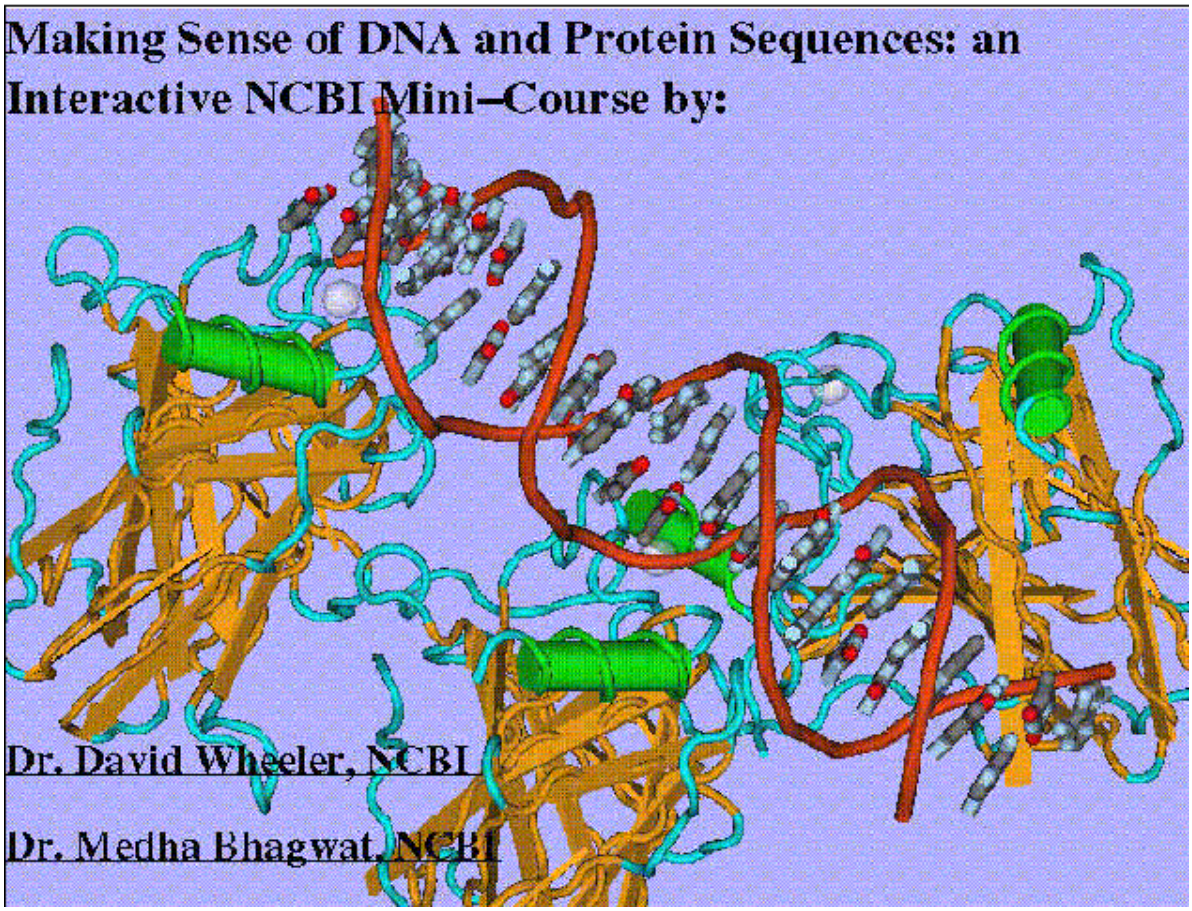
Making Sense of DNA and Protein Sequences: an Interactive NCBI Mini–Course by:



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

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

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: 27-Mar-107 Time: 13:53:15

Sequence 13:53:11 : 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):

>13:53:11|GENSCAN_predicted_peptide_1|143_aa
MPRTLPWTTVFTAVASSARAKSMEKLTVVFLLRMHSALVVSQPSMATRVNLPVFDPQSLN
SRAPAKTTSAAQAITAYLSIFFHLIELQGKRIGWLFRWLSPLSASSQRYESTKSGESPKT
TQSFRMNGKQLRAATQKKAFFDD

>13:53:11|GENSCAN_predicted_peptide_2|424_aa
MSQICKRGLLISNRLAPAALRCKSTWFSEVQMGPPDAILGVTEAFKKDTNPKKINLGAGA
YRDDNTQPFVLPSVREAEKRVVSRSLDKEYATIIGIPEFYNKAIELALGKGSKRLAAKHN
VTAQSISGTGALRIGAAFLAKFWQGNREIYIPSPSWGNHVAIFEHAGLPVNRYRYYDKDT
CALDFGGLIEDLKKIPEKSIVLLHACAHNPTGVDPTLEQWREISALVKKRNLYPFIDMAY
QGFATGDIDRDAQAVRTFEADGHDFCLAQSFAKNMGLYGERAGAFTVLCSDEEEAARVMS
QVKILIRGLYSNPPVHGARIAAEILNNEDLRAQWLKDVKLMADRIIDVRTKLKDNLIKLG
SSQNWDHIVNQIGMFCFTGLKPEQVQKLIKDHSVYLTNDGRVSMAGVTSKNVEYLAESIH
KVTK

>13:53:11|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:

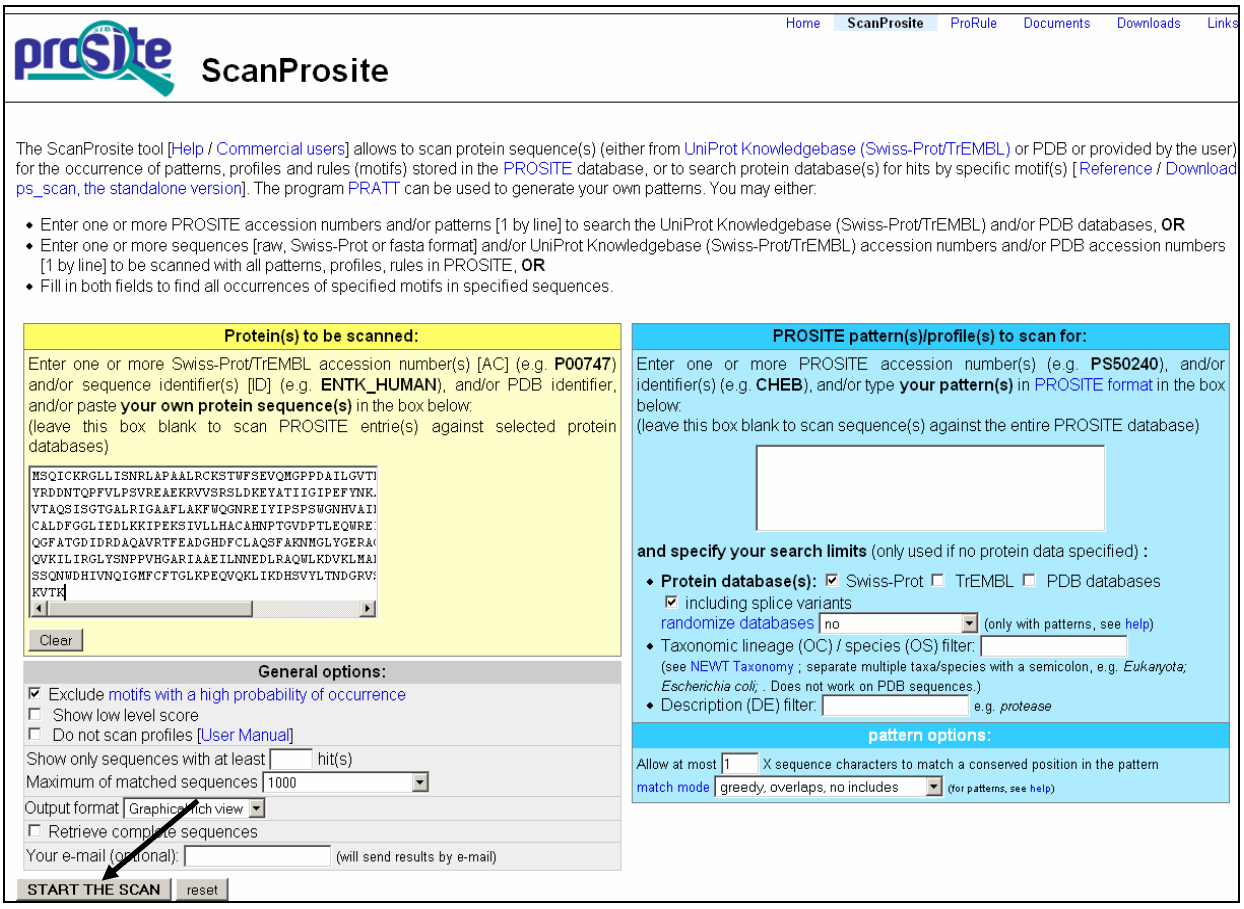
ScanProsite

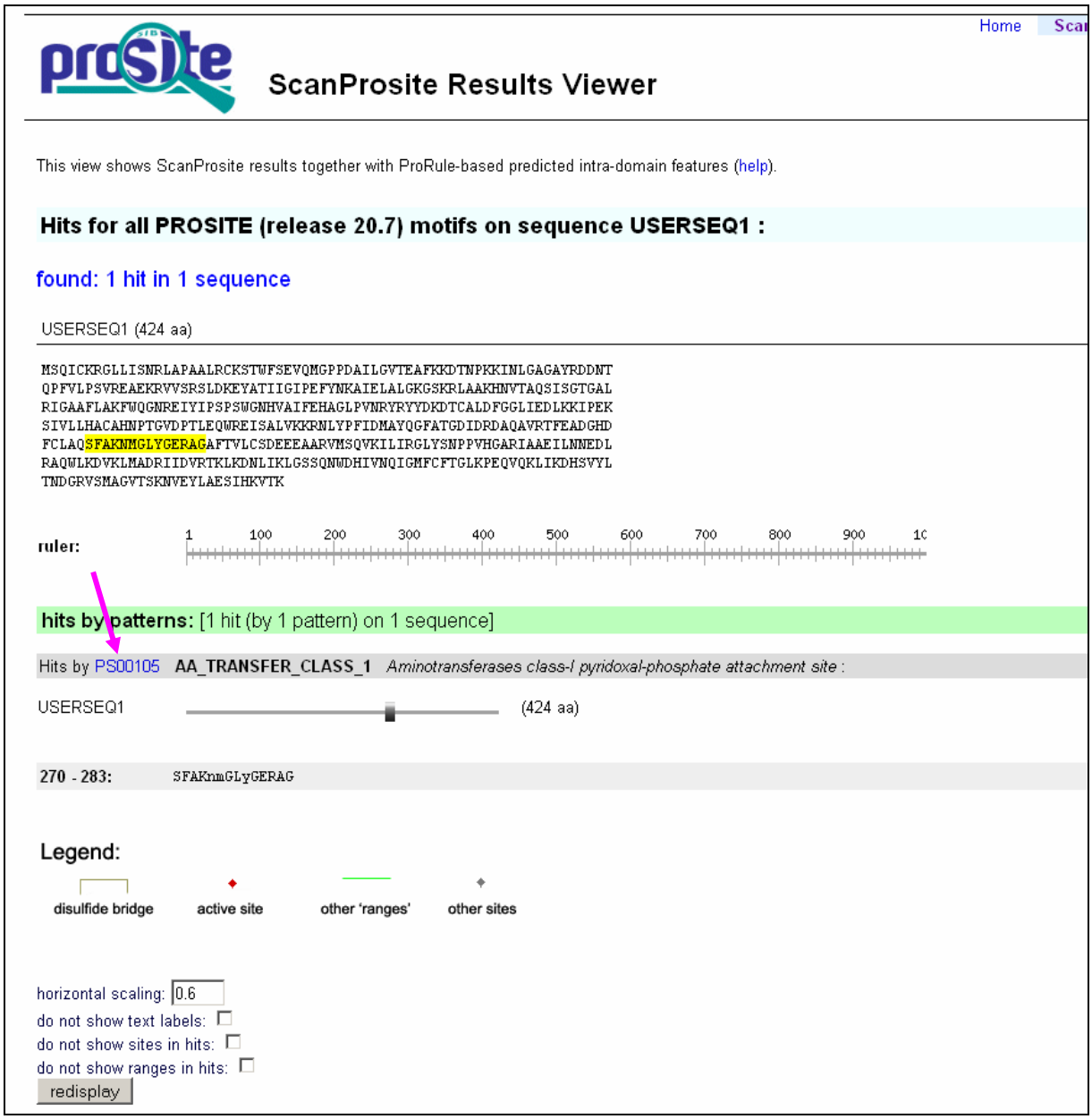
Paste the raw (leave off the fasta defline) protein sequence from GenScan into the ScanProsite input box, choose to ***Exclude patterns with a high probability of occurrence***, and press the ***"Start the Scan"*** button. Paste the ScanProsite hit into your notebook. To see the Prosite summary for the hit, click on the PDOCxxxx number.

Hit from ScanProsite

Prosite pattern

Prosite Summary


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)
MSQICKRGLLISNRLAPAALRCKSTWFSEVQMGPPDAILGVTI
YRDDNTQPFVLPSVREAEKRVVSRSLDKEYATIIGIPEFYNK
VTAQSISGTGALRIGAAFLAKFWQGNREIYIPSPSWGNHVAI
CALDFGGLIEDLKKIPEKSIVLLHACAHNPTGVDPTLEQWRE
QGFATGDIDRDAQAVRTFEADGHDFCLAQSFAKNMGLYGERA
QVKILIRGLYSNPPVHGARIAAEILNNEDLRAQWLKDVKLMA
SSQNWDHIVNQIGMFCFTGLKPEQVQKLIKDHSVYLTNDGRV
KVTK
Clear
General options:
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): Swiss-Prot TrEMBL PDB databases
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 Results Viewer
This view shows ScanProsite results together with ProRule-based predicted intra-domain features (help).
Hits for all PROSITE (release 20.7) motifs on sequence USERSEQ1 :
found: 1 hit in 1 sequence
USERSEQ1 (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 :
USERSEQ1
(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)

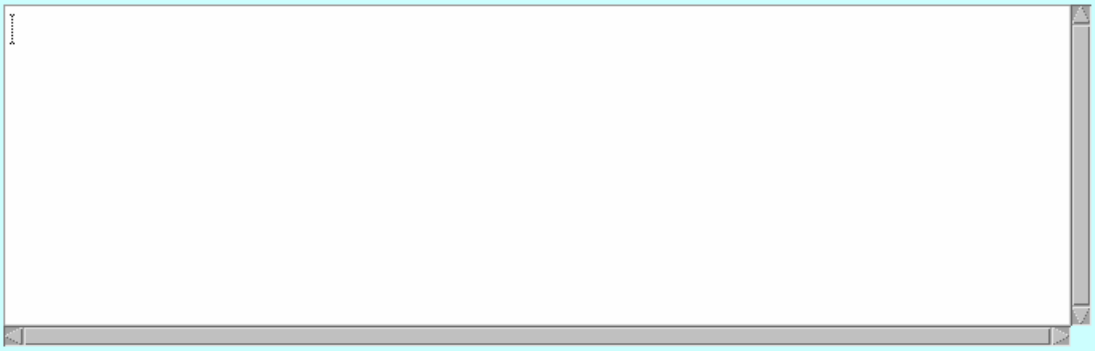
Consensus pattern:	**[GS] - [LIVMFYTAC] - [GSTA] - K - x(2) - [GSALVN] - [LIVMFA] - x - [GNAR] - {V} - R - [LIVMA] - [GA]** *K is the pyridoxal-P attachment site*
Sequences known to belong to this class detected by the pattern:	ALL
Other sequence(s) detected in Swiss-Prot:	1

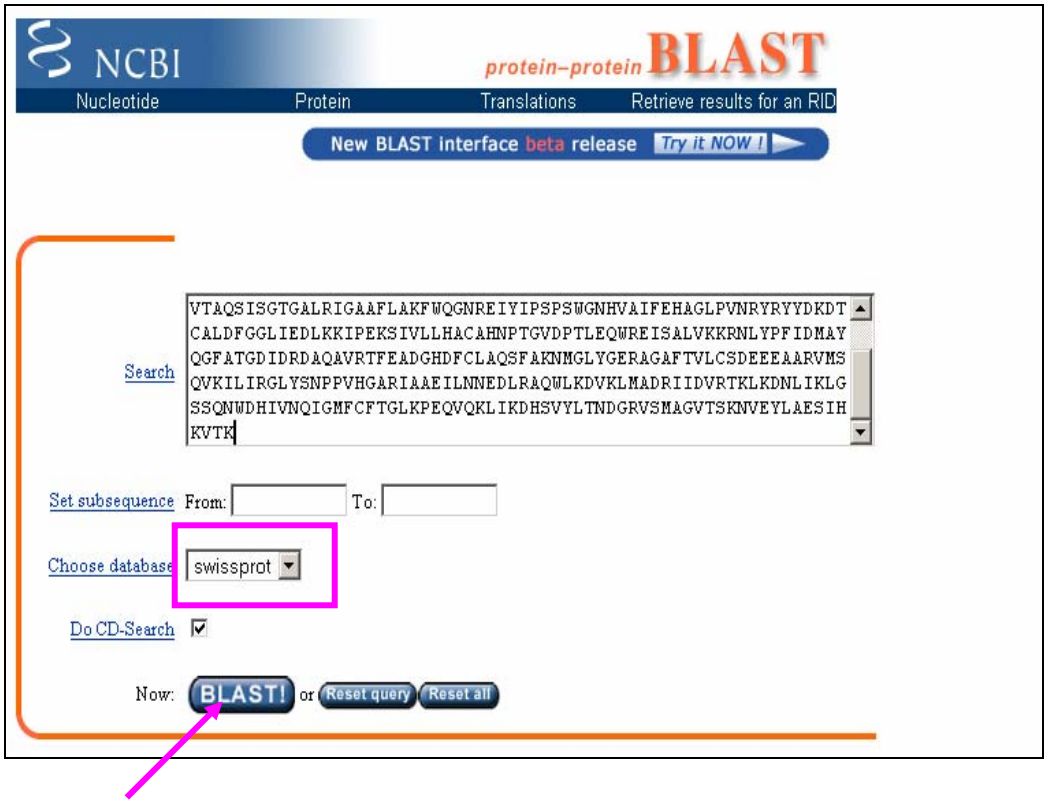
To search for proteins with similar sequences, use:

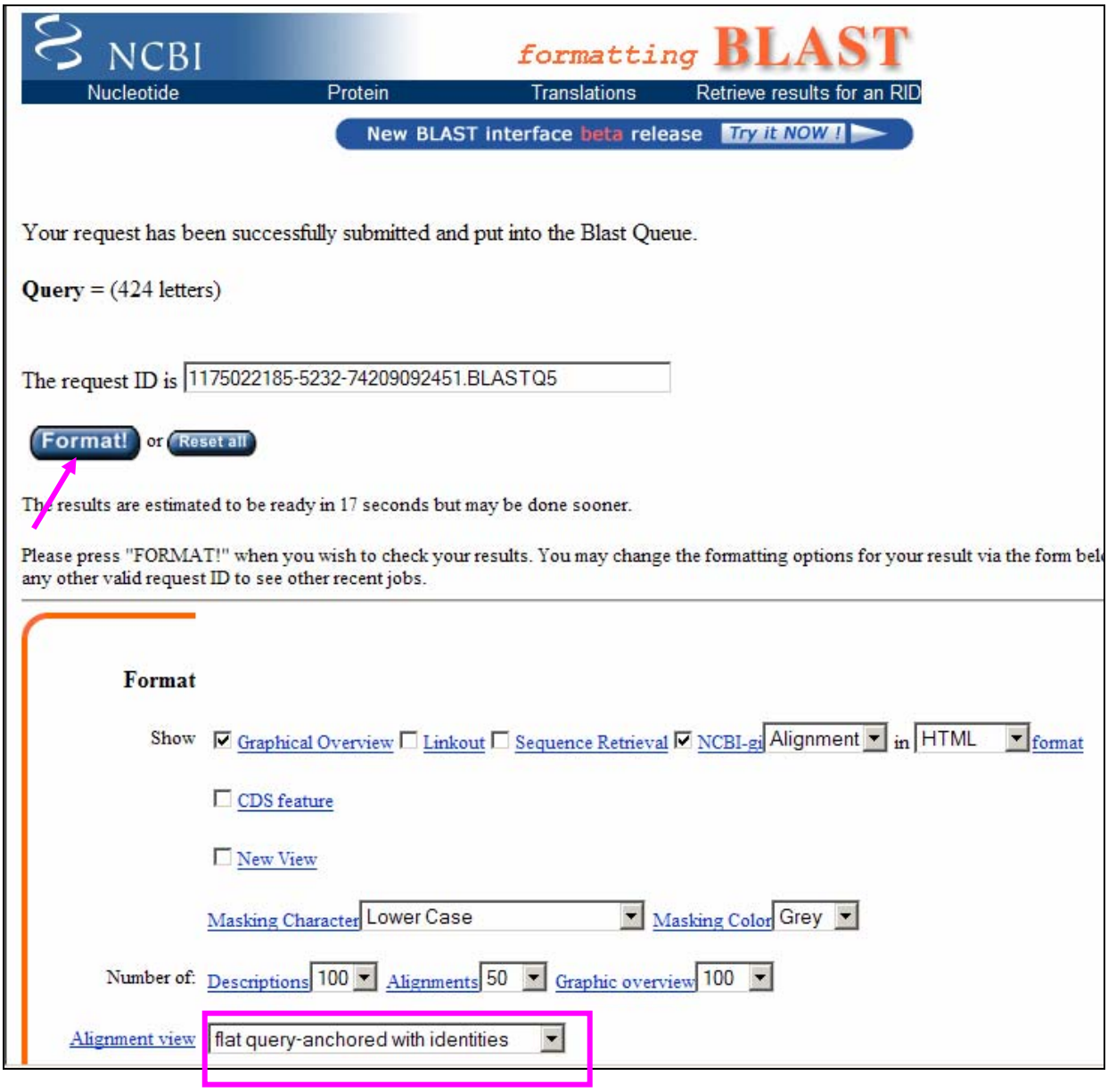
BLAST

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.

