Lesson Plan

Course: MBMG 501 (Genetic Engineering)

Lecture: Analysis of Protein Structure

Lecturer/Instructor: Duangrudee Tanramluk, Ph.D.

Contact Information: Duangrudee.tan@mahidol.ac.th

Academic Program: M.Sc. program in Molecular Genetics and Genetic Engineering and Ph.D. program in Molecular Genetics and Genetic Engineering

Date/Time: 9 November 2017 (1.30 pm –3.30 pm)

Room: C405

Learning Objectives:

At the end of this lecture, student should be able to

1. Understand the amino acid structures and their properties
2. Describe the details of protein structure from various visualization styles
3. Describe information that can be derived from a protein structure model

Content Outlines:

1. Introduction to protein structure
    - 1.1 Hierarchical organization and size of proteins
    - 1.2 The 20 amino acid sidechains
2. Protein structure representation
    - 2.1 PDB file format
    - 2.2 Structure representation styles
3. Techniques for protein structure determination
    - 3.1 X–ray crystallography
    - 3.2 Cryo-Electron Microscopy
    - 3.3 Nuclear Magnetic Resonance
4. Information that can be derived from protein structure
    - 4.1 Example of structures and their usefulness
    - 4.2 Information derived from a PDB file of protein structure.

References:

1. **Introduction to Protein Structure** by Carl Branden & John Tooze
2. **Crystallography made crystal clear** by Gale Rhodes (Chapter 2: An Overview of Protein Crystallography)
3. Saibil, H. R., **Macromolecular structure determination by cryo-electron microscopy.** *Acta Cryst*. (2000), **D56**, 1215-1222.
4. **Computational Structural Biology Methods and Applications** by Torsten Schwede and Manuel Peitsch (Chapter 27: Molecular Graphics in Structural Biology)
5. **Structural Bioinformatics** by Philip E. Bourne and Helge Weissig (Chapter 2: Fundamentals of Protein Structure, Chapter 19: Inferring Protein Function from Structure)

Suggested Reading:

Subramaniam S, *et al.* **Resolution advances in Cryo-EM enable application to drug discovery.** *Curr Opin Struct Biol* (2016), **41:**194-202.

Learning Organization:

1. Lecture
2. Kahoot Quiz
3. Questions and answers

Learning materials:

1. Powerpoint slide
2. Handout
3. Internet/Website (https://ajarnwi.wordpress.com/courses/mbmg501/)

Student Assessment:

Written examination in colored answer sheets (2 questions)

Last modified: 24/10/2017

Analysis of Protein Structure

Dr. Duangrudee Tanramluk

Contents:

- Introduction to protein structure
- Protein structure representation
- Techniques for protein structure determination
- Information that can be derived from protein structure

Objectives:

At the end of this lecture, student should be able to

1. Describe the details of protein structure from various visualization styles
2. Understand the principle of protein structure determination methods
3. Describe information that can be derived from a protein structure model.

Introduction

Proteins are essential biological molecules which play crucial roles in several cellular processes, e.g. catalyzing chemical reactions, transporting small molecules, serving as the sub-cellular structures, reacting against foreign bodies, and regulating other proteins in response to cellular environment.

The protein structure can be considered from several levels. The primary structure of a protein composes of amino acid sequence which constitutes a polypeptide chain. Some region of the polypeptide chain can form a local structure via series of hydrogen bonds from the repeating backbone torsion angles referred as the secondary structure. Combination of several of secondary structures gives rise to a three dimensional fold called the tertiary structure. Within the fold space, common structural motifs and domains can be recognised. For many proteins, a larger organization called the quaternary structure is required for the protein function. One can use the three dimensional structure of a protein to rationalize the molecular interaction underlying its biological process after obtaining the functional information from biochemical experiments. Several tools have been developed to visualize the three dimensional structures and assist in molecular design. With the advent of structural genomics, structural biologists are working on the structure of all the proteins that a genome could produce without knowing their function. Functional annotation becomes an important step toward characterization of cellular and physiological roles of gene products.

