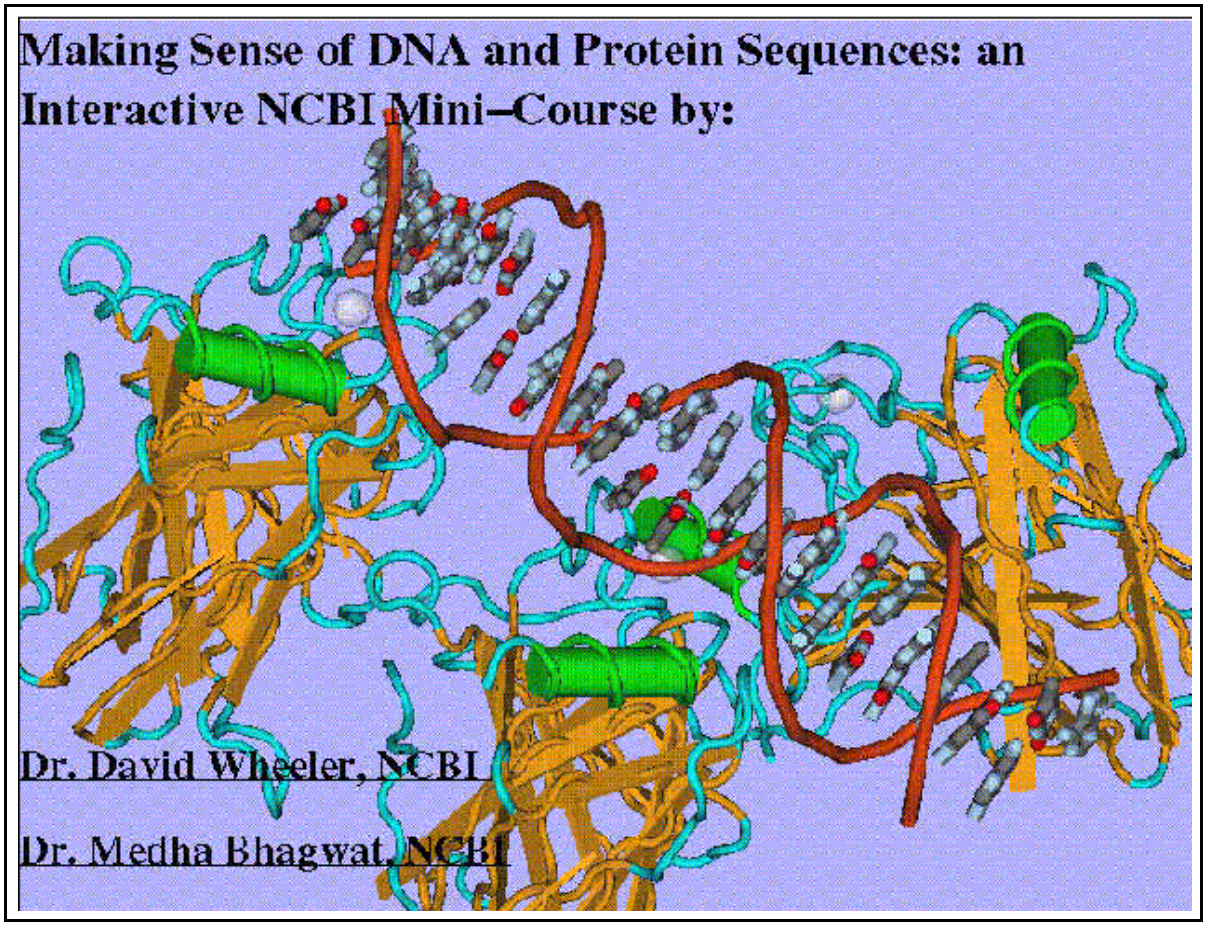

Making Sense of DNA and Protein Sequences: an Interactive NCBI Mini–Course by:
Dr. David Wheeler, NCBI
Dr. Medha Bhagwat, NCBI

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 electonic 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

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

For information about Genscan, click here

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):

```
238741 ggcgggagct gaatggaata attacataat ttacataact tgaatatgat tgcccag
238801 aacttcttac cctcgaccac ttccttcacg tggatggcat ggtgcagctc ttccatg
238861 tgcttgtcct cggttatgta taatattgct gttagcggtt cagtcagttt gactggg
238921 tgtcgtcggt tttgcctgca aagtgatata atttcaatat gttcccaaag gattata
238981 gagttaaaat tacgtgagca gcacgggctc atccaggcag ccgccatcgc aagcgta
239041 gctcctgtct gaacgatcca gcgcgatgtt gatcaccgcc ttgttgtcat catccac
239101 gatggtgacg ttgacatggg ttttgttgcc gtaattgagg cgcctgaagt ggaagtc
239161 atgcagaaac ttcggatgga tgttgcactc cagatgctga ttgctaaaac gaaagct
239221 aaagctcagg accatggctt ggatgactcc cgaagctccc gctcgaataa gattgtg
239281 gccctgtttc tccaacgtta gcatccacga cgtaaccagg gagttaagct gctgaag
239341 agacatctcg cgccaaagat tctccgcctg cagagtgtga tacgaatcgt agcagcc
```

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

Run GENSCAN Clear Input

GENSCAN 1.0 Date run: 13-Nov-101 Time: 12:40:08

Sequence 12:40:08 : 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):

>12:40:08|GENSCAN_predicted_peptide_1|143_aa
MPRTLPWTTVFTAVASSARAKSMEKLTVVFLLRMHSALVVSQPSMATRVNLPVFDPQSLN
SRAPAKTTSAAQAITAYLSIFFHLIELQGKRIGWLFRWLSPLSASSQRYESTKSGESPKT
TQSFRMNGKQLRAATQKKAFFDD

>12:40:08|GENSCAN_predicted_peptide_2|424_aa
MSQICKRGLLISNRLAPAALRCKSTWFSEVQMGPPDAILGVTEAFKKDTNPKKINLGAGA
YRDDNTQPFVLPSVREAEKRVVSRSLDKEYATIIGIPEFYNKAIELALGKGSKRLAAKHN
VTAQSISGTGALRIGAAFLAKFWQGNREIYIPSPSWGNHVAIFEHAGLPVNRYRYYDKDT
CALDFGGLIEDLKKIPEKSIVLLHACAHNPTGVDPTLEQWREISALVKKRNLYPFIDMAY
QGFATGDIDRDAQAVRTFEADGHDFCLAQSFAKNMGLYGERAGAFTVLCSDEEEAARVMS
QVKILIRGLYSNPPVHGARIAAEILNNEDLRAQWLKDVKLMADRIIDVRTKLKDNLIKLG
SSQNWDHIVNQIGMFCFTGLKPEQVQKLIKDHSVYLTNDGRVSMAGVTSKNVEYLAESIH
KVTK

>12:40:08|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 raw (leave off the fasta defline) protein sequence from GenScan into the ScanProsite input box, choose to ***Exclude patterns with a high probability of occurrence***, 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

Prosite Summary

ScanProsite

This program allows to scan a protein sequence (either from SWISS-PROT or TrEMBL or provided by the user) for the occurrence of patterns and profiles stored in the PROSITE database, or to search protein databases with a user-entered pattern. You may either:

- enter a PROSITE accession number or pattern to search the SWISS-PROT/TrEMBL and/or PDB databases with a pattern, **OR**
- enter a sequence or a SWISS-PROT/TrEMBL accession number to scan the sequence with all patterns, profiles and rules in PROSITE, **OR**
- fill in both fields to find all occurrences of a pattern or profile in a sequence.

