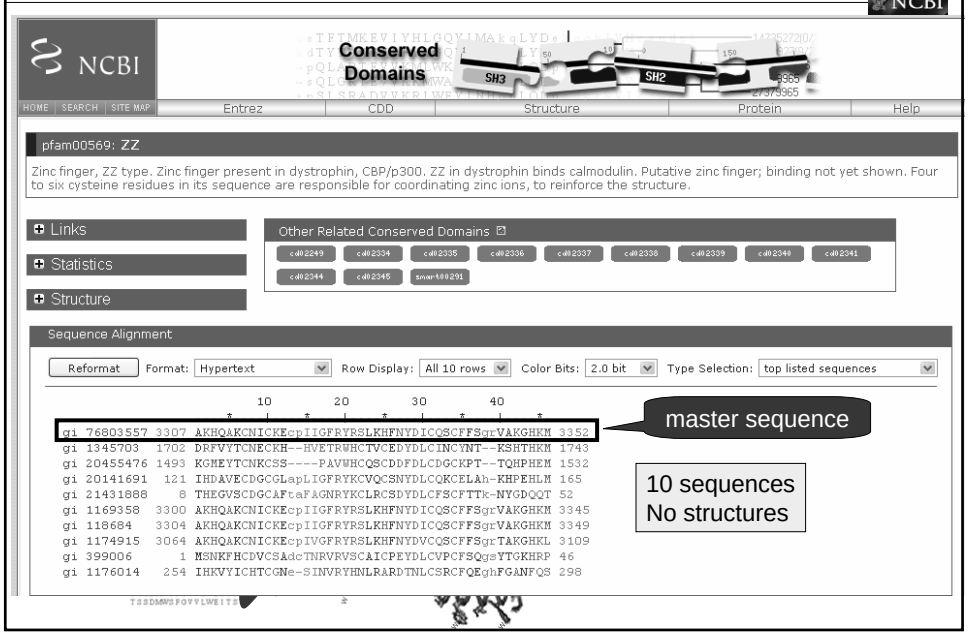
Protein Families

COG	COG0123 4873 (21%)	NCBI	BLAST based alignments derived from complete proteomes of prokaryotes
Protein Clusters	PRK01234 5952 (25%)		

CD Content

1. A protein multiple sequence alignment
   - Alignment generated from Pfam, SMART, or NCBI records
   - Each aligned sequence is present in Entrez Protein
2. Sequences with 3D structural data (whenever possible)
   - Original sequences are replaced with identical ones having 3D data, if available
3. A consensus sequence (*not shown*)
   - Contains columns populated in >50% of the aligned sequences
   - Defines the length of the alignment and PSSM
   - The consensus residue for a column is that with the highest frequency
4. A master sequence
   - The best representative protein for the alignment
   - Serves as an anchor for the alignment
5. A PSSM (Position Specific Scoring Matrix)
   - Quantifies the residue conservation in the alignment

Uncurated CD Summary



Understanding PSSMs

NCBI

- 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

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

$f(\text{observed}) > f(\text{expected})$ $\quad s > 0$

$f(\text{observed}) < f(\text{expected})$ $\quad s < 0$

C only

D,N,H

C only

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

Reading PSSMs

PSSM Viewer

NCBI

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

The NCBI Scoremat

NCBI

ASN.1 representation of a PSSM in two formats

SM(f) only substitution frequencies

used by blastpgp

produced by Cn3D 4.2

Pos 1

```
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 },
```

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

SM(s) final substitution scores

available via ftp

Scores scaled by 100

```
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,
```

