

Structural Analysis Quick Start

An NCBI Mini-Course

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's 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

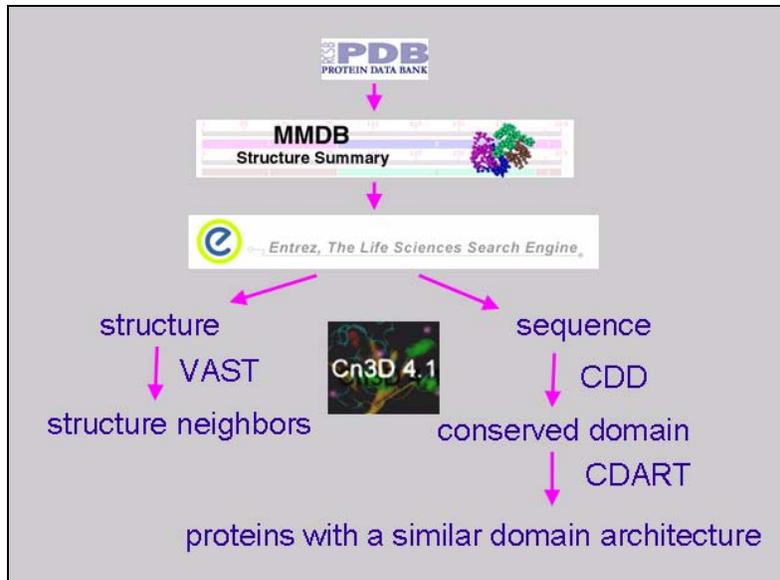
- Identify a conserved domain present in the query protein using **CDD**
- Search for other proteins containing similar domain(s) using **CDART**
- Explore a 3D modeling template for the query sequence using **CDD**
- Find similar structures using **VAST**
- Visualize and annotate the 3D protein structures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

The following handout includes the screen shots of the exercise demonstrated in the mini-course.

URL: <http://www.ncbi.nlm.nih.gov/Class/minicourses/quickstructure.html>

Instructor:

Dr. Medha Bhagwat, NCBI
bhagwat@ncbi.nlm.nih.gov



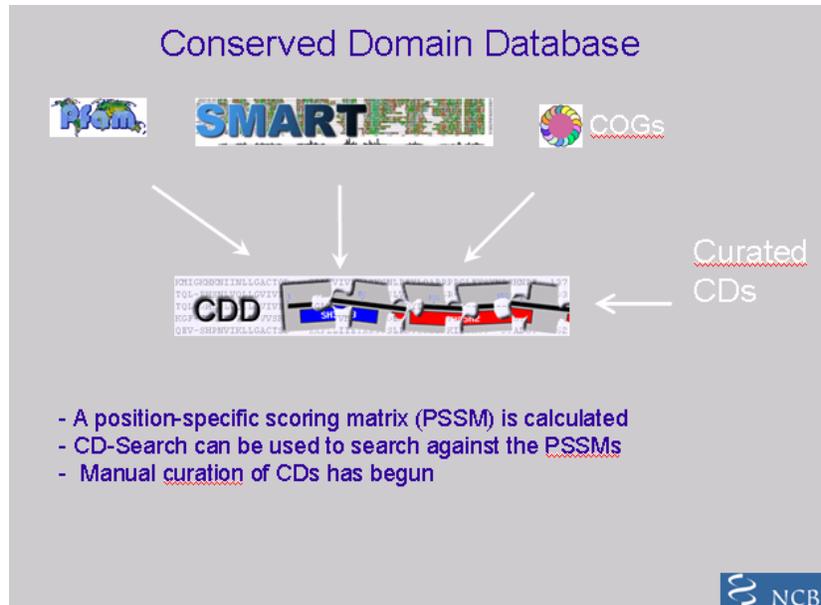
CDD

<http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml>

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

NCBI



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.

```
MDPIALTAAVGADLLGDGRPETLWLGIGTLLMLIGTFYFIVKGWG
SMFFGIGLTEVQVGSEMLDIYARYADWLFTPLLLLDLALLAKV
HTPLARYTWWLFSTICMIVVLYFLATSLRAAAKERGPEVASTFN
VGLGIETLLFMVLDVTAKVGFGFILLRSRAILGDTEAPEPSAGAE
```

A. What is the domain present in this protein?

Obtain more information about the domain by searching in

[NCBI's Bookshelf](#)

B. Obtain a list of proteins with a similar domain architecture by clicking on the "Show" button.

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 listbox to specify "up to 5" sequences and click on the "All Atoms" radio button. 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 1E12_A of the similar protein from the CD-Browser page. Then click on the structure link on the top right side, then on 1E12, and finally on the chain A graphic.

