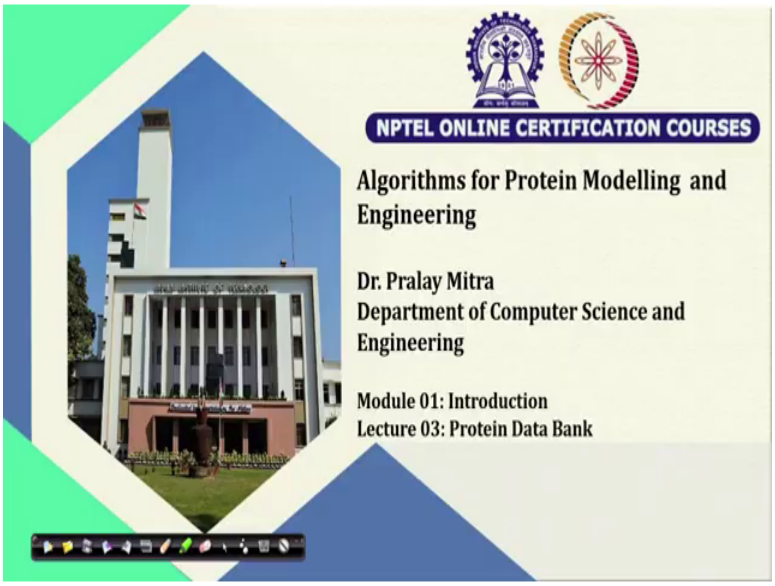
Algorithms for protein modelling and engineering
Professor Pralay Mitra
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur
Lecture: 03
Protein Data Bank

(Refer Slide Time: 00:15)



NPTEL ONLINE CERTIFICATION COURSES

Algorithms for Protein Modelling and Engineering

Dr. Pralay Mitra
Department of Computer Science and Engineering

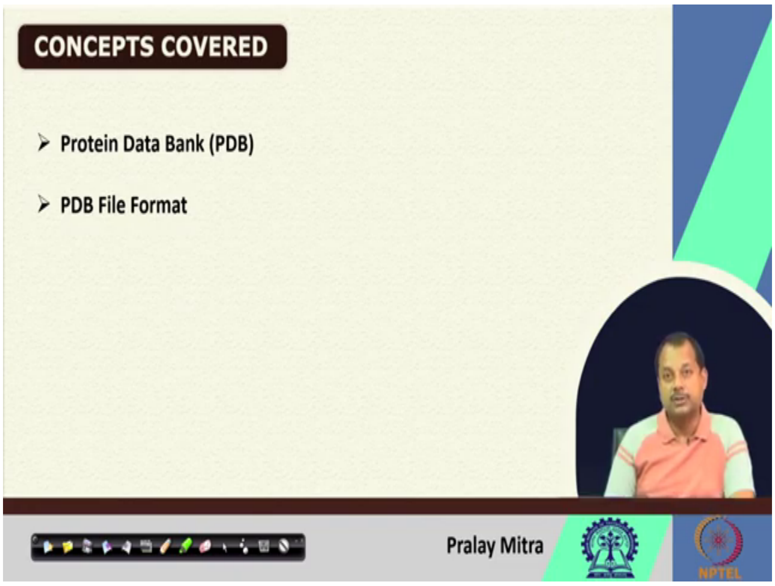
Module 01: Introduction
Lecture 03: Protein Data Bank

Welcome back. Today in this third lecture of the introduction module, we plan to discuss the protein structural database. There is only one database we shall be discussing and that is the protein data bank and throughout the course, that protein data bank is enough for our discussion.

While discussing this protein data bank, you should also note down that when I say it is the structural database, then obviously, the sequence information will be available to us. But additionally, we shall have atomic coordinate information and the details of the experimental techniques through which the coordinate information has been identified.

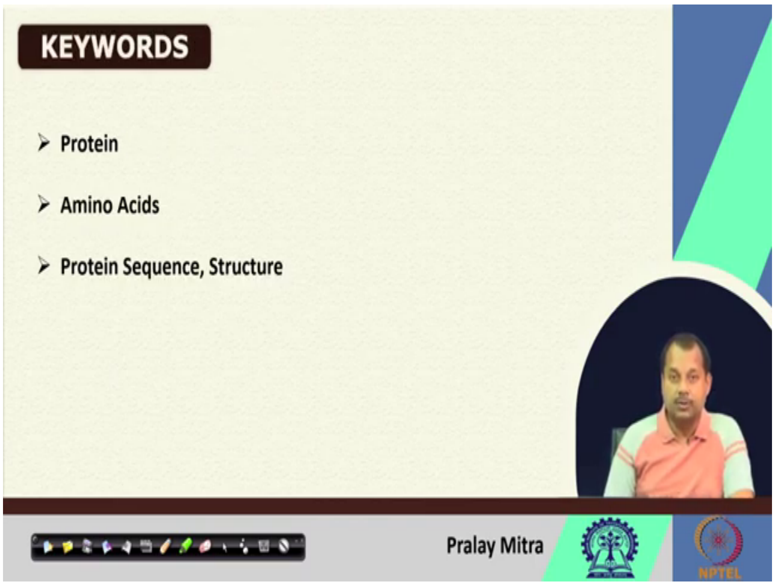
In the context of the protein data bank, mostly, it is the X-ray crystallography technique that is being used. Apart from that experimental techniques like nuclear magnetic resonance (NMR) or cryo-EM or solid-state NMR are also used to get the coordinates or the structure of the protein at the atomic level resolution. OK! Let us start now.

