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
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 (ScanProsise)
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 Mb) 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 GENSAN email server. If your browser (e.g., Lynx) does not support file upload or multipart forms, use the older version.

Organism: Vertebrate
Sub-optimal 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):

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 GENSAN  Clear Input
## Predicted genes/exons:

<table>
<thead>
<tr>
<th>Gen.Ex Type</th>
<th>S. Begin ... End Len Fr Ph Yf/cf Do/Cf CodRg P. ... Tscr...</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.01 Snak</td>
<td>1.01 Split + 27 458 432 2 0 48 49 303 0.447 24.60</td>
</tr>
<tr>
<td>1.02 Plya</td>
<td>1.02 Split + 409 494 6 1.95</td>
</tr>
<tr>
<td>2.00 Prom</td>
<td>2.00 Split + 338 469 40 0.286</td>
</tr>
<tr>
<td>2.01 Init</td>
<td>2.01 Split + 1069 160 60 2 2 53 84 03 0.670 3.80</td>
</tr>
<tr>
<td>2.02 Inttr</td>
<td>2.02 Split + 274 278 160 2 1 72 105 204 0.980 28.49</td>
</tr>
<tr>
<td>2.03 Inttr</td>
<td>2.03 Split + 377 277 100 1 0 10 86 251 0.599 17.47</td>
</tr>
<tr>
<td>2.04 Inttr</td>
<td>2.04 Split + 29 95 210 200 0 0 72 100 586 0.999 55.93</td>
</tr>
<tr>
<td>2.05 Term</td>
<td>2.05 Split + 3253 3948 670 0 0 90 49 1324 0.999 122.25</td>
</tr>
<tr>
<td>2.06 Pulv</td>
<td>2.06 Split + 4113 4125 6 1.95</td>
</tr>
<tr>
<td>3.04 Plya</td>
<td>3.04 Split + 4182 4137 6 0.45</td>
</tr>
<tr>
<td>3.05 Term</td>
<td>3.05 Split + 4943 4261 133 0 2 37 42 95 0.922 2.55</td>
</tr>
<tr>
<td>3.02 Inttr</td>
<td>3.02 Split + 4135 4511 125 2 2 44 90 91 0.949 5.13</td>
</tr>
<tr>
<td>3.01 Init</td>
<td>3.01 Split + 9046 4691 353 0 2 66 43 485 0.997 38.43</td>
</tr>
</tbody>
</table>

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):

>14:16:27 | GENSCAN_predicted_peptide_11423.png

![](https://example.com/peptide1.png)

>14:16:27 | GENSCAN_predicted_peptide_11424.png

![](https://example.com/peptide2.png)

>14:16:27 | GENSCAN_predicted_peptide_11421.png

![](https://example.com/peptide3.png)
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 define) 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 PDGxxxx number.

**Hit from ScanProsite**

**Prosite pattern**

**Prosite Summary**
The ScanProsite tool [Help] 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 paste your own protein sequence(s) in the box below (leave this box blank to scan PROSITE entry(s) against selected protein databases).

**PROSITE pattern/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

**General options:**
- Exclude motifs with a high probability of occurrence
- Show low level scores
- 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)

**Pattern options:**
- Allow at most [ ] X sequence characters to match a conserved position in the pattern
- Match mode [greedy, overlaps, no includes] (see patterns, see help)

START THE SCAN  reset
ScanProsite Results Viewer

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

**Hits for all PROSITE (release 19.22) motifs on sequence USERSEQ1:**

**found: 1 hit in 1 sequence**

USERSEQ1 (424 aa)

<p>| | | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

Hits by PS00165  AA_TRANSFER_CLASS_1  Aminotransferases class-I pyridoxal phosphate attachment site:

USERSEQ1 (424 aa)

270 - 283:  [PAMxxSl2yGxR]G

**Legend:**

- disulfide bridge
- active site
- other 'ranges'
- other sites
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 isoforms: 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-amino cyclopropane-1-carboxylate synthase (EC 4.4.1.14) (ACC synthase) from plants. ACC synthase catalyzes the first step in ethylene biosynthesis.
- Pseudomonas denitrificans FabC, which is involved in cobalamin biosynthesis.
- Yeast hypothetical protein YUL060w.

