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)
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


Slides:

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

Pfam  SMART  COG  KOG  PRK

Curated CDs

- Structure template
- Structure neighbors
- Sequence
- Conserved domain
- Proteins with a similar domain architecture

VAST  CDD  CDART
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.

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 "Structure View" button.
The structure of the *Halobacterium salinarum* bacteriorhodopsin mutant 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, 1S52_B, on the CD-Browser page. Then click Links → Structure on the top right, then on 1S52 again in the Entrez Structure page, and finally on the chain A graphic. Select one or more of the check boxes next to the structure neighbors and view by clicking on the "View 3D Structure" button.

Screen images:
Many Integral Proteins Contain Multiple Transmembrane a Helices

Although Figure 3-34 depicts glycosinolase as a monomer with a single a helix spanning the bilayer, this protein is present in eukaryotic membranes as a dimer of two identical polypeptide chains. The two membrane-spanning a helices of glycosinolase are thought to form a coiled-coil structure (see Figure 3-9a) stabilized by specific interactions between the amino acid side chains at the interface of the two helices. It is now known that many other transmembrane proteins contain two or more membrane-spanning a helices. For instance, the bacterial photosynthetic reaction center (PRC) 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 a helices (see Figure 16-40).

A large and important class of transmembrane proteins is defined by the presence of seven transmembrane a helices. More than 150 such "seven-helices" membrane proteins have been identified. This class of integral protein is targeted by the bacterial photosynthetic reaction center (PRC) and is composed of 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 a helices (see Figure 16-40).

A large and important class of transmembrane proteins is defined by the presence of seven transmembrane a helices. More than 150 such "seven-helices" membrane proteins have been identified. This class of integral protein is targeted by the bacterial photosynthetic reaction center (PRC) and is composed of 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 a helices (see Figure 16-40).

Other seven-spanning membrane proteins include the opsin (eye proteins that absorb light), cell-surface receptors for many hormones, and receptors for odorant molecules. Amino acid sequence analysis of these proteins has shown that no amino acids are found in the same position in all of them, and only a few residues are conserved in even a substantial number of them. Nonetheless, each of these proteins contains seven stretches of hydrophobic amino acids long enough (>22 amino acids) to span the phospholipid bilayer. Though direct evidence is lacking, it is thought that all of these proteins adopt a similar conformation in the membrane similar to that of bacteriorhodopsin. This is one of several examples of how investigators can predict the orientation of proteins in a membrane from the amino acid sequence alone. 

Figure 3-34
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, 99]
Molecular components in the MMDB structure are listed below. The icons indicate macromolecular chains, 3D domains, protein classifications, and ligands. Please hold the mouse over each icon for more information on the component. You may also click the thumbnails below to view corresponding chains and domains in Cn3D.

VAST related structures for MMDB 26602, 1SS2 sequence 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 related structure search controls. The second section is the VAST related structure list itself.

Move the mouse over the red alignment footprints in the graphics below and click; you will obtain a structure-based sequence alignment.

Total related structures: 165; 26 representatives from the Medium redundancy subset displayed.

Click to:
- Check All
- Uncheck All

| Chain | 26602 | 1SS2 | Related MBP
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Chain B</td>
<td>Chain B</td>
<td>Chain B</td>
<td></td>
</tr>
<tr>
<td>1SBD A</td>
<td>1SBD A</td>
<td>1SBD A</td>
<td></td>
</tr>
<tr>
<td>1SS2 A</td>
<td>1SS2 A</td>
<td>1SS2 A</td>
<td></td>
</tr>
<tr>
<td>1SS2 B</td>
<td>1SS2 B</td>
<td>1SS2 B</td>
<td></td>
</tr>
<tr>
<td>1SS2 C</td>
<td>1SS2 C</td>
<td>1SS2 C</td>
<td></td>
</tr>
<tr>
<td>1SS2 E</td>
<td>1SS2 E</td>
<td>1SS2 E</td>
<td></td>
</tr>
<tr>
<td>1SS2 F</td>
<td>1SS2 F</td>
<td>1SS2 F</td>
<td></td>
</tr>
<tr>
<td>1SS2 G</td>
<td>1SS2 G</td>
<td>1SS2 G</td>
<td></td>
</tr>
<tr>
<td>1SS2 H</td>
<td>1SS2 H</td>
<td>1SS2 H</td>
<td></td>
</tr>
</tbody>
</table>
Total related structures: 165; 26 representatives from the Medium redundancy subset displayed.

<table>
<thead>
<tr>
<th>PDB Code</th>
<th>All Len</th>
<th>Score</th>
<th>E-Val</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>1VGO A</td>
<td>227</td>
<td>15.7</td>
<td>10e-16.3</td>
<td>1.1</td>
<td>56.4</td>
<td>10/2005</td>
<td>0.3</td>
<td>0.5</td>
<td>Crystal Structure Of Archaeorhodopsin-2y</td>
</tr>
<tr>
<td>1C3W A</td>
<td>222</td>
<td>16.0</td>
<td>10e-17.3</td>
<td>0.8</td>
<td>99.5</td>
<td>03/2001</td>
<td>0.8</td>
<td>0.4</td>
<td>Bacteriorhodopsin</td>
</tr>
<tr>
<td>1E12 A</td>
<td>220</td>
<td>15.2</td>
<td>10e-15.1</td>
<td>1.5</td>
<td>34.1</td>
<td>03/2001</td>
<td>2.2</td>
<td>0.7</td>
<td>Halorhodopsin, A Light-Driven Chloride Pump</td>
</tr>
<tr>
<td>1H2S A</td>
<td>216</td>
<td>15.3</td>
<td>10e-13.2</td>
<td>1.1</td>
<td>28.7</td>
<td>11/2002</td>
<td>1.5</td>
<td>0.5</td>
<td>Molecular Basis Of Transmembrane Signalling By Sensory Rhodopsin Il-Transducer Complex</td>
</tr>
<tr>
<td>1X1O A</td>
<td>208</td>
<td>11.4</td>
<td>10e-9.8</td>
<td>1.6</td>
<td>29.3</td>
<td>11/2004</td>
<td>2.5</td>
<td>0.8</td>
<td>Arachadonate Sensory Rhodopsin</td>
</tr>
<tr>
<td>1JFP A</td>
<td>170</td>
<td>6.7</td>
<td>0.0351</td>
<td>3.9</td>
<td>8.8</td>
<td>11/2001</td>
<td>10.8</td>
<td>2.4</td>
<td>Structure Of Bovine Rhodopsin (Dark Adapted)</td>
</tr>
<tr>
<td>1OW8 C</td>
<td>135</td>
<td>4.8</td>
<td>0.0353</td>
<td>4.4</td>
<td>7.4</td>
<td>11/2003</td>
<td>21.0</td>
<td>3.3</td>
<td>Paxillin L02 Motif Bound To The Focal Adhesion Targeting (Fak) Domain Of The Focal Adhesion Kinase</td>
</tr>
<tr>
<td>2P7V A</td>
<td>120</td>
<td>5.7</td>
<td>0.0372</td>
<td>2.7</td>
<td>8.3</td>
<td>11/2007</td>
<td>NA</td>
<td>2.3</td>
<td>Crystal Structure Of The Escherichia Coli Regulator Of Sigma 70, Rsd, In Complex With</td>
</tr>
</tbody>
</table>
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.

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. Click on the “Full Report” radio button. 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 "Aligned Rows" list box pull down menu 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 "Structure View" 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 kinase 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 conserved 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. Type 1JNK_A in the search box and click on the Go button. 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 Alignment" button.