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
GENSCAN 1.0  Date run: 19-Jan-107  Time: 08:28:26
Sequence 08:28:21 : 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 |
Predicted peptide sequence(s):

>08:28:21|GENSCAN_predicted_peptide_1|143_aa
MPRTLPWTTVFTAVASSARAKSMELTVFLLRMH5ALVVSQSAPSMATRVNLVPFDQSLN
SRAPAKTTSAAQAIAYLSIFHLIELQGKRIGWLFRWLSPLSASSQRYESTKSGESPKT
TOQFRMGKQLRAATQKKAFFD

>08:28:21|GENSCAN_predicted_peptide_2|424_aa
MSQ1CKRGLLINSNRLAPAALRCKSFWDEVQMGFPFDAILGVTEAFKDTNPKKINLGA
YRDCTQFPLPSVREAEKRVSRLSDKLEYATIGIEPEYNKAIIELALKGSKRALAKHN
VTAQSIISGTGALRIGAALKFWQGNIREYFSPSWGNHVAIFEHAGLPVRNYRYDKDT
CALDPFFGLIEDLKKIEPSVLLHACAHNPVTGDPDLEQWREISALVKKRNLYFMDAY
QGFATGDIDRDAFAQAVTFEADGHDFCLAQSFAKNMGLYGERAGAFTVLCSDEDDAEARVM
QVQILRGYSNPPVHGARIAAEILNNETDRLAQWRLKVDKLMADDIIDVRKDKDNLKL
SSQNWDHVQIOMFCFTGLPKKLPQVQLIKDHSVYLTNGDRVSMAGVTSKNVEYLAESIH
KVTK

>08:28:21|GENSCAN_predicted_peptide_3|221_aa
MSNLQQNLNLSTSWMLTELQGCHNLRAGASPQIQAMVLSFGSRFSNQHLCNIHPKF
LHRDFHFRRLYNNGNKTHVNVTLVDDDNKAVINIALDSSDRSYYACGGCLDFPVLLTQN
RRQFPVKLTLPIVTLITYEDKQHMEELHHAIVKVEAPAEQQHQLALHRHGQQLGL
PTLVWSVCA1II1VFIFLCKL1IKEYCEPSDCKLRYNYKF
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 PDOCxxxx number.

**Hit from ScanProsite**

**Prosite pattern**

**Prosite Summary**
The ScanProsite tool [Help] allows to scan protein sequence(s) (either from UniProt Knowledgebase (Swiss-Prot/EMBL) 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/EMBL) and/or PDB databases, OR
- Enter one or more sequences [raw, Swiss-Prot or fasta format] and/or UniProt Knowledgebase (Swiss-Prot/EMBL) 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/EMBL accession number(s) [AC] (e.g. P00747) and/or sequence identifier(s) [ID] (e.g. ENTK_HUMAN) and/or PDB identifier, and/or paste your own protein sequence(s) in the box below. [Leave this box blank to scan PROSITE entries against selected protein databases]

**PROSITE pattern(s)/profile(s) to scan for:**

Enter one or more PROSITE accession number(s) (e.g. PS50240), and/or identifier(s) (e.g. CHEB), and/or type your pattern(s) in PROSITE format in the box below. [Leave this box blank to scan sequence(s) against the entire PROSITE database]

**General options:**

- Exclude motifs with a high probability of occurrence
- Show low level score
- Do not scan profiles [User Manual]

Show only sequences with at least [n] hits(s)

Maximum of matched sequences

1000

Output format: [Graphical, rich view]

Retrieve complete sequences

Your e-mail (optional): [will send results by e-mail]

[START THE SCAN] [reset]
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)

<table>
<thead>
<tr>
<th>Hit by patterns:</th>
<th>[1 hit (by 1 pattern) on 1 sequence]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hits by PS00105</td>
<td>AA_TRANSFER_CLASS_1  Aminotransferase class-I pyridoxal-phosphate attachment site :</td>
</tr>
<tr>
<td>USERSEQ1</td>
<td>270 - 283: SFARPGSg2gRAG</td>
</tr>
</tbody>
</table>
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 aapC).
- 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 ppcC, which is involved in cobaltamin 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:
April 2005 / Pattern and text revised.

Technical section:

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

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequences known to belong to this class detected by the pattern</td>
<td>ALL</td>
</tr>
<tr>
<td>Other sequence(s) detected in Swiss-Prot</td>
<td>1</td>
</tr>
</tbody>
</table>
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 = 09:35:17/GENSCAN_predicted_peptide_2/424_aa(424 letters)**

The request ID is [1170341203-21962-19203141397]BLASTQ4

The results are estimated to be ready in 13 seconds but may be done sooner.

Please press "FORMAT!" when you wish to check your results. You may change the formatting options for your result view again. You may also request results of a different search by entering any other valid request ID to see other recent jobs.