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:
November 1995 / Pattern and text revised.

Technical section:

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

AA_TRANSFER_CLASS_1, PS01085: Aminotransferases class-I pyridoxal-phosphate attachment site (PATTERN)

Consensus pattern:


K is the pyridoxal phosphate 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)
Your request has been successfully submitted and put into the Blast Queue.

**Query** = 14:10:32\textsc{\textregistered}GENSCAN\_predicted\_peptide\_2424\_aa (424 letters)

The request ID is: 100566746-7811-27625

Format:

- **Alignment view**:
  - flat query--anchored with identities

Please press "FORMAT" when you wish to check your results. You may change the formatting options for your result via the form below and press "FORMAT" again. You may also request results of a different search by entering any other valid request ID to see other recent jobs.
**Distribution of 61 Blast Hits on the Query Sequence**

Mouse over to see the details; click to show alignments.

<table>
<thead>
<tr>
<th>Query</th>
<th>Color key for alignment scores</th>
<th>&lt;40 &lt;br&gt;40-60 &lt;br&gt;60-80 &lt;br&gt;80-200 &lt;br&gt;&gt;=200</th>
</tr>
</thead>
</table>

Sequences producing significant alignments:

| gi|112983|sp|P00505|AATX HUMAN | Aspartate aminotransferase, mi... | 563 | 3e-160 |
| gi|112983|sp|P00520|AATX HUMAN | Aspartate aminotransferase, mi... | 562 | 7e-160 |
| gi|113987|sp|P00507|AATX RAT | Aspartate aminotransferase, mi... | 561 | 1e-159 |
| gi|112982|sp|P00507|AATX HORE | Aspartate aminotransferase, mi... | 554 | 3e-157 |
| gi|113985|sp|P00506|AATX PIG | Aspartate aminotransferase, mi... | 552 | 7e-157 |