Finding CDs Using PSSMs

NCBI

- 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

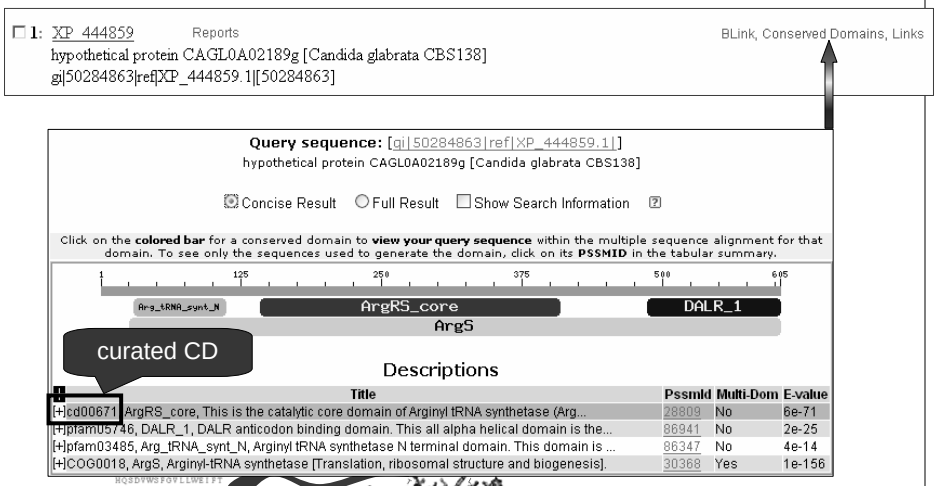
Attractor = sequence

PSSM PSSM PSSM PSSM PSSM PSSM

Precomputed RPS-BLAST

NCBI

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



CDD Curation

NCBI

- 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: Structural Neighbors

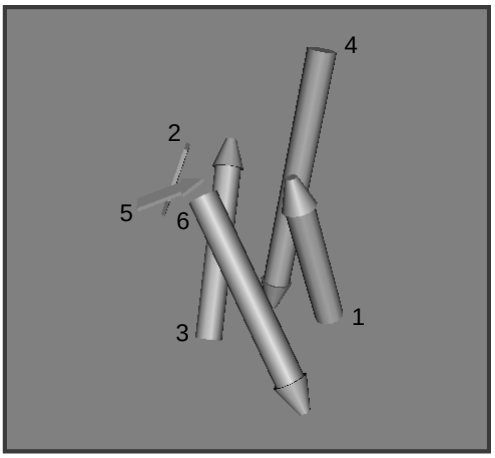
NCBI

Vector Alignment Search Tool

For each 3D domain,

locate SSEs (secondary structure elements),

and represent them as individual vectors.

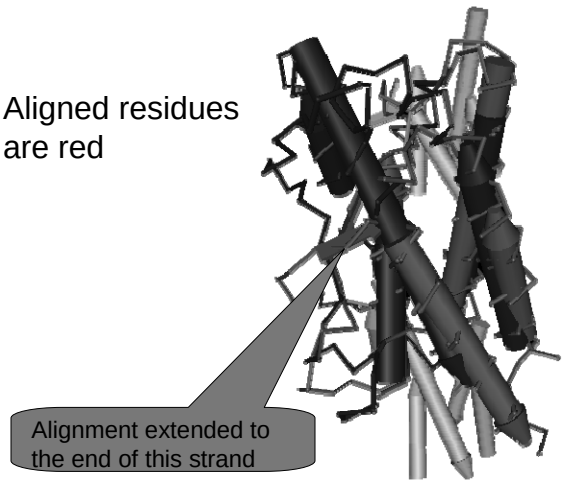


Human IL-4

VAST: Refinement

NCBI

Aligned residues are red



Cα atoms are added to the aligned SSEs

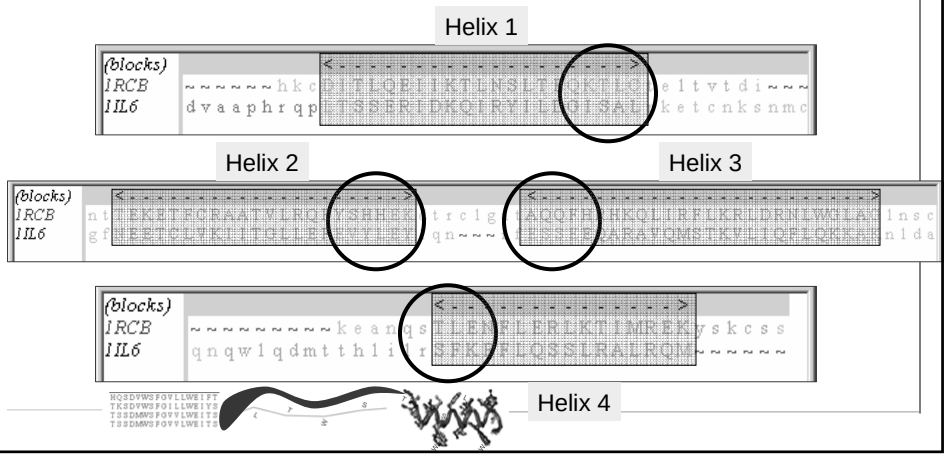
Alignments are allowed to extend beyond SSE boundaries

All atoms are added to the models, and the detailed backbone and sidechain positions are refined

