Bioinformatics Lunch Seminar (Summer 2014)



- Every other Friday at 1 pm. 20-30 minutes plus discussion
- Informal, ask questions anytime, start discussions
- Content will be based on feedback
- Targeted at broad audience of various levels of backgrounds and education
- Emphasis on Genomics Center

Contact: Raymond Hovey Genomics Center - SFS <u>rhovey@uwm.edu</u> 414-382-1774 http://www.greatlakesgenomics.uwm.edu/

Proteins with similar structures often have similar functions.

Protein structure is divided into these 4 levels:

1. Primary Structure- Polypeptide backbone

2. Secondary Structure- Local Hydrogen bonds along the backbone

3. Tertiary structure- Long distance bonding involving the AA side chains

4. Quaternary structure- Protein-Protein interactions leading to formation of dimers, tetramers, etc.

Motifs and **domains** are combinations of secondary structures. Motifs only consist out of few secondary structures. They may but need not have a function. A domain is more complex. It is usually defined as a modular functional unit folding independently.

Chapter 4.2: Three dimensional protein structure

Primary amino acid sequence determines the 3D structure of proteins

Secondary structure elements



 $\alpha\text{-helix}$



 β -sheet

Major structures: $H = \alpha$ -helix E = β -sheet (extended strands)

Minor structures: G = 3-helix I = 5-helix T = H-bonded turn S = bend

Crambin (1CRN)

TTCCPSIVARSNFNVCRLPGTPEAICATYTGCIIIPGATCPGDYAN EE SSHHHHHHHHHHHTT HHHHHHHS EE SSS TTS





Tertiary structure (folding of helices and β -sheets and loops)

Quaternary structure (two or more subunits)

Tertiary structure is formed by long distance interactions of the amino acid side chains. There are several types of bonds:

- A) lonic bonds
- B) Hydrogen bonds
- C) Hydrophobic interactions
- D) Disulfide bonds -- weak covalent bond between 2 Cysteins (R-Cys-S-S-Cys-R)



Hydrogen Bonds in Tertiary Structure

Hydrogen bonds in tertiary structure involve polar non-charged amino acid side chains: Alcohols -- Ser Thr Tyr Amides --- Asn Gln

Ionic bonds in Tertiary Structure

lonic bonds formed between polar, charged AA side chains: Negative -- Glu Asp Positive -- Lys Arg His

Hydrophobic Interactions in Tertiary Structure

Hydrophobic interactions involve the side chains of Non-Polar amino acids: Hydrocarbon -- Ala Val Leu lle Pro Met Aromatics ----- Phe Trp (Tyr)

Bonds in quaternary structure (between subunits): Commonly found:

- Ionic Bonds
- Hydrogen Bonds
- Less Commonly found:
- Hydrophobic interfaces
- Interchain Disulfides

- Protein Structure is mostly stabilized by weak, non-covalent bonds. The energy difference between the denatured and native structure of proteins is small (important—many proteins change conformation during function; but also make them easy to unfold or even denature)

- This is one reason why the prediction of a protein structure is so difficult (no very low energy status)
- There are some structures in the secondary structure that can be predicted with certain probability

-But how the entire structure looks like is unpredictable-possibilities of interaction of amino acids are infinite

- Levinthal's paradox: a 100 aa protein has 3²⁰⁰ possible backbone configurations - many orders of magnitude beyond the capacity of the fastest computers

- Strategy -> explore the 3D structure of protein classes/ domains by X-ray defraction structure analysis and predict structure of other proteins by modeling

How to find the 3D structure of a protein of interest



http://www.ncbi.nlm.nih.gov/Structure/MMDB/docs/mmdb_search.html

Because 3D structure is strongly conserved in molecular evolution, one can often make valid predictions concerning structure and function by examining the structures of related proteins.

To find a structure for a protein or related proteins, use the following procedure:

Find the protein of interest using an accession number or keyword search in Entrez's structure database. Use blastp if the sequence is new and not in Entrez.



Click to open in Cn3D software

S NCBI



Download Cn3D 4.1 for PC, Mac and Unix

help new users get started!

Cn3D 4.1 Homepage

PubMed

Cn3D is a helper application for your web browser that allows you 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.