**Format**

- Show: Graphical Overview, Linkout, Sequence Retrieval, NCBI-gs (alignment) in HTML, format
  - CDS feature
  - New View

<table>
<thead>
<tr>
<th>Masking Character</th>
<th>Lower Case</th>
<th>Masking Color</th>
<th>Grey</th>
</tr>
</thead>
</table>

**Number of:** Descriptions: 100, Alignments: 50, Graphic overview: 100

**Alignment view**

- Pairwise
- Pairwise with identities

**Format for PSI-BLAST**

- Query-anchored without identities
- Flat query-anchored with identities
- Flat query-anchored without identities

**Limit results by selected query**

- AND: All organisms
Taxonomy reports

Query: 13:39:01|OXOSCAN_predicted_peptide_2|424
Length: 424

Distribution of 59 Blast Hits on the Query Sequence

Mouse-over to show define and scores, click to show alignments

Distance tree of results

Sequences producing significant alignments:

<table>
<thead>
<tr>
<th>Score</th>
<th>E Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>563</td>
<td>4e-160</td>
</tr>
<tr>
<td>562</td>
<td>9e-160</td>
</tr>
</tbody>
</table>
Clusters of Orthologous Groups of proteins (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>Proteins in COG</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Archaeoglobus fulgidus</td>
<td>2420</td>
</tr>
<tr>
<td>O</td>
<td>Halobacterium sp. NRC-1</td>
<td>2605</td>
</tr>
<tr>
<td>M</td>
<td>Methanothermobacter thermautotrophicum</td>
<td>1786</td>
</tr>
<tr>
<td>C</td>
<td>Thermoplasma acidophilum</td>
<td>1482</td>
</tr>
<tr>
<td>E</td>
<td>Thermoplasma vulcanii</td>
<td>1439</td>
</tr>
<tr>
<td>K</td>
<td>Pyrococcus horikoshii</td>
<td>1800</td>
</tr>
<tr>
<td>D</td>
<td>Pyrococcus abyssi</td>
<td>1768</td>
</tr>
<tr>
<td>Z</td>
<td>Aeropyrum pernix</td>
<td>1841</td>
</tr>
<tr>
<td>X</td>
<td>Bacillus amyloliquefaciens</td>
<td>5955</td>
</tr>
<tr>
<td>Q</td>
<td>Agrobacterium radiobacter</td>
<td>1560</td>
</tr>
<tr>
<td>V</td>
<td>Thermotoga maritima</td>
<td>1858</td>
</tr>
<tr>
<td>D</td>
<td>Desulfovibrio desulfuricans</td>
<td>3187</td>
</tr>
<tr>
<td>L</td>
<td>Mycobacterium tuberculosis</td>
<td>3972</td>
</tr>
<tr>
<td>E</td>
<td>Mycobacterium leprae</td>
<td>1605</td>
</tr>
<tr>
<td>K</td>
<td>Lactococcus lactis</td>
<td>2267</td>
</tr>
<tr>
<td>F</td>
<td>Streptococcus pyogenes</td>
<td>1697</td>
</tr>
<tr>
<td>E</td>
<td>Bacillus subtilis</td>
<td>4118</td>
</tr>
<tr>
<td>C</td>
<td>Bacillus halodurans</td>
<td>4066</td>
</tr>
<tr>
<td>A</td>
<td>Sarcina</td>
<td>3167</td>
</tr>
<tr>
<td>E</td>
<td>Bacillus sp. PK7</td>
<td>4275</td>
</tr>
<tr>
<td>E</td>
<td>Bacteroides sp. D157</td>
<td>5315</td>
</tr>
<tr>
<td>F</td>
<td>Pseudomonas aeruginosa</td>
<td>5567</td>
</tr>
<tr>
<td>G</td>
<td>Pseudomonas chlororaphis</td>
<td>3833</td>
</tr>
<tr>
<td>H</td>
<td>Haemophilus influenzae</td>
<td>1714</td>
</tr>
<tr>
<td>H</td>
<td>Pasteurella 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 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 Mauve 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".
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

Sequence data

Cut and paste your sequences here below:

Available files:
- Sequence Input file
- Cluster file
- Results as a fasta file
- Results as a text page (nont)
- Results as a postscript page(s) with ESTrree (protein only)
- Alignment and tree description (html) Get a better view of your protein family, phylogenetic tree, pruned tree and subtrees, summarized coloured alignment and subalignments.
- Results as a html page (needs to enable style sheets)
- Results as a text page with colour indications (need a text editor)
- Results as a pdf 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 ALFDATAYQGF ATGDLDKDAY AVRLGV.... E KLSTVSPVVF CQSFAKNMAGM
tyrB  PFDLIAAYQGF GAG.MEEDAY AIRAIA.... S AGLF.ALV SNSFSKIFSL
ZtyrB  PFDLIAAYQGF GAG.MEEDAY AIRAIA.... S AGLF.ALV SNSFSKIFSL
NMB1678 PFMDIAAYQGF GGD.LSDDAY AVRKAV.... E MELP.LFV SNSFSKINSL
NMA1937 PFMDIAAYQGF GGD.LSDDAY AVRKAV.... E MELP.LFV SNSFSKINSL
PA3139 PFMDIAAYQGF GGM.LEEADA AVRLFA.... Q SGLS.FFV SNSFSKIFSL
XFO0396 PCIDLAYQGF NQQ.IDADAAY AIRLLA.... E EGSNYYV SNSFSKIFSL
 aspC  PLDFAYQGF ARG.LEEADA GLRAFA.... G E.MKELIV SNSFSKIFSL
ZaspC  PLDFAYQGF ARG.LEEADA GLRAFA.... G E.MKELIV SNSFSKIFSL
VC1293 PLDFAYQGF ARG.LEEADA GLRAFA.... G E.MKELIV SNSFSKIFSL
HI1617 PLDFAYQFL AGN.LEADAY ARAF.... Q E.EKELLV SNSFSKIFSL
PM0621 PLDFAYQGF GGD.LSDDAY AVRKAV.... E MELP.LFV SNSFSKINSL
NMB1678 PLDFAYQGF GGD.LSDDAY AVRKAV.... E MELP.LFV SNSFSKINSL
NMA1937 PLDFAYQGF GGD.LSDDAY AVRKAV.... E MELP.LFV SNSFSKINSL
CT637 PLDFAYQGF GGD.LSDDAY AVRKAV.... E M.NTELLI SNSFSKIFSL
NMAO719 PLDFAYQGF GGD.LSDDAY AVRKAV.... E M.NTELLI SNSFSKIFSL
VCA0513 PFVIDAYQGF GGD.LEDAA GLRYMA.... E R.MEELIV SNSFSKIFSL
m11040 PFVDIAYQGF GGD.LEDADAY GLRLLA.... A K.VPKMVL SNSFSKIFSL
PA0870 PLDFAYQGF GGD.LEEADA GLRAFA.... Q E.EKELIV SNSFSKIFSL
NSN1878 PLDFAYQGF GGD.LEEADA GLRAFA.... Q E.EKELIV SNSFSKIFSL
NMA0719 PLDFAYQGF GGD.LEEADA GLRAFA.... Q E.EKELIV SNSFSKIFSL
PA0870 PLDFAYQGF GGD.LEEADA GLRAFA.... Q E.EKELIV SNSFSKIFSL
CT637 PLDFAYQGF GPD.LSDDAY AVRKAV.... E E.MKELIV SNSFSKIFSL
NMA0719 PLDFAYQGF GPD.LSDDAY AVRKAV.... E E.MKELIV SNSFSKIFSL
YKL106w PIVDMAYQGF ESGLNKADAY LLRLCLAVNK YPNWSNSIFL CQSFAKNMAGM
Consensus  Pf.D.AYQGF ..G.le.Da. ..Rl.a.... ..........v a.S.sKngf$
301

19:17:19|GENSCAN_pre  YGERAGAFTV LCSDE....... EEE.... AARVMSQVKI LIRGLYSNPP
 YLR027c YGERVGCFHL ALTKQ....... AQNKTI KPAVTSQKLK IIRSESVSNPP
tyrB  YGERVGGGSV MCEDA....... EECMA  AGRVGQLQKA TVRNYSSNP
ZtyrB  YGERVGGGSV MCEDA....... EECMA  AGRVGQLQKA TVRNYSSNP
NMB1678 YGERVGGGSV VCPNK....... EEE.... ADLVFGQLKE TVRRIYSSPN
NMA1937 YGERVGGGSV VCPNK cc ...... EEE.... ADLVFGQLKF
TVRRIIYSSPN
PA3139 YGERVGLSI VTESR....... ETDE.... SARVLSQVKR VIRTYSNPP
XFO0396 YGERVGLSI VAQNT....... EEO.... AQAIAQSVQRK IIRLIYSSPS
 aspC  YNERVGACTL VAADS....... EET.... VDRAFSQLMK AIARANYSNPP
ZaspC  YNERVGACTL VAADS....... EET.... VDRAFSQLMK AIARANYSNPP
VC1293 YNERVGAFTL VAAS.... EET.... AEASFSQVK AAIIRYSNPP
HI1617 YNERVGAFTL VAENA....... EI.... ASISLTVKS IIRLYSNPA
PM0621 YNERVGAFTL VAETE....... EIQ.... AATALTQKV7 IIRLYSNPA
NMB0540 YNERVGAFTL VADEE....... EIA.... AARAHSVQVT IIRLYSNPA
NMA0719 YNERVGAFTL VADEE....... EIA.... AARAHSVQVT IIRLYSNPA
VCA0513 YERTGAAIV IGKNQ....... QE.... VTRNARKMLT LARSTYMPQ
m110405 YRDRVGAAMV LARDS....... AQ.... ADVAMSQMLS AARAYSNPP
PA0870 YRDRVGAAMV LARDS....... AQ.... ADVAMSQMLS AARAYSNPP
CT637 YGSVGGFFGA IHQDK....... EEO.... LNRILSFLEL KIRGEYSSPA
CPr0740 YGERVGFAYF HSHFT....... EDE.... LRVKSHSFLEL KIRGEYSSQ
YKL106w YGERVGLSV ITQTAANMK FNPIQORKSL QQNIDSLQKLK IVRGMYSSPP
Consensus  Yg#RvGa...v ............... sQk. 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 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. 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 structure 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.