VAST: Alignment of Sequence

NCBI

- *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*

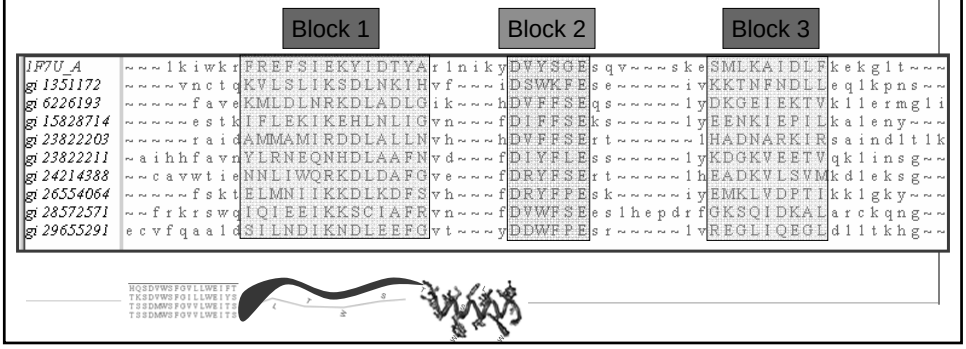


Cn3D

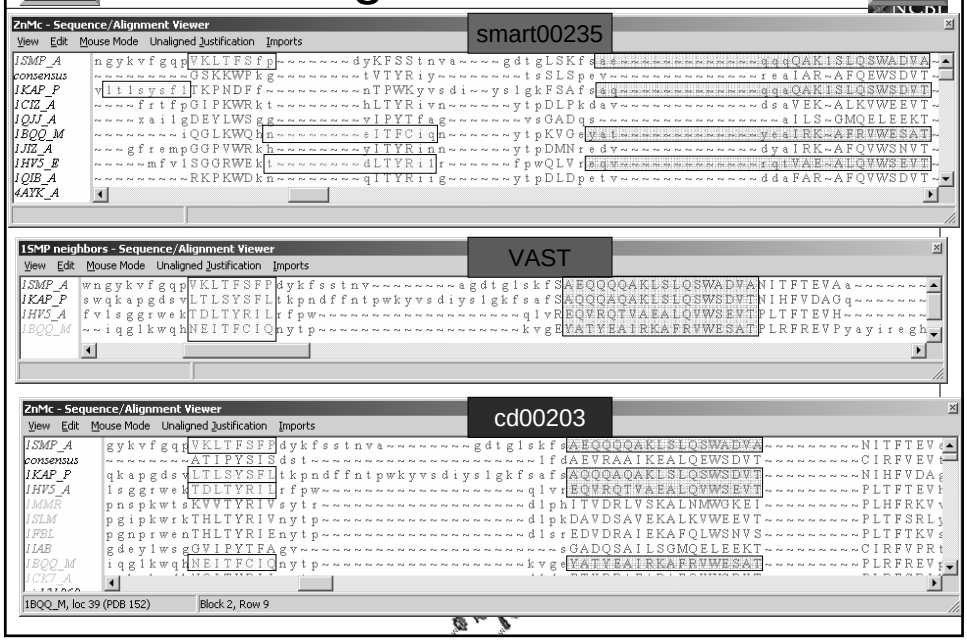
The CDD Alignment Model

NCBI

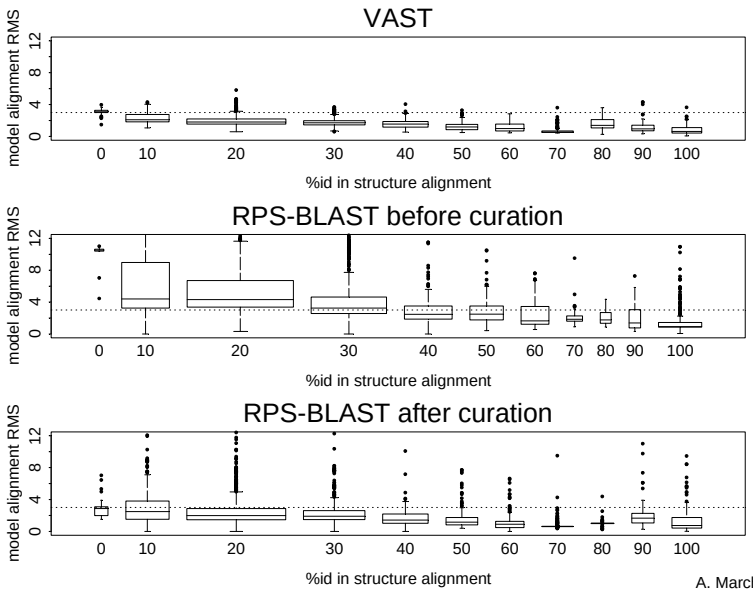
- Each sequence is aligned pairwise to the master sequence
- Aligned blocks represent secondary structure elements
- Aligned blocks have no internal gaps
- Aligned sequences have a residue at each column in the block
- Residues in the same column occupy the same position in space



