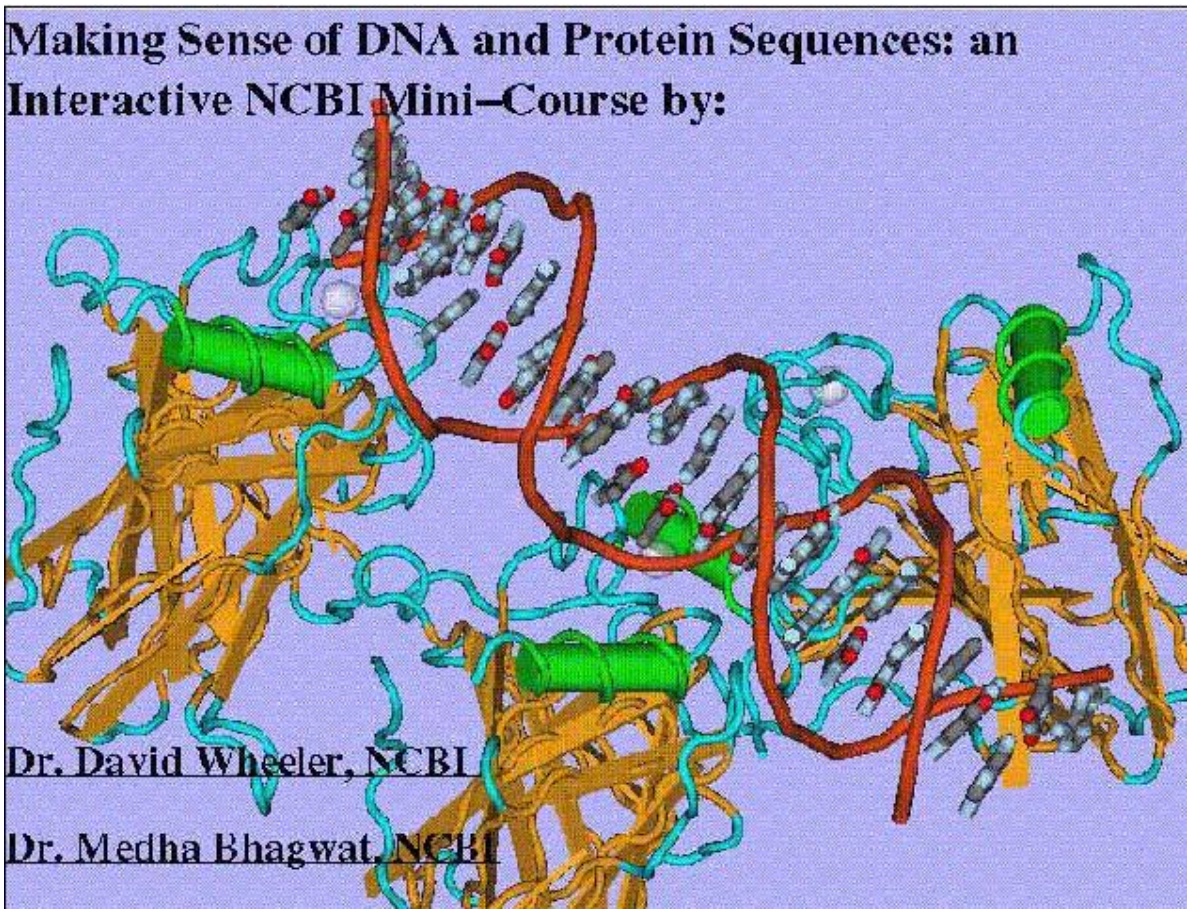
Making Sense of DNA and Protein Sequences: an Interactive NCBI Mini–Course by:



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 encodes 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:

Class Page: http://www.ncbi.nlm.nih.gov/Class/minicourses/
GenScan: http://genes.mit.edu/GENSCAN.html
ScanProsite: http://www.expasy.org/tools/scanprosite/
BLASTP: http://www.ncbi.nlm.nih.gov/BLAST
COGs: http://www.ncbi.nlm.nih.gov/COG/old/
MultAlin: http://prodes.toulouse.inra.fr/multalin/multalin.html
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: 2-Jan-108 Time: 08:08:13

Sequence 08:08:12 : 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):

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

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

>08:08:12|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 [help]

Sequence(s) to be scanned:

Enter:

- UniProtKB(Swiss-Prot and TrEMBL) AC and/or ID *(e.g. P00747, ENTK_HUMAN)*
- PDB identifier(s)
- your own protein sequence(s)

```
VTAQSISGTGALRIGAAFLAKFWQGNREIYIPSPSWGNHVAIFEHAGLPV
NRYRYYDKDT
CALDFGGLIEDLKKIPEKSIVLLHACAHNPTGVDPTLEQWREISALVKKR
NLYPFIDMAY
QGFATGDIDRDAQAVRTFEADGHDFCLAQSFAKNMGLYGERAGAFTVLCS
DEEEAARVMS
QVKILIRGLYSNPPVHGARIAAEILNNEDLRAQWLKDVKLMADRIIDVRT
KLKDNLIKLG
SSQNWDHIVNQIGMFCFTGLKPEQVQKLIKDHSVYLTNDGRVSMAGVTSK
NVEYLAESIH
KVTK
```

Clear

☑ Exclude motifs with a high probability of occurrence

☐ Do not scan profiles

Motif(s) to scan for:

Enter:

- PROSITE AC and/or ID *(e.g. PS50240, CHEB)*
- your own pattern(s)

Clear

Protein database(s):

☑ UniProtKB/Swiss-Prot ☑ including splice variants ☐ UniProtKB/TrEMBL ☐ PDB

randomize databases no

Filter(s):

- On taxonomy: *(e.g.Eukaryota; Escherichia coli;)*
- On description: *(e.g.protease)*

Pattern option(s):

- Allow at most 0 X sequence characters to match a conserved position in the pattern
- Match mode greedy, overlaps, no includes

Output:

- Format Graphical rich view
- Show only sequences with at least hit(s)

Maximum of matched sequences 1000

☐ Show low level score

☐ Retrieve complete sequences

Your e-mail:

START THE SCAN reset

Home | ScanProsite | ProRule | Documents | Downloads | Links

ScanProsite Results Viewer

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

Hits for all PROSITE (release 20.24) motifs on sequence 08-08-12-GENSCAN_predicted_peptide_2-424_aa :

found: 1 hit in 1 sequence

08-08-12-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* :

08-08-12-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

Home ScanProsite ProRule Documents Downloads Links

PROSITE documentation PDOC00098 [for PROSITE entry PS00105]

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 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

- Retrieve an alignment of Swiss-Prot true positive hits:
  Clustal format, color, condensed view / Clustal format, color / Clustal format, plain text / Fasta format
- Taxonomic tree view of all Swiss-Prot/TrEMBL entries matching PS00105
- Retrieve a list of all Swiss-Prot/TrEMBL entries matching PS00105
- Scan Swiss-Prot/TrEMBL entries against PS00105
- view ligand binding statistics

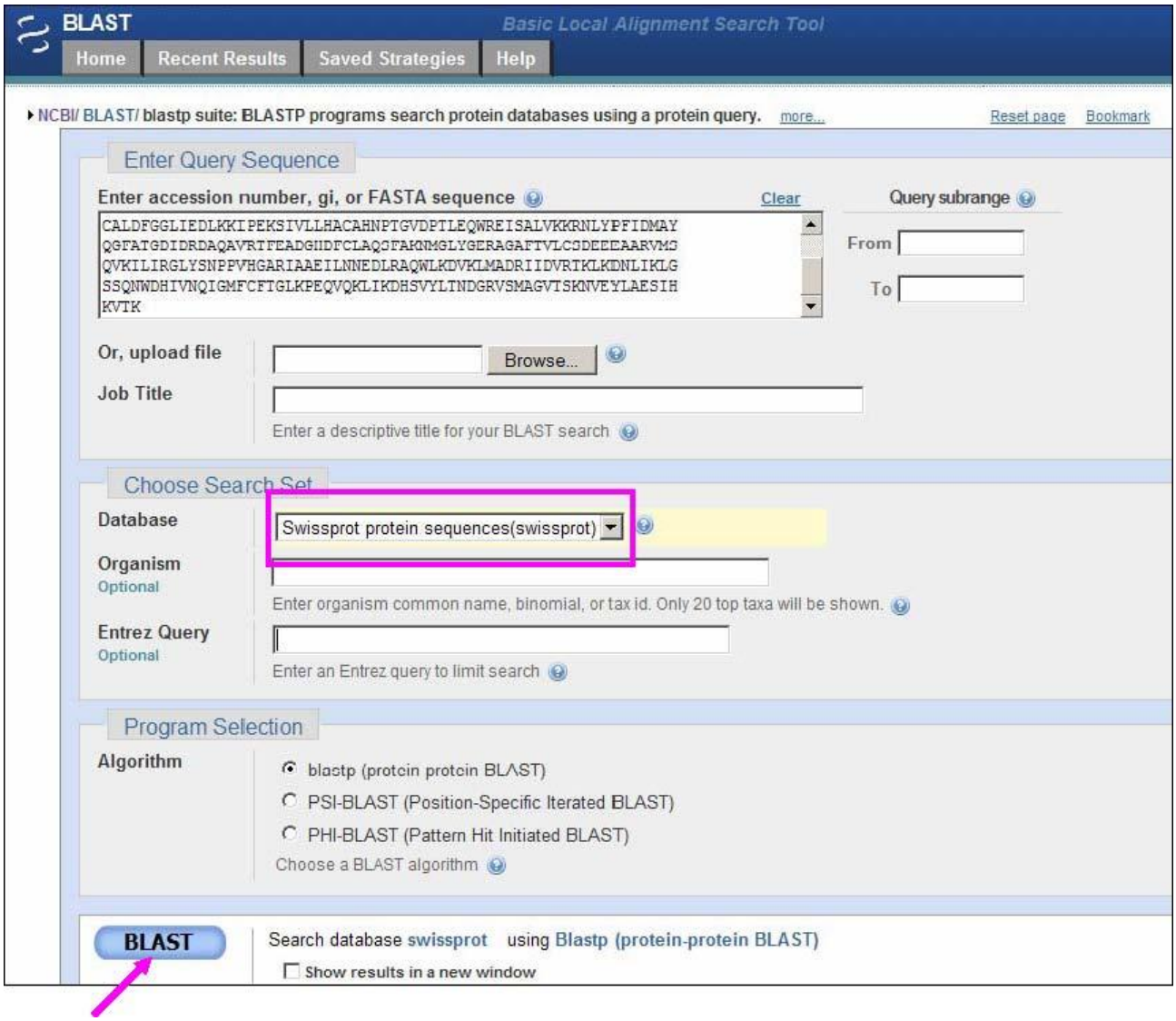
Matching PDB structures: 1AAM 1AAT 1AAW 1AHE ... [ALL]