Below is a relatively simple sample of what Cn3D can do. There

are many more examples in the Tutorial, along with instructions to

Cn3D Tutorial

Cn3D feature highlights

Cn3D FAQ

Frequently Asked Questions

Cn3D Install

Installation and Configuration

MMDB

NCBI's structure



To visualize the 3D structure we need a helper program = Cn3D. Download it if it is not installed already

_

Cn3D is NCBI's 3D structure viewer. It allows you to interactively view 3D structures, sequences, and sequence alignments. Cn3D is available for Windows, MacOS, (and Unix until V4.1).

http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml



8 1CLF - Cn3D 4.1

Eile View Show/Hide Style Window CDD Help



_ 🗆 🛛

Show rotation by mouse click

ICLF - Sequence/Alignment Viewer

View Edit Mouse Mode Unaligned Justification Imports

ICLF_A aykiadscvscgacasecpvnaisqgdsifvidadtcide

Example: Ferredoxin of Clostridium Pasteurianum with 2 oxidized [4Fe-4S] (MMDB-ID 55778)

Structure viewer

Sequence and alignment viewer

ρ

Many opportunities to change 3D view of the protein

Tutorial at: http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3dtut.shtml





The style panel contains detailed controls for all drawing styles, colors, and labels





Enlarge by zoom in (view menu)

Righlight single amino acids (double click; one of the cysteine residues That coordinate the FeS cluster)

See amino acid in the sequence alignment viewer

1CLF - Sequence/Alignment Viewer

View Edit Mouse Mode Unaligned Justification Imports

ICLF aykiadscvscgacasecpvnaisqgdsifvidadtcidcgncanv<mark>c</mark>p

If you are working with an unknown protein do the following:

Standard protein-protein BLAST [blastp]". Then select "pdb" in the "Database" menu (to search against known structures),

Enter	Query	Sequence
-------	-------	----------

Enter accession number, gi, or FASTA sequen to a construct of the sequences (nr) Ouery subrange (construction of the sequences (nr) MAPSSARPLLQTLTCICLATLCLALFLPCLRLIALGGSFYT, VACLINILSCINLTMCRRSC From AYTLIWALWRVCLDGWRLIPPLALPAGTCLWVFCPUTCGRM DSQSHVFINSTVLGCMCGCL From AYTLIWALWRVCLDGWRLIPPLALPAGTCLWVFCPUTCGRM DSQSHVFINSTVLGCMCGCL From DLPAPGENSRGRFFSFKATPIKICNRLFFCTPHRDVVALDA TCKEVWHYRPGCBFGANIYQ To DIFGATDCTHRIISTDSGSPPTLFALDALTGQLCHSFCHIC TIDLRDCMCTIPPCFHFITSP() To (*) Im Im Or, upload file Browse Im Job Title Enter a descriptive title for your BLAST search (consection of the sequences (nr) Im Protein Data Bank protein s(pdb) Im Im Im Organism Non-redundant protein sequences (nr) Suggested Suggested Only 20 top taxa will be shown. (path 20 top taxa
MAPSSARPLLQTLTGICLAILGLALFLPGLRLIALGGSFYY VAGLLMILSGIMLTHGRRSG AYTLIWALWRVGLDGWRLIPPLALPAGIGLWVFGWIGGRMDS9SHVRINSTVLGCHGGL From CGWYITGQRYQQFNPFPGSTGASLSGSDQMAANDWQFHGGT AGD RFAAPTQINAQMAINLK DLPPGENSRGRFFSFKATPIKIGNRLFFCTPHEDVVALDA TGKEWWHYRPGGEFGANIYQ To To Or, upload file Browse Job Title Enter a descriptive title for you'r BLAST search @ Choose Search Set Organism Optional Non-redundant protein sequences (nr) Reference proteins (refset_protein) suggested Only 20 top taxa will be shown. @
Choose Search Set Database Organism Optional Entrez Query Protein Data Bank proteins(pdb) Image: Choose Search Set Image: Choose Search Set Protein Data Bank proteins(pdb) Image: Choose Search Set Image: Choose Search Set Image: Choose Search Set
Choose Search Set Database Protein Data Bank proteins(pdb) ▼ Organism Optional Non-redundant protein sequences (nr) Reference proteins (refseq_protein) Swissprot protein sequences(swissprot) Patented protein sequences(pat) suggested Entrez Query Protein Data Bank proteins(ndb) Only 20 top taxa will be shown. ()
Database Protein Data Bank proteins(pdb) Image: Comparison of the second s
Organism Optional Protein Data Bank proteins(pdb) Image: Constraint of the second
Organism Non-redundant protein sequences (nr) Optional Reference proteins (refseq_protein) Swissprot protein sequences(swissprot) Suggested Only 20 top taxa will be shown. (Patented protein Data Bank proteins(ndb) Only 20 top taxa will be shown. (Patented Protein Data Bank proteins(ndb)
Optional Reference proteins (refseq_protein) suggested Swissprot protein sequences(swissprot) Only 20 top taxa will be shown. (a) Entrez Query Protein Data Bank proteins(ndb)
Entrez Query Protein Data Bank proteins(ndb)
Entrez Query Protein Data Bank proteins(ndb)
Ortical
Environmental samples(env_nr)
Program Selection
Algorithm
C blastp (protein-protein BLAST)
PSI-BLAST (Position-Specific Iterated BLAST)
O PHI-BLAST (Pattern Hit Initiated BLAST)