Curating CDs with VAST



CD-Curation: Effect on model alignment accuracy



A. Marchler-Bauer

Curated CD: cd00671

NCBI

cd00671: **ArgRS_core, with user query added**

This is the catalytic core domain of Arginyl tRNA synthetase (ArgRS). This class I enzyme is a monomer, which aminoacylates the 2'-OH of the nucleotide at the 3' of the appropriate tRNA. The core domain is based on the Rossman fold and is responsible for the ATP-dependent formation of the enzyme bound aminoacyl-adenylate. There are at least three subgroups of ArgRS. One type contains both characteristic class I HIGH and KMSKS motifs, which are involved in ATP binding. The other subtype lacks the KMSKS motif; however, it has a lysine N-terminal to the HIGH motif, which serves as the functional counterpart to the second lysine of the KMSKS motif. A third group, which is found primarily in archaea and a few bacteria, lacks both the KMSKS motif and the HIGH loop lysine.

Links

- **Source:** Pfam
- **Taxonomy:** cellular organisms
- **PubMed:** 3 links
- **Book:** 1 link
- **Protein:** Related Protein, Related Structure, Architectures, Representatives
- **Related CDs:** 7 links

active site | HIGH motif | KMSK motif

Feature 1: active site

Evidence:

- **Structure:** 1F7U_A; ArgRS bound to L-arg and tRNA from paper - View structure with Cn3D
- **Comment:** amino acid activation is tRNA dependent
- **Citation:** PMID 11106639

Download Cn3D for Viewing 3D Structure | Scroll to Sequence Alignment Display

cd00671 is part of a hierarchy of related CD models.
Use the graphical representation to navigate this hierarchy.

Statistics

- **PSSM-Id:** 28809
- **View PSSM:** cd00671
- **Aligned:**
- **Status:**
- **Created:**
- **Updated:**

active site residues

```
Feature 1            #  # #     #  #
1F7U_A       144 KVIIEFSSPNIAKPFHAGHLRSTIIGGFLANLYEKLGWEVIRMNYLGDWGKQFGLLAVGFER.[10].PIHHLFDVYVRI 227
query        193 KVIVDFSSPNIAKEMHVGHLRSTIIGESISRLFEFAGYDVLRLNHVGDWGTQFGMLIAHLQD.[10].PIGDLQVFYKES 276
1IQ0_A       104 VVLVEHTSVNPNKELHVGHLRNIALGDAIARILAYAGREVLVLNYIDDTGRQAAETLFALRH.[10].YDHFAGRAYVRL 187
gi 2501046   111 KIFLEFVSANPTGPLHIGHARGAVFGDSLAKIARFLGHEVLCEYYVNDMGSQIRLLGLSVWL.[14].EVFYKGEYIIEI 198
gi 7674354   121 TIVVEHTSANPIHPLHLGHARNSSLGDALARLLEARGHRVNRRFYVDDVGRQAVVASLGFKL.[15].VDHAVGWVYAVT 209
gi 16081423  109 RVSVEHTSANPTGPLHIGRARNSIIGDSIYRILSRYGYRTVRQYFVNDSGKQMISLYTAYIK.[ 5].TIENLLENYQKI 187
gi 16127589  115 RVLIDYAGPNVAKPMHVGHLRASIIGESVKRLYRFRGDDVVGDAHFGDWGFQMGLLISAIMD.[35].TLADLDRIYPAA 223
gi 20178140  197 RVLVDFSSPNIAKQMHVGHLRSTIIGESLCRLLEFLQHDVIRINHLGDWGTQFGMLIAHLED.[10].PISDLQLFYKES 280
gi 20178331  193 KVIVDFSSPNIAKEMHVGHLRSTIIGESISRLFEFAGYDVLRLNHVGDWGTQFGMLIAHLQD.[10].PIGDLQVFYKES 276
gi 23474500  117 KVQIEYVSANPTGPLHIGHGRGAAIGDSLARMMRFAGFETSTEYYINDAGRQMLLLGASVYY.[13].EEYYRGDYIRDI 203
```

CD Hierarchies

NCBI

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

cd00671 is part of a hierarchy of related CD models.
Use the graphical representation to navigate this hierarchy.