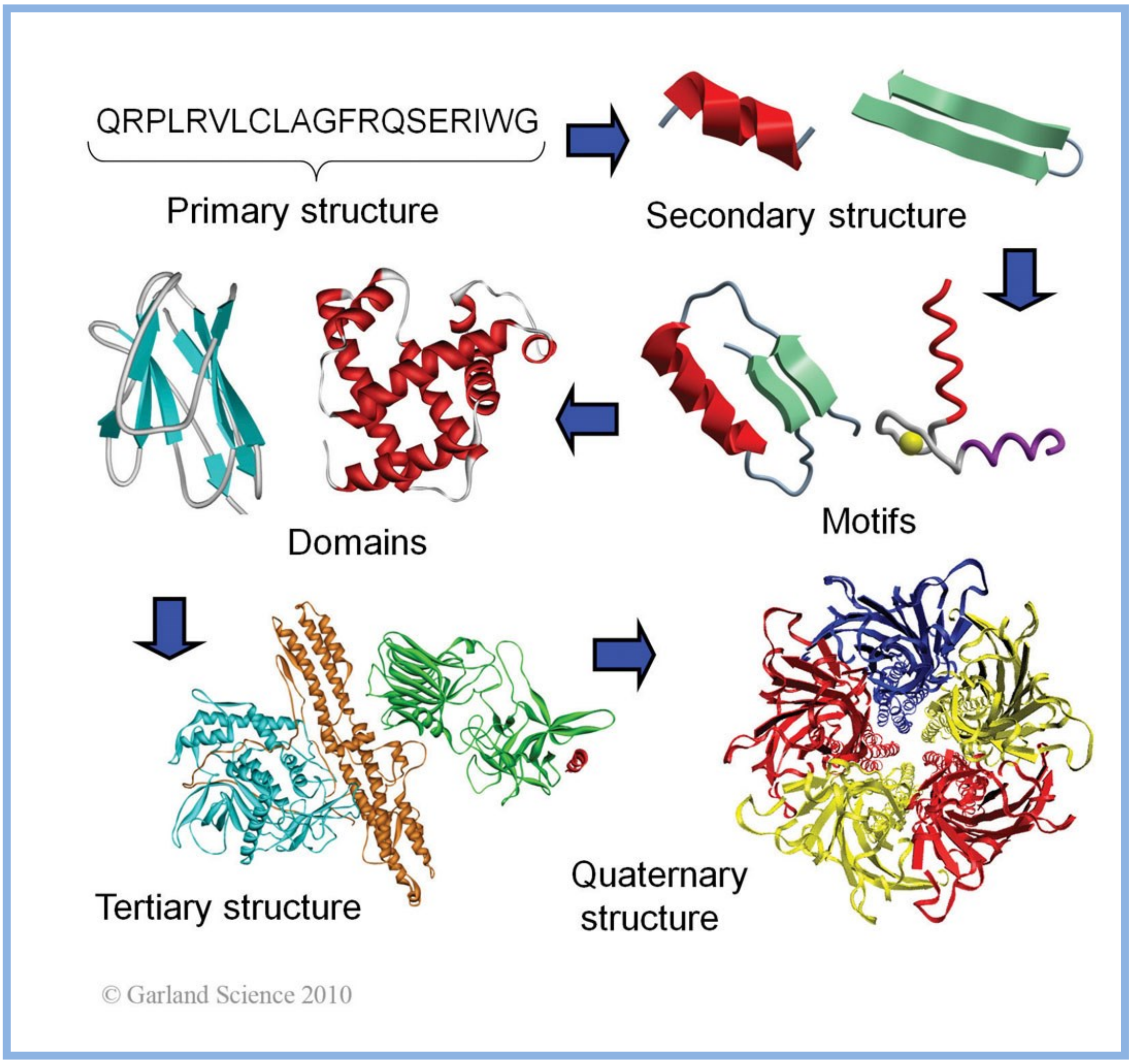
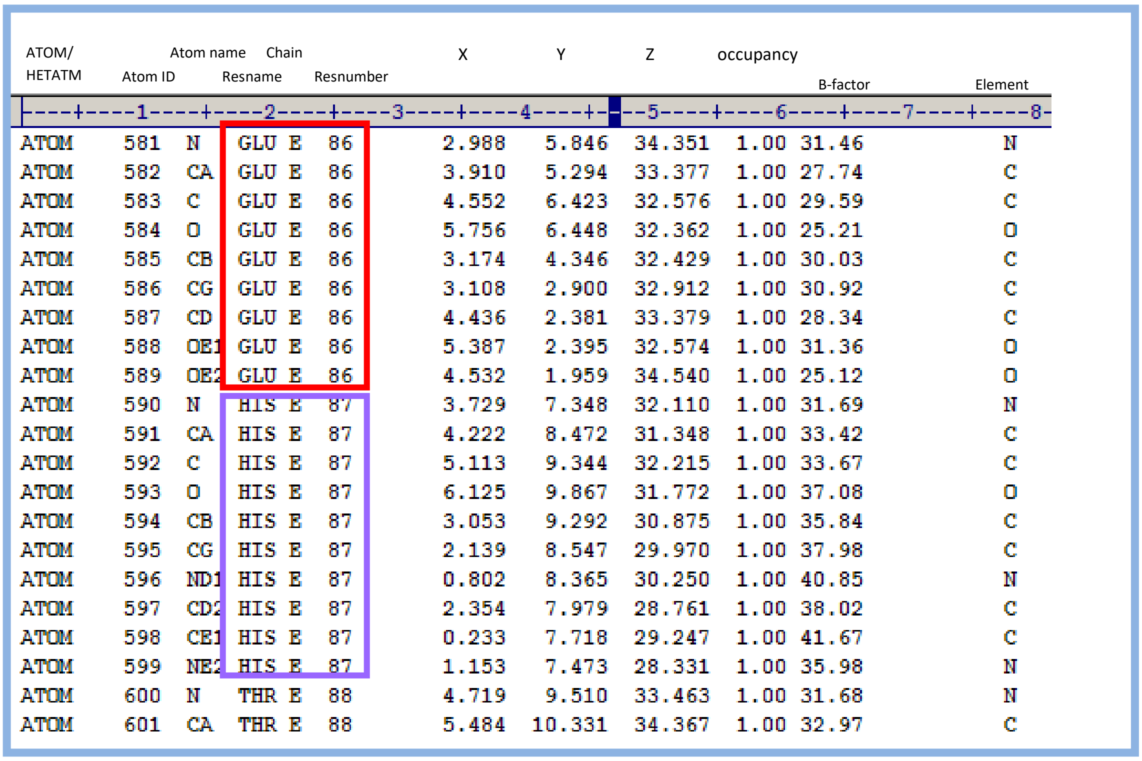


Figure 1 Protein structure hierarchy (taken from Bernhard Rupp, **Biomolecular Crystallography: Principles, Practice, and Applications to Structural Biology** (2010), Garland Science)

Protein Structure Representation

Upon publishing the protein structure in a scientific journal, structural biologist usually deposits the structural data in a repository called the Protein Data Bank (PDB: http://www.pdb.org). Each file in the PDB is given a 4-character identification number (PDB ID) and has to conform to a textual format called the PDB format. You can download the (.pdb) file and visualize molecules in a molecular viewer (e.g. Pymol, Discovery Studio). Figure 2 shows the ATOM record which is a section of the PDB file that most molecular viewers can read in order to draw atoms and bonds. Notice that we can hardly locate the position of hydrogen atom. Therefore, most of the structure model was built without hydrogen atoms and hence we usually omitted them in the PDB file.



```
|----+----1----+----2----+----3----+----4----+-█--5----+----6----+----7----+----8-
ATOM    581  N   GLU E  86       2.988   5.846  34.351  1.00 31.46           N
ATOM    582  CA  GLU E  86       3.910   5.294  33.377  1.00 27.74           C
ATOM    583  C   GLU E  86       4.552   6.423  32.576  1.00 29.59           C
ATOM    584  O   GLU E  86       5.756   6.448  32.362  1.00 25.21           O
ATOM    585  CB  GLU E  86       3.174   4.346  32.429  1.00 30.03           C
ATOM    586  CG  GLU E  86       3.108   2.900  32.912  1.00 30.92           C
ATOM    587  CD  GLU E  86       4.436   2.381  33.379  1.00 28.34           C
ATOM    588  OE1 GLU E  86       5.387   2.395  32.574  1.00 31.36           O
ATOM    589  OE2 GLU E  86       4.532   1.959  34.540  1.00 25.12           O
ATOM    590  N   HIS E  87       3.729   7.348  32.110  1.00 31.69           N
ATOM    591  CA  HIS E  87       4.222   8.472  31.348  1.00 33.42           C
ATOM    592  C   HIS E  87       5.113   9.344  32.215  1.00 33.67           C
ATOM    593  O   HIS E  87       6.125   9.867  31.772  1.00 37.08           O
ATOM    594  CB  HIS E  87       3.053   9.292  30.875  1.00 35.84           C
ATOM    595  CG  HIS E  87       2.139   8.547  29.970  1.00 37.98           C
ATOM    596  ND1 HIS E  87       0.802   8.365  30.250  1.00 40.85           N
ATOM    597  CD2 HIS E  87       2.354   7.979  28.761  1.00 38.02           C
ATOM    598  CE1 HIS E  87       0.233   7.718  29.247  1.00 41.67           C
ATOM    599  NE2 HIS E  87       1.153   7.473  28.331  1.00 35.98           N
ATOM    600  N   THR E  88       4.719   9.510  33.463  1.00 31.68           N
ATOM    601  CA  THR E  88       5.484  10.331  34.367  1.00 32.97           C
```

Figure 2 Some part of the ATOM record in the PDB ID: 1STC. This PDB file 1stc.pdb was opened with text editor EditPlus to show the atomic coordinates of glutamate (GLU), atom ID 581-589 from residue number 86 of chain E, and histidine (HIS), atom ID 590-599 from residue number 87.