(Refer Slide Time: 01:28)



CONCEPTS COVERED

- Protein Data Bank (PDB)
- PDB File Format



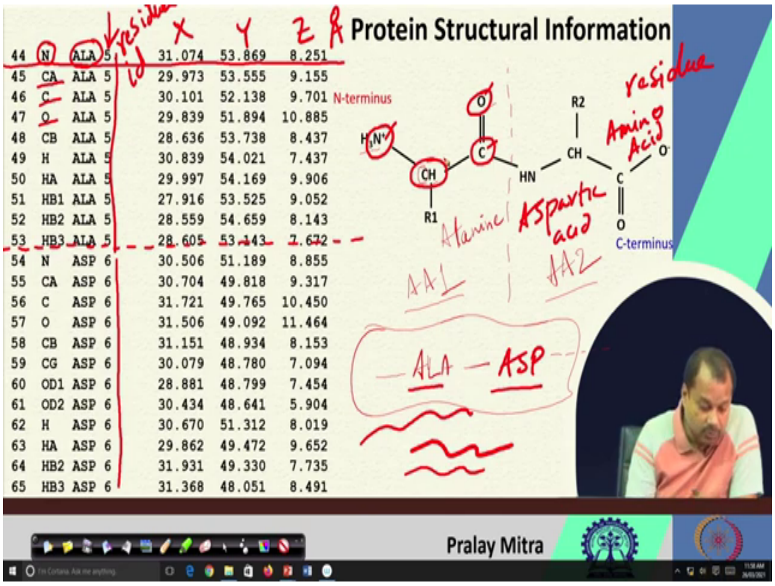
KEYWORDS

- Protein
- Amino Acids
- Protein Sequence, Structure

The topics I am planning to cover today are the protein data bank (PDB) and the PDB file format. Today, we shall start the discussion on the PDB file format, and shall continue with PDB file parsing etcetera.

The keyword is the same as the previous lecture - protein amino acid, protein sequence, and structure. Because, from this protein data bank, we shall get the sequence as well as the structural information regarding a protein molecule.

(Refer Slide Time: 02:07)



Protein Structural Information

			residue id	X	Y	Z Å
44	N	ALA	5	31.074	53.869	8.251
45	CA	ALA	5	29.973	53.555	9.155
46	C	ALA	5	30.101	52.138	9.701
47	O	ALA	5	29.839	51.894	10.885
48	CB	ALA	5	28.636	53.738	8.437
49	H	ALA	5	30.839	54.021	7.437
50	HA	ALA	5	29.997	54.169	9.906
51	HB1	ALA	5	27.916	53.525	9.052
52	HB2	ALA	5	28.559	54.659	8.143
53	HB3	ALA	5	28.605	53.143	7.672
54	N	ASP	6	30.506	51.189	8.855
55	CA	ASP	6	30.704	49.818	9.317
56	C	ASP	6	31.721	49.765	10.450
57	O	ASP	6	31.506	49.092	11.464
58	CB	ASP	6	31.151	48.934	8.153
59	CG	ASP	6	30.079	48.780	7.094
60	OD1	ASP	6	28.881	48.799	7.454
61	OD2	ASP	6	30.434	48.641	5.904
62	H	ASP	6	30.670	51.312	8.019
63	HA	ASP	6	29.862	49.472	9.652
64	HB2	ASP	6	31.931	49.330	7.735
65	HB3	ASP	6	31.368	48.051	8.491

Now, let me show you this diagram. Here you can see that if I pick a color say red. On the right-hand side, you know that this is the borderline, and left-hand side this is amino acid 1 and on the right-hand side amino acid 2. Two different amino acids are there. Thus you can consider them as a di-peptide - two amino acids connected and forming one peptide bond.

If I keep on adding the amino acids one after another that way we will get the complete protein sequence and that particular sequence will give us the protein molecule. And when I look inside this one then I will identify that they are atoms like carbon, nitrogen, hydrogen, here, here and then oxygen is here. Carbon is also present here etcetera.

When I say it is the protein structure data, which means the coordinate information of each of the atoms at the 3-dimensional space is present. By coordinate information, I mean the atomic coordinate information of the amino acids. Corresponding to a protein molecule if I zoom in then I will have the list of amino acids, say amino acid 1, amino acid 2 like that way I will have the amino acids. OK!

These amino acids - you remember that there is a full form or the complete name of those amino acids like alanine, cysteine, aspartic acid, glutamic acid etcetera. Also, those names can be represented by three-character letters. Also, those names can be represented by one single character taken from the English alphabet. In the protein data bank, the three-character letters are stored.

Hence, I shall identify one amino acid. Now, if I assume that this is my alanine. And say this is my cysteine. This, I am assuming without any loss of generality, is just for this discussion. Then it will be ALA-CYS that establishes a connection/covalent bond. Right? From a protein point of view, this is alanine and cysteine. If I look inside, this is di-peptide (only two amino acids are there), but some others will also be connected on the left-hand side or right-hand side.

There are two amino acids in this case alanine and cysteine. OK! Now, this alanine and cysteine corresponding to each amino acid will have one central carbon, then nitrogen, then oxygen, then this carbon. Those atoms will be there. So, we need to represent the coordinate information of all of them.

If I look on the left-hand side here, then you will see that these three columns indicate three coordinates X, Y, and Z in angstrom (denoted as Å is 10^{-8} meters). Each line represents information regarding one atom. What is that information? The name of the atom - it is the nitrogen belongs to amino acid say alanine. And its coordinate is X, Y, Z with values 31.071, 53.869, 8.251, respectively.

When I have those amino acids say alanine, cysteine in this case. Here in this example, it is alanine and aspartic acid. To keep parity, with this I can say erase this one and I can write say aspartic acid. Then, it will be ASP and now you see that this ASP is on the left-hand side and ALA. In a protein sequence, those amino acids are there.

If I assume that it is starting from the left-hand side, so 1, 2, 3, 4, 5, 6, so each amino acid will have some numbering. That numbering is here. And this is called the residue ID. You should remember one thing residue and this amino acid are used interchangeably. When I say amino acids, it means residue and when I say residue then it means amino acid. In the context of this protein molecule, they are used interchangeably.

