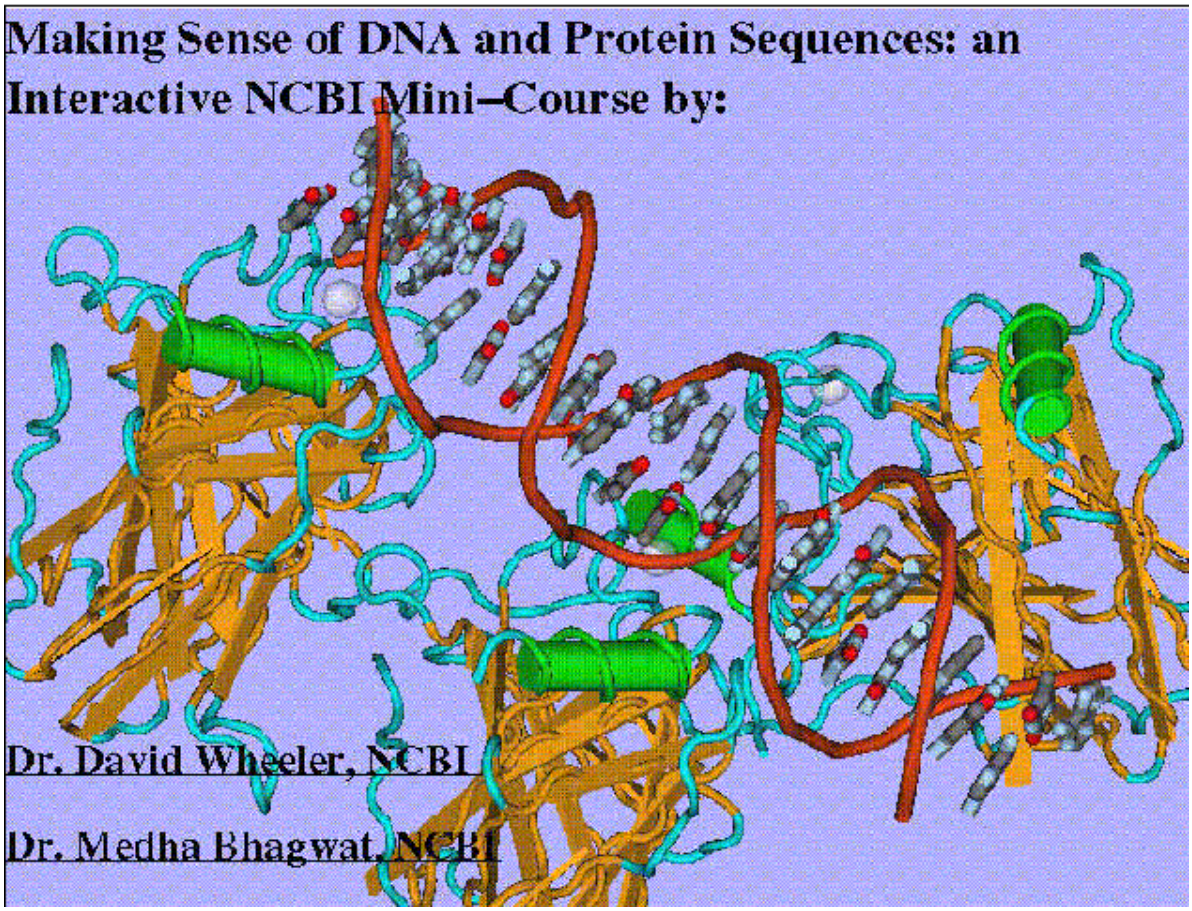
Making Sense of DNA and Protein Sequences: an Interactive NCBI Mini–Course by:



Dr. David Wheeler, NCBI

Dr. Medha Bhagwat, NCBI

Introduction:

In this course (http://www.ncbi.nlm.nih.gov/Class/minicourses/), we will first try to make sense of the DNA sequence by determining whether it codes for a protein. If it does, then we will use this protein sequence to search for the presence of any motifs or structural domains present in it and also to predict its function. Finally, we will map the protein sequence onto the structure of a protein with similar sequence.

We recommend beginning with the uncharacterized *Drosophila melanogaster* genomic sequence from the GenBank record AE003584 found in the first electronic notebook, however, you can use another uncharacterized *Drosophila melanogaster* genomic sequence by choosing another notebook from the list below.

Electronic Notebook for Protein Sequence Analysis

The electronic notebook is a tutorial and analysis web-form consisting of a set of links to protein analysis tools combined with areas into which results and personal notes can be recorded. All the analysis tools open into a second "tools" window from which the results of an analysis can be pasted into the electronic notebook. The "Cheat now!" links open a third window in which a complete set of results have already been recorded. The electronic notebook can also be used to analyze a new DNA sequence by substituting the new sequence the original sequence found in the DNA sequence text area. The electronic notebooks used in this course are publicly accessible over the internet.

URLs Used:

1. **Class Page:** http://www.ncbi.nlm.nih.gov/Class/minicourses/
2. **GenScan:** http://genes.mit.edu/GENSCAN.html
3. **ScanProsite:** http://www.expasy.org/tools/scanprosite/
4. **BLASTP:** http://www.ncbi.nlm.nih.gov/BLAST
5. **COGs:** http://www.ncbi.nlm.nih.gov/COG/old/
6. **MultAlin:** http://prodes.toulouse.inra.fr/multalin/multalin.html
7. **CDD:** http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi

NoteBooks:

http://www.ncbi.nlm.nih.gov/Class/minicourses/x1a.html
http://www.ncbi.nlm.nih.gov/Class/minicourses/x2a.html

Outline

Making Sense of DNA and Protein Sequences
Eukaryotic DNA query (Drosophila genome)
Predict coding region/exons (GenScan)
Obtain protein product (GenScan)
Identify motif/site (ScanProsite)
Search for similar sequences (BLASTp)
Predict function (COG)
Perform multiple sequence alignment (Multalin)
Obtain 3-D structural template (CDD)

To identify any exons in the DNA sequence and generate a predicted protein sequence, click here:

GenScan

Paste your DNA sequence into the GenScan input window. Press the *"Run Genscan"* button. Select the protein translation with the highest exon P–values and paste this FASTA formatted output into your notebook.

Protein Sequence from Genscan

The New GENSCAN Web Server at MIT

Identification of complete gene structures in genomic DNA

```
                \\|//
                (o o)
-. .-.   .-oOOo~(_)~oOOo-.   .-. .-.   .-. .-.   .-. .-.   .-. .-.   .-. .-.   .-. .-.   .-. .-.   .-
||X|||\ /|||X|||\ /|||X|||\ /|||X|||\ /|||X|||\ /|||X|||\ /|||X|||\ /|||X|||\ /|||X|||\ /|||X|||\ /||
|/ \|||X|||/ \|||X|||/ \|||X|||/ \|||X|||/ \|||X|||/ \|||X|||/ \|||X|||/ \|||X|||/ \|||X|||/ \|||X|||/
'   `-' `-'   `-' `-'   `-' `-'   `-' `-'   `-' `-'   `-' `-'   `-' `-'   `-' `-'   `-' `-'   `-' `-'
```

For information about Genscan, click here

This server provides access to the program Genscan for predicting the locations and exon-intron structures of genes in genomic sequences from a variety of organisms.

This server can accept sequences up to 1 million base pairs (1 Mbp) in length. If you have trouble with the web server or if you have a large number of sequences to process, request a local copy of the program (see instructions at the bottom of this page) or use the GENSCAN email server. If your browser (*e.g.*, Lynx) does not support file upload or multipart forms, use the older version.

Organism: Vertebrate Suboptimal exon cutoff (optional): 1.00

Sequence name (optional):

Print options: Predicted peptides only

Upload your DNA sequence file (one-letter code, upper or lower case, spaces/numbers ignored):

Browse...

Or paste your DNA sequence here (one-letter code, upper or lower case, spaces/numbers ignored):

```
   239041 gctcctgtct gaacgatcca gcgcgatgtt gatcaccgcc ttgttgtcat
catccacaat
   239101 gatggtgacg ttgacatggg ttttgttgcc gtaattgagg cgcctgaagt
ggaagtccct
   239161 atgcagaaac ttcggatgga tgttgcactc cagatgctga ttgctaaaac
gaaagctgcc
   239221 aaagctcagg accatggctt ggatgactcc cgaagctccc gctcgaataa
gattgtggca
   239281 gccctgtttc tccaacgtta gcatccacga cgtaaccagg gagttaagct
gctgaagatt
   239341 agacatctcg cgccaaagat tctccgcctg cagagtgtga tacgaatcgt
agcagccttc
```

To have the results mailed to you, enter your email address here (optional):

Run GENSCAN Clear Input

GENSCAN 1.0 Date run: 20-Jun-2007 Time: 09:48:28

Sequence 09:48:24 : 5100 bp : 46.29% C+G : Isochore 2 (43 - 51 C+G%)

Parameter matrix: HumanIso.smat

Predicted genes/exons:

Gn.Ex	Type	S	.Begin	...End	.Len	Fr	Ph	I/Ac	Do/T	CodRg	P....	Tscr..
1.01	Sngl	+	27	458	432	2	0	48	49	383	0.447	24.68
1.02	PlyA	+	489	494	6							1.05
2.00	Prom	+	830	869	40							-6.86
2.01	Init	+	1002	1069	68	2	2	53	89	83	0.970	3.88
2.02	Intr	+	2549	2708	160	2	1	72	105	284	0.980	28.49
2.03	Intr	+	2771	2872	102	1	0	10	86	251	0.999	17.47
2.04	Intr	+	2935	3183	249	0	0	73	100	586	0.999	55.93
2.05	Term	+	3253	3948	696	0	0	90	49	1324	0.999	122.25
2.06	PlyA	+	4120	4125	6							1.05
3.04	PlyA	-	4162	4157	6							-0.45
3.03	Term	-	4448	4261	188	0	2	37	42	95	0.922	-2.55
3.02	Intr	-	4635	4511	125	2	2	44	90	91	0.949	5.13
3.01	Init	-	5046	4694	353	0	2	66	43	485	0.897	38.43

Click here to view a PDF image of the predicted gene(s)

Click here for a PostScript image of the predicted gene(s)

Predicted peptide sequence(s):

>09:48:24|GENSCAN_predicted_peptide_1|143_aa
MPRTLPWTTVFTAVASSARAKSMEKLTVVFLLRMHSALVVSQPSMATRVNLPVFDPQSLN
SRAPAKTTSAAQAITAYLSIFFHLIELQGKRIGWLFRWLSPLSASSQRYESTKSGESPKT
TQSFRMNGKQLRAATQKKAFFDD

>09:48:24|GENSCAN_predicted_peptide_2|424_aa
MSQICKRGLLISNRLAPAALRCKSTWFSEVQMGPPDAILGVTEAFKKDTNPKKINLGAGA
YRDDNTQPFVLPSVREAEKRVVSRSLDKEYATIIGIPEFYNKAIELALGKGSKRLAAKHN
VTAQSISGTGALRIGAAFLAKFWQGNREIYIPSPSWGNHVAIFEHAGLPVNRYRYYDKDT
CALDFGGLIEDLKKIPEKSIVLLHACAHNPTGVDPTLEQWREISALVKKRNLYPFIDMAY
QGFATGDIDRDAQAVRTFEADGHDFCLAQSFAKNMGLYGERAGAFTVLCSDEEEAARVMS
QVKILIRGLYSNPPVHGARIAAEILNNEDLRAQWLKDVKLMADRIIDVRTKLKDNLIKLG
SSQNWDHIVNQIGMFCFTGLKPEQVQKLIKDHSVYLTNDGRVSMAGVTSKNVEYLAESIH
KVTK

>09:48:24|GENSCAN_predicted_peptide_3|221_aa
MSNLQQLNSLVTSWMLTLEKQGCHNLIRAGASGVIQAMVLSFGSFRFSNQHLECNIHPKF
LHRDFHFRRLNYGNKTHVNVTIIVDDDNKAVINIALDRSDRSYYACDGGCLDEPVLLTQN
RRQFPVKLTEPLTAILYITEDKQHMEELHHAIHVKEVVEAPAHEQHLIALHRHGHQLGGL
PTLFWVSVCAIIIVFHIFLCKLIIKEYCEPSDKLRYRYNKP

To scan the protein sequence for the occurrence of motifs/patterns found in the PROSITE database, use:

ScanProsite

Paste the protein sequence from GenScan into the ScanProsite input box and press the *"Start the Scan"* button. Paste the ScanProsite hit into your notebook. To see the Prosite summary for the hit, click on the PDOCxxxx number.