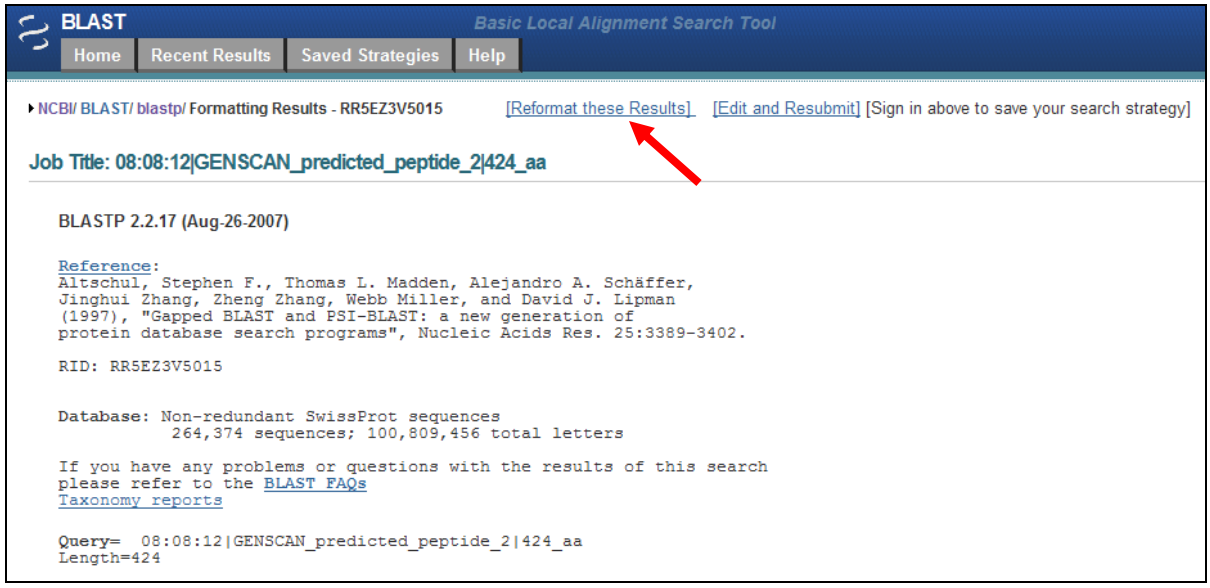
To search for proteins with similar sequences, use BLAST:

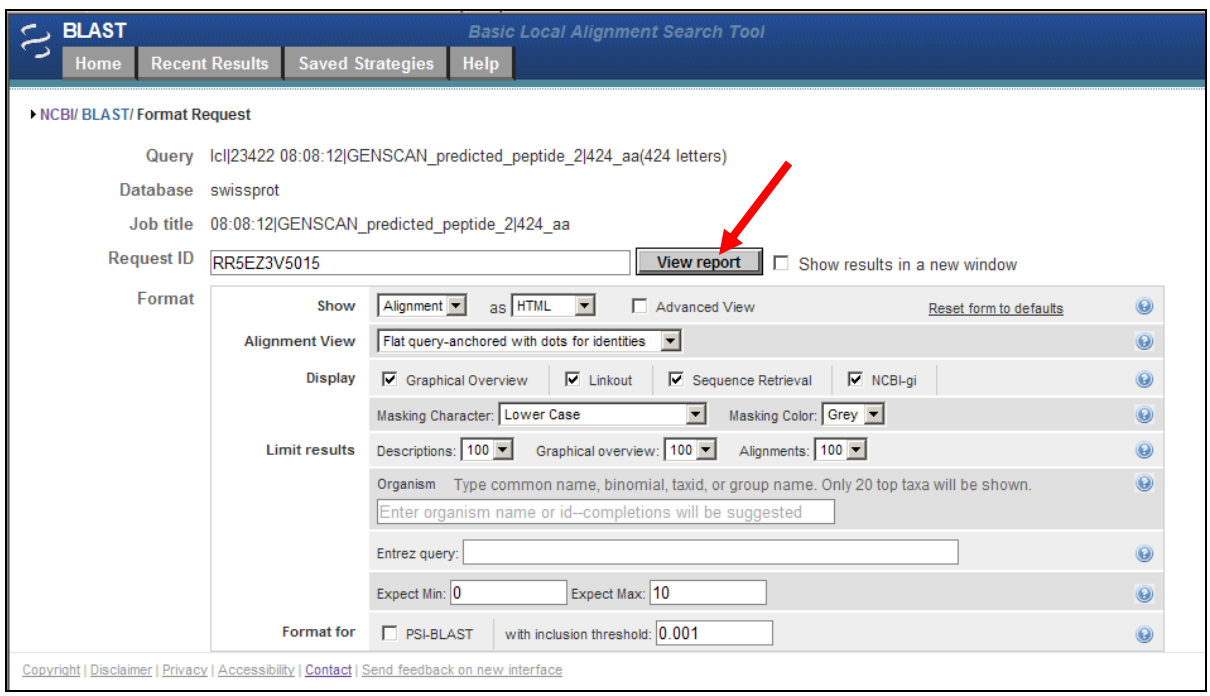
BLAST

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
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
NCBI/ BLAST/ blastp/ Formatting Results - RR5EZ3V5015
[Reformat these Results]
[Edit and Resubmit] [Sign in above to save your search strategy]
Job Title: 08:08:12|GENSCAN_predicted_peptide_2|424_aa
BLASTP 2.2.17 (Aug-26-2007)
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.
RID: RR5EZ3V5015
Database: Non-redundant SwissProt sequences
264,374 sequences; 100,809,456 total letters
If you have any problems or questions with the results of this search
please refer to the BLAST FAQs
Taxonomy reports
Query= 08:08:12|GENSCAN_predicted_peptide_2|424_aa
Length=424


BLAST
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
NCBI/ BLAST/ Format Request
Query lcl|23422 08:08:12|GENSCAN_predicted_peptide_2|424_aa(424 letters)
Database swissprot
Job title 08:08:12|GENSCAN_predicted_peptide_2|424_aa
Request ID RR5EZ3V5015
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

NCBI/ BLAST/ blastp/ Formatting Results - RR5EZ3V5015 [Reformat these Results] [Edit and Resubmit] [Sign in above to save your search strategy]

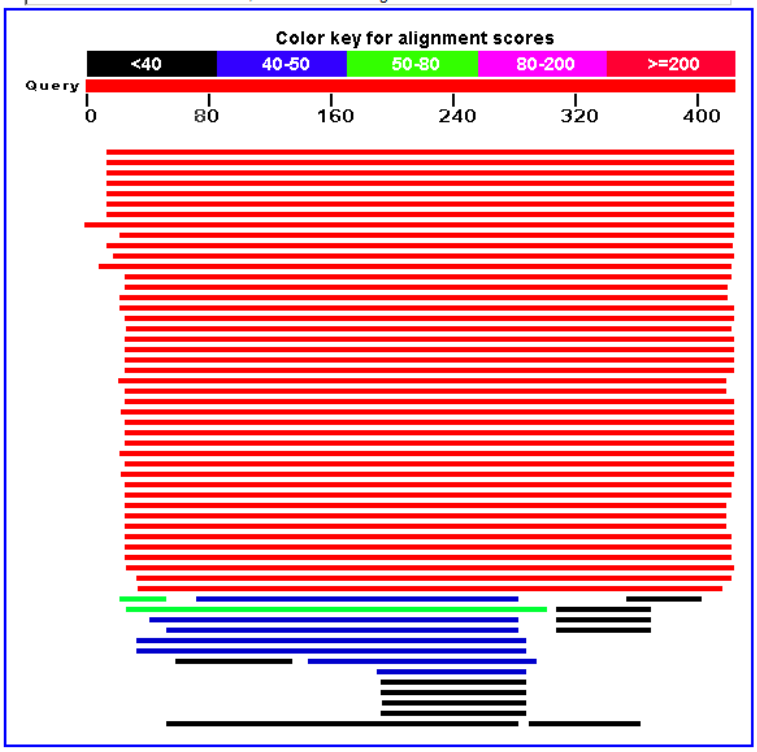
Job Title: 08:08:12|GENSCAN_predicted_peptide_2|424_aa

BLASTP 2.2.17 (Aug-26-2007)

```
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.

RID: RR5EZ3V5015


Database: Non-redundant SwissProt sequences
           264,374 sequences; 100,809,456 total letters

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

Query=  08:08:12|GENSCAN_predicted_peptide_2|424_aa
Length=424
```