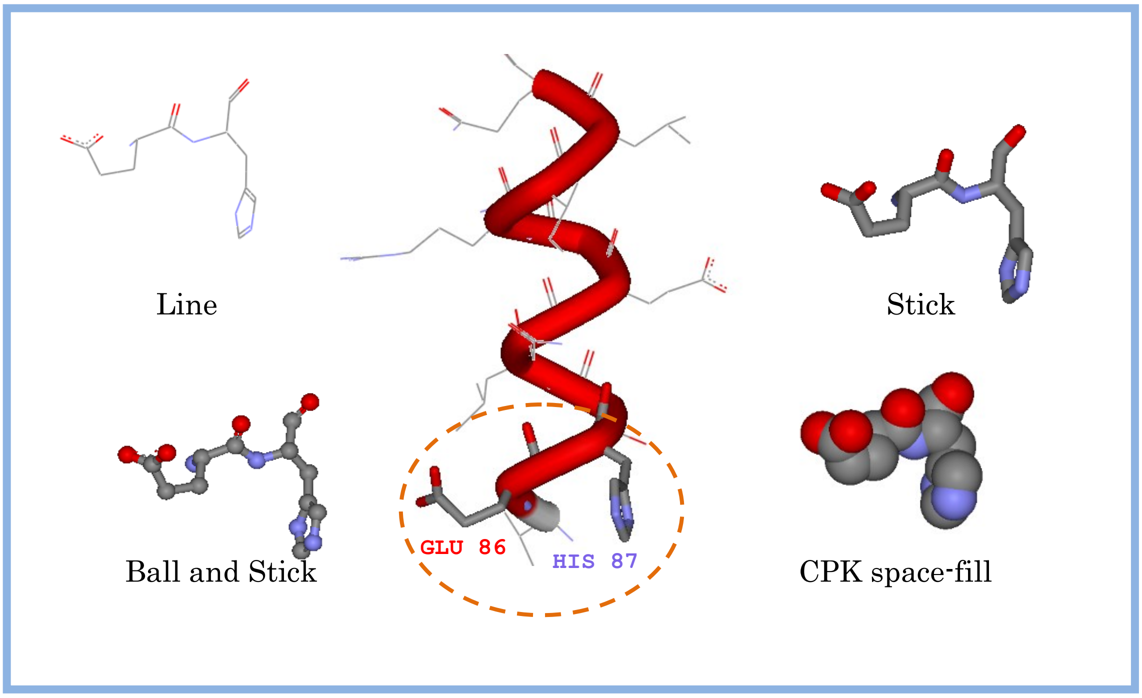
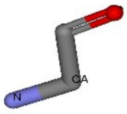
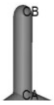
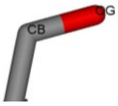
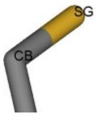
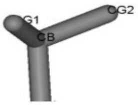
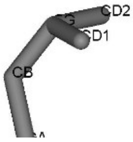
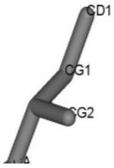
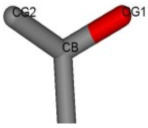
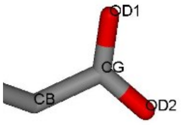
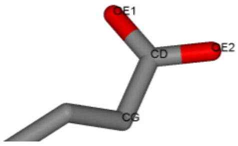
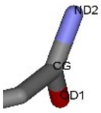
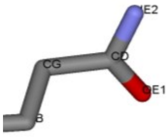
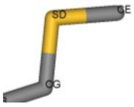
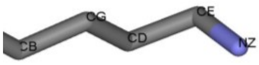
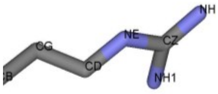
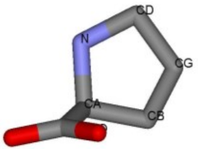
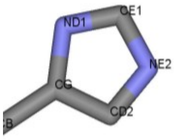
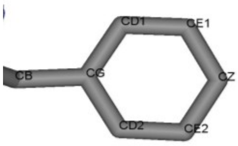
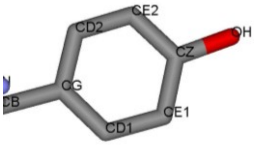
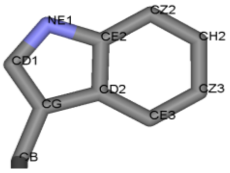


Figure 3 Atomic representations of the residue number 86 and 87 which are parts of a helix from the PDB ID: 1STC. The peptide backbone is displayed as a large spiral tube, portraying its helical character. (Figures created using Discovery Studio package).

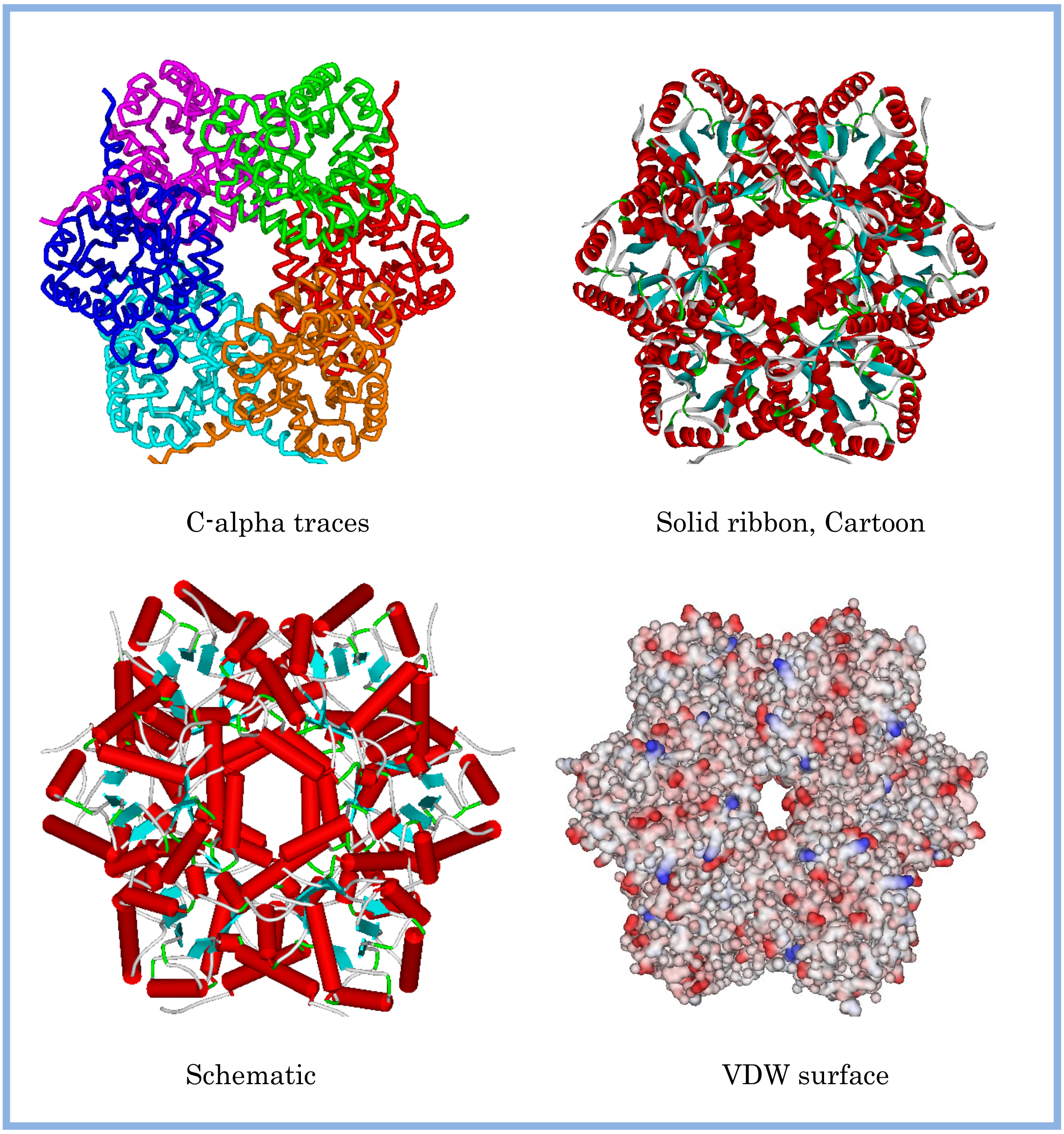
When coloring atom by element type, carbon atoms are usually colored in grey, oxygen atoms are in red, nitrogen atoms are in blue, and sulphur atoms are in yellow. The Corey-Pauling Koltun (CPK) model uses this coloring scheme and displays each atom as a sphere. Coloring atoms by their electrostatic charge is useful for visualizing binding interaction. For instance, glutamate and aspartate can form similar interaction due to their partial negative charge characters, so we can represent them as a part of the molecular surface with the same color.