Statistics

- **PSSM-Id:** 28809
- **View PSSM:** cd00671
- **Aligned:** 44 rows
- **Status:** curated CD
- **Created:** 24-Sep-2002
- **Updated:** 5-Jan-2006

Structure

Structure View

- **Program:** Cn3D
- **Drawing:** All Atoms
- **Aligned Rows:** up to 10

Download Cn3D

Hierarchy

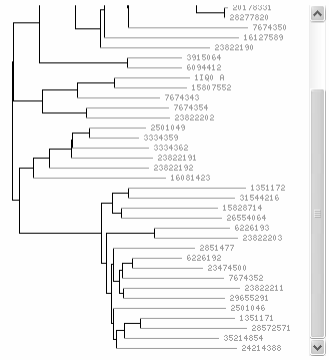
Interactive Display

- **Display:** cd00671 Branch
- Add Query Sequence:

Download CDTree

cd00671 Sequence Cluster

Zoom Out | Detailed View



Sub-family Hierarchy

Interactive Display with CDTree

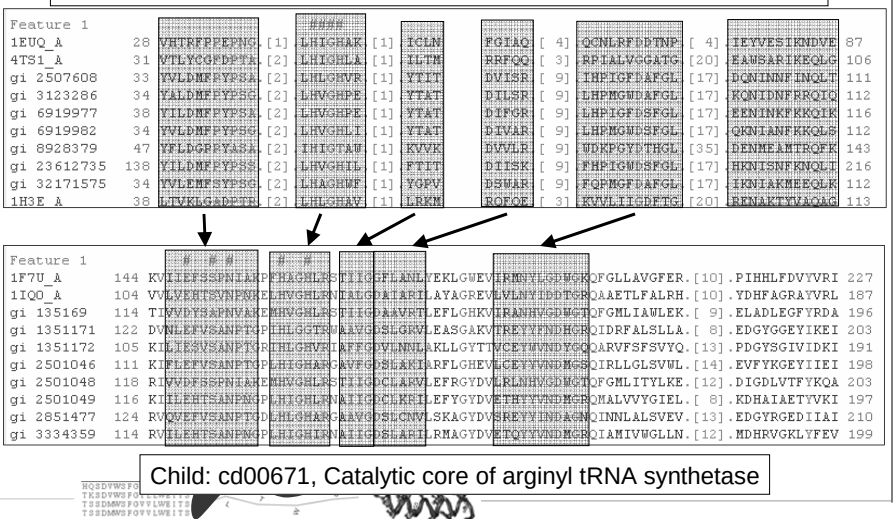
- cd02156 nt_trans
  - cd00802 class_I_aaRS_core
    - cd00395 Tyr_Trp_RS_core
      - cd00805 TyrRS_core
      - cd00806 TrpRS_core
    - cd00418 GlxRS_core
      - cd00807 Gln_GluRS_non_core
      - cd00808 GluRS_core
    - cd00668 Ile_Leu_Val_MetRS_core
      - cd00812 LeuRS_core
      - cd00814 MetRS_core
      - cd00817 ValRS_core
      - cd00818 IleRS_core
    - cd00671 ArgRS_core
    - cd00672 CysRS_core
    - cd00674 LysRS_core_class_I
  - cd02039 cutidulultransferase_like

CD Family Values

NCBI

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

NCBI

Parent: cd00802

Children are more opinionated than their parents!

