A protein domain is considered to be a distinct functional and/or structural unit. A domain in a structural context refers to a segment of a polypeptide chain that can fold into an independent three-dimensional structure. It may interact with other domains of the protein or may simply be joined to other domains by a polypeptide chain. A domain in a sequence context refers to a long sequence pattern that is shared by other proteins having a common evolutionary origin. A domain may include all of the protein sequence or a part of it. A conserved domain is a recurring unit in molecular evolution whose extents can be determined by sequence and structure analysis.

The Conserved Domain Database (CDD) contains domains derived from the Smart, Pfam and Clusters of Orthologous Groups (COGs) databases. Conserved domains can be represented as multiple sequence alignments. Source alignments are processed by NCBI as follows:

- Sequences in the alignment for which a link can not be provided to a protein in Entrez are removed.
- If possible, a closely related sequence with a known structure is substituted.
- A representative sequence, preferably with a structure link, is chosen from among those in the alignment.
- A consensus sequence is made.
- A position-specific scoring matrix (PSSM) is constructed.

The Conserved Domain search (CD-search) compares a protein sequence to the PSSMs in the CDD database to identify conserved domains within it and to identify a 3-D modeling template. Since the PSSMs are the "subject", instead of the query as in PSI-Blast, the CD-search is a form of Reverse Position-Specific Blast (RPS-Blast).

The Conserved Domain Architecture Retrieval Tool (CDART) can be used to identify proteins containing the domain(s) present in the query sequence. Conserved domain(s) present in all sequences within Entrez proteins are identified using CD-search during routine NCBI processing. These pre-computed results are accessed through CDART.

The Vector Alignment Search Tool (VAST) is a computer algorithm developed at NCBI to detect similar protein 3-dimensional structures. The "structure neighbors" for every structure in NCBI' Molecular Modeling Database (MMDB)
are pre-computed. These neighbors can be used to identify distant homologs that cannot be recognized by sequence comparison alone. A VAST-search can be used for determining the structure neighbors for recently solved structures not yet in MMDB.

Cn3D is a helper application for web browsers to view 3-dimensional structures from NCBI's Entrez retrieval service. Cn3D runs on Windows, Macintosh, and Unix. Cn3D simultaneously displays structure, sequence, and alignment, and now has powerful annotation and alignment editing features.

In this course, we will learn to

- Identify a conserved domain present in the query protein using CDD
- Search for other proteins containing similar domain(s) using CDART
- Explore a 3D modeling template for the query sequence using CDD
- Find similar structures using VAST
- Visualize and annotate the 3D protein structures using Cn3D

The remainder of the handout includes the introductory slides and the screen shots of the exercise demonstrated in Problem 1.


Course developed by: Dr. Medha Bhagwat (bhagwat@ncbi.nlm.nih.gov)

Slides
Conserved Domains

• recurring unit in molecular evolution;
  extent can be determined by
  sequence and structure analysis;
• performs a defined function;
• represented as a multiple local sequence
  alignment of proteins containing the domain
  (position specific scoring matrix, pssm).

CDD Search page:

Conserved Domain Database

• A database of position-specific scoring matrices (PSSM)
• CD-Search can be used to search against the PSSMs
• Manual curation of CDs is ongoing (cd12345.version)
Position-Specific Score Matrix