Table 1 The 20 amino acid side chains in stick representation. The three-letter amino acid name is on the top with one-letter abbreviations in parentheses. Note that these side chains can adopt several conformations. (Figures created using Discovery Studio package).

GLY (G)*	ALA (A)	SER (S)	CYS (C)
			
VAL (V)	**LEU (L)**	**ILE (I)**	**THR (T)**
			
ASP (D)	**GLU (E)**	**ASN (N)**	**GLN (Q)**
			
MET (M)	**LYS (K)**	**ARG (R)**	**PRO (P)**
			
HIS (H)	**PHE (F)**	**TYR (Y)**	**TRP (W)**
			 NE1 CZ2 CE2 CH2 CD1 CD2 CZ3 CG CE3 CB

(*Glycine has no side chain, so its main chain is shown instead)

When viewing a protein structure by showing every atoms and bonds, we can hardly see the polypeptide chain that those atoms belong to. Interactions among different chains can be observed by coloring carbon atoms from each polypeptide chain as one color per one chain. In order to see the organization of secondary structure elements, alpha helices are shown with spiral shape ribbons and beta strands are shown with thick arrows. Schematic styles can show the polypeptide chain direction from the head of the cylinder rods and the arrows. Some side chains that involve in making interactions and also the ligand can be shown with stick model and the rest of the side chains are usually omitted for clarity.



C-alpha traces

Solid ribbon, Cartoon

Schematic

VDW surface

Figure 4 Amino acid representation of the PDB ID: 2P10. In the schematic style, alpha helices are displayed as rods, beta strands are represented by arrows, and turns are drawn as traces. In the VDW surface style, each atom is represented as a sphere with its van der Waals radius. (Figures created using Discovery Studio package).

Techniques for Protein Structure Determination

Various techniques can be used to determine the three dimensional structure of a protein. The majority of structural data are obtained from X-ray diffraction, nuclear magnetic resonance spectroscopy (NMR), or cryo-electron microscopy (cryo-EM) experiments. The choice of the method depends on several considerations. For example, the concentration of protein sample for both NMR and X-ray diffraction should be at least 1 mg/ml, while the protein sample for cryo-EM is more diluted and hence require a much smaller amount of sample. X-ray diffraction gives more precise value of atomic coordinates and its molecular size can range from a very small molecule to mega-daltons of protein complex. However, not all protein can be crystallized and only some proteins can diffract X-ray. Structures obtained from cryo-EM are usually in the near native physiological form and the molecule can be very large, but it often has poor resolution. Some of the strengths and weaknesses are summarized in Table 1.

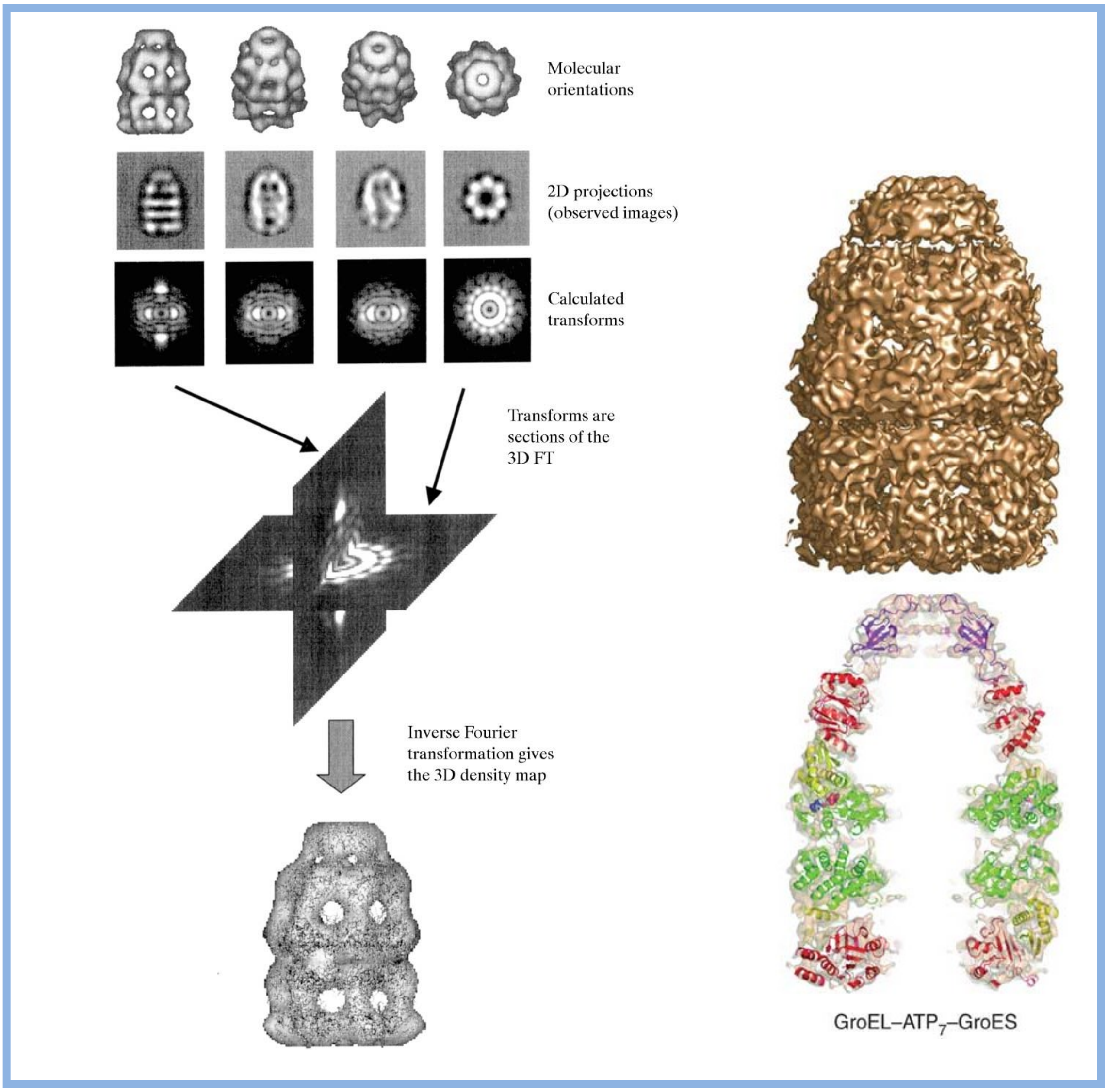

Molecular orientations
2D projections (observed images)
Calculated transforms
Transforms are sections of the 3D FT
Inverse Fourier transformation gives the 3D density map
GroEL–ATP$_7$–GroES