Hit from ScanProsite

Prosite pattern

Cheat Now!

Home | ScanProsite | ProRule | Documents | Downloads | Links

ScanProsite

The ScanProsite tool [Help / Commercial users] allows to scan protein sequence(s) (either from UniProt Knowledgebase (Swiss-Prot/TrEMBL) or PDB or provided by the user) for the occurrence of patterns, profiles and rules (motifs) stored in the PROSITE database, or to search protein database(s) for hits by specific motif(s) [Reference / Download ps_scan, the standalone version]. The program PRATT can be used to generate your own patterns. You may either:

- Enter one or more PROSITE accession numbers and/or patterns [1 by line] to search the UniProt Knowledgebase (Swiss-Prot/TrEMBL) and/or PDB databases, **OR**
- Enter one or more sequences [raw, Swiss-Prot or fasta format] and/or UniProt Knowledgebase (Swiss-Prot/TrEMBL) accession numbers and/or PDB accession numbers [1 by line] to be scanned with all patterns, profiles, rules in PROSITE, **OR**
- Fill in both fields to find all occurrences of specified motifs in specified sequences.

Protein(s) to be scanned:

Enter one or more Swiss-Prot/TrEMBL accession number(s) [AC] (e.g. **P00747**) and/or sequence identifier(s) [ID] (e.g. **ENTK_HUMAN**), and/or PDB identifier, and/or paste **your own protein sequence(s)** in the box below:
(leave this box blank to scan PROSITE entrie(s) against selected protein databases)

```
MSQICKRGLLISNRLAPAALRCKSTWFSEVQMGPPDAILGVT
YRDDNTQPFVLPSVREAEKRVVSRSLDKEYATIIGIPEFYNK.
VTAQSISGTGALRIGAAFLAKFWQGNREIYIPSPSWGNHVAI
CALDFGGLIEDLKKIPEKSIVLLHACAHNPTGVDPTLEQWRE
QGFATGDIDRDAQAVRTFEADGHDFCLAQSFAKNMGLYGERA
QVKILIRGLYSNPPVHGARIAAEILNNEDLRAQWLKDVKLMA
SSQNWDHIVNQIGMFCFTGLKPEQVQKLIKDHSVYLTNDGRV
KVTK
```

Clear

General options:

- [x] Exclude motifs with a high probability of occurrence
- [ ] Show low level score
- [ ] Do not scan profiles [User Manual]

Show only sequences with at least [ ] hit(s)
Maximum of matched sequences [1000]
Output format [Graphical rich view]
[ ] Retrieve complete sequences
Your e-mail (optional): [ ] (will send results by e-mail)

START THE SCAN | reset

PROSITE pattern(s)/profile(s) to scan for:

Enter one or more PROSITE accession number(s) (e.g. **PS50240**), and/or identifier(s) (e.g. **CHEB**), and/or type **your pattern(s)** in PROSITE format in the box below:
(leave this box blank to scan sequence(s) against the entire PROSITE database)

and specify your search limits (only used if no protein data specified) :

- **Protein database(s):** [x] Swiss-Prot [ ] TrEMBL [ ] PDB databases
  [x] including splice variants
  randomize databases [no] (only with patterns, see help)
- Taxonomic lineage (OC) / species (OS) filter: [ ]
  (see NEWT Taxonomy ; separate multiple taxa/species with a semicolon, e.g. *Eukaryota; Escherichia coli;* . Does not work on PDB sequences.)
- Description (DE) filter: [ ] e.g. *protease*

pattern options:

Allow at most [1] X sequence characters to match a conserved position in the pattern
match mode [greedy, overlaps, no includes] (for patterns, see help)

Home ScanProsite ProRule Do

ScanProsite Results Viewer

This view shows ScanProsite results together with ProRule-based predicted intra-domain features (help).

Hits for all PROSITE (release 20.10) motifs on sequence 10-10-45-GENSCAN_predicted_peptide_2-424_aa :

found: 1 hit in 1 sequence

10-10-45-GENSCAN_predicted_peptide_2-424_aa (424 aa)

```
MSQICKRGLLISNRLAPAALRCKSTWFSEVQMGPPDAILGVTEAFKKDTNPKKINLGAGAYRDDNT
QPFVLPSVREAEKRVVSRSLDKEYATIIGIPEFYNKAIELALGKGSKRLAAKHNVTAQSISGTGAL
RIGAAFLAKFWQGNREIYIPSPSWGNHVAIFEHAGLPVNRYRYYDKDTCALDFGGLIEDLKKIPEK
SIVLLHACAHNPTGVDPTLEQWREISALVKKRNLYPFIDMAYQGFATGDIDRDAQAVRTFEADGHD
FCLAQSFAKNMGLYGERAGAFTVLCSDEEEAARVMSQVKILIRGLYSNPPVHGARIAAEILNNEDL
RAQWLKDVKLMADRIIDVRTKLKDNLIKLGSSQNWDHIVNQIGMFCFTGLKPEQVQKLIKDHSVYL
TNDGRVSMAGVTSKNVEYLAESIHKVTK
```

ruler: 1 100 200 300 400 500 600 700 800 900 10

hits by patterns: [1 hit (by 1 pattern) on 1 sequence]

Hits by PS00105 **AA_TRANSFER_CLASS_1** *Aminotransferases class-I pyridoxal-phosphate attachment site* :

10-10-45-GENS~ (424 aa)

270 - 283: SFAKnmGLyGERAG

Legend:

disulfide bridge active site other 'ranges' other sites

horizontal scaling: 0.6

do not show text labels:

do not show sites in hits:

do not show ranges in hits:

redisplay

Aminotransferases class-I pyridoxal-phosphate attachment site

Description:

Aminotransferases share certain mechanistic features with other pyridoxal-phosphate dependent enzymes, such as the covalent binding of the pyridoxal-phosphate group to a lysine residue. On the basis of sequence similarity, these various enzymes can be grouped [1,2] into subfamilies. One of these, called class-I, currently consists of the following enzymes:

- Aspartate aminotransferase (AAT) (EC 2.6.1.1). AAT catalyzes the reversible transfer of the amino group from L-aspartate to 2-oxoglutarate to form oxaloacetate and L-glutamate. In eukaryotes, there are two AAT isozymes: one is located in the mitochondrial matrix, the second is cytoplasmic. In prokaryotes, only one form of AAT is found (gene aspC).
- Tyrosine aminotransferase (EC 2.6.1.5) which catalyzes the first step in tyrosine catabolism by reversibly transferring its amino group to 2- oxoglutarate to form 4-hydroxyphenylpyruvate and L-glutamate.
- Aromatic aminotransferase (EC 2.6.1.57) involved in the synthesis of Phe, Tyr, Asp and Leu (gene tyrB).
- 1-aminocyclopropane-1-carboxylate synthase (EC 4.4.1.14) (ACC synthase) from plants. ACC synthase catalyzes the first step in ethylene biosynthesis.
- Pseudomonas denitrificans cobC, which is involved in cobalamin biosynthesis.
- Yeast hypothetical protein YJL060w.

The sequence around the pyridoxal-phosphate attachment site of this class of enzyme is sufficiently conserved to allow the creation of a specific pattern.

Last update:
April 2006 / Pattern and text revised.

Technical section:

PROSITE method (with tools and information) covered by this documentation:

AA_TRANSFER_CLASS_1, PS00105; **Aminotransferases class-I pyridoxal-phosphate attachment site** (PATTERN)

Consensus pattern:	**[GS] - [LIVMFYTAC] - [GSTA] - K - x(2) - [GSALVN] - [LIVMFA] - x - [GNAR] - {V} - R - [LIVMA] - [GA]** *K is the pyridoxal-P attachment site*
Sequences known to belong to this class detected by the pattern:	ALL
Other sequence(s) detected in Swiss-Prot:	1

To search for proteins with similar sequences, use BLAST:

BLAST

Run a BLASTp search against the Swiss-Prot database by pasting the protein sequence from GenScan into the input box on the BLASTp page. Choose the SwissProt database from the database listbox, then press the ***"BLAST"*** button. Format your results as "Flat-query anchored with dots for identities" by selecting the "Reformat these Results" link on the results page and paste this alignment into your notebook.

BLASTP Alignment (against SwissProt)

BLAST Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

NCBI/ BLAST/ blastp suite: BLASTP programs search protein databases using a protein query. more... Reset page Bookmark

Enter Query Sequence

Enter accession number, gi, or FASTA sequence Clear

```
CALDFGGLIEDLKKIPEKSIVLLHACAHNPTGVDPTLEQWREISALVKKRNLYPFIDMAY
QGFATGDIDRDAQAVRTFEADGHDFCLAQSFAKNMGLYGERAGAFTVLCSDEEEAARVMS
QVKILIRGLYSNPPVHGARIAAEILNNEDLRAQWLKDVKLMADRIIDVRTKLKDNLIKLG
SSQNWDHIVNQIGMFCFTGLKPEQVQKLIKDHSVYLTNDGRVSMAGVTSKNVEYLAESIH
KVTK
```

Query subrange

From

To

Or, upload file Browse...

Job Title

Enter a descriptive title for your BLAST search

Choose Search Set

Database Swissprot protein sequences(swissprot)

Organism
Optional

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Entrez Query
Optional

Enter an Entrez query to limit search

Program Selection

Algorithm

- blastp (protein-protein BLAST)
- PSI-BLAST (Position-Specific Iterated BLAST)
- PHI-BLAST (Pattern Hit Initiated BLAST)

Choose a BLAST algorithm

BLAST Search database swissprot using Blastp (protein-protein BLAST)

Show results in a new window

BLAST Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

My NCBI [Sign In] [Register]

NCBI/ BLAST/ blastp/ Formatting Results - 6GM3GGM4014 [Formatting options]

Job Title: 14:09:30|GENSCAN_predicted_peptide_2|424_aa

Putative conserved domains have been detected, click on the image below for detailed results.

1 75 150 225 300 375 424

TyrB

Request ID	6GM3GGM4014
Status	Searching
Submitted at	Thu Jun 7 14:37:36 2007
Current time	Thu Jun 7 14:37:41 2007
Time since submission	00:00:04

This page will be automatically updated in 6 seconds

BLAST Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

NCBI/ BLAST/ Format Request

Query Protein sequence(424 letters)

Database swissprot

Job title Protein sequence(424 letters)

Request ID 7073S5X6014 View report Show results in a new window

Format

Show Alignment as HTML Advanced View Reset form to defaults

Alignment View Flat query-anchored with dots for identities

Display Graphical Overview Linkout Sequence Retrieval NCBI-gi

Masking Character: Lower Case Masking Color: Grey

Limit results Descriptions: 100 Graphical overview: 100 Alignments: 100

Organism Type common name, binomial, taxid, or group name. Only 20 top taxa will be shown.

