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:
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 (ScanProsit)
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.
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 [ ]
## Predicted genes/exons:

<table>
<thead>
<tr>
<th>Gn.Ex</th>
<th>Type</th>
<th>S</th>
<th>Begin</th>
<th>...End</th>
<th>Len</th>
<th>Fr</th>
<th>Ph</th>
<th>I/Ac</th>
<th>Do/T</th>
<th>CodRg</th>
<th>P....</th>
<th>Tscr..</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.01</td>
<td>Sngl</td>
<td>+</td>
<td>27</td>
<td>458</td>
<td>432</td>
<td>2</td>
<td>0</td>
<td>48</td>
<td>49</td>
<td>383</td>
<td>0.447</td>
<td>24.68</td>
</tr>
<tr>
<td>1.02</td>
<td>PlyA</td>
<td>+</td>
<td>489</td>
<td>494</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.05</td>
</tr>
<tr>
<td>2.00</td>
<td>Prom</td>
<td>+</td>
<td>830</td>
<td>869</td>
<td>40</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-6.86</td>
</tr>
<tr>
<td>2.01</td>
<td>Init</td>
<td>+</td>
<td>1002</td>
<td>1069</td>
<td>68</td>
<td>2</td>
<td>2</td>
<td>53</td>
<td>89</td>
<td>83</td>
<td>0.970</td>
<td>3.88</td>
</tr>
<tr>
<td>2.02</td>
<td>Intr</td>
<td>+</td>
<td>2549</td>
<td>2708</td>
<td>160</td>
<td>2</td>
<td>1</td>
<td>72</td>
<td>105</td>
<td>284</td>
<td>0.980</td>
<td>28.49</td>
</tr>
<tr>
<td>2.03</td>
<td>Intr</td>
<td>+</td>
<td>2771</td>
<td>2872</td>
<td>102</td>
<td>1</td>
<td>0</td>
<td>10</td>
<td>86</td>
<td>251</td>
<td>0.999</td>
<td>17.47</td>
</tr>
<tr>
<td>2.04</td>
<td>Intr</td>
<td>+</td>
<td>2935</td>
<td>3183</td>
<td>249</td>
<td>0</td>
<td>0</td>
<td>73</td>
<td>100</td>
<td>586</td>
<td>0.999</td>
<td>55.93</td>
</tr>
<tr>
<td>2.05</td>
<td>Term</td>
<td>+</td>
<td>3253</td>
<td>3948</td>
<td>696</td>
<td>0</td>
<td>0</td>
<td>90</td>
<td>49</td>
<td>1324</td>
<td>0.999</td>
<td>122.25</td>
</tr>
<tr>
<td>2.06</td>
<td>PlyA</td>
<td>+</td>
<td>4120</td>
<td>4125</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.05</td>
</tr>
<tr>
<td>3.04</td>
<td>PlyA</td>
<td>-</td>
<td>4162</td>
<td>4157</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.45</td>
</tr>
<tr>
<td>3.03</td>
<td>Term</td>
<td>-</td>
<td>4448</td>
<td>4261</td>
<td>188</td>
<td>0</td>
<td>2</td>
<td>37</td>
<td>42</td>
<td>95</td>
<td>0.922</td>
<td>-2.55</td>
</tr>
<tr>
<td>3.02</td>
<td>Intr</td>
<td>-</td>
<td>4635</td>
<td>4511</td>
<td>125</td>
<td>2</td>
<td>2</td>
<td>44</td>
<td>90</td>
<td>91</td>
<td>0.949</td>
<td>5.13</td>
</tr>
<tr>
<td>3.01</td>
<td>Init</td>
<td>-</td>
<td>5046</td>
<td>4694</td>
<td>353</td>
<td>0</td>
<td>2</td>
<td>66</td>
<td>43</td>
<td>485</td>
<td>0.897</td>
<td>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):**