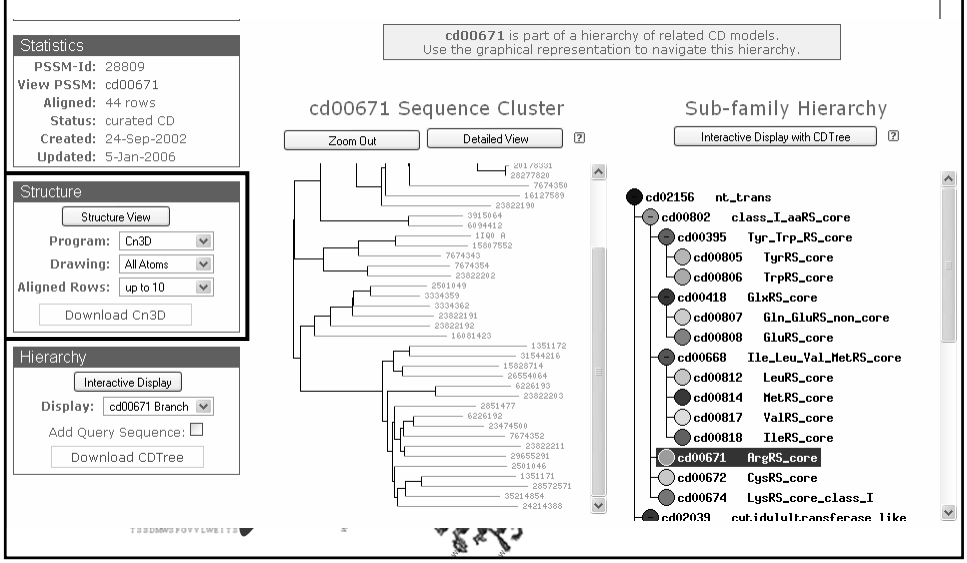
Child: cd00671

P	C	Master	A	G	I	L	V	M	F	W	P	C	S	T	Y	N	Q	H	K	R	D	E
21	A	48 - I	3	-2	1	0	0	2	1	1	-3	0	0	1	3	0	-2	-1	0	-2	-4	-3
22	L	49 - C	-1	-4	3	2	2	3	0	-2	-3	0	0	3	0	0	-2	-3	-3	-3	-2	-3
23	L	50 - L	0	-3	3	3	0	0	3	-2	-1	-3	-2	0	0	-3	0	-1	1	0	-5	-2
24	N	51 - N	1	0	0	0	0	0	-3	-5	0	0	1	0	-3	4	1	-2	2	-2	-1	-2
25	D	52 - F	-2	-2	-2	-3	-2	-2	2	5	-4	-5	-2	-3	2	-1	-1	-1	0	0	6	0
26	V	53 - G	0	-1	3	0	3	0	1	3	-4	-1	-1	0	0	-2	-1	0	-3	0	0	-1
27	L	54 - I	-1	-5	3	2	1	0	5	1	-5	-2	-3	-2	4	-3	-2	0	-2	-1	-3	-3
28	A	55 - A	4	-1	0	0	1	0	-2	-5	-1	0	0	0	-1	0	3	-1	-1	-1	-4	-3
29	R	56 - Q	-2	-2	-4	-3	-3	-3	-5	-5	-4	-3	-2	-2	-4	0	3	0	3	7	0	0
30	Y	57 - D	1	-3	-2	-1	-2	0	4	3	-4	-4	-1	-2	6	0	0	2	0	0	-1	0
P	C	Master	A	G	I	L	V	M	F	W	P	C	S	T	Y	N	Q	H	K	R	D	E

23	A	166 - T	5	-1	0	-2	0	-2	-3	-4	-2	-2	2	4	-3	0	-2	-3	-2	-2	0	-2
24	I	167 - I	3	-3	5	0	3	0	0	-3	-3	0	0	-1	0	-4	-3	-3	-3	-3	-4	-3
25	I	168 - I	-2	-5	6	2	2	0	1	-3	-4	-2	-3	0	2	-1	-3	-3	-4	-4	-4	-4
26	G	169 - G	-1	7	-5	-5	-4	-4	-4	-4	-3	-4	0	-3	-4	-1	-3	-3	-3	-3	-2	-3
27	D	170 - G	-1	-1	-4	-5	-4	-4	5	5	-2	-4	0	-2	-4	0	1	-2	-1	-2	7	3
28	S	171 - F	2	-2	0	-2	1	-2	0	-3	-2	4	4	2	-2	-2	-2	-3	-2	-2	-2	-2
29	L	172 - L	-1	-5	3	5	1	4	-1	-3	-4	-2	-3	-2	-2	-4	-3	-4	-3	-3	-5	-4
30	A	173 - A	5	0	-2	-3	0	0	-3	-3	-2	3	2	-1	0	0	-2	-2	0	0	-2	-1
P	C	Master	A	G	I	L	V	M	F	W	P	C	S	T	Y	N	Q	H	K	R	D	E
31	R	174 - N	-2	-3	-4	-4	-4	-3	-4	-4	-3	-4	0	-2	-3	4	0	-1	0	7	-1	0
32	I	175 - L	-1	-4	5	3	2	2	-1	-3	-4	0	-3	0	-2	-4	-3	-4	-3	-4	-4	-4