Scan a protein for PROSITE matches

Enter a SWISS-PROT/TrEMBL accession number (AC) (for example **P05130**) or a sequence identifier (ID) (for example **NOTC_DROME**), or paste your own **protein** sequence in the box below:

```
MSQICKRGLLISNRLAPAALRCKSTWFSEVQMGPPDAILG
YRDDNTQPFVLPSVREAEKRVVSRSLDKEYATIIGIPEFY
VTAQSISGTGALRIGAAFLAKFWQGNREIYIPSPSWGNHV
CALDFGGLIEDLKKIPEKSIVLLHACAHNPTGVDPTLEQW
QGFATGDIDRDAQAVRTFEADGHDFCLAQSFAKNMGLYGE
QVKILIRGLYSNPPVHGARIAAEILNNEDLRAQWLKDVKL
SSQNWDHIVNQIGMFCFTGLKPEQVQKLIKDHSVYLTNDG
KVTK
```

Clear

and specify which motifs to use:

Scan ☑ patterns ☑ profiles ☑ rules [User Manual] (You may also specify a PROSITE entry in the box to the left)

☐ Exclude patterns with a high probability of occurrence

Your e-mail (optional): (will send results by e-mail)

☐ plain text output

START THE SCAN RESET

Search SWISS-PROT with a PROSITE entry

Enter a PROSITE accession number (for example **PS01253**), type your pattern in PROSITE format:
(leave this box blank to scan a sequence with the entire PROSITE database)

and specify your search limits:

- The ☑ SWISS-PROT ☐ TrEMBL ☐ TrEMBLnew ☐ PD
  new databases (You may also specify a protein in the box the right)
  ☑ including splice variants
- The following taxons:
  (see SWISS-PROT list of organisms; separate multiple taxons with a semicolon, e.g. *Homo sapiens; Drosophila*. Not availabl PDB.)
- Sequences with at least hits
- At most 1000 matches

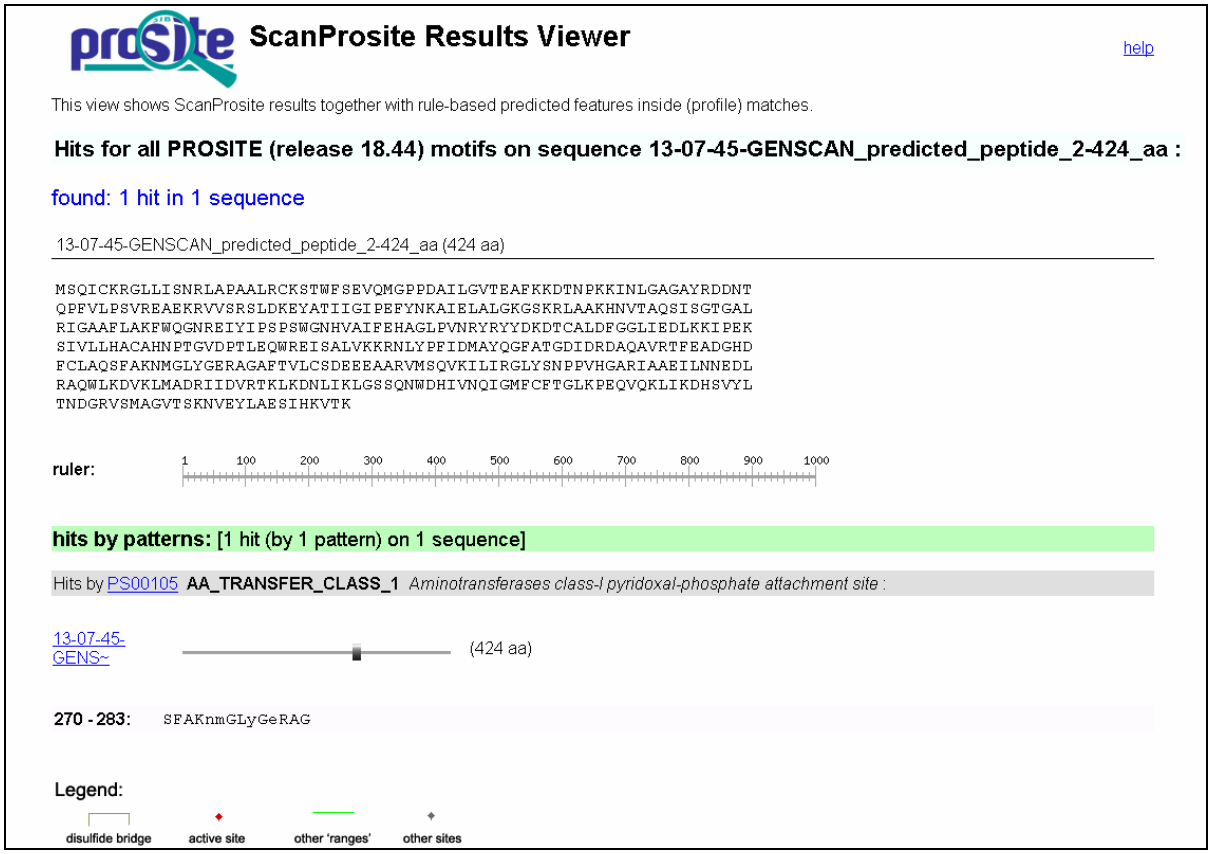
Advanced options: ☐ FASTA output

allow at most 1 X sequence characters to matcl conserved position in the pattern

match mode greedy, overlaps, no includes

(for patterns, see help)

randomize databases no (to t



ScanProsite Results Viewer

help

This view shows ScanProsite results together with rule-based predicted features inside (profile) matches.

Hits for all PROSITE (release 18.44) motifs on sequence 13-07-45-GENSCAN_predicted_peptide_2-424_aa :

found: 1 hit in 1 sequence

13-07-45-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 1000

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* :

13-07-45-GENS~ (424 aa)

270 - 283: SFAKnmGLyGeRAG

Legend:

disulfide bridge | active site | other 'ranges' | other sites

Aminotransferases class-I pyridoxal-phosphate attachment site

PROSITE cross-reference(s)

PS00105; AA_TRANSFER_CLASS_1	**Retrieve an alignment of Swiss-Prot true positive hits:** [Clustal format, color, condensed view] [Clustal format, color] [Clustal format, plain text] [Fasta format] **Retrieve a list of all Swiss-Prot/TrEMBL entries matching PS00105**

Documentation

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.

Description of pattern(s) and/or profile(s)

Consensus pattern	[GS]-[LIVMFYTAC]-[GSTA]-K-x(2)-[GSALVN]-[LIVMFA]-x-[GNAR]- x-R-[LIVMA]-[GA] [K is the pyridoxal-P attachment site]

To search for proteins with similar sequences, use:

BLAST

Run a BLASTp search against the SwissProt database by pasting the protein sequence from GenScan into the input box on the Advanced BLAST page. Choose the SwissProt database from the database listbox and the ***"blastp"*** program from the program listbox, then press the ***"Submit"*** button. Format your results as "Flat query anchored with identities" and paste this alignment into your notebook.

BLASTP Alignment (against SwissProt)

NCBI BLAST

Nucleotide | Protein | Translations | Retrieve results for an RID

Search

```
>14:10:32|GENSCAN_predicted_peptide_2|424_aa
MSQICKRGLLISNRLAPAALRCKSTWFSEVQMGPPDAILGVTEAFKKDTNPKKINLGAGA
YRDDNTQPFVLPSVREAEKRVVSRSLDKEYATIIGIPEFYNKAIELALGKGSKRLAAKHN
VTAQSISGTGALRIGAAFLAKFWQGNREIYIPSPSWGNHVAIFEHAGLPVNRYRYYDKDT
CALDFGGLIEDLKKIPEKSIVLLHACAHNPTGVDPTLEQWREISALVKKRNLYPFIDMAY
```

Set subsequence From: To:

Choose database swissprot

Do CD-Search

Now: BLAST! or Reset query Reset all

Options for advanced blasting

Limit by entrez query AND (none)

Composition-based statistics

Choose filter Low complexity Mask for lookup table only Mask lower case

Expect 10

NCBI

formatting **BLAST**

Nucleotide | Protein | Translations | Retrieve results for an RID

Your request has been successfully submitted and put into the Blast Queue.