Serine/Threonine protein kinases catalytic loop

PSSM scores

| A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
| 435 K | -1 | 0 | 0 | -1 | -1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 436 E | 0 | 1 | 0 | 2 | -1 | 0 | 2 | -1 | 0 | 0 | 0 | 0 | -1 | 0 | -1 | 0 | -1 | 0 | -1 | 0 |
| 437 S | 0 | 0 | -1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 438 N | -1 | 0 | -1 | -1 | 1 | 0 | -1 | 3 | 3 | -1 | -1 | -1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 439 K | -2 | 1 | -1 | -2 | 0 | 1 | -2 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 440 P | -2 | 2 | -1 | -2 | -3 | -2 | 2 | -1 | 0 | 3 | 7 | 1 | 2 | 3 | -1 | -1 | -1 | -1 | -1 | -1 |
| 441 A | 3 | 2 | 1 | -2 | 0 | 1 | -2 | -2 | 2 | 2 | 1 | 0 | 3 | 2 | 3 | 1 | 0 | 3 | 2 | 3 |
| 442 M | -3 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |
| 443 A | -4 | -2 | -1 | -3 | -5 | -2 | -4 | -6 | -6 | -5 | -3 | -2 | -3 | -4 | -5 | -3 | -5 | -7 | -5 | -7 |
| 444 H | -4 | -1 | -3 | -5 | -2 | -4 | -6 | -6 | -6 | -5 | -3 | -2 | -3 | -4 | -5 | -5 | 0 | 5 | 0 | 5 |
| 445 R | -4 | -4 | -1 | -1 | -2 | -3 | -5 | -6 | -6 | -6 | -6 | -6 | -6 | -6 | -6 | -6 | -6 | -6 | -6 | -6 |
| 446 S | -3 | 4 | 1 | 0 | -4 | -2 | 0 | 1 | -2 | 0 | 1 | -2 | 0 | 1 | -2 | 0 | 1 | -2 | 0 | 1 |
| 447 N | -4 | -5 | -6 | -6 | -6 | -5 | -6 | -6 | -6 | -5 | -6 | -6 | -6 | -6 | -6 | -6 | -6 | -6 | -6 | -6 |
| 448 K | 0 | 0 | 1 | 3 | 5 | -1 | -1 | -3 | -5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| 449 S | -3 | -2 | -3 | 0 | -2 | -2 | -3 | -3 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 | -4 |
| 450 K | 0 | 3 | 0 | 1 | -5 | 0 | 0 | 4 | -1 | -4 | -3 | -4 | -3 | -4 | -3 | -4 | -3 | -4 | -3 | -4 |
| 451 N | -4 | -3 | 8 | -1 | -5 | 2 | -3 | 1 | -6 | -6 | -6 | -6 | -6 | -6 | -6 | -6 | -6 | -6 | -6 | -6 |
| 452 I | -3 | -5 | -5 | -6 | 0 | -5 | -5 | -6 | 5 | 2 | -2 | -2 | -5 | -4 | -3 | -5 | -5 | 3 | -3 | -5 |
| 453 N | -4 | -1 | -3 | -5 | -2 | -3 | -5 | -6 | -6 | -6 | -6 | -6 | -6 | -6 | -6 | -6 | -6 | -6 | -6 | -6 |
| 454 V | -3 | -3 | -5 | -6 | -3 | 4 | -5 | -6 | 5 | 3 | 3 | 4 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 455 K | -2 | 1 | 1 | -4 | -5 | 0 | -1 | -2 | 1 | -4 | -2 | 1 | -2 | 1 | -2 | 1 | -2 | 1 | -2 | 1 |
| 456 N | 1 | 1 | 3 | 0 | -4 | -1 | 1 | 0 | -3 | -4 | -4 | -3 | -2 | -5 | -2 | -2 | -5 | -4 | -4 | -4 |
| 457 D | -3 | -2 | 5 | 5 | -1 | -1 | 1 | 0 | -5 | -4 | 0 | -2 | -5 | 1 | -2 | 0 | -6 | -4 | -5 | -5 |
| 458 L | -3 | -1 | 0 | -3 | 0 | -3 | -2 | 3 | -4 | -2 | 3 | 0 | 1 | 1 | -2 | 2 | 3 | 0 | 1 | 1 |

Catalytic loop
Problem 1

In this problem, we will follow these steps:

A. Identify conserved domain(s) present in a protein.
B. Search for other proteins containing similar domain(s).
C. Explore a 3D modeling template for the query sequence.
D. Find distant sequence homologs that may not be identified by BLAST.

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. Click on the CDD link provided below,

CDD

Paste the following protein sequence in the CD-Search query box and run the search.

```
MDPALTAAVGADLLGDRPETLWLGIGTLMLIIGTFYFV/KGWGSMFI/FGIGTEVQ/GSEMLD/YY/ADWLFTTPL/LD/LALLAKHTLARYTWLFLSTDM/VLYFLATS/LRAAAAG/PEVASTFN/VGLIETLFM/LDVTAKV/GFGILLRSRAILGOTEAPESAGAE/
```

A. What is the domain present in this protein? Obtain more information about the domain by searching in NCBI's Bookshelf

B. Go back to the CD-Search results page. Obtain a list of proteins with similar domain architecture by clicking on the "Search for similar domain architectures" button. To display the records, click on the link to the sequences and from there on the “Look up Sequences in Entrez”. Change the display from “Summary” to “FASTA”.

C. Go back to the CD-Search results page. Generate a multiple sequence alignment for the top 10 sequences representative of the conserved domain hit by clicking on the graphic of the domain. Use the "Row Display" list box pull down menu to specify "up to 5" sequences and reformat sequence alignment. Extend the “Structure” display 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.
The structure of the *Halobacterium salinarum* halorhodopsin protein and its sequence alignment with our query protein are displayed. For a better view of the backbone, remove the side chains globally (Style--Edit global style--Protein side chains). The query protein contains a bacterial rhodopsin signature (FMVLDVTAKVGF) where K is the retinal binding site. Identify these residues in the query protein and highlight the corresponding lysine residue in the halorhodopsin protein sequence.

