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

Dr. David Wheeler, NCBI
Dr. Medha Bhagwat, NCBI

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

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

Electronic Notebook for Protein Sequence Analysis

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

**URLs Used:**
2. GenScan: http://genes.mit.edu/GENSCAN.html

**NoteBooks:**
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 optional: Predicted peptides only

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

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

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

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

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

Parameter matrix: HumanIso.smat

Predicted genes/exons:

| Gn.Ex | Type | S | Begin | ...End | Len | Fr | Ph | I/Ac | Do/T | CodRg | P.... | Tscr.. |
|-------|------|---|-------|--------|-----|----|----|------|------|-------|-------|--------|--------|
| 1.01  | Sngl | + | 27    | 458    | 432 | 0  | 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  | Intr | + | 3253  | 3948   | 696 | 0 | 0  | 90   | 49   | 1324  | 0.999 | 122.25 |
| 2.06  | PlyA | + | 4120  | 4125   | 6   |   |    |      |      |       |       | 1.05   |
| 3.04  | PlyA | - | 4162  | 4157   | 6   |   |    |      |      |       |       | -0.45  |
| 3.03  | Term | - | 4448  | 4261   | 188 | 0 | 2  | 37   | 42   | 95    | 0.922 | -2.55  |
| 3.02  | Intr | - | 4635  | 4511   | 125 | 2 | 2  | 44   | 90   | 91    | 0.949 | 5.13   |
| 3.01  | Init | - | 5046  | 4694   | 353 | 0 | 2  | 66   | 43   | 485   | 0.897 | 38.43  |

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

Predicted peptide sequence(s):

>09:48:24|GENSCAN_predicted_peptide_1|143_aa
MPRTLPWTFTAVASSARAKSMEKLTVVFLRMHSALVVSPSMATRVLFPFDQSLN
SRAPAKTTSAAAQTAYLSIFFHIELQGRIGWLFRWLSPLSASSQRYESSTKSGESPKT
TQSFRMNQKQLRAATQKKAFFDD

>09:48:24|GENSCAN_predicted_peptide_2|424_aa
MSQICKRGLLISNRCAAPALRECKTWFSEVQMGPPDAIIWLGEAFAKDTNPPKINLGAGAYRDDQPPFVLSPVRESAEKRVRSRSLQKDKEYATGIGIPEYNKAIELALGKGSKRLAACKHVTAQSIISGTGALRIGAANLAFKFWQGNREYIFPSWGNHVAIFEHAGLPVNRYYDKDTCALDPGGLIENDLKKIPEKSIIVLHHACAHNPTGVDPTLEQWREISALARVKRNLYFIDMAYQGFATGDIDRDAQVRAFTEDGHDFCLAQSFKANMGLYGERAGAFTVLCDEEAARVMSQVKLIRGLYNSPuVHARIAAEILNEDLRACLWLDVKLMAADRIIDVRTKLKDNLIKLGSSQNDHIVNQQGFCFTGLKEPQVQKLIKHDHSVYLTNDRVSMAGVTSKNEVLAESIH
KVTK

>09:48:24|GENSCAN_predicted_peptide_3|221_aa
MSNLQQLNSLVTLSMLTELQGCHNLIRAGASGVIQAMVLSFGSRFSNQHLECNHPKFLHRDHFRRLNYGKNKTHVNTTIIIVDDEDKAVINIALDRSRRSYACDGGCDEPIVLLTQNRRQFPVLTEPLTAILYITEDQHMEELSSHAIHVKEVVEAPAEQHLIALHRHGQLGLFTLFWVSVCAIIIVFHIIFLCKLIIKEYCEPSDQLRYYNKF
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**
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.
This view shows ScanProsite results together with ProRule-based predicted intra-domain features (help).

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

**found: 1 hit in 1 sequence**

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

**Legend:**

- disulfide bridge
- active site
- other ranges
- other sites

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

Hits by P300105 AA_TRANSFER_CLASS_1 Aminotransferases class-I pyridoxal-phosphate attachment site:

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

270 - 283: YFAHnNG/GERAG

**Legend:**

- disulfide bridge
- active site
- other ranges
- other sites

horizontal scaling: [ ]

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 putida aspC, which is involved in cobalamin biosynthesis.
- Yeast hypothetical protein YUL005w.

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 2008 / Pattern and text revised.