>13:14:50|GENSCAN_predicted_peptide_1|143_aa
MPRTLPWTTVFATASSARAKMEKLTVFLLRMSALVSVQPSMATRNVFLVPFDQSLN
SRAPAKTTSAAAITAYSLIIHILEQGKRIGLWFRLWLSPLSASSQRYESTKSGESPKT
TQSFRMNGKQLRAATQKKAFFDD

>13:14:50|GENSCAN_predicted_peptide_2|424_aa
MSQICRKGLLLSNRLAPALRCKSTWFSEVQMGFDAILGVTEAFKKDTNKINLGAAGA
YRDNTQFVLPSVREAERKVRSRLDKKEYATIIGIPEYKIAELGKGSRLAAKH
VTAQSISETGALRIGAFLAKFWQGNREIYIPSPSWGNHVAIFEPHAGLFPVNRYYDKDT
CALDFGLLIEDLKKILEKSVIPLHACAHNPTGVDPTLEQWREISALVKKRNLYPFIDMAY
QQFATGDIIDRDAQAVRTFEADGHDFCLAQSFANNGKLGERAGAFVTVLCSDEEAAARVMS
QVUILIIGLSNPPVHGARIAAEILNNEDLRAQWLKDVKLMADRIIDVRTKLDNLIKLG
SSQNDHIVNQIMFCFTGLKPEQVQKLKIDHHSVYLTNDGRVSMAGVTSKNVEYLAESIH
KVT

>13:14:50|GENSCAN_predicted_peptide_3|221_aa
MSNLQQLNLVTSWMLTLEKQGCHNLIRAGASGVRIAMVLSFSFRFNSQHLECNIHPKF
LHRDFHFRRNLNYGNKTHVNTIIVDDDNKAIVIALDRSDRYYACDGGCLDEPVILTQN
RRQFPVKLPELTAILYTEDKQHEELHHAHVEAPAEQHQLIALHRHGHQQLGGL
PTLFWVSVCIIIVFHIFLCLKLIIKEYCEPSDKLRYRNKF
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**
ScanProsite Results Viewer

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

**Hits for all PROSITE (release 20.19) motifs on sequence 13-14-50-GENSCAN_predicted_peptide_2-424_aa :**

found: 1 hit in 1 sequence

13-14-50-GENSCAN_predicted_peptide_2-424_aa (424 aa)

<table>
<thead>
<tr>
<th>Rule</th>
<th>1</th>
<th>100</th>
<th>200</th>
<th>300</th>
<th>400</th>
<th>500</th>
<th>600</th>
<th>700</th>
<th>800</th>
<th>900</th>
<th>1000</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>
</tr>
</tbody>
</table>

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

Hits by **AA_TRANSFER_CLASS_1** Annonexanteferases class I pyridoxal-5-phosphate attachment site:

13-14-50-GENSCAN_predicted_peptide_2-424_aa (424 aa)

270 - 283: RSFAGSG/GERA6

Legend:
- disulphide bridge
- active site
- other 'ranges'
- other sites

horizontal scaling: [0.6]
do not show text labels: []
do not show ranges in hits: []
do not show ranges in hits: []
redo query
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 * tyrE*).
- 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 cobalamnin biosynthesis.
- Yeast hypothetical protein YUL050w

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_PS00105, Aminotransferases class-I pyridoxal-phosphate attachment site (PATTERN)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Consensus pattern:</strong></td>
</tr>
<tr>
<td>K is the pyridoxal-P attachment site</td>
</tr>
</tbody>
</table>

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**:

**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)**
NCBI BLAST/blastp suite: BLASTP programs search protein databases using a protein query.

**Enter Query Sequence**
- Enter accession number, gi, or FASTA sequence
- Query subrange

**Choose Search Set**
- Database: Swissprot protein sequences
- Organism: Optional
- Entrez Query: Optional

**Program Selection**
- Algorithm: blastp (protein-protein BLAST)
- PSI-BLAST (Position-Specific Iterated BLAST)
- PHI-BLAST (Pattern Hit Initiated BLAST)

[BLAST] Search database swissprot using Blastp (protein-protein BLAST)

Show results in a new window
Clusters of Orthologous Groups of proteins (COGs) were delineated by comparing protein sequences encoded in 43 complete genomes. COGs consist of individual proteins or groups of paralogs from at least 3 genomes.