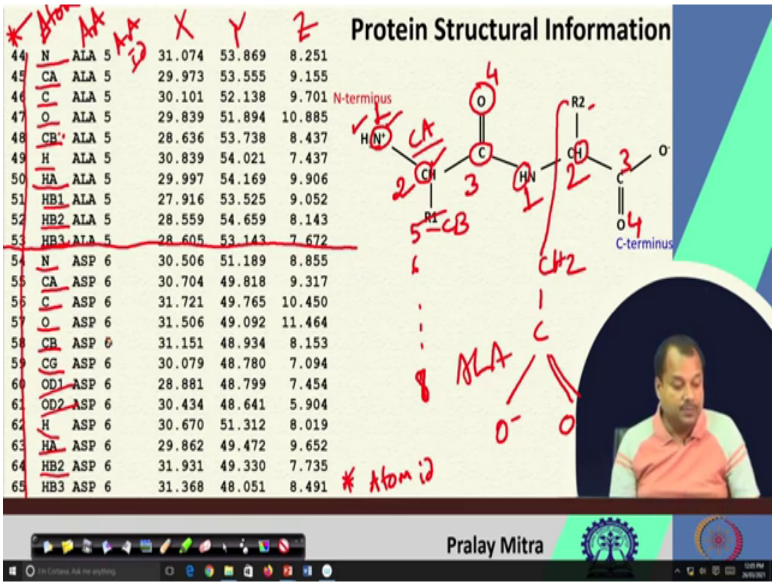
This ID you can consider residue ID or amino acid ID and that is why you see alanine up to here, it is alanine. Its residue number is 5, then it is aspartic acid, its residue number is 6 that way it is changing. It is a good thing you can consider that as if that is the serial number corresponding to that amino acid. Now, this is the name of the atom inside the residue or the amino acid.

In the case of alanine, first of all, corresponding to each amino acid the basic structure is the same. At the center, there is one carbon and on one side that is nitrogen, another side is carbon and this carbon is attached with oxygen. Occasionally, we can get the hydrogen if we are doing NMR or if we are using some high-resolution X-ray crystallography or X-ray diffraction technique or we are using some sophisticated software or computationally we are attaching the hydrogen molecule. Otherwise, because of the small size of the hydrogen atoms most of the time during X-ray diffraction, we are not able to detect the hydrogen atom.

That’s why you may see that hydrogen atoms are missing in most of the protein structures, that’s fine. If a hydrogen atom is required then using some computational techniques (there are some computational techniques, I will discuss in the context), we can attach the hydrogen atoms with the protein molecule. Otherwise, if it is there, it is fine. If it is not there, then also it is fine.

There is a pattern in this protein data bank on how the atoms will be stored corresponding to an amino acid. First of all, you can consider that. First amino acid, second amino acid, third, fourth, fifth, sixth, like that it will go. For the first amino acid, there are several atoms. If there are many atoms, then how do they will order? Also, regarding the ordering, there is one pattern.

(Refer Slide Time: 10:32)



Protein Structural Information

*	Atom	AA	AA id	X	Y	Z
44	N	ALA	5	31.074	53.869	8.251
45	CA	ALA	5	29.973	53.555	9.155
46	C	ALA	5	30.101	52.138	9.701
47	O	ALA	5	29.839	51.894	10.885
48	CB	ALA	5	28.636	53.738	8.437
49	H	ALA	5	30.839	54.021	7.437
50	HA	ALA	5	29.997	54.169	9.906
51	HB1	ALA	5	27.916	53.525	9.052
52	HB2	ALA	5	28.559	54.659	8.143
53	HB3	ALA	5	28.605	53.143	7.672
54	N	ASP	6	30.506	51.189	8.855
55	CA	ASP	6	30.704	49.818	9.317
56	C	ASP	6	31.721	49.765	10.450
57	O	ASP	6	31.506	49.092	11.464
58	CB	ASP	6	31.151	48.934	8.153
59	CG	ASP	6	30.079	48.780	7.094
60	OD1	ASP	6	28.881	48.799	7.454
61	OD2	ASP	6	30.434	48.641	5.904
62	H	ASP	6	30.670	51.312	8.019
63	HA	ASP	6	29.862	49.472	9.652
64	HB2	ASP	6	31.931	49.330	7.735
65	HB3	ASP	6	31.368	48.051	8.491

So, let me go to that pattern. This N, N-terminus, or amino N will occur first that you can see here. After that one this C which is also called the CA or C alpha this will occur. After that one, this C which is the carbon from the carboxylic group will appear here, after that the oxygen which is attached with this carbon inside the carboxyl group, will be there. After that, first, this is 1, this is 2, this is 3, this is 4, then, this will start 5, 6, 7, 8, 9, 10 based upon how many side chains are there and it varies from amino acid to amino acid.

For glycine there is the only hydrogen here, for alanine there is only one carbon here - that is CH3 will come here. Likewise that it will go, and when I say this is CA, this A also stands for alpha, then C beta will be the next atom just after here that is this one. That is why for Alanine, you see CB, B stands for beta C stands for carbon that is the atom.

If you consider only the first character in this column, then you will get what is the atom. If you consider the next onwards then you will see what is their position if this is N. After N there is nothing that means it is part of this amino group. If there is A which means this is part of C alpha if after C nothing is there, then it is part of the carboxyl group, if after O nothing is there then it is part of the carboxyl group that way will go.

Now, CB indicates this guy. This R1 is for CB because we are assuming this is my ALA or Alanine. Now after CB this hydrogen will come, this hydrogen is here. Now, this HA will come. HA means this hydrogen is attached with CA then HB 1, B 2, B 3 that you understand.