Figure 5 Obtaining a cryo-EM structure (taken from Saibil, H. R., *Acta Cryst.* (2000). D56, 1215-1222 and Ranson N. A., et al. *Nat Struct Mol Biol* (2006), 13, 147 – 152)

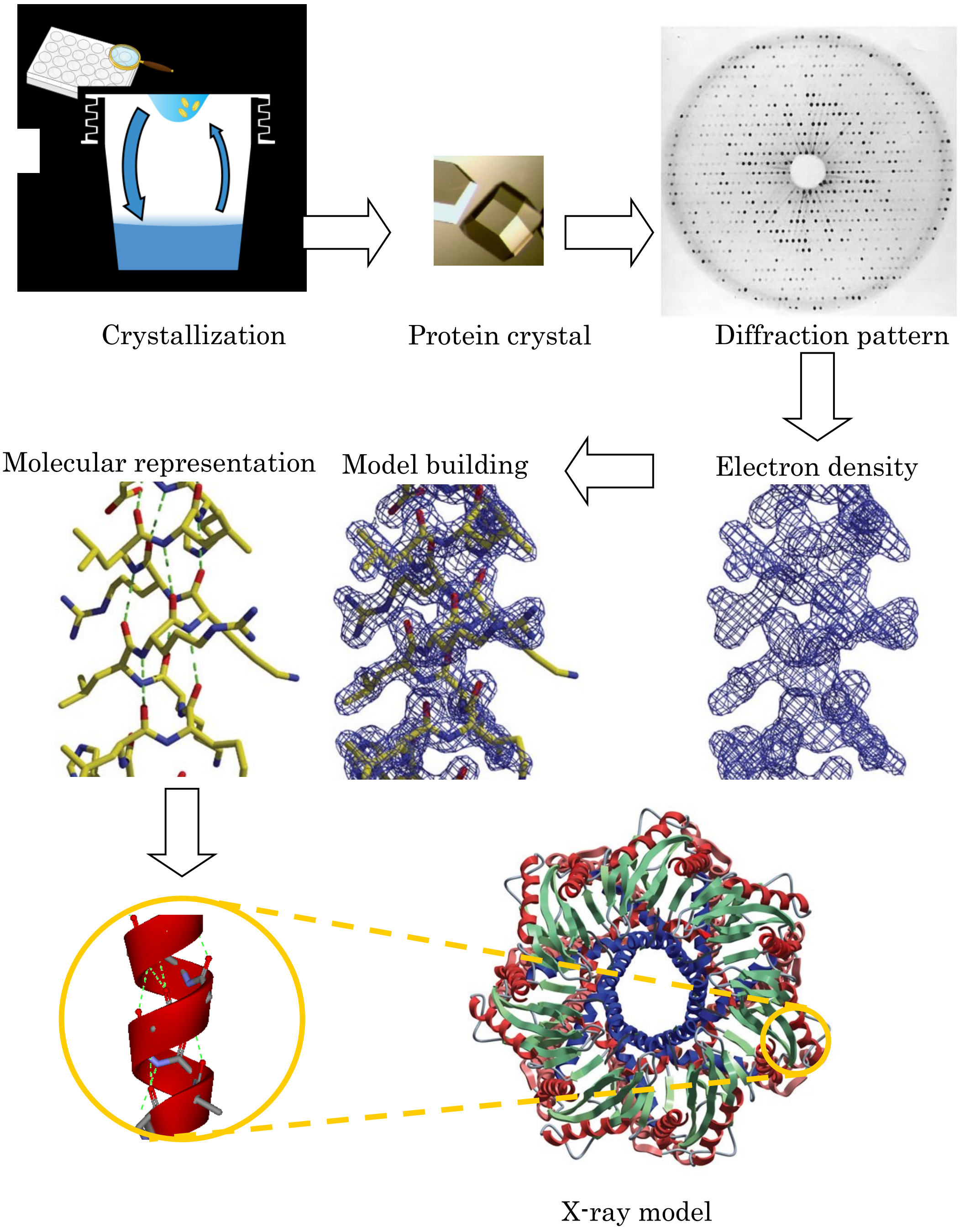


Figure 6 X-ray crystallography (pictures were adapted from Bernhard Rupp, **Biomolecular Crystallography: Principles, Practice, and Applications to Structural Biology** (2010), Garland Science and content of the figures were created using Discovery Studio package).

Table 2 Advantages and limitations of protein structure determination methods

Factor	X-ray	NMR	Cryo-EM
Sample	Crystal (from ~2 to 10 mg/ml homogenous protein)	Homogenous, monodisperse solution (~0.2 mM)	Thin cross section (<2,000 Å), crystals, ordered assemblies or isolated particles usually in native, hydrated state.
Size	Mega-dalton (MDa) range is possible	Usually < 40 kDa, up to 1 MDa in rare cases	More than 300 kDa
Data collection	Diffraction pattern	Peaks that can be interpreted to restraints	Low contrast images
Resolution	0.5 - 5.0 Å (median ~2.0 Å)	-	Usually >3.0 Å, about 10 Å is normal. An extremely rare case is at < 2.0 Å resolution.
Conformation	As in crystal lattice	Dynamic processes	Can also trap transient state by rapid freezing
Final structure	One model of atomic coordinates with B-factor	An ensemble of atomic coordinates	EM electron density map, Cα trace.