Query = 14:10:32|GENSCAN_predicted_peptide_2|424_aa (424 letters)

The request ID is 1005667746-7811-27625

Format! or Reset all

The results are estimated to be ready in 8 seconds but may be done sooner.

Please press "FORMAT!" when you wish to check your results. You may change the formatting options for your result via the form below and press "FORMAT!" again. You may also request results of a different search by entering any other valid request ID to see other recent jobs.

Format

Show ☐ Graphical Overview ☐ NCBI-gi Alignment in HTML format

Number of: Descriptions 100 Alignments 50

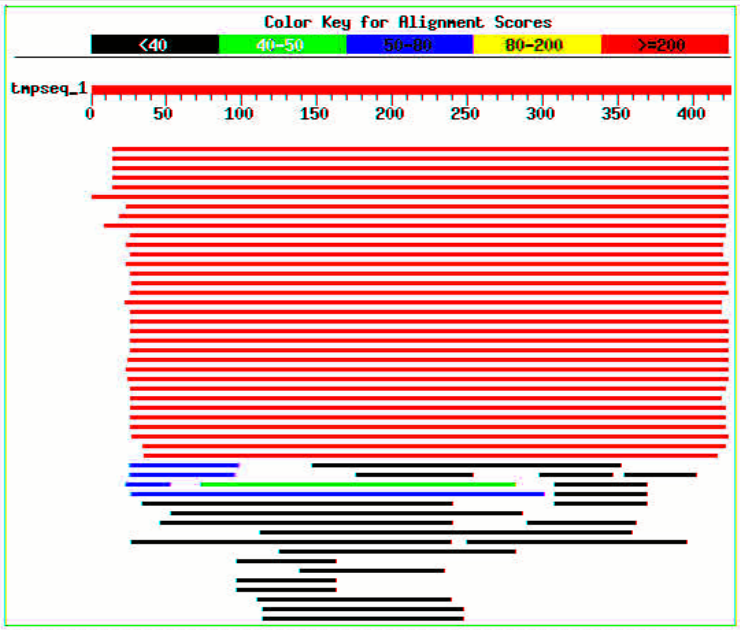
Alignment view flat query-anchored with identities

Format for PSI-BLAST ☐ with inclusion threshold: 0.001

Limit results by entrez query AND (none)

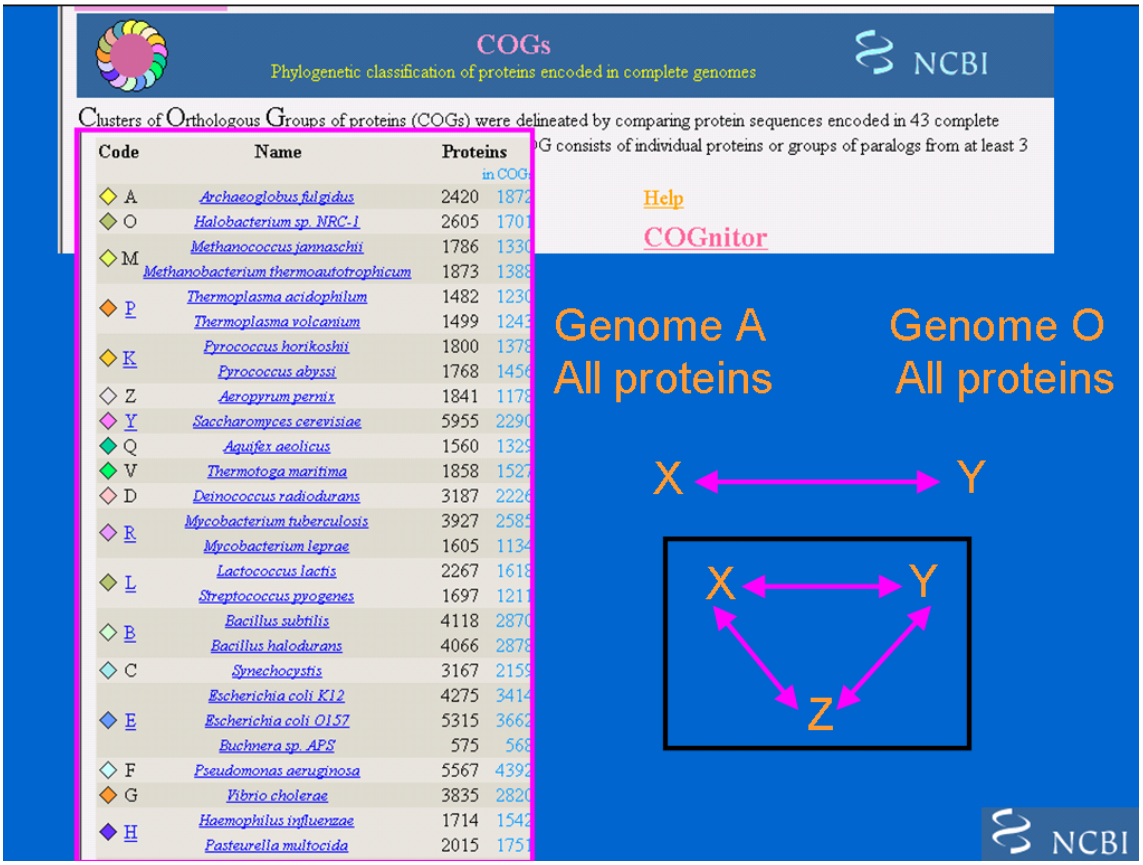
Distribution of 60 Blast Hits on the Query Sequence