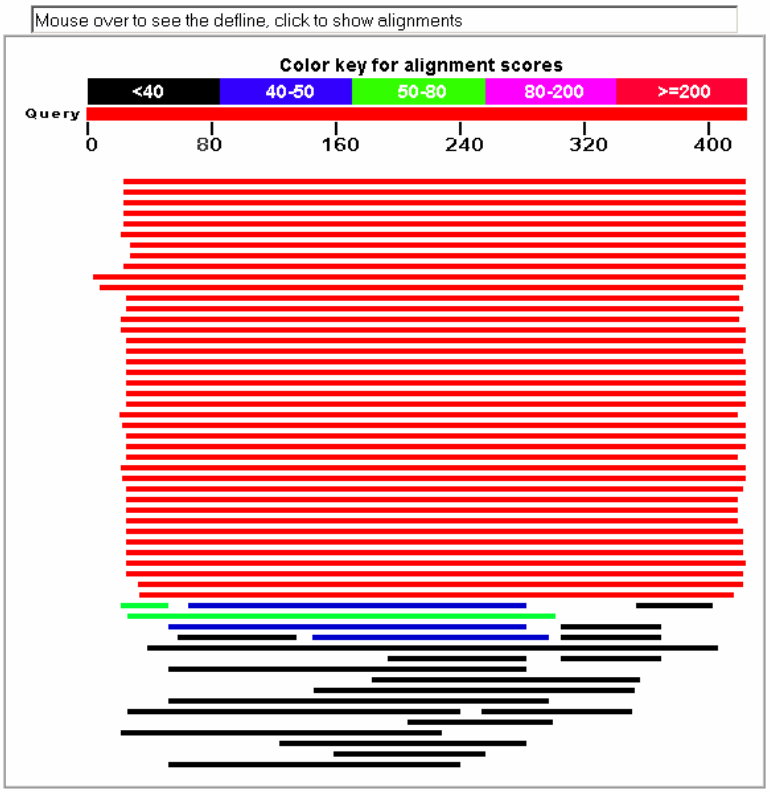
BLASTP Alignment (against SwissProt)





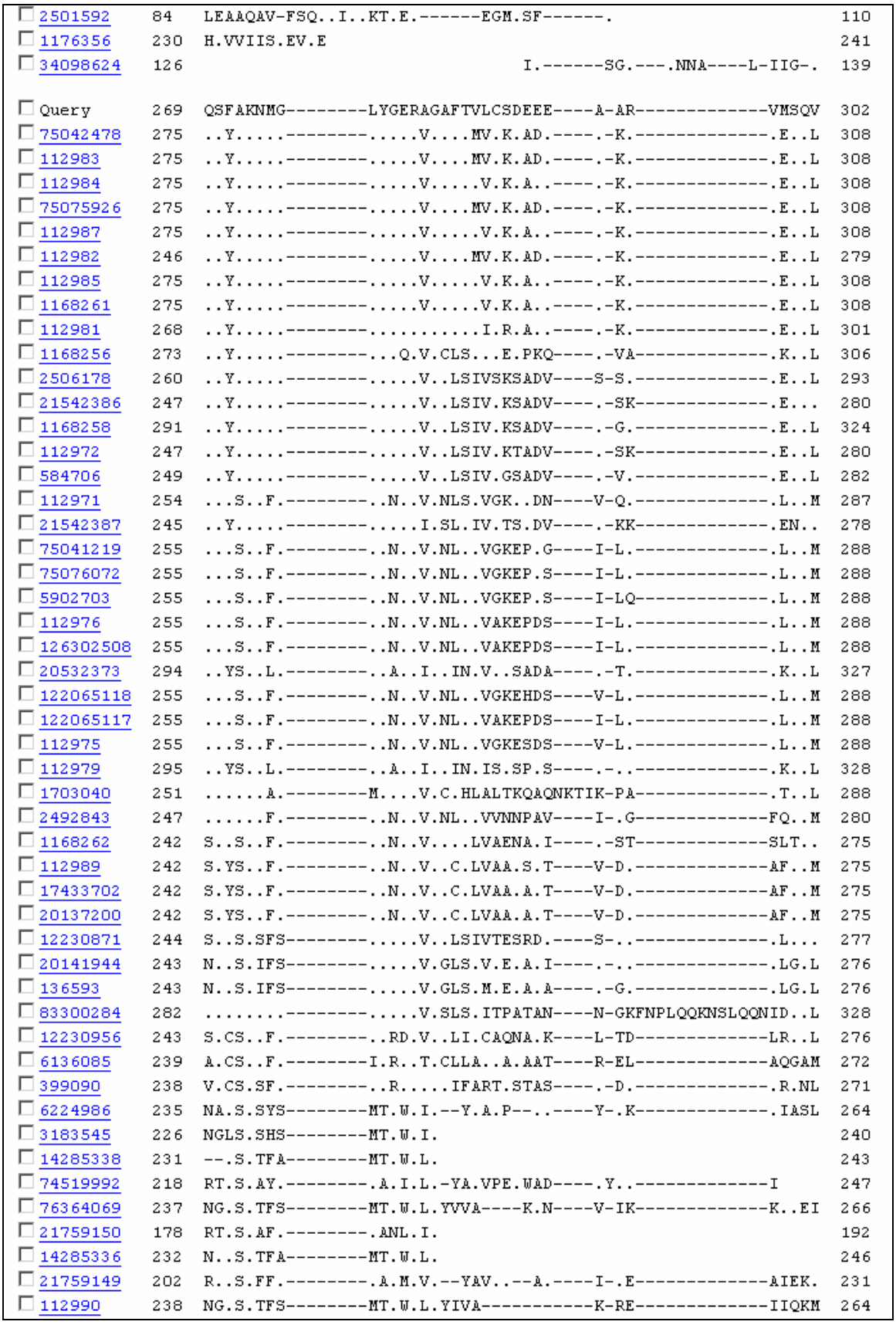

NCBI
formatting BLAST
Nucleotide
Protein
Translations
Retrieve results for an RID
New BLAST interface beta release
Try it NOW !
Your request has been successfully submitted and put into the Blast Queue.
Query = (424 letters)
The request ID is 1175022185-5232-74209092451.BLASTQ5
Format!
or
Reset all
The results are estimated to be ready in 17 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 bel
any other valid request ID to see other recent jobs.
Format
Show
Graphical Overview
Linkout
Sequence Retrieval
NCBI-gi
Alignment
in
HTML
format
CDS feature
New View
Masking Character
Lower Case
Masking Color
Grey
Number of:
Descriptions
100
Alignments
50
Graphic overview
100
Alignment view
flat query-anchored with identities

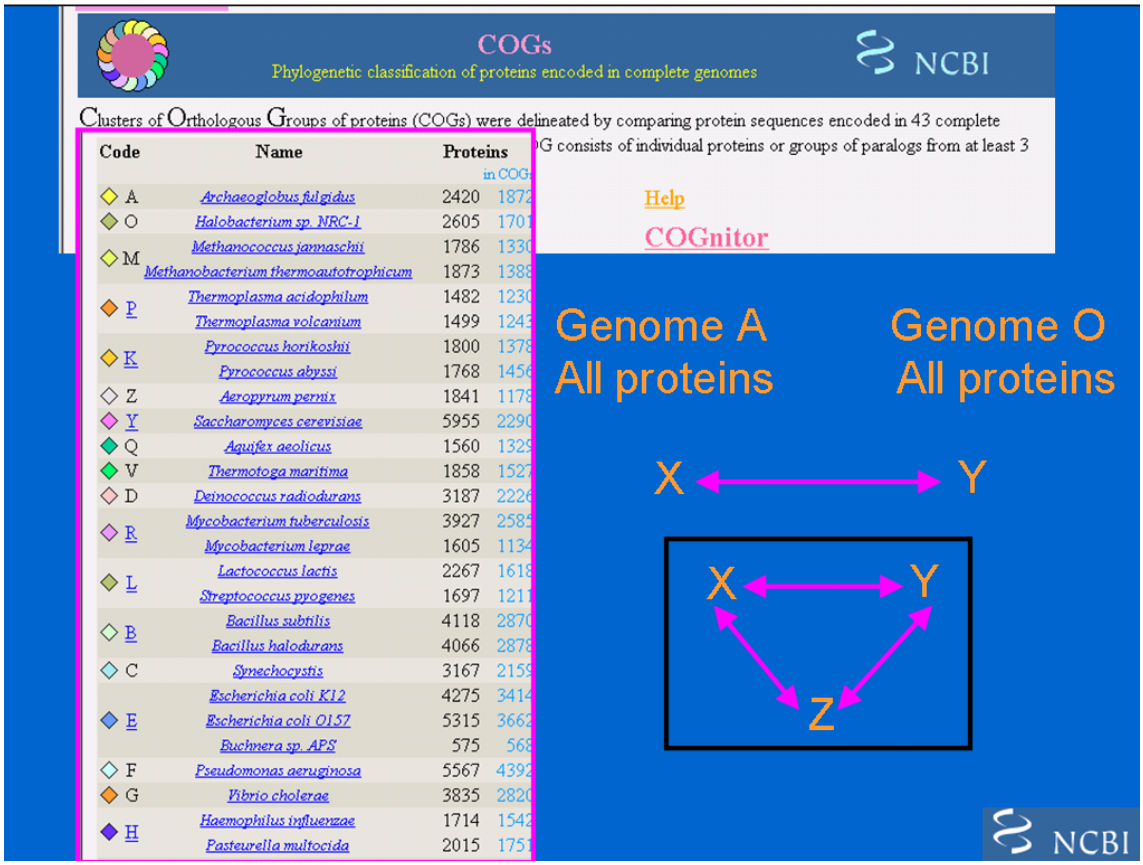
Distribution of 63 Blast Hits on the Query Sequence