Information that can be derived from 3D structure

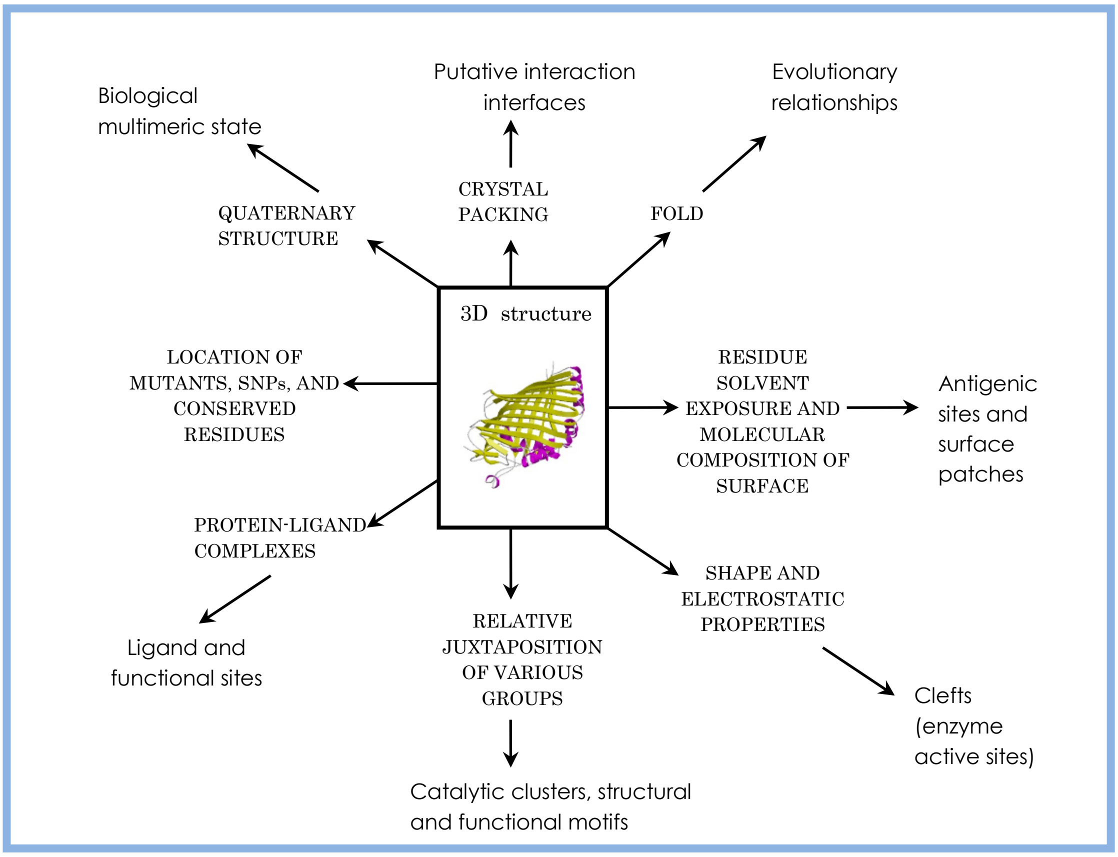

Biological multimeric state
Putative interaction interfaces
Evolutionary relationships
QUATERNARY STRUCTURE
CRYSTAL PACKING
FOLD
3D structure
LOCATION OF MUTANTS, SNPs, AND CONSERVED RESIDUES
RESIDUE SOLVENT EXPOSURE AND MOLECULAR COMPOSITION OF SURFACE
Antigenic sites and surface patches
PROTEIN-LIGAND COMPLEXES
SHAPE AND ELECTROSTATIC PROPERTIES
Ligand and functional sites
RELATIVE JUXTAPOSITION OF VARIOUS GROUPS
Clefts (enzyme active sites)
Catalytic clusters, structural and functional motifs

Figure 5 From structure to function: A summary of information that can be derived from 3D structure relating to biological function (reproduced from Philip E. Bourne and Helge Weissig, **Structural Bioinformatics** (2003) John Wiley & Sons, Inc.)

If you have a PDB identification number (PDB ID), a protein name, or its amino acid sequence, you can search for a PDB ID of the structure from various websites. In this lecture, I will show you how to use a website called **PDBsum** in order to start exploring your structure. It provides several links to visualize the structure. To display a structure from PDBsum on a Java-enabled web browser, you can just activate the program **Jmol** on the left of Figure 7 or Figure 9. The URL for PDBsum is:

http://www.ebi.ac.uk/pdbsum

For each PDB code, this server extracts most of the information the can be found before the ATOM record of a PDB file.

For example:

- Title

  The name of the protein with ligand and/or nucleic acid in the complex is shown.

- Chains

  It can identify all the polypeptide chains that are associated with the structure. The chain length (number of amino acids in the chain) is also provided.

- Details about the protein and the source organism

  The ***Uni***versal ***prot***ein resource (Uniprot) is the gold standard for annotation of amino acid sequence. The UniprotKB entry is in an abbreviation form of protein_organism. If you enter PDB ID: 1B39 in PDBsum, you will obtain Uniprot ID: CDK2_HUMAN which is for the cell division protein kinase 2 from human.

- Resolution

  It is the uncertainty of atomic position in a model. If the value, in angstrom (Å), is smaller, the electron density from the map is more accurate. Cryo-EM density map cannot be used for analyzing atomic interaction because its resolution is so poor that it cannot accurately define the inter-atomic distance in the model. A quick look at resolution, i.e. high resolution is < 2 Å, can imply the good model quality.

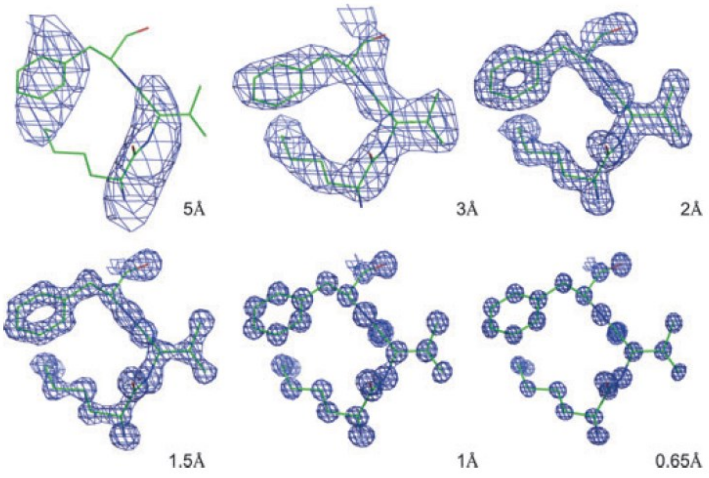
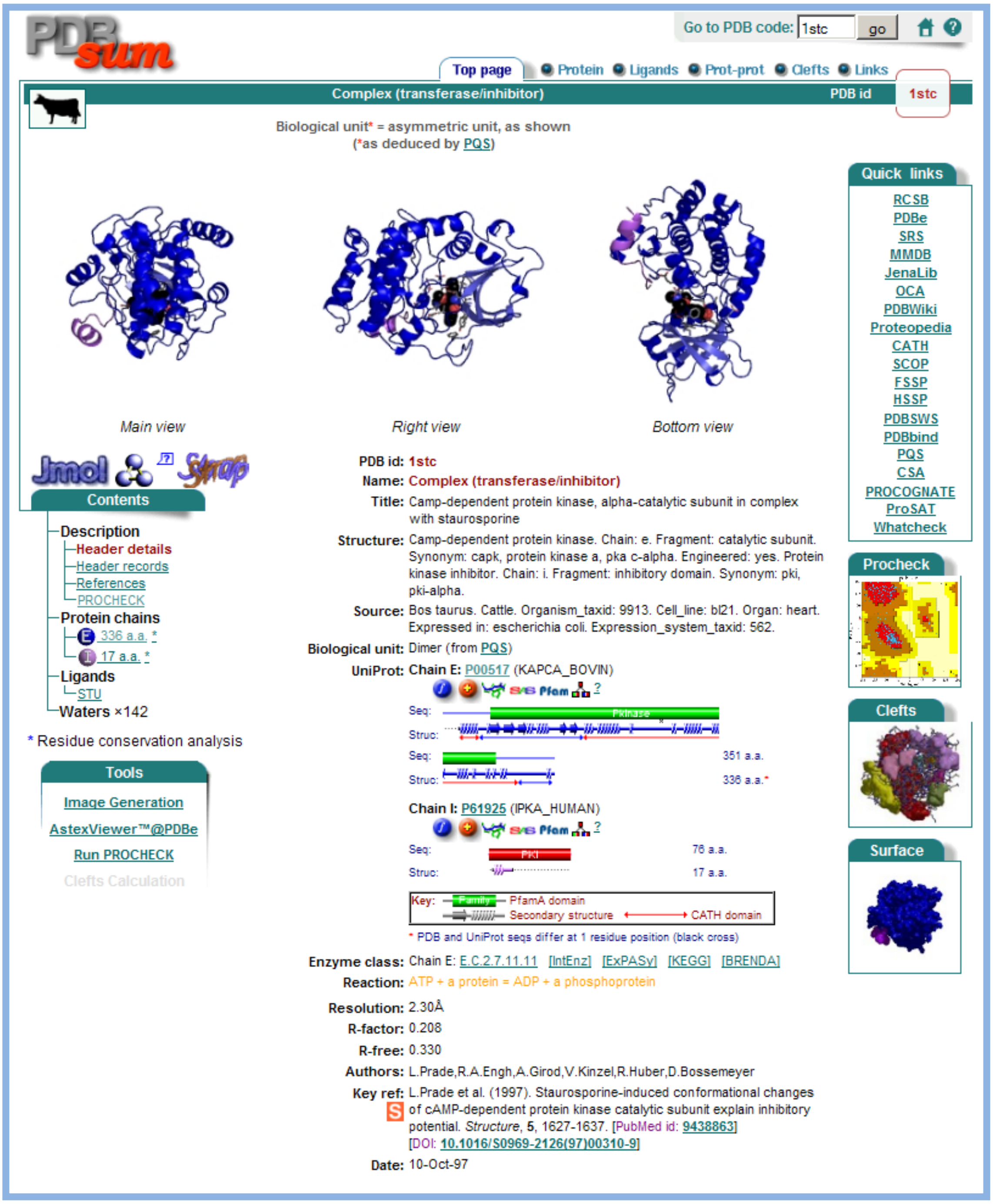