Mouse-over to show defline and scores. Click to show alignments



Sequences producing significant alignments:		Score (bits)	E Value
sp|P00505|AATM_HUMAN	ASPARTATE AMINOTRANSFERASE, MITOCHOND...	549	e-156
sp|P05202|AATM_MOUSE	ASPARTATE AMINOTRANSFERASE, MITOCHOND...	549	e-156
sp|P00507|AATM_RAT	ASPARTATE AMINOTRANSFERASE, MITOCHONDRI...	548	e-156
sp|P00506|AATM_PIG	ASPARTATE AMINOTRANSFERASE, MITOCHONDRIA...	543	e-154
sp|P12344|AATM_BOVIN	ASPARTATE AMINOTRANSFERASE, MITOCHONDR...	542	e-154
sp|P00508|AATM_CHICK	ASPARTATE AMINOTRANSFERASE, MITOCHONDR...	538	e-153
sp|P08907|AATM_HORSE	ASPARTATE AMINOTRANSFERASE, MITOCHONDR...	538	e-152
sp|P46643|AAT1_ARATH	ASPARTATE AMINOTRANSFERASE, MITOCHONDR...	434	e-121
sp|P28011|AAT1_MEDSA	ASPARTATE AMINOTRANSFERASE 1 (TRANSAMI...	422	e-118
sp|P46644|AAT3_ARATH	ASPARTATE AMINOTRANSFERASE, CHLOROPLAS...	409	e-114
sp|P28734|AATC_DAUCA	ASPARTATE AMINOTRANSFERASE, CYTOPLASMI...	409	e-114

```
1          249 DR------DAQAVR------TFE-ADGHD-FCL-AQSFAKNMGLYGERAGAFTVLCSDEE 293
112983     255 .K------..W...------H.I-EQ.IN-V..-C..Y..........V....MV.K.AD 299
112984     255 .K------..W...------H.I-EQ.IN-V..-C..Y..........V.....V.K.A. 299
112987     255 .K------..W...------H.I-EQ.IN-V..-C..Y..........V.....V.K.A. 299
112982     226 NK------..W...------Y.I-EQ.IN-V..-C..Y..........V....MV.K.AD 270
112985     255 NK------..W...------H.I-EQ.IN-V..-C..Y..........V.....V.K.A. 299
1168261    255 NK------..W...------H.I-EQ.IN-V..-C..Y..........V.....V.K.A. 299
112981     248 N.------..W.L.------H.I-EQ.I.-VV.-S..Y................I.R.A. 292
1168256    253 A.------..KSI.------I.L-E...H-IGI-S..Y........Q.V.CLS...E.PK 297
2506178    240 .A------...P..------L.V-...GE-LLV-...Y..........V..LSIVSKSAD 284
112972     227 .A------...S..------I.V-...GE-CLA-...Y..........V..LSIV.KTAD 271
1168258    271 .T------..KPI.------M.V-...GE-CLV-...Y..........V..LSIV.KSAD 315
1168257    227 .T------...S..------..V-...GE-CLI-...YT.........V..LSIV.KSAD 271
584706     229 .Q------...S..------M.V-...GE-LLM-...Y..........V..LSIV.GSAD 273
112971     234 .K------..W...------Y.V-SE.FE-LFC-....S..F...N..V.NLS.VGK..D 278
1168259    225 .A------......------M.V-...GE-CLI-...Y..........I.SL.IV.TS.D 269
5902703    235 E.------..W.I.------Y.V-SE.FE-.FC-....S..F...N..V.NL..VGKEP. 279
112976     235 EK------..W.I.------Y.V-SE.FE-LFC-....S..F...N..V.NL..VAKEPD 279
1168260    274 .E------..AS..------L.A-ER.ME-.FV-...YS..L...A..I..IN.V..SAD 318
112973     234 ..------..W...------Y.V-SE.FE-LFC-....S..F...N..V.NL..VAKEPD 278
112975     235 EK------..W.I.------Y.V-SE.FE-LFC-....S..F...N..V.NL..VGKESD 279
416566     235 EK------..W.I.------Y.V-SE.FE-LFC-....S..F...N..V.NL..VAKEPD 279
112979     275 .E------..AS..------L.V-.R.LE-VLV-...YS..L...A..I..IN.IS.SP. 319
112978     235 EK------..W.I.------Y.V-SE.FE-LFC-P...S..F...N..V.NL..VGKEHD 279
1703040    228 .K------..Y...LGVE--KLS-TVSPV-.-V-C......A.M....V.C.HLALTKQA 275
2492843    227 AA------..W.I.------Y.V-DQ.ME-MVV-S......F...N..V.NL..VVNNPA 271
1168262    222 .E------..YGL.------A.A-.NHKE-LLV-.S..S..F...N..V....LVAENA. 266
112989     222 EE------..EGL.------A.A-.MHKE-LIV-.S.YS..F...N..V..C.LVAA.S. 266
12230871   224 EE------..A...------L.A-QS.LS-.FV-SS..S.SFS.....V..LSIVTESRD 268
2492844    223 .D------..YVI.------AIA-SA.LP-ALV-SN..S.IFS.....V.GLS.V.E.A. 267
136593     223 EE------..Y.I.------AIA-SA.LP-ALV-SN..S.IFS.....V.GLS.M.E.A. 267
398985     256 LK------..YLL.LCLNVNKYP-NWSNG-IF.-C............V.SLS.ITPATA 306
12230956   223 EE------..W...------L.A-GELPE-VLV-TS.CS..F...RD.V..LI.CAQNA. 267
6136085    219 EE------..AGT.------LIA-SRIPE-VLI-.A.CS..F.I.R..T.CLLA..A.AA 263
399090     218 .Q------.VAGL.------HLL-GVVPE-ALV-.V.CS.SF...R.....IFART.STA 262
6224986    216 KF------VSP.SF------SD.-VKNIT-.TV--NA.S.SYSMT.W.I.--Y.A.P--. 255
3183545    208 ..------PHYSIA------.YL-R.--Q-TIV-INGLS.SHSMT.W.I.           240
14285338   211 KHYSIASL.GMFE.------.IT-VN.----------.S.TFAMT.W.L.           243
14285336   212 YE------..KHYP------MIK-FAPEN-TI.-.N..S.TFAMT.W.L.           246
112990     218 --------EGKMRS------.L.DS.WR.-.LIYVNG.S.TFSMT.W.L.YIVA------ 256
14285339   213 FT------SILD.----------SE.F.RIVY-ING.S.SHSMT.W.V.---Y.I.S.K 251
3122214    211 NT------.------------LL-P.HEN-LFI-LRT.S.V...A.M.I.           239
13432184   184 .F------GGETCL------PLL-KKYDN-LVV-V.T.S.SRS.A.I.L.----VAYGSA 224
2501592    102 E.------EGM.                                                 107
```


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
Help
COGnitor
Code
Name
Proteins
in COG
A Archaeoglobus fulgidus 2420
O Halobacterium sp. NRC-1 2605
M Methanococcus jannaschii 1786
Methanobacterium thermoautotrophicum 1873
P Thermoplasma acidophilum 1482
Thermoplasma volcanium 1499
K Pyrococcus horikoshii 1800
Pyrococcus abyssi 1768
Z Aeropyrum pernix 1841
Y Saccharomyces cerevisiae 5955
Q Aquifex aeolicus 1560
V Thermotoga maritima 1858
D Deinococcus radiodurans 3187
R Mycobacterium tuberculosis 3927
Mycobacterium leprae 1605
L Lactococcus lactis 2267
Streptococcus pyogenes 1697
B Bacillus subtilis 4118
Bacillus halodurans 4066
C Synechocystis 3167
E Escherichia coli K12 4275
Escherichia coli O157 5315
Buchnera sp. APS 575
F Pseudomonas aeruginosa 5567
G Vibrio cholerae 3835
H Haemophilus influenzae 1714
Pasteurella multocida 2015
Genome A
All proteins
Genome O
All proteins
X
Y
X
Y
Z
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

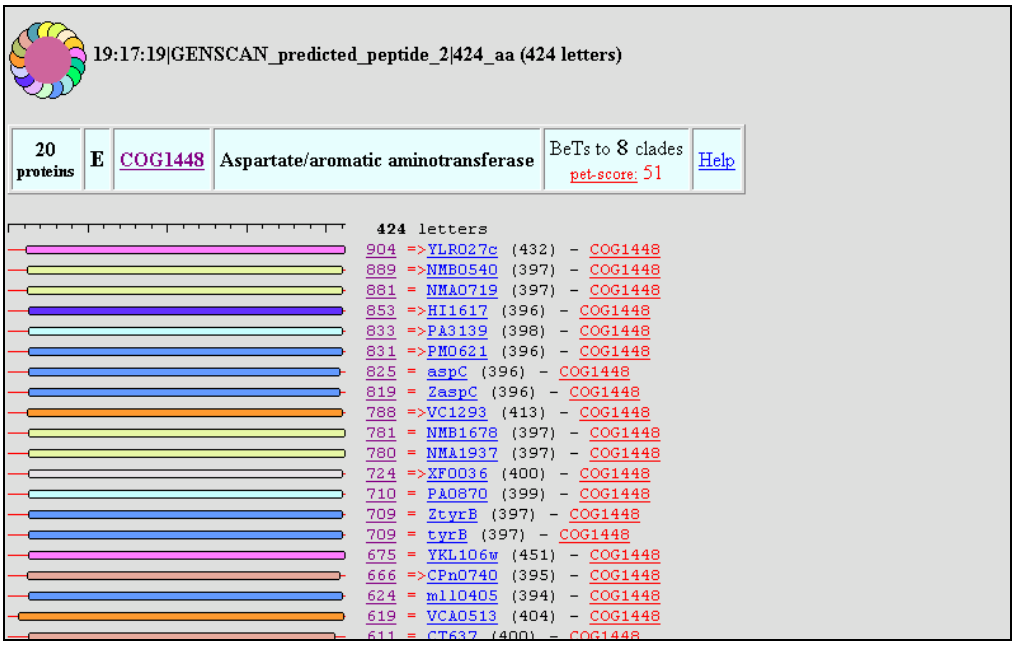
compare to COGs | Clear | BeTs to 3 clades