S NCBI	LSYPOTDVFI LSYPDTDVI F Struc LSYSDTOCI	TURE NF	SSFENV DSLENV ASFEN ESLIN DSLENV		KTP FILVER VP VP
PubMed Entre	z BLAST	OMIM	Books	TaxBrowser	Entrez Structure
Search Entrez Stru	cture	🕶 for			Go
Cn3D 4.1 Homepage					
Cn3D Tutorial This chapter illustrates some more advanced alignment features in Cn3D: importing sequences and structures, and visualizing sequence conservation. Cn3D Tutorial Importing sequences and structures				iment features in <u>sualizing</u>	

Live Demonstration:

You have a gene encoding a hypothetical protein that shows homologies to a certain protein family. Check this assumption by comparison the predicted 3D-structure

Example is a gene from *Gluconobacter oxydans* that encodes a putative PQQ dependent dehydrogenase (GOX1441)

Distribution of 17 Blast Hits on the Query Sequence



Distance tree of results NEW Related Structures

Sequences p	roducing	significant alignments:	Score (Bits)	E Value	
pdb 1KB0 A	Chain A,	Crystal Structure Of Quinohemoprotein Alc	102	3e-22 S	
pdb/1YIQ/A	Chain A,	Molecular Cloning And Structural Analysis	<u>95.9</u>	4e-20 🔼	
pdb 1KV9 A	Chain A,	Structure At 1.9 A Resolution Of A Quinoh	88.2	9e-18 🔼	
pdb 4AAH A	Chain A,	Methanol Dehydrogenase From Methylophilus	77.0	2e-14 🔼	
pdb 1G72 A	Chain A,	Catalytic Mechanism Of Quinoprotein Metha	77.0	2e-14 🗧	
pdb 2D0V A	Chain A,	Crystal Structure Of Methanol Dehydrogena	74.3	1e-13 🔼	
pdb lLRW A	Chain A,	Crystal Structure Of Methanol Dehydrogena	<u>69.7</u>	3e-12 🔼	
male LIFE CLA	Choin a	Cruatel Structure Of The Ouiperrotein Fth	67 0	10 11 S	

Select the one at the top with the best score and E-value-- leads to the alignment, from which the "Structure" link on the right leads ultimately to the 3d file.