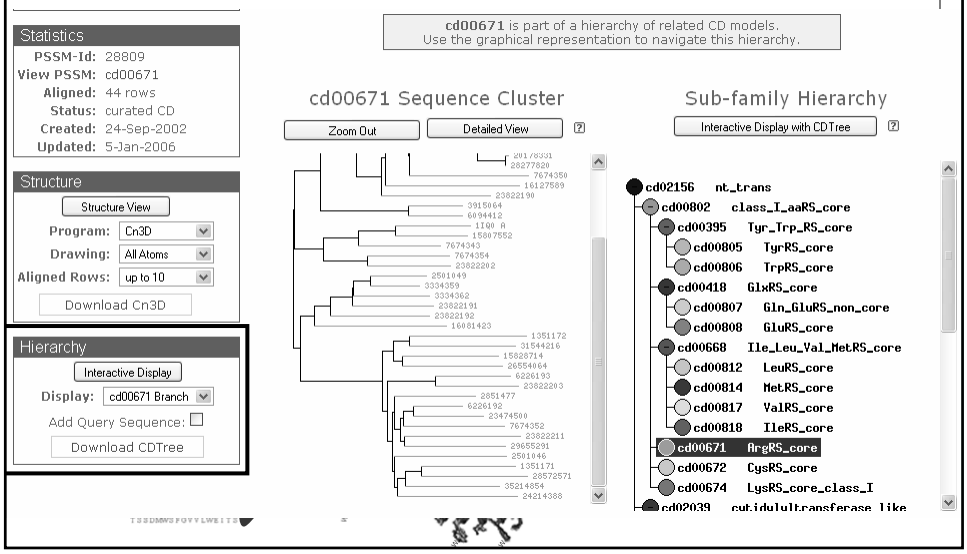
Viewing CDs in Cn3D

cd00671: Catalytic core of arginyl tRNA synthetase



Viewing CDs in CD-Tree

cd00671: Catalytic core of arginyl tRNA synthetase



CDTree

Viewing CDs in CD-Tree

NCBI

Displays a taxonomic tree

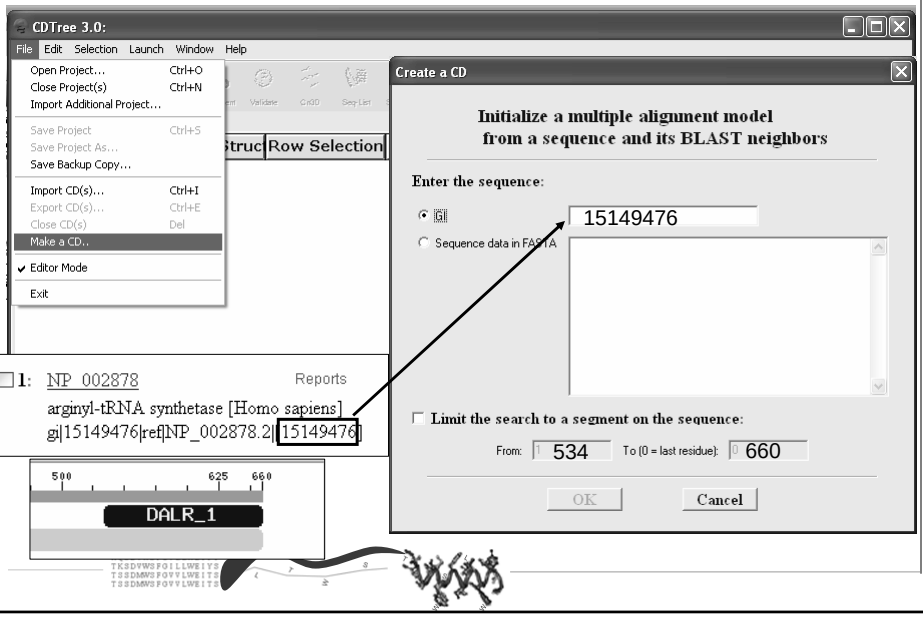
Displays a phylogenetic tree

View and sort sequences; remaster the alignment

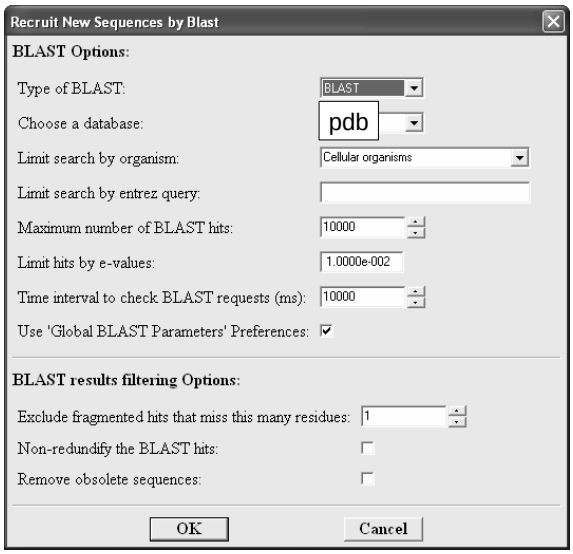
View structures and align imported sequences

Creating CDs with CD-Tree

NCBI



Set BLAST Parameters



Aligning Imported Structures

1. Align the structure(s) found by BLAST to the alignment model, changing the alignment model if necessary

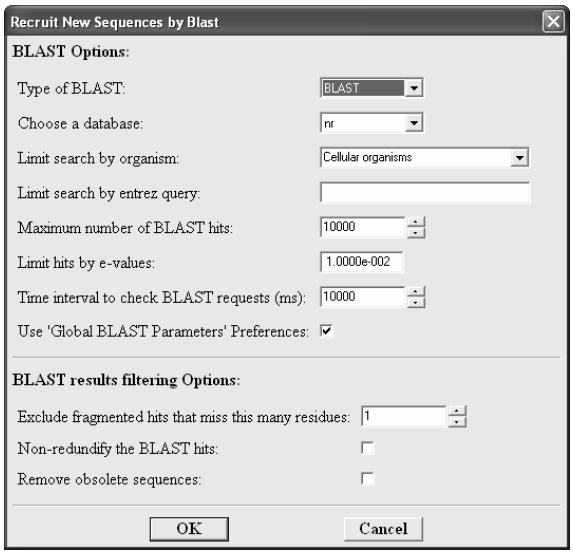
2. Make one of the structures the “master” sequence (remaster)