NCBI Conserved Domain Search

RPS-BLAST 2.2.10 [Oct-19-2004]
 Query= local sequence: lcl|tmpseq_0
 gi|114807|sp|P19585|BAC1_HALS1 ARCHAERHODOPSIN 1 PRECURSOR (AR 1)
 (260 letters)

Database: cdd.v2.03

Click on boxes for multiple alignments

show Domain Relatives

PSSMs producing significant alignments:

	Score	E
	(bits)	value
gnl CDD 1587 pfam01036, Bac_rhodopsin, Bacteriorhodopsin..	185	5e-48
gnl CDD 14608 COG5524, COG5524, Bacteriorhodopsin [General function predicti...	134	1e-32

[gnl|CDD|1587](#) pfam01036, Bac_rhodopsin, Bacteriorhodopsin..

CD-Length = 232 residues, 89.2% aligned
 Score = 185 bits (470), Expect = 5e-48

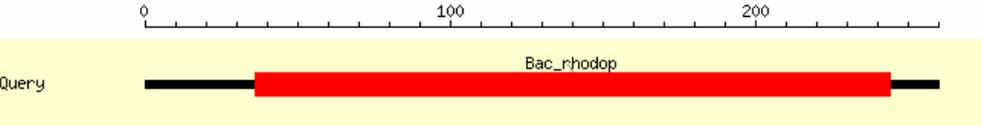
Query: 37 FYFIVKGVGVDKREYYSITILVPGIASAAYLSMFFGIGLTEVQVGSMLDIYYARYA 96
 Sbjct: 17 LLFIYMGRLBDEPARKIYAATLILPGIAIVAYLGMALGIGVTVEMPAEH--IYWAYRL 74

Query: 97 DWLFTTPLLDDLALLAKVDRVSIQTLVGDALMIVTGLVGALS-HTPLARYTWLWLFSTI 155
 Sbjct: 75 DWLFTTPLLFLGLLAGARRTIFTLVVADALMIVTGLAALTTSSGLLRWVLFGISTA 134

 **CDART: Conserved Domain Architecture Retrieval Tool**

[New Query](#) [Overview](#) [PubMed](#) [Nucleotide](#) [Protein](#) [Structure](#) [Taxonomy](#) [Help?](#)

[About CDART](#)



Query Bac_rhodop

[Similar domain architectures](#)

[251 Sequences](#)
cellular organisms
Blue-light absorbi

 **NCBI Conserved Domain Search**

[New Search](#) [PubMed](#) [Nucleotide](#) [Protein](#) [Structure](#) [CDD](#) [Taxonomy](#)

RPS-BLAST 2.2.10 [Oct-19-2004]
 Query= local sequence: lcl|tmpseq_0
 gi|114807|sp|P19585|BAC1_HALS1 ARCHAERHODOPSIN 1 PRECURSOR (AR 1)
 (260 letters)

Database: cdd.v2.03

Click on boxes for multiple alignments



Domain Relatives

PSSMs producing significant alignments:

	Score	E (bits) value
gnl CDD 1587 pfam01036, Bac_rhodopsin, Bacteriorhodopsin..	185	5e-48
gnl CDD 14608 COG5524, COG5524, Bacteriorhodopsin [General function predicti...	134	1e-32

[gnl|CDD|1587](#) pfam01036, Bac_rhodopsin, Bacteriorhodopsin..

CD-Length = 232 residues, 89.2% aligned
 Score = 185 bits (470), Expect = 5e-48

Query: 37 FYEIVKGVTDKEAREYYSTITLVPGIASAAYLSMFFGIGLTVQVGSEMLDIYYARYA 96
 Sbjct: 17 LLEFIYMGRLSDPEARKIYAATILIPGIAIVAYLGMALGIGVTTVEMPAEH--IYWARYL 74

Query: 97 DWLFTPLLLLDLALLAKVDRVSIPTLVGVDALMIVTGLVGLS-HTPLARYTWWLFSI 155
 Sbjct: 75 DWLFTPLLLFLGLLAGADRRITPTLVVADALMIVTGLAAALTSSGLLRWVLFGI STA 134

NCBI Conserved Domains

Entrez CDD Structure Protein Help

pfam01036.11 Bac_rhodopsin, with user query added

Links: Bacteriorhodopsin.