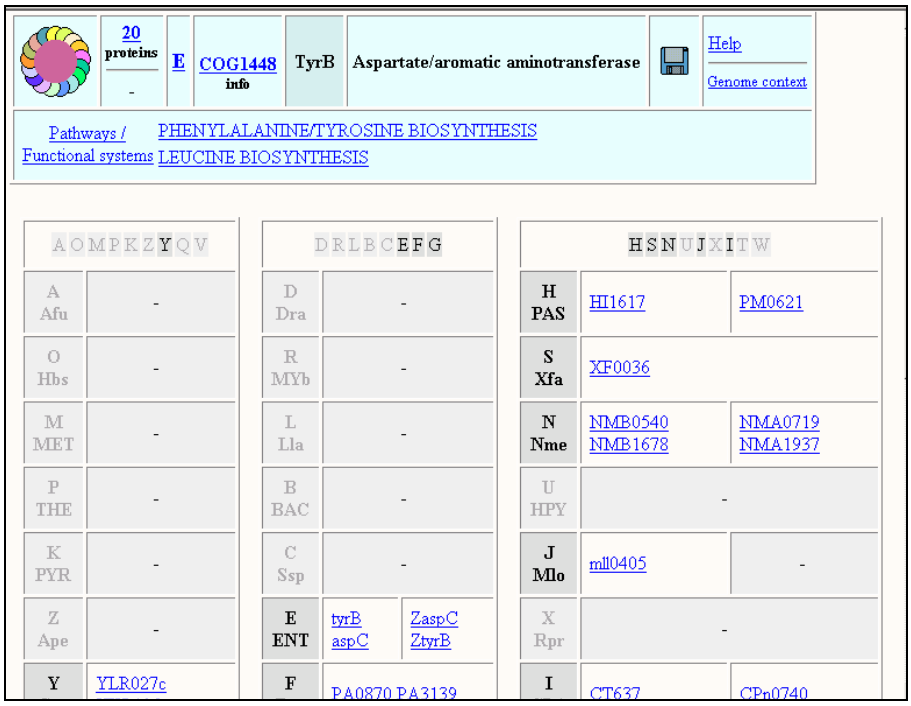
Help

Example

Paste your sequence and press the button above.

```
>13:07:45|GENSCAN_predicted_peptide_2|424_aa
MSQICKRGLLISNRLAPAALRCKSTWFSEVQMGPPDAILGVTEAFKKDTNPKKINLGAGA
YRDDNTQPFVLPSVREAEKRVVSRSLDKEYATIIGIPEFYNKAIELALGKGSKRLAAKHN
VTAQSISGTGALRIGAAFLAKFWQGNREIYIPSPSWGNHVAIFEHAGLPVNRYRYYDKDT
CALDFGGLIEDLKKIPEKSIVLLHACAHNPTGVDPTLEQWREISALVKKRNLYPFIDMAY
QGFATGDIDRDAQAVRTFEADGHDFCLAQSFAKNMGLYGERAGAFTVLCSDEEEAARVMS
QVKILIRGLYSNPPVHGARIAAEILNNEDLRAQWLKDVKLMADRIIDVRTKLKDNLIKLG
SSQNWDHIVNQIGMFCFTGLKPEQVQKLIKDHSVYLTNDGRVSMAGVTSKNVEYLAESIH
KVTK
```


19:17:19|GENSCAN_predicted_peptide_2|424_aa (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

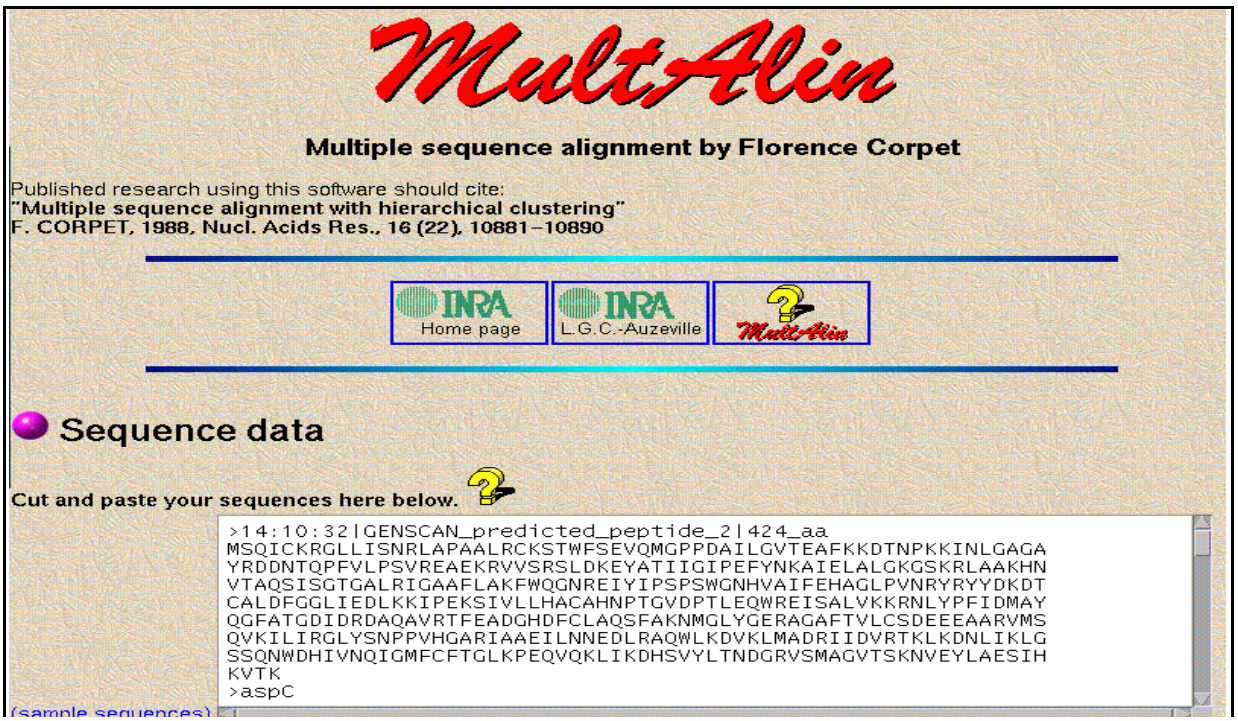

20 proteins
E
COG1448
info
TyrB
Aspartate/aromatic aminotransferase
Help
Genome context
Pathways / Functional systems
PHENYLALANINE/TYROSINE BIOSYNTHESIS
LEUCINE BIOSYNTHESIS
A O M P K Z Y Q V
D R L B C E F G
H S N U J X I T W
A Afu -
O Hbs -
M MET -
P THE -
K PYR -
Z Ape -
Y YLR027c
D Dra -
R MYb -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




```
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
19:17:19|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
19:17:19|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
```

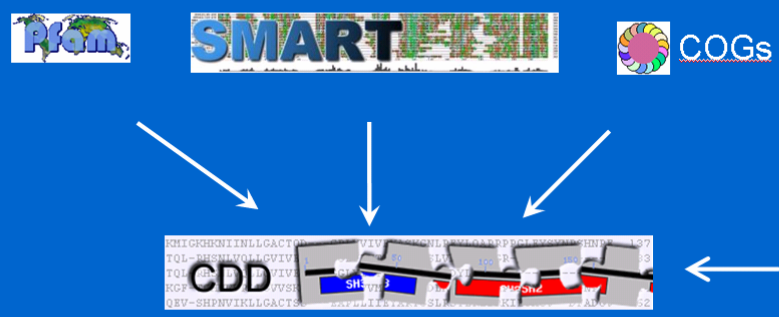
http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml

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



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 a library of conserved protein domains, 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. Paste this alignment into your notebook. Before viewing a structure with Cn3D, use the listbox to specify "up to 5" sequences and "All Atoms". Invoke Cn3D with a display of a 3D modeling template, and a multiple sequence alignment including your query sequence, by pressing the "View 3D Structure" button. Residues identical in your sequence and the structura 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

NCBI Conserved Domain Search

New Search | PubMed | Nucleotide | Protein | Structure | CDD | Taxonomy | Help?