<table>
<thead>
<tr>
<th>Code</th>
<th>Name</th>
<th>Proteins in COG</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Archaeoglobus fulgidus</td>
<td>2420</td>
<td>1873</td>
</tr>
<tr>
<td>O</td>
<td>Halobacterium sp. NRC-1</td>
<td>2605</td>
<td>1706</td>
</tr>
<tr>
<td>M</td>
<td>Methanosarcina mazei</td>
<td>1706</td>
<td>1356</td>
</tr>
<tr>
<td></td>
<td>Methanosarcina thermophila</td>
<td>1873</td>
<td>1356</td>
</tr>
<tr>
<td>F</td>
<td>Thermoplasma acidophilum</td>
<td>1482</td>
<td>1217</td>
</tr>
<tr>
<td>T</td>
<td>Thermoplasma volantium</td>
<td>1499</td>
<td>1248</td>
</tr>
<tr>
<td>E</td>
<td>Pyrococcus horikoshii</td>
<td>1800</td>
<td>1370</td>
</tr>
<tr>
<td>K</td>
<td>Pyrococcus abyssi</td>
<td>1768</td>
<td>1459</td>
</tr>
<tr>
<td>Z</td>
<td>Anacystis myxosha</td>
<td>1841</td>
<td>1177</td>
</tr>
<tr>
<td>Y</td>
<td>Saccharomyces cerevisiae</td>
<td>5955</td>
<td>2290</td>
</tr>
<tr>
<td>Q</td>
<td>Aquifex aeolicus</td>
<td>1560</td>
<td>1372</td>
</tr>
<tr>
<td>V</td>
<td>Thermotoga maritima</td>
<td>1858</td>
<td>1522</td>
</tr>
<tr>
<td>D</td>
<td>Drosophila melanogaster</td>
<td>3187</td>
<td>2226</td>
</tr>
<tr>
<td>E</td>
<td>Mycobacterium tuberculosis</td>
<td>3927</td>
<td>2589</td>
</tr>
<tr>
<td>E</td>
<td>Mycobacterium leprae</td>
<td>1605</td>
<td>1134</td>
</tr>
<tr>
<td>L</td>
<td>Lactoceps lactis</td>
<td>2287</td>
<td>1616</td>
</tr>
<tr>
<td>R</td>
<td>Streptococcus pyogenes</td>
<td>1697</td>
<td>1231</td>
</tr>
<tr>
<td>T</td>
<td>Escherichia coli K12</td>
<td>4118</td>
<td>2877</td>
</tr>
<tr>
<td>E</td>
<td>Escherichia coli O157</td>
<td>4066</td>
<td>2877</td>
</tr>
<tr>
<td>C</td>
<td>E. coli</td>
<td>3167</td>
<td>2152</td>
</tr>
<tr>
<td>F</td>
<td>Pseudomonas aeruginosa</td>
<td>4275</td>
<td>3416</td>
</tr>
<tr>
<td>G</td>
<td>Vibrio cholerae</td>
<td>5567</td>
<td>4362</td>
</tr>
<tr>
<td>H</td>
<td>Haemophilus influenzae</td>
<td>1714</td>
<td>1544</td>
</tr>
<tr>
<td>I</td>
<td>Pasteurella multocida</td>
<td>2015</td>
<td>1725</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 COGmator to compare the protein sequence to the COGs database.

Paste the FASTA formatted protein sequence from GenScan into the COGmator 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".
Clusters of Orthologous Groups (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.