Source: Pfam
 Taxonomy: Halobacteriaceae
 PubMed: 1 link
 Protein: pfam01036 related architectures representatives

Related CD: 1 link
 This domain model appears to be related to other CDs:
 pfam01036 --- C005524
 [mouse over icons to display CD accession/name and number of common hits]

Statistics:
 PSSM-Id: 1587
 Aligned: 14 rows
 PSSM: 232 columns
 Status: Alignment from source
 Created: 13-Dec-2003
 Updated: 13-Dec-2003

Structure:
 Show Structure
 Program: Cn3D
 Drawing: All Atoms
 Aligned Rows: up to 5
 (download Cn3D)

Show Alignment Format: Compact Hypertext Row Display: up to 5 Color Bits: 2.0 bits
 Type Selection: the most similar members

1E12_A 8 . [16] . L V F V Y M . [1] . R T I R P G R P R L I W G A T L M I P L V S I S S Y L G L L S G L T V G M I E M P . [11] . S Q W G R Y L T W A L S T P M I 98

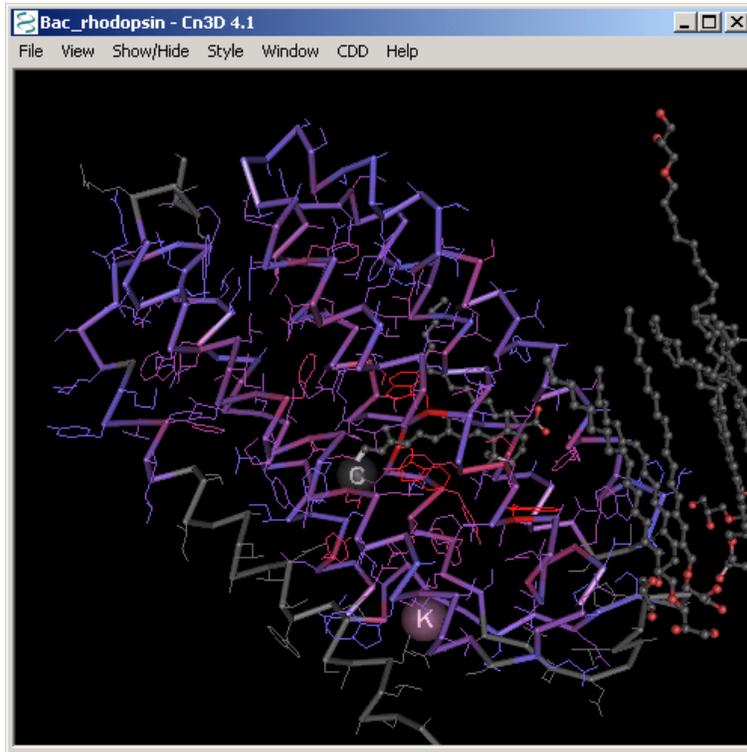
CDD Descriptive Items

Name: Bac_rhodopsin

Bacteriorhodopsin.

Structure summary:
 PDB 1E12 (MMDB 13348)
 1E12_A: gi 8569313 ([Halobacterium salinarum] Chain A, Halorhodopsin, A Light-Driven Chloride Pump)