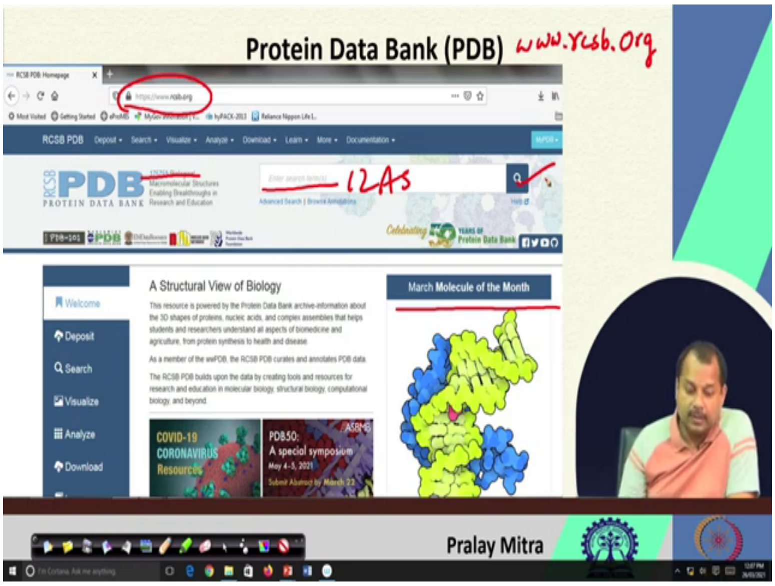
Hydrogen 1 is attached with CB, hydrogen 2 is attached with CB, hydrogen 3 is attached with CB that will come, and that completes my information regarding alanine residue. The coordinate information as I mentioned is X, this is Y, and this is Z. This is an amino acid in three-letter characters, this is amino acid ID, this is atom information and this one that I am writing here is atom ID.

Similar to an amino acid, you also need to store the specific ID corresponding to each atom. Those atom numbers are going here. Apart from this is it is not the structure of the PDB that I will go. I shall discuss that later. But this is the information you need. If I say I am interested to store the protein molecule's structure at the atomic level resolution. Then I need to correspond to each atom the coordinate information that particular item belongs to which amino acid what is the ID of that amino acid because if there are multiple alanines then which alanine I am talking about, I have to specify that one.

Again, inside that alanine, if there are two carbons - which carbon, I am talking about? I have to specify that one. That is why as ID I have to mention the position. And these are the minimum information that I need to store so that uniquely I can identify one particular atom that belongs to one particular amino acid inside a protein molecule. This information is required. The same thing is repeating in ASP or this aspartic acid. Now, this is N 1, this is 2, this is 3, this is 4, and then CB hereafter the CG, Beta, Gamma, then delta oxygen D1, D2 because there is a bifurcation.

You will see that if the situation says something like this then it is somewhat oxygen here and here. So, it will be this one. If it is like this, then it will be OD1, OD2. Now, this H you know that this H attached with nitrogen then HA attached with C alpha, then HB2 is H2 here and this is R2, HB2, HB3 that is my aspartic acid. This is the minimum information I need to store. With that aim, I will go to the protein data bank and I will extract that information from that protein database. But before that one, let me give you one introduction to this protein data bank which will be useful for you.

(Refer Slide Time: 16:53)



Protein Data Bank (PDB) www.rcsb.org

A Structural View of Biology

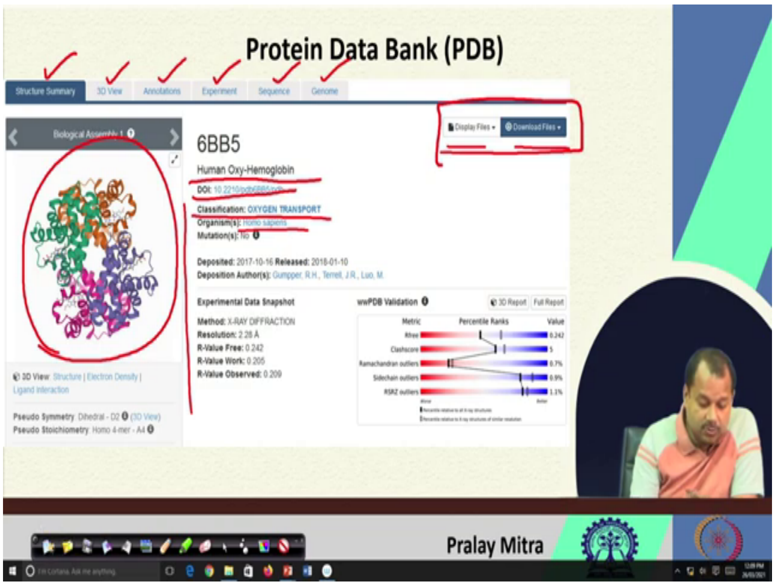
This resource is powered by the Protein Data Bank archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease.

As a member of the wwPDB, the RCSB PDB curates and annotates PDB data.

The RCSB PDB builds upon the data by creating tools and resources for research and education in molecular biology, structural biology, computational biology, and beyond.

Pralay Mitra

https://www.rcsb.org
RCSB PDB
Deposit
Search
Visualize
Analyze
Download
Learn
More
Documentation
PDB
PROTEIN DATA BANK
Macromolecular Structures Enabling Breakthroughs in Research and Education
Enter search term(s)
12AS
Advanced Search | Browse Annotations
Help
Celebrating 50 Years of Protein Data Bank
Welcome
Deposit
Search
Visualize
Analyze
Download
COVID-19 CORONAVIRUS Resources
PDB50: A special symposium May 4-5, 2021 Submit Abstract by March 22
March Molecule of the Month



Protein Data Bank (PDB)

Structure Summary | 3D View | Annotations | Experiment | Sequence | Genome

Display Files ▾ | Download Files ▾

Biological Assembly 1

6BB5

Human Oxy-Hemoglobin

DOI: 10.2210/pdb6BB5/pdb

Classification: OXYGEN TRANSPORT

Organism(s): Homo sapiens

Mutation(s): No

Deposited: 2017-10-16 Released: 2018-01-10

Deposition Author(s): Gumpper, R.H., Terrell, J.R., Luo, M.