Distance tree of results NEW

Sequences producing significant alignments:		Score (Bits)	E Value	
gi|75042478|sp|Q5REB0|AATM_PONPY	Aspartate aminotransferase, ...	564	3e-160	
gi|112983|sp|P00505|AATM_HUMAN	Aspartate aminotransferase, mi...	563	4e-160	G
gi|112984|sp|P05202|AATM_MOUSE	Aspartate aminotransferase, mi...	562	9e-160	G
gi|75075926|sp|Q4R559|AATM_MACFA	Aspartate aminotransferase, ...	561	1e-159	
gi|112987|sp|P00507|AATM_RAT	Aspartate aminotransferase, mito...	561	1e-159	G
gi|112982|sp|P08907|AATM_HORSE	Aspartate aminotransferase, mi...	554	2e-157	
gi|112985|sp|P00506|AATM_PIG	Aspartate aminotransferase, mito...	552	9e-157	G
gi|1168261|sp|P12344|AATM_BOVIN	Aspartate aminotransferase, m...	550	3e-156	G
gi|112981|sp|P00508|AATM_CHICK	Aspartate aminotransferase, mi...	546	6e-155	G
gi|1168256|sp|P46643|AAT1_ARATH	Aspartate aminotransferase, m...	447	5e-125	G
gi|2506178|sp|P28011|AAT1_MEDSA	Aspartate aminotransferase 1 (Tr	434	5e-121	
gi|21542386|sp|P46645|AAT2_ARATH	Aspartate aminotransferase, ...	420	4e-117	G
gi|1168258|sp|P46644|AAT3_ARATH	Aspartate aminotransferase, c...	420	4e-117	G
gi|112972|sp|P28734|AATC_DAUCA	Aspartate aminotransferase, cytop	420	6e-117	
gi|584706|sp|P37833|AATC_ORYSA	Aspartate aminotransferase, cytop	418	2e-116	


2501592 84 LEAAQAV-FSQ..I..KT.E.------EGM.SF------. 110
1176356 230 H.VVIIS.EV.E 241
34098624 126 I.------SG.---.NNA----L-IIG-. 139
Query 269 QSFAKNMG--------LYGERAGAFTVLCSDEEE----A-AR----------VMSQV 302
75042478 275 ..Y.....--------.....V....MV.K.AD.----.-K.----------.E..L 308
112983 275 ..Y.....--------.....V....MV.K.AD.----.-K.----------.E..L 308
112984 275 ..Y.....--------.....V.....V.K.A..----.-K.----------.E..L 308
75075926 275 ..Y.....--------.....V....MV.K.AD.----.-K.----------.E..L 308
112987 275 ..Y.....--------.....V.....V.K.A..----.-K.----------.E..L 308
112982 246 ..Y.....--------.....V....MV.K.AD.----.-K.----------.E..L 279
112985 275 ..Y.....--------.....V.....V.K.A..----.-K.----------.E..L 308
1168261 275 ..Y.....--------.....V.....V.K.A..----.-K.----------.E..L 308
112981 268 ..Y.....--------..........I.R.A..----.-K.----------.E..L 301
1168256 273 ..Y.....--------...Q.V.CLS...E.PKQ----.-VA----------.K..L 306
2506178 260 ..Y.....--------.....V..LSIVSKSADV----S-S.----------.E..L 293
21542386 247 ..Y.....--------.....V..LSIV.KSADV----.-SK----------.E... 280
1168258 291 ..Y.....--------.....V..LSIV.KSADV----.-G.----------.E..L 324
112972 247 ..Y.....--------.....V..LSIV.KTADV----.-SK----------.E..L 280
584706 249 ..Y.....--------.....V..LSIV.GSADV----.-V.----------.E..L 282
112971 254 ...S..F.--------..N..V.NLS.VGK..DN----V-Q.----------.L..M 287
21542387 245 ..Y.....--------.....I.SL.IV.TS.DV----.-KK----------.EN.. 278
75041219 255 ...S..F.--------..N..V.NL..VGKEP.G----I-L.----------.L..M 288
75076072 255 ...S..F.--------..N..V.NL..VGKEP.S----I-L.----------.L..M 288
5902703 255 ...S..F.--------..N..V.NL..VGKEP.S----I-LQ----------.L..M 288
112976 255 ...S..F.--------..N..V.NL..VAKEPDS----I-L.----------.L..M 288
126302508 255 ...S..F.--------..N..V.NL..VAKEPDS----I-L.----------.L..M 288
20532373 294 ..YS..L.--------..A..I..IN.V..SADA----.-T.----------.K..L 327
122065118 255 ...S..F.--------..N..V.NL..VGKEHDS----V-L.----------.L..M 288
122065117 255 ...S..F.--------..N..V.NL..VAKEPDS----I-L.----------.L..M 288
112975 255 ...S..F.--------..N..V.NL..VGKESDS----V-L.----------.L..M 288
112979 295 ..YS..L.--------..A..I..IN.IS.SP.S----.-..----------.K..L 328
1703040 251 ......A.--------M....V.C.HLALTKQAQNKTIK-PA----------.T..L 288
2492843 247 ......F.--------..N..V.NL..VVNNPAV----I-.G----------FQ..M 280
1168262 242 S..S..F.--------..N..V....LVAENA.I----.-ST----------SLT.. 275
112989 242 S.YS..F.--------..N..V..C.LVAA.S.T----V-D.----------AF..M 275
17433702 242 S.YS..F.--------..N..V..C.LVAA.A.T----V-D.----------AF..M 275
20137200 242 S.YS..F.--------..N..V..C.LVAA.A.T----V-D.----------AF..M 275
12230871 244 S..S.SFS--------.....V..LSIVTESRD.----S-..----------.L... 277
20141944 243 N..S.IFS--------.....V.GLS.V.E.A.I----.-..----------.LG.L 276
136593 243 N..S.IFS--------.....V.GLS.M.E.A.A----.-G.----------.LG.L 276
83300284 282 ........--------.....V.SLS.ITPATAN----N-GKFNPLQQKNSLQQNID..L 328
12230956 243 S.CS..F.--------..RD.V..LI.CAQNA.K----L-TD----------LR..L 276
6136085 239 A.CS..F.--------I.R..T.CLLA..A.AAT----R-EL----------AQGAM 272
399090 238 V.CS.SF.--------..R.....IFART.STAS----.-D.----------.R.NL 271
6224986 235 NA.S.SYS--------MT.W.I.--Y.A.P--..----Y-.K----------.IASL 264
3183545 226 NGLS.SHS--------MT.W.I. 240
14285338 231 --.S.TFA--------MT.W.L. 243
74519992 218 RT.S.AY.--------.A.I.L.-YA.VPE.WAD----.Y..----------I 247
76364069 237 NG.S.TFS--------MT.W.L.YVVA----K.N----V-IK----------K..EI 266
21759150 178 RT.S.AF.--------.ANL.I. 192
14285336 232 N..S.TFA--------MT.W.L. 246
21759149 202 R..S.FF.--------.A.M.V.--YAV..--A.----I-.E----------AIEK. 231
112990 238 NG.S.TFS--------MT.W.L.YIVA----------K-RE----------IIQKM 264


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 COG
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
Escherichia coli K12 4275 3414
E 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
X
Y
X
Y
Z
NCBI

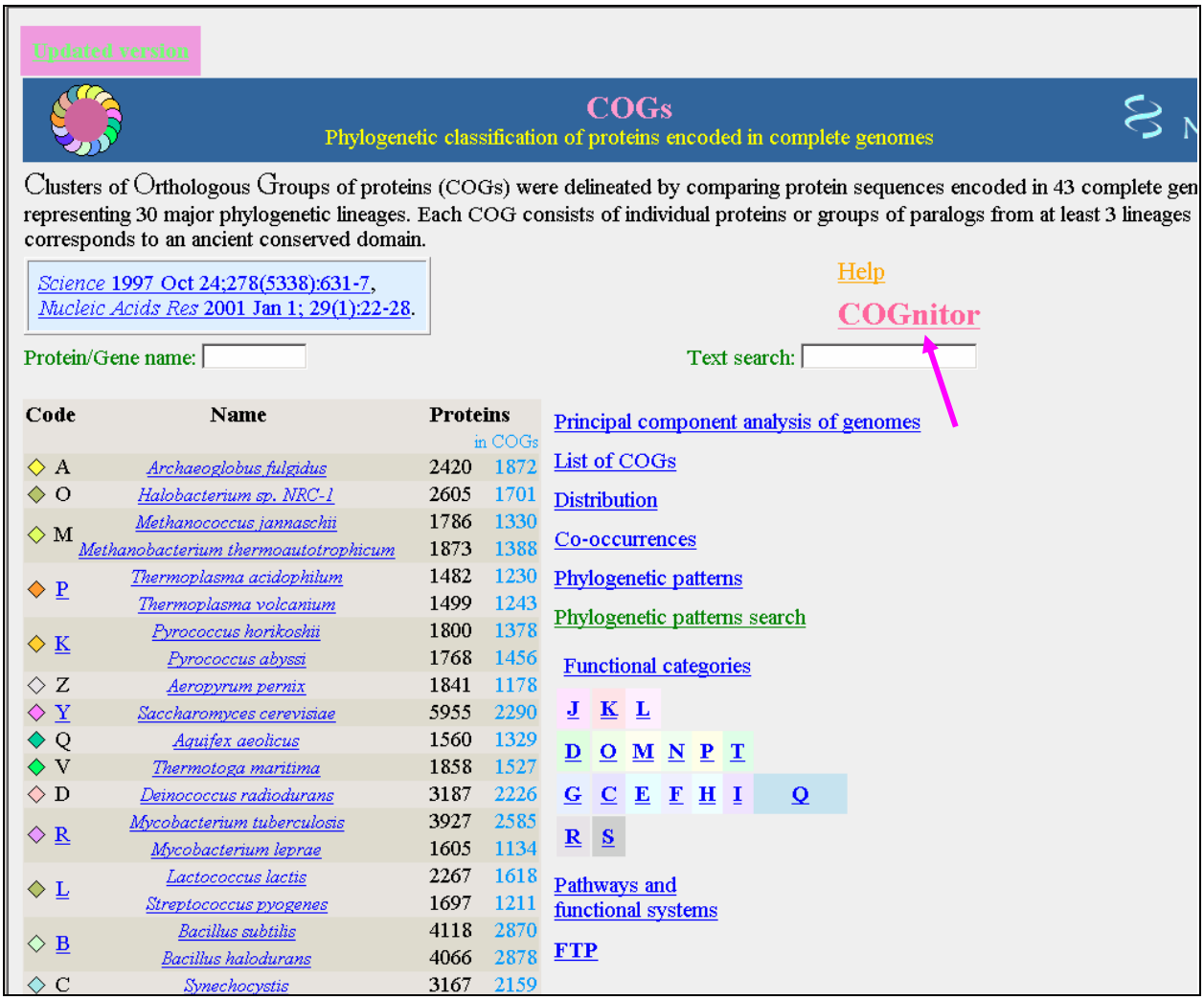
To search against the COGs database, click here:

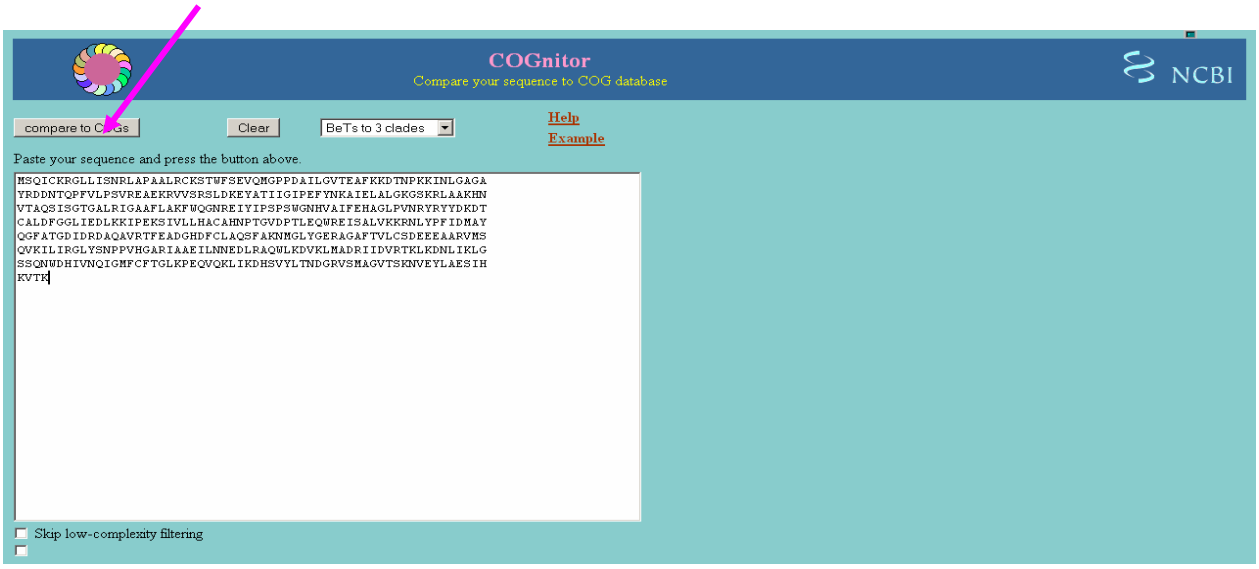
COGs

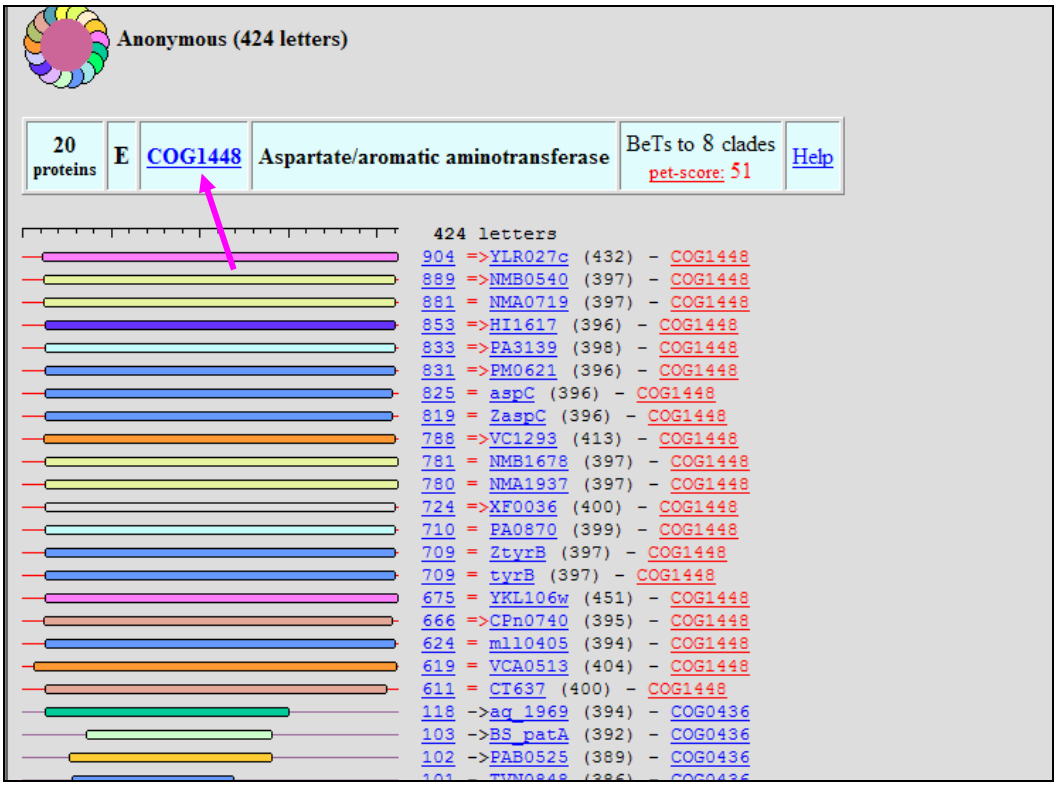
Clusters of Orthologous Groups of proteins (COGs) were delineated by comparing protein sequences encoded in 43 complete genomes, representing 30 major phylogenetic lineages. Each COG consists of individual proteins or groups of paralogs from at least 3 lineages and thus corresponds to an ancient conserved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

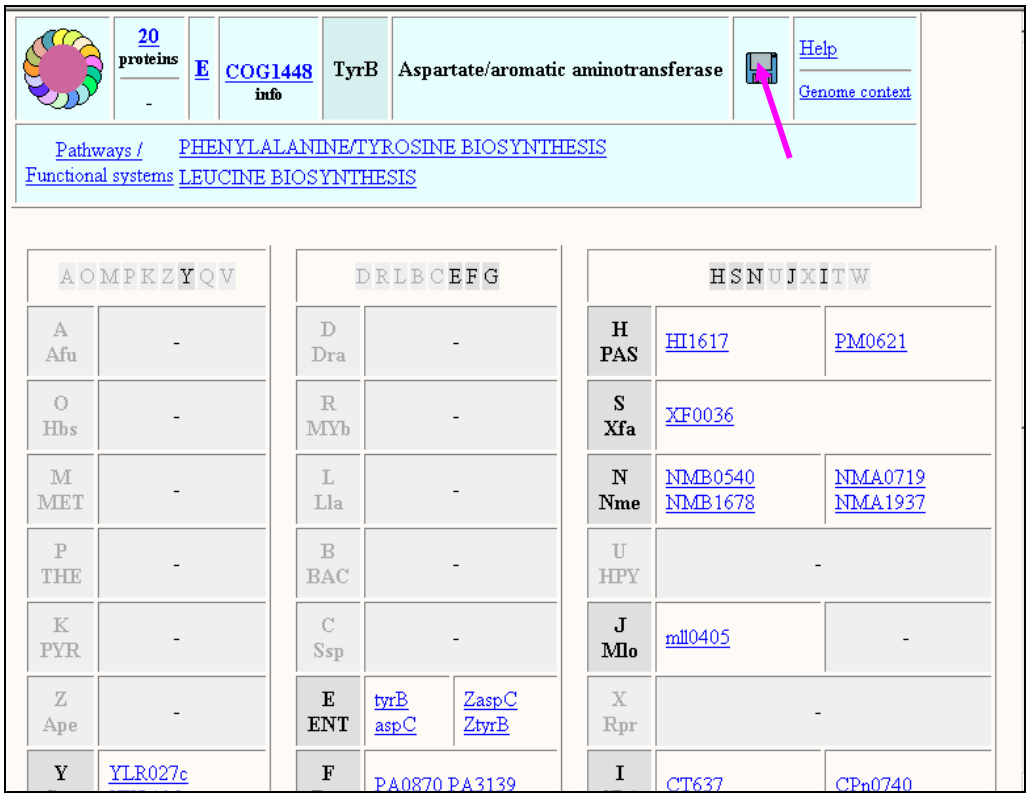
Paste the FASTA formatted protein sequence from GenScan into the COGnitor input box and press the ***"compare to COGs"*** button. Click on the link to the highest-scoring COG and click on the disk icon to save the sequences in the COG to a local file on your desktop to be used as input to Multalin below. 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
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

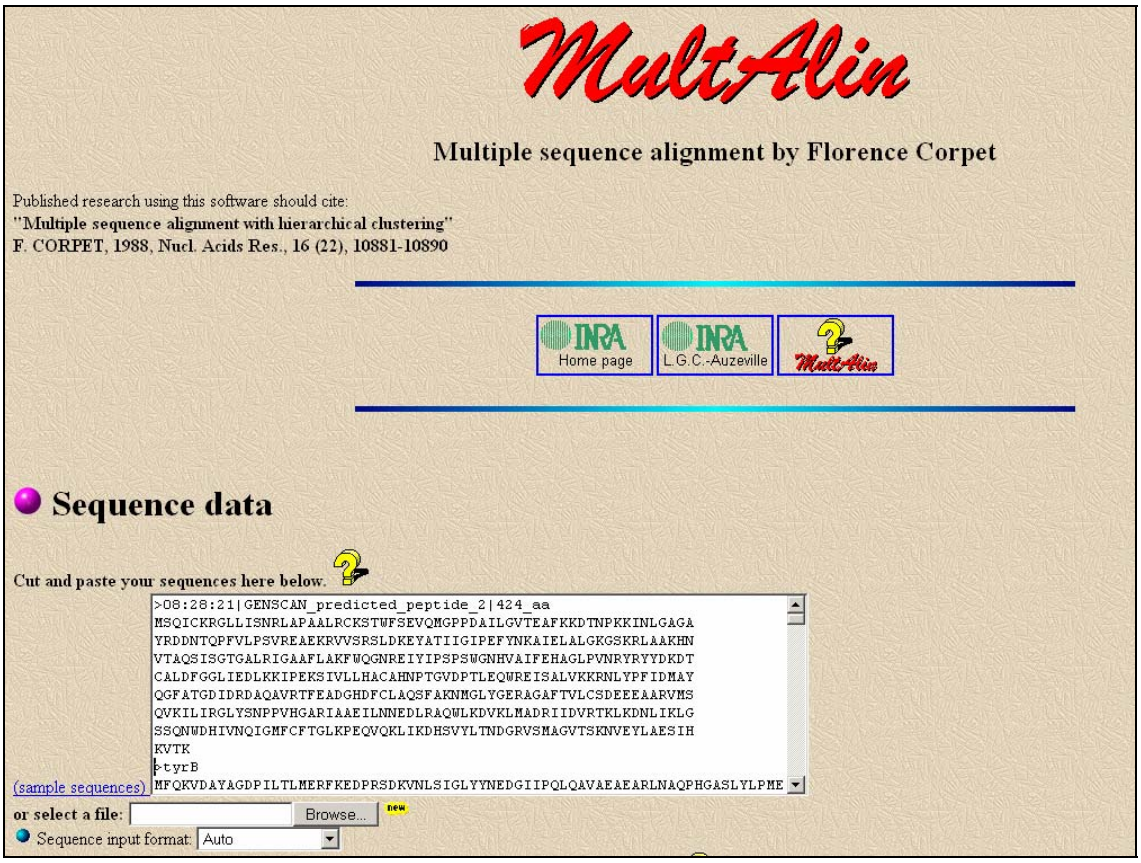

20 proteins
E
COG1448 info
TyrB
Aspartate/aromatic aminotransferase
Help
Genome context
Pathways / Functional systems
PHENYLALANINE/TYROSINE BIOSYNTHESIS
LEUCINE BIOSYNTHESIS
A O M P K Z Y Q V
D R L B C E F G
H S N U J X I T W
A Afu -
O Hbs -
M MET -
P THE -
K PYR -
Z Ape -
Y YLR027c
D Dra -
R MYb -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