**Protein/Genome name:**

<table>
<thead>
<tr>
<th>Code</th>
<th>Name</th>
<th>Proteins in COGs</th>
<th>Principal component analysis of genomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Archaeoglobus sulfidoeae</td>
<td>2420 1872</td>
<td>List of COGs</td>
</tr>
<tr>
<td>O</td>
<td>Halobacterium sp. NCIB-1</td>
<td>2605 1701</td>
<td>Distribution</td>
</tr>
<tr>
<td>M</td>
<td>Methanococcus jannaschii</td>
<td>1786 1330</td>
<td>Co-occurrences</td>
</tr>
<tr>
<td>P</td>
<td>Methanobacterium thermoautotrophicum</td>
<td>1873 1388</td>
<td>Phylogenetic patterns</td>
</tr>
<tr>
<td>K</td>
<td>Pyrococcus horikoshii</td>
<td>1482 1230</td>
<td>Phylogenetic patterns search</td>
</tr>
<tr>
<td>Z</td>
<td>Zymobacterium termitis</td>
<td>1800 1378</td>
<td>Functional categories</td>
</tr>
<tr>
<td>Y</td>
<td>Saccharomyces cerevisiae</td>
<td>1768 1456</td>
<td>J K L</td>
</tr>
<tr>
<td>Q</td>
<td>Aspergillus niger</td>
<td>1841 1178</td>
<td>D O M N P T</td>
</tr>
<tr>
<td>V</td>
<td>Thermotoga nitroprofeta</td>
<td>5955 2290</td>
<td>G C E F H I Q</td>
</tr>
<tr>
<td>D</td>
<td>Drosophila melanogaster</td>
<td>1560 1329</td>
<td>R S</td>
</tr>
<tr>
<td>B</td>
<td>Mycobacterium tuberculosis</td>
<td>1858 1527</td>
<td>Pathways and functional systems</td>
</tr>
<tr>
<td>L</td>
<td>Lactococcus lactis</td>
<td>3187 2226</td>
<td>FTP</td>
</tr>
<tr>
<td>E</td>
<td>Streptomyces coelicoflavus</td>
<td>3927 2585</td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>Bacillus subtilis</td>
<td>1697 1211</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>Bacillus halodurans</td>
<td>4118 2870</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Synechocystis</td>
<td>4066 2878</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3167 2159</td>
<td></td>
</tr>
</tbody>
</table>
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
Sequence data

Cut and paste your sequence here below:

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Format</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

or select a file

Sequence input format: Auto

<table>
<thead>
<tr>
<th>Alignment</th>
<th>Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score 1</td>
<td>Score 2</td>
</tr>
<tr>
<td>Score 3</td>
<td>Score 4</td>
</tr>
</tbody>
</table>

21
Consensus levels: high=90% low=50%

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

14:13:11|GENSCAN pre  PFIDMAYQGF ATGDIDRDQA AVRTFE.... ..ADGHDFCL AQSFAKNMGL
       YLR027c  ALFDTAYQGF ATGDLDKDAY AVRLGV...E KLSTVSPVFV CQSFAKNAGM
tyrB      PFLDIAYQGF GAG.MEEDAY AIRAIA...S ..AGLP.ALV SNSFSDIFSLL
ZtyrB     PFLDIAYQGF GAG.MEEDAY AIRAIA...S ..AGLP.ALV SNSFSDIFSLL
NMB1678   PFMDIAYQGF GGD.LDSDAY AVRKAV...E ..MELP.LFV SNSFSDIFSLL
NMA1937   PFMDIAYQGF GGD.LDSDAY AVRKAV...E ..MELP.LFV SNSFSDIFSLL
PA3139    PFDIAYQGF GNG.IEEDAY GLRIF...K ..Y.NSEILV ASSFSKNFGL
aspC      PLDFAYQGF ARG.LEEAE GLRAFA...A ..M.HKELIV ASSFSKNFGL
ZaspC     PLDFAYQGF ARG.LEEAE GLRAFA...A ..M.HKELIV ASSFSKNFGL
VC1293    PLDFAYQGF ASG.VEEAA GLRIFA...K ..Y.NSEILV ASSFSKNFGL
H11617    PLDFAYQQL ASG.LEEAY GLRAFA...A ..N.IKELLY ASSFSKNFGL
PM0621    PLDFAYQGF ARG.LEEAF GLRTFA...K ..N.IKELLY ASSFSKNFGL
NMB0540   PLDFAYQGF GNG.LEEAF GLRFV...K ..H.NTEILL ASSFSKNFGL
NMA0719   PLDFAYQGF GNG.LEEAF GLRFV...K ..H.NTEILL ASSFSKNFGL
VCA0513   PFVDAAYQGF GGD.LEQDAQ GLRYMA...E ..R.MEELI TTSCSKNFGL
mll0405   PFVDAAYQGF GGD.LEQDAQ GLRLA...A ..K.VPEMVV ASSCSKNFGL
PA0870    PLIDFAYQGF ARG.LEEAF GLRDA...G ..E.LPEVVL TSSCSKNFGL
CT637     PFFDMayLGF ASG.IEDDR PVRDLC...E ..AGVTFVV AGGASKNFSL
CPn0740   PFFDAYQGF AGH.IELDRK PIESIF...S ..EGNTVVL AASSSKNFSL
YKL106w   PIVDMAYQGL ESNLLKDAY LLRLCNVVK YPNWSNGLF CQSFANMGL
Consensus  Pf.D.AYQGf ...G.le.Da.. ...Rl.a.... ........v a.S.sKnfg$

14:13:11|GENSCAN_pre  YGERAGAFTV LCSDE....... ....EE.... AARVMSQVKI LIRGLYSNPP
YLR027c  YGERVCGFHL ALTKQ........ ....AQNKTI KPANTSQMLK IIRSEVSNPP
tyrB  YGERVGLLSV MCEDA........ ....EA.... AGRVGLQNLK TVRRNYSSPP
ZtyrB  YGERVGLLSV LCEDA........ ....EA.... AGRVGLQNLK TVRRNYSSPP
NMB1678  YGERVGLLSV VCPNK........ ....EE.... ADLVFGQLKF TVRRIYSSPP
NMA1937  YGERVGLLSV VCPNK........ ....EE.... ADLVFGQLKF TVRRIYSSPP
PA3139  YGERVLS1 VTESR........ ....DE.... SARVLSQVK KIVRTYSNPP
XFM036  YGERVGLLS1 VASNT........ ....EQ.... AQAIQSQVK IIRTIYSSPS
aspC  YNERVGACTL VAADS........ ....ET.... VDRAFSQMK AIAIYSNPP
ZaspC  YNERVGACTL VAADS........ ....ET.... VDRAFSQMK AIAIYSNPP
VC1293  YNERVGAFTL VAADS........ ....TV.... ASTSLTQVK IIRTLYSNPA
HI1617  YNERVGAFTL VAADS........ ....EI.... AATAALTQVK IIRTLYSNPA
PM0621  YNERVGAFTL VAADS........ ....ET.... AARAHSQVK IIRTLYSNPA
NMB0540  YNERVGAFTL VAADS........ ....ET.... AARAHSQVK IIRTLYSNPA
NMA0719  YNERVGAFTL VAADS........ ....AT.... AARAHSQVK IIRTLYSNPA
VCA0513  YRERTGAIV IGSNQ........ ....QE.... VTNARGKML LTARSTYMP
ml10405  YRDRVGAAMV LARDS........ ....AQ.... ADVAMSQML AARAMYSNPP
PA0870  YRDRVGAAMV LARDS........ ....EK.... LTDLRSQAF AIARLSYPMP
CT637  YGRVGGFFGA IHQDK........ ....LD.... LNRILSFLEE QIRGEYSPQ
Cpn0740  YGERVGYFAV HSTFT........ ....DE.... LVIKHSFLEE KIRGEYSPQ
YKL106w  YGERVGLSSV ITPTAANNK FNPLQQKNSL QQNITDQLK KVGGMYSPQ
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

- Pfam
- SMART
- COG
- KOG
- PRK

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

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.
User Annotations

Annotation Control

Available

Displayed

Turn On

Turn Off

Move Up

Move Down

Selection:

Description:

New

Show

Edit

Move

Delete

Done

Edit Annotation

Edit Annotation:

Name: lvs

Description:

Edit Style

OK

Cancel