3D View: Structure | Electron Density | Ligand Interaction

Pseudo Symmetry: Dihedral - D2 (3D View)

Pseudo Stoichiometry: Homo 4-mer - A4

Experimental Data Snapshot

Method: X-RAY DIFFRACTION

Resolution: 2.28 Å

R-Value Free: 0.242

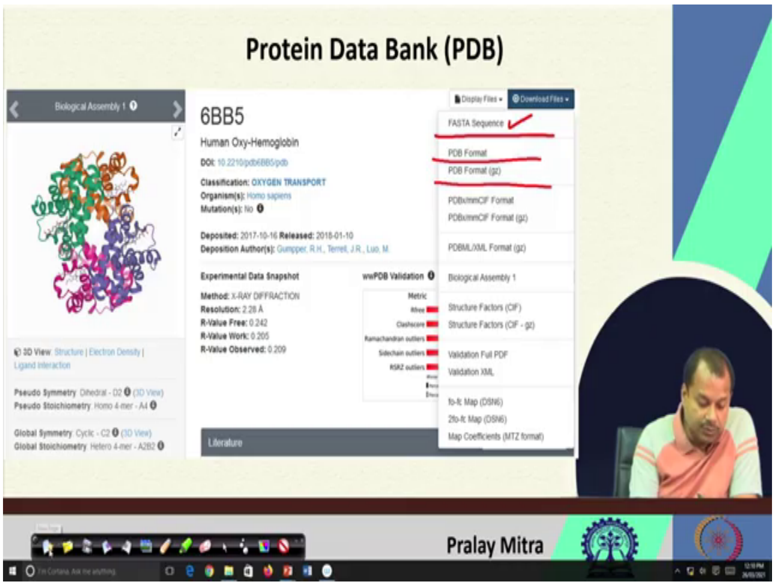
R-Value Work: 0.205

R-Value Observed: 0.209

wwPDB Validation | 3D Report | Full Report

Metric	Percentile Ranks	Value
Rfree		0.242
Clashscore		5
Ramachandran outliers		0.7%
Sidechain outliers		0.9%
RSRZ outliers		1.1%

Worse
Better
Percentile relative to all X-ray structures
Percentile relative to X-ray structures of similar resolution
Pralay Mitra



Protein Data Bank (PDB)

Display Files ▾ | Download Files ▾

Biological Assembly 1

6BB5

Human Oxy-Hemoglobin

DOI: 10.2210/pdb6BB5/pdb

Classification: OXYGEN TRANSPORT

Organism(s): Homo sapiens

Mutation(s): No

Deposited: 2017-10-16 Released: 2018-01-10

Deposition Author(s): Gumpper, R.H., Terrell, J.R., Luo, M.

3D View: Structure | Electron Density | Ligand Interaction

Pseudo Symmetry: Dihedral - D2 (3D View)

Pseudo Stoichiometry: Homo 4-mer - A4

Global Symmetry: Cyclic - C2 (3D View)

Global Stoichiometry: Hetero 4-mer - A2B2

Experimental Data Snapshot

Method: X-RAY DIFFRACTION

Resolution: 2.28 Å

R-Value Free: 0.242

R-Value Work: 0.205

R-Value Observed: 0.209

wwPDB Validation

Metric: Rfree, Clashscore, Ramachandran outliers, Sidechain outliers, RSRZ outliers

Download Files menu:

- FASTA Sequence ✓
- PDB Format
- PDB Format (gz)
- PDBx/mmCIF Format
- PDBx/mmCIF Format (gz)
- PDBML/XML Format (gz)
- Biological Assembly 1
- Structure Factors (CIF)
- Structure Factors (CIF - gz)
- Validation Full PDF
- Validation XML
- fo-fc Map (DSN6)
- 2fo-fc Map (DSN6)
- Map Coefficients (MTZ format)

Literature

This is the homepage of the protein data bank, and as you can identify that its link is www.rcsb.org or Google, if you just type protein data bank, you will get this one. On the homepage, you will get the March molecule of the month. People used to experiment they get the structure, they submit or deposit that structure in the protein data bank and whose structure looks better or say featured structure; those are represented in the front page.

This line is important here. We shall see as of now, how much data is deposited in the protein data bank. Here you will see enter search terms the PDB ID that is the unique ID corresponding to one protein molecule. If you say, experimentally get some structure and deposit that one in protein data bank following their norms. Then after the submission, you will also get one ID. IDs are exactly four alphanumeric characters.

As of now, the first one is digit after that one of three alphanumeric characters may be a digit or maybe a character from the English alphabets. And, it is not case sensitive means it can be smaller case or uppercase. I can enter the ID as 12AS and I can hit the search then if there is any protein molecule with this PDB ID will appear here. This is the front page of that entry.

Now, I searched for 6BB5, because I already demonstrated to you the structure of human Oxy-Hemoglobin. Also, I mentioned that the PDB ID is 6BB5. If you search for that one, then you will get in the homepage, the information and on the top side, not actually on the top, but on the upper side, you will have different tabs like structure summary, 3D view, annotations, experiment, sequence, genome all those information you will get.