Figure 6 Effects of resolution. The electron density map (mesh) obtained at 0.65Å resolution agrees with the position of atoms in the line model better than the one obtained at 5Å resolution. (Taken from Wlodawer, A., et al. *FEBS J.* (2008), 275, 1–21.)

- Bound ligand and metal ion

In the PDB file, a heterogen is a molecule, a prosthetic group, a solvent, or an ion that is not a part of the protein and its atoms are listed in the HETATM section of the PDB file. A small molecule is usually identified with its heterogen three-letter code, while an ion is normally identified by its element symbol.

- Authors, references, and abstract of the published literature



Complex (transferase/inhibitor)

Biological unit* = asymmetric unit, as shown
(*as deduced by PQS)

PDB id: 1stc
Name: Complex (transferase/inhibitor)
Title: Camp-dependent protein kinase, alpha-catalytic subunit in complex with staurosporine
Structure: Camp-dependent protein kinase. Chain: e. Fragment: catalytic subunit. Synonym: capk, protein kinase a, pka c-alpha. Engineered: yes. Protein kinase inhibitor. Chain: i. Fragment: inhibitory domain. Synonym: pki, pki-alpha.
Source: Bos taurus. Cattle. Organism_taxid: 9913. Cell_line: bl21. Organ: heart. Expressed in: escherichia coli. Expression_system_taxid: 562.
Biological unit: Dimer (from PQS)
UniProt: Chain E: P00517 (KAPCA_BOVIN)
351 a.a.
336 a.a.*
Chain I: P61925 (IPKA_HUMAN)
76 a.a.
17 a.a.

Key: 🟩 Family PfamA domain
Secondary structure ←→ CATH domain

* PDB and UniProt seqs differ at 1 residue position (black cross)

Enzyme class: Chain E: E.C.2.7.11.11 [IntEnz] [ExPASy] [KEGG] [BRENDA]
Reaction: ATP + a protein = ADP + a phosphoprotein
Resolution: 2.30Å
R-factor: 0.208
R-free: 0.330
Authors: L.Prade,R.A.Engh,A.Girod,V.Kinzel,R.Huber,D.Bossemeyer
Key ref: S L.Prade et al. (1997). Staurosporine-induced conformational changes of cAMP-dependent protein kinase catalytic subunit explain inhibitory potential. *Structure*, **5**, 1627-1637. [PubMed id: 9438863] [DOI: 10.1016/S0969-2126(97)00310-9]
Date: 10-Oct-97

Figure 7 PDBsum screenshot upon searching for PDB ID: 1STC from **www.ebi.ac.uk/pdbsum.**

In addition to the information that can be found in the head of the PDB file, there are also some ‘derived’ data that are included in PDBsum. For example:

- Protein-ligand interactions

In order to bind protein well, the ligand usually has complementary electrostatic potentials with the amino acids in the protein binding site. Hydrophobic interactions can also govern protein packing and the hydrogen bonding interactions are important for ligand binding.

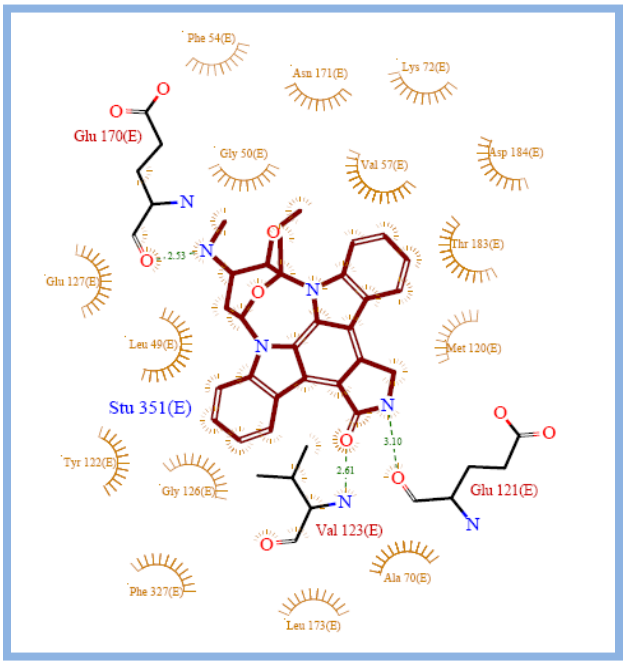


Figure 8 Ligand interactions obtained from the ‘Ligands’ tab from the top of Figure 7. The program LIGPLOT shows a ligand called staurosporine (heterogen name:STU) interacts with various amino acids in the pocket of the cAMP-dependent protein kinase (PDB ID: 1STC). Hydrogen bonds are shown with the dash lines. To see more detailed analysis of the ligand interaction with series of proteins, the server MANORAA.org can be used.