Search the Conserved Domain Database with Reverse Position Specific BLAST

Search Database: CDD v2.00 - 11382 PSSMs

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

```
>19:17:19|GENSCAN_predicted_peptide_2|424_aa
MSQICKRGLLISNRLAPAALRCKSTWFSEVQMGPPDAILGVTEAFKKDTNPKKINLGAGA
YRDDNTQPFVLPSVREAEKRVVSRSLDKEYATIIGIPEFYNKAIELALGKGSKRLAAKHN
VTAQSISGTGALRIGAAFLAKFWQGNREIYIPSPSWGNHVAIFEHAGLPVNRYRYYDKDT
CALDFGGLIEDLKKIPEKSIVLLHACAHNPTGVDPTLEQWREISALVKKRNLYPFIDMAY
QGFATGDIDRDAQAVRTFEADGHDFCLAQSFAKNMGLYGERAGAFTVLCSDEEEAARVMS
QVKILIRGLYSNPPVHGARIAAEILNNEDLRAQWLKDVKLMADRIIDVRTKLKDNLIKLG
SSQNWDHIVNQIGMFCFTGLKPEQVQKLIKDHSVYLTNDGRVSMAGVTSKNVEYLAESIH
```

Please read about FASTA format description

Advanced options for the BLAST server:

Expect 0.01 Filter ☑ Low complexity. Search mode: Multiple hits 1-pass

Output formatting options:

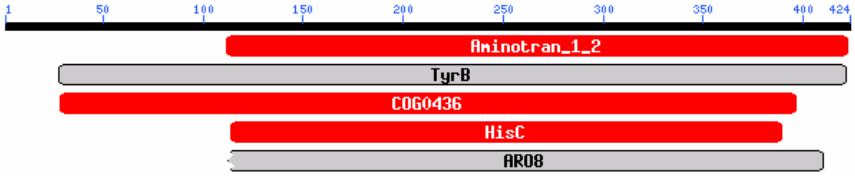
Display up to 25 hits with Extended Graphic Overview in Color Scheme 3

Print Graphics using Default Width

```
Query= local sequence:
19:17:19|GENSCAN_predicted_peptide_2|424_aa
         (424 letters)

Database: cdd.v2.00
           11,382 PSSMs; 2,824,437 total columns
```