Distribution of 91 Blast Hits on the Query Sequence

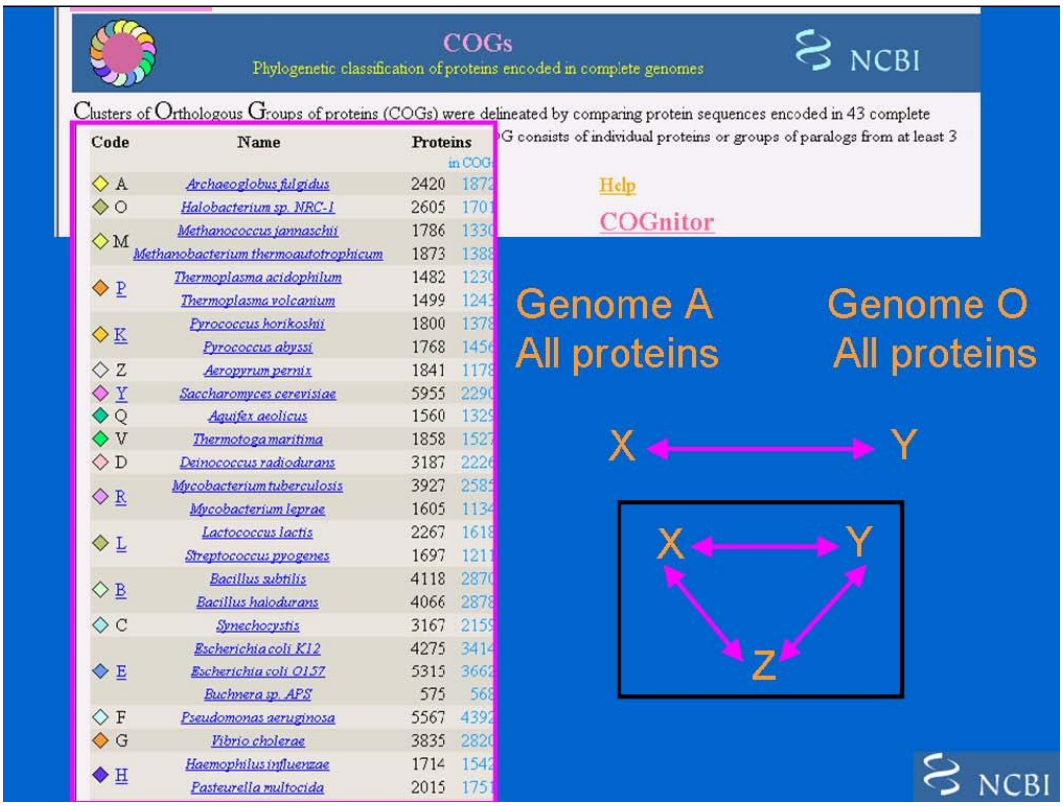
Mouse over to see the define, click to show alignments



Distance tree of results NEW

Sequences producing significant alignments:	Score (Bits)	E Value	
gi|75042478|sp|Q5REB0.1|AATM_PONPY Aspartate aminotransferase...	545	2e-154	
gi|112983|sp|P00505.2|AATM_HUMAN Aspartate aminotransferase, ...	543	4e-154	G
gi|112984|sp|P05202.1|AATM_MOUSE Aspartate aminotransferase, ...	543	5e-154	G
gi|112987|sp|P00507.2|AATM_RAT Aspartate aminotransferase, mi...	543	7e-154	G
gi|75075926|sp|Q4R559.1|AATM_MACFA Aspartate aminotransferase...	542	1e-153	

```
Query       265  FCL-AQ--SFAKNMGLYGERAGAFTVLCSDEEE----AAR-------------VMSQVK-  303
75042478    271  V..-C.--.Y..........V....MV.K.AD.----.K.--------------.E..L.-  309
112983      271  V..-C.--.Y..........V....MV.K.AD.----.K.--------------.E..L.-  309
112984      271  V..-C.--.Y..........V.....V.K.A..----.K.--------------.E..L.-  309
112987      271  V..-C.--.Y..........V.....V.K.A..----.K.--------------.E..L.-  309
75075926    271  V..-C.--.Y..........V....MV.K.AD.----.K.--------------.E..L.-  309
112985      271  V..-C.--.Y..........V.....V.K.A..----.K.--------------.E..L.-  309
1168261     271  V..-C.--.Y..........V.....V.K.A..----.K.--------------.E..L.-  309
112981      264  VV.-S.--.Y................I.R.A..----.K.--------------.E..L.-  302
112982      242  V..-C.--.Y..........V....MV.K.AD.----.K.--------------.E..L.-  280
74582855    276  ML.-C.--..............C.SI.AN.A..----...-------------IE..T.-  314
1168256     269  IGI-S.--.Y........Q.V.CLS...E.PKQ----.VA--------------.K..LQ-  307
2506178     256  LLV-..--.Y..........V..LSIVSKSADV----SS.--------------.E..L.-  294
1168258     287  CLV-..--.Y..........V..LSIV.KSADV----.G.--------------.E..L.-  325
21542386    243  CLI-..--.Y..........V..LSIV.KSADV----.SK-------------.E....-  281
112972      243  CLA-..--.Y..........V..LSIV.KTADV----.SK-------------.E..L.-  281
584706      245  LLM-..--.Y..........V..LSIV.GSADV----.V.--------------.E..L.-  283
112971      250  LFC-..--..S..F...N..V.NLS.VGK..DN----VQ.--------------.L..ME-  288
21542387    241  CLI-..--.Y..........I.SL.IV.TS.DV----.KK-------------.EN..L-  279
75041219    251  .FC-..--..S..F...N..V.NL..VGKEP.G----IL.--------------.L..ME-  289
156630457   251  .FC-..--..S..F...N..V.NL..VGKEP.S----ILQ-------------.L..ME-  289
75076072    251  .FC-..--..S..F...N..V.NL..VGKEP.S----IL.--------------.L..ME-  289
5902703     251  .FC-..--..S..F...N..V.NL..VGKEP.S----ILQ-------------.L..ME-  289
20532373    290  .FV-..--.YS..L...A..I..IN.V..SADA----.T.--------------.K..L.-  328
112979      291  VLV-..--.YS..L...A..I..IN.IS.SP.S----...-------------.K..L.-  329
112976      251  LFC-..--..S..F...N..V.NL..VAKEPDS----IL.--------------.L..ME-  289
122065118   251  LFC-..--..S..F...N..V.NL..VGKEHDS----VL.--------------.L..ME-  289
126302508   251  LFC-..--..S..F...N..V.NL..VAKEPDS----IL.--------------.L..MQ-  289
122065117   251  LFC-..--..S..F...N..V.NL..VAKEPDS----IL.--------------.L..ME-  289
112975      251  LFC-..--..S..F...N..V.NL..VGKESDS----VL.--------------.L..ME-  289
1703040     247  VFV-C.--.....A.M....V.C.HLALTKQAQNKTIKPA-------------.T..LA-  289
74582081    247  .FV-C.--............T.CMHYVAK.AST----KNK--------------.L..LC-  285
2492843     243  MVV-S.--.....F...N..V.NL..VVNNPAV----I.G-------------FQ..MS-  281
1168262     238  LLV-.S--..S..F...N..V....LVAENA.I----.ST-------------SLT...-  276
12230871    240  .FV-SS--..S.SFS.....V..LSIVTESRD.----S..-------------.L....-  278
112989      238  LIV-.S--.YS..F...N..V..C.LVAA.S.T----VD.-------------AF..M.-  276
17433702    238  LIV-.S--.YS..F...N..V..C.LVAA.A.T----VD.-------------AF..M.-  276
20137200    238  LIV-.S--.YS..F...N..V..C.LVAA.A.T----VD.-------------AF..M.-  276
20141944    239  ALV-SN--..S.IFS.....V.GLS.V.E.A.I----...-------------.LG.L.-  277
12230956    239  VLV-TS--.CS..F...RD.V..LI.CAQNA.K----LTD-------------LR..LA-  277
```


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 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
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
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