Enter organism name or id--completions will be suggested

Entrez query:

Expect Min: 0 Expect Max: 10

Format for PSI-BLAST with inclusion threshold: 0.001

Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback on new interface

BLAST — Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

My NCBI [Sign In] [Register]

NCBI/ BLAST/ blastp/ Formatting Results - 7JCYSGUJ014 [Reformat these Results] [Edit and Resubmit] [Sign in above to save your search strategy]

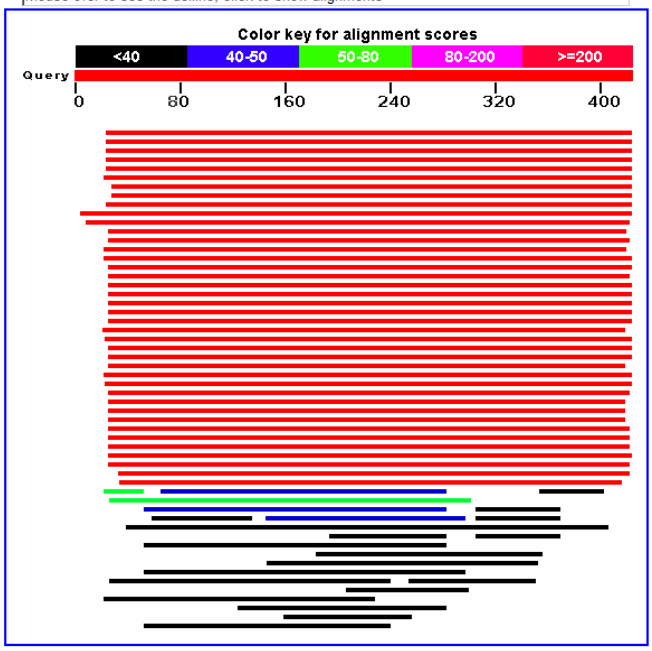
Job Title: Protein sequence(424 letters)

Show Conserved Domains

```
BLASTP 2.2.16 (Mar-25-2007)

Reference:
Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman
(1997), "Gapped BLAST and PSI-BLAST: a new generation of
protein database search programs", Nucleic Acids Res. 25:3389-3402.

Reference:
Schäffer, Alejandro A., L. Aravind, Thomas L. Madden, Sergei
Shavirin, John L. Spouge, Yuri I. Wolf, Eugene V. Koonin, and
Stephen F. Altschul (2001), "Improving the accuracy of PSI-BLAST
protein database searches with composition-based statistics
and other refinements", Nucleic Acids Res. 29:2994-3005.

RID: 7JCYSGUJ014

Database: Non-redundant SwissProt sequences
          247,743 sequences; 93,466,512 total letters

If you have any problems or questions with the results of this search
please refer to the BLAST FAQs
Taxonomy reports

Query=
Length=424
```

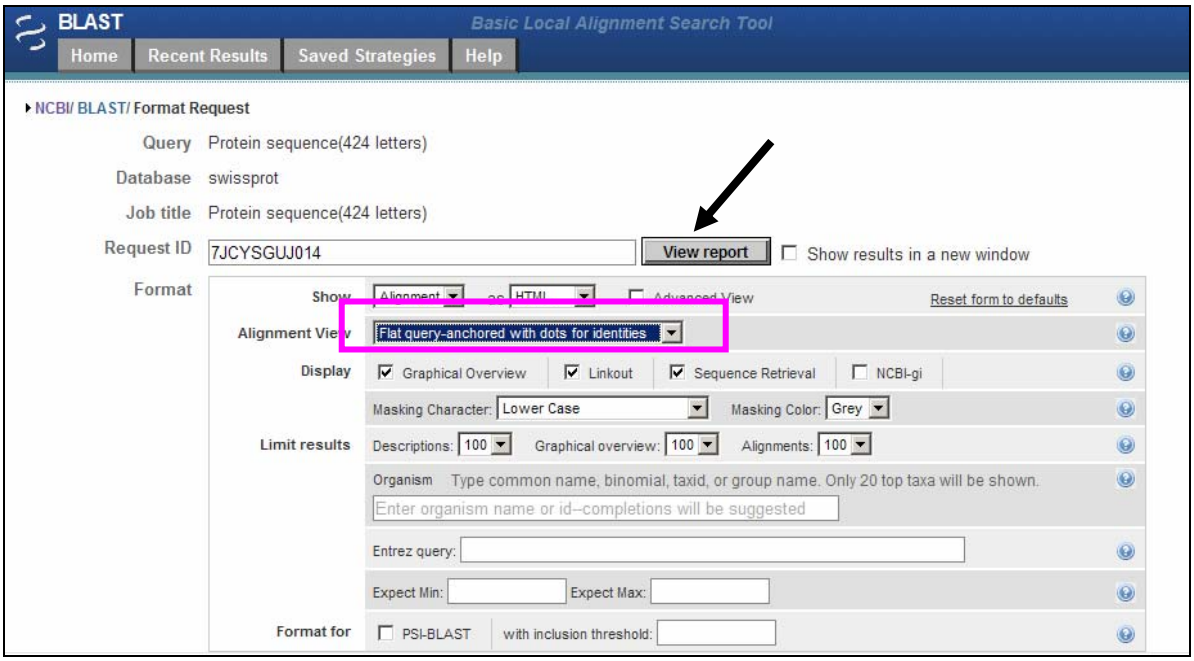
Distribution of 63 Blast Hits on the Query Sequence

Mouse over to see the defline, click to show alignments



Distance tree of results NEW

Sequences producing significant alignments:		Score (Bits)	E Value	
sp|Q5REB0|AATM_PONPY	Aspartate aminotransferase, mitochondria...	564	3e-160	
sp|P00505|AATM_HUMAN	Aspartate aminotransferase, mitochondria...	563	4e-160	G
sp|P05202|AATM_MOUSE	Aspartate aminotransferase, mitochondria...	562	1e-159	G
sp|Q4R559|AATM_MACFA	Aspartate aminotransferase, mitochondria...	561	1e-159	
sp|P00507|AATM_RAT	Aspartate aminotransferase, mitochondrial ...	561	1e-159	G
sp|P08907|AATM_HORSE	Aspartate aminotransferase, mitochondria...	554	3e-157	
sp|P00506|AATM_PIG	Aspartate aminotransferase, mitochondrial ...	552	9e-157	G
sp|P12344|AATM_BOVIN	Aspartate aminotransferase, mitochondria...	550	3e-156	G
sp|P00508|AATM_CHICK	Aspartate aminotransferase, mitochondria...	546	6e-155	G
sp|P46643|AAT1_ARATH	Aspartate aminotransferase, mitochondria...	447	5e-125	G
sp|P28011|AAT1_MEDSA	Aspartate aminotransferase 1 (Transaminase	434	5e-121	
sp|P46645|AAT2_ARATH	Aspartate aminotransferase, cytoplasmic	420	4e-117	G




```
24418724   273   ------V                                                  273

Query      251   ----DAQAVR------TFE---ADGH----D-FCL-AQ--SFAKNMGLYGERAGAFTVLC  289
75042478   257   ----..W...------H.I---EQ.I----N-V..-C.--.Y..........V....MV.  295
112983     257   ----..W...------H.I---EQ.I----N-V..-C.--.Y..........V....MV.  295
112984     257   ----..W...------H.I---EQ.I----N-V..-C.--.Y..........V.....V.  295
112987     257   ----..W...------H.I---EQ.I----N-V..-C.--.Y..........V.....V.  295
75075926   257   ----..W...------H.I---EQ.I----N-V..-C.--.Y..........V....MV.  295
112985     257   ----..W...------H.I---EQ.I----N-V..-C.--.Y..........V.....V.  295
1168261    257   ----..W...------H.I---EQ.I----N-V..-C.--.Y..........V.....V.  295
112981     250   ----..W.L.------H.I---EQ.I----.-VV.-S.--.Y................I.  288
112982     228   ----..W...------Y.I---EQ.I----N-V..-C.--.Y..........V....MV.  266
1168256    255   ----..KSI.------I.L---E...----H-IGI-S.--.Y........Q.V.CLS...  293
2506178    242   ----...P..------L.V---...G----E-LLV-..--.Y..........V..LSIVS  280
1168258    273   ----..KPI.------M.V---...G----E-CLV-..--.Y..........V..LSIV.  311
21542386   229   ----...S..------..V---...G----E-CLI-..--.Y..........V..LSIV.  267
112972     229   ----...S..------I.V---...G----E-CLA-..--.Y..........V..LSIV.  267
584706     231   ----...S..------M.V---...G----E-LLM-..--.Y..........V..LSIV.  269
112971     236   ----..W...------Y.V---SE.F----E-LFC-..--..S..F...N..V.NLS.VG  274
21542387   227   ----......------M.V---...G----E-CLI-..--.Y..........I.SL.IV.  265
75041219   237   ----..W.I.------Y.V---SE.F----E-.FC-..--..S..F...N..V.NL..VG  275
75076072   237   ----..W.I.------Y.V---SE.F----E-.FC-..--..S..F...N..V.NL..VG  275
5902703    237   ----..W.I.------Y.V---SE.F----E-.FC-..--..S..F...N..V.NL..VG  275
20532373   276   ----..AS..------L.A---ER.M----E-.FV-..--.YS..L...A..I..IN.V.  314
112979     277   ----..AS..------L.V---.R.L----E-VLV-..--.YS..L...A..I..IN.IS  315
112976     237   ----..W.I.------Y.V---SE.F----E-LFC-..--..S..F...N..V.NL..VA  275
122065118  237   ----..W.I.------Y.V---SE.F----E-LFC-..--..S..F...N..V.NL..VG  275
126302508  237   ----..W...------Y.V---SE.F----E-LFC-..--..S..F...N..V.NL..VA  275
122065117  237   ----..W.I.------Y.V---SE.F----E-LFC-..--..S..F...N..V.NL..VA  275
112975     237   ----..W.I.------Y.V---SE.F----E-LFC-..--..S..F...N..V.NL..VG  275
1703040    230   ----..Y...------LGVEKLSTVS----P-VFV-C.--.....A.M....V.C.HLAL  271
2492843    229   ----..W.I.------Y.V---DQ.M----E-MVV-S.--.....F...N..V.NL..VV  267
1168262    224   ----..YGL.------A.A---.NHK----E-LLV-.S--..S..F...N..V....LVA  262
```

COGs

Phylogenetic classification of proteins encoded in complete genomes

NCBI

Clusters of Orthologous Groups of proteins (COGs) were delineated by comparing protein sequences encoded in 43 complete
G consists of individual proteins or groups of paralogs from at least 3

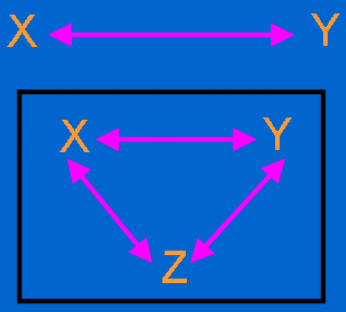
Code	Name	Proteins	in COGs
A	*Archaeoglobus fulgidus*	2420	1872
O	*Halobacterium sp. NRC-1*	2605	1701
M	*Methanococcus jannaschii*	1786	1330
	Methanobacterium thermoautotrophicum	1873	1388
P	*Thermoplasma acidophilum*	1482	1230
	Thermoplasma volcanium	1499	1243
K	*Pyrococcus horikoshii*	1800	1378
	Pyrococcus abyssi	1768	1456
Z	*Aeropyrum pernix*	1841	1178
Y	*Saccharomyces cerevisiae*	5955	2290
Q	*Aquifex aeolicus*	1560	1329
V	*Thermotoga maritima*	1858	1527
D	*Deinococcus radiodurans*	3187	2226
R	*Mycobacterium tuberculosis*	3927	2584
	Mycobacterium leprae	1605	1134
L	*Lactococcus lactis*	2267	1618
	Streptococcus pyogenes	1697	1211
B	*Bacillus subtilis*	4118	2870
	Bacillus halodurans	4066	2878
C	*Synechocystis*	3167	2159
E	*Escherichia coli K12*	4275	3414
	Escherichia coli O157	5315	3662
	Buchnera sp. APS	575	568
F	*Pseudomonas aeruginosa*	5567	4392
G	*Vibrio cholerae*	3835	2820
H	*Haemophilus influenzae*	1714	1542
	Pasteurella multocida	2015	1751

Help

COGnitor

Genome A
All proteins

Genome O
All proteins



NCBI

To search against the COGs database, click here:

COGs

Clusters of Orthologous Groups of proteins (COGs) were delineated by comparing protein sequences encoded in 43 complete genomes, representing 30 major phylogenetic lineages. Each COG consists of individual proteins or groups of paralogs from at least 3 lineages and thus corresponds to an ancient conserved domain. Use the COGnitor to compare the protein sequence to the COGs database.

Paste the FASTA formatted protein sequence from GenScan into the COGnitor input box and press the ***"compare to COGs"*** button. Click on the link to the highest-scoring COG and click on the disk icon to save the sequences in the COG to a local file on your desktop to be used as input to Multalin below. Drag this file from your desktop onto your "tools" browser window to display the sequences. Then copy and paste these into your notebook under "COGs FASTA Sequences".