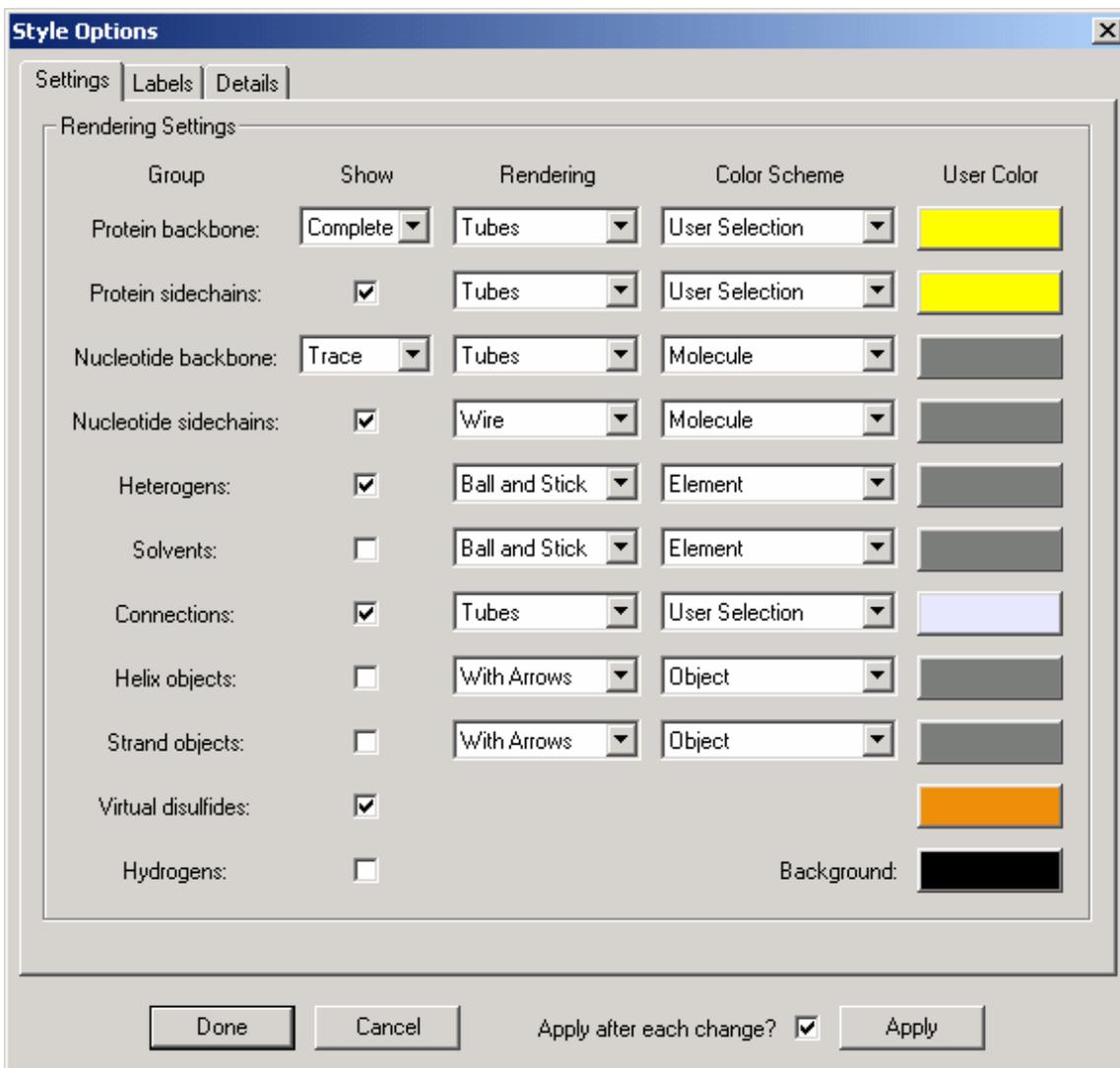
Show Annotations Panel Show References Panel Dismiss