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
(sample sequences)
MFQKVDAYAGDPILTLMERFKEDPRSDKVNLSIGLYYNEDGIIPQLQAVAEAEARLNAQPHGASLYLPME
or select a file: Browse... new
Sequence input format: Auto


PA0870 ALAPHGLAERFAHVGAQRGMFSYTGLSPQQVARLRDEHAVYLVSSGRANVAGLDARRLDRLAQAIAQVCAD
CT637 AMRNV-AGHSFDFIASQKGFFGYPGFSKEQVLFLREELGIYTTAGGRFNLNGITDKNINRVTHGFAQAYEYPRSVS
CPn0740 ALRKV-AGHTFDFLLSQHGFFAYPGFSDKQVLFLREQHAVYTTAGGRMNLNGITEKNIDHVVQSFIQAYEL
YKL106w RL----GWPDLVNFAQQHGMFYYTRFSPKQVEILRNNYFVYLTGDGRLSLSGVNDSNVDYLCESLEAVSKMDKLA
Consensus .L.....fdfi..q.GmFsy.gls..QV.rLr.#..!Y.v..GR..vagl...n..yla.ai..v.....
Available files:
-Sequence Input file
-Cluster file
-Results as a fasta file
-Results as a text page (msf)
-Results as postscript page(s) with ESPript (protein only) new
-Alignment and tree description (rfd) new Get a better view of your protein family : phylogenetic tree, pruned tree and subtrees, s
coloured alignment and subalignments.
- Results as an html page (needs to enable style sheets) new
-Results as a text page with colour indications (need a text editor)
-Results as a gif image
Add one sequence to the alignment new
Cut and paste your sequence here below (FASTA/MULTALIN FORMAT ONLY).

```
      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

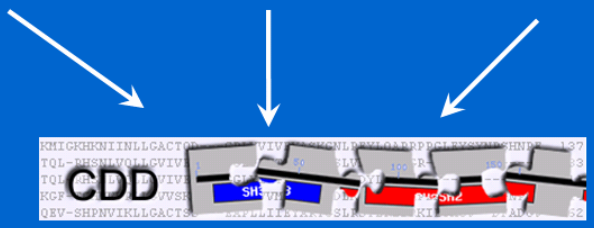
Conserved Domain

- 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 COGs

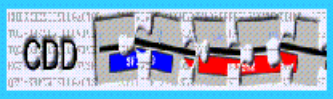


Curated CDs

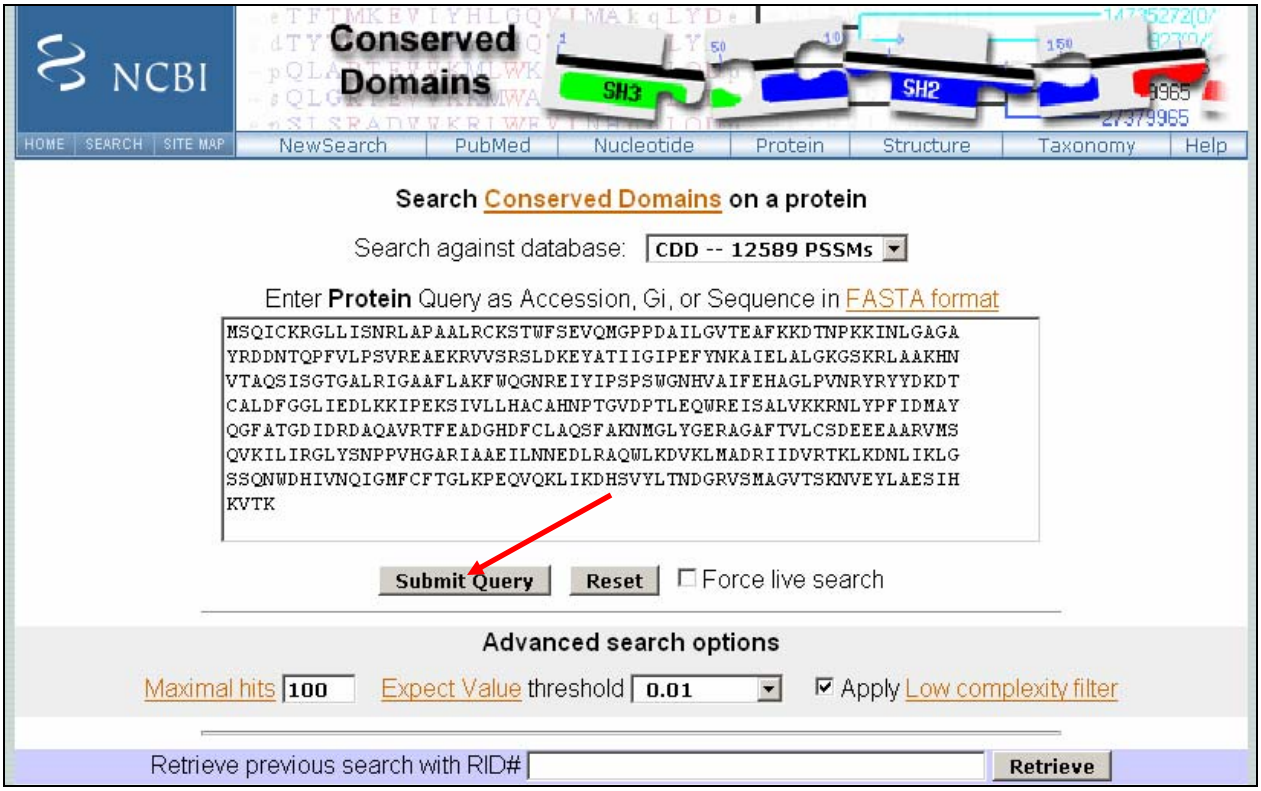
- A position-specific scoring matrix (PSSM) is calculated
- CD-Search can be used to search against the PSSMs
- Manual curation of CDs has begun

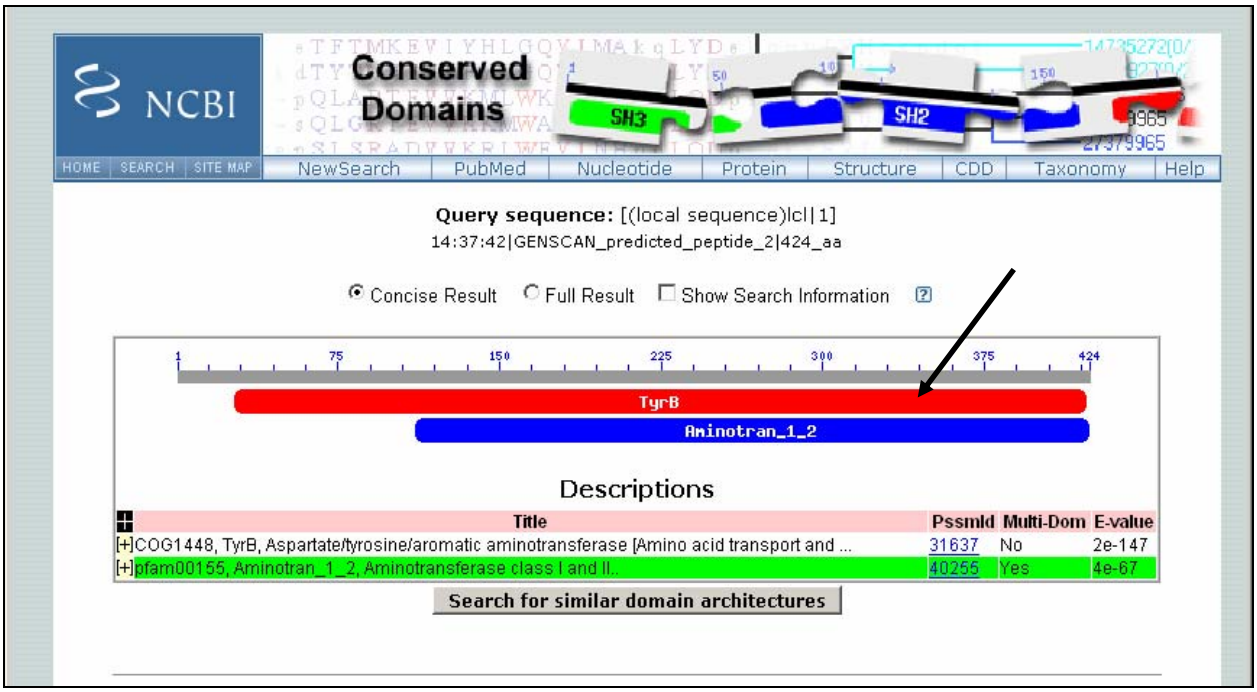
NCBI

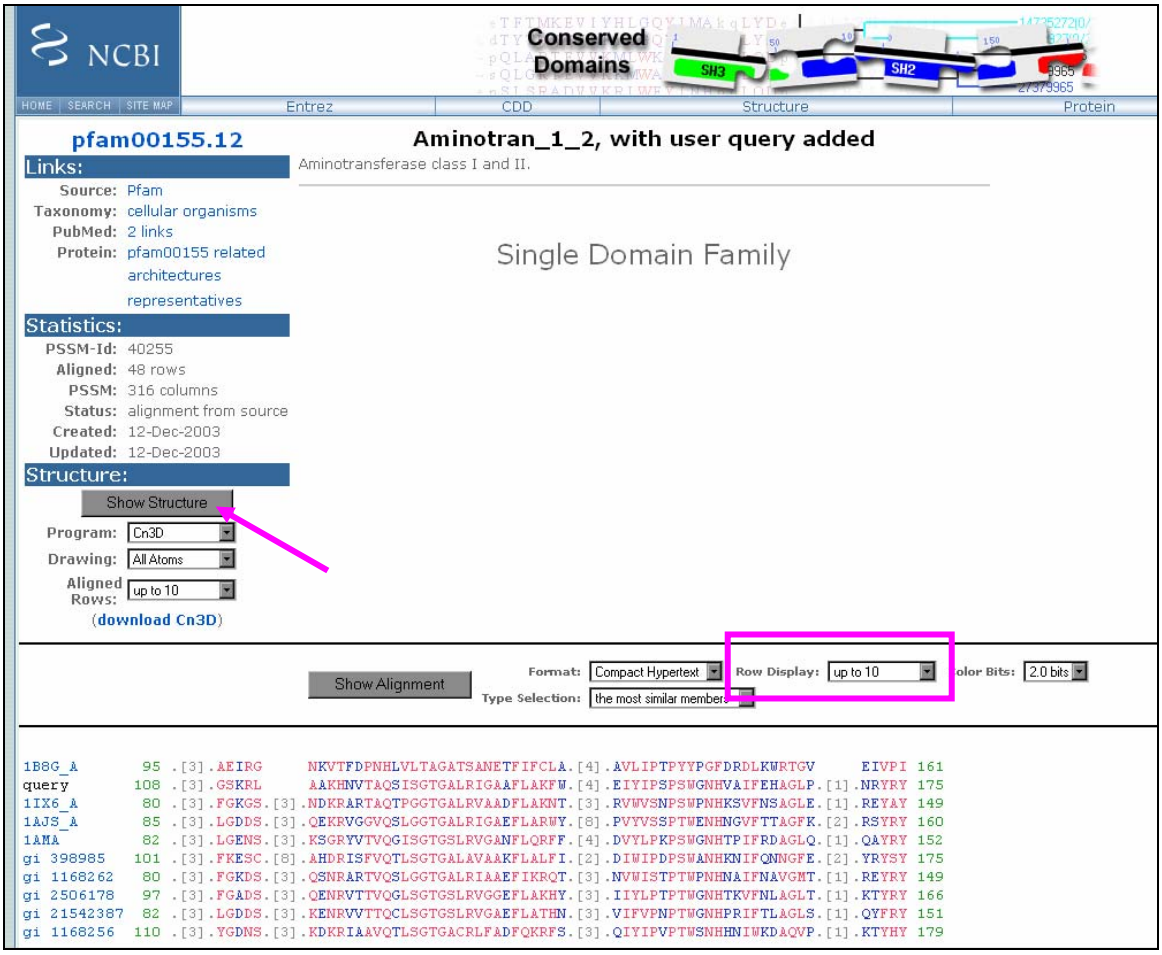