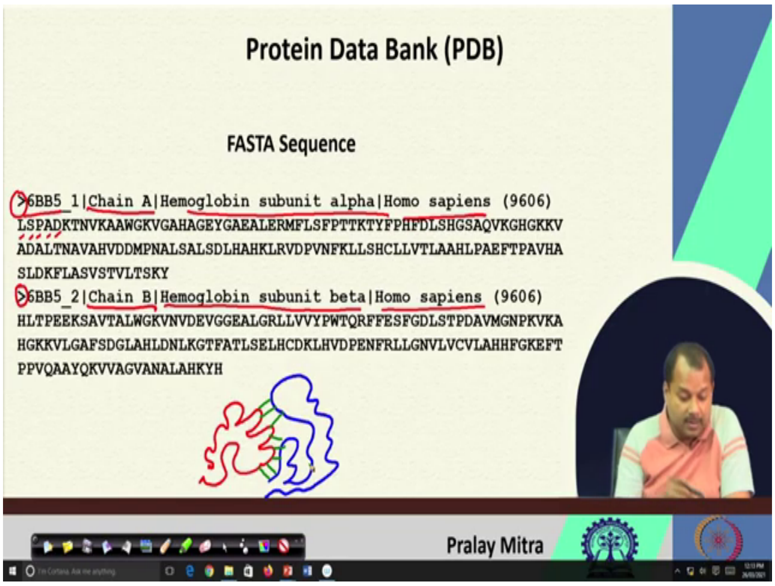
Here, I am not going to discuss each and everything you can explore by yourself. But here in the structure summary is what you will get. If this submission is corresponding to any publication that DOI information of that publication you will get also and then classification, this organism from where it is extracted. This information is also available inside the protein data bank file that we call a PDB file.

This is the structure, but you remember this 6BB5 consists of two chains, two chains mean two connected components - two molecules shortly it will be clearer to you. Here it shows four they are colored by their chain - color one is green, another is purple, another is pink, another is this golden. There are two chains, but using symmetry operation on that the biologically relevant or functional form that has been identified is the tetrameric form.

On the right side, you will locate two buttons - one is for display file, another is for download file. What are those files? Let us see those. If you click on display for a download file, then you will see that FASTA sequence you are familiar with this FASTA sequence. It will give you the sequence then PDB format so that directly you can download or in the zip format if you are suffering from the bandwidth then you can download the zip format and then you can unzip that one.

The same thing is also there in the display file, the display file indicates in your browser one tab will be opened and it will be displayed there and the download file means it will allow you to download in your local system. Then we can process it further. Let us assume we downloaded or we displayed the FASTA format file and let us see how will it look in the case of PDB.

(Refer Slide Time: 21:44)



Protein Data Bank (PDB)

FASTA Sequence

```
>6BB5_1|Chain A|Hemoglobin subunit alpha|Homo sapiens (9606)
LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKV
ADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHA
SLDKFLASVSTVLTSKY
>6BB5_2|Chain B|Hemoglobin subunit beta|Homo sapiens (9606)
HLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKA
HGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFT
PPVQAAYQKVVAGVANALAHKYH
```

This is the FASTA sequence. Regarding the sequence, you are familiar with. Now, I am giving you more details. In this case, as I mentioned 6BB5 contains two chains, designated as chain A and chain B. Also, I mentioned in the FASTA format, if any line starts with greater than symbol (>) that means it is a comment part of the amino acid sequence or not part of the protein sequence.

These two lines are comments in that way. Now, first what I am writing is PDB ID underscore 1 underscore 2 indicating there are two chains. What is the name or ID of that chain? This regarding chain, I will discuss later. This is chain B and chain A. What is this Hemoglobin subunit alpha? Homo sapiens are the organism. Hemoglobin subunit beta and Homo sapiens is the organism that is mentioned.

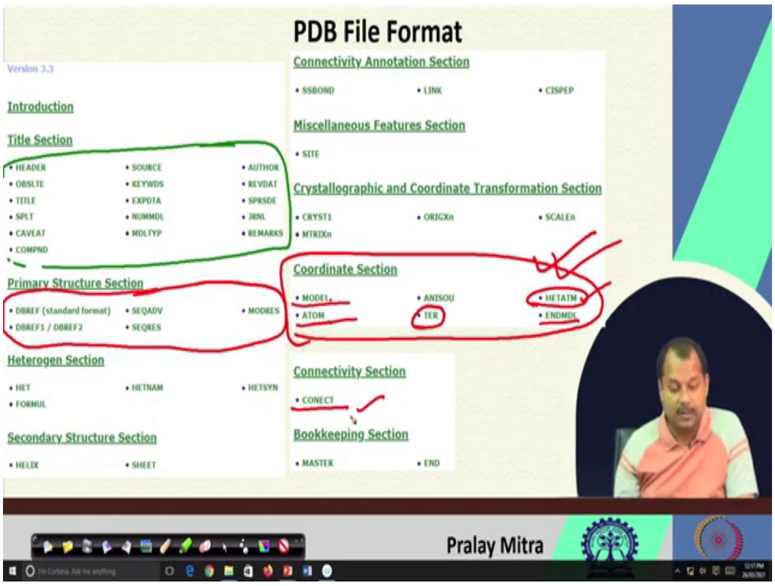
After that, I am getting the sequence and these three lines consist of the amino acid in single-letter characters. L indicates Leucine, S indicates Serine, P Proline, A Alanine, D Aspartic acid like that way you are having the list of amino acids which make the protein molecule. Now, when I say Chain A, it means one connected component of the protein molecule.

When I have one sequence, the sequence is like this, like this, like this. I mean the sequence when folds they may take a structure something like this. And in this case, you see this is N-terminus, this is C-terminus. I am assuming that one. A protein molecule if it is this one, then it is only one single molecule and if you remember in our introductory class, I mentioned that this only one protein molecule cannot have any function. It is only one and to have a function it must interact with somebody else. Let us assume there is another one and it is also placed in proximity with each other.

With this situation, we are having one region with the green color that indicates the region where they will interact or they will participate in the interaction. And because of that interaction, the function will occur. Now, you look at the red color that consists of one connected component. And blue color consists of another connected component.

I am calling this red color one chain and the blue color. I am calling as another chain. Hence, there are two chains. That’s why you are having two different chain IDs - chain A and chain B and those are something like this. So, they are separate.