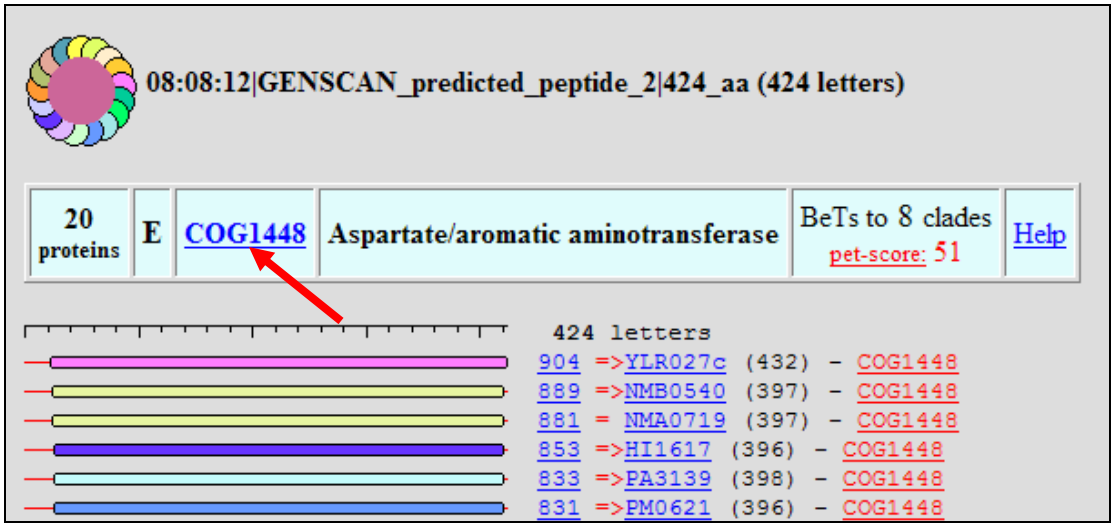
Help

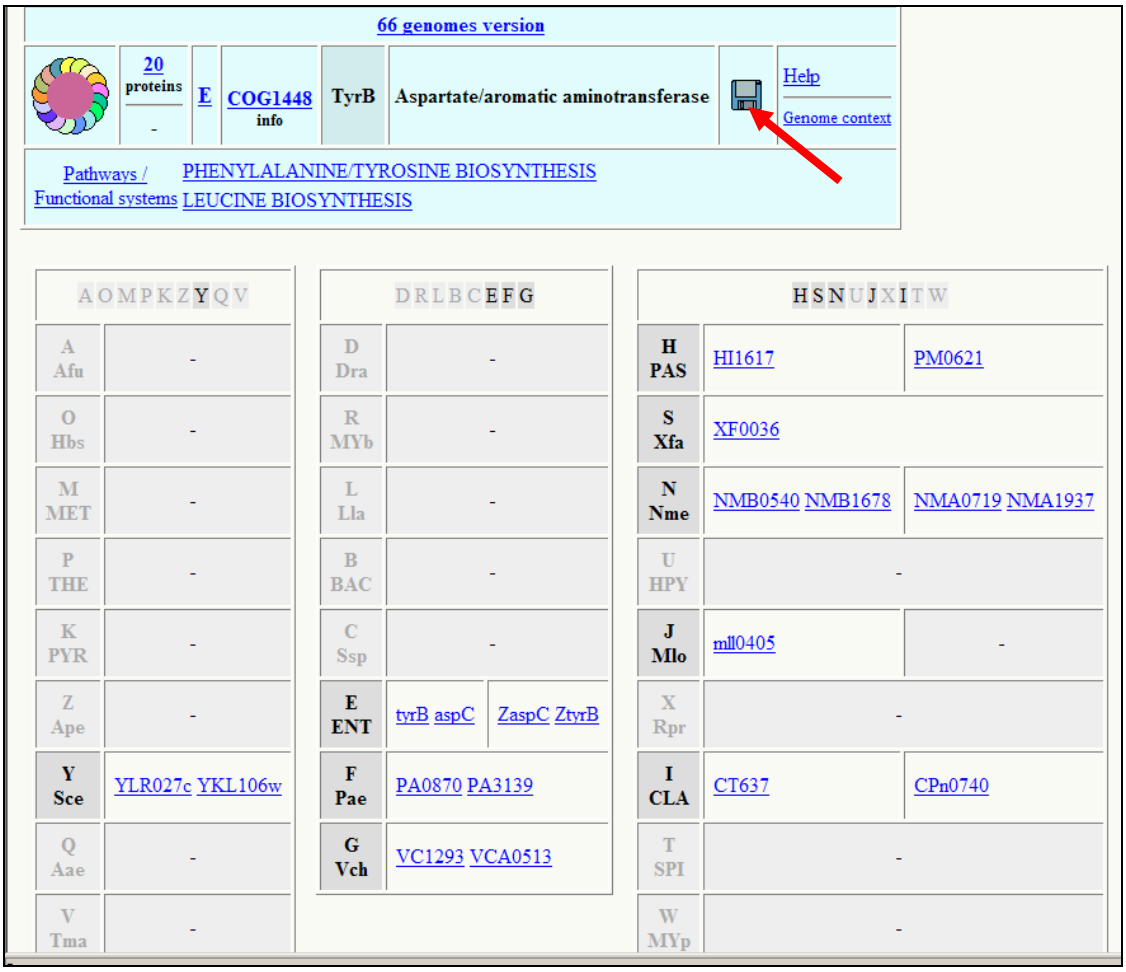
Example

Paste your sequence and press the button above.

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

Skip low-complexity filtering


08:08:12|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

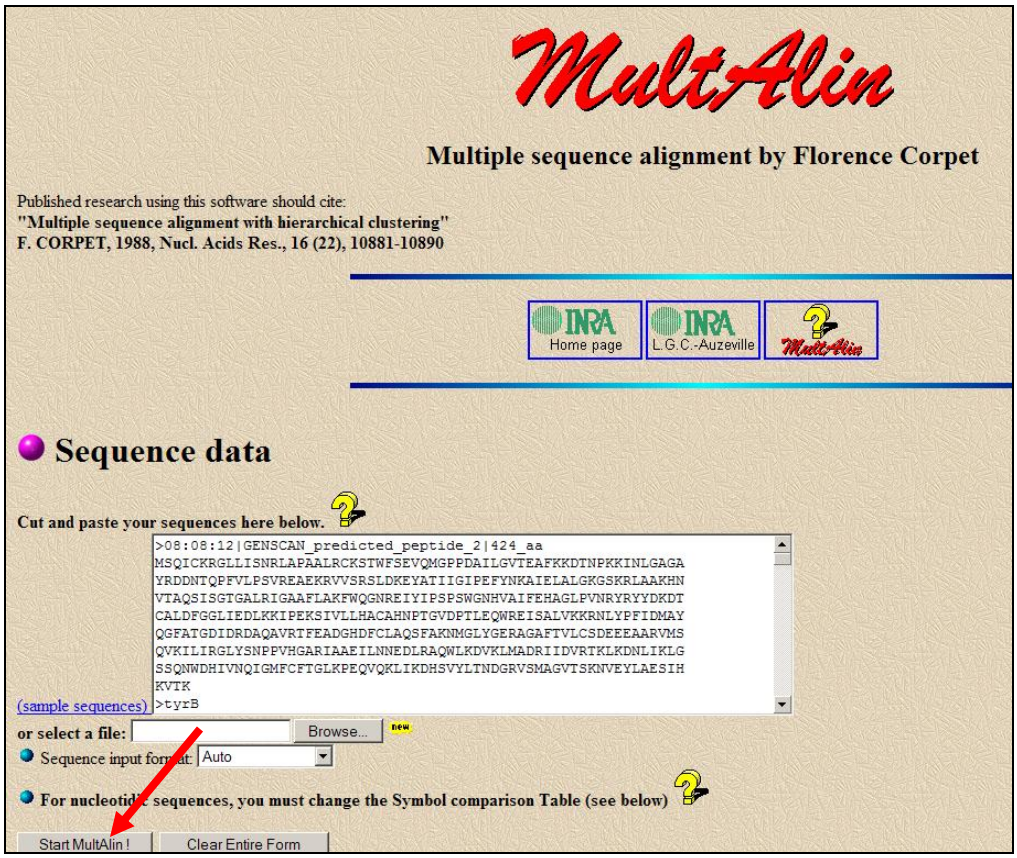

66 genomes version
20 proteins
E
COG1448 info
TyrB
Aspartate/aromatic aminotransferase
Help
Genome context
Pathways / Functional systems
PHENYLALANINE/TYROSINE BIOSYNTHESIS
LEUCINE BIOSYNTHESIS
A O M P K Z Y Q V
A Afu -
O Hbs -
M MET -
P THE -
K PYR -
Z Ape -
Y Sce YLR027c YKL106w
Q Aae -
V Tma -
D R L B C E F G
D Dra -
R MYb -
L Lla -
B BAC -
C Ssp -
E ENT tyrB aspC ZaspC ZtyrB
F Pae PA0870 PA3139
G Vch VC1293 VCA0513
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 CLA CT637 CPn0740
T SPI -
W MYp -

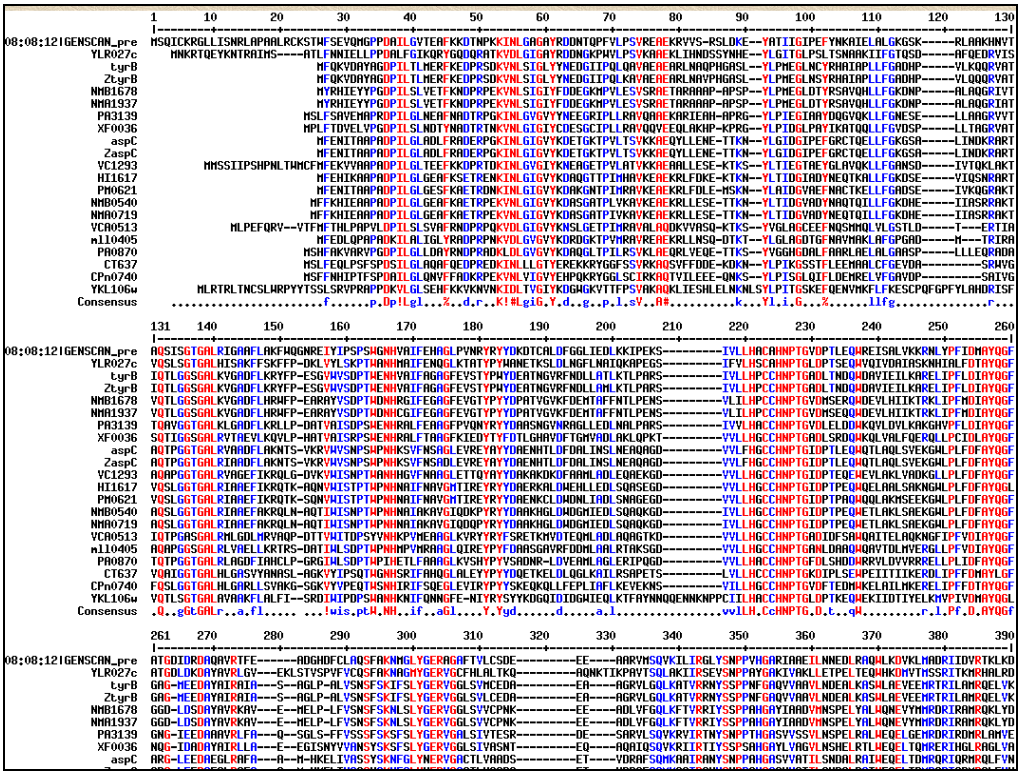
To generate a multiple sequence alignment, use:

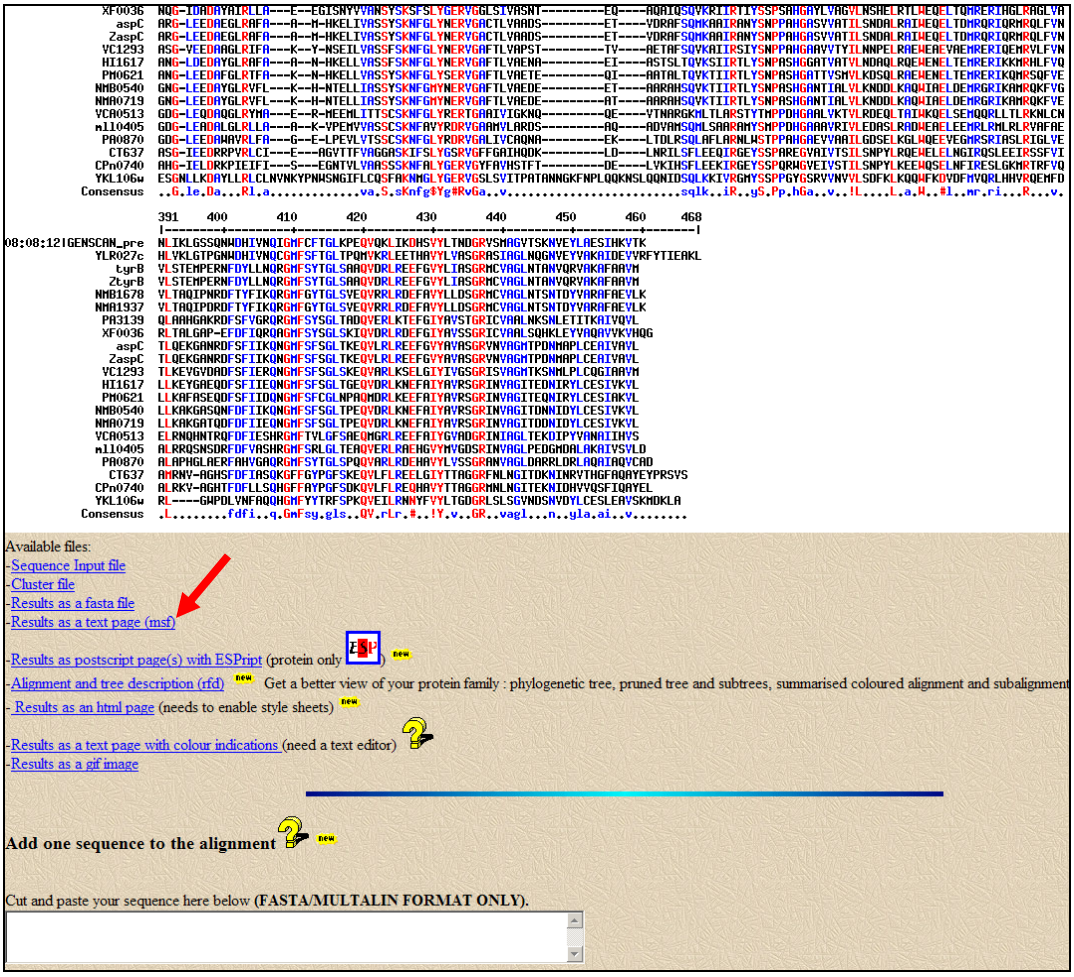
MultAlin

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.
>08:08:12|GENSCAN_predicted_peptide_2|424_aa
MSQICKRGLLISNRLAPAALRCKSTWFSEVQMGPPDAILGVTEAFKKDTNPKKINLGAGA
YRDDNTQPFVLPSVREAEKRVVSRSLDKEYATIIGIPEFYNKAIELALGKGSKRLAAKHN
VTAQSISGTGALRIGAAFLAKFWQGNREIYIPSPSWGNHVAIFEHAGLPVNRYRYYDKDT
CALDFGGLIEDLKKIPEKSIVLLHACAHNPTGVDPTLEQWREISALVKKRNLYPFIDMAY
QGFATGDIDRDAQAVRTFEADGHDFCLAQSFAKNMGLYGERAGAFTVLCSDEEEAARVMS
QVKILIRGLYSNPPVHGARIAAEILNNEDLRAQWLKDVKLMADRIIDVRTKLKDNLIKLG
SSQNWDHIVNQIGMFCFTGLKPEQVQKLIKDHSVYLTNDGRVSMAGVTSKNVEYLAESIH
KVTK
(sample sequences)
>tyrB
or select a file:
Browse...
Sequence input format: Auto
For nucleotide sequences, you must change the Symbol comparison Table (see below)
Start MultAlin !
Clear Entire Form


08:08:12|GENSCAN_pre
YLR027c
tyrB
ZtyrB
NMB1678
NMA1937
PA3139
XF0036
aspC
ZaspC
YC1293
HI1617
PM0621
NMB0540
NMA0719
YCA0513
mll0405
PA0870
CT637
CPn0740
YKL106w
Consensus


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)
-Alignment and tree description (rfd) Get a better view of your protein family : phylogenetic tree, pruned tree and subtrees, summarised coloured alignment and subalignment
- Results as an html page (needs to enable style sheets)
-Results as a text page with colour indications (need a text editor)
-Results as a gif image
Add one sequence to the alignment
Cut and paste your sequence here below (FASTA/MULTALIN FORMAT ONLY).

```
                    251                                                300
08:08:12|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
08:08:12|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
```