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

<table>
<thead>
<tr>
<th>Code</th>
<th>Name</th>
<th>Genomes</th>
<th>Proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Archaeoglobus fulgidus</td>
<td>2420</td>
<td>187</td>
</tr>
<tr>
<td>O</td>
<td>Halobacterium sp. NRC-1</td>
<td>2605</td>
<td>170</td>
</tr>
<tr>
<td>M</td>
<td>Methanococcus jannaschii</td>
<td>1786</td>
<td>136</td>
</tr>
<tr>
<td></td>
<td>Methanothermobacter thermautotrophicum</td>
<td>1873</td>
<td>138</td>
</tr>
<tr>
<td>F</td>
<td>Thermoplasma acidophilum</td>
<td>1482</td>
<td>123</td>
</tr>
<tr>
<td></td>
<td>Thermoplasma nodosum</td>
<td>1499</td>
<td>124</td>
</tr>
<tr>
<td>K</td>
<td>Pyrococcus horikoshii</td>
<td>1800</td>
<td>137</td>
</tr>
<tr>
<td>L</td>
<td>Pyrococcus abyssi</td>
<td>1768</td>
<td>143</td>
</tr>
<tr>
<td>Z</td>
<td>Aeropyrum pernix</td>
<td>1841</td>
<td>117</td>
</tr>
<tr>
<td>Y</td>
<td>Saccharomyces cerevisiae</td>
<td>5955</td>
<td>2275</td>
</tr>
<tr>
<td>C</td>
<td>Agrobacterium radiobacter</td>
<td>1560</td>
<td>1325</td>
</tr>
<tr>
<td>V</td>
<td>Thermotoga maritima</td>
<td>1858</td>
<td>1522</td>
</tr>
<tr>
<td>D</td>
<td>Deinococcus radiodurans</td>
<td>3187</td>
<td>2229</td>
</tr>
<tr>
<td>B</td>
<td>Mycobacterium tuberculosis</td>
<td>3927</td>
<td>2590</td>
</tr>
<tr>
<td>A</td>
<td>Mycobacterium leprae</td>
<td>1605</td>
<td>1195</td>
</tr>
<tr>
<td>L</td>
<td>Lactococcus lactis</td>
<td>2267</td>
<td>1625</td>
</tr>
<tr>
<td>E</td>
<td>Streptococcus pneumoniae</td>
<td>1697</td>
<td>1215</td>
</tr>
<tr>
<td>E</td>
<td>Bacillus subtilis</td>
<td>4118</td>
<td>2870</td>
</tr>
<tr>
<td>C</td>
<td>Bacillus halodurans</td>
<td>4066</td>
<td>2870</td>
</tr>
<tr>
<td>C</td>
<td>Sarcina pasteurii</td>
<td>3167</td>
<td>2150</td>
</tr>
<tr>
<td>E</td>
<td>Escherichia coli K-12</td>
<td>4275</td>
<td>3415</td>
</tr>
<tr>
<td>E</td>
<td>Escherichia coli O157</td>
<td>5315</td>
<td>3665</td>
</tr>
<tr>
<td>C</td>
<td>Pasteurella pestis</td>
<td>575</td>
<td>56</td>
</tr>
<tr>
<td>P</td>
<td>Pseudomonas aeruginosa</td>
<td>5567</td>
<td>4395</td>
</tr>
<tr>
<td>G</td>
<td>Pseudomonas putidae</td>
<td>3835</td>
<td>2825</td>
</tr>
<tr>
<td>K</td>
<td>Haemophilus influenzae</td>
<td>1714</td>
<td>1541</td>
</tr>
<tr>
<td>H</td>
<td>Pasteurella multocida</td>
<td>2015</td>
<td>1751</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 COGntor to compare the protein sequence to the COG database.

Paste the FASTA formatted protein sequence from GenScan into the COGntor 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".
<table>
<thead>
<tr>
<th>Protein</th>
<th>E</th>
<th>COG1448</th>
<th>Aspartate/aromatic aminotransferase</th>
<th>Help</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>424 letters</td>
<td></td>
</tr>
<tr>
<td>00001</td>
<td>YLB0276</td>
<td>(422)</td>
<td>COG1448</td>
<td></td>
</tr>
<tr>
<td>00002</td>
<td>YMB0940</td>
<td>(397)</td>
<td>COG1448</td>
<td></td>
</tr>
<tr>
<td>00003</td>
<td>YMB0713</td>
<td>(397)</td>
<td>COG1448</td>
<td></td>
</tr>
<tr>
<td>00004</td>
<td>YMB1517</td>
<td>(396)</td>
<td>COG1448</td>
<td></td>
</tr>
<tr>
<td>00005</td>
<td>YMB1339</td>
<td>(395)</td>
<td>COG1448</td>
<td></td>
</tr>
<tr>
<td>00006</td>
<td>YMB0611</td>
<td>(396)</td>
<td>COG1448</td>
<td></td>
</tr>
<tr>
<td>00007</td>
<td>yaucA</td>
<td>(396)</td>
<td>COG1448</td>
<td></td>
</tr>
<tr>
<td>00008</td>
<td>yaucC</td>
<td>(396)</td>
<td>COG1448</td>
<td></td>
</tr>
<tr>
<td>00009</td>
<td>YMB1293</td>
<td>(415)</td>
<td>COG1448</td>
<td></td>
</tr>
<tr>
<td>00010</td>
<td>YMB1478</td>
<td>(397)</td>
<td>COG1448</td>
<td></td>
</tr>
<tr>
<td>00011</td>
<td>YMB1937</td>
<td>(397)</td>
<td>COG1448</td>
<td></td>
</tr>
<tr>
<td>00012</td>
<td>YMB0038</td>
<td>(402)</td>
<td>COG1448</td>
<td></td>
</tr>
<tr>
<td>00013</td>
<td>FAM0070</td>
<td>(395)</td>
<td>COG1448</td>
<td></td>
</tr>
<tr>
<td>00014</td>
<td>FAM0478</td>
<td>(397)</td>
<td>COG1448</td>
<td></td>
</tr>
<tr>
<td>00015</td>
<td>FAM0406</td>
<td>(394)</td>
<td>COG1448</td>
<td></td>
</tr>
<tr>
<td>00016</td>
<td>FAM0061</td>
<td>(404)</td>
<td>COG1448</td>
<td></td>
</tr>
<tr>
<td>00017</td>
<td>FAM0540</td>
<td>(402)</td>
<td>COG1448</td>
<td></td>
</tr>
</tbody>
</table>