COGs FASTA Sequences

Updated version

COGs

Phylogenetic classification of proteins encoded in complete genomes

Clusters of Orthologous Groups of proteins (COGs) were delineated by comparing protein sequences encoded in 43 complete gen
representing 30 major phylogenetic lineages. Each COG consists of individual proteins or groups of paralogs from at least 3 lineages
corresponds to an ancient conserved domain.

Science 1997 Oct 24;278(5338):631-7,
Nucleic Acids Res 2001 Jan 1; 29(1):22-28.

Help

COGnitor

Protein/Gene name:

Text search:

Code	Name	Proteins	in COGs
A	*Archaeoglobus fulgidus*	2420	1872
O	*Halobacterium sp. NRC-1*	2605	1701
M	*Methanococcus jannaschii*	1786	1330
	Methanobacterium thermoautotrophicum	1873	1388
P	*Thermoplasma acidophilum*	1482	1230
	Thermoplasma volcanium	1499	1243
K	*Pyrococcus horikoshii*	1800	1378
	Pyrococcus abyssi	1768	1456
Z	*Aeropyrum pernix*	1841	1178
Y	*Saccharomyces cerevisiae*	5955	2290
Q	*Aquifex aeolicus*	1560	1329
V	*Thermotoga maritima*	1858	1527
D	*Deinococcus radiodurans*	3187	2226
R	*Mycobacterium tuberculosis*	3927	2585
	Mycobacterium leprae	1605	1134
L	*Lactococcus lactis*	2267	1618
	Streptococcus pyogenes	1697	1211
B	*Bacillus subtilis*	4118	2870
	Bacillus halodurans	4066	2878
C	*Synechocystis*	3167	2159

Principal component analysis of genomes

List of COGs

Distribution

Co-occurrences

Phylogenetic patterns

Phylogenetic patterns search

Functional categories

J K L

D O M N P T

G C E F H I Q

R S

Pathways and functional systems

FTP

COGnitor

Compare your sequence to COG database

NCBI

compare to COGs | Clear | BeTs to 3 clades

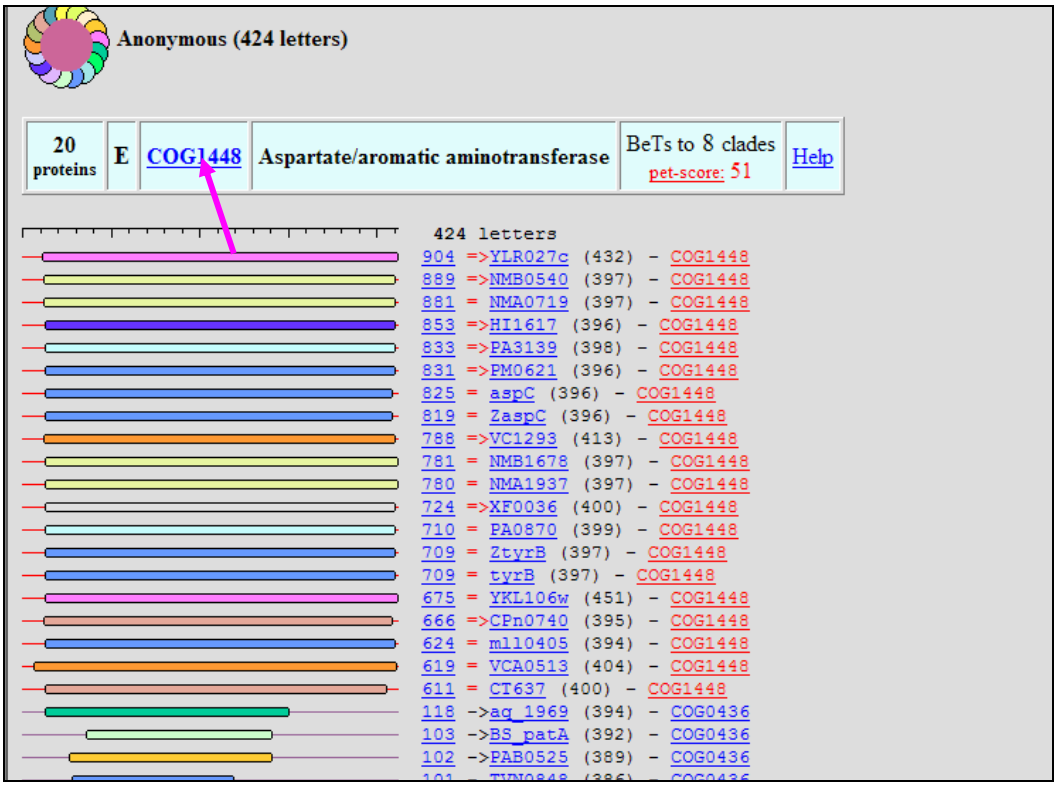
Help

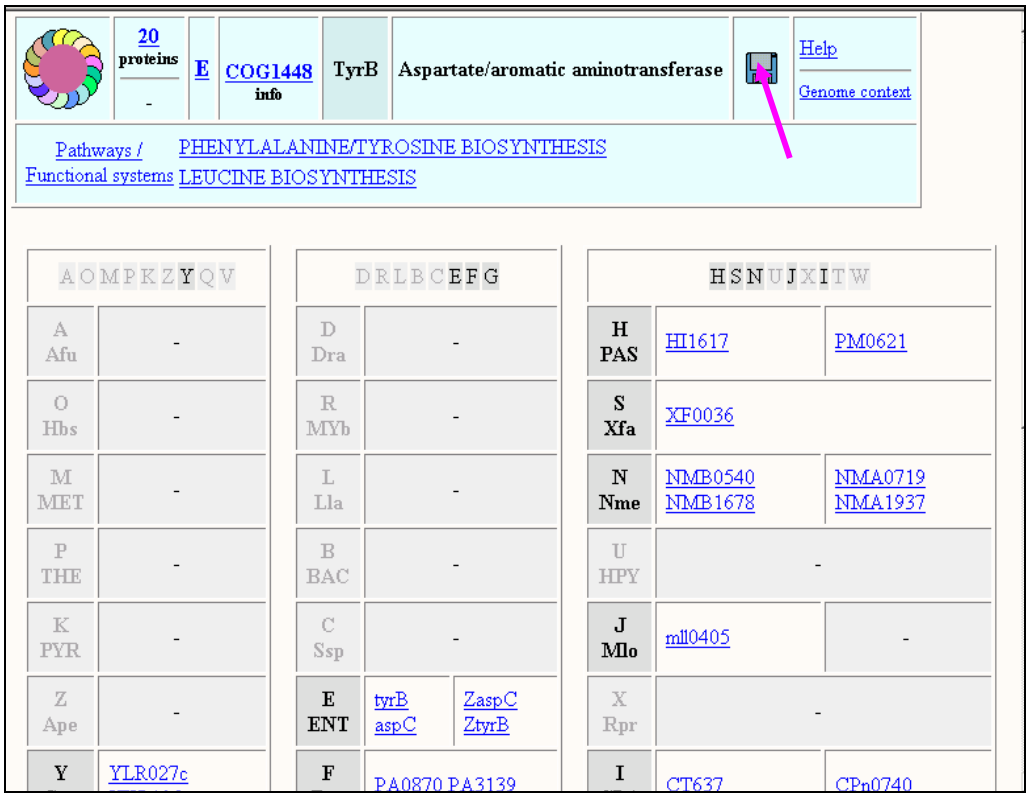
Example

Paste your sequence and press the button above.

```
MSQICKRGLLISNRLAPAALRCKSTWFSEVQMGPPDAILGVTEAFKKDTNPKKINLGAGA
YRDDNTQPFVLPSVREAEKRVVSRSLDKEYATIIGIPEFYNKAIELALGKGSKRLAAKHN
VTAQSISGTGALRIGAAFLAKFWQGNREIYIPSPSWGNHVAIFEHAGLPVNRYRYYDKDT
CALDFGGLIEDLKKIPEKSIVLLHACAHNPTGVDPTLEQWREISALVKKRNLYPFIDMAY
QGFATGDIDRDAQAVRTFEADGHDFCLAQSFAKNMGLYGERAGAFTVLCSDEEEAARVMS
QVKILIRGLYSNPPVHGARIAAEILNNEDLRAQWLKDVKLMADRIIDVRTKLKDNLIKLG
SSQNWDHIVNQIGMFCFTGLKPEQVQKLIKDHSVYLTNDGRVSMAGVTSKNVEYLAESIH
KVTK
```

Skip low-complexity filtering


Anonymous (424 letters)
20 proteins
E
COG1448
Aspartate/aromatic aminotransferase
BeTs to 8 clades
pet-score: 51
Help
424 letters
904 =>YLR027c (432) - COG1448
889 =>NMB0540 (397) - COG1448
881 = NMA0719 (397) - COG1448
853 =>HI1617 (396) - COG1448
833 =>PA3139 (398) - COG1448
831 =>PM0621 (396) - COG1448
825 = aspC (396) - COG1448
819 = ZaspC (396) - COG1448
788 =>VC1293 (413) - COG1448
781 = NMB1678 (397) - COG1448
780 = NMA1937 (397) - COG1448
724 =>XF0036 (400) - COG1448
710 = PA0870 (399) - COG1448
709 = ZtyrB (397) - COG1448
709 = tyrB (397) - COG1448
675 = YKL106w (451) - COG1448
666 =>CPn0740 (395) - COG1448
624 = mll0405 (394) - COG1448
619 = VCA0513 (404) - COG1448
611 = CT637 (400) - COG1448
118 ->aq_1969 (394) - COG0436
103 ->BS_patA (392) - COG0436
102 ->PAB0525 (389) - COG0436


20 proteins
E
COG1448
info
TyrB
Aspartate/aromatic aminotransferase
Help
Genome context
Pathways / Functional systems
PHENYLALANINE/TYROSINE BIOSYNTHESIS
LEUCINE BIOSYNTHESIS
A O M P K Z Y Q V
A Afu -
O Hbs -
M MET -
P THE -
K PYR -
Z Ape -
Y YLR027c
D R L B C E F G
D Dra -
R MYb -
L Lla -
B BAC -
C Ssp -
E ENT tyrB aspC ZaspC ZtyrB
F PA0870 PA3139
H S N U J X I T W
H PAS HI1617 PM0621
S Xfa XF0036
N Nme NMB0540 NMB1678 NMA0719 NMA1937
U HPY -
J Mlo mll0405 -
X Rpr -
I CT637 CPn0740

To generate a multiple sequence alignment, use:

MultAlin

Paste the sequences from your best–hit COG, saved in your "COGs FASTA Sequences" notebook area, into the input box of Multalin. Also paste in the protein sequence derived from GenScan to include your unknown sequence in this alignment and press the ***"Start Multalin!"*** button. Display these results in text form by clicking on the "–Results as a text page (msf) " link. Paste this Multalin display into your notebook.