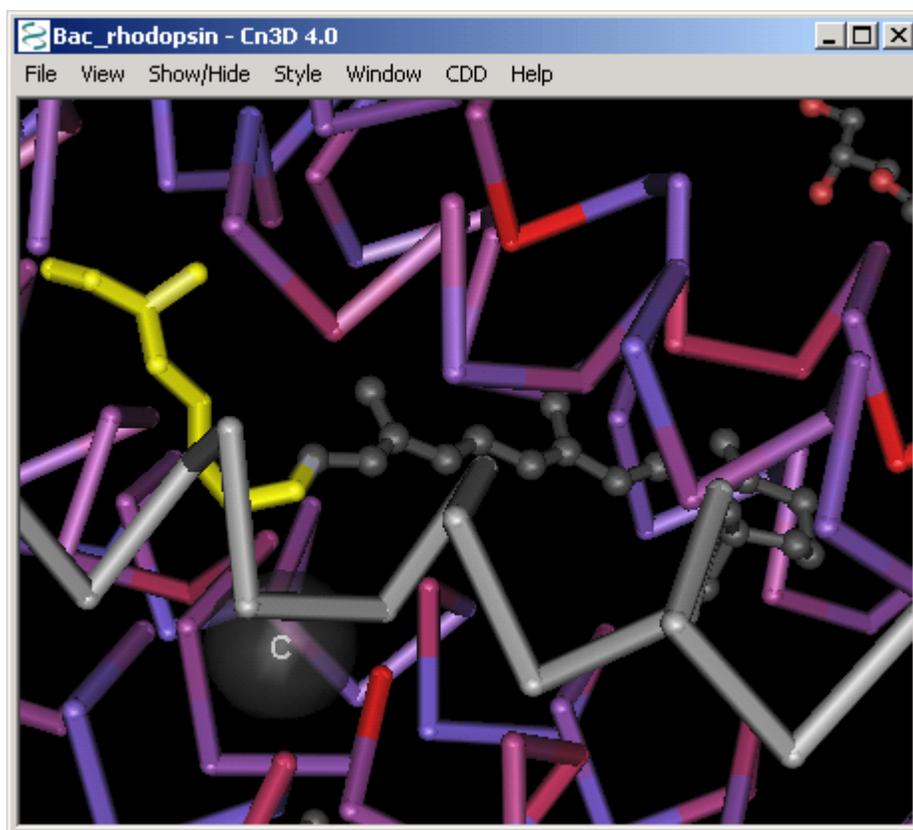


Bac_rhodopsin - Sequence/Alignment Viewer

View Edit Mouse Mode Unaligned Justification Imports

<i>IB12_A</i>	LIWGATLMIPLVSISSYLGLLSGLTVGMIEMPAGH a l a g e m v r S Q W G R Y L T W A L S T P M I L L A L G L L A D V D L G S L F T V I A A D I G
query	EYYSITILVPGIASAAYLSMFFGI GLTEVQVGSEMI ~~~~~ d I Y Y A R Y A D W L F T T P L L L L D L A L L A K V D R V S I G T L V G V D A L
<i>IUAZ_A</i>	EYYSITILVPGIASAAYLSMFFGI GLTEVQVGSEMI ~~~~~ d I Y Y A R Y A D W L F T T P L L L L D L A L L A K V D R V S I G T L V G V D A L
<i>IMOK_A</i>	KFYA I T T L V P A I A F T M Y L S M L L G Y G L T M V P F G G E Q n ~~~~~ p I Y W A R Y A D W L F T T P L L L L D L A L L V D A D Q G T I L A L V G A D G I
gi 2499387	KFYIATIMIAAIAFVNLYLSMALGFGVTTI ELGGEE r ~~~~~ a I Y W A R Y T D W L F T T P L L L Y D L A L L A G A D R N T I Y S L V G L D V L





Display GenPept Send all to file

Range: from begin to end Features: SNP CDD MGC HPRD STS

1: [1E12A](#). Reports Chain A, Halorhod...[gi:8569313] BLink, Conserved Domains, Links

LOCUS 1E12_A 253 aa linear
 DEFINITION Chain A, Halorhodopsin, A Light-Driven Chloride Pump
 ACCESSION 1E12_A
 VERSION 1E12_A GI:8569313
 DBSOURCE pdb: molecule 1E12, chain 65, release Apr 6, 2000; deposition: Apr 6, 2000;

Links
 ▶ Related Structure
 ▶ Related Sequences
 ▶ 3D Domains
 ▶ Domain Relatives
 ▶ PubMed
 ▶ Structure
 ▶ Taxonomy

NCBI Entrez Structure

Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books

Search Structure for [] Go Clear

Limits Preview/Index History Clipboard Details

Display Summary Show: 20 Send to Text

1: [1E12](#) Halorhodopsin, A Light-Driven Chloride Pump [mmdbId:13348] MMDB, Links

NCBI MMDB Structure Summary

PubMed BLAST Structure Taxonomy OMIM Help? Cn3d

Description: Halorhodopsin, A Light-Driven Chloride Pump.
Deposition: M.Kolbe, H.Besir, L.-O.Essen & D.Oesterhelt, 6-Apr-00
Taxonomy: [Halobacterium salinarum](#)
Reference: [PubMed](#) MMDB: [13348](#) PDB: [1E12](#)

View 3D Structure of Best Model with Cn3D Display **NEW** [Get Cn3D 4.0!](#)

Protein Chain A
 CDs Bac_rhodopsin





[PubMed](#) [BLAST](#) [Structure](#) [Taxonomy](#) [OMIM](#) [Help?](#) [Cn3D](#)

Query: MMDB [13348](#), 1E12 chain A
Description: Halorhodopsin, A Light-Driven Chloride Pump

of with NEW [Get Cn3D 4.1!](#)

using for VAST neighbors

subset sorted by in

MMDB or PDB ids: or 3D-Domain ids:

104 neighbors found. 13 representatives from the [Medium redundancy](#) subset displayed.