Download - Gentein Graphica	Viext &
Chain A, Crystal Structure Of Quinohemoprotein Alcohol Dehydrogenase From Comamonas Testosteroni	
Sequence ID: <u>pdb11KH0IA</u> Length: 677 Number of Matches: 1	Related Information
tango 1: 28 to 354 GanPast Graphics Third Hatch & Previous Hatch	Structure 30 structure date
Score Espect Method Identifies Punitives Gaps	SURGER - 30 BURGER & Sub
110 bits(274) 1e-24 Compositional matrix adjust. 168/653(26%) 249/653(38%) 154/653(23%)	
Query 174 DWOFHOUTPAGDRFAAPTQINACMAHULWYAWYYESGLPRPGKNSRGREFSFWAIFIWI 213	
sejet 28 ซึ่งครามพบกนิยรพันธ์ตามผู้ไม่ผิมผิดหลังไม่ไม่มีหรักนี้นี้รัรได้ดีหรีนี้โร้งหวั 76	
Query 234 GMRLFFCTPHRDVVALGATIGKEVNHYRPGGEFGAMIYQACRGV9YADIPGATDCTHRII 293	
++ V A+D TG +W Y P Q B + G D +B + SEjet 17 DOINYVSASHSVVHAIDTRIGNRIWIYDPQUERSTOFKOCCDVVNROV 124	
Query 294 STDSGSPFILFALDAETO-QLCRSFORDOWIDLADOWSTIPOGFEFITSPOWFL 346	
* G L ALDA IG ++ H +G+ G II P V 5bjct 125 ALMKSKVYVGAMDGBLIALDAAIGKEVWHQBIFEGQKGSLIIIGAFKVF 173	
Ouery 347 HURIVISGWVYDDOSVDEPSGVIRAFDRITSOLANANIMGRVPRUSELGFNEIFIRGT 404	-
+++ + + + G I A+DA IG+ W W VF + P F +E+ R Shict 174 MEMVIIDENDEALYOVADVIIAVDAEIGERKMENFSVPDD-PSKFFEDESMERAA 226	
Ouery 405 PNU	
P+G W I DA LM +YV G +P + V+ + D Y S+ Sbjot 227 RIMDESGHAMDENTFOALLNIMYVGTGNSSEMENKVBSENGGUNLYLASIV 286	
Query 449 ALDLTIGEEEMHFQAVHHDLWDFDLFVGPSLVDLPDASGILTPALVQTTMQSELFVLDRB 508	
ALD TG+ +MH+Q D WD+ L D+ A G ++ K G FVLDR Bbjct 287 ALDPDTGKYEMHYQEIFGODWDYTSTQEMILADIKIA-GHPRAVILHAPRDGFFFVLORI 345	
Query 509 TOOPFYRVQEKFVP60DIPHERYSFTOPYSVTMPRIERPDITEDDIWGATFFDOLACRIA 568	
Sbjot 346 NGK	
Query 569 KHIMBERGLYIPPERQGTIGFEAFDGVALWYGGTIDFINGVMYIMITFIFFIMIDVPHIB 628	
+ B + P G +W+ + +P G++I+ + F + L+ Sbjot 373 IAAARDESKFQDAVPGPYGAHNKHEMSENPQTELVYLFAQKVFVNLMDD 421	
Query 629 AvQesLFKFWBSKNSTFFFVFTHNEGESLFVAAVIKPWLSLFSA-BCLAPFWSKNSAID 687	
Sbjet 422HOME-FNGAGPGHPGSSTGWNTAKFNNAPPKSKPFSBLLAND 443	
Query 608 LVHERVINERALGIIHONGPINHLRHEVGLETSIYONGGSVTIPHGLVENGALADQSTHI 747	
Sbjot 464 PVAQKAAMFVENVSPWNGGILITAGNVVYQGT-ADGRLVA 502	
Query 748 LDGHDGHTLFRTELDAGGNAT9HTYHGEDGRQYVVLAVGGHGGLRTRNSDE 798	
G L G A P-TYM DERCYV 4AVG G GL R ++ Shver 503 YEAAFDERLYKAPTGYVAAABTTVM-VDGRUVYVAAABAFER 554	

In our case the sequence of the conserved hypothetical protein showed homologies the a PQQdependent ethanol dehydrogenase with known 3D structure from Comamonas testosteroni

S N	CB1	Related Structures	
Struc	tures rela	ted to [si]58002297[gb]AAW61191] detydrigerau 3 [Gleenebadar toydan t216])	Bachgelener BLAST BOD
Subset: Lo	wordundancy 🔳	Sort by: RLAST E-value Display: Grach a Structures per page 20 Ratroch Conglay [Total structure (accessed)]	res:1]
Refated structures	Beery any Specific hits Imperfamilies miti-desains IMBR_0	International and the state of the state	1.30x34
View 3 Alter Doo View Satur	Structure and mont in Ca3D 7nd instaled initial interview or skipn data or skipn data	41 101	click

This is a very powerful general method for finding structures whose proteins are related to the query closely enough so that structural properties can be inferred by homology.



Many opportunities to modify the alignment

Aligned domains are shown in red and blue, unaligned in gray. Colors are consistent in sequence and structure windows.

号 1KBO - Cn3D 4.1

File View Show/Hide Style Window CDD Help



Style-> Coloring Shortcuts -> Aligned

Now aligned is red and unaligned purple.

_ 🗆 ×

PQQ and active center well aligned. So our protein is a PQQ dependent dehydrogenase

It does not contain a heme group and the corresponding heme-binding domain

Bioinformatics Lunch Seminar (Summer 2014)



- Every other Friday at 1 pm. 20-30 minutes plus discussion
- Informal, ask questions anytime, start discussions
- Content will be based on feedback
- Targeted at broad audience of various levels of backgrounds and education
- Emphasis on Genomics Center

Contact: Raymond Hovey Genomics Center - SFS <u>rhovey@uwm.edu</u> 414-382-1774 http://www.greatlakesgenomics.uwm.edu/