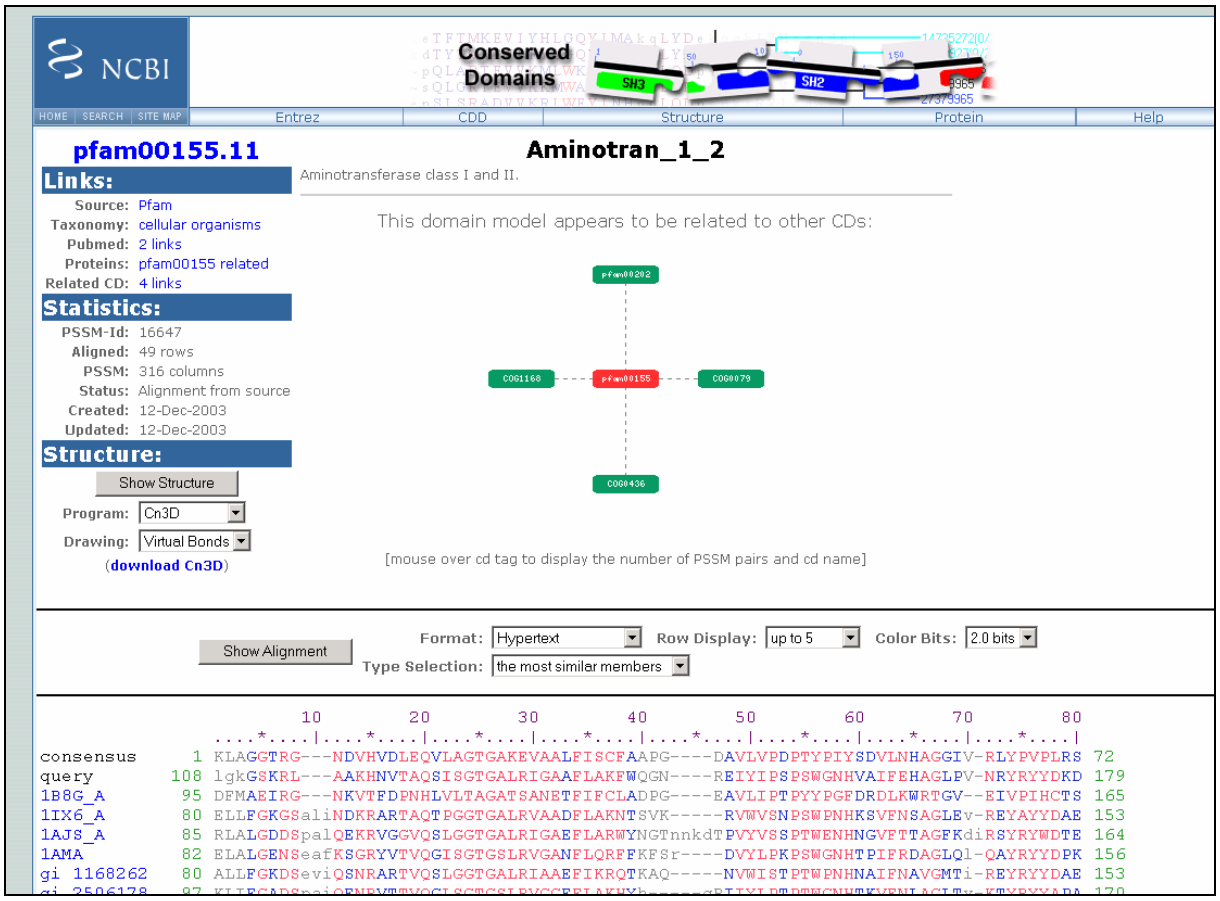
Click on boxes for multiple alignments

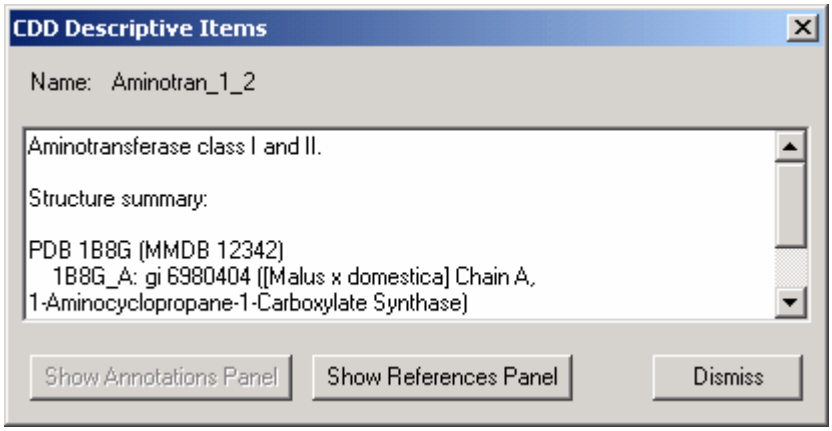


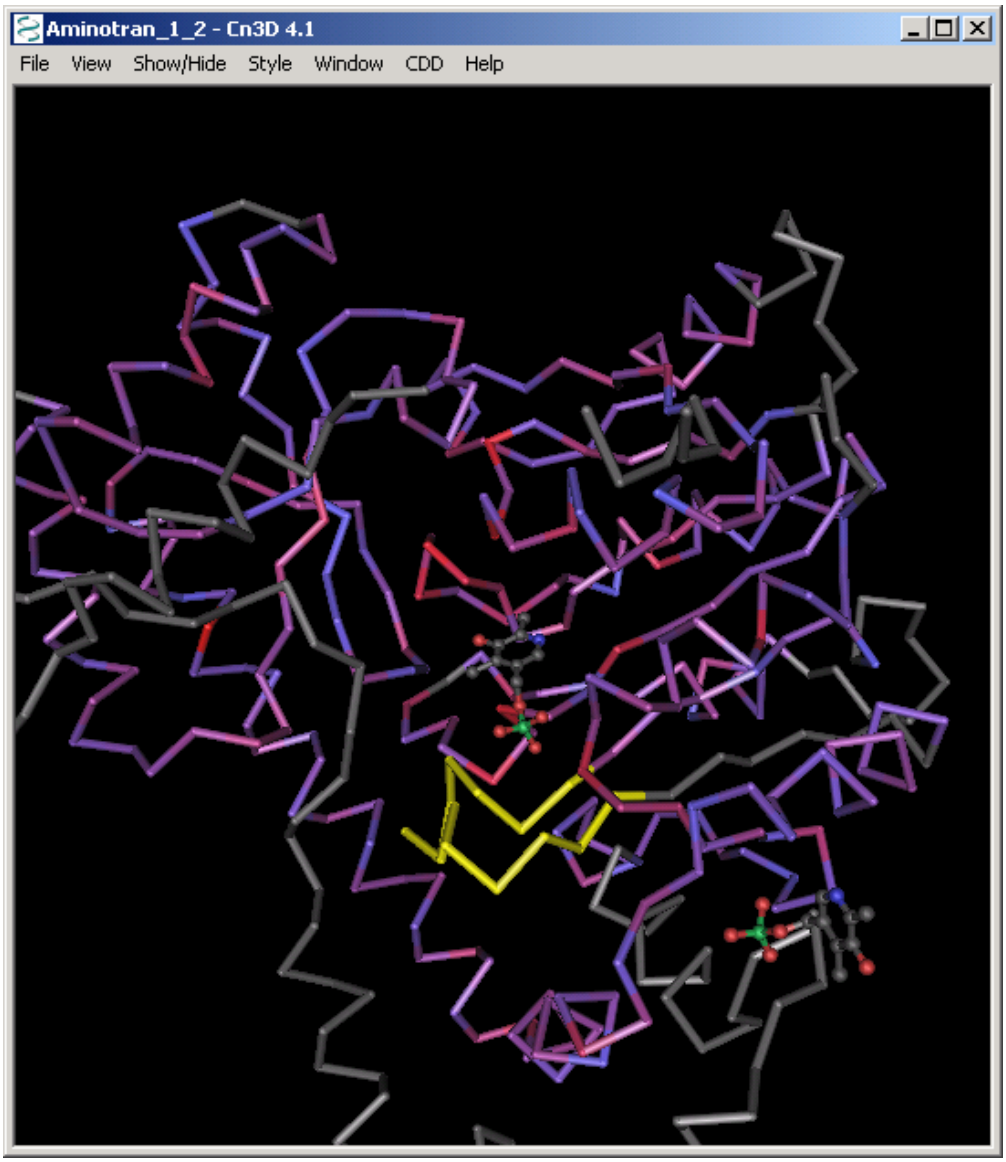
Show Domain Relatives

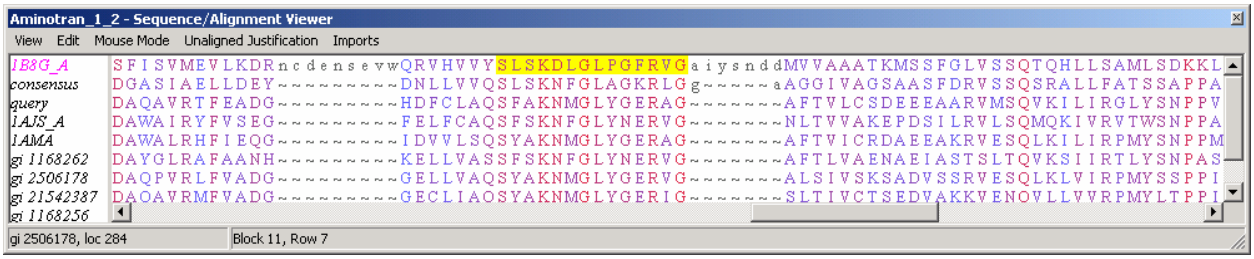
.. This CD alignment includes 3D structure. To display structure, download Cn3D!

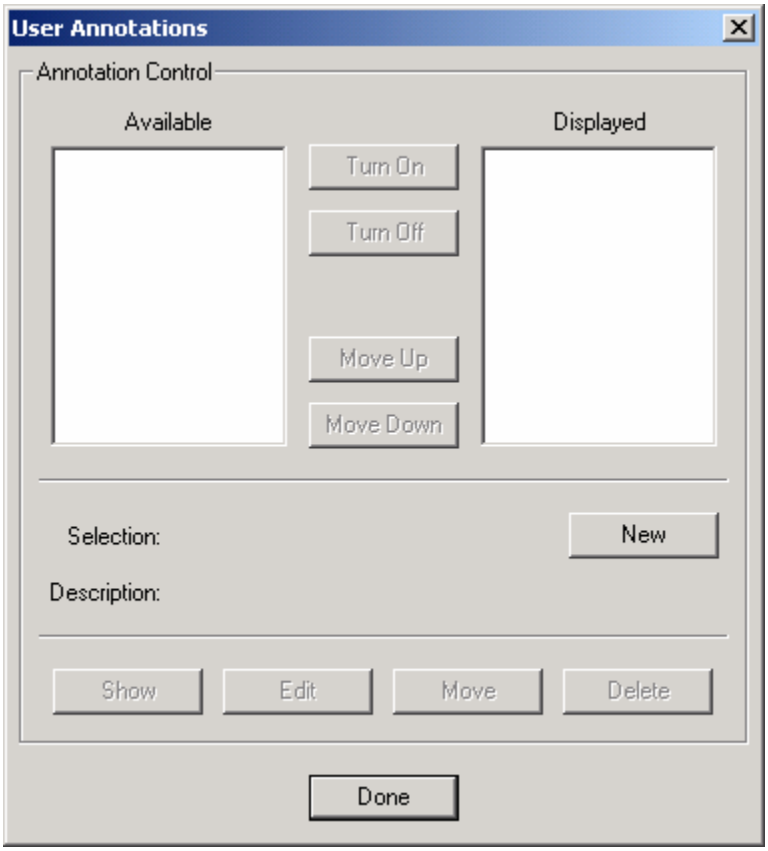
PSSMs producing significant alignments:		Score (bits)	E value
gnl|CDD|16647	pfam00155, Aminotran_1_2, Aminotransferase class I and II	249	4e-67
gnl|CDD|11162	COG1448, TyrB, Aspartate/tyrosine/aromatic aminotransferase [A...	516	2e-147
gnl|CDD|10310	COG0436, COG0436, Aspartate/tyrosine/aromatic aminotransferase...	81.9	1e-16
gnl|CDD|9954	COG0079, HisC, Histidinol-phosphate/aromatic aminotransferase ...	46.1	8e-06
gnl|CDD|10886	COG1167, ARO8, Transcriptional regulators containing a DNA-bin...	37.7	0.003