Query	Chain	CDs	Residues
1E12 A 3d Dom.	Chain A	Bac_rhodopsin	1-253
<input type="checkbox"/> 1H2S A			219
<input type="checkbox"/> 1C3M A			217
<input type="checkbox"/> 1XI0 A			209
<input type="checkbox"/> 1ST6 A 8			115
<input type="checkbox"/> 1ST6 A 7			98

Query: MMDB [13348](#), 1E12 chain A
Description: Halorhodopsin, A Light-Driven Chloride Pump

of with NEW [Get Cn3D 4.1!](#)

using for VAST neighbors

subset sorted by page in

MMDB or PDB ids: or 3D-Domain ids:

-
-
-




[PubMed](#) [BLAST](#) [Structure](#) [Taxonomy](#) [OMIM](#) [Help?](#) [Cn3D](#)

Query: MMDB [13348](#), 1E12 chain A
Description: Halorhodopsin, A Light-Driven Chloride Pump

of with [NEW Get Cn3D 4.1!](#)

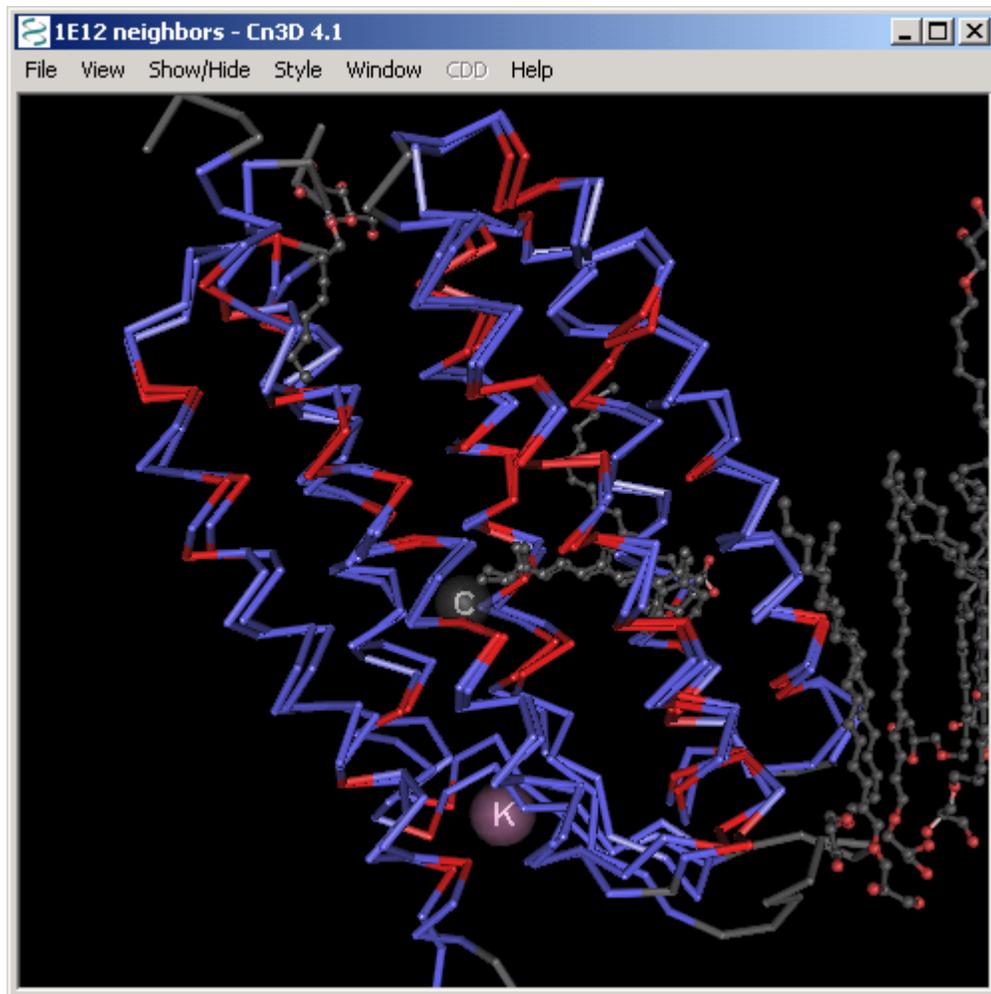
using for VAST neighbors

subset sorted by in

MMDB or PDB ids: or 3D-Domain ids:

104 neighbors found. 13 representatives from the [Medium redundancy](#) subset displayed.