**Technical section:**

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

<table>
<thead>
<tr>
<th>AA_TRANSFER_CLASS_1, PS01005, Aminotransferases class-I pyridoxal-phosphate attachment site (PATTERN)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Consensus pattern:</strong></td>
</tr>
<tr>
<td>[GS] · [LIVMFY]T [AC] · [GSTA] · K · x(2) · [GSALVN] · [LIVMFA] · x · [GNAR] · (V) · R · [LIVMA]</td>
</tr>
<tr>
<td>[GA]</td>
</tr>
<tr>
<td>K is the pyridoxal-P attachment site</td>
</tr>
<tr>
<td>Sequences known to belong to this class detected by the pattern:</td>
</tr>
<tr>
<td>Other sequence(s) detected in Swiss-Prot:</td>
</tr>
</tbody>
</table>
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)
### COGs

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

<table>
<thead>
<tr>
<th>Code</th>
<th>Name</th>
<th>Genes in COG</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Archaeoglobus fiordia</td>
<td>2420</td>
</tr>
<tr>
<td>O</td>
<td>Halobacterium sp. NRC-1</td>
<td>2605</td>
</tr>
<tr>
<td>M</td>
<td>Methanococcus jannaschii</td>
<td>1766</td>
</tr>
<tr>
<td>M</td>
<td>Methanobacterium thermoautotrophicum</td>
<td>1873</td>
</tr>
<tr>
<td>E</td>
<td>Thermoplasma acidophilum</td>
<td>1498</td>
</tr>
<tr>
<td>E</td>
<td>Thermoplasma vulcanii</td>
<td>1499</td>
</tr>
<tr>
<td>K</td>
<td>Pyrococcus horikoshii</td>
<td>1800</td>
</tr>
<tr>
<td>Z</td>
<td>Pyrococcus abyssi</td>
<td>1768</td>
</tr>
<tr>
<td>Y</td>
<td>Anoxybacillus pasteurii</td>
<td>1841</td>
</tr>
<tr>
<td>Y</td>
<td>Bacillus calmettei</td>
<td>5955</td>
</tr>
<tr>
<td>Q</td>
<td>Aquifex aeolicus</td>
<td>1560</td>
</tr>
<tr>
<td>T</td>
<td>Thermotoga maritima</td>
<td>1858</td>
</tr>
<tr>
<td>D</td>
<td>Desulfovibrio desulfuricans</td>
<td>3187</td>
</tr>
<tr>
<td>E</td>
<td>Mycobacterium tuberculosis</td>
<td>3927</td>
</tr>
<tr>
<td>E</td>
<td>Mycobacterium leprae</td>
<td>1605</td>
</tr>
<tr>
<td>L</td>
<td>Lactococcus lactis</td>
<td>2267</td>
</tr>
<tr>
<td>E</td>
<td>Streptococcus pyogenes</td>
<td>1697</td>
</tr>
<tr>
<td>E</td>
<td>Bacillus subtilis</td>
<td>4118</td>
</tr>
<tr>
<td>E</td>
<td>Bacillus licheniformis</td>
<td>4056</td>
</tr>
<tr>
<td>C</td>
<td>Propionibacterium</td>
<td>3167</td>
</tr>
<tr>
<td>E</td>
<td>Escherichia coli K12</td>
<td>4275</td>
</tr>
<tr>
<td>E</td>
<td>Escherichia coli O157</td>
<td>5315</td>
</tr>
<tr>
<td>C</td>
<td>Bacillus sp. ASP</td>
<td>575</td>
</tr>
<tr>
<td>F</td>
<td>Pseudomonas aeruginosa</td>
<td>5567</td>
</tr>
<tr>
<td>G</td>
<td>Vibrio cholerae</td>
<td>3835</td>
</tr>
<tr>
<td>H</td>
<td>Haemophilus influenzae</td>
<td>1714</td>
</tr>
<tr>
<td>K</td>
<td>Pasturella multocida</td>
<td>2015</td>
</tr>
</tbody>
</table>
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 COG editor to compare the protein sequence to the COGs database.

Paste the FASTA formatted protein sequence from GenScan into the COG editor 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
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 corresponding to an ancient conserved domain.

<table>
<thead>
<tr>
<th>Code</th>
<th>Name</th>
<th>Proteins in COGs</th>
<th>in COGs</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Archaeoglobus sulfidifrons</td>
<td>2420</td>
<td>1872</td>
</tr>
<tr>
<td>O</td>
<td>Halobacterium sp. NRC-1</td>
<td>2605</td>
<td>1701</td>
</tr>
<tr>
<td>M</td>
<td>Methanothermobacter thermautotrophicum</td>
<td>1873</td>
<td>1388</td>
</tr>
<tr>
<td>P</td>
<td>Thermoplasma acidocaldarium</td>
<td>1482</td>
<td>1230</td>
</tr>
<tr>
<td>K</td>
<td>Pyrococcus horikoshii</td>
<td>1800</td>
<td>1378</td>
</tr>
<tr>
<td>Z</td>
<td>Azotobacter vinelandii</td>
<td>1841</td>
<td>1178</td>
</tr>
<tr>
<td>Y</td>
<td>Saccharomyces cerevisiae</td>
<td>5955</td>
<td>2280</td>
</tr>
<tr>
<td>Q</td>
<td>Aspergillus niger</td>
<td>1560</td>
<td>1329</td>
</tr>
<tr>
<td>V</td>
<td>Thermotoga maritima</td>
<td>1858</td>
<td>1527</td>
</tr>
<tr>
<td>D</td>
<td>Desulfovibrio desulfuricans</td>
<td>3187</td>
<td>2226</td>
</tr>
<tr>
<td>R</td>
<td>Mycobacterium tuberculosis</td>
<td>3927</td>
<td>2585</td>
</tr>
<tr>
<td>L</td>
<td>Lactococcus lactis</td>
<td>2267</td>
<td>1618</td>
</tr>
<tr>
<td>B</td>
<td>Bacillus subtilis</td>
<td>4118</td>
<td>2870</td>
</tr>
<tr>
<td>C</td>
<td>Bacillus halodurans</td>
<td>4066</td>
<td>2878</td>
</tr>
</tbody>
</table>

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

**COGntor**

Compare your sequence to COG database

**NCBI**
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.
Multiple sequence alignment by Florence Corpet

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

Sequence data

Cut and paste your sequence here below.

```
# Sample sequence
PPQVDAVATCPIEILTWKDKFPRDTTFKVL5SRF YIREEQ S1 PQLGQVYERKARSLGNYAMCPYAGKAYCTLYFY
```

or select a file

Sequence input format: Auto

```
14:10:3216/HUMAN_pre
YLR047c
angC
Hel11A7
tyrG
Cbf70A
YBL166w
Consensus

14:10:3216/HUMAN_pre
YLR047c
angC
Hel11A7
tyrG
Cbf70A
YBL166w
Consensus

14:10:3216/HUMAN_pre
YLR047c
angC
Hel11A7
tyrG
Cbf70A
YBL166w
Consensus

14:10:3216/HUMAN_pre
YLR047c
angC
Hel11A7
tyrG
Cbf70A
YBL166w
Consensus
```

...
Consensus levels: high=90% low=50%

Consensus symbols:

! is anyone of IV
$ is anyone of LM
% is anyone of FY
# is anyone of NDQEBZ/

Consensus:

251

14:13:11 GENSCAN_pre PFIDMAYQGF ATGDIDRDAQ AVRTFEE.... ADGDHDFCL AQSFAKNMGL
YLR027c ALFDTAYQGF ATGDLDKDAY AVRLGV....E KLSTVSPVFV CQSFAKNAGM
tyB PFLDIAQGF GAG.MEEDAY AIRAIA....S ...AGLP.ALV SNSFSKIFSL
ZtyrB PFLDIAQGF GAG.MEEDAY AIRAIA....S ...AGLP.ALV SNSFSKIFSL
NMB1678 PFMDIAQGF GGD.LDSDAY AVRKAV....E ...MELP.LFV SNSFSKIFSL
NMA1937 PFMDIAQGF GGD.LDSDAY AVRKAV....E ...MELP.LFV SNSFSKIFSL
PA3139 PFLDIAQGF GNG.IEEDAY AVRLFA....Q ...SGLF.FV SNSFSKIFSL
XF0036 PCIDLAYQGF NQG.IDADAY AIRLLA....E ...EGISNVVV ANSYSKSFSL
aspC PFLDFAYQGF ARG.LEEDEE GLRAFA....A ...M.HKELIV ASSYSKNFGL
ZaspC PFLDFAYQGF ARG.LEEDEE GLRAFA....A ...M.HKELIV ASSYSKNFGL
VC1293 PFLDFAYQGF ASG.VEEDAA GLRAFA....K ...Y.NSEILV ASSYSKNFGL
HI1617 PFLDFAYQGL ANG.IEEDAY GLRAFA....A ...N.HKELIV ASSYSKNFGL
PM0621 PLDFAYQGF ARG.LEEDEE GLRTFA....K ...N.HKELIV ASSYSKNFGL
NMB0540 PFLDFAYQGF GAG.MEEDAY GLRVFLE....K ...H.NTELLI ASSYSKNFGL
NMA0719 PFLDFAYQGF GAG.MEEDAY GLRVFLE....K ...H.NTELLI ASSYSKNFGL
VCA0513 PFLDFAYQGF GDG.LEQDAQ GLRYMA....E ...R.MEELL TSSCSKNFGL
m10405 PFVIAQGF GDG.LEEADAL GLRLLA....A ...K.VPEMVV ASSCSKNFAV
PA0870 PLDFAYQGF GDG.LEEADAL AVRLEPA....G ...E.LPEVLL TSSCSKNFGL
CT637 PFFDMAYQGF ASG.VEEDAA GLRAFA....K ...Y.NSEILV ASSYSKNFGL
CPn0740 PFFDTAYQGF AHG.IEEDAR PVRLCFL PIEIFIP....S ...EVTTEVV ASSCSKNFGL
YKL106w PIVDMAYQQL ESGNLLKDAY LRLCLNVTK YPNWSNGIFL CQSFAKNMGL
Consensus Pf.D.AYQGf ..G.le.Da. ..Rl. a..... ........ v a.S.sKnfg$ 301 350

14:13:11|GENSCAN_pre YGERAGFRTV LCSDE. .... EE. .... AARVMSQVKI LIRGLYNSPP
YLR027c YGERVCGFH'L ALTQK. .... AQNKTI KPAVTSQLAK IIRSEVSNPP
tyrB YGERVGGLSV MCEDA. .... EA. .... AGRVGLQGLA TVRNNYSIPP
ZtyrB YGERVGGGLSV LCEDA. .... EA. .... AGRVGLQGLA TVRNNYSIPP
NMB1678 YGERVGGGLSV VCINK. .... EE. .... ADLVFGQLKF TVRRYSSPP
NMA1937 YGERVGGGLSV VCPNK. .... EE. .... ADLVFGQLKF TVRRYSSPP
PA3139 YGERVGLS1 VTESR. .... DE. .... SARVLSQVKR VIRTNYSIPP
XF0036 YGERVGLS1 VSAN. .... EQ. .... AQAISQVKR IIRTIVYSSPS
aspC YNERVGAFTL VAADS. .... ET. .... VDRAFSQMKAAIARNYSNPP
ZaspC YNERVACTL VAADS. .... ET. .... VDRAFSQMKAAIARNYSNPP
VC1293 YNERVGAFTL VAPST. .... TV. .... AETAFSQVKAIITIYSNPP
HI1617 YNERVAFITL VAEN. .... EI. .... ASTSLTQVKS IIRTLYSNPA
PM0621 YSERVGAFTL VAETE. .... QI. .... AAALTQVKT IIRTLYSNPA
NMB0540 YNERVGAFTL VAEE. .... ET. .... AARAHSQVKT IIRTLYSNPA
NMA0719 YNERVGAFTL VAEE. .... AT. .... AARAHSQVKT IIRTLYSNPA
VCA0513 YRERTGAIV IGTKN. .... QE. .... VTNARGKMLT LARSTYTMPP
ml10405 YRDRVGAAMV LARDS. .... AQ. .... ADVAMSQMLS AARAMYNSPP
PA0870 YRDRVVALIV CAQNA. .... EK. .... LTLDRSQLAF LARNLWSTPP
CT637 YGRRYGGFAV ITQDK. .... LD. .... LNRILSFLE QIRGEYSSPA
CP0740 YGERVGYFAV HSTFT. .... DE. .... LVKIHSEQL KIRGEYSSQP
YKL106w YGERVGSLSV ITPAANNKG FNPLQQKNSL QQNISQMLK IVRGMYSSPP
Consensus YgRVGao..v ........ ........ sqlk. .IR..ys.Pp


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
Conserved Domain Database

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

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 representative 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.
Aspartate/tyrosine/aromatic aminotransferase [Amino acid transport and metabolism]

Structure summary:
PDB 3TAT (MMDB 11042)
3TAT_A: gi 58225244 ([Escherichia coli] Chain A, Tyrosine Aminotransferase)