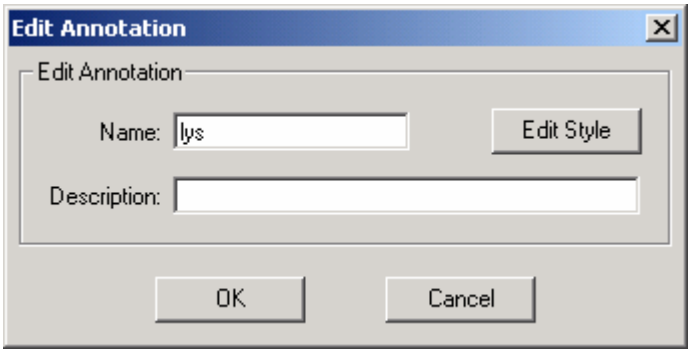

NCBI
Conserved Domains
SH3
SH2
HOME
SEARCH
SITE MAP
Entrez
CDD
Structure
Protein
Help
pfam00155.11
Aminotran_1_2
Aminotransferase class I and II.
Links:
Source: Pfam
Taxonomy: cellular organisms
Pubmed: 2 links
Proteins: pfam00155 related
Related CD: 4 links
Statistics:
PSSM-Id: 16647
Aligned: 49 rows
PSSM: 316 columns
Status: Alignment from source
Created: 12-Dec-2003
Updated: 12-Dec-2003
Structure:
Show Structure
Program: Cn3D
Drawing: Virtual Bonds
(download Cn3D)
This domain model appears to be related to other CDs:
[mouse over cd tag to display the number of PSSM pairs and cd name]
Show Alignment
Format: Hypertext
Row Display: up to 5
Color Bits: 2.0 bits
Type Selection: the most similar members
consensus
query
1B8G_A
1IX6_A
1AJS_A
1AMA
gi 1168262

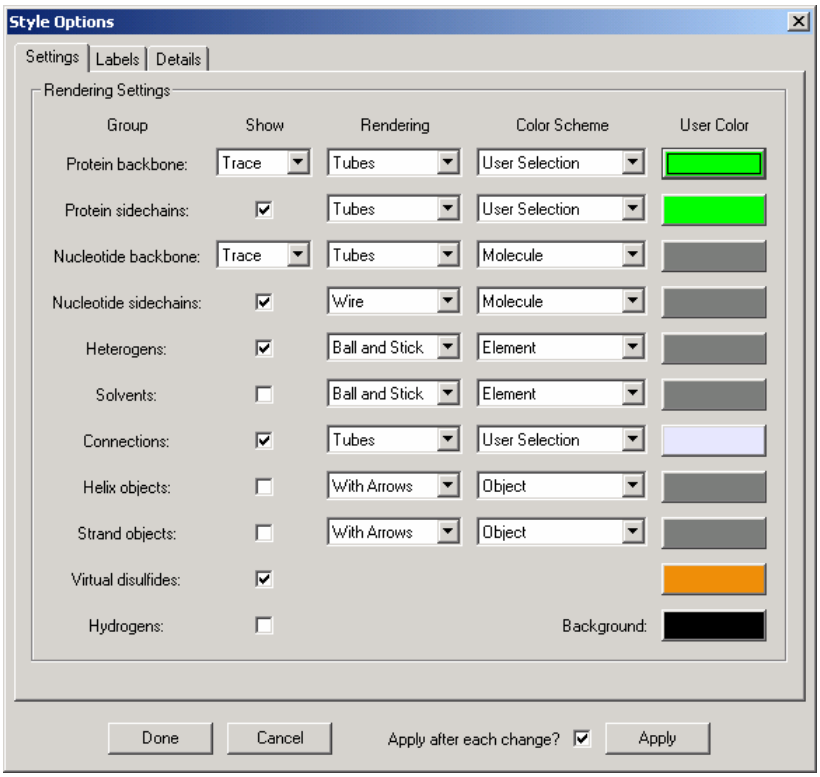

CDD Descriptive Items
Name: Aminotran_1_2
Aminotransferase class I and II.
Structure summary:
PDB 1B8G (MMDB 12342)
1B8G_A: gi 6980404 ([Malus x domestica] Chain A, 1-Aminocyclopropane-1-Carboxylate Synthase)
Show Annotations Panel
Show References Panel
Dismiss

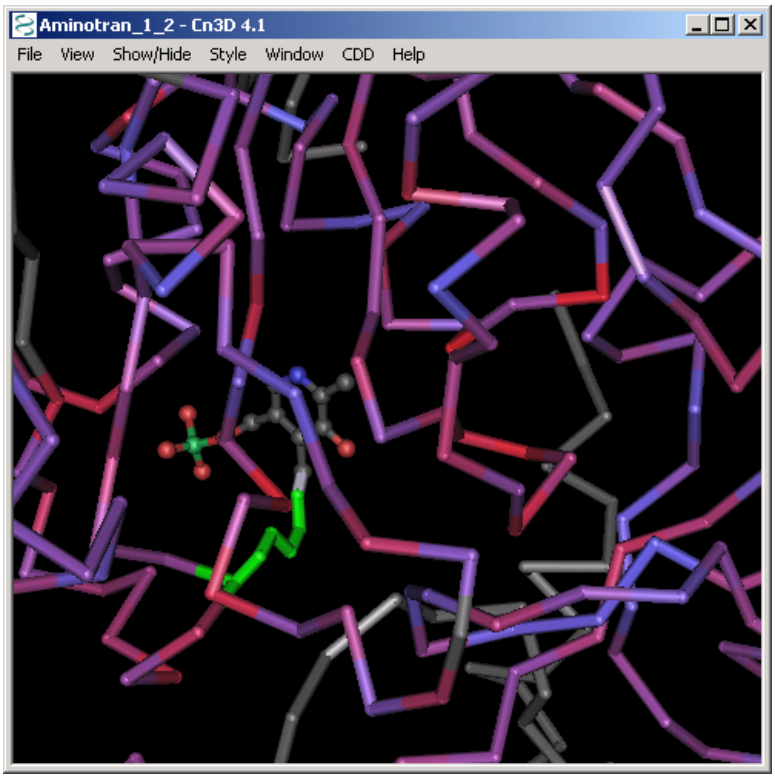

Aminotran_1_2 - Cn3D 4.1
File View Show/Hide Style Window CDD Help


Aminotran_1_2 - Sequence/Alignment Viewer
View Edit Mouse Mode Unaligned Justification Imports
1B8G_A SFISVMEVLKDRncdensevwQRVHVVYSLSKDLGLPGFRVGaiysnddMVVAAATKMSSFGLVSSQTQHLLSAMLSDKKL
consensus DGASIAELLDEY~~~~~~~~~~DNLLVVQSLSKNFGLAGKRLGg~~~~~aAGGIVAGSAASFDRVSSQSRALLFATSSAPPA
query DAQAVRTFEADG~~~~~~~~~~HDFCLAQSFAKNMGLYGERAG~~~~~~~AFTVLCSDEEEAARVMSQVKILIRGLYSNPPV
1AJS_A DAWAIRYFVSEG~~~~~~~~~~FELFCAQSFSKNFGLYNERVG~~~~~~~NLTVVAKEPDSILRVLSQMQKIVRVTWSNPPA
1AMA DAWALRHFIEQG~~~~~~~~~~IDVVLSQSYAKNMGLYGERAG~~~~~~~AFTVICRDAEEAKRVESQLKILIRPMYSNPPM
gi 1168262 DAYGLRAFAANH~~~~~~~~~~KELLVASSFSKNFGLYNERVG~~~~~~~AFTLVAENAEIASTSLTQVKSIIRTLYSNPAS
gi 2506178 DAQPVRLFVADG~~~~~~~~~~GELLVAQSYAKNMGLYGERVG~~~~~~~ALSIVSKSADVSSRVESQLKLVIRPMYSSPPI
gi 21542387 DAQAVRMFVADG~~~~~~~~~~GECLIAQSYAKNMGLYGERIG~~~~~~~SLTIVCTSEDVAKKVENQVLLVVRPMYLTPPI
gi 1168256
gi 2506178, loc 284
Block 11, Row 7


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


Aminotran_1_2 - Cn3D 4.1
File
View
Show/Hide
Style
Window
CDD
Help