(Refer Slide Time: 25:21)



PDB File Format

Version 3.3

Introduction

Title Section

- HEADER
- OBSLTE
- TITLE
- SPLT
- CAVEAT
- COMPND
- SOURCE
- KEYWDS
- EXPDTA
- NUMMDL
- MDLTYP
- AUTHOR
- REVDAT
- SPRSDE
- JRNL
- REMARKS

Primary Structure Section

- DBREF (standard format)
- DBREF1 / DBREF2
- SEQADV
- SEQRES
- MODRES

Heterogen Section

- HET
- FORMUL
- HETNAM
- HETSYN

Secondary Structure Section

- HELIX
- SHEET

Connectivity Annotation Section

- SSBOND
- LINK
- CISPEP

Miscellaneous Features Section

- SITE

Crystallographic and Coordinate Transformation Section

- CRYST1
- MTRIXn
- ORIGXn
- SCALEn

Coordinate Section

- MODEL
- ATOM
- ANISOU
- TER
- HETATM
- ENDMDL

Connectivity Section

- CONECT

Bookkeeping Section

- MASTER
- END

This is the PDB file format that you can go through on the RCSB website and you can look by yourself. The different components are listed here. The title section will contain a lot of information like header, source, author, keywords, title, etcetera. Primary structure section, I also mentioned that is nothing but the protein sequence that consists of sequence SEQRES, MODRES all those information.

Currently, all are not much important for us. Only which are important for us is this coordinate section where we have the atom information between two chains. There is a terminator with the keyword TER. If I am using NMR, or if I have several copies or several instances of one protein molecule that I got experimentally then different model information will also be present as a different model. Also, we find some hetero atom information.

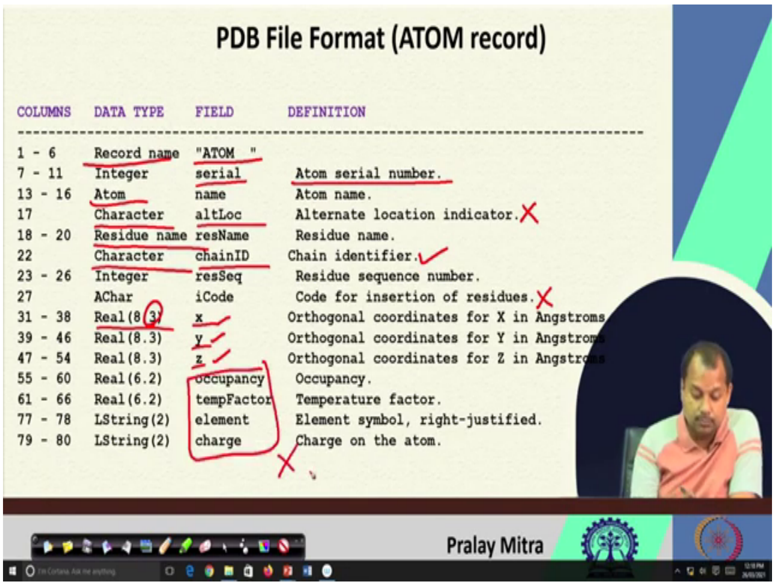
As hetero atom, all the atoms belong to some molecule and are not part of the essential amino acids, like water molecule or other small molecule or moiety etcetera, will be represented. This CONECT is a keyword that indicates connectivity information. As you understand that for amino acids, the connections are known. Connection means that covalent bonds. This bond information is known, but there may be a variety of hetero atoms like the atoms from the water molecule, and some other small molecules.

It is very difficult to remember that what will be the connectivity information I mean the covalent bond for all those molecules. So, the best thing is that no need to remember, you are representing those as an atom. Now, along with that atom, you have the atom ID. Use that atom ID to mention which atom is connected with what using this CONECT information.

However, if it is standard amino acids from the list of 20, you know their structure is the same. At the center there is carbon then on one side hydrogen, another side chain, and other amino and carboxyl and also amino carboxyl and the sidechain those are limited information that I can code and I can remember for my purpose but not for the entire protein molecule.

That is why for the standard amino acids, this CONECT information is not required and it is not represented. However, all those molecules, which are not part of this amino acid are denoted as hetero atoms and for them, this CONECT information is required.

(Refer Slide Time: 28:44)



PDB File Format (ATOM record)

COLUMNS	DATA TYPE	FIELD	DEFINITION
1 - 6	Record name	"ATOM "	
7 - 11	Integer	serial	Atom serial number.
13 - 16	Atom	name	Atom name.
17	Character	altLoc	Alternate location indicator. ✗
18 - 20	Residue name	resName	Residue name.
22	Character	chainID	Chain identifier. ✓
23 - 26	Integer	resSeq	Residue sequence number.
27	AChar	iCode	Code for insertion of residues. ✗
31 - 38	Real(8.3)	x ✓	Orthogonal coordinates for X in Angstroms
39 - 46	Real(8.3)	y ✓	Orthogonal coordinates for Y in Angstroms
47 - 54	Real(8.3)	z ✓	Orthogonal coordinates for Z in Angstroms
55 - 60	Real(6.2)	occupancy	Occupancy.
61 - 66	Real(6.2)	tempFactor	Temperature factor.
77 - 78	LString(2)	element	Element symbol, right-justified.
79 - 80	LString(2)	charge	Charge on the atom.