Multalin Alignment

MultAlin

Multiple sequence alignment by Florence Corpet

Published research using this software should cite:
"Multiple sequence alignment with hierarchical clustering"
F. CORPET, 1988, Nucl. Acids Res., 16 (22), 10881-10890

INRA Home page | INRA L.G.C.-Auzeville | MultAlin

Sequence data

Cut and paste your sequences here below.

```
>08:28:21|GENSCAN_predicted_peptide_2|424_aa
MSQICKRGLLISNRLAPAALRCKSTWFSEVQMGPPDAILGVTEAFKKDTNPKKINLGAGA
YRDDNTQPFVLPSVREAEKRVVSRSLDKEYATIIGIPEFYNKAIELALGKGSKRLAAKHN
VTAQSISGTGALRIGAAFLAKFWQGNREIYIPSPSWGNHVAIFEHAGLPVNRYRYYDKDT
CALDFGGLIEDLKKIPEKSIVLLHACAHNPTGVDPTLEQWREISALVKKRNLYPFIDMAY
QGFATGDIDRDAQAVRTFEADGHDFCLAQSFAKNMGLYGERAGAFTVLCSDEEEAARVMS
QVKILIRGLYSNPPVHGARIAAEILNNEDLRAQWLKDVKLMADRIIDVRTKLKDNLIKLG
SSQNWDHIVNQIGMFCFTGLKPEQVQKLIKDHSVYLTNDGRVSMAGVTSKNVEYLAESIH
KVTK
>tyrB
MFQKVDAYAGDPILTLMERFKEDPRSDKVNLSIGLYYNEDGIIPQLQAVAEAEARLNAQPHGASLYLPME
```

(sample sequences)

or select a file: Browse... new

Sequence input format: Auto

```
                    1       10        20        30        40        50        60        70        80        90       100       110       120       130
                    |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
14:10:32|GENSCAN_pre MSQICKRGLLISNRLAPAALRCKSTWFSEVQMGPPDAILGVTEAFKKDTNPKKINLGAGAYRDDNTQPFVLPSVREAEKRVVS-RSLDKE--YATIIGIPEFYNKAIELALGKGSKR-----LAAKHNVT
          YLR027c     MNKRTQEYKNTRAIMS----ATLFNNIELLPPDALFGIKQRYGQDQRATKVDLGIGAYRDDNGKPWVLPSVKAAEKLIHNDSSYNHE--YLGITGLPSLTSNAAKIIFGTQSDA-----FQEDRVIS
                aspC                         MFENITAAPADPILGLADLFRADERPGKINLGIGVYKDETGKTPVLTSVKKAEQYL-LENETTKN--YLGIDGIPEFGRCTQELLFGKGSAL-----INDKRART
             HI1617                          MFEHIKAAPADPILGLGEAFKSETRENKINLGIGVYKDAQGTTPIMHAVKEAEKRL-FDKEKTKN--YLTIDGIADYNEQTKALLFGKDSEV-----IQSNRART
               tyrB                          MFQKVDAYAGDPILTLMERFKEDPRSDKVNLSIGLYYNEDGIIPQLQAVAEAEARLNAQPHGASL--YLPMEGLNCYRHAIAPLLFGADHPV-----LKQQRVAT
               CT637                       MSLFEQLPSFSPDSILGLAQAFQEDPREDKINLLLGTYEREKKRYGGFSSVRKAQS-VFFDDEKDKN--YLPIKGSSTFLEEMAALCFGE---------VDASRWVG
             CPn0740                       MSFFNHIPTFSPDAILGLQNVFFADKRPEKVNLVIGVYEHPQKRYGGLSCIRKAQT-VILEEEQNKS--YLPISGLQIFLDEMRELVFGA---------VDPSAIVG
              YKL106w          MLRTRLTNCSLWRPYYTSSLSRVPRAPPDKVLGLSEHFKKVKNVNKIDLTVGIYKDGWGKVTTFPSVAKAQKLIESHLELNKNLSYLPITGSKEFQENVMKFLFKESCPQFGPFYLAHDRISF
             Consensus .........................f..v...ppD.ilgl.#.%k.d.r..K!#L.iG.Y.d..g....l.s!.kA#k......e..k...Ylpi.G...f.......llfg..........l...r...

                    131     140       150       160       170       180       190       200       210       220       230       240       250       260
                    |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
14:10:32|GENSCAN_pre AQSISGTGALRIGAAFLAKFWQGNREIYIPSPSWGNHVAIFEHAGLPVNRYRYYDKDTCALDFGGLIEDLKKIPEKS---------IVLLHACAHNPTGVDPTLEQWREISALVKKRNLYPFIDMAYQGF
          YLR027c VQSLSGTGALHISAKFFSKFFP-DKLVYLSKPTWANNMAIFENQGLKTATYPYWANETKSLDLNGFLNAIQKAPEGS---------IFVLHSCAHNPTGLDPTSEQWVQIVDAIASKNHIALFDTAYQGF
                aspC AQTPGGTGALRVAADFLAKNTS-VKRVWVSNPSWPNHKSVFNSAGLEVREYAYYDAENHTLDFDALINSLNEAQAGD---------VVLFHGCCHNPTGIDPTLEQWQTLAQLSVEKGWLPLFDFAYQGF
             HI1617 VQSLGGTGALRIAAEFIKRQTK-AQNVWISTPTWPNHNAIFNAYGMTIREYRYYDAERKALDWEHLLEDLSQASEGD---------VVLLHGCCHNPTGIDPTPEQWQELAALSAKNGWLPLFDFAYQGL
               tyrB IQTLGGSGALKVGADFLKRYFP-ESGVWVSDPTWENHVAIFAGAGFEVSTYPWYDEATNGVRFNDLLATLKTLPARS---------IVLLHPCCHNPTGADLTNDQWDAVIEILKARELIPFLDIAYQGF
               CT637 VQAIGGTGALHLGASVYAN-ASLAGKVYIPSQTWGNHSRIFAHQGLALEYYPYYDQETKELDLQGLKAILRSAPETS---------LVLLHCCCHNPTGKDIPLSEWPEIITIIKERDLIPFFDMAYLGF
             CPn0740 FQSLGGTGALHLGARLLSV-AKGSGKVYYPEQTWSNHIRIFSQEGLEVIRYPYYSKEQKQLLFEPLIAFLKEVEKNS---------VILLHGCCHNPTGVDFTEDMWKELAILMKERELIPFFDTAYQGF
              YKL106w VQTLSGTGALAVAAKFLALFI--SRDIWIPDPSWANHKNIFQNNGFE-NIYRYSYYKDGQIDIDGWIEQLKTFAYNNQQENNKNPPCIILHACCHNPTGLDPTKEQWEKIIDTIYELKMYPIVDMAYQGL
             Consensus vQ.lgGtGAL...A.fla........!wip.ptW.NH..IF...Gle...Y.ygd..t..ld..gli..lk..p..s..........vlLH.CcHNPTG.Dpt.eqW..i.....e....p..D.AYqGf

                    261     270       280       290       300       310       320       330       340       350       360       370       380       390
                    |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
14:10:32|GENSCAN_pre ATGDIDRDAQAVRTFE---AD---GHDFCLAQSFAKNMGLYGERAGAFTVLCSDE---------EE----AARVMSQVKILIRGLYSNPPVHGARIAAEILNNEDLRAQWLKDVKLMADRIIDVRTKLKD
          YLR027c ATGDLDKDAYAVRLGVEKLST---VSPVFVCQSFAKNAGMYGERVGCFHLALTKQ---------AQNKTIKPAVTSQLAKIIRSEVSNPPAYGAKIVAKLLETPELTEQWHKDMVTMSSRITKMRHALRD
                aspC ARG-LEEDAEGLRAFA---AM---HKELIVASSYSKNFGLYNERVGACTLYAADS---------ET----VDRAFSQMKAAIRANYSNPPAHGASVVATILSNDALRAIWEQELTDMRQRIQRMRQLFVN
             HI1617 ANG-LDEDAYGLRAFA---AN---HKELLVASSFSKNFGLYNERVGAFTLVAENA---------EI----ASTSLTQVKSIIRTLYSNPASHGGATVATVLNDAQLRQEWENELTEMRERIKKMRHLFVQ
               tyrB GAG-MEEDAYAIRAIA---SA---GLPALVSNSFSKIFSLYGERVGGLSVMCEDA---------ER----AGRVLGQLKATVRRNYSSPPNFGAQVVAAVLNDEALKASWLAEVEEMRTRILAMRQELVK
               CT637 ASG-IEEDRRPVRLCI---EA---GVTTFVAGGASKIFSLYGSRVGFFGAIHQDK---------LD----LNRILSFLEEQIRGEYSSPAREGYAIVTSILSNPYLRQEWELELNGIRQSLEEIRSSFVI
             CPn0740 AHG-IELDRKPIEIFI---SE---GNTVLVAASSSKNFALYGERVGYFAVHSTFT---------DE----LVKIHSFLEEKIRGEYSSPQRWGVEIVSTILSNPYLKEEWQSELNFIRESLGKMRTRFVQ
              YKL106w ESGNLLKDAYLLRLCLNVNKYPNWSNGIFLCQSFAKNMGLYGERVGSLSVITPATANNGKFNPLQQKNSLQQNIDSQLKKIVRGMYSSPPGYGSRVVNVVLSDFKLKQQWFKDVDFMVQRLHHVRQEMFD
             Consensus a.G.l..Day..r................v..sfsKnfg$YgeRvG.f.v.........................sqlk..!Rg.ySsPp..G..!va..Ls...L...W..#v..mr.ri..mR...v.

                    391     400       410       420       430       440       450       460       468
                    |--------+---------+---------+---------+---------+---------+---------+-------|
14:10:32|GENSCAN_pre NLIKLGSSQNWDHIVNQIGMFCFTGLKPEQVQKLIKDHSVYLTNDGRVSMAGVTSKNVEYLAESIHKVTK
          YLR027c HLVKLGTPGNWDHIVNQCGMFSFTGLTPQMVKRLEETHRVYLVASGRASIAGLNQGNVEYVAKAIDEVVRFYTIEAKL
                aspC TLQEKGANRDFSFIIKQNGMFSFSGLTKEQVLRLREEFGVYAVASGRVNVAGMTPDNMAPLCEAIVAYL
             HI1617 LLKEYGAEQDFSFIIEQNGMFSFSGLTGEQVDRLKNEFAIYAVRSGRINVAGITEDNIRYLCESIVKVL
               tyrB VLSTEMPERNFDYLLNQRGMFSYTGLSAAQVDRLREEFGVYLIASGRMCVAGLNTANVQRVAKAFAAVM
               CT637 AMRNV-AGHSFDFIASQKGFFGYPGFSKEQVLFLREELGIYTTAGGRFNLNGITDKNINRVTHGFAQAYEYPRSVS
             CPn0740 ALRKV-AGHTFDFLLSQHGFFAYPGFSDKQVLFLREQHAVYTTAGGRMNLNGITEKNIDHVVQSFIQAYEL
              YKL106w RL----GWPDLVNFAQQHGMFYYTRFSPKQVEILRNNYFVYLTGDGRLSLSGVNDSNVDYLCESLEAVSKMDKLA
             Consensus .$........fd.i..Q.GmF.%tgls..qV..Lre...!Ylta.GR...aG.n..N!.yl.es...v..........
```

```
      Consensus levels: high=90% low=50%
Consensus symbols:
 ! is anyone of IV
 $ is anyone of LM
 % is anyone of FY
 # is anyone of NDQEBZ//
```

```
                        251                                                300
14:13:11|GENSCAN_pre  PFIDMAYQGF ATGDIDRDAQ AVRTFE.... ..ADGHDFCL AQSFAKNMGL
             YLR027c  ALFDTAYQGF ATGDLDKDAY AVRLGV...E KLSTVSPVFV CQSFAKNAGM
                tyrB  PFLDIAYQGF GAG.MEEDAY AIRAIA...S ..AGLP.ALV SNSFSKIFSL
               ZtyrB  PFLDIAYQGF GAG.MEEDAY AIRAIA...S ..AGLP.ALV SNSFSKIFSL
             NMB1678  PFMDIAYQGF GGD.LDSDAY AVRKAV...E ..MELP.LFV SNSFSKNLSL
             NMA1937  PFMDIAYQGF GGD.LDSDAY AVRKAV...E ..MELP.LFV SNSFSKNLSL
              PA3139  PFLDIAYQGF GNG.IEEDAA AVRLFA...Q ..SGLS.FFV SSSFSKSFSL
              XF0036  PCIDLAYQGF NQG.IDADAY AIRLLA...E ..EGISNYVV ANSYSKSFSL
                aspC  PLFDFAYQGF ARG.LEEDAE GLRAFA...A ..M.HKELIV ASSYSKNFGL
               ZaspC  PLFDFAYQGF ARG.LEEDAE GLRAFA...A ..M.HKELIV ASSYSKNFGL
              VC1293  PLFDFAYQGF ASG.VEEDAA GLRIFA...K ..Y.NSEILV ASSFSKNFGL
              HI1617  PLFDFAYQGL ANG.LDEDAY GLRAFA...A ..N.HKELLV ASSFSKNFGL
              PM0621  PLFDFAYQGF ANG.LEEDAF GLRTFA...K ..N.HKELLV ASSYSKNFGL
             NMB0540  PLFDFAYQGF GNG.LEEDAY GLRVFL...K ..H.NTELLI ASSYSKNFGM
             NMA0719  PLFDFAYQGF GNG.LEEDAY GLRVFL...K ..H.NTELLI ASSYSKNFGM
             VCA0513  PFVDIAYQGF GDG.LEQDAQ GLRYMA...E ..R.MEEMLI TTSCSKNFGL
             mll0405  PFVDIAYQGF GDG.LEADAL GLRLLA...A ..K.VPEMVV ASSCSKNFAV
              PA0870  PLIDFAYQGF GDG.LEEDAW AVRLFA...G ..E.LPEVLV TSSCSKNFGL
               CT637  PFFDMAYLGF ASG.IEEDRR PVRLCI...E ...AGVTTFV AGGASKIFSL
             CPn0740  PFFDTAYQGF AHG.IELDRK PIEIFI...S ...EGNTVLV AASSSKNFAL
             YKL106w  PIVDMAYQGL ESGNLLKDAY LLRLCLNVNK YPNWSNGIFL CQSFAKNMGL
```

```
           Consensus  Pf.D.AYQGf ..G.le.Da. ..Rl.a.... .........v a.S.sKnfg$

                      301                                                350
14:13:11|GENSCAN_pre  YGERAGAFTV LCSDE..... ....EE.... AARVMSQVKI LIRGLYSNPP
             YLR027c  YGERVGCFHL ALTKQ..... ....AQNKTI KPAVTSQLAK IIRSEVSNPP
                tyrB  YGERVGGLSV MCEDA..... ....EA.... AGRVLGQLKA TVRRNYSSPP
               ZtyrB  YGERVGGLSV LCEDA..... ....EA.... AGRVLGQLKA TVRRNYSSPP
             NMB1678  YGERVGGLSV VCPNK..... ....EE.... ADLVFGQLKF TVRRIYSSPP
             NMA1937  YGERVGGLSV VCPNK..... ....EE.... ADLVFGQLKF TVRRIYSSPP
              PA3139  YGERVGALSI VTESR..... ....DE.... SARVLSQVKR VIRTNYSNPP
              XF0036  YGERVGGLSI VASNT..... ....EQ.... AQAIQSQVKR IIRTIYSSPS
                aspC  YNERVGACTL VAADS..... ....ET.... VDRAFSQMKA AIRANYSNPP
               ZaspC  YNERVGACTL VAADS..... ....ET.... VDRAFSQMKA AIRANYSNPP
              VC1293  YNERVGAFTL VAPST..... ....TV.... AETAFSQVKA IIRSIYSNPP
              HI1617  YNERVGAFTL VAENA..... ....EI.... ASTSLTQVKS IIRTLYSNPA
              PM0621  YSERVGAFTL VAETE..... ....QI.... AATALTQVKT IIRTLYSNPA
             NMB0540  YNERVGAFTL VAEDE..... ....ET.... AARAHSQVKT IIRTLYSNPA
             NMA0719  YNERVGAFTL VAEDE..... ....AT.... AARAHSQVKT IIRTLYSNPA
             VCA0513  YRERTGAAIV IGKNQ..... ....QE.... VTNARGKMLT LARSTYTMPP
             mll0405  YRDRVGAAMV LARDS..... ....AQ.... ADVAMSQMLS AARAMYSMPP
              PA0870  YRDRVGALIV CAQNA..... ....EK.... LTDLRSQLAF LARNLWSTPP
               CT637  YGSRVGFFGA IHQDK..... ....LD.... LNRILSFLEE QIRGEYSSPA
             CPn0740  YGERVGYFAV HSTFT..... ....DE.... LVKIHSFLEE KIRGEYSSPQ
             YKL106w  YGERVGSLSV ITPATANNGK FNPLQQKNSL QQNIDSQLKK IVRGMYSSPP
           Consensus  Yg#RvGa..v .......... .......... .....sqlk. .iR..yS.Pp
```