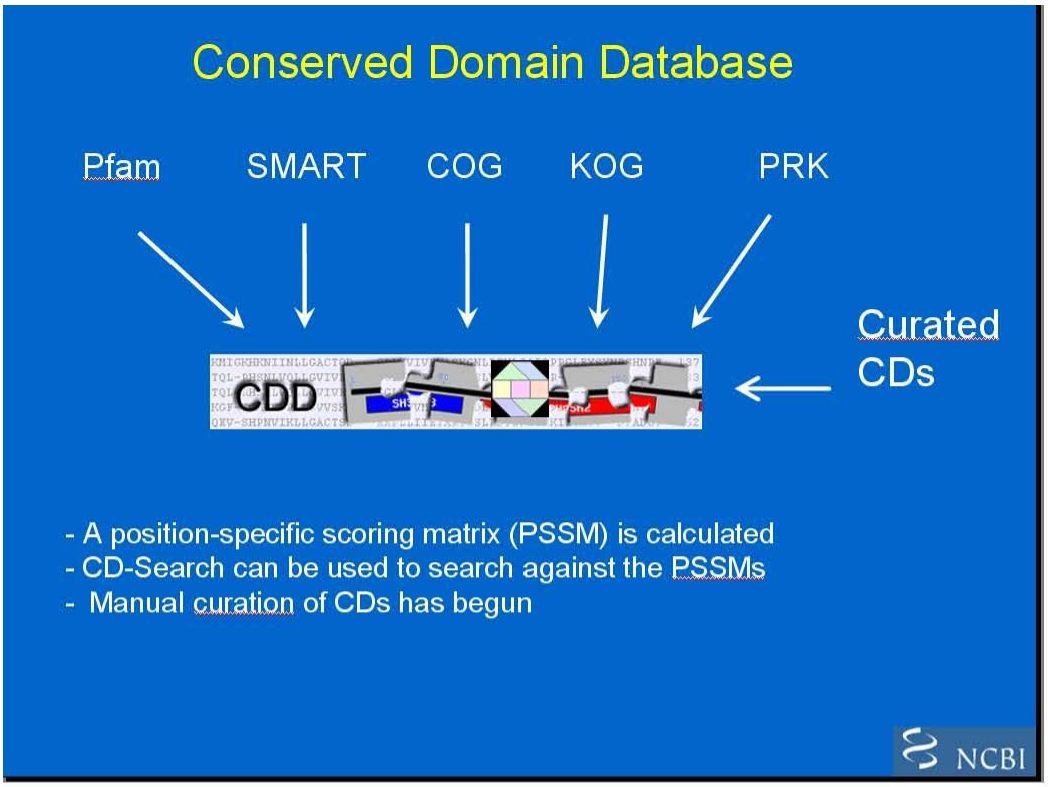
CDD

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

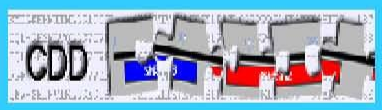
Conserved Domain

- 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

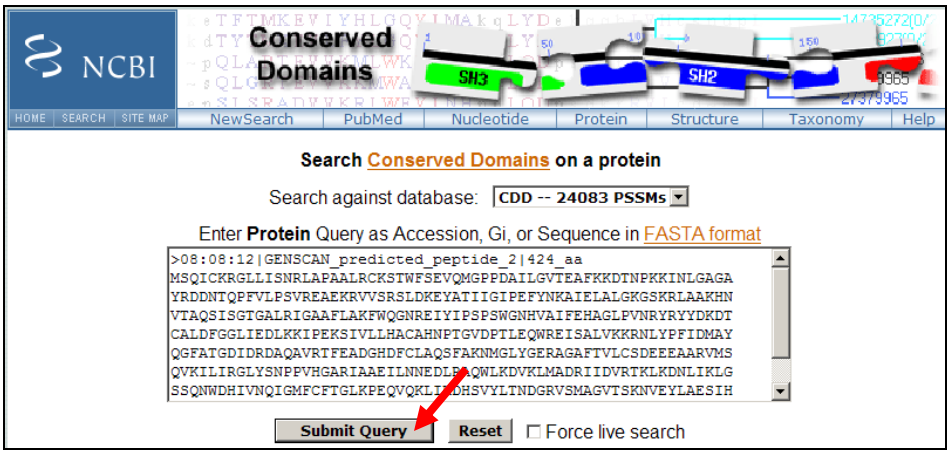
NCBI

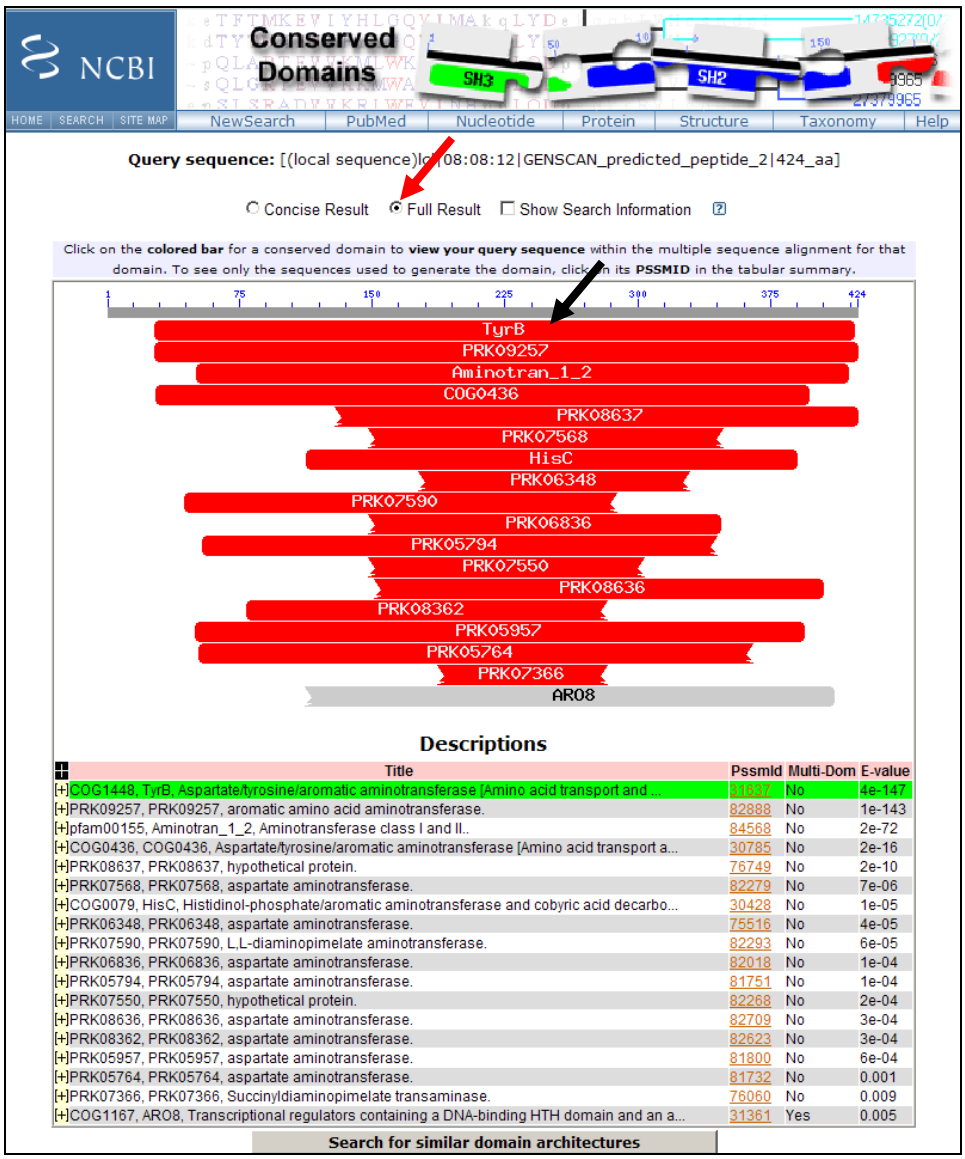


To search for protein domains and view a model structure for your protein, click here:

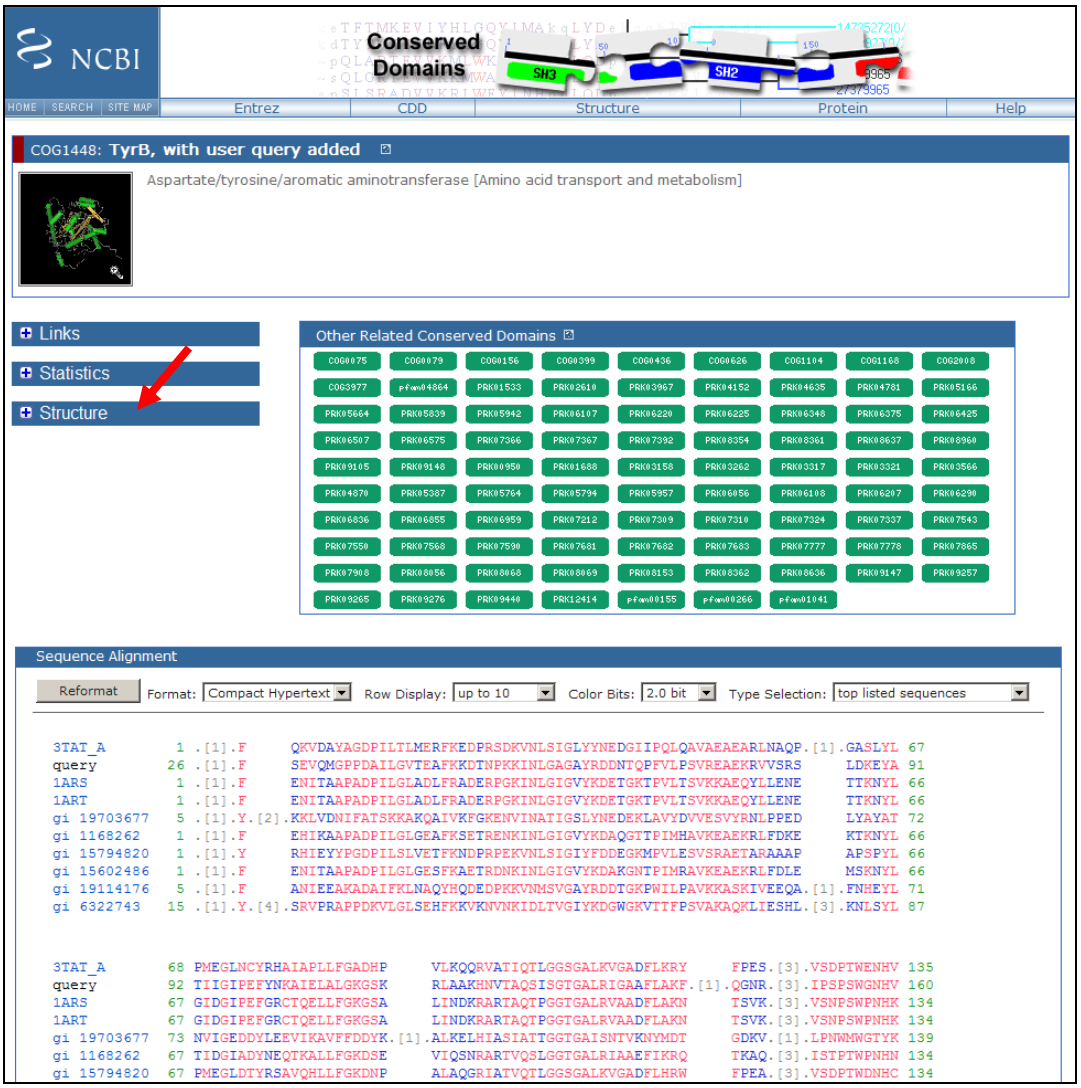


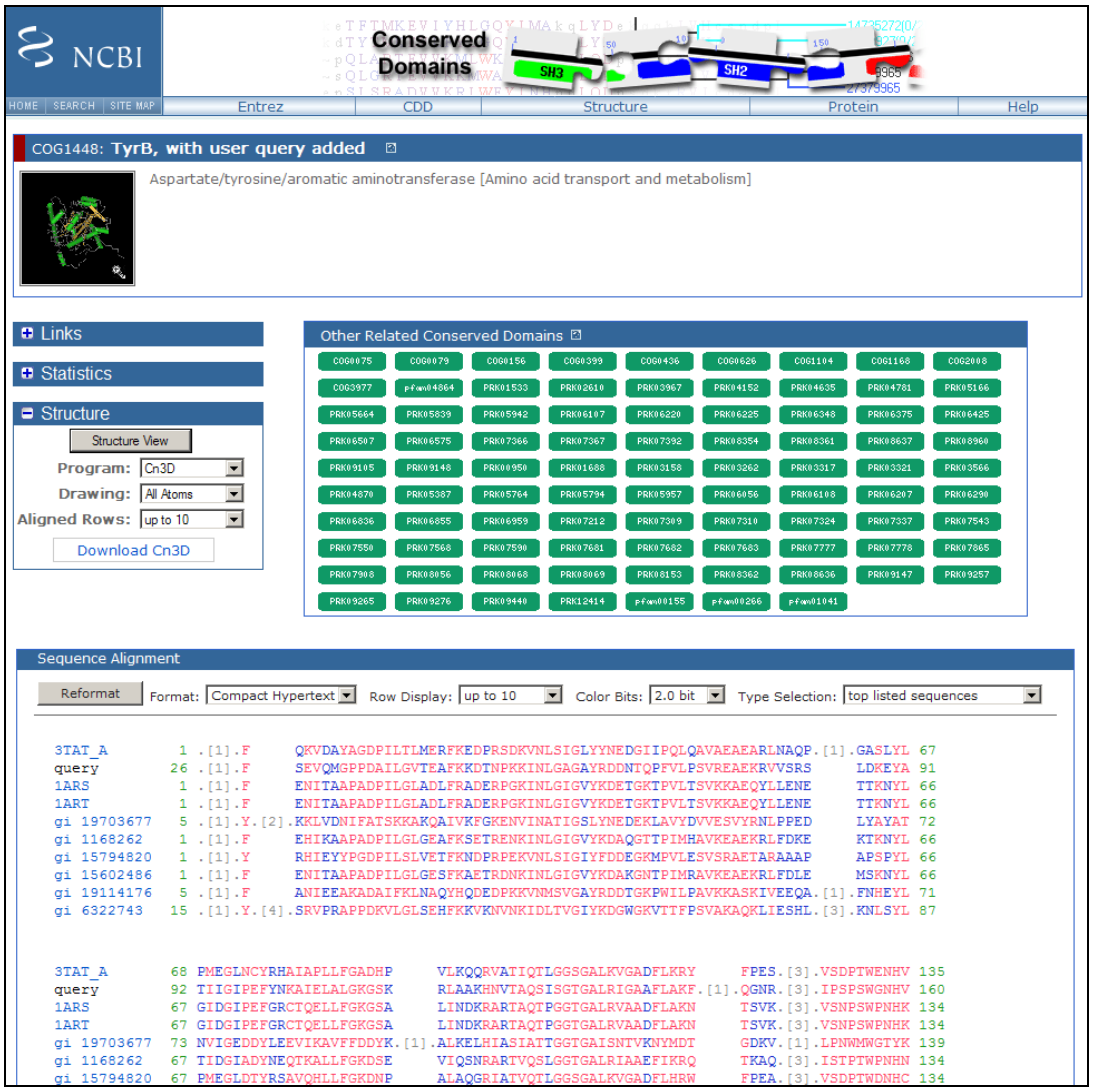
NCBI's Conserved Domain Search allows you to match your protein sequence to conserved protein domains in the Conserved Domain Database, generate a multiple sequence alignment based on this match, and explore 3D modeling templates for your sequence.
Paste your protein sequence from GenScan into the CD-Search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