	PDB C D	Ali. Len.	SCORE	P-VAL	RMSD	%Id	MMDB Date	Description
<input type="checkbox"/>	1H2S A	219	15.6	10e-15.9	1.4	27.4	11/2002	Molecular Basis Of Transmembrane Signalling By Sensory Rhodopsin Li-Transducer Complex
<input type="checkbox"/>	1C3W A	217	15.2	10e-15.1	1.6	33.6	03/2001	BacteriorhodopsinLIPID COMPLEX AT 1.55 A RESOLUTION
<input type="checkbox"/>	1XIO A	209	11.9	10e-11.1	1.7	26.3	11/2004	Anabaena Sensory Rhodopsin
<input type="checkbox"/>	1ST6 A B	115	6.1	0.0086	2.9	7.0	08/2004	Crystal Structure Of A Cytoskeletal Protein



1E12 neighbors - Sequence/Alignment Viewer

View Edit Mouse Mode Unaligned Justification Imports

1E12_A	a v r e N A L L S S L W V N V A L A G I A I L V F V Y M G R T I R P G r P R L I W G A T L M I P L V S I S S Y L G L L S G L T V G M I E m p a g h a l a g e M V R S Q W
1H25_A	~ ~ ~ ~ M V G L T T L F W L G A I G M L V G T L A F A W A G R D A G S G ~ E R R Y Y V T L V G I S G I A A V A Y V V M A L G V G W V P V A ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ E R T V F A

Problem 2

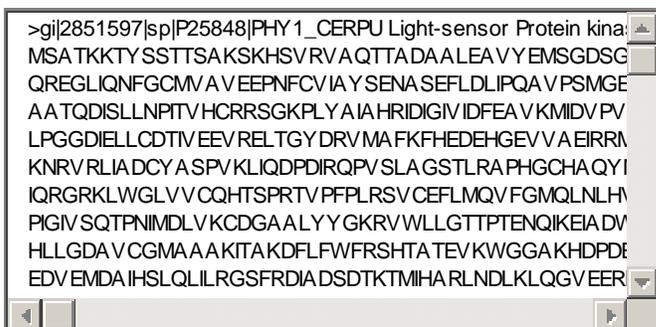
In this problem, we will follow these steps:

- Identify conserved domain(s) present in a protein.
- Search for other proteins containing similar domain(s).
- Explore a 3D modeling template for the query sequence.
- 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.



```
>gj|2851597|sp|P25848|PHY1_CERPU Light-sensor Protein kina:
MSATKKTY SSTTSA KSKHSV RVA QTTA DAALEA VY EMSGDSG
QREGLIQNF GCMV A V EEPNFCV IA Y SENA SEFLDLIPQA V PSMGE
AA TQDISLLN PTV HCRRSGK PLY A IA HRIDIGIV IDFEA V KMIDV PV
LPGGDIELL CDTIV EEV RELTGY DRV MAFKFHEDEHGEV V A EIRRN
KNRV RLIA DCY A SPVKLIQDPDIRQPV SLA GSTLRA PHGCHA QYI
IQRGRKLWGLV V CQHTSPRTV PFPLRSV CEFLMQV FGMQLNLH
PIGIV SQTPNIMDLV KCDGAA LYY GKRV WLLGTTPTENQIKEIA DV
HLLGDA VCGMA AAKITAKDFLFWFRSHTA TEV KWGGAKHDPDE
EDVEMDA IHSLQLILRGSFRDIA DSDTKTMIHARLNDLKLQGV EER
```

- What are the domains present in this protein?

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

[NCBI's Bookshelf](#)

- Obtain a list of proteins with a similar domain architecture by clicking on the "Show" button.

- 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|17776). Use the list box to specify "up to 5" sequences and click on the "All Atoms" radio button. 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.