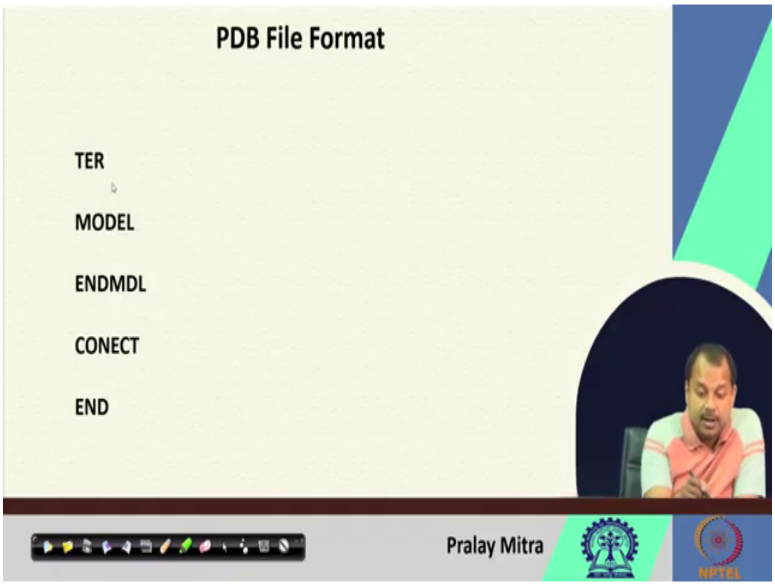
✗

This is the PDB file format. The first column indicates the column information. Here, the data type is the record name, sorry, the data type is the record name. The only atom will be represented here then the serial number that is the atom serial number you remember the first column in our example, then atom.

Atom name will be there then the character that is for altering location. This part we shall discuss later not right now. Then, residue names like ALA, ASP will be there. Again after this one character that is the chain ID. For 6BB5 we have two chains - chain A and chain B. That chain identifier may be a single character or it can be an integer. Anyway, whatever it may be it will be a single column of information. OK!

Then this AChar of this code for insertion of residue these we shall not discuss now. Later, after X, Y, Z coordinate, we shall discuss. Please remember that after the decimal place there are three points. That’s why three in total. There will be eight spaces for storing the coordinate information X, Y, and Z. These things we will not discuss now, we will discuss sometime later. OK!

(Refer Slide Time: 30:18)



PDB File Format

TER

MODEL

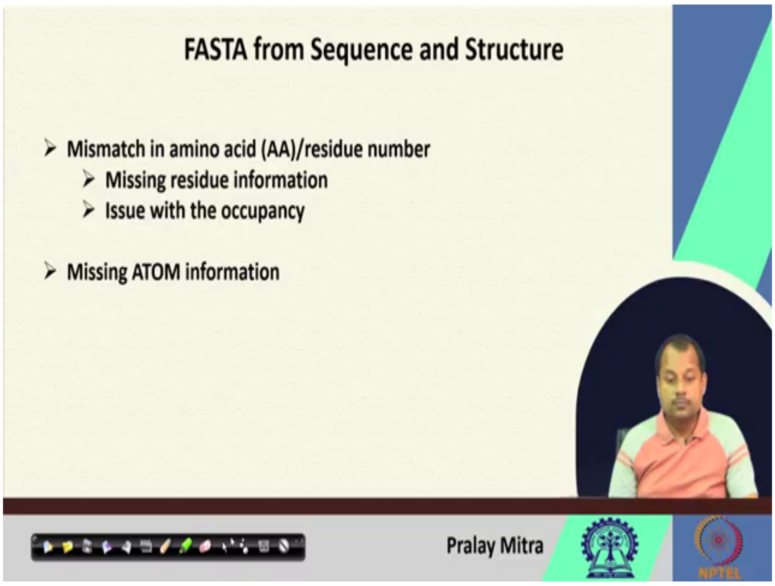
ENDMDL

CONECT

END

Now, in the PDB file format apart from atom few things are also there TER, MODEL, ENDMOL, CONECT, and END. When we will do the parsing, we will discuss this one.

(Refer Slide Time: 30:36)



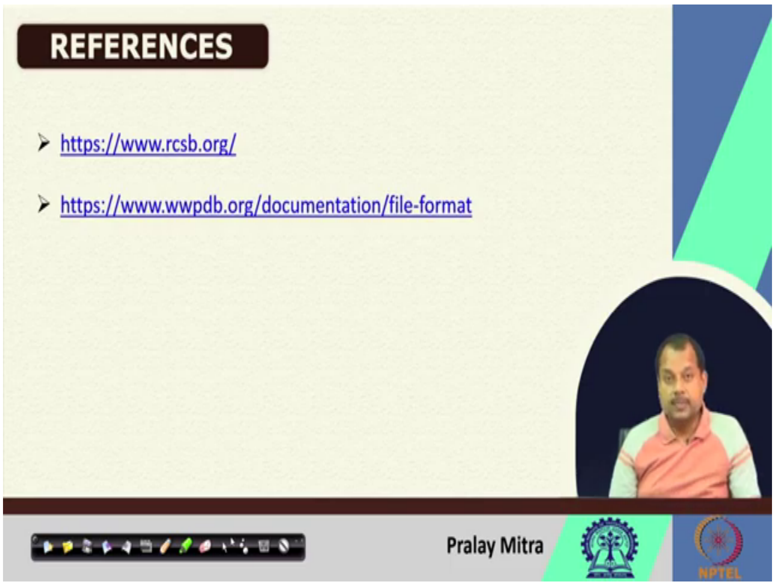
FASTA from Sequence and Structure

- Mismatch in amino acid (AA)/residue number
  - Missing residue information
  - Issue with the occupancy
- Missing ATOM information

One thing you should remember is the FASTA sequence that you will get from the PDB for a protein molecule. I mean that will say you are looking at the PDB then there is a space for SEQRES from where the sequence information is given and also when you are scanning the structure then from that structure you can extract that sequence information.

Now, this two-sequence information may not match. You should remember about this one. The primary reason is that when there is an experiment during the modeling process some information may be lost. That lost information or missed information from the structure is not available because of that one if you read the sequence from the protein structure that may not match with the protein sequence corresponding to a protein molecule. You should be careful about this when you are working with the protein structure. These things we shall discuss in detail during the algorithm development also.

(Refer Slide Time: 31:53)



REFERENCES

- https://www.rcsb.org/
- https://www.wwpdb.org/documentation/file-format

This is the link for the RCSB that you noted and if you do a Google search, you will get this information. That’s it. Thank you.