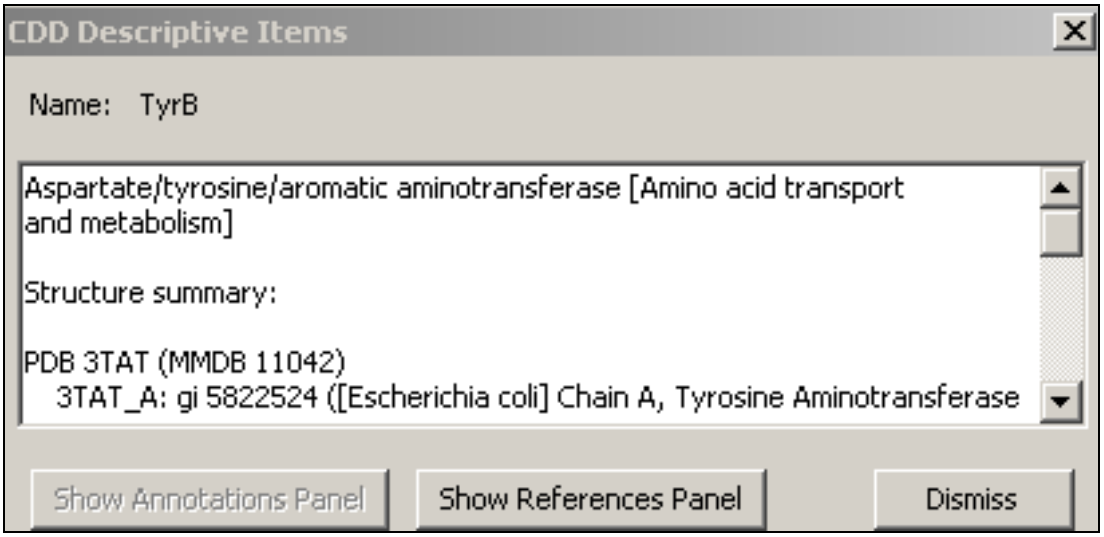


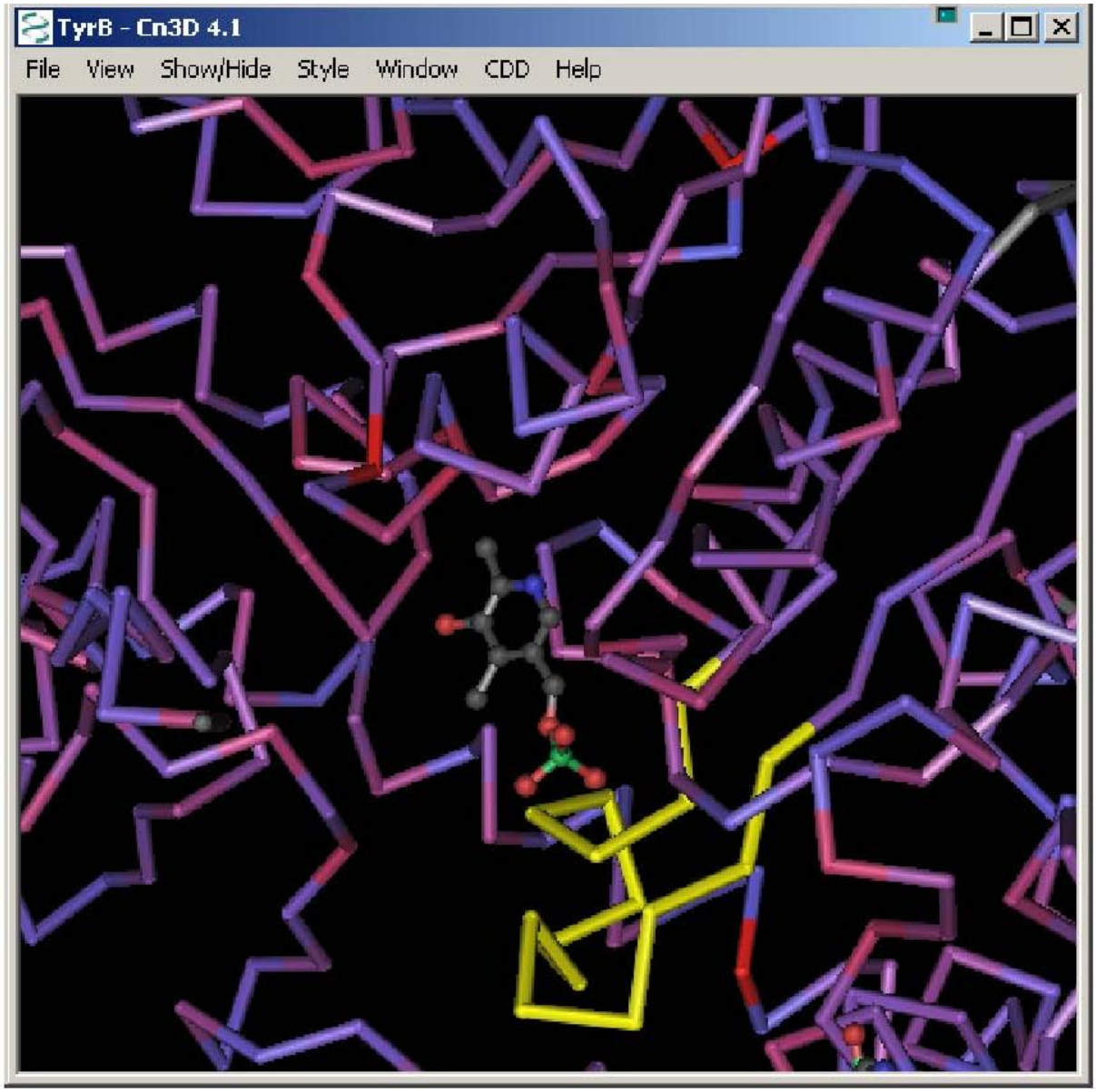


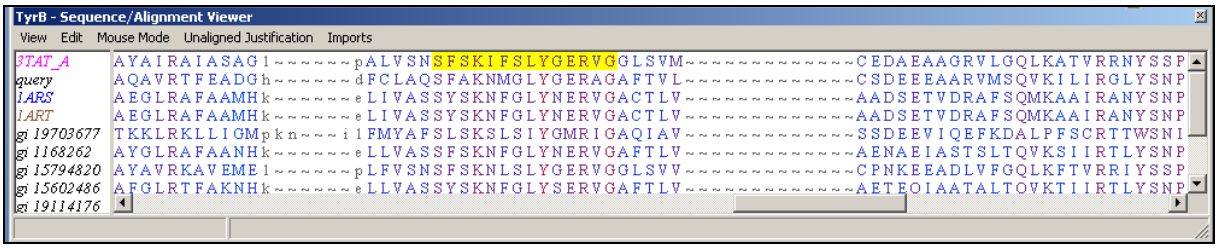
Title	PssmId	Multi-Dom	E-value
[+]COG1448, TyrB, Aspartate/tyrosine/aromatic aminotransferase [Amino acid transport and ...	81837	No	4e-147
[+]PRK09257, PRK09257, aromatic amino acid aminotransferase.	82888	No	1e-143
[+]pfam00155, Aminotran_1_2, Aminotransferase class I and II..	84568	No	2e-72
[+]COG0436, 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.	82279	No	7e-06
[+]COG0079, HisC, Histidinol-phosphate/aromatic aminotransferase and cobyric acid decarbo...	30428	No	1e-05
[+]PRK06348, PRK06348, aspartate aminotransferase.	75516	No	4e-05
[+]PRK07590, PRK07590, L,L-diaminopimelate aminotransferase.	82293	No	6e-05
[+]PRK06836, PRK06836, aspartate aminotransferase.	82018	No	1e-04
[+]PRK05794, PRK05794, aspartate aminotransferase.	81751	No	1e-04
[+]PRK07550, PRK07550, hypothetical protein.	82268	No	2e-04
[+]PRK08636, PRK08636, aspartate aminotransferase.	82709	No	3e-04
[+]PRK08362, PRK08362, aspartate aminotransferase.	82623	No	3e-04
[+]PRK05957, PRK05957, aspartate aminotransferase.	81800	No	6e-04
[+]PRK05764, PRK05764, aspartate aminotransferase.	81732	No	0.001
[+]PRK07366, PRK07366, Succinyldiaminopimelate transaminase.	76060	No	0.009
[+]COG1167, ARO8, Transcriptional regulators containing a DNA-binding HTH domain and an a...	31361	Yes	0.005

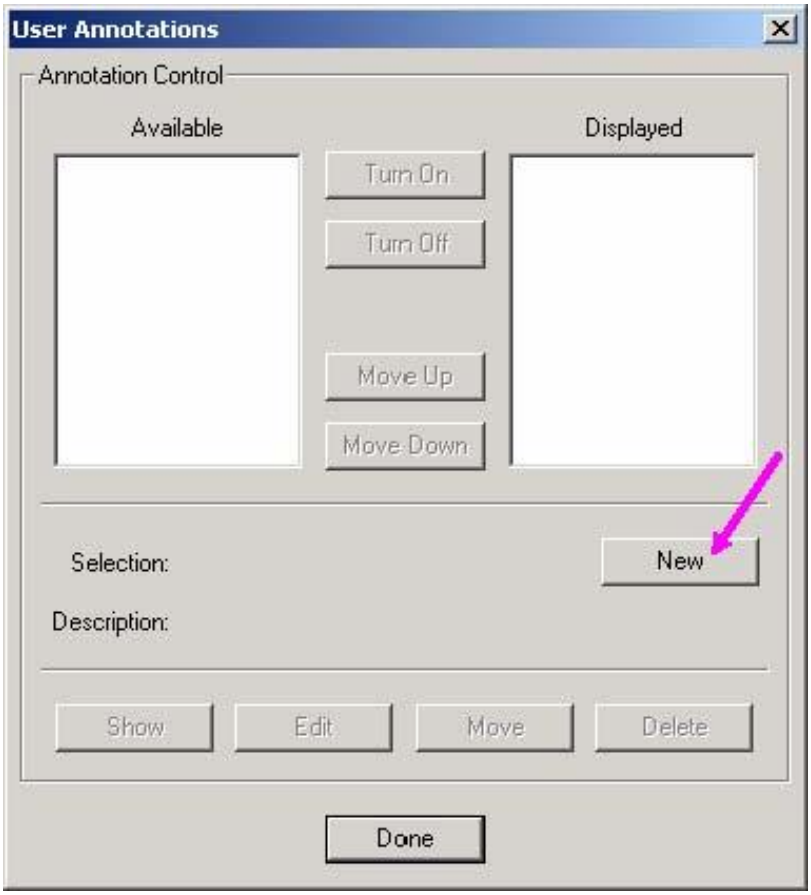

NCBI
Conserved Domains
SH3
SH2
HOME
SEARCH
SITE MAP
Entrez
CDD
Structure
Protein
Help
COG1448: TyrB, with user query added
Aspartate/tyrosine/aromatic aminotransferase [Amino acid transport and metabolism]
Links
Statistics
Structure
Other Related Conserved Domains
Sequence Alignment
Reformat
Format: Compact Hypertext
Row Display: up to 10
Color Bits: 2.0 bit
Type Selection: top listed sequences