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

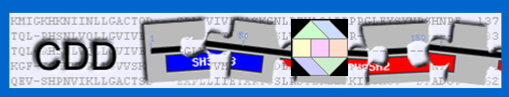
Conserved Domain

- recurring unit in molecular evolution, whose extents can be determined by sequence and structure analysis
- performs a particular function
- represented as a multiple local sequence alignment of proteins containing the domain

NCBI

Conserved Domain Database

Pfam SMART COG KOG PRK



Curated CDs

- A position-specific scoring matrix (PSSM) is calculated
- CD-Search can be used to search against the PSSMs
- Manual curation of CDs has begun

NCBI

To search for protein domains and view a model structure for your protein, click here:

NCBI's Conserved Domain Search allows you to match your protein sequence to conserved protein domains in the Conserved Domain Database, generate a multiple sequence alignment based on this match, and explore 3D modeling templates for your sequence.
Paste your protein sequence from GenScan into the CD-Search query box and run the search. From the search results page, generate a multiple sequence alignment for the top 10 sequences representitive of the conserved domain hit by clicking on the cartoon of the domain. To view a structure with Cn3D, click on the "+Structure" link, use the listbox to specify "up to 5" sequences and invoke Cn3D with a display of a 3D modeling template, and a multiple sequence alignment including your query sequence, by pressing the "Show Structure" button. Residues identical in your sequence and the structural template are shown in red. Locate the Prosite Motif you found earlier within the Cn3D alignment window by using View--Find Pattern. Use Style--Annotate from the Cn3D window to color the highlighted residues and show their side chains.

NCBI

Conserved Domains

HOME | SEARCH | SITE MAP

NewSearch | PubMed | Nucleotide | Protein | Structure | Taxonomy | Help

Search Conserved Domains on a protein

Search against database: CDD -- 23523 PSSMs

Enter **Protein** Query as Accession, Gi, or Sequence in FASTA format

```
>14:42:25|GENSCAN_predicted_peptide_2|424_aa
MSQICKRGLLISNRLAPAALRCKSTWFSEVQMGPPDAILGVTEAFKKDTNPKKINLGAGA
YRDDNTQPFVLPSVREAEKRVVSRSLDKEYATIIGIPEFYNKAIELALGKGSKRLAAKHN
VTAQSISGTGALRIGAAFLAKFWQGNREIYIPSPSWGNHVAIFEHAGLPVNRYRYYDKDT
CALDFGGLIEDLKKIPEKSIVLLHACAHNPTGVDPTLEQWREISALVKKRNLYPFIDMAY
QGFATGDIDRDAQAVRTFEADGHDFCLAQSFAKNMGLYGERAGAFTVLCSDEEEAARVMS
QVKILIRGLYSNPPVHGARIAAEILNNEDLRAQWLKDVKLMADRIIDVRTKLKDNLIKLG
SSQNWDHIVNQIGMFCFTGLKPEQVQKLIKDHSVYLTNDGRVSMAGVTSKNVEYLAESIH
KVTK
```

Submit Query | Reset | Force live search

Advanced search options

Maximal hits 100 | Expect Value threshold 0.01 | Apply Low complexity filter

Retrieve previous search with RID# | Retrieve

Help | Disclaimer | Write to the Help Desk

NCBI | NLM | NIH

NCBI

Conserved Domains

HOME | SEARCH | SITE MAP

NewSearch | PubMed | Nucleotide | Protein | Structure | Taxonomy | Help

Query sequence: [(local sequence)lcl|14:42:25|GENSCAN_predicted_peptide_2|424_aa]

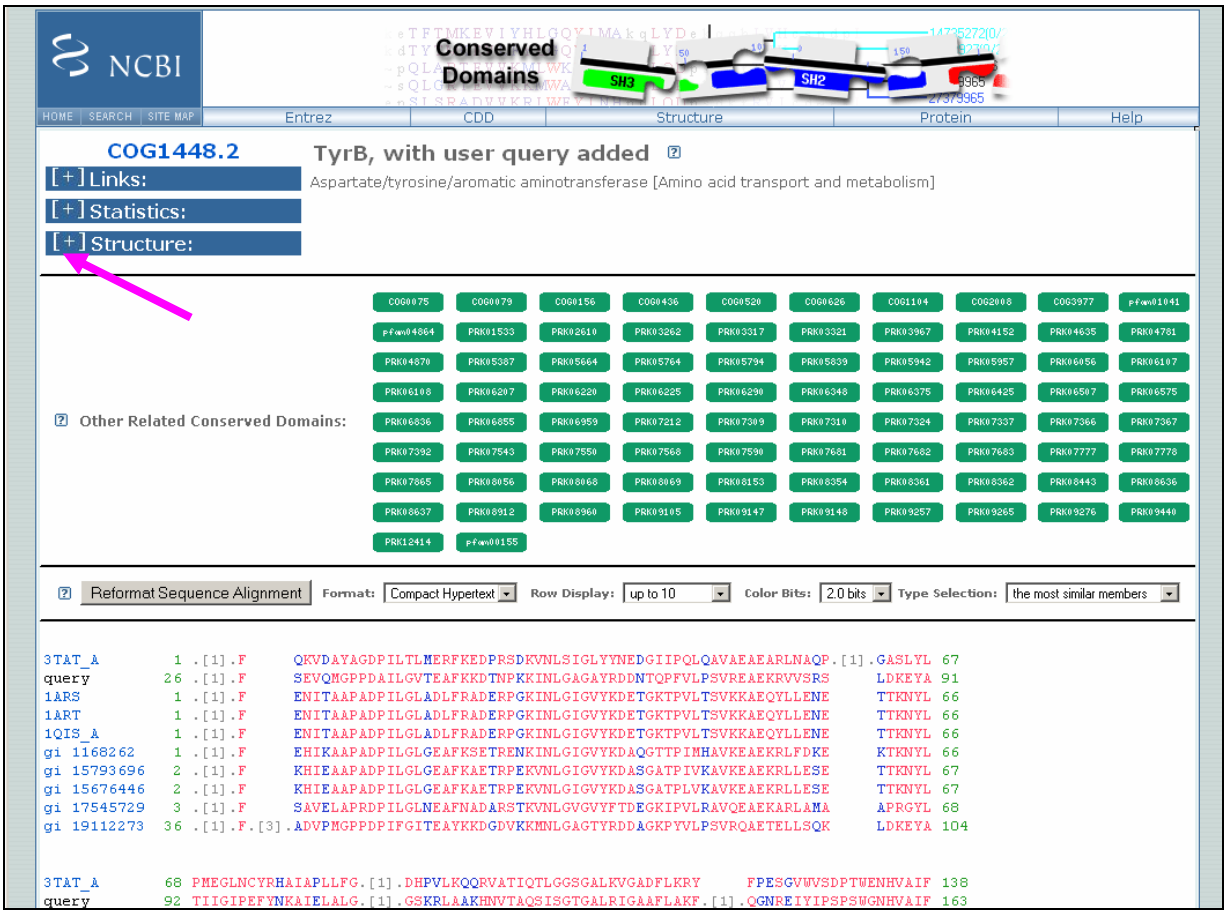
Concise Result | Full Result | Show Search Information