To search for protein domains and view a model structure for your protein, click here:

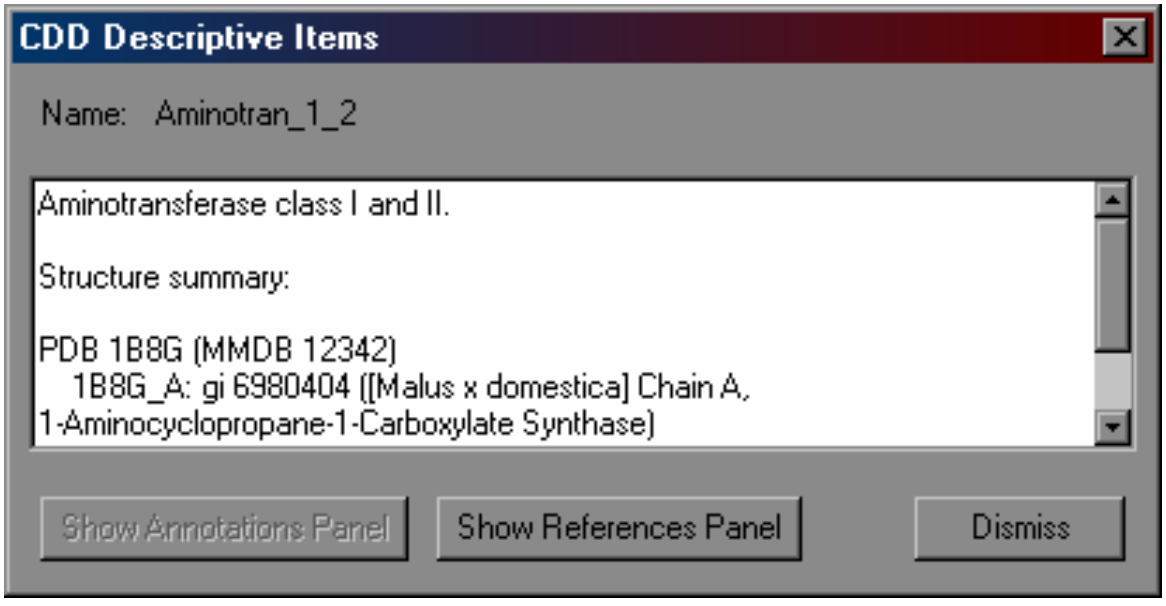


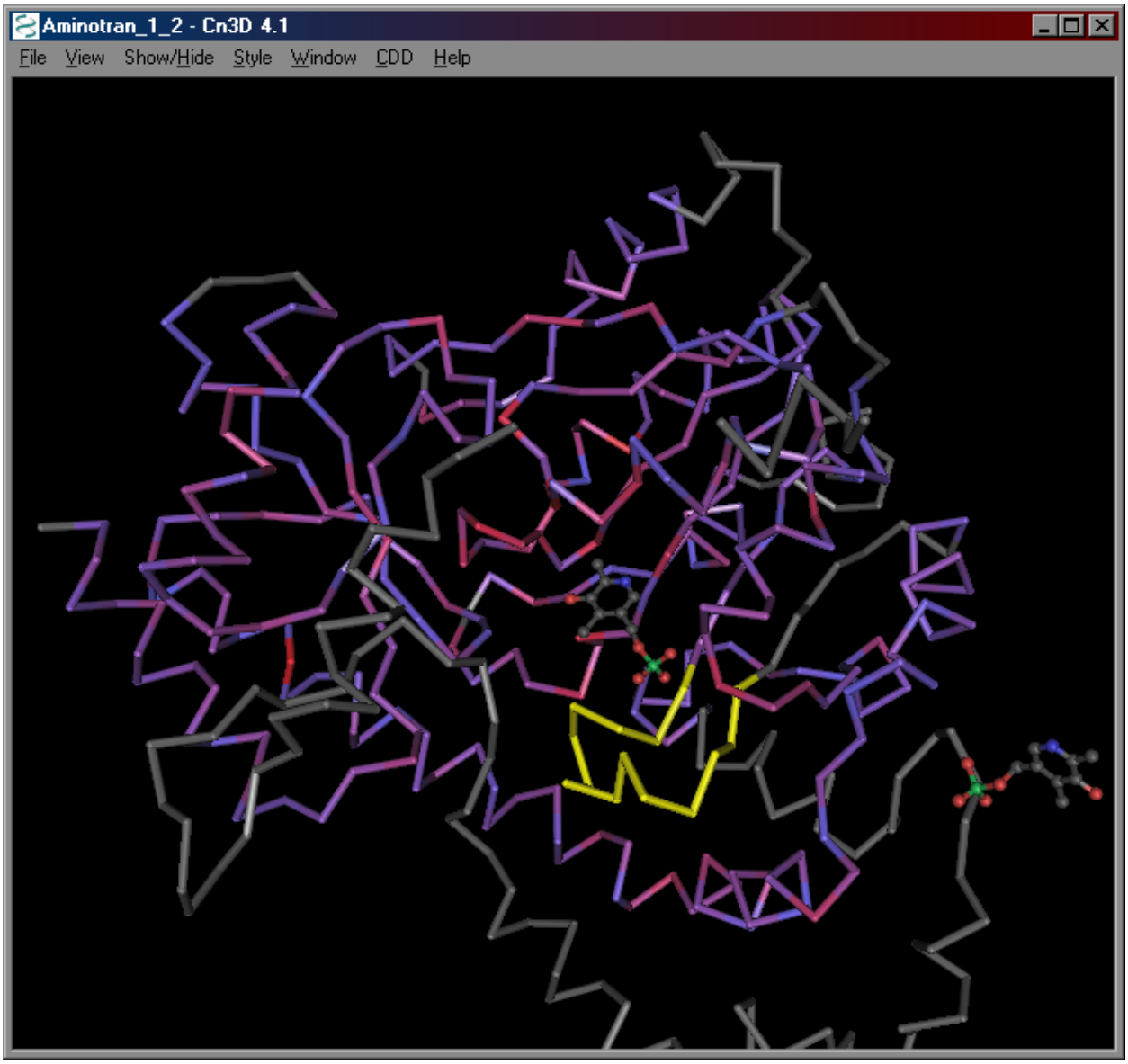
NCBI's Conserved Domain Search allows you to match your protein sequence to a library of conserved protein domains, generate a multiple sequence alignment based on this match, and explore 3D modeling templates for your sequence.
Paste your protein sequence from GenScan into the CD–Search query box and 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.



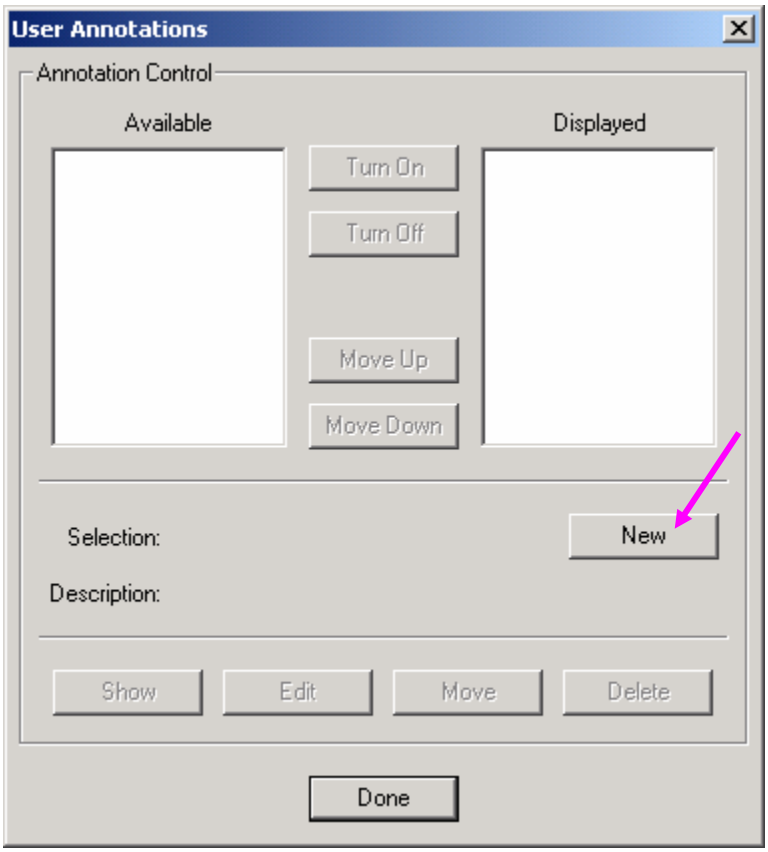

NCBI
Conserved Domains
SH3
SH2
HOME
SEARCH
SITE MAP
NewSearch
PubMed
Nucleotide
Protein
Structure
CDD
Taxonomy
Help
Query sequence: [(local sequence)lcl|1]
14:37:42|GENSCAN_predicted_peptide_2|424_aa
Concise Result
Full Result
Show Search Information
1
75
150
225
300
375
424
TyrB
Aminotran_1_2
Descriptions
Title
Pssmid
Multi-Dom
E-value
[+]COG1448, TyrB, Aspartate/tyrosine/aromatic aminotransferase [Amino acid transport and ...
31637
No
2e-147
[+]pfam00155, Aminotran_1_2, Aminotransferase class I and II..
40255
Yes
4e-67
Search for similar domain architectures

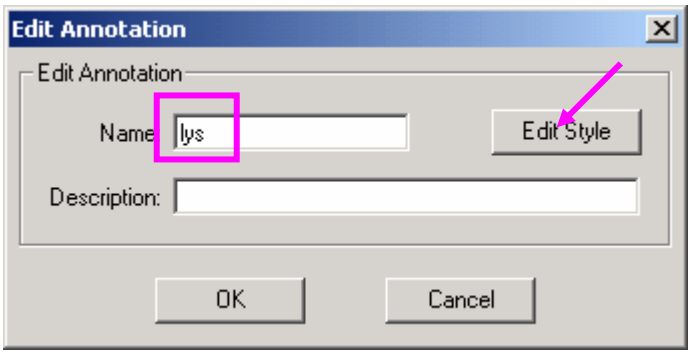

NCBI
Conserved Domains
SH3
SH2
HOME | SEARCH | SITE MAP
Entrez
CDD
Structure
Protein
pfam00155.12
Aminotran_1_2, with user query added
Aminotransferase class I and II.
Links:
Source: Pfam
Taxonomy: cellular organisms
PubMed: 2 links
Protein: pfam00155 related architectures representatives
Statistics:
PSSM-Id: 40255
Aligned: 48 rows
PSSM: 316 columns
Status: alignment from source
Created: 12-Dec-2003
Updated: 12-Dec-2003
Structure:
Show Structure
Program: Cn3D
Drawing: All Atoms
Aligned Rows: up to 10
(download Cn3D)
Single Domain Family