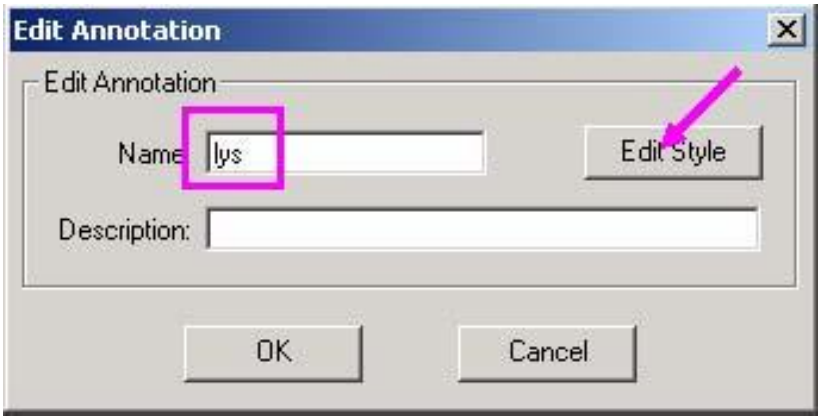

NCBI
HOME | SEARCH | SITE MAP
Conserved Domains
SH3
SH2
Entrez
CDD
Structure
Protein
Help
COG1448: TyrB, with user query added
Aspartate/tyrosine/aromatic aminotransferase [Amino acid transport and metabolism]
Links
Statistics
Structure
Structure View
Program: Cn3D
Drawing: All Atoms
Aligned Rows: up to 10
Download Cn3D
Other Related Conserved Domains
Sequence Alignment
Reformat
Format: Compact Hypertext
Row Display: up to 10
Color Bits: 2.0 bit
Type Selection: top listed sequences

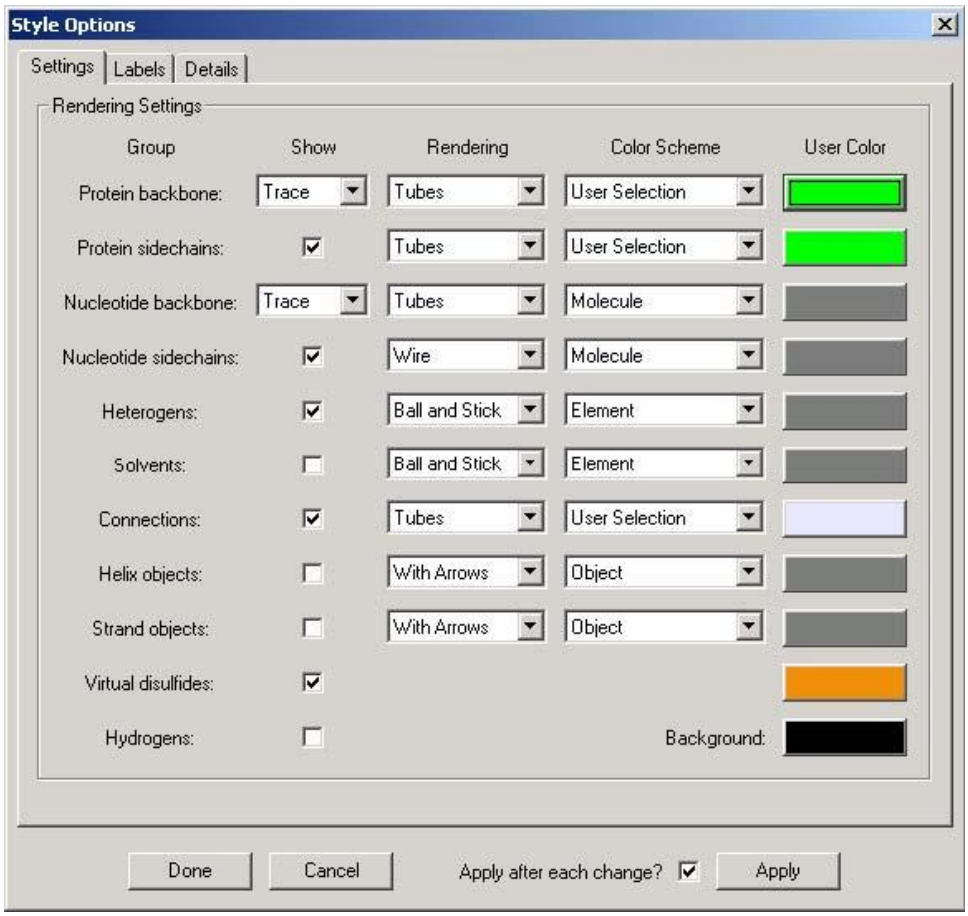
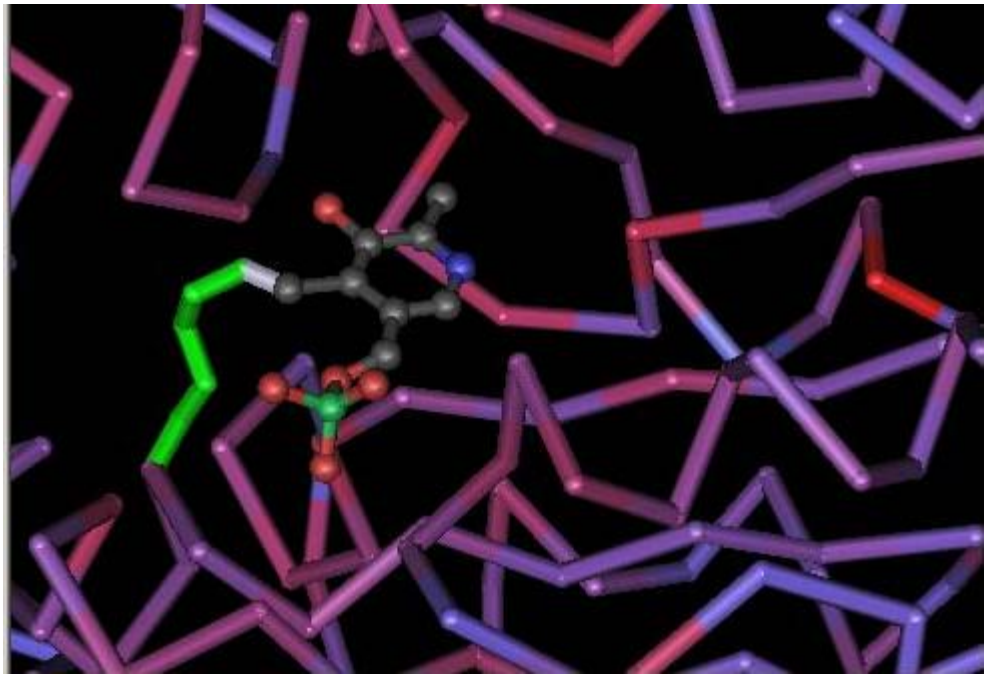

CDD Descriptive Items
Name: TyrB
Aspartate/tyrosine/aromatic aminotransferase [Amino acid transport and metabolism]
Structure summary:
PDB 3TAT (MMDB 11042)
3TAT_A: gi 5822524 ([Escherichia coli] Chain A, Tyrosine Aminotransferase
Show Annotations Panel
Show References Panel
Dismiss


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


TyrB - Sequence/Alignment Viewer
View Edit Mouse Mode Unaligned Justification Imports
STAT_A AYAIRAIASAG1~~~~~~pALVSNSFSKIFSLYGERVGGLSVM~~~~~~~~~~~~~~CEDAEAAGRVLGQLKATVRRNYSSP
query AQAVRTFEADGh~~~~~~dFCLAQSFAKNMGLYGERAGAFTVL~~~~~~~~~~~~~~CSDEEEAARVMSQVKILIRGLYSNP
1ARS AEGLRAFAAMHk~~~~~~eLIVASSYSKNFGLYNERVGACTLV~~~~~~~~~~~~~~AADSETVDRAFSQMKAAIRANYSNP
1ART AEGLRAFAAMHk~~~~~~eLIVASSYSKNFGLYNERVGACTLV~~~~~~~~~~~~~~AADSETVDRAFSQMKAAIRANYSNP
gi 19703677 TKKLRKLLIGMpkn~~~i1FMYAFSLSKSLSIYGMRIGAQIAV~~~~~~~~~~~~~~SSDEEVIQEFKDALPFSCRTTWSNI
gi 1168262 AYGLRAFAANHk~~~~~~eLLVASSFSKNFGLYNERVGAFTLV~~~~~~~~~~~~~~AENAEIASTSLTQVKSIIRTLYSNP
gi 15794820 AYAVRKAVEME1~~~~~~pLFVSNSFSKNLSLYGERVGGLSVV~~~~~~~~~~~~~~CPNKEEADLVFGQLKFTVRRIYSSP
gi 15602486 AFGLRTFAKNHk~~~~~~eLLVASSYSKNFGLYSERVGAFTLV~~~~~~~~~~~~~~AETEQIAATALTQVKTIIRTLYSNP
gi 19114176


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