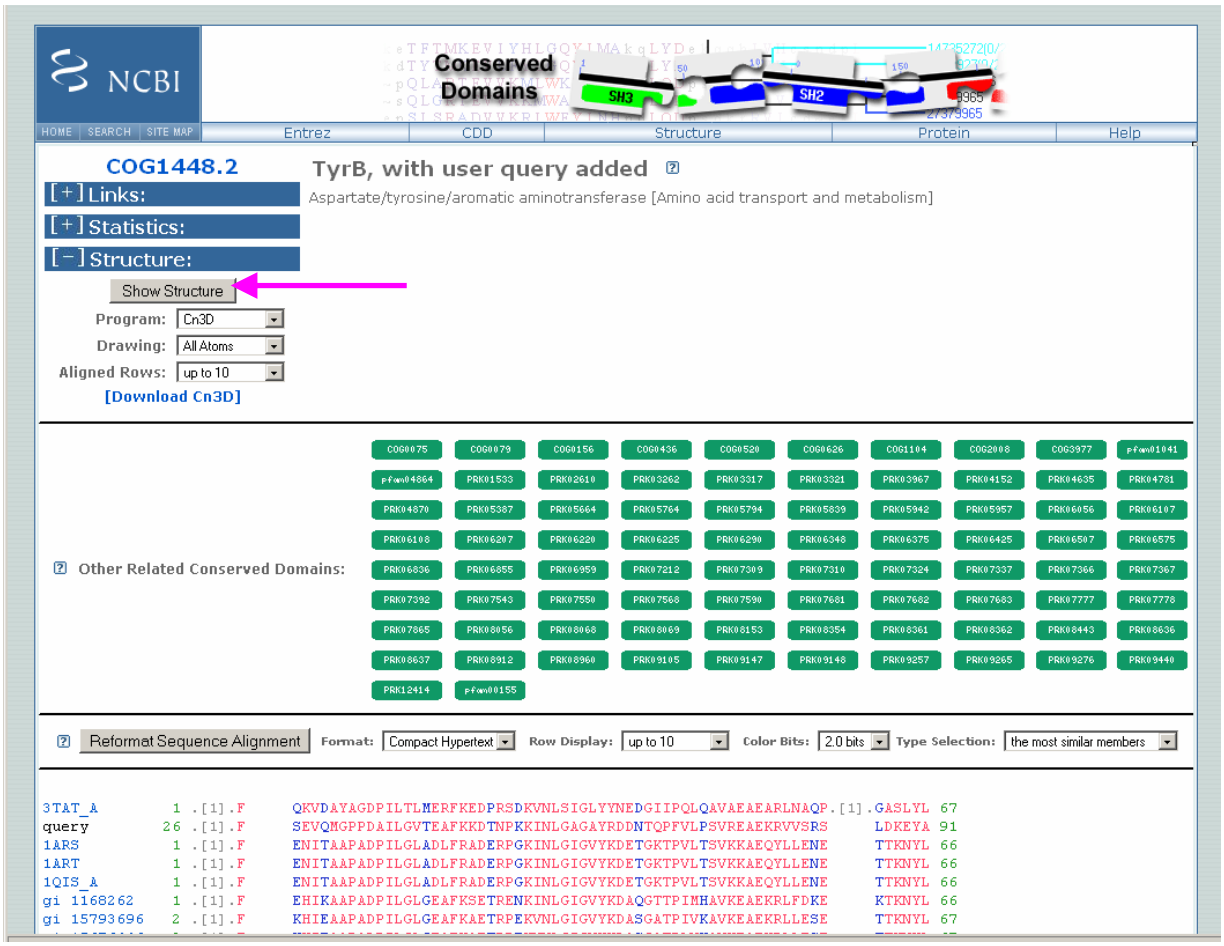
1 75 150 225 300 375 424

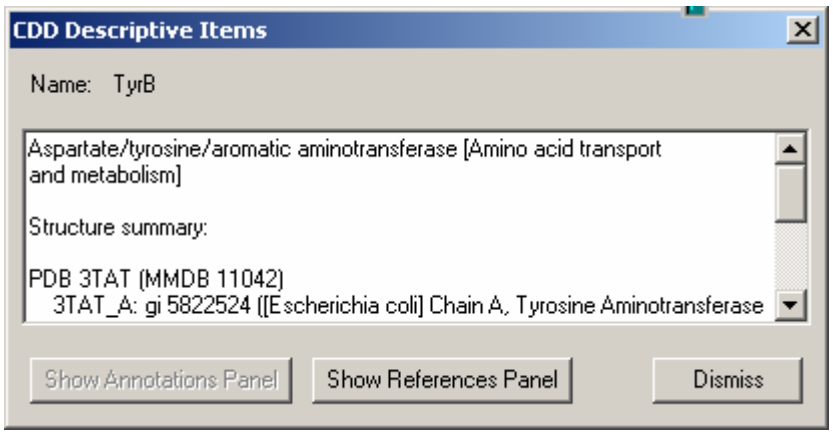
TyrB
PRK09257
Aminotran_1_2
COG0436
PRK08637
PRK07568
HisC
PRK07590
PRK06348
PRK08362
PRK06836
PRK05957
PRK08636
PRK07550
PRK05794
PRK05764
PRK07366
ARO8

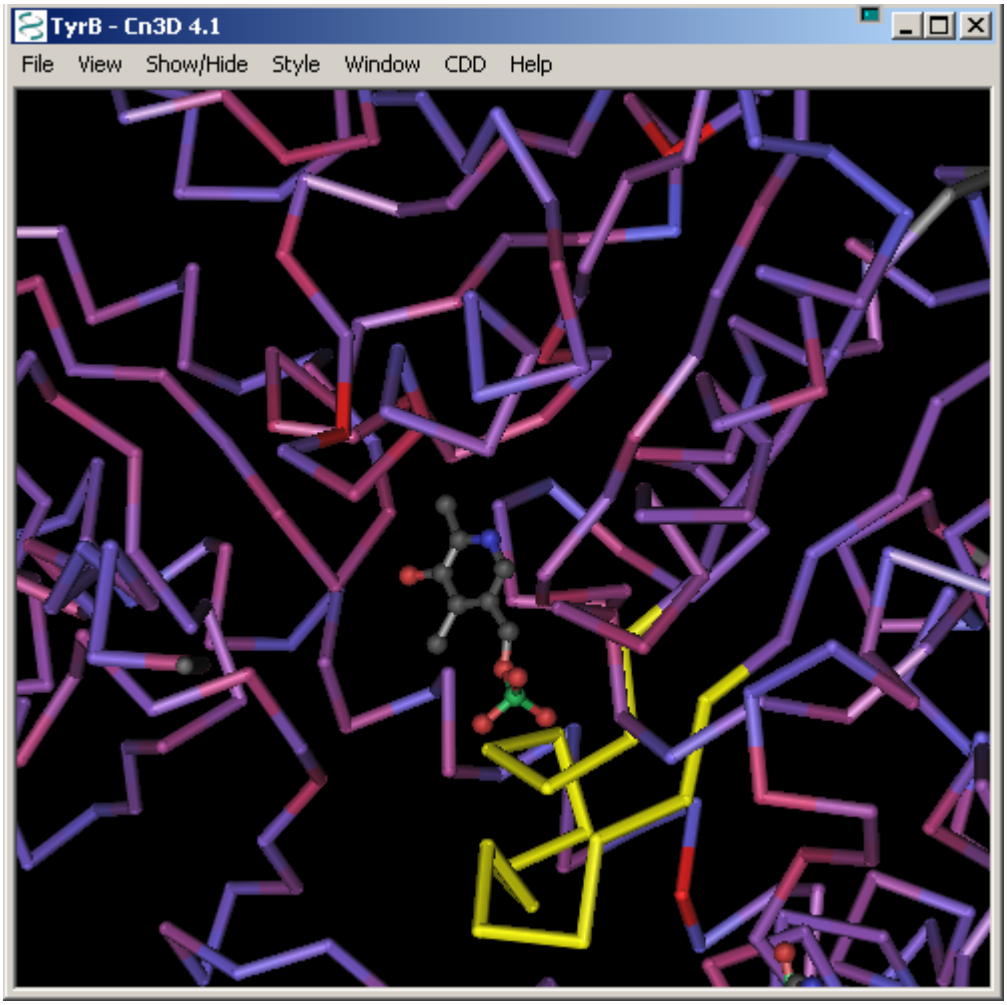
Descriptions

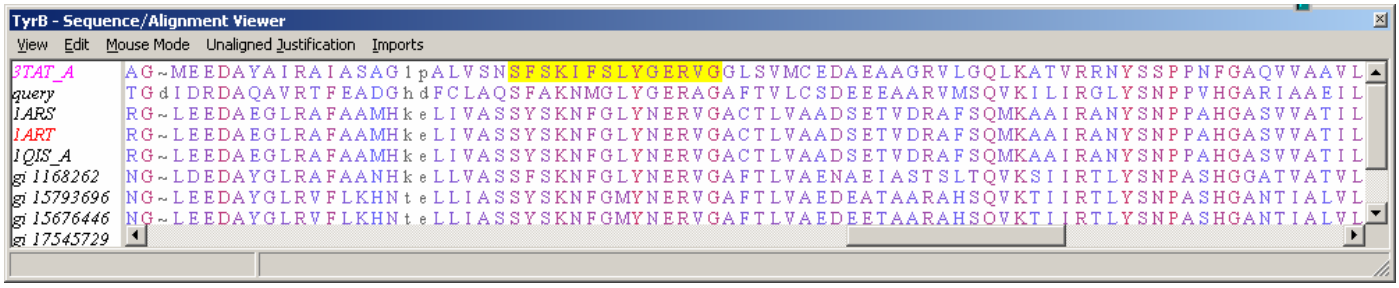
Title	PssmId	Multi-Dom	E-value
[+]COG1448, TyrB, Aspartate/tyrosine/aromatic aminotransferase [Amino acid transport and ...	31037	No	4e-147
[+]PRK09257, PRK09257, aromatic-amino-acid aminotransferase.	77018	No	1e-142
[+]pfam00155, Aminotran_1_2, Aminotransferase class I and II..	79492	No	5e-75
[+]COG0436, COG0436, Aspartate/tyrosine/aromatic aminotransferase [Amino acid transport a...	30785	No	2e-16
[+]PRK08637, PRK08637, hypothetical protein.	76749	No	2e-10
[+]PRK07568, PRK07568, aspartate aminotransferase.	76171	No	4e-06
[+]COG0079, HisC, Histidinol-phosphate/aromatic aminotransferase and cobyric acid decarbo...	30428	No	1e-05
[+]PRK07590, PRK07590, L,L-diaminopimelate aminotransferase.	76190	No	3e-05