Show Alignment
Format: Compact Hypertext
Row Display: up to 10
Color Bits: 2.0 bits
Type Selection: the most similar members
1B8G_A 95 .[3].AEIRG NKVTFDPNHLVLTAGATSANETFIFCLA.[4].AVLIPTPYYPGFDRDLKWRTGV EIVPI 161
query 108 .[3].GSKRL AAKHNVTAQSISGTGALRIGAAFLAKFW.[4].EIYIPSPSWGNHVAIFEHAGLP.[1].NRYRY 175
1IX6_A 80 .[3].FGKGS.[3].NDKRARTAQTPGGTGALRVAADFLAKNT.[3].RVWVSNPSWPNHKSVFNSAGLE.[1].REYAY 149
1AJS_A 85 .[3].LGDDS.[3].QEKRVGGVQSLGGTGALRIGAEFLARWY.[8].PVYVSSPTWENHNGVFTTAGFK.[2].RSYRY 160
1AMA 82 .[3].LGENS.[3].KSGRYVTVQGISGTGSLRVGANFLQRFF.[4].DVYLPKPSWGNHTPIFRDAGLQ.[1].QAYRY 152
gi 398985 101 .[3].FKESC.[8].AHDRISFVQTLSGTGALAVAAKFLALFI.[2].DIWIPDPSWANHKNIFQNNGFE.[2].YRYSY 175
gi 1168262 80 .[3].FGKDS.[3].QSNRARTVQSLGGTGALRIAAEFIKRQT.[3].NVWISTPTWPNHNAIFNAVGMT.[1].REYRY 149
gi 2506178 97 .[3].FGADS.[3].QENRVTTVQGLSGTGSLRVGGEFLAKHY.[3].IIYLPTPTWGNHTKVFNLAGLT.[1].KTYRY 166
gi 21542387 82 .[3].LGDDS.[3].KENRVVTTQCLSGTGSLRVGAEFLATHN.[3].VIFVPNPTWGNHPRIFTLAGLS.[1].QYFRY 151
gi 1168256 110 .[3].YGDNS.[3].KDKRIAAVQTLSGTGACRLFADFQKRFS.[3].QIYIPVPTWSNHHNIWKDAQVP.[1].KTYHY 179

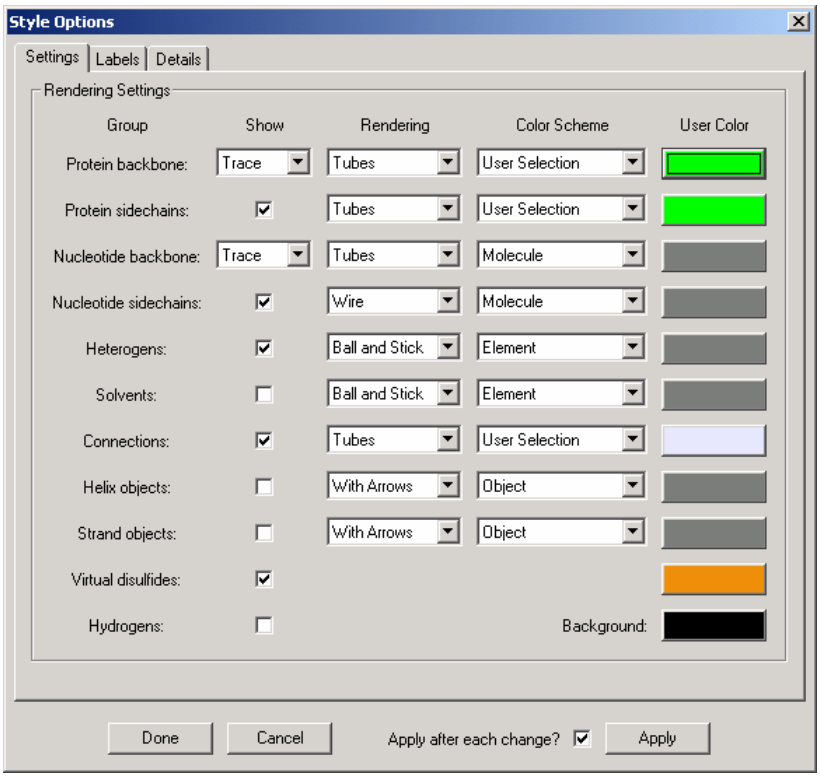

CDD Descriptive Items
Name: Aminotran_1_2
Aminotransferase class I and II.
Structure summary:
PDB 1B8G (MMDB 12342)
1B8G_A: gi 6980404 ([Malus x domestica] Chain A, 1-Aminocyclopropane-1-Carboxylate Synthase)
Show Annotations Panel
Show References Panel
Dismiss

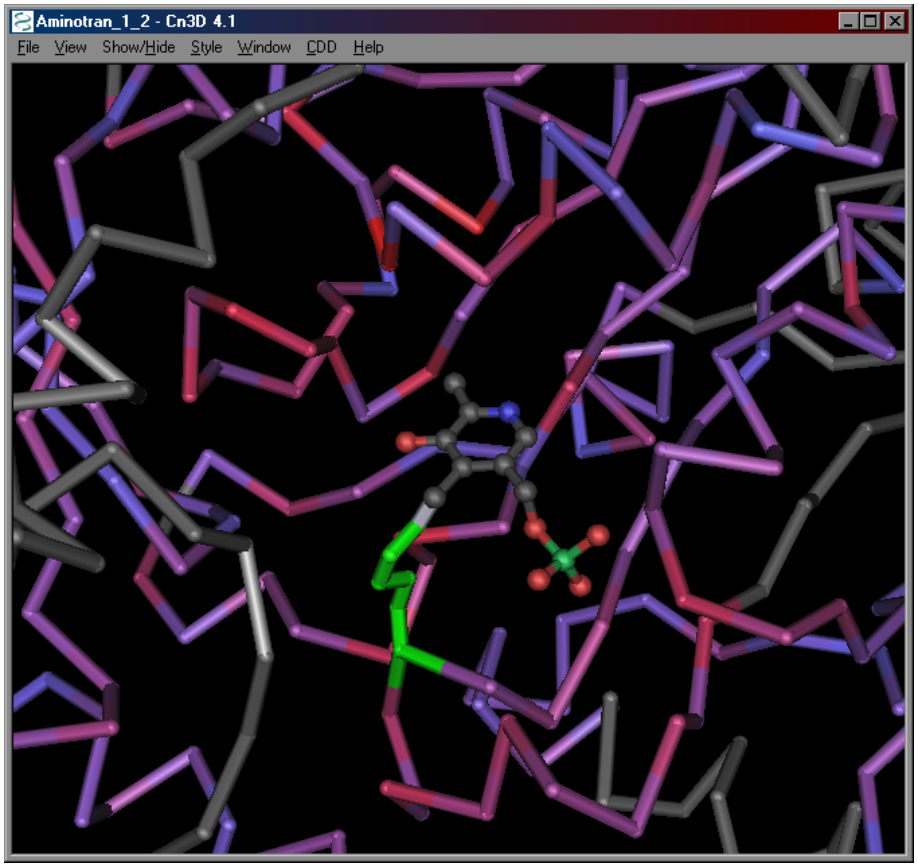

Aminotran_1_2 - Cn3D 4.1
File View Show/Hide Style Window CDD Help


Aminotran_1_2 - Sequence/Alignment Viewer
View Edit Mouse Mode Unaligned Justification Imports
1B8G_A
query
1IX6_A
1AJS_A
1AMA
gi 398985
gi 1168262
gi 2506178
gi 21542387


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