Pathways / PHEYLALANINE/TYROSINE BIOSYNTHESIS
Functional systems LEUCINE BIOSYNTHESIS

A 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
|     | A  | B  | C  | D  | E  | F  | G  | H  | I  | J  | K  | L  | M  | N  | O  | P  | Q  | R  | S  | T  | U  |
|-----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
|     |   Afp |   Dva |   RvA |   MYb |   Lla |   HAC |   Cnp |   mBR |   zacC |   ZnpC |   ZnpB |   Epy |   N  |   X  |   Y |   Z |   T  |   P  |   K  |   J  |   I  |
|     | Afn | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  |
|     |   H3 | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  |
|     |   H1617 | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  |
|     |   X00621 | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  |
|     |   X00136 | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  |
|     | NMB0940 | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  |
|     | NMB1671 | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  |
|     | NMB1937 | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  |
|     | S Mla | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  |
|     | U HPY | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  |
|     | J Mio | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  |
|     | X Hpr | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  |
|     | I CT0377 | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  |
|     |   CT0740 | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  | -  |
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 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
Multiple sequence alignment by Florence Corpet

Available files:
- Sequence input file
- Results as a fasta file
- Results as a text page
- Results as a postscript page(s) with EScript (protein only)
- Alignment and tree description file
  
Get a better view of your protein family - phylogenetic tree, pruned tree and subtrees, summarized coloured alignment and subalignments.
- Results as an html page (needs to enable style sheets)
- Results as a text page with colour indications (needs 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).
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
19:17:19|GENSCAN_pre  PFIDMAYQGF ATGDIDRDAQ AVRTFE.... ADGHDFCL AQSFAKNMGL
YLR027c  ALFDTAYQGF ATGDLDKDAY AVRLGV.... EKLSSTVSPVF CQSFAKNAGM
tyrB  PFLDIAYQGF GAG.MEEDAY AIRAIA.... SAGLP.ALV SNSFSKFISL
ZtyrB  PFLDIAYQGF GAG.MEEDAY AIRAIA.... SAGLP.ALV SNSFSKFISL
NMB1678  PFMDIAYQGF GGD.LDSDAY AVRKAV.... EMLPF.LVF SNSFSKNSL
NMA1937  PFMDIAYQGF GGD.LDSDAY AVRKAV.... EMLPF.LVF SNSFSKNSL
PA3139  PFLDIAYQGF GNG.LEEADA VRALPA.... Q SGLS.FFV SSSFSKFSL
XF0036  PCIDLAYQGF NQG.IDADAY AIRLIA.... EGISNYVV ANSYSFSKL
aspC  PLDFAYQGF ARG.LEEADA GLRAFA.... A.HKELIV ASSYSKFGL
ZaspC  PLDFAYQGF ARG.LEEADA GLRAFA.... A.HKELIV ASSYSKFGL
VC1293  PLDFAYQGF ASG.IEEDAA GLRIFA.... K Y.NSEILV ASSFSKFGL
H11617  PLDFAYQFL ANG.IEEDAA GLRIFA.... K N.HKELIV ASSYSKFGL
PM0621  PLDFAYQGF ANG.IEEDAA GLRIFA.... K N.HKELIV ASSYSKFGL
NMB0540  PLDFAYQGF GNG.LEEADA GLRVFL.... K H.NTELLI ASSYSKNFGM
NMA0719  PLDFAYQGF GGD.LEDSDAY GLRIVA.... K H.NTELLI ASSYSKNFGM
VCA0513  PFVDIAYQGF GGD.LEQDAQ GLRYMA.... E.R.MEELI TTSCSKNFGL
ml10405  PFVDIAYQGF GGD.LEEADAL GLRLLA.... A K.VFGMVV ASSCSKFNF
PA0870  PLIDFAYQGF GNG.LEEADA GLRVFL.... G E.LPEVLV TSSCSKNFG
CT637  PFFDAMAYFQFF GGD.LEEADAL GLRVFL.... K H.NTELLI ASSYSKNFGM
Cpn0740  PFFDAMAYQGF AHG.IEADAL GLRLLA.... A K.VPAMLV ASSCSKFNF
YKL106w  PIVDMAYQGF ESGNLLKDAY LLRLCLNVNK YPNWSNMQFL CQSFAKNAGM
Consensus  Pf.D.AYQGF..G..ie.Da....Rl.a....v a.S.sKnfg$