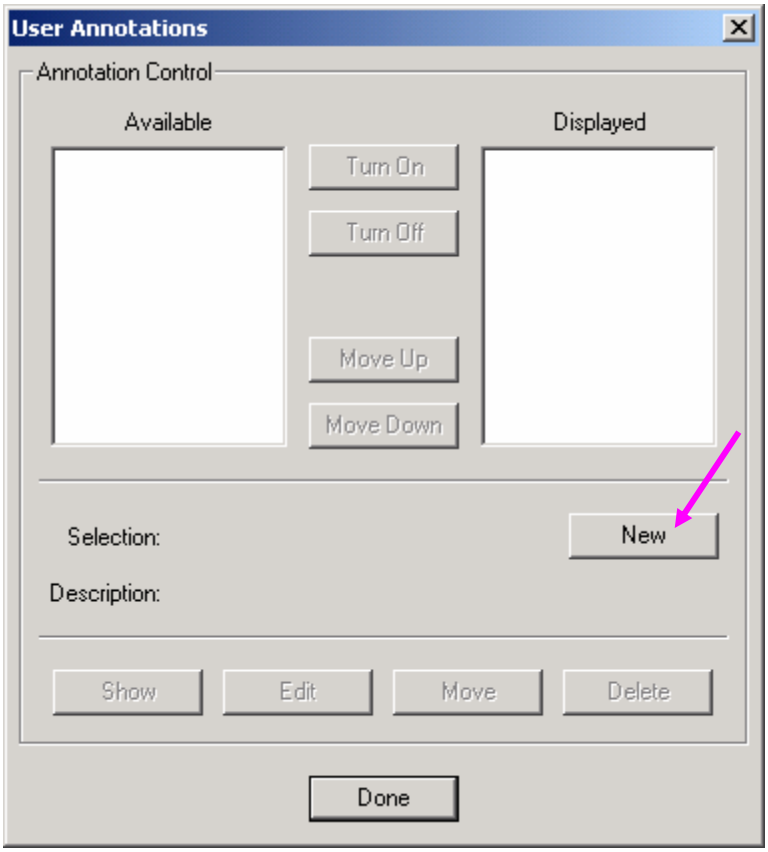

NCBI
Conserved Domains
SH3
SH2
HOME
SEARCH
SITE MAP
Entrez
CDD
Structure
Protein
Help
COG1448.2
TyrB, with user query added
Aspartate/tyrosine/aromatic aminotransferase [Amino acid transport and metabolism]
[+] Links:
[+] Statistics:
[+] Structure:
Other Related Conserved Domains:
Reformat Sequence Alignment
Format: Compact Hypertext
Row Display: up to 10
Color Bits: 2.0 bits
Type Selection: the most similar members
3TAT_A 1 .[1].F QKVDAYAGDPILTLMERFKEDPRSDKVNLSIGLYYNEDGIIPQLQAVAEAEARLNAQP.[1].GASLYL 67
query 26 .[1].F SEVQMGPPDAILGVTEAFKKDTNPKKINLGAGAYRDDNTQPFVLPSVREAEKRVVSRS LDKEYA 91
1ARS 1 .[1].F ENITAAPADPILGLADLFRADERPGKINLGIGVYKDETGKTPVLTSVKKAEQYLLENE TTKNYL 66
1ART 1 .[1].F ENITAAPADPILGLADLFRADERPGKINLGIGVYKDETGKTPVLTSVKKAEQYLLENE TTKNYL 66
1QIS_A 1 .[1].F ENITAAPADPILGLADLFRADERPGKINLGIGVYKDETGKTPVLTSVKKAEQYLLENE TTKNYL 66
gi 1168262 1 .[1].F EHIKAAPADPILGLGEAFKSETRENKINLGIGVYKDAQGTTPIMHAVKEAEKRLFDKE KTKNYL 66
gi 15793696 2 .[1].F KHIEAAPADPILGLGEAFKAETRPEKVNLGIGVYKDASGATPIVKAVKEAEKRLLESE TTKNYL 67
gi 15676446 2 .[1].F KHIEAAPADPILGLGEAFKAETRPEKVNLGIGVYKDASGATPLVKAVKEAEKRLLESE TTKNYL 67
gi 17545729 3 .[1].F SAVELAPRDPILGLNEAFNADARSTKVNLGVGVYFTDEGKIPVLRAVQEAEKARLAMA APRGYL 68
gi 19112273 36 .[1].F.[3].ADVPMGPPDPIFGITEAYKKDGDVKKMNLGAGTYRDDAGKPYVLPSVRQAETELLSQK LDKEYA 104
3TAT_A 68 PMEGLNCYRHAIAPLLFG.[1].DHPVLKQQRVATIQTLGGSGALKVGADFLKRY FPESGVWVSDPTWENHVAIF 138
query 92 TIIGIPEFYNKAIELALG.[1].GSKRLAAKHNVTAQSISGTGALRIGAAFLAKF.[1].QGNREIYIPSPSWGNHVAIF 163

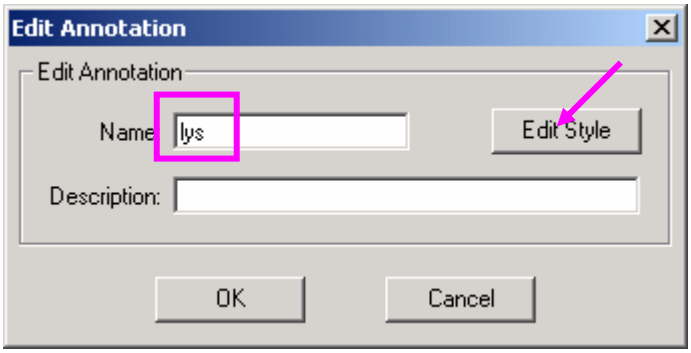

NCBI
Conserved Domains
SH3
SH2
HOME SEARCH SITE MAP
Entrez
CDD
Structure
Protein
Help
COG1448.2
TyrB, with user query added
Aspartate/tyrosine/aromatic aminotransferase [Amino acid transport and metabolism]
[+] Links:
[+] Statistics:
[−] Structure:
Show Structure
Program: Cn3D
Drawing: All Atoms
Aligned Rows: up to 10
[Download Cn3D]
Other Related Conserved Domains:
Reformat Sequence Alignment
Format: Compact Hypertext
Row Display: up to 10
Color Bits: 2.0 bits
Type Selection: the most similar members
3TAT_A 1 .[1].F QKVDAYAGDPILTLMERFKEDPRSDKVNLSIGLYYNEDGIIPQLQAVAEAEARLNAQP.[1].GASLYL 67
query 26 .[1].F SEVQMGPPDAILGVTEAFKKDTNPKKINLGAGAYRDDNTQPFVLPSVREAEKRVVSRS LDKEYA 91
1ARS 1 .[1].F ENITAAPADPILGLADLFRADERPGKINLGIGVYKDETGKTPVLTSVKKAEQYLLENE TTKNYL 66
1ART 1 .[1].F ENITAAPADPILGLADLFRADERPGKINLGIGVYKDETGKTPVLTSVKKAEQYLLENE TTKNYL 66
1QIS_A 1 .[1].F ENITAAPADPILGLADLFRADERPGKINLGIGVYKDETGKTPVLTSVKKAEQYLLENE TTKNYL 66
gi 1168262 1 .[1].F EHIKAAPADPILGLGEAFKSETRENKINLGIGVYKDAQGTTPIMHAVKEAEKRLFDKE KTKNYL 66
gi 15793696 2 .[1].F KHIEAAPADPILGLGEAFKAETRPEKVNLGIGVYKDASGATPIVKAVKEAEKRLLESE TTKNYL 67

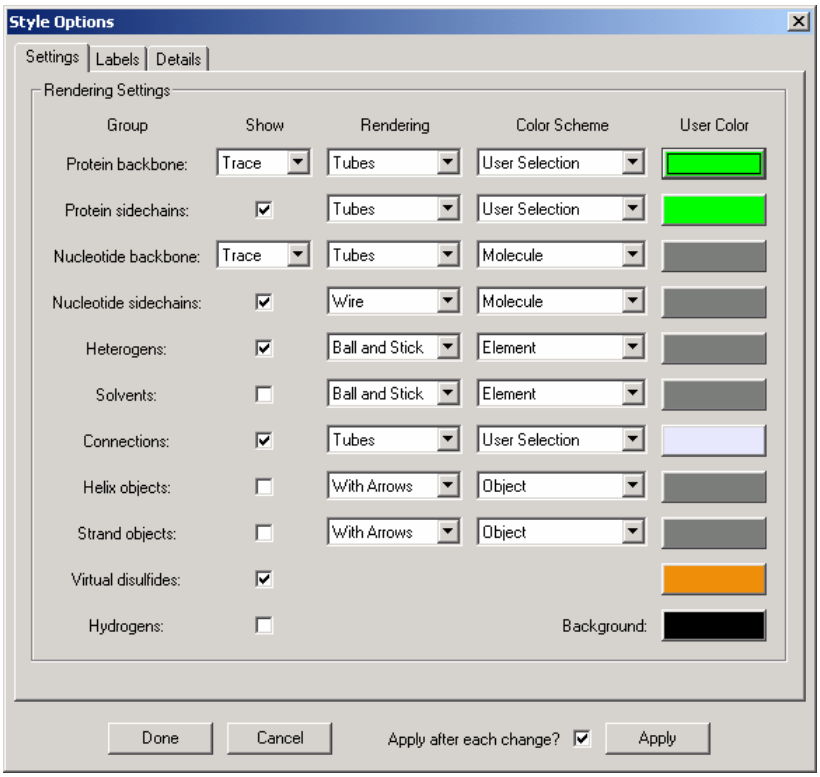
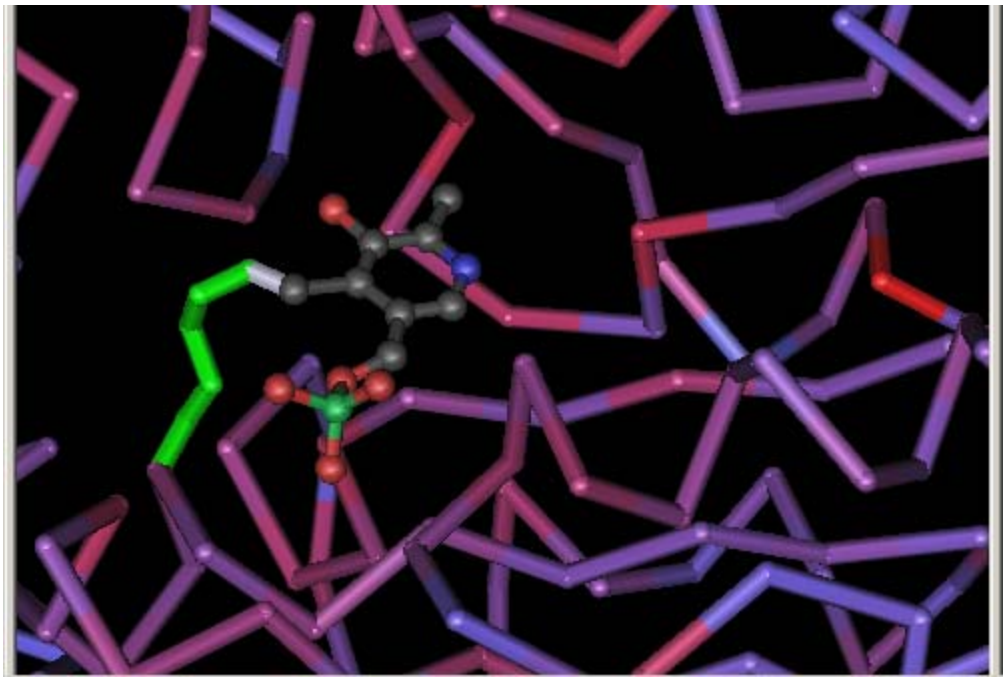

CDD Descriptive Items
Name: TyrB
Aspartate/tyrosine/aromatic aminotransferase [Amino acid transport and metabolism]
Structure summary:
PDB 3TAT (MMDB 11042)
3TAT_A: gi 5822524 ([Escherichia coli] Chain A, Tyrosine Aminotransferase
Show Annotations Panel
Show References Panel
Dismiss


TyrB - Cn3D 4.1
File View Show/Hide Style Window CDD Help


TyrB - Sequence/Alignment Viewer
View Edit Mouse Mode Unaligned Justification Imports
3TAT_A AG~MEEDAYAIRAIASAG1pALVSNSFSKIFSLYGERVGGLSVMCEDAEAAGRVLGQLKATVRRNYSSPPNFGAQVVAAVL
query TGdIDRDAQAVRTFEADGhdFCLAQSFAKNMGLYGERAGAFTVLCSDEEEAARVMSQVKILIRGLYSNPPVHGARIAAEIL
1ARS RG~LEEDAEGLRAFAAMHkeLIVASSYSKNFGLYNERVGACTLVAADSETVDRAFSQMKAAIRANYSNPPAHGASVVATIL
1ART RG~LEEDAEGLRAFAAMHkeLIVASSYSKNFGLYNERVGACTLVAADSETVDRAFSQMKAAIRANYSNPPAHGASVVATIL
1QIS_A RG~LEEDAEGLRAFAAMHkeLIVASSYSKNFGLYNERVGACTLVAADSETVDRAFSQMKAAIRANYSNPPAHGASVVATIL
gi 1168262 NG~LDEDAYGLRAFAANHkeLLVASSFSKNFGLYNERVGAFTLVAENAEIASTSLTQVKSIIRTLYSNPASHGGATVATVL
gi 15793696 NG~LEEDAYGLRVFLKHNteLLIASSYSKNFGMYNERVGAFTLVAEDEATAARAHSQVKTIIRTLYSNPASHGANTIALVL
gi 15676446 NG~LEEDAYGLRVFLKHNteLLIASSYSKNFGMYNERVGAFTLVAEDEETAARAHSQVKTIIRTLYSNPASHGANTIALVL
gi 17545729


User Annotations
Annotation Control
Available
Displayed
Turn On
Turn Off
Move Up
Move Down
Selection:
New
Description:
Show
Edit
Move
Delete
Done


Edit Annotation
Edit Annotation
Name
lys
Edit Style
Description:
OK
Cancel


Style Options
Settings
Labels
Details
Rendering Settings
Group
Show
Rendering
Color Scheme
User Color
Protein backbone:
Trace
Tubes
User Selection
Protein sidechains:
Tubes
User Selection
Nucleotide backbone:
Trace
Tubes
Molecule
Nucleotide sidechains:
Wire
Molecule
Heterogens:
Ball and Stick
Element
Solvents:
Ball and Stick
Element
Connections:
Tubes
User Selection
Helix objects:
With Arrows
Object
Strand objects:
With Arrows
Object
Virtual disulfides:
Hydrogens:
Background:
Done
Cancel
Apply after each change?
Apply