CDTree

Cn3D

Recruit New Sequences

NCBI



Recruit Structures with PSI-BLAST

NCBI

- Given the new multiple sequence alignment, recruit more sequences using the new PSSM.
- First, try to recruit more structures.
- Next, import the VAST alignments of any new structures, and use these as guides to merge the new structures into the alignment

Import Final CDTree

Creating CDs from mFASTA

NCBI

fa2cd.exe

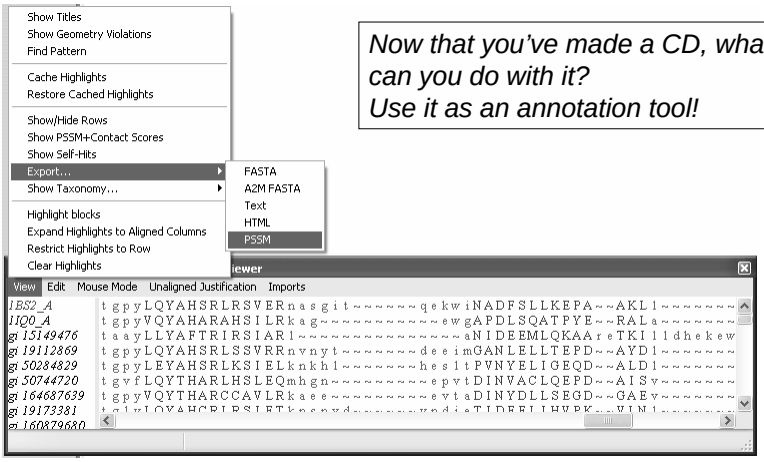
- command-line conversion utility
- comes with CD-Tree
- Input: mFASTA
- Output: CD

```
fa2cd –i infile –n cdname –o output_cd
```

Generating Custom Scoremats

NCBI

Export a PSSM from Cn3D 4.2



Now that you've made a CD, what can you do with it?
Use it as an annotation tool!

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
```

name of new database

scaling parameters

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

text file containing scoremat file names

Standalone RPS-BLAST

NCBI

rpsblast

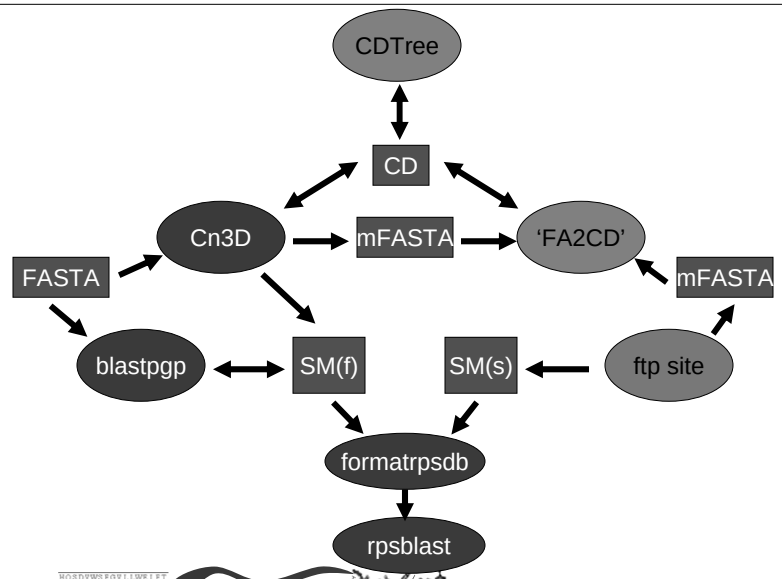
ftp://ftp.ncbi.nih.gov/blast/executables/LATEST

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

mFASTA

- Why run RPS-BLAST on your own machine?
  - annotate functional domains for many proteins
  - ultimate flexibility
  - customized databases
  - parsing output
  - speed

The CDD Data Tree



For More Information…

E-mail addresses *or…stop by our booth!*

• Eric Sayers	sayers@ncbi.nlm.nih.gov
• General Help	info@ncbi.nlm.nih.gov
• BLAST	blast-help@ncbi.nlm.nih.gov

The (free!) NCBI Newsletter

http://www.ncbi.nih.gov/About/newsletter.html

The NCBI Handbook

Follow the link from the NCBI Home Page

The NCBI Education Page

http://www.ncbi.nih.gov/Education/index.html