- Protein-protein interactions

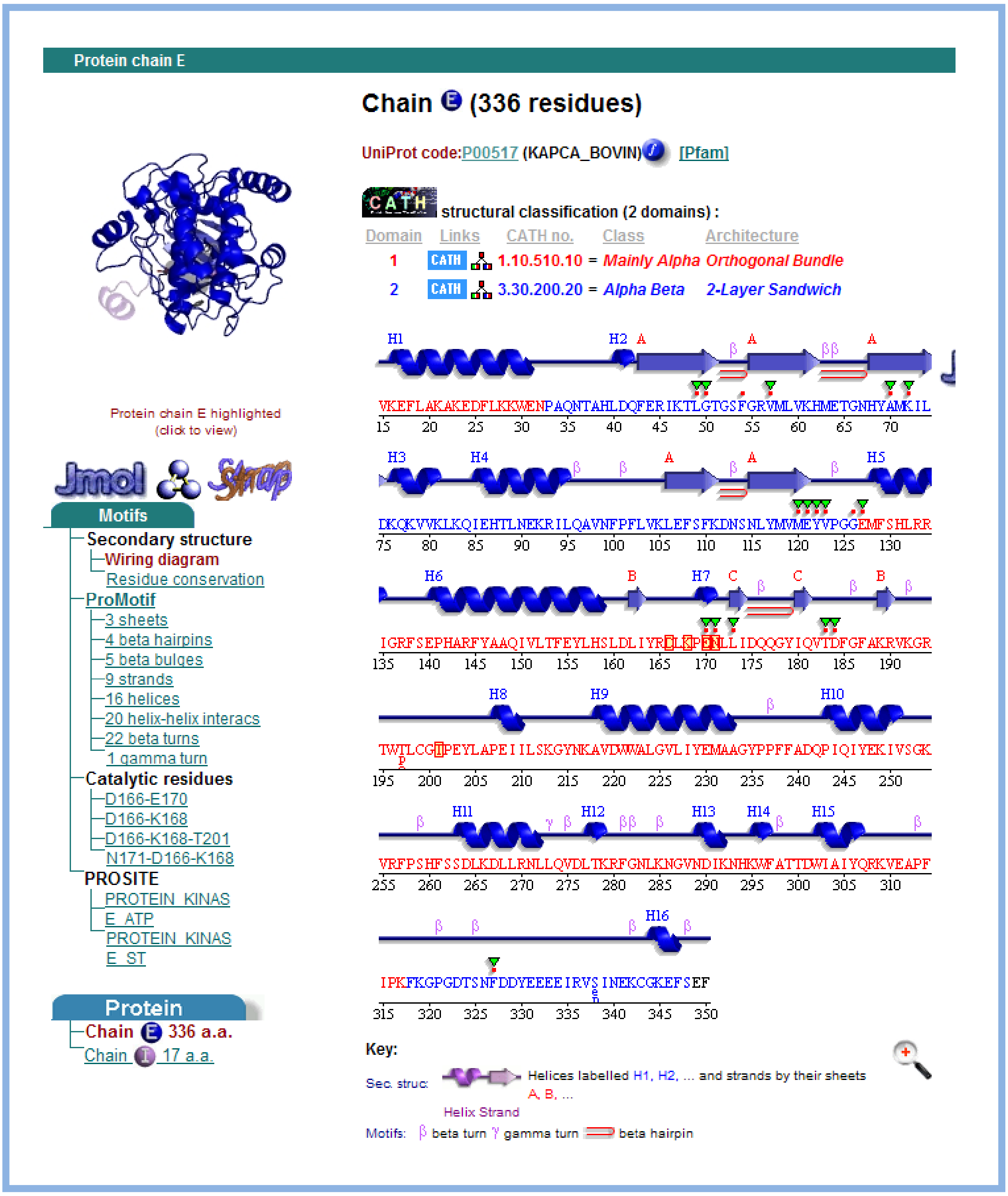
There is a ‘Prot-prot’ tab to show interacting interfaces between this protein and other proteins on the top of Figure 7.

- Cleft positions

The largest cleft is a possible active site and may be useful for inhibitor design.

- Ramachandran plot

PROCHECK is the underlying program that calculate Ramachandran plot for PDBsum. You can assess whether the structure has a bad model quality by observing the phi-psi angles whether they are distributed in the usual location. There are usual locations where alpha helices, beta sheets and left handed helices occurred with glycine scattered around. If these angles deviated too much from the standard locations, it may mean that the structure was not correctly built or was fabricated.



Protein chain E

Chain E (336 residues)

UniProt code: P00517 (KAPCA_BOVIN) [Pfam]

CATH structural classification (2 domains) :

Domain	Links	CATH no.	Class	Architecture
1	CATH	1.10.510.10 =	Mainly Alpha	Orthogonal Bundle
2	CATH	3.30.200.20 =	Alpha Beta	2-Layer Sandwich

```
VKEFLAKAKEDFLKKWENPAQNTAHLDQFERIKTLGTGSFGRVMLVKHMETGNHYAMKIL
15   20   25   30   35   40   45   50   55   60   65   70

DKQKVVKLKQIEHTLNEKRILQAVNFPFLVKLEFSFKDNSNLYMVMEYVPGGEMFSHLRR
75   80   85   90   95   100  105  110  115  120  125  130

IGRFSEPHARFYAAQIVLTFEYLHSLDLIYRDLKPENLLIDQQGYIQVTDFGFAKRVKGR
135  140  145  150  155  160  165  170  175  180  185  190

TWTLCGTPEYLAPEIILSKGYNKAVDWWALGVLIYEMAAGYPPFFADQPIQIYEKIVSGK
195  200  205  210  215  220  225  230  235  240  245  250

VRFPSHFSSDLKDLLRNLLQVDLTKRFGNLKNGVNDIKNHKWFATTDWIAIYQRKVEAPF
255  260  265  270  275  280  285  290  295  300  305  310

IPKFKGPGDTSNFDDYEEEEIRVSINEKCGKEFSEF
315  320  325  330  335  340  345  350
```

Key:

Sec. struc: Helix Strand — Helices labelled H1, H2, ... and strands by their sheets A, B, ...

Motifs: β beta turn γ gamma turn beta hairpin

Motifs
- Secondary structure
  - Wiring diagram
  - Residue conservation
- ProMotif
  - 3 sheets
  - 4 beta hairpins
  - 5 beta bulges
  - 9 strands
  - 16 helices
  - 20 helix-helix interacs
  - 22 beta turns
  - 1 gamma turn
- Catalytic residues
  - D166-E170
  - D166-K168
  - D166-K168-T201
  - N171-D166-K168
- PROSITE
  - PROTEIN_KINASE_ATP
  - PROTEIN_KINASE_ST

Protein
- Chain E 336 a.a.
- Chain I 17 a.a.

Figure 9 By clicking on the protein chain, one can obtain the amino acid sequence, the sequence diagram and the CATH identification number. The protein class and the architecture of this chain can be recognized from the CATH ID. This picture also shows residues that are in contact with the ligand (dots) and active site residues (tiny rectangles). Secondary structures are shown on top of sequence (cartoon). Source: www.ebi.ac.uk/pdbsum.

- Fold classification

CATH (*C*lass, *A*rchitecture, *T*opology, *H*omologous superfamily) is a server for protein fold classification. PDBsum provides a link to CATH from the 'Protein' tab. According to CATH, a protein can be primarily classified as mainly alpha, mainly beta, mixed alpha-beta, or few secondary structures. If two proteins have similar sequences and adopt the same fold, they can be evolutionary related. With a CATH ID, you can look for structures that have the same protein fold with your queried PDB ID.

- Enzyme Commission number (E.C.)

Every enzyme code can be identified with the letter "EC" followed by four numbers separated by periods. If your protein is an enzyme, it will have its own E.C. number to identify that it is an oxidoreductase, transferase, hydrolase, lyase, isomerase, or ligase.

- Pathways

The Kyoto Encyclopedia of Genes and Genomes (KEGG) is one of the most comprehensive interaction pathway resources. A pathway can be retrieved if your protein is known to take part in metabolism, genetic information processing, environmental information processing, cellular processes, organismal systems, or human diseases. This is how you can discover the role of your protein in relation with other proteins in the pathway.

This reading material introduced the concepts of the protein structure representation, the differences among major structure determination methods, and also an overview of structural information that can be obtained from the PDB file. These key points and details of the figures can provide students with background knowledge to understand literatures on structural biology and the relating fields.