Display the side chains of this residue (Use Style--Annotate--New--Edit Style. Change the protein backbone Rendering to Tubes, Color Scheme to User Selection and User Color to choose the color for the highlighted residue, for example yellow. Repeat these steps for the Protein Side chains row and click the Protein Side chains on. Click on the “Done” button. To zoom in, press z on the keyboard. Identify the cofactor near the lysine residue.

D. To obtain the structural neighbors for the halorhodopsin protein, first click on the structure entry link, 1TNO_A, on the CD-Browser page. Then click Links → Structure on the top right, then on 1TNO again in the Entrez Structure page, and finally on the chain A graphic. To view neighbors with 1TNO_A, select one or more of the check boxes next to the structure neighbors and view by clicking on the "View 3D Structure" button.

**Screenshots**
Many Integral Proteins Contain Multiple Transmembrane α-Helices

Although Figure 3-33 depicts glycolysis as a monomer with a single α-helix spanning the bilayer, this protein is present in erythrocyte membranes as a dimer of two identical polypeptide chains. The two membrane-spanning α-helices of glycolytic proteins are thought to form a coiled-coil structure (see Figure 3-33). Stabilized by specific interactions between the amino and side chains at the interface of the two helices, it is now known that many other transmembrane proteins contain two or more membrane-spanning α-helices. For instance, the bacterial photosynthetic reaction center (RC) comprises four subunits and several prosthetic groups, including four chlorophyll molecules. In this complex protein, three of the four subunits span the membrane, two of these subunits (L and M) each contain five membrane-spanning α-helices (see Figure 16-40).

A large and important family of integral proteins is defined by the presence of seven membrane-spanning α-helices. More than 150 such “seven-spanning” membrane proteins have been identified. This class of useful proteins is typified by bacteriorhodopsin, a protein found in a photosynthetic bacterium (Figure 3-34). Absorption of light by the retinal group attached to bacteriorhodopsin causes a conformational change in the protein that results in pumping of protons from the cytoplasm across the bacterial membrane into the extracellular space. The proton concentration gradient thus generated across the membrane is used to synthesize ATP, as discussed in Chapter 16. Both the overall arrangement of the seven α-helices in bacteriorhodopsin and the identity of most of the amino acids can be resolved by computer analysis of micrographs of two-dimensional crystals of the membrane-embedded protein taken at various angles to the electron beam.

Figure 3-34. Overall structure of bacteriorhodopsin as deduced from electron diffraction analyses of two-dimensional crystals of the protein in the bacterial membrane. The seven membrane-spanning α-helices are labeled A–G. The retinal pigment is covalently attached to lysine 216 in helix G. The approximate position of the protein in the phospholipid bilayer is indicated. (Adapted from R. Henderson et al., 1990, J. Mol. Biol. 213:659.)
### VAST neighbors for MMDB 29943, 1TN0 A

Overview: There are two main sections to this page. The first section consists of the alignment view controls, the list controls, and the advanced neighbor search controls. The second section is the VAST neighbor list itself.

#### Advanced neighbor search

Total neighbors: 177; 28 representatives from the Medium redundancy subset displayed.