300
19:17:19|GENSCAN_pre  YGERAGAFTV LCSDE..... EE..... AARVMSSQVI LIRGLYSNPP
YLR027c  YGERVCQFHL ALTKQ..... AQNKT I.PAVTSQALK IIPSIVSNPP
tyrB  YGERVGGGLSV MCEDA..... EA..... AARVgLQGLA TTVRNYSSPP
ZtyrB  YGERVGGGLSV MCEDA..... EA..... AARVgLQGLA TTVRNYSSPP
NMB1678  YGERVGGGLSV VCPNK..... EE..... ADLVgQLKFA TVRBIYSSPP
NMA1937  YGERVGGGLSV VCPNK..... EE..... ADLVgQLKFA TVRBIYSSPP
PA3139  YGERVGALSI VTESR..... DE..... SARVLSQVRK VTRNYSNPP
XF0036  YGERVGGGLSI VASNT..... EQ..... QAASQSQVKR IIRTIYSSPP
aspC  YNERVAGCTL VAADS..... ET..... VDRASQMKQA AIRANYSNPP
ZaspC  YNERVAGCTL VAADS..... ET..... VDRASQMKQA AIRANYSNPP
VC1293  YNERVAGFTL VAPST..... TV..... AETAFSQVKR IIRSIYSNPP
H11617  YNERVAGFTL VAENA..... EI..... ASTSLTQVKL IIRTYLNSPA
PM0621  YSERVAFTL VADEE..... QT..... AATALTQKTL IIRTYLNSPA
NMB0540  YNERVAGFTL VAEDA..... ET..... AARASQVKTL IIRTYLNSPA
NMA0719  YERRTAIVGQV5 IGKNO..... QE..... VTNARGKMFL LARSTYTMPP
ML10405  YRDRVMGFALV QNQ..... QN..... ADVAMSQMFL ARAAMYSNPP
PA0870  YRDRVAMGALV CAOQA..... EK..... LTLDRSQAFL LARNLWTPP
CT637  YGERVQFSLV HPQDK..... LD..... LNLIFSFLKQ IQRGEYSSPP
Cpn0740  YGERVQFSLV HPQDK..... LD..... LNLIFSFLKQ IQRGEYSSPP
YKL106w  YGERVQFSLV HPQDK..... LD..... LNLIFSFLKQ IQRGEYSSPP
Consensus  Yg#RvGa...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 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 CDD 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. 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 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.
CD Search Reference:

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)