<table>
<thead>
<tr>
<th>PDB</th>
<th>CD</th>
<th>All. Len</th>
<th>Score</th>
<th>EV</th>
<th>Rmsd</th>
<th>%Id</th>
<th>MMDB Date</th>
<th>LHM</th>
<th>GSP</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2JAE</td>
<td>A</td>
<td>225</td>
<td>15.3</td>
<td>1e-14.9</td>
<td>1.7</td>
<td>33.3</td>
<td>01/2007</td>
<td>2.2</td>
<td>0.8</td>
<td>Ground State Of Halorhodopsin T203v</td>
</tr>
<tr>
<td>1G3W</td>
<td>A</td>
<td>222</td>
<td>16.0</td>
<td>1e-16.8</td>
<td>0.8</td>
<td>99.5</td>
<td>03/2001</td>
<td>0.8</td>
<td>0.4</td>
<td>BacteriorhodopsinLIDP COMPLEX AT 1.55 A RESOLUTION</td>
</tr>
<tr>
<td>1E12</td>
<td>A</td>
<td>220</td>
<td>15.5</td>
<td>1e-15.3</td>
<td>1.5</td>
<td>34.1</td>
<td>03/2001</td>
<td>2.1</td>
<td>0.7</td>
<td>Halorhodopsin, A Light-Driven Chloride Pump</td>
</tr>
<tr>
<td>1H2S</td>
<td>A</td>
<td>216</td>
<td>15.3</td>
<td>1e-14.8</td>
<td>1.1</td>
<td>29.2</td>
<td>11/2002</td>
<td>1.5</td>
<td>0.5</td>
<td>Molecular Basis Of Transmembrane Signalling By Sensory Rhodopsin II-Transducer Complex</td>
</tr>
<tr>
<td>2F93</td>
<td>A</td>
<td>214</td>
<td>15.5</td>
<td>1e-15.2</td>
<td>1.0</td>
<td>29.4</td>
<td>05/2006</td>
<td>1.4</td>
<td>0.5</td>
<td>K Intermediate Structure Of Sensory Rhodopsin ITRANSCLUDER Complex In Combination With The Ground State Structure</td>
</tr>
<tr>
<td>1X1Q</td>
<td>A</td>
<td>208</td>
<td>11.4</td>
<td>1e-8.9</td>
<td>1.6</td>
<td>29.3</td>
<td>11/2004</td>
<td>2.5</td>
<td>0.8</td>
<td>Anabaena Sensory Rhodopsin</td>
</tr>
</tbody>
</table>

Click to: Check All, Uncheck All
Problem 2

In this problem, we will follow these steps:

A. Identify conserved domain(s) present in a protein.
B. Search for other proteins containing similar domain(s).
C. Explore a 3D modeling template for the query sequence.
D. Find distant sequence homologs that may not be identified by BLAST.

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. Click on the CDD link provided below,

CDD

paste the following protein sequence in the CD-Search query box and run the search.

```
>gi|2851597|sp|P25848|PHY1_CERPU Light-sensor Protein kinase
MSATKKTYSSSETTSAKSHSVRRAQTTADALEAVYREMSGDSG
QRQEGIAGFMAGVAYEIPCGCVYSENASEFELDLPOAVPSMGEE
AATQDISLNPTRHCRRSGKPLYAAEHIDINGIDFEAVKMDVNAV
LPGGDELLCDTVEEVRELTVDRVAHKHHEHHHEVVAEERN
KNRRVRLACDYSAPVKLOQDPDRQPSLALGSTRAPNGCHARYY
IQRGKRLWGLVVCQHTSRRTVPPFLRSVCERLMOQVFAMQLNHL
PIGNQSQTPNMDLVKCDGALEYGKRVWLLGTTPTECQIKEADY
HLLDGAVCMAAACKTKDFLFWFRSH2TATEVKKWGAJKHDPDE
EDVBMADAIHSOQLILRSGFRDAEELKTDKTHMAHLKLNLKLQGV</pre>
```

A. What are the domains present in this protein?
(Select the “Full Result” radio button to display all of the domains.)

-Suppose, we are interested in the serine/threonine protein kinase domain. Obtain more information about it by searching in NCBI's Bookshelf

B. Go back to the CD-Search results page. Obtain a list of proteins with similar domain architecture by clicking on the "Search for similar domains architectures" button. To display the records, click on the links to the subsets of sequences and from there on the “Look up Sequences in Entrez”. Change the display from “Summary” to “FASTA”.

C. Go back to the CD-Search results page. Generate a multiple sequence alignment for the top 10 sequences representative of the conserved domain hit by clicking on the graphic representation of the serine/threonine kinase domain from CDD (CDD[00180]). Use the "Row Display" list box pull down menu to specify "up to 5" sequences and reformat sequence alignment. 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.

To show only one top structure, click on the down arrow key. For better view of the backbone, remove the side chains globally (Style--Edit global style--Protein side chains). The query protein contains a serine/threonine protein kinases active-site signature (IIHRDLKSMNILV) where K is the ATP binding site. Identify these residues in the query protein and highlight the corresponding lysine residue in the first protein sequence.

Display the side chains of this residue (Use Style--Annotate--New--Edit Style. Change the protein backbone Rendering to Tubes, Color Scheme to User Selection and User Color to choose the color for the highlighted residue, for example yellow. Repeat these steps for the Protein Side chains row and click the Protein Side chains on. Click on the "Done" button. To zoom in, press z on the keyboard. Note the heterogen near the lysine residue.

D. To obtain the structural neighbors for the serine/threonine protein kinase protein, first click on the structure entry link 1JNK of the similar protein from the CD-Browser page. Then click on the structure link on the top right side, then on 1JNK, and finally on the chain graphic. Select one or more of the check boxes next to the structure neighbors and download the structures by clicking